Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
τομεας σηματων, ελεγχου και ρομποτικης

Enhancing Contrastive Language-Vision Pre-training with Generative Dialogue

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Τσαπραζλή Ευθύμιου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Γεώργιος Σμυρνής
Διδακτορικός Ερευνητής University of Texas at Austin.

ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΛΟΓΟΥ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ
Αθήνα, Ιούλιος 2024

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Enhancing Contrastive Language-Vision Pre-training with Generative Dialogue

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Τσαπραζλή Ευθύμιου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Γεώργιος Σμυρνής
Διδακτορικός Ερευνητής University of Texas at Austin.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Ιουλίου, 2024.

.....	.....	.....
Πέτρος Μαραγκός	Αθανάσιος Ροντογιάννης	Ιωάννης Κορδώνης
Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

.....
ΕΥΘΥΜΙΟΣ ΤΣΑΠΡΑΖΛΗΣ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα πολυτροπικά μοντέλα εικόνας-κειμένου έχουν γίνει θεμελιώδη στη μηχανική μάθηση, οδηγώντας στην ανάπτυξη αρκετών state-of-the-art αρχιτεκτονικών, όπως τα CLIP, DALL-E και Stable Diffusion, μεταξύ άλλων. Αυτά τα βασικά μοντέλα μπορούν να εφαρμοστούν σε διάφορες εργασίες, συχνά διαφορετικές από τους αρχικούς στόχους εκπαίδευσής τους. Αυτή η ευελιξία προκύπτει από την ικανότητά τους να χρησιμοποιούν μια τροπικότητα για να εξάγουν γνώση για την άλλη, επιτρέποντάς τους να λειτουργούν χωρίς δεδομένα συγκεκριμένα για την εκάστοτε εργασία. Ωστόσο, αυτές οι αρχιτεκτονικές είναι πολύ κοστοβόρες στην εκπαίδευση τους, απαιτώντας συνήθως εκατομμύρια ζεύγη εικόνας-κειμένου. Ως εκ τούτου, η χρήση αυτών των εκτενών θεμελιωδών μοντέλων συχνά περιλαμβάνει την **λεπτομερή προσαρμογή** τους αντί της εκπαίδευσής τους από το μηδέν. Επιπλέον, αυτά τα μοντέλα μπορούν να χρησιμοποιηθούν απευθείας για εξαγωγή χαρακτηριστικών ή προβλέψεων, λειτουργώντας ως μια βάση πάνω στην οποία μπορούν να χτιστούν πιο σύνθετες αρχιτεκτονικές. Επιπλέον, τα πολυτροπικά παραγωγικά μοντέλα (multimodal generative models) που χρησιμοποιούνται ευρέως σήμερα, προσφέρουν εξαιρετικές δυνατότητες στη λεκτική περιγραφή των εικόνων. Αυτά τα μοντέλα μπορούν να παράγουν ουσιαστικές περιγραφές και υψηλής ποιότητας απαντήσεις σχετικά με οπτικές εισόδους που τους δίνονται. Πιστεύουμε ότι ο διάλογος μεταξύ ενός χρήστη και ενός γενετικού μοντέλου για μια εικόνα εισόδου μπορεί να προσφέρει μια επιπλέον προοπτική στο ζεύγος εικόνας-κειμένου.

Αρχικά, μελετάμε το πρόβλημα της εκπαίδευσης ενός τρίτου πύργου για μια νέα μορφή δεδομένων χρησιμποιώντας ως βάση ένα προεκπαιδευμένο μοντέλο CLIP. Αυτό το πρόσθετο στοιχείο μπορεί να χρησιμοποιηθεί για την ενσωμάτωση καινούριων τροπικοτήτων στο μοντέλο βάσης. Στην εργασία μας, που ονομάζεται **CLIP-3Modal**, εξετάζουμε τη χρήση ενός γεννετικού μοντέλου όπως το BLIP-2, το οποίο μας παρέχει έναν διάλογο με επίκεντρο την εικόνα. Αξιολογούμε το μοντέλο μας στο πλαίσιο της ανάκτησης εικόνας και κειμένου, και το συγκρίνουμε με το μοντέλο βάσης που χρησιμοποιήσαμε. Στη συνέχεια, εγκαταλείπουμε την προσέγγιση του τρίτου πύργου και επικεντρωνόμαστε στη λεπτομερή προσαρμογή του αρχικού CLIP σε κειμενικές εισόδους τύπου ερώτησης-απάντησης. Προτείνουμε το **DRAFT (Dual Representation Adaptive Fine-Tuning)**, μια μέθοδο βασισμένη στη μάθηση με αντιδιαστολή και την ευθυγράμμιση κατανομών, σχεδιασμένη να προσαρμόζει μοντέλα τύπου CLIP σε κειμενικές περιγραφές εκτός κατανομής, όπως ο διάλογος. Διεξάγαμε εκτεταμένα πειράματα και αφαιρετικές μελέτες για να αποδείξουμε τα πλεονεκτήματα και τα οφέλη της μεθόδου μας σε σχέση με το βασικό μοντέλο. Τα πειράματά μας επικεντρώθηκαν κυρίως σε απάντηση οπτικών ερωτήσεων, όπου η μέθοδός μας βελτίωσε σημαντικά την απόδοση του CLIP.

Λέξεις Κλειδιά — Μάθηση με Αντιδιαστολή, Αυτό-επιβλεπόμενη Μάθηση, Πολυτροπική Μάθηση, Παραγωγική Τεχνητή Νοημοσύνη, Βαθιά Μάθηση, Νευρωνικά Δίκτυα, Μηχανική Μάθηση

Abstract

Image-text models have become essential in machine learning, leading to the development of several state-of-the-art architectures, such as CLIP, DALL-E, and Stable Diffusion, among others. These foundational models can be applied to various tasks, often different from their initial training objectives. This versatility arises from their ability to use each modality to infer knowledge about the other, enabling them to function without specific task-related data. However, these architectures are also very expensive to train, typically requiring millions of image-text pairs. Therefore, the use of these extensive foundational models often involves **finetuning** them rather than training them from scratch. Additionally, these models can be used directly for inference, serving as the foundation upon which more complex pipelines can be built. Moreover, multimodal generative models have taken the world by storm, providing exceptional capabilities for image captioning. These models can generate meaningful descriptions and high-quality responses regarding visual concepts. We believe that a dialogue between a user and a generative model about a given image can offer an additional perspective on the image-text pair.

Initially, we study the problem of training a third tower for a new modality given a pre-trained CLIP model. This additional component can be used to incorporate other modalities into the model pipeline. In our framework, called **CLIP-3Modal**, we consider the use of a model such as BLIP-2, which provides us with a dialogue centered around the image. We evaluate our model in the setting of image and text retrieval, and compare it against the regular image and text based one. Then, we abandon the third tower approach and focus on fine-tuning the original CLIP to adapt to question-answer style textual inputs. We introduce **DRAFT (Dual Representation Adaptive Fine-Tuning)**, a method based on contrastive learning and distribution alignment, designed to adapt CLIP-like models to out-of-distribution textual descriptions, such as dialogue. We conducted extensive experimentation and ablation studies to demonstrate the advantages and benefits of our method over the baseline model. Our experiments primarily focused on Visual Question Answering tasks, where our method significantly improved CLIP's performance.

Keywords — Contrastive Learning, Self-supervised Learning, Multimodal Learning, Generative AI, Deep Learning, Neural Networks, Machine Learning

Ευχαριστίες

Με αφορμή αυτή τη διπλωματική θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Πέτρο Μαραγκό για την εμπιστοσύνη που μου έδειξε και τις ευκαιρίες που μου έδωσε κατά την εκπόνηση αυτής της εργασίας. Μέσω αυτής της διπλωματικής εργασίας κατάφερα να συνεργαστώ με ένα πρωτοπόρο πανεπιστήμιο των ΗΠΑ, το University of Texas at Austin, και να βάλω γερά θεμέλια για την μετέπειτα ερευνητική μου καριέρα και τον ευχαριστώ για αυτό.

Επίσης, θα ήθελα να ευχαριστήσω και τον Γιώργο Σμύρνη ο οποίος συνεπίβλεψε αυτή την εργασία. Παρά τις σημαντικές του υποχρεώσεις , ο Γιώργος αφιέρωσε πολύ από το χρόνο του δίνοντας χρήσιμες συμβουλές και στάθηκε ως μέντορας μου στα πρώτα μου ερευνητικά βήματα.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, την αδερφή μου και την οικογένεια μου για τη στήριξη και τη κατανόηση τους καθόλη τη διάρκεια των σπουδών μου. Επίσης θα ήθελα ευχαριστήσω τους φίλους και κοντινούς μου ανθρώπους για όλες τις στιγμές που ήταν εκεί για εμένα και έκαναν το ταξίδι των σπουδών μου ακόμη πιο όμορφο.

Τσαπραζλής Ευθύμιος
Ιούλιος 2024

Contents

List of Figures

List of Tables

Color Code

Theorem 0.0.1: title

Theorem

Definition 0.0.2: title

Definition

Lemma 0.0.3: title

Lemma

Example 0.0.4: title

Example

Εκτεταμένη Περίληψη στα Ελληνικά

Αυτο-επιβλεόμενη Μάθηση

Η Αυτο-επιβλεπόμενη Μάθηση στοχεύει στην παροχή πλούσιων αναπαραστάσεων και χαρακτηριστικών βαθιάς εκμάθησης αποφεύγοντας τη χρήση επισημειωμένων δεδομένων όπως στην Επιβλεπόμενη Μάθηση. Η χρήση των επισημειωμένων δεδομένων είναι αυτή που καθορίζει τη δυσκολία της πρακτικής εφαρμογής της Βαθιάς Μάθησης σήμερα. Αυτές οι μέθοδοι σημείωσαν ραγδαία πρόοδο τα τελευταία χρόνια, καταφέρνοντας έτσι να συγκριθούν, ως προς την απόδοση, με τις πλήρως επιβλεπόμενες μεθόδους προ-εκπαίδευσης. Αρκετές φορές μάλιστα, τα αποτελέσματά τους ξεπερνούν την απόδοσή των επιβλεπόμενων μεθόδων [42, 24].

Η έρευνα για την αυτο-επίβλεψη ξεκίνησε από το προβληματισμό που υπήρχε ως προς το κόστος της χειροκίνητης επισήμανσης των συνόλων δεδομένων. Με τις αυτο-επιβλεπόμενες μεθόδους μπορούμε να σχεδιάσουμε προεπιλεγμένες εργασίες (pretext tasks) για να αποκτήσουμε ετικέτες προς χρήση κατά τη διάρκεια της εποπτείας στην εκπαίδευση ενός βαθιού νευρωνικού δικτύου. Συγκεκριμένα, μπορούμε να εκπαιδεύσουμε βαθιές αναπαραστάσεις χαρακτηριστικών χωρίς τη χρήση ετικετών, έτσι ώστε να εκπαιδεύσουμε ένα βαθύ νευρωνικό δίκτυο στο να λύσει μια μεταγενέστερη εργασία (downstream task) χρησιμοποιώντας συγκριτικά λιγότερα επισημειωμένα δεδομένα σε σχέση με την επιβλεπόμενη μάθηση.

Όπως αναφέραμε παραπάνω, με την αυτο-επιβλεπόμενη μάθηση μαθαίνουμε εργασίες που απαιτούν την πρόβλεψη ενός μέρους της εισόδου ή την εξαγωγή μιας ετικέτας από άλλο μέρος της εισόδου. Μπορούμε να δούμε ότι η αυτο-επίβλεψη έρχεται σε αντίθεση με την επιβλεπόμενη μάθηση. Με την επιβλεπόμενη μάθηση προσπαθούμε να προβλέψουμε ενα στόχο-έξοδο που μας δίνεται. Με άλλα λόγια, εκπαιδεύουμε ένα μοντέλο να εκτιμά την πυκνότητα των εισερχόμενων δεδομένων ή να μαθαίνει να παράγει αυτά. Κάθε μέθοδος αυτο-επίβλεψης διαφέρει στη στρατηγική εξαγωγής ετικετών. Αυτή η επιλογή της προεπιλεγμένης εργασίας καθορίζει πόσο αποτελεσματικές θα είναι και οι προκύπτουσες αναπαραστάσεις σε διάφορες μεταγενέστερες εργασίες. Ορισμένα αποτελέσματα υποδηλώνουν ότι πέρα από την προεπιλεγμένη εργασία, η ποιότητα της αναπαράστασης είναι επίσης μια λογαριθμική συνάρτηση του μεγέθους των μη επισημασμένων δεδομένων. Εάν αυτή η υπόθεση ισχύει, τότε μπορούμε να επιτύχουμε καλύτερη απόδοση, απλά, με τη χρήση μεγαλύτερων συνόλων προ-εκπαίδευσης. Κάτι τέτοιο θα μπορούσε να πραγματοποιηθεί από τις βελτιώσεις στη συλλογή δεδομένων και την αναβάθμιση της υπολογιστικής ισχύος [29].

Ορισμοί

Αρχικά πρέπει να ορίσουμε το πρόβλημα που προσπαθεί να λύσει η αυτο-επιβλεπόμενη μάθηση και να το συγκρίνουμε με τις ήδη γνωστές μεθόδους μάθησης όπως η επιβλεπόμενη και η μη-επιβλεπόμενη μάθηση [24].

Definition: Επιβλεπόμενη Μάθηση

Έστω ένα επισημασμένο σύνολο δεδομένων $D_i = \{x_i^{(t)}, y_i^{(t)}\}_{i=1}^N$ όπου N είναι ο αριθμός των δειγμάτων. Υποθέτουμε ότι έχουμε ένα πρόβλημα, το οποίο καταφέρνουμε να λύσουμε στο D_t δημιουργώντας ένα προβλεπτικό μοντέλο, $\hat{y} = f(x)$ που εκτιμά την ετικέτα y. Για εφαρμογές βαθιάς μάθησης όπως η δική μας, το προβλεπτικό μοντέλο αποτελείται από μια λειτουργία εξαγωγής αναπαραστάσεων h_θ και μια λειτουργία ταξινόμησης ή παλινδρόμησης g_φ:

$$f(x) = g_\varphi(h_\theta(x)) \tag{0.0.1}$$

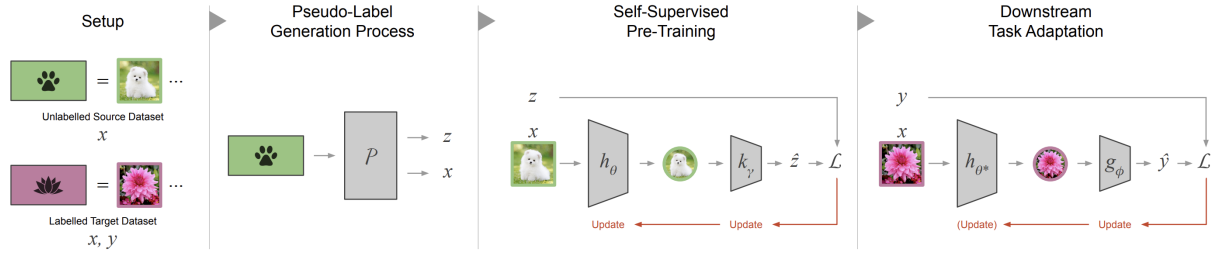


Figure 0.0.1: **Αυτο-επιβλεπόμενη Μάθηση.** Η ροή εργασίας της αυτο-εποπτευόμενης μάθησης ξεκινά με ένα μη επισημασμένο αρχικό σύνολο δεδομένων και ένα επισημασμένο σύνολο δεδομένων στόχου. Στη συνέχεια, μέσω της προεπιλεγμένης εργασίας δημιουργούνται ψευδο-ετικέτες προγραμματιστικά από το μη επισημασμένο σύνολο. Οι προκύπτουσες είσοδοι, x και οι ψευδο-ετικέτες z χρησιμοποιούνται για την προ-εκπαίδευση του μοντέλου $k_\gamma(h_\theta(\cdot))$ – που αποτελείται από τον εξαγωγέα χαρακτηριστικών h_θ και τις εξόδους k_γ – για την επίλυση της προεπιλεγμένης εργασίας. Μετά την ολοκλήρωση της προ-εκπαίδευσης, τα βάρη θ^* που έμαθε ο εξαγωγέας χαρακτηριστικών h_{θ^*} μεταφέρονται και χρησιμοποιούνται μαζί με ένα νέο μοντέλο εξόδου g_ϕ για την επίλυση της μεταγενέστερης εργασίας στόχου.[24]

Έτσι, εκπαιδεύουμε το μοντέλο μας να ελαχιστοποιεί μια συνάρτηση απώλειας $\mathcal{L}$:

$$\underset{\theta,\varphi}{arg\,min} \sum_{(x_i^{(t)},y_i^{(t)})\in D_i} \mathcal{L}(g_\varphi(h_\theta(x_i^{(t)}), y_i^{(t)})) \tag{0.0.2}$$

Το βασικό πρόβλημα αυτής της προσέγγισης είναι ότι το h_θ μπορεί να έχει εκατοντάδες εκατομμύρια παραμέτρους προς εκπαίδευση, το οποίο απαιτεί εκατομμύρια επισημασμένα δείγματα στο σύνολο δεδομένων D_t.

Definition: Μη Επιβλεπόμενη Μάθηση

Σε αντίθεση με την επιβλεπόμενη μάθηση, η μη επιβλεπόμενη μάθηση προσπαθεί να μάθει από μη επισημασμένα δεδομένα εκτιμώντας μια συνάρτηση πυκνότητας πιθανότητας πάνω στα εισερχόμενα δεδομένα ή δημιουργώντας ένα γενετικό μοντέλο. Αυτές οι τεχνικές ποικίλουν από τα Μείγματα Γκαουσιανών (Gaussian Mixture Models) [33] έως τα Γενετικά Ανταγωνιστικά Δίκτυα (GANs) και τους Αυτοκωδικοποιητές Μεταβολών (Variational Autoencoders) [27]. Άλλες προσεγγίσεις επικεντρώνονται στην εκμάθηση λανθανουσών αναπαραστάσεων όπως η συσταδοποίηση (clustering) και οι αυτοκωδικοποιητές (autoencoders) [27]. Ακολουθώντας τη προηγούμενη σημειογραφία μας, υποθέτουμε ότι έχουμε έναν αυτοκωδικοποιητή, ο στόχος μας είναι να ελαχιστοποιήσουμε την απώλεια ανακατασκευής:

$$\underset{\theta,\varphi}{arg\,min} \sum_{(x_i^{(t)})\in D_i} \mathcal{L}(g_\varphi(h_\theta(x_i^{(t)}), x_i^{(t)})) \tag{0.0.3}$$

όπου ο h_θ εξάγει μια αναπαράσταση χαρακτηριστικών και ο g_φ ανακατασκευάζει την είσοδο x δεδομένης της αναπαράστασης $h_\theta(x)$.

Τώρα, μπορούμε να δούμε την αυτο-επιβλεπόμενη μάθηση ως μια ειδική περίπτωση μη επιβλεπόμενης μάθησης όπου δεν χρειάζεται να ανακατασκευάσουμε την είσοδο ή να εκτιμήσουμε μια πυκνότητα πιθανότητας. Αντίθετα, δημιουργούμε μια προεπιλεγμένη εργασία $\mathcal{P}$ που εκμεταλλεύεται τη γνώση σχετικά με τα δεδομένα.

Definition: Αυτο-επιβλεπόμενη Μάθηση

Έστω $D_s = \{x_i^{(s)}\}_{i=1}^M$ ένα μη επισημασμένο σύνολο δεδομένων *πηγής*, όπου $M \gg N$ (το μη επισημασμένο σύνολο δεδομένων είναι σημαντικά μεγαλύτερο από το επισημασμένο), ο στόχος μας είναι να χρησιμοποιήσουμε το D_t και το D_s μαζί για να μάθουμε ένα προβλεπτικό μοντέλο $f(x) = g_\varphi(h_\theta(x))$.

Δημιουργούμε μια διαδικασία $\mathcal{P}$ ως την προεπιλεγμένη εργασία μας έτσι ώστε να παράγουμε τις ψευδο-ετικέτες και ένα στόχο για να καθοδηγήσουμε την εκμάθηση. Δεδομένου του συνόλου δεδομένων πηγής D_s παράγουμε ψευδο-ετικέτες z και σημεία δεδομένων $\mathcal{P}(D_s) = \{x_i, z_i\}_{i=1}^M$.
Έστω $\bar{D}_s = \mathcal{P}(D_s) = \{x_i, z_i\}_{i=1}^M$ το νέο ψευδο-επισημασμένο σύνολο δεδομένων και k_γ το προεπιλεγμένο μοντέλο. Προσπαθούμε να βελτιστοποιήσουμε τον αυτο-επιβλεπόμενο στόχο στο νέο σύνολο δεδομένων $\bar{D}_s$:

$$\theta^* = arg\min_{\theta,\gamma} \sum_{(x_i,z_i)\in\mathcal{P}(D_s)} \mathcal{L}(k_\gamma(h_\theta(x_i), z_i)) \tag{0.0.4}$$

Τέλος, αφαιρούμε τη λειτουργία προεπιλογής k_γ και κρατάμε μόνο τη βελτιστοποιημένη λειτουργία αναπαράστασης h_{θ^*}. Χρησιμοποιούμε το h_{θ^*} ως μερική λύση για να λύσουμε το πρόβλημα-στόχο χρησιμοποιώντας το μοντέλο $g_\varphi(h_{\theta^*}(\cdot))$.

Επειδή οι παράμετροι θ^* είναι καλά προσαρμοσμένοι, πρέπει να μάθουμε μόνο μια μικρή ποσότητα παραμέτρων για να λύσουμε την μεταγενέστερη εργασία. Υπάρχουν δύο κοινές μέθοδοι για να λύσουμε το παραπάνω πρόβλημα χρησιμοποιώντας το θ^*, η ***λεπτομερής προσαρμογή (fine-tuning)*** και η ***γραμμική ανάγνωση (linear-readout)***.

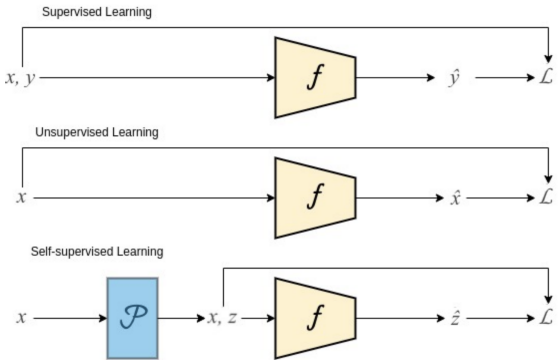


Figure 0.0.2: **Μια επεξηγηματική σύγκριση μεταξύ της Αυτο-επιβλεπόμενης Μάθησης, της Επιβλεπόμενης Μάθησης και της Μη Επιβλεπόμενης Μάθησης.**

Προεπιλεγμένες Εργασίες (Pretext Tasks)

Όπως αναφέραμε, η επιλογή της προεπιλεγμένης εργασίας είναι κρίσιμη για την ποιότητα της αναπαράστασης. Οι προεπιλεγμένες εργασίες μπορούν να χωριστούν σε τέσσερις κύριες κατηγορίες: πρόβλεψη μετασχηματισμού, πρόβλεψη με μάσκα, διάκριση παραδείγματος και συσταδοποίηση.

Πρόβλεψη Μετασχηματισμού Αυτές οι μέθοδοι προσπαθούν να προβλέψουν έναν γνωστό μετασχηματισμό που εφαρμόζεται στην είσοδο. Παραδείγματα περιλαμβάνουν την πρόβλεψη περιστροφών, την ανακατασκευή εικόνων ή την πρόβλεψη των χρωμάτων μιας gray-scale εικόνας.

Πρόβλεψη με μάσκα Εδώ, η εργασία είναι το μοντέλο να προβλέψει τις κρυμμένες (ή μάσκαρισμένες) τιμές σε μια είσοδο. Αυτή η μέθοδος είναι δημοφιλής στην επεξεργασία φυσικής γλώσσας (π.χ. BERT) και στις εφαρμογές όρασης.

Διάκριση Παραδείγματος Η διάκριση παραδείγματος περιλαμβάνει την πρόβλεψη αν ένα ζεύγος εισόδων ανήκει στην ίδια κατηγορία αντί να προβλέπει την ακριβή κατηγορία που ανήκουν τα δύο δείγματα εισόδου.

Συσταδοποίηση Οι μέθοδοι συσταδοποίησης παραδοσιακά επικεντρώνονται στη διαίρεση των εισόδων σε ομάδες με υψηλή ομοιότητα εντός της ομάδας και χαμηλή μεταξύ των ομάδων (π.χ. K-Means). Ωστόσο, στις

εφαρμογές αυτο-επιβλεπόμενης μάθησης, στοχεύουν στην εκμάθηση ενός καλού εξαγωγέα χαρακτηριστικών αντί για αναθέσεις συσταδοποίησης.

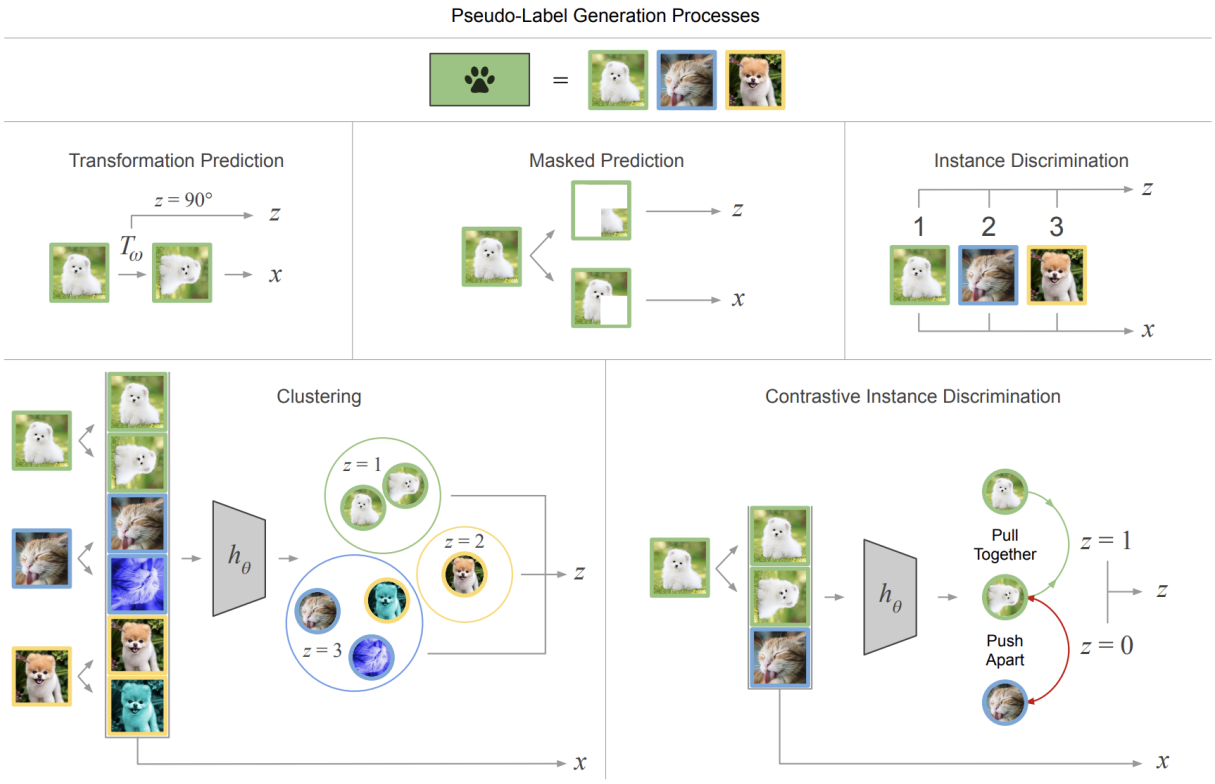


Figure 0.0.3: **Διαδικασία Παραγωγής Ψευδο-ετικετών.** Παραδείγματα παραγωγής ψευδο-ετικετών στις τέσσερις οικογένειες προεπιλεγμένων εργασιών: πρόβλεψη μετασχηματισμού, πρόβλεψη με μάσκα, διάκριση παραδείγματος και συσταδοποίηση. Περιλαμβάνεται μια πρόσθετη απεικόνιση της δημοφιλούς περίπτωσης της διάκρισης παραδείγματος χρησιμοποιώντας μάθηση με αντιδιαστολή. Τα τετράγωνα αναπαριστούν τις εισόδους x ενώ οι κύκλοι απεικονίζουν τους διανυσματικούς χώρους των εισόδων αυτών, $h_\theta(x)$.[24]

Θεωρητικό Υπόβαθρο

Μάθηση με Αντιδιαστολή

Η μάθηση με αντιδιαστολή, ως ένα υποσύνολο της αυτο-επιβλεπόμενης μάθησης, έχει αποτελέσει μια σημαντική τεχνική που χρησιμοποιείται σε πρόσφατα state-of-the-art μοντέλα μηχανικής μάθησης, όπως το CLIP [64]. Βασίζεται στη χρήση πολλαπλών σημασιολογικά σχετιζόμενων δειγμάτων των οποίων οι αναπαραστάσεις θα πρέπει να γίνονται όσο το δυνατόν πιο παρόμοιες. Αυτά ονομάζονται θετικά δείγματα, και συχνά προέρχονται από το ίδιο δείγμα, αλλά έχουν υποστεί διαφορετικούς μετασχηματισμούς. Εκτός από τις όψεις του ίδιου δείγματος, οι μέθοδοι με αντιδιαστολή χρησιμοποιούν επίσης αρνητικά δείγματα, τα οποία θεωρούνται σημασιολογικά διαφορετικά από το αρχικό. Αυτά συχνά λαμβάνονται από τυχαία δείγματα μέσα στην ίδια παρτίδα (batch) [13] ή από ένα ιστορικό δειγμάτων του μοντέλου [36]. Χρησιμοποιώντας αυτά τα αρνητικά δείγματα, το μοντέλο μπορεί να βελτιώσει τις αναπαραστάσεις του, μαθαίνοντας να διακρίνει μεταξύ δειγμάτων που είναι διαφορετικά μεταξύ τους. Αυτή η μορφή εκμάθησης αναπαραστάσεων έχει γίνει εξαιρετικά δημοφιλής σε μοντέλα εικόνας-κειμένου, αποτελώντας ένα αναπόσπαστο στοιχείο πολλών state-of-the-art δουλειών σε αυτό το τομέα [64, 97].

Ορισμός

Η μάθηση με αντιδιαστολή έχει γίνει δημοφιλής τεχνική για τη μη επιβλεπόμενη μάθηση τα τελευταία χρόνια. Ο όρος 'μάθηση με αντιδιαστολή' εισήχθη αρχικά από τον Arora [6] ως οι αλγόριθμοι που θυμίζουν το γνωστό ***word2vec*** [58], το οποίο εκμεταλλεύεται τη δημιουργία θετικών ζευγών που περιέχουν σημασιολογικά παρόμοιες αναπαραστάσεις και των αρνητικών δειγμάτων. Υποθέτουμε την ύπαρξη ενός θετικού ζεύγους $(x, x^+) \sim D_{sim}$ και k ανεξάρτητων και ταυτόσημων (i.i.d) αρνητικών δειγμάτων $x_1^-, x_2^-, ..., x_k^- \sim D_{neg}$ που προφανώς δεν σχετίζονται με το x. Τα θετικά ζεύγη μπορούν να προκύψουν λαμβάνοντας δύο ανεξάρτητες τυχαία ενισχυμένες προβολές (augmented views) του ίδιου δείγματος, π.χ. δύο αποκοπές της ίδιας εικόνας. Ο κύριος στόχος της μάθησης με αντιδιαστολή είναι να μεγιστοποιήσει τη σημασιολογική ομοιότητα του θετικού ζεύγους και να την ελαχιστοποιήσει για όλα τα άλλα ζεύγη που σχηματίζονται από τα αρνητικά δείγματα. Η *cosine similarity* έγινε η πιο κοινή συνάρτηση ομοιότητας για αυτές τις εργασίες λόγω της απλότητάς της.

$$CosSim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \cdot \mathbf{y}}{||\mathbf{x}|| \cdot ||\mathbf{y}||} \tag{0.0.1}$$

Μια άλλη όψη της μάθησης με αντιδιαστολή δίνεται πάλι από [6], στην οποία σχηματίζουμε λανθάνουσες κατηγορίες και κάθε προβολή που δημιουργούμε ανήκει στην κατηγορία που ορίζει το αρχικό δείγμα. Στη συνέχεια, ο στόχος για αυτή την εργασία είναι να ελαχιστοποιηθεί η απώλεια διασταυρούμενης εντροπίας σε όλες τις κατηγορίες.

Συνάρτηση Απώλειας με Αντιδιαστολή (Contrastive Loss)

Με τον ίδιο τρόπο, ορίζουμε την συνάρτησ απώλειας με αντιδιαστολή ως μια παραλλαγή της δημοφιλούς απώλειας διασταυρούμενης εντροπίας (cross-entropy loss). Μια ευρέως χρησιμοποιούμενη συνάρτηση απώλειας για μάθηση με αντιδιαστολή που χρησιμοποιήθηκε σε αρκετές δουλειές ([13, 79, 92, 59]) είναι η *NT-Xent* (Κανονικοποιημένη Απώλεια Διασταυρούμενης Εντροπίας με Θερμοκρασία). Ας υποθέσουμε ότι έχουμε μια μικροπαρτίδα μεγέθους N και ένα θετικό ζεύγος (x_i, x_j) και κάθε άλλο ζεύγος υπολογίζεται ως αρνητικό, τότε η απώλεια ορίζεται ως:

$$l_{i,j}^{NT-Xent} = -\log \frac{\exp(sim(x_i, x_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} \exp(sim(x_i, x_k)/\tau)} \tag{0.0.2}$$

όπου τ είναι μια παράμετρος θερμοκρασίας. Η παραπάνω απώλεια μπορεί να θεωρηθεί ως μια παραλλαγή μιας άλλης δημοφιλούς απώλειας για μάθηση με αντιδιαστολή, την *InfoNCE*, η οποία προτάθηκε από [59] και εφαρμόστηκε από [100] για μάθηση αναπαράστασης ζευγών *(εικόνας, κειμένου)* με αντιδιαστολή σε ιατρική απεικόνιση. Η συνάρτηση απώλειας *InfoNCE* ορίζεται ως:

$$l_{i,j}^{InfoNCE} = -\log \frac{\exp(sim(x_i, x_j)/\tau)}{\sum_{k=1}^{N} \exp(sim(x_i, x_k)/\tau)} \tag{0.0.3}$$

Πάλι υποθέτουμε ότι έχουμε μια μικροπαρτίδα μεγέθους N και το (x_i, x_j) είναι ένα θετικό ζεύγος ενώ όλα τα άλλα ζεύγη θεωρούνται αρνητικά.

Ευθυγράμμιση & Ομοιομορφία

Στη δουλειά τους, οι Wang και Isola [91] εισήγαγαν δύο βασικές ιδιότητες της απώλειας με αντιδιαστολή: 1) την ***ευθυγράμμιση***, 2) την ***ομοιομορφία***. Αυτές οι δύο είναι ποσοτικοποιημένα μέτρα της ποιότητας της αναπαράστασης. Συγκεκριμένα, η *ευθυγράμμιση* διασφαλίζει ότι οι αναπαραστάσεις και των δύο δειγμάτων του θετικού ζεύγους θα αντιστοιχούν σε κοντινά χαρακτηριστικά και η *ομοιομορφία* διασφαλίζει ότι τα διανύσματα χαρακτηριστικών θα διατηρούν όσο το δυνατόν περισσότερες πληροφορίες των δεδομένων, καθώς πρέπει να κατανέμονται περίπου ομοιόμορφα στην μοναδιαία υπερσφαίρα.

Definition: Τέλεια Ευθυγράμμιση

> Μπορούμε να πούμε ότι ένας κωδικοποιητής f είναι ***τέλεια ευθυγραμμισμένος*** αν $f(x) = f(y)$ σχεδόν σίγουρα πάνω από $(x, y) \sim D_{pos}$.

Definition: Τέλεια Ομοιομορφία

Μπορούμε να πούμε ότι ένας κωδικοποιητής f είναι ***τέλεια ομοιόμορφος*** αν η κατανομή του $f(x)$ για $x \sim D_{pos}$ είναι η ομοιόμορφη κατανομή σ_{m-1} στη $\mathcal{S}^{m-1}$, όπου η $\mathcal{S}^{m-1}$ είναι ένας χώρος Borel.

Βλέπουμε ότι δεν μπορούμε να έχουμε τέλεια ευθυγράμμιση και τέλεια ομοιομορφία ταυτόχρονα, καθώς αυτό το σενάριο υποδηλώνει ότι κάθε ενισχυμένη προβολή ενός σημείου δεδομένων θα πρέπει να έχει το ίδιο διάνυσμα χαρακτηριστικών για να έχει τέλεια ευθυγράμμιση. Ωστόσο, αυτό δεν σχηματίζει μια ομοιόμορφη κατανομή, έτσι δεν ικανοποιούμε τον περιορισμό της ομοιομορφίας. Ποσοτικοποιώντας τις παραπάνω ιδιότητες, μπορούμε να αποκτήσουμε τους ακόλουθους δύο μετρικούς δείκτες:

Definition: Απώλεια Ευθυγράμμισης

Η απώλεια ευθυγράμμισης ορίζεται ως η αναμενόμενη απόσταση μεταξύ των θετικών ζευγών (x, x^+):

$$l_{align}(f;a) \triangleq \mathop{\mathbb{E}}_{(x,x^+)\sim D_{pos}} [||f(x) - f(x^+)||_2^a], \quad a > 0 \tag{0.0.4}$$

όπου f είναι ένας κωδικοποιητής και το D_{pos} η κατανομή πάνω στα θετικά ζεύγη.

Definition: Απώλεια Ομοιομορφίας

Η απώλεια ομοιομορφίας ορίζεται ως ο λογάριθμος της συνάρτησης του μέσου διμερούς Γκαουσιανού δυναμικού:

$$l_{uniform}(f;t) \triangleq \log \mathop{\mathbb{E}}_{\substack{i.i.d. \\ (x,y)\sim D_{data}}} \left[e^{-t||f(x)-f(y)||_2^2}\right], \quad t > 0 \tag{0.0.5}$$

όπου f είναι ένας κωδικοποιητής, το D_{data} είναι η κατανομή πάνω στα δεδομένα και ο Γκαουσιανός πυρήνας (Radial Basis Function (RBF) kernel) ορίζεται ως $G_t : \mathcal{S}^d \times \mathcal{S}^d \to \mathbb{R}_+$ [17, 9]:

$$G_t(u,v) \triangleq e^{-t|u-v||_2^2} = e^{2t\cdot u^\top v - 2t} \tag{0.0.6}$$

Οι συγγραφείς απέδειξαν ότι η απώλεια με αντιδαστολή βελτιστοποιεί ασυμπτωτικά ταυτόχρονα για ευθυγράμμιση και ομοιομορφία και βρήκαν ισχυρή συμφωνία μεταξύ των δύο μετρικών και της απόδοσης στη μεταγενέστερη εργασία. Έστω $\mathcal{L}_{contrastive}$ η απώλεια με αντιδιαστολή, τ μια παράμετρος θερμοκρασίας, f ένας κωδικοποιητής και το εσωτερικό γινόμενο $(\cdot)$ η συνάρτηση ομοιότητας. Τότε, ασυμπτωτικά για M αρνητικά δείγματα έχουμε:

$$\lim_{M\to\infty} \mathcal{L}_{contrastive} - \log M = -\frac{1}{\tau} \mathop{\mathbb{E}}_{(x,x^+)\sim D_{pos}} \left[f(x)^\top f(x^+)\right] + \mathop{\mathbb{E}}_{x\sim D_{data}} \left[\log \mathop{\mathbb{E}}_{x\sim D_{data}} e^{f(x^-)^\top f(x)/\tau}\right] \tag{0.0.7}$$

όπου,

$$\mathcal{L}_{contrastive}(f;\tau,M) = \mathop{\mathbb{E}}_{\substack{(x,x^+)\sim D_{pos} \\ \{x_i^-\}_{i=1}^M \sim^{i.i.d.} D_{data}}} \left[-\log \frac{\exp\left(f(x)^\top f(x^+)/\tau\right)}{\exp\left(f(x)^\top f(x^+)/\tau\right) - \sum_i \exp\left(f(x_i^-)^\top f(x_i^+)/\tau\right)}\right] \tag{0.0.8}$$

Βλέπουμε ότι ο πρώτος όρος ελαχιστοποιείται αν και μόνον αν ο f είναι τέλεια ευθυγραμμισμένος και ο δεύτερος όρος αν και μόνον αν ο f είναι ένας τέλεια ομοιόμορφος κωδικοποιητής.

Πολυτροπική Μάθηση

Η έρευνα πάνω στη πολυτροπική μάθηση είναι στο προσκήνιο για πάνω από 50 χρόνια προσπαθώντας να σχεδιάσει πράκτορες με έξυπνες ικανότητες όπως η κατανόηση, η λογική και η μάθηση με τον ίδιο τρόπο που ένας άνθρωπος θα εκτελούσε τις ίδιες διεργασίες. Η πολυτροπική μάθηση είναι ένα πολύ ενεργό ερευνητικό πεδίο και διευρύνεται σε πολλές επιστήμες. Η βαθιά μάθηση κατέκτησε τον κόσμο την περασμένη δεκαετία, και έκανε ξανά

δημοφιλή την έρευνα στον πολυτροπικό τομέα, δίνοντας τα εργαλεία για να ξεκλειδώσουμε νέες δυνατότητες και προοπτικές των στοιχείων που χρησιμοποιούνται στη μηχανική μάθηση. Οι κύριες τροπικότητες (modalities) που χρησιμοποιούνται για την έρευνα στη μηχανική μάθηση είναι: γλώσσα, όραση, ακουστική, αφή, φυσιολογικά σήματα και κινητικότητα. Διάφορες εφαρμογές όπως τα αυτόνομα αυτοκίνητα [93], οι τεχνολογίες κειμένου-σε-ομιλία [5], η κατανόηση βίντεο [81], η παραγωγή εικόνας και κειμένου [67, 77], οι ενσωματωμένοι πράκτορες [10], η πολυαισθητηριακή συγχώνευση [47] μας φέρνουν πιο κοντά στον στόχο μας για έξυπνα συστήματα σε τομείς όπως η ρομποτική, η υγειονομική περίθαλψη και η αλληλεπίδραση ανθρώπου-υπολογιστή.

Κάθε τροπικότητα έχει 3 βασικές αρχές: *ετερογένεια*, *συνδέσεις* και *αλληλεπιδράσεις* [52] που έχουν οδηγήσει σε επακόλουθες καινοτομίες. Οι τροπικότητες είναι ετερογενείς λόγω της ποικιλίας της ποιότητας και της δομής της πληροφορίας. Για παράδειγμα, η όραση συχνά καταγράφεται ως εικόνες ενώ η γλώσσα ως κείμενο δημιουργημένο από ένα σύνολο χαρακτήρων. Συνδεδεμένες είναι οι τροπικότητες που συχνά σχετίζονται και μοιράζονται κοινά στοιχεία, π.χ. η ακουστική και η γλώσσα είναι συνδεδεμένες όταν έχουμε ένα δείγμα ομιλίας και τις απομαγνητοφωνήσεις του. Τέλος, οι τροπικότητες αλληλεπιδρούν για να αποδώσουν νέες πληροφορίες για την εξαγωγή συμπερασμάτων. Αυτές οι αρχές οδήγησαν τους ερευνητές σε μια ταξινόμηση [52] του πεδίου που περιέχει τις 6 προκλήσεις στην σύγχρονη πολυτροπική μάθηση: *αναπαράσταση*, *ευθυγράμμιση*, *λογική*, *παραγωγή*, *μεταφορά* και *ποσοτικοποίηση*.

Μάθηση Αναπαραστάσεων

Η αναπαράσταση των δεδομένων, σε μορφή που να μπορεί να χρησιμοποιηθεί από ένα υπολογιστικό μοντέλο, ήταν πάντα μια πρόκληση στην κοινότητα της μηχανικής μάθησης. Η πολυτροπική μάθηση αναπαραστάσεων μελετά τους τρόπους που μπορεί να αναπαρασταθεί και να συγκροτηθεί η πολυτροπική πληροφορία ώστε να αντικατοπτρίζει την ετερογένεια και τις διασυνδέσεις μεταξύ των τροπικοτήτων. Ωστόσο, αυτή η πρόκληση παρουσιάζει πολλές δυσκολίες όπως το πώς να συνδυάσουμε διαφορετικές τροπικότητες, πώς να αντιμετωπίσουμε τον θόρυβο ή πώς να διαχειριστούμε τα ελλιπή δεδομένα. Η ικανότητα να αναπαρασταθούν τα δεδομένα με έναν ουσιαστικό τρόπο που περιέχει κρίσιμες πληροφορίες για τη φύση της κάθε οντότητας αποτελεί τη ραχοκοκαλιά κάθε πολυτροπικού μοντέλου μάθησης. Για να καλύψουμε την παραπάνω πρόκληση, διαιρούμε το πεδίο σε 3 υπο-προκλήσεις [52]: *συγχώνευση αναπαραστάσεων* όπου ο αριθμός των τροπικοτήτων είναι μεγαλύτερος από τον αριθμό των ξεχωριστών αναπαραστάσεων, *συντονισμός αναπαραστάσεων* όπου διατηρούμε τον ίδιο αριθμό αναπαραστάσεων αλλά ενθαρρύνουμε τη αλληλεπίδραση μεταξ΄ύ των τροπικοτήτων και *διαίρεση αναπαραστάσεων* όπου ο αριθμός των αναπαραστάσεων είναι μεγαλύτερος από τις δοσμένες τροπικότητες και προσπαθεί να καταγράψει τη δομική πληροφορία των δεδομένων. Στο Σχήμα 0.0.4 παρουσιάζουμε σχηματικά τις παραπάνω υπο-προκλήσεις.

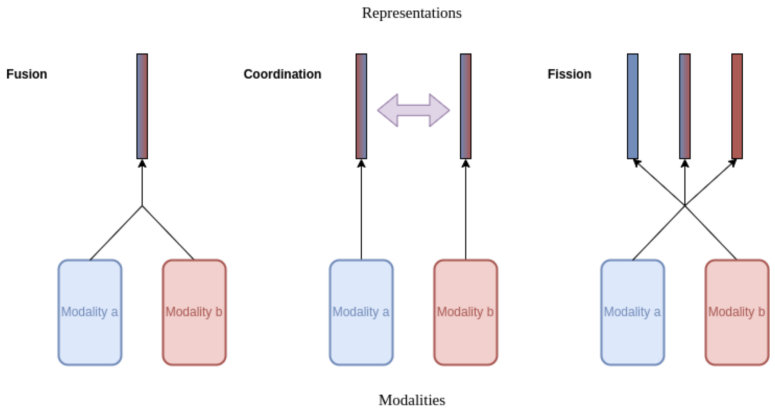


Figure 0.0.4: **Πολυτροπική Μάθηση Αναπαραστάσεων:** Συγχώνευση, Συντονισμός, Διαίρεση

Σχετική Βιβλιογραφία

CLIP

Πρόσφατες δουλειές έχουν αποδείξει ότι η βαθιά μάθηση έχει φέρει επανάσταση στην όραση υπολογιστών. Ωστόσο, εξακολουθούν να υπάρχουν σημαντικά προβλήματα, καθώς αυτές οι προσεγγίσεις απαιτούν τεράστια σύνολα δεδομένων με χειροκίνητη επισημείωση. Ταυτόχρονα, τα τυπικά μοντέλα όρασης είναι καλά σε μία μόνο εργασία και απαιτούν σημαντική προσπάθεια για να προσαρμοστούν σε μια νέα. Για να αντιμετωπιστούν αυτά τα προβλήματα, ο Radford και οι συνεργάτες του εισήγαγαν το CLIP (Contrastive Language-Image Pretraining) [64]. Το παραπάνω νευρωνικό δίκτυο εκπαιδεύεται σε μια μεγάλη ποικιλία εικόνων με εποπτεία από φυσική γλώσσα που είναι διαθέσιμη από το διαδίκτυο. Συγκεκριμένα, για να μάθει από το ακατέργαστο κείμενο, η εργασία προ-εκπαίδευσής του είναι να προβλέψει ποια λεζάντα αντιστοιχεί σε ποια εικόνα. Αποδείχθηκε ότι η παραπάνω πρόταση είναι ένας αποτελεσματικός και επεκτάσιμος τρόπος για να μάθουμε την SOTA αναπαράσταση μιας εικόνας. Μετά την προ-εκπαίδευση, η φυσική γλώσσα χρησιμοποιείται για την αναφορά διαφορετικών οπτικών εννοιών, επιτρέποντας έτσι τη μεταφορά γνώσης του μοντέλου χωρις την εκθεσή του σε δεδομένα σχετικά με την εργασία που εκτελεί. Το CLIP αξιολογήθηκε σε περισσότερα από 30 σύνολα δεδομένων υπολογιστικής όρασης και μια ποικιλία εργασιών όπως OCR (αναγνώριση χαρακτήρων αντικειμένων), αναγνώριση δράσεων σε βίντεο, γεωεντοπισμός και πολλούς τύπους λεπτομερών ταξινομήσεων.

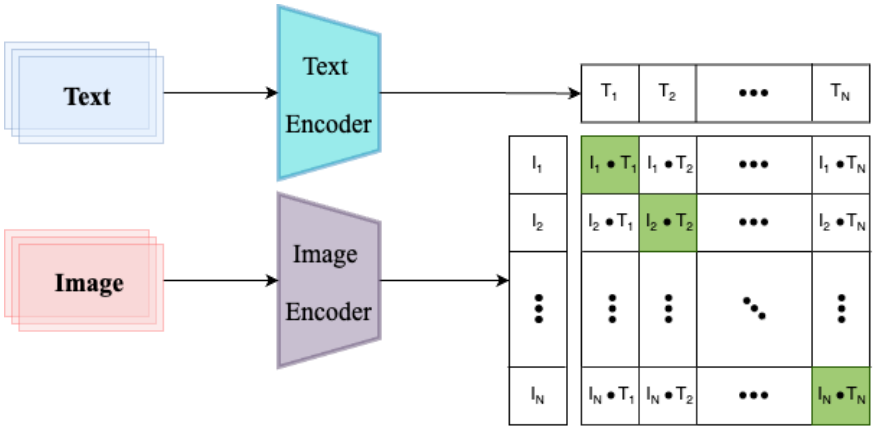


Figure 0.0.5: **CLIP:** Contrastive Language-Image Pre-training

Επίβλεψη μεσω Φυσικής Γλώσσας

Η προ-εκπαίδευση με ακατέργαστο κείμενο έφερε επανάσταση στο τομέα του NLP τα τελευταία χρόνια. Διάφοροι στόχοι εκπαίδευσης, όπως η πρόβλεψη ενός μασκαρισμένου περιεχομένου και η αυτόματη γλωσσική μοντελοποίηση, έγιναν πιο επεκτάσιμοι. Επιπλέον, οι αρχιτεκτονικές χωρίς εξειδίκευση εργασιών εξάλειψαν την ανάγκη για εξειδικευμένες εξόδους στα νευρωνικά δίκτυα μετά την ανάπτυξη του "κείμενο-προς-κείμενο" ως διεπαφή εισόδου-εξόδου. Για παράδειγμα, το GPT-3 [11] έγινε ένα ευρέως δημοφιλές σύστημα αποδεικνύοντας ότι αυτά τα συστήματα είναι τώρα ανταγωνιστικά σε πολλές εργασίες απαιτώντας λίγα έως καθόλου εξειδικευμένα δεδομένα για εκπαίδευση. Σε άλλους τομείς όπως η υπολογιστική όραση, είναι ακόμα μια κοινή πρακτική να εκπαιδεύονται μοντέλα με σύνολα δεδομένων που σημειώνονται χειροκίνητα. Η προτεινόμενη ιδέα ήταν, αν τα συστήματα υπολογιστικής όρασης μπορούν να μάθουν απευθείας από κείμενο του διαδικτύου με τον ίδιο τρόπο όπως στο NLP. Υπάρχουν ενθαρρυντικές εργασίες και αξιοσημείωτα πλεονεκτήματα αυτής της ιδέας. Πρώτον, είναι πιο εύκολο να επεκταθεί η εποπτεία φυσικής γλώσσας σε σύγκριση με τις ετικέτες που προέρχονται από ανθρώπους για την ταξινόμηση εικόνων, καθώς η φυσική γλώσσα δεν απαιτεί σημείωση. Επιπλέον, η εποπτεία με φυσικής γλώσσας αποδίδει καλύτερες αναπαραστάσεις από τις εποπτευόμενες και αυτο-εποπτευόμενες μεθόδους, διότι συνδέει την αναπαράσταση με τη γλώσσα και επιτρέπει ευέλικτη μεταφορά της γνώσης.

Μέθοδος

Το CLIP ακολουθεί την εργασία του CMC [83], στην οποία αποδείχθηκε ότι ένας αντιθετικός στόχος αποδίδει καλύτερες αναπαραστάσεις από έναν προβλεπτικό στόχο. Συγκεκριμένα, το CLIP χρησιμοποιεί μάθηση με

αντιδιαστολη και προσπαθεί να προβλέψει ποιο κείμενο "ως σύνολο" είναι συζευγμένο με μια εικόνα αντί να προβλέπει τις "ακριβείς" λέξεις, που είναι μια δύσκολη εργασία. Δεδομένης μιας παρτίδας N ζευγών εικόνας-κείμενου, το CLIP προσπαθεί να προβλέψει ποιοι από τους $N \times N$ συνδυασμούς είναι οι σωστοί. Για να το κάνει αυτό, το CLIP μαθαίνει έναν πολυτροπικό χώρο χαρακτηριστικών με κοινή εκπαίδευση ενός κωδικοποιητή εικόνας και ενός κωδικοποιητή κειμένου. Ο σκοπός του είναι να μεγιστοποιήσει την ομοιότητα συνημιτόνου των αναπαραστάσεων της εικόνας και του κειμένου των N πραγματικών ζευγών, ενώ ταυτόχρονα να ελαχιστοποιήσει την ομοιότητα συνημιτόνου των $N(2-N)$ εσφαλμένων ζευγών. Ως στόχο πρέπει να βελτιστοποιήσει μια συμμετρική απώλεια διασταυρούμενης εντροπίας πάνω σε αυτούς τους δείκτες ομοιότητας, παρόμοια με την προσαρμογή της απώλειας InfoNCE [59] για ζεύγη (εικόνα, κείμενο) στο ConVIRT [100]. Συγκεκριμένα, για μια παρτίδα N ζευγών εικόνας-κείμενου $(v_i^T, v_i^I), i = 1, ..., N$, υπολογίζουμε την απώλεια CLIP όπως φαίνεται παρακάτω:

$$l_i^{(I \to T)} = -\log \frac{\exp\left(v_i^I \cdot v_i^T / \tau\right)}{\sum_{k=1}^{N} \exp\left(v_i^I \cdot v_k^T / \tau\right)} \tag{0.0.1}$$

$$l_i^{(T \to I)} = -\log \frac{\exp\left(v_i^T \cdot v_i^I / \tau\right)}{\sum_{k=1}^{N} \exp\left(v_i^T \cdot v_k^I / \tau\right)} \tag{0.0.2}$$

$$\mathcal{L}_{CLIP} = \frac{1}{2N} \sum_{i=1}^{N} \left(l_i^{(I \to T)} + l_i^{(T \to I)}\right) \tag{0.0.3}$$

Κατά τη διάρκεια της εκπαίδευσης, η μόνη ενίσχυση δεδομένων που εφαρμόζεται είναι μια τυχαία τετραγωνική περικοπή από εικόνες με αλλαγμένες διαστάσεις. Κατά την ώρα της αξιολόγησης, ο εκπαιδευμένος κωδικοποιητής κειμένου συνθέτει έναν γραμμικό ταξινομητή μηδενικής βολής (zero-shot) μέσω της αναπαράστασης των κλάσεων των εικόνων στόχων. Στα Σχήματα 0.0.5 και 0.0.6 υπάρχουν σχηματικές αναπαραστάσεις της εκπαίδευσης και της αξιολόγησης του CLIP αντίστοιχα.

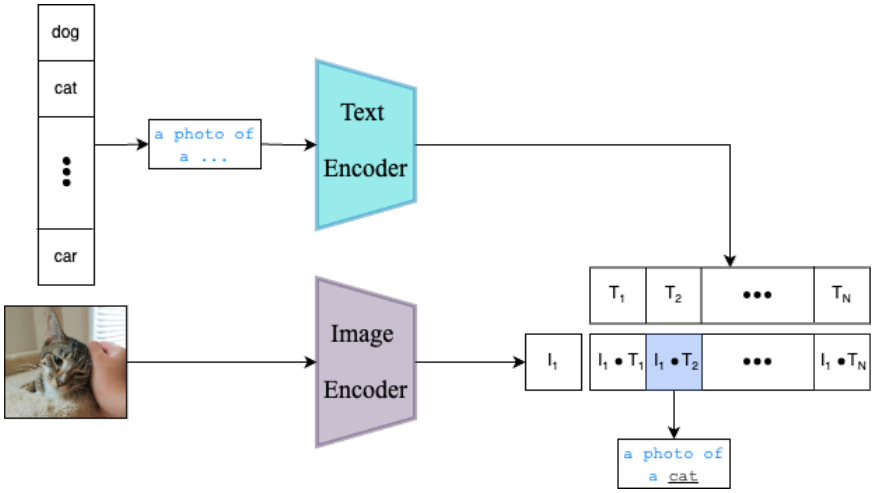


Figure 0.0.6: **Αξιολόγηση πρόβλεψης μηδενικής βολής του CLIP.** Μετατρέπουμε όλες τις κλάσεις ενός συνόλου δεδομένων σε λεζάντες όπως "μια φωτογραφία μιας γάτας" και προβλέπουμε την κλάση της λεζάντας που το CLIP εκτιμά ότι ταιριάζει καλύτερα με μια δεδομένη εικόνα εισόδου.

Transformers

Οι Transformers είναι μοντέλα που βασίζονται μόνο σε μηχανισμούς αυτοπροσοχής (self-attention) [88], χωρίς να χρησιμοποιούν CNN ή RNN. Μπορούμε να τους ταξινομήσουμε σε 3 βασικές κατηγορίες: *1) autoencoding Transformers*, που είναι μια στοίβα από κωδικοποιητές, *2) auto-regressive Transformers*, που είναι μια στοίβα από αποκωδικοποιητές, και *3) sequence-to-sequence Transformers*, που είναι μια στοίβα από κωδικοποιητές

ακολουθούμενη από μια στοίβα από αποκωδικοποιητές. Η είσοδος ενός Transformer είναι μια ενσωμάτωση των λέξεων (word embeddings), ωστόσο στην περίπτωση seq-to-seq χρειαζόμαστε και μια ενσωμάτωση της θέσης (positional embedding). Οι κωδικοποιητές αποτελούνται από ένα στρώμα αυτοπροσοχής και ένα νευρωνικό δίκτυο. Κάθε κωδικοποιητής περνά την έξοδό του ως είσοδο στον επόμενο κωδικοποιητή στη στοίβα. Οι αποκωδικοποιητές, από την άλλη πλευρά, περιέχουν ένα στρώμα αυτοπροσοχής, ακολουθούμενο από ένα στρώμα προσοχής κωδικοποιητή-αποκωδικοποιητή και ένα νευρωνικό δίκτυο. Αυτή η αρχιτεκτονική έχει ως αποτέλεσμα SOTA εφαρμογές και γρηγορότερους χρόνους εκπαίδευσης από τις παραδοσιακές αρχιτεκτονικές CNN και RNN. Το BERT [20] (Bidirectional Encoder Representations from Transformers) πρέπει να θεωρείται μια από τις πιο χαρακτηριστικές αρχιτεκτονικές Transformer στη βιβλιογραφία. Το BERT είναι μια γενικής χρήσης αρχιτεκτονική βασισμένη σε Transformer που επιτυγχάνει SOTA αποτελέσματα σε διάφορες εργασίες και σύνολα δεδομένων Επεξεργασίας Φυσικής Γλώσσας. Επιπλέον, μπορούμε να θεωρήσουμε το BERT ως έναν βαθύ αμφίδρομο autoencoding Transformer. Πιο συγκεκριμένα, κάθε εισερχόμενη λέξη αναπαρίσταται ως μια ενσωμάτωση τοκεν (token embedding), μια ενσωμάτωση τμήματος (segment embedding) και μια ενσωμάτωση θέσης (positional embedding). Κατά την εκπαίδευση, το BERT καλύπτει τυχαία ένα μικρό τμήμα των εισερχόμενων ενσωματώσεων και ο στόχος είναι να συμπληρώσει τις καλυμμένες λέξεις εκπαιδεύοντας τους κωδικοποιητές αυτοπροσοχής. Τέλος, η αναπαράσταση εξόδου αποτελείται από την κρυφή κατάσταση του τοκεν ταξινόμησης που χρησιμεύει ως είσοδος στην κεφαλή ταξινόμησης που προσαρμόζεται πάνω στο BERT.

Μηχανισμοί Προσοχής (Attention)

Αυτοπροσοχή

Τα μοντέλα Transformer πραγματοποιούν υπολογισμούς παράλληλα, χρησιμοποιώντας μπλοκ αυτοπροσοχής [88], σε αντίθεση με τα παραδοσιακά ακολουθιακά μοντέλα που πραγματοποιούν όλους τους υπολογισμούς διαδοχικά. Σε αυτό το μέρος, πρέπει να ορίσουμε πώς λειτουργεί ο μηχανισμός Αυτοπροσοχής. Ας υποθέσουμε ότι έχουμε μια πρόταση που περιέχει n λέξεις. Ξεκινάμε αναπαριστώντας τις n λέξεις στην πρόταση χρησιμοποιώντας ενσωματώσεις λέξεων, έτσι το αποτέλεσμα πρέπει να είναι ένα d-διάστατο διάνυσμα $x_i \in \mathbb{R}^d$. Η πρότασή μας θα είναι ο πίνακας $n \times d$ $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{(n \times d)}$. Επειδή αυτή η αναπαράσταση δεν λαμβάνει υπόψη τη γειτονιά κάθε λέξης στην πρόταση, η αυτοπροσοχή υπολογίζει παράλληλα n αναπαραστάσεις αυτοπροσοχής $A_1, \ldots, A_n$ για τις n λέξεις. Για κάθε ενσωματωμένη λέξη $x_i \in \mathbb{R}^d$ υπολογίζουμε ένα ερώτημα $q_i \in \mathbb{R}^{d_q}$, ένα κλειδί $k_i \in \mathbb{R}^{d_k}$ και μια τιμή $v_i \in \mathbb{R}^{d_v}$ που αναπαρίστανται ως γραμμές του πίνακα X, Q, K και V, με γραμμική προβολή του X στον d-διάστατο χώρο των ερωτημάτων, κλειδιών και τιμών όπως φαίνεται παρακάτω:

$$Q = XW_Q,\ K = XW_K,\ V = XW_V \tag{0.0.4}$$

και

$$q_i = x_i W_Q,\ h_i = x_i W_K,\ v_i = x_i W_V \tag{0.0.5}$$

όπου W_Q, W_K και W_V είναι εκμαθημένοι πίνακες $d \times d$ και οι q_i, k_i και v_i είναι οι i-ες γραμμές των πινάκων Q, K και V $n \times d$. Στη συνέχεια υπολογίζουμε την αναπαράσταση αυτοπροσοχής A_i ως το softmax του εσωτερικού γινομένου μεταξύ του q_i και του k_j (και οι δύο έχουν διάσταση d_k) για $j = 1, ..., n$ πολλαπλασιασμένο με τις τιμές v_i (διάστασης d_v).

$$A_i(q_i, K, V) = \sum_{i=1}^{n} \frac{\exp(q_i k_i)}{\sum_j \exp(q_i k_j)} v_i \tag{0.0.6}$$

αν υπολογίσουμε το άθροισμα για όλες τις n λέξεις, έχουμε μια κλιμακούμενη προσοχή με παράγοντα κλιμάκωσης $\frac{1}{\sqrt{d_k}}$ όπως φαίνεται παρακάτω:

$$A(X) = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{0.0.7}$$

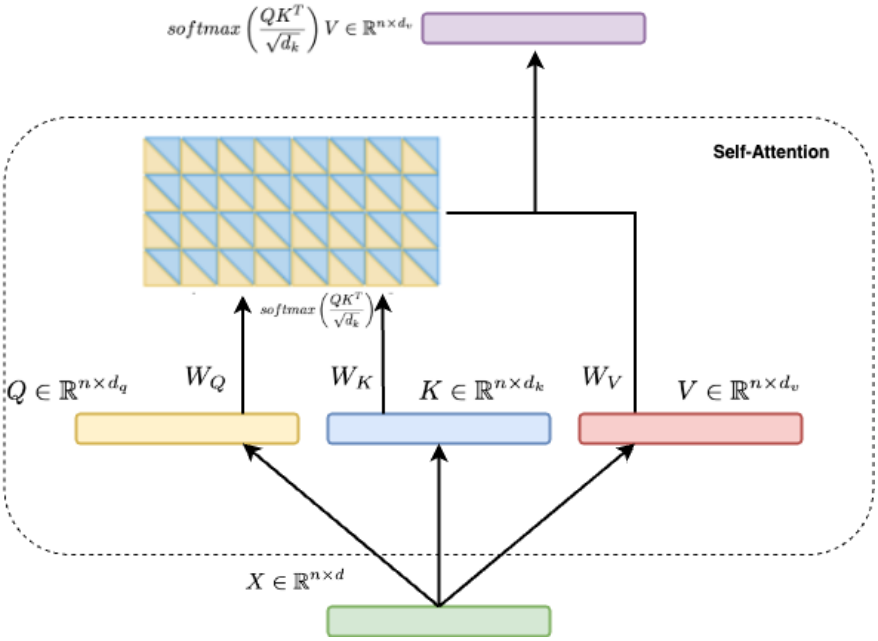


Figure 0.0.7: **Μηχανισμός Αυτοπροσοχής**

Πολυκέφαλη Προσοχή

Η πολυκέφαλη προσοχή πραγματοποιεί αυτοπροσοχή πολλές φορές. Αντί για εσωτερικά γινόμενα, τα διανύσματα ερωτημάτων, κλειδιών και τιμών πολλαπλασιάζονται με πίνακες W_h^Q, W_h^K, W_h^V για κάθε κεφαλή $h = 1, ..., m$ και οι αναπαραστάσεις πολυκέφαλης προσοχής είναι:

$$A_h(X) = \text{Attention}_h(Q, K, V) = \text{Attention}_h(W_h^Q Q, W_h^K K, W_h^V V) = \text{softmax}\left(\frac{W_h^Q Q W_h^K K^\top}{\sqrt{d_k}}\right) V \quad (0.0.8)$$

για κάθε ενσωματωμένη λέξη i. Οι m αναπαραστάσεις πολυκέφαλης προσοχής $A_h(X)$ για $h = 1, ..., m$ συγκεντρώνονται και πολλαπλασιάζονται:

$$\text{MultiHead}(X) = \text{Concat}\left(A_1(X), ..., A_m(X)\right) W^O \quad (0.0.9)$$

όπου το W^O είναι ένας εκμαθημένος πίνακας $md \times d$. Για να υπολογίσουμε την πολυκέφαλη προσοχή, πρώτα υπολογίζουμε το άθροισμα της ενσωματωμένης ακολουθίας X διαστάσεων $n \times d$ και της κωδικοποίησης θέσης P ως είσοδο. Στη συνέχεια, για κάθε κεφαλή προσοχής, υπολογίζουμε τα ερωτήματα Q, τα κλειδιά K και τις τιμές V που αναπαρίστανται από πίνακες $d \times d$, οι οποίοι εισέρχονται στο επίπεδο της πολυκέφαλης προσοχής του οποίου η έξοδος είναι διαστάσεων $n \times d$.

Αρχιτεκτονική

Δεδομένου ότι έχουμε προτάσεις αποτελούμενες από ενσωματωμένες λέξεις, ένας Transformer μπορεί να χρησιμοποιηθεί για διάφορες εργασίες όπως η κατανόηση φυσικής γλώσσας, η παραγωγή κειμένου και η μετάφραση μιας πρότασης από μία γλώσσα σε μία άλλη. Ένας Transformer αποτελείται από μπλοκ κωδικοποιητή ή αποκωδικοποιητή ή και τα δύο.

Κωδικοποίηση Θέσης

Η θέση των λέξεων απαιτείται για τον υπολογισμό του σκορ προσοχής. Οι θέσεις των λέξεων στην πρόταση αναπαρίστανται ως μια κωδικοποίηση θέσης από ημιτονοειδείς και συνημιτονοειδείς συναρτήσεις και προστίθενται στο X [88]. Πιο συγκεκριμένα, ο πίνακας κωδικοποίησης θέσης P είναι ένας πίνακας $n \times d$ και ορίζεται από:

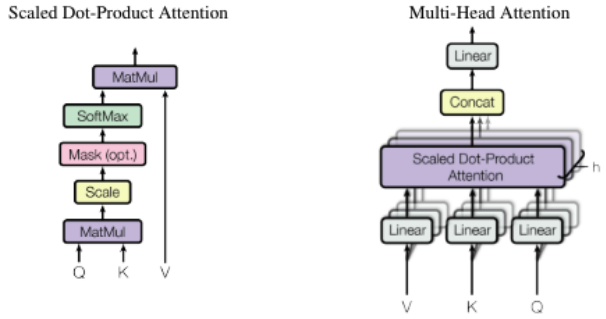


Figure 0.0.8: **Προσοχή vs Πολυκέφαλη Προσοχή.** Αριστερά: Η Κλιμακούμενη Προσοχή Εσωτερικού Γινομένου. Δεξιά: Το Μοντέλο Πολυκέφαλης Προσοχής. [88]

$$P_{pos,2i} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right),\ P_{pos,2i+1} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{0.0.10}$$

όπου pos είναι η θέση της λέξης στην πρόταση, το d είναι η διάσταση των ενσωματώσεων λέξεων και το $i = 1, ..., d$ είναι ο δείκτης διάστασης. Προσθέτοντας την κωδικοποίηση θέσης P στο X επιτρέπουμε στο μοντέλο να μάθει να επικεντρώνει τη προσοχή σε σχετικές θέσεις.

Κωδικοποιητής

Το μπλοκ κωδικοποιητή παίρνει ως είσοδο έναν πίνακα X ενσωματωμένων λέξεων. Η κωδικοποίηση θέσης προστίθεται στις ενσωματωμένες λέξεις για να σχηματίσει την είσοδο $X + P$. Στη συνέχεια, υπολογίζουμε τα ερωτήματα Q, τα κλειδιά K και τις τιμές V και τα περνάμε μέσα από μια στρώση πολυκέφαλης προσοχής της οποίας η έξοδος τροφοδοτείται σε ένα νευρωνικό δίκτυο προώθησης. Για να δημιουργήσουμε έναν κωδικοποιητή, το μπλοκ που αποτελείται από τη πολυκέφαλη προσοχή και το νευρωνικό δίκτυο προώθησης επαναλαμβάνεται πολλές φορές.

Αποκωδικοποιητής

Η έξοδος του κωδικοποιητή τροφοδοτείται στο μπλοκ αποκωδικοποιητή, το οποίο προβλέπει την μεταφρασμένη πρόταση. Ο αποκωδικοποιητής επίσης αποτελείται από πολλαπλά μπλοκ πολυκέφαλης προσοχής, που τροφοδοτούνται σε νευρωνικά δίκτυα προώθησης και προσθέτουν κωδικοποιήσεις θέσης στις εισόδους. Τόσο ο κωδικοποιητής όσο και ο αποκωδικοποιητής μπορεί να αποτελούνται από υπολειμματικές συνδέσεις μεταξύ των μπλοκ και να προσθέτουν στρώσεις κανονικοποίησης πριν από τα νευρωνικά δίκτυα προώθησης. Η έξοδος του αποκωδικοποιητή τροφοδοτείται μέσω μιας γραμμικής προβολής ακολουθούμενης από μια στρώση softmax. Κατά τη διάρκεια της παραγωγής, ο αποκωδικοποιητής προβλέπει νέες λέξεις, ενώ κατά τη διάρκεια της εκπαίδευσης, ο αποκωδικοποιητής προβλέπει καλυμμένες λέξεις από την είσοδο. Η αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή φαίνεται στο Σχήμα 0.0.9.

Εκπαίδευση

Η προ-εκπαίδευση ενός Transformer είναι υπολογιστικά δαπανηρή και συχνά περιλαμβάνει τεράστια μεγέθη μη επισημασμένων δεδομένων. Οι πιο συνηθισμένοι στόχοι βελτιστοποίησης για την προ-εκπαίδευση γλωσσικών μοντέλων είναι *1)* η πρόβλεψη καλυμμένων λέξεων, που είναι η πρόβλεψη μιας τυχαίας διεγραμμένης λέξης σε μια πρόταση ή η πρόβλεψη της επόμενης λέξης και *2)* η πρόβλεψη εάν δύο προτάσεις ακολουθούν η μία την άλλη ή όχι. Αυτό το υπολογιστικά δαπανηρό βήμα συνήθως πραγματοποιείται μία φορά, ακολουθούμενο από ένα σχετικά γρήγορο βήμα προσαρμογής. Κατά την προσαρμογή, το προ-εκπαιδευμένο μοντέλο προσαρμόζεται σε ένα συγκεκριμένο σύνολο δεδομένων και συγκεκριμένη εργασία. Η προσαρμογή μπορεί να πραγματοποιηθεί πολύ αποδοτικά σε ένα σχετικά μικρό σύνολο δεδομένων για μια συγκεκριμένη χρήση. Η προ-εκπαίδευση ακολουθούμενη από την προσαρμογή αναφέρεται ως μεταφορά μάθησης.

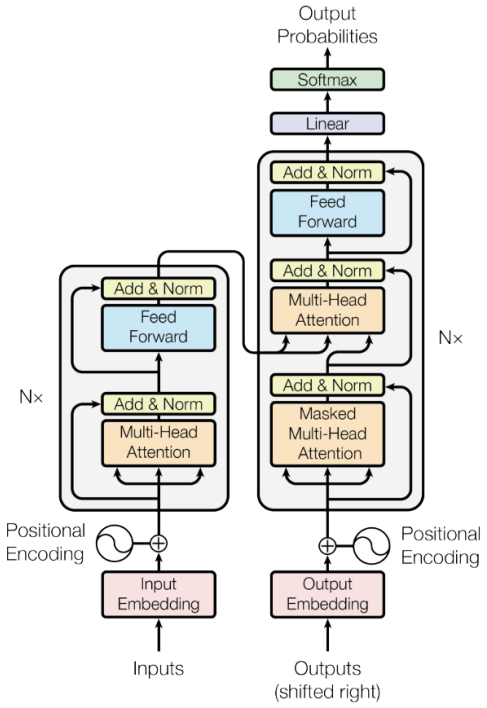


Figure 0.0.9: **Η αρχιτεκτονική του μοντέλου Transformer** [88]

Vision Transformers

Οι Vision Transformers [22] (ViTs) αποτελούν μια σημαντική πρόοδο στον τομέα της υπολογιστικής όρασης, εφαρμόζοντας τις αρχές των αρχιτεκτονικών transformer από την επεξεργασία φυσικής γλώσσας (NLP) πάνω σε εικόνες. Με την αξιοποίηση των μηχανισμών αυτοπροσοχής, οι ViTs έχουν αναδιαμορφώσει τις δυνατότητες και την απόδοση των μοντέλων αναγνώρισης εικόνας, επιτυγχάνοντας έτσι εκπληκτικά αποτελέσματα σε διάφορα benchmarks και εφαρμογές.

Αρχιτεκτονική

Η βασική καινοτομία των Vision Transformers είναι ο μηχανισμός αυτοπροσοχής, ο οποίος επιτρέπει στο μοντέλο να ζυγίζει τη σημασία των διαφόρων μερών των εισερχόμενων δεδομένων δυναμικά. Για μια εικόνα που χωρίζεται σε τμήματα (patches), ο μηχανισμός αυτοπροσοχής υπολογίζει τις σχέσεις μεταξύ των τμημάτων για να κατανοήσει την εικόνα ως σύνολο. Στο Σχήμα 0.0.10 παρουσιάζεται η διαδικασία που περιγράφεται παραπάνω.

Δεδομένης μιας εισόδου εικόνας $x \in \mathbb{R}^{H\times W\times C}$, η εικόνα χωρίζεται σε $N = \frac{HW}{P^2}$ τμήματα μεγέθους $P \times P$. Κάθε τμήμα x_p στη συνέχεια μετατρέπεται σε διάνυσμα $x_p \in \mathbb{R}^{P^2\cdot C}$. Αυτά τα τμήματα μετατρέπονται γραμμικά σε ενσωματώματα

$$z_0^i = W_p x_p^i + b_p, \tag{0.0.11}$$

όπου $W_p \in \mathbb{R}^{D\times P^2 C}$ και $b_p \in \mathbb{R}^D$. Ένα τόκεν ταξινόμησης z_0^0 προστίθεται στη ακολουθία των ενσωματωμένων τμημάτων. Στη συνέχεια προστίθενται ενσωματώματα θέσης $E_{pos} \in \mathbb{R}^{(N+1)\times D}$ για να σχηματιστεί έτσι το

$$z_0 = [z_0^0; z_0^1; \ldots; z_0^{N-1}] + E_{pos} \tag{0.0.12}$$

Η ακολουθία z_0 επεξεργάζεται από L στρώματα ενός κωδικοποιητή Transformer. Κάθε στρώμα αποτελείται από μια Κανονικοποίηση Στρώματος (Layer Normalization), μια Πολυκέφαλη Αυτοπροσοχή (MHSA), και ένα Δίκτυο Προώθησης (FFN). Συγκεκριμένα, η επεξεργασία μέσα σε κάθε στρώμα ορίζεται ως:

$$\hat{z}_{\ell-1} = \text{LayerNorm}(z_{\ell-1}), \tag{0.0.13}$$

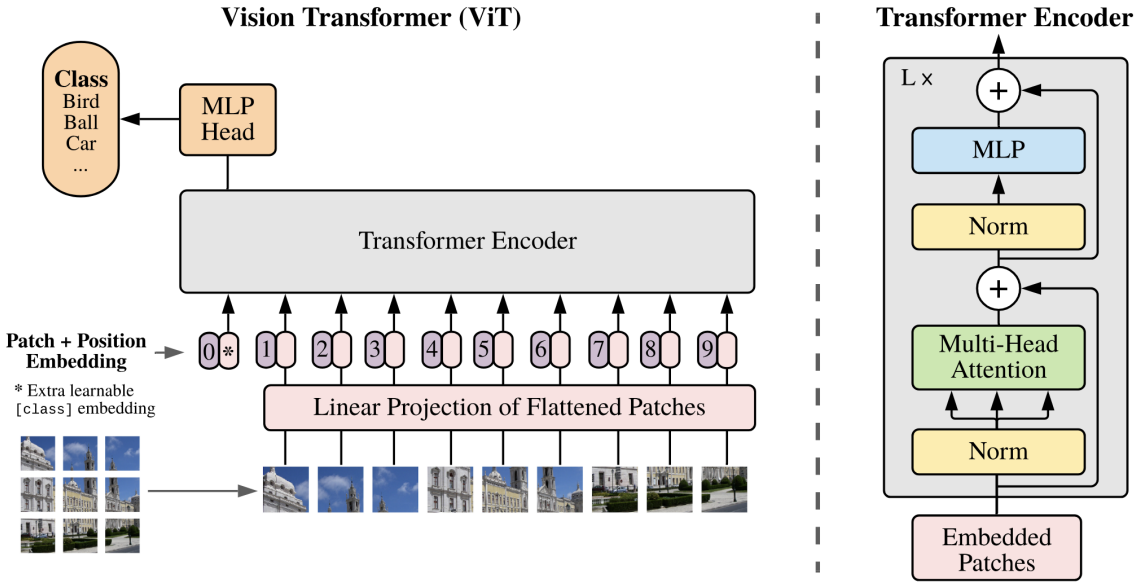


Figure 0.0.10: **ViT.** Επισκόπηση του μοντέλου [22]

$$z'_\ell = \hat{z}_{\ell-1} + \text{MHSA}(\hat{z}_{\ell-1}), \tag{0.0.14}$$

και

$$z_\ell = z'_\ell + \text{FFN}(\text{LayerNorm}(z'_\ell)) \tag{0.0.15}$$

Μετά από τα L στρώματα κωδικοποιητή Transformer, το τόκεν ταξινόμησης z_0^0 χρησιμοποιείται για την τελική ταξινόμηση. Το αποτέλεσμα λαμβάνεται εφαρμόζοντας Κανονικοποίηση Στρώματος ακολουθούμενη από ένα νευρωνικό δίκτυο (μια MLP κεφαλή):

$$\hat{z}_L = \text{LayerNorm}(z_L), \tag{0.0.16}$$

και

$$y = \text{MLP}(\hat{z}_L^0), \tag{0.0.17}$$

όπου

$$\text{MLP}(x) = xW_{head} + b_{head} \tag{0.0.18}$$

Συνοπτικά, ο Vision Transformer μπορεί να διατυπωθεί ως

$$y = \text{MLP}(\text{LayerNorm}(z_L^0)), \tag{0.0.19}$$

όπου z_L είναι το αποτέλεσμα μετά από τα L στρώματα κωδικοποιητή Transformer που εφαρμόζονται στην ακολουθία των τμημάτων της εικόνας μαζί με τα προστιθέμενα ενσωματώματα θέσης.

Λεκτική Περιγραφή Εικόνων

Η λεκτική περιγραφή εικόνων είναι μεταξύ των βασικών χρήσεων των βασικών μοντέλων που χρησιμοποιούν τόσο μοντέλα εικόνας όσο και κειμένου.

Definition: Λεκτική Περιγραφή Εικόνων

> *Λεκτική Περιγραφή Εικόνων* είναι η διαδικασία η οποία περιγράφουμε το οπτικό περιεχόμενο μιας εικόνας σε φυσική γλώσσα, χρησιμοποιώντας ένα σύστημα οπτικής κατανόησης μαζί με ένα γλωσσικό μοντέλο που μπορεί να παράγει συντακτικά σωστές και νοηματικά ουσιαστικές προτάσεις για τη δεδομένη έννοια.

Ο κύριος στόχος αυτού του πεδίου είναι να βρεθεί η πιο αποτελεσματική διαδικασία η οποία μπορεί να κατανοήσει τις οπτικές πληροφορίες που περιέχει η εικόνα, να τις αναπαραστήσει και να τις μετατρέψει σε μια ακολουθία κειμένου που να αποτυπώνει τις συνδέσεις μεταξύ των οπτικών και κειμενικών εννοιών.

Τα τελευταία χρόνια, δύο μοντέλα έχουν διακριθεί σε αυτό το τομέα. Το πρώτο είναι το Flamingo [1], ένα μοντέλο που λειτουργεί ενσωματώνοντας εναλλάξ εικόνα και κείμενο. Αυτό επιτρέπει στο μοντέλο να παράγει περιγραφές με μία καθοδηγούμενη προτροπή (prompt), συνδυάζοντας τα στοιχεία της εικόνας με εκείνα της προτροπής σε μία ενιαία ακολουθία. Το δεύτερο μοντέλο είναι το BLIP-2 [50], το οποίο χρησιμοποιεί ένα επιπλέον transformer μοντέλο (που ονομάζεται Querying Transformer) για να συνδυάσει τις οπτικές και κειμενικές τροπικότητες των υποκείμενων μοντέλων, επιτρέποντας έτσι την παραγωγή περιγραφών μέσω της μεταφοράς πληροφοριών από το οπτικό μοντέλο στο γλωσσικό μοντέλο. Αυτή η διαδικασία ακολουθεί το παράδειγμα των θεμελιωδών μοντέλων, καθώς τμήματα των μεγάλων προεκπαιδευμένων μοντέλων μπορούν να χρησιμοποιηθούν ως κομμάτια σε αρχιτεκτονικές για διάφορες εργασίες, με μόνο ένα μικρό κομμάτι (πχ Quering Transformer) να προστίθεται για να επιτρέψει την αλληλεπίδραση μεταξύ των δύο. Επιπλέον, αξίζει να αναφερθούμε και στο LLaVA [55], το οποίο συνδέει έναν κωδικοποιητή όρασης και ένα μεγάλο γλωσσικό μοντέλο και τα εκπαιδεύει με προσαρμογή οδηγιών (instruction tuning) πάνω σε παραγόμενα δεδομένα. Το LLaVA χρησιμοποιείται για τη γενικού σκοπού κατανόηση γλώσσας και όρασης και παρουσιάζει εξαιρετικές ικανότητες πολυτροπικής συνομιλίας.

Απάντηση Οπτικών Ερωτήσεων

Η Απάντηση Οπτικών Ερωτήσεων (Visual Question Answering - VQA) είναι ένα διεπιστημονικό πεδίο που συνδυάζει την υπολογιστική όραση και την επεξεργασία φυσικής γλώσσας (NLP) για τη δημιουργία συστημάτων ικανών να απαντούν σε ερωτήσεις σχετικά με εικόνες. Ο στόχος του VQA είναι να επιτρέψει στις μηχανές να κατανοούν και να συλλογίζονται σχετικά με το οπτικό περιεχόμενο μιας εικόνας, με τρόπο που να προσεγγίζει την ανθρώπινη κατανόηση. Αυτό περιλαμβάνει όχι μόνο την αναγνώριση αντικειμένων και σκηνών μέσα σε μια εικόνα, αλλά και την ερμηνεία αυτού του οπτικού πληροφοριακού περιεχομένου ώστε να παρέχουν ακριβείς απαντήσεις σε κειμενικές ερωτήσεις.

Definition: Απάντηση Οπτικών Ερωτήσεων (VQA)

> *Απάντηση Οπτικών Ερωτήσεων* είναι η διαδικασία της ανάπτυξης συστημάτων τεχνητής νοημοσύνης που μπορούν να παρέχουν ακριβείς και σχετικές απαντήσεις σε ερωτήσεις που τίθενται σχετικά με εικόνες.

Η ανάπτυξη των συστημάτων VQA έχει υποκινηθεί από τις προόδους στη βαθιά μάθηση, ιδιαίτερα στα συνελικτικά νευρωνικά δίκτυα (CNNs) για την επεξεργασία εικόνων και τα αναδραστικά νευρωνικά δίκτυα (RNNs) ή τους transformers για τη διαχείριση του κειμένου. Τα πρώτα μοντέλα VQA βασίζονταν κυρίως στην εξαγωγή χαρακτηριστικών από εικόνες και ερωτήσεις ανεξάρτητα, πριν συγχωνεύσουν αυτές τις πληροφορίες για να παράγουν απαντήσεις. Με τον καιρό, αναπτύχθηκαν πιο εξελιγμένες αρχιτεκτονικές οι οποίες χρησιμοποιούν μηχανισμούς προσοχής, πολυτροπικά ενσωματώματα και στρατηγικές από κοινού εκπαίδευσης για να καταγράψουν καλύτερα τις περίπλοκες σχέσεις μεταξύ οπτικών και κειμενικών δεδομένων.

Αρχιτεκτονική Συστημάτων VQA

Ένα σύστημα VQA αποτελείται από διάφορα στάδια: την **εξαγωγή χαρακτηριστικών εικόνας**, τη **κωδικοποίηση ερωτήματος**, τη **πολυτροπική συγχώνευση** και τη **πρόβλεψη απάντησης**. Το πρώτο βήμα σε ένα σύστημα VQA περιλαμβάνει την εξαγωγή σημαντικών χαρακτηριστικών από την εισαγόμενη εικόνα. Αυτό επιτυγχάνεται συνήθως με τη χρήση προκαθορισμένων κωδικοποιητών εικόνας, όπως τα CNNs (ResNet[34], VGG [76]) ή οι Transformers (ViT [22]), τα οποία παράγουν μια πλούσια αναπαράσταση της εικόνας. Επιπλέον, το κειμενικό ερώτημα επεξεργάζεται για να δημιουργηθεί μια αναπαράσταση χαρακτηριστικών που να συλλαμβάνει τη σημασιολογική του έννοια. Πολλές εργασίες χρησιμοποιούν RNNs (όπως τα LSTMs [39], GRUs [16]) ή μοντέλα βασισμένα στους transformers (όπως το BERT [20]), τα οποία μπορούν να κωδικοποιήσουν το ερώτημα σε ένα σταθερού μήκους διάνυσμα ή μια ακολουθία ενσωματώσεων. Στη συνέχεια, η βασική πρόκληση στο VQA είναι η αποτελεσματική συνένωση των οπτικών και κειμενικών χαρακτηριστικών για να δημιουργηθεί μια συνεκτική αναπαράσταση που μπορεί να χρησιμοποιηθεί για την εξαγωγή της απάντησης. Έχουν προταθεί διάφορες τεχνικές συγχώνευσης, συμπεριλαμβανομένων των:

- **Συνένωση**: Άμεση συνένωση των διανυσμάτων χαρακτηριστικών από την εικόνα και το ερώτημα.
- **Στοιχειώδης Πρόσθεση/Πολλαπλασιασμός**: Εκτέλεση στοιχειωδών πράξεων για τη συγχώνευση των χαρακτηριστικών.
- **Μηχανισμοί Προσοχής**: Εφαρμογή προσοχής για την εστίαση στα σημαντικά μέρη της εικόνας με βάση το ερώτημα, ενισχύοντας την αλληλεπίδραση μεταξύ των τρόπων.
- **Πολυτροπικοί Transformers**: Χρήση αρχιτεκτονικών transformers που επεξεργάζονται και ενσωματώνουν ταυτόχρονα οπτικές και κειμενικές πληροφορίες.

Τέλος, η συγχωνευμένη πολυτροπική αναπαράσταση περνάει μέσα από ένα ή περισσότερα πλήρως συνδεδεμένα στρώματα ενός νευρωνικού δικτύου για την πρόβλεψη της απάντησης. Το στρώμα εξόδου χρησιμοποιεί συνήθως μια συνάρτηση ενεργοποίησης softmax για να παράγει μια κατανομή πιθανότητας σε ένα προκαθορισμένο σύνολο πιθανών απαντήσεων.

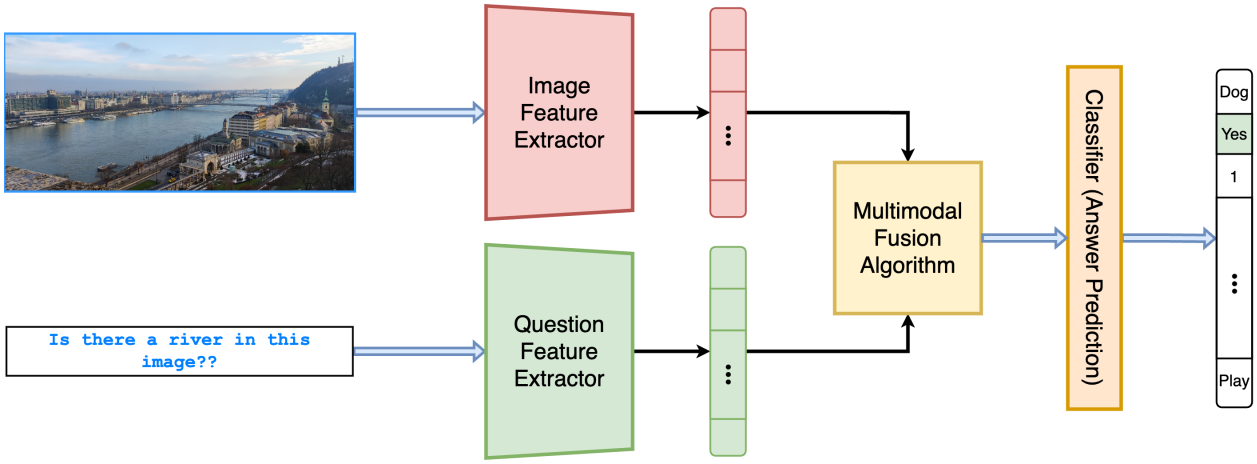


Figure 0.0.11: Μια τυπική **αρχιτεκτονική VQA**.

Προτεινόμενες Μέθοδοι

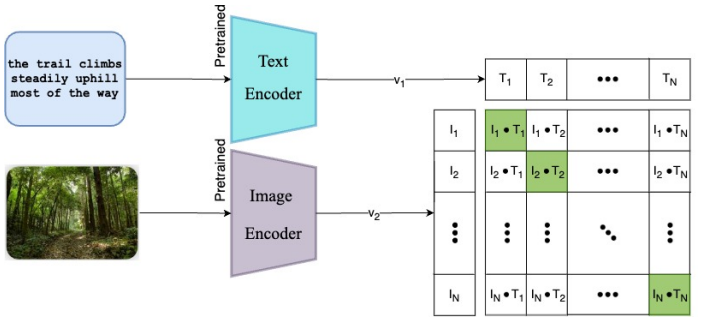
Τα μοντέλα εικόνας-κειμένου έχουν γίνει θεμελιώδη στη μηχανική μάθηση, δημιουργώντας αρκετές sota αρχιτεκτονικές, όπως τα CLIP, DALL-E και Stable Diffusion, μεταξύ άλλων [64, 67, 68, 71]. Αυτά τα θεμελιώδη μοντέλα μπορούν να χρησιμοποιηθούν σε μια ποικιλία εργασιών, συνήθως διαφορετικών από εκείνες για τις οποίες εκπαιδεύτηκαν. Αυτό συμβαίνει επειδή χρησιμοποιούν την κάθε τροπικότητα για να εξάγουν γνώση για την προηγούμενη, επιτρέποντάς τους έτσι να λειτουργούν χωρίς να έχουν δει δεδομένα για τη συγκεκριμένη εργασία. Ταυτόχρονα, αυτές οι αρχιτεκτονικές είναι επίσης πολύ κοστοβόρες για να εκπαιδευτούν. Η εκπαίδευση γίνεται συχνά σε εκατομμύρια ζεύγη εικόνας-κειμένου. Ως εκ τούτου, για τη χρήση αυτών των μεγάλων θεμελιωδών μοντέλων συχνά γίνεται ζήτημα η **λεπτή προσαρμογή (fine-tuning)** τους, αντί να τα εκπαιδεύουμε από την αρχή. Αυτά τα μοντέλα μπορούν επίσης να χρησιμοποιηθούν απλά για εξαγωγή πληροφορίας και να χρησιμοποιηθούν ως το σημείο εκκίνησης γύρω από το οποίο μπορεί να κατασκευαστεί ένα μεγαλύτερο μοντέλο.

Ενίσχυση του CLIP με Διάλογο: Προσέγγιση με Τρίτο Κωδικοποιητή

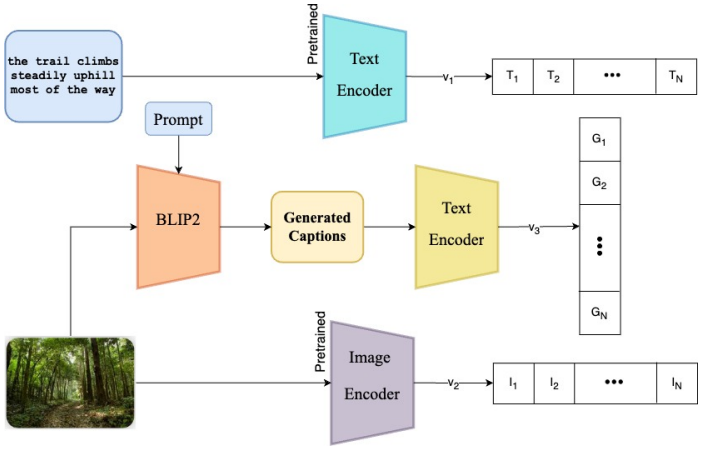
Το κίνητρο πίσω από τη δουλειά μας προέρχεται από την υπολογιστική όραση και τη χρήση επιπρόσθετων προβολών [51, 94] του ίδιου αντικειμένου για την απόκτηση περισσότερων πληροφοριών σχετικά με τα χαρακτηριστικά του. Αυτή η ιδέα προσαρμόστηκε στη μάθηση με αντιδιαστολή από το CMC [84], όπου χρησιμοποιήθηκαν επιπρόσθετες αισθητηριακές προβολές της ίδιας εικόνας για την ενίσχυση της αντιθετικής εκπαίδευσης. Με τον ίδιο τρόπο, για τη περίπτωση των ζευγών εικόνας-κειμένου βρίσκουμε μια επιπρόσθετη προβολή στους διαλόγους που δημιουργούνται από παραγωγικά μοντέλα. Αυτά τα επιπρόσθετα συνθετικά δεδομένα θα είναι σε μορφή κειμένου, αλλά διαφορετικής φύσης σε σύγκριση με τις λεζάντες που χρησιμοποιήθηκαν από τον κωδικοποιητή κειμένου κατά την εκπαίδευση του CLIP. Έτσι, θεωρούμε αυτά τα βοηθητικά σύνολα

δεδομένων κειμένου, όπως τα μεταδεδομένα, τους διαλόγους σχετικά με μια εικόνα ή τις λεπτομέρειες προϊόντων σε μια βάση δεδομένων ως επιπρόσθετες τροπικότητες.

Σε αυτό το πλαίσιο, εξετάζουμε τη χρήση ενός **τρίτου κωδικοποιητή** σε αυτές τις αρχιτεκτονικές εικόνας-κειμένου. Η προσθήκη του τρίτου μπορεί να φανεί στο Σχήμα 0.0.12. Εστιάζουμε τη μελέτη μας σε ένα μοντέλο CLIP εκπαιδευμένο από το OpenCLIP [41], το οποίο ενισχύουμε μέσω ενός επιπρόσθετου κωδικοποιητή, με αποτέλεσμα την αρχιτεκτονική **CLIP-3Modal** [86]. Αυτός ο κωδικοποιητής χρησιμεύει στη ενσωμάτωση επιπρόσθετων τροπικοτήτων στην είσοδο, επαναχρησιμοποιώντας στοιχεία από θεμελιώδη μοντέλα. Σε αντίθεση με προηγούμενες δουλειές που εξετάζουν τη χρήση ενός τρίτου κωδικοποιητή στην αρχιτεκτονική [44], θεωρούμε ρητά τον τρίτο κωδικοποιητή μας ως λειτουργούντα σε μια διαφορετική τροπικότητα, πέρα από τις συνήθεις εικόνα και κείμενο. Στη δική μας περίπτωση, θεωρούμε την επιπρόσθετη τροπικότητα ως το διάλογο ενός χρήστη με ένα μοντέλο λεκτικής περιγραφής εικόνων όπως το BLIP-2 [50]. Με αυτόν τον τρόπο, στοχεύουμε να ενισχύσουμε τις πληροφορίες από τα δεδομένα εισόδου μας με τις εξόδους ενός θεμελιώδους μοντέλου.



(a) CLIP



(b) CLIP-3Modal

Figure 0.0.12: **Η προτεινόμενη αρχιτεκτονική μας CLIP-3Modal.** Προτείνουμε την ενσωμάτωση ενός τρίτου κωδικοποιητή στο μοντέλο CLIP, ο οποίος επεκτείνει τους υπάρχοντες κωδικοποιητές εικόνας και κειμένου. Αυτός ο επιπρόσθετος κωδικοποιητής μπορεί να χρησιμοποιηθεί κατά την αξιολόγηση του μοντέλου, μαζί με τους υπάρχοντες κωδικοποιητές.

Χρήση του BLIP-2 για Λεκτική Περιγραφή

Για να ενσωματώσουμε τον τρίτο πύργο στην αρχιτεκτονική μας, πρέπει να συμπεριλάβουμε μια τρίτη τροπικότητα στα δεδομένα εισόδου μας. Για να το κάνουμε αυτό, χρησιμοποιούμε το BLIP-2 [50] έτσι ώστε να επεκτείνουμε ένα σύνολο δεδομένων εικόνας-κειμένου με μια επιπρόσθετη τροπικότητα. Επιλέγουμε το CC3M [74] ως το σύνολο δεδομένων που εμπλουτίζουμε με λεκτικλες περιγραφές που δημιουργούνται από το BLIP-2. Έτσι, παρέχουμε κάθε εικόνα ως είσοδο στο μοντέλο BLIP-2. Στη συνέχεια, δίνουμε τις ακόλουθες δύο ερωτήσεις, με τη σειρά, στο μοντέλο:

"Τι βλέπεις σε αυτήν την εικόνα;"

"Τι κάνει αυτήν την εικόνα μοναδική;"

Αυτό μπορεί επίσης να φανεί στο Σχήμα 0.0.13. Αυτό το ζευγάρι ερωτήσεων παρέχει μια βασική μορφή διαλόγου μεταξύ ενός χρήστη και του μοντέλου. Επιπλέον, όταν παρέχουμε τη δεύτερη ερώτηση στο μοντέλο λεκτικών περιγραφών, του δίνουμε επίσης και την απάντηση της πρώτης ερώτησης ως είσοδο. Αυτό επιτρέπει στην απάντηση της δεύτερης ερώτησης να εκμεταλλευτεί τα συμφραζόμενα από την απάντηση του μοντέλου στην πρώτη ερώτηση.

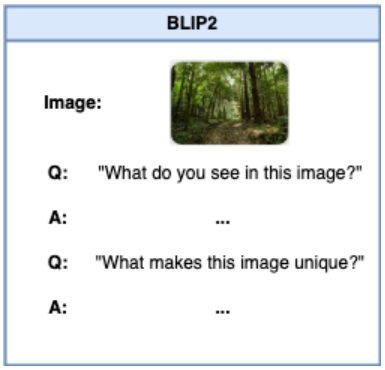


Figure 0.0.13: **Χρήση του μοντέλου BLIP-2.** Οι ερωτήσεις με τις οποίες προτρέπεται το μοντέλο παρέχουν μια βασική μορφή διαλόγου, η οποία χρησιμοποιείται ως μια τρίτη τροπικότητα.

Σημειώνουμε εδώ ότι το μοντέλο που δημιουργεί την τρίτη τροπικότητα δεν λαμβάνει ως είσοδο το κείμενο από τα ζεύγη εικόνας-κειμένου. Αυτό σημαίνει ότι, παρόλου που η τρίτη τροπικότητα είναι και αυτή σε μορφή κειμένου, το πραγματικό περιεχόμενο δεν εξαρτάται άμεσα από την υπάρχουσα λεζάντα της εικόνας. Αυτό επιτρέπει στο μοντέλο BLIP-2 να παρέχει νέες πληροφορίες για αυτό το συγκεκριμένο δείγμα.

Εκπαίδευση του Τρίτου Κωδικοποιητή

Τώρα στοχεύουμε να εκπαιδεύσουμε μια αρχιτεκτονική CLIP που ενσωματώνει έναν τρίτο κωδικοποιητή στην αρχιρτεκτονική της, χρησιμοποιώντας περιγραφές παραγώμενες από το BLIP-2 ως είσοδο. Για να το κάνουμε αυτό, ξεκινάμε από ένα προεκπαιδευμένο μοντέλο CLIP που παρέχεται από το OpenCLIP. Οι κωδικοποιητές εικόνας και κειμένου από αυτό το μοντέλο γίνονται οι κωδικοποιητές εικόνας και κειμένου για την αρχιτεκτονική μας. Για να κατασκευάσουμε τον τρίτο κωδικοποιητή της αρχιτεκτονικής μας, ξεκινάμε κάνοντας ένα αντίγραφο του προεκπαιδευμένου κωδικοποιητή κειμένου. Στη συνέχεια παγώνουμε τους αρχικούς κωδικοποιητές εικόνας και κειμένου, και εκπαιδεύουμε μόνο τον τρίτο κωδικοποιητή στο εκτεταμένο σύνολο δεδομένων CC3M. Η συνάρτηση απώλειας μας είναι παρόμοια με αυτή που χρησιμοποιείται στην κανονική εκπαίδευση του CLIP:

$$\begin{aligned}\mathcal{L} = &-\frac{a}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top G(z_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top G(z_i)/\tau)} \\ &-\frac{(1-a)}{N}\sum_{i=1}^{N}\log\frac{\exp(T(y_i)^\top G(z_i)/\tau)}{\sum_{j=1}^{N}\exp(T(y_j)^\top G(z_i)/\tau)},\ a\in[0,1]\end{aligned} \tag{0.0.1}$$

όπου (x_i, y_i, z_i) είναι ένα από τα δείγματά μας και I, T, G είναι οι κωδικοποιητές εικόνας, κειμένου και παραγόμενου διαλόγου αντίστοιχα. Στο παραπάνω, το a είναι μια παράμετρος ανάμειξης μεταξύ των δύο απωλειών. Θεωρητικά, θέλουμε το a να είναι αρκετά υψηλό για να ενθαρρύνει τη σωστή συμπεριφορά της αναπαράστασης της παραγόμενης περιγραφής σε σχέση με την αντίστοιχη αναπαράσταση της εικόνας. Ταυτόχρονα, δεν θέλουμε μια πολύ υψηλή τιμή του a, καθώς αυτό θα οδηγήσει απλώς στην αντικατάσταση του προεκπαιδευμένου κωδικοποιητή κειμένου με τον τρίτο κωδικοποιητή. Η προσεκτική ανάθεση της παραμέτρου a μπορεί να οδηγήσει τον τρίτο κωδικοποιητή στο να λαμβάνει υπόψη και τις δύο αρχικές τροπικότητες.

Ενίσχυση του CLIP με Διάλογο: Προσέγγιση με Προσαρμογή Τομέα

Το CLIP-3Modal [86] παρείχε κάποια πολύ καλά αποτελέσματα, καθώς βελτίωσε τα αποτελέσματα σε αξιολογήσεις με zero-shot retrieval του βασικού μοντέλου CLIP (δείτε Αξιολόγηση του Τρίτου Κωδικοποιητή). Παρά την επιτυχία του στην ενίσχυση του CLIP με περισσότερη γνώση μέσω καινούριων παραγόμενων κειμενικών σημάτων, το CLIP-3Modal έδειξε πολύ χαμηλή απόδοση στην απάντηση οπτικών ερωτήσεων. Πιστεύουμε ότι οι κατανομές των λεζαντών του CC3M και του παραγόμενου διαλόγου έχουν μια σημαντική μετατόπιση και η αρχιτεκτονική με τον τρίτο κωδικοποιητή, μαζί με την απώλεια αντιδιαστολής, δεν καταφέρνουν να προσαρμοστούν σε αυτήν. Σε αυτό το πλαίσιο, εγκαταλείπουμε την προσέγγιση με τον τρίτο πύργο και διατηρούμε την αρχιτεκτονική του CLIP όπως είναι. Ο κύριός μας στόχος τώρα είναι να προσαρμόσουμε το CLIP σε διάλογο (χωρίς να βλάψουμε τη γενίκευση) ξεκινώντας από κειμενικές περιγραφές. Προτείνουμε το **DRAFT (Dual Representation Adaptive Fine-Tuning)** ως μια νέα τεχνική προσαρμογής αρχιτεκτονικών παρόμοιων με το CLIP σε κειμενικές εισόδους τύπου διαλόγου.

Όλοι πρέπει να συμφωνούν ότι το *Κείμενο* και ο *Διάλογος* είναι και τα δύο μέρη της Επεξεργασίας Φυσικής Γλώσσας, αλλά είναι διαφορετικά. Συγκεκριμένα, ο διάλογος είναι πιο δυναμικός από το κείμενο. Η χρονική διάταξη στον διάλογο είναι κρίσιμη και πολύ σημαντική, αλλά όχι στο κείμενο, το οποίο είναι στατικό και όχι τόσο ακολουθιακό. Επιπλέον, ο διάλογος έχει αλληλεπίδραση. Αυτή η αλληλεπίδραση επιβάλλει στόχους και μια ακολουθία, που όμως δεν υπάρχουν στο κείμενο. Τέλος, ο διάλογος μπορεί να είναι περιορισμένος στο περιεχόμενο, αλλά είναι πλουσιότερος στα συμφραζόμενα από το απλό κείμενο. Αυτές οι ιδιότητες μπορούν να είναι ορατές στα παραδείγματα που παρουσιάζονται στο Σχήμα 0.0.15.

Θεωρούμε αυτό το πρόβλημα ως ένα πρόβλημα προσαρμογής τομέα, όπου το *πηγαίο* πεδίο είναι οι *Λεζάντες* και το πεδίο *στόχος* είναι ο *Διάλογος*. Οι περισσότερες δουλειές που ασχολούνται με τη προσαρμογή τομέα για το CLIP εστιάζουν σε μετατοπίσεις στις κατανομές των εικόνων, κάτι που καθιστά το πρόβλημά μας αρκετά μοναδικό. Για παράδειγμα, το CLIPood [75] προσπαθεί μέσω του κειμένου να γενικεύσει σε εικόνες εκτός κατανομής εκμεταλλευόμενο τις σημασιολογικές σχέσεις μεταξύ των κατηγοριών. Επιπλέον, χωρίς τη χρήση του τρίτου κωδικοποιητή, εστιάζουμε τη δουλειά μας στη λεπτομερή προσαρμογή του υπάρχοντος κωδικοποιητή κειμένου τόσο στις λεζάντες όσο και στις εισόδους διαλόγου. Με αυτό το τρόπο, προσπαθούμε να προσαρμόσουμε τον κωδικοποιητή κειμένου στα δεδομένα διαλόγου χωρίς να βλάψουμε την απόδοση του προεκπαιδευμένου μοντέλου σε πρβλήματα εντός κατανομής.

Χρήση LLaVA για την παραγωγή Ερωτήσεων

Ακολουθώντας τη δουλειά μας στο CLIP-3Modal, χρειάζεται να περιλάβουμε μια τρίτη τροπικότητα στα δεδομένα εισόδου μας. Ωστόσο, αντί να χρησιμοποιούμε το BLIP-2, τώρα χρησιμοποιούμε το LLaVA v1.5 [55, 54] έτσι ώστε να επεκτείνουμε το υπάρχον σύνολο δεδομένων CC3M [74] δημιουργώντας πολλαπλές ερωτήσεις μαζί με τις απαντήσεις τους για κάθε είσοδο. Πιο συγκεκριμένα, παρέχουμε κάθε εικόνα ξεχωριστά ως είσοδο στο μοντέλο LLaVA και στη συνέχεια του ζητάμε να παράξει 3 θεμελιώδεις ερωτήσεις σχετικά με αυτήν, όπως φαίνεται στο Σχήμα 0.0.14. Αυτά τα δείγματα διαλόγου φαίνεται να παρέχουν καλύτερες περιγραφές για την εικόνα εισόδου από τις υπάρχουσες λεζάντες και παράλληλα προσθέτουν νέες πληροφορίες από αυτές που ήδη υπάρχουν. Έχουμε μια αναλογία 1:3 μεταξύ λεζαντών και διαλόγων, οπότε με την επέκτασή μας το σύνολο εκπαίδευσης τριπλασιάστηκε σε μέγεθος. Ορισμένα παραδείγματα παρουσιάζονται στο Σχήμα 0.0.15.

Σημειώνουμε πάλι ότι το μοντέλο παραγωγής λεκτικών περιγραφών δεν λαμβάνει την αρχική λεζάντα ως είσοδο κατά τη δημιουργία των παραπάνω ερωτήσεων.

Εκπαίδευση για Προσαρμογή Τομέα με τη μέθοδο DRAFT

Μάθηση με Αντιδιαστολή Τώρα θέλουμε να προσαρμόσουμε ένα προεκπαιδευμένο μοντέλο CLIP σε εισόδους κειμενικής περιγραφής τύπου ερωτήσεων-απαντήσεων. Όπως στην Προσέγγιση με τον Τρίτο Κωδικοποιητή, ξεκινάμε από ένα προεκπαιδευμένο μοντέλο CLIP που παρέχεται από το OpenCLIP. Στη συνέχεια, παγώνουμε τα βάρη του κωδικοποιητή εικόνας ενώ κρατάμε τον κωδικοποιητή κειμένου πλήρως εκπαιδεύσιμο. Εκπαιδεύουμε έτσι μόνο τον κωδικοποιητή κειμένου στο επεκτεταμένο σύνολο δεδομένων μας. Παρέχουμε τόσο τις λεζάντες όσο και τα ζεύγη ερωτήσεων-απαντήσεων ως είσοδο στον κωδικοποιητή κειμένου και την εικόνα ως είσοδο στον παγωμένο κωδικοποιητή εικόνας. Αρχικά, ορίζουμε μια εργασία αντιδιαστολής, παρόμοια με αυτή του CLIP, με συνάρτηση απώλειας:

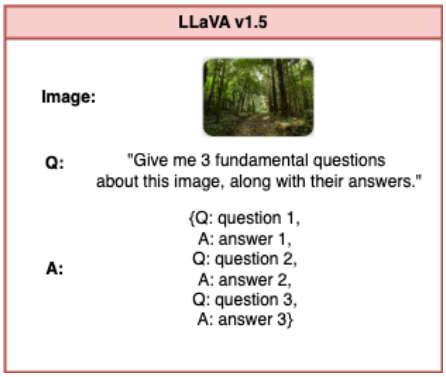


Figure 0.0.14: **Χρήση του μοντέλου LLaVA-1.5**. Τα ζευγάρια ερωτήσεων-απαντήσεων που δημιουργεί το μοντέλο παρέχουν καινούριες κειμενικές υποδείξεις και χρησιμοποιούνται ως τρίτη τροπικότητα στην προσέγγιση προσαρμογής τομέα.

$$\begin{aligned}\mathcal{L}_{contrastive} = &-\frac{a}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top T(z_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top T(z_i)/\tau)} \\ &-\frac{(1-a)}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top T(y_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top T(y_i)/\tau)},\ a\in[0,1]\end{aligned} \tag{0.0.2}$$

όπου (x_i, y_i, z_i) είναι ένα από τα δείγματά μας και I, T είναι οι κωδικοποιητές εικόνας και κειμένου αντίστοιχα. Το πρώτο μέρος είναι η απώλεια αντιδιαστολής μεταξύ της εικόνας και του δείγματος διαλόγου, ενώ το δεύτερο μέρος είναι η απώλεια αντιδιαστολής μεταξύ εικόνας και κειμένου (απώλεια CLIP). Επιπλέον, χρησιμοποιούμε έναν παράγοντα ανάμειξης a για να ισορροπήσουμε τις δύο απώλειες. Φυσικά, θέλουμε ο a να είναι αρκετά υψηλός για να ενθαρρύνουμε το μοντέλο να παράγει χρήσιμες αναπαραστάσεις για τον διάλογο σε αντιστοιχία με τις αναπαραστάσεις των εικόνων. Με τη δεύτερη απώλεια, προσπαθούμε να μην αφήσουμε τον κωδικοποιητή κειμένου να ξεχάσει την προεκπαίδευσή του και να βλάψει τη γενίκευση που έχει ήδη αποκτήσει. Με προσεκτική ανάθεση της υπερπαραμέτρου a, μπορούμε να προσαρμόσουμε το κωδικοποιητή κειμένου σε εισόδους διαλόγου και ταυτόχρονα να διατηρήσουμε τη γενίκευση του προεκπαιδευμένου μοντέλου CLIP.

Μέγιστη Μέση Απόκλιση (MMD) Έχει αποδειχθεί ότι η απώλεια αντιδιαστολής εξασφαλίζει ότι οι κατανομές των 2 τροπικοτήτων που συμμετέχουν στην απώλεια (εικόνα-λεζάντα ή εικόνα-διάλογος) θα ευθυγραμμιστούν (βλ. Ευθυγράμμιση & Ομοιομορφία). Ωστόσο, αυτή η ιδιότητα δεν εγγυάται ότι οι κατανομές των λεζάντων και του διαλόγου θα ευθυγραμμιστούν επίσης. Αποφεύγουμε να χρησιμοποιήσουμε απώλεια αντιδιαστολής μεταξύ τους. Η διαίσθησή μας υποδηλώνει ότι μια επιπλέον απώλεια αντιδιαστολής θα βλάψει τον χώρο των αναπαραστάσεων καθώς μια λεζάντα δεν μοιράζεται απαραίτητα την ίδια σημασιολογική σχέση με το διάλογο. Σε αυτό το πλαίσιο, στόχος μας είναι να ευθυγραμμίσουμε την κατανομή ομοιότητας εικόνας-διαλόγου με την κατανομή ομοιότητας εικόνας-λεζάντας ελαχιστοποιώντας μια απόσταση πάνω σε αυτές τις κατανομές. Συγκεκριμένα, προσπαθούμε να ελαχιστοποιήσουμε τη **Μέγιστη Μέση Απόκλιση (MMD)** (Maximum Mean Discrepancy), η οποία υπολογίζει την απόκλιση των μέσων των κατανομών σε ένα χώρο Hilbert και ευθυγραμμίζει τις οριακές κατανομές των δύο εισόδων. Η απόσταση MMD διατυπώνεται ως εξής:

$$\mathrm{MMD}^2(\mathcal{X}_s, \mathcal{X}_t) = \left\| \frac{1}{n_s}\sum_{i=1}^{n_s}\phi(x_i^s) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi(x_j^t) \right\|_{\mathcal{H}}^2 \tag{0.0.3}$$

όπου $\mathcal{X}_s$ και $\mathcal{X}_t$ είναι τα δείγματα μιας κατανομής πηγής και μιας κατανομής στόχου, αντίστοιχα. n_s και n_t είναι ο αριθμός των αντίστοιχων δειγμάτων και η ϕ είναι η αντιστοιχία των χαρακτηριστικών που ενσωματώνει τα δείγματα σε έναν αναπαραγόμενο χώρο Hilbert (RKHS) $\mathcal{H}$.

Caption: all the details on hosting a party

QA #1: What is the main focus of the image?
Answer: The main focus of the image is a table filled with various snacks, including bags of popcorn and cookies.

QA #2: What is the table setting for?
Answer: The table setting appears to be for a Halloween party, as indicated by the presence of Halloween decorations and the overall festive atmosphere.

QA #3: What is the table covered with?
Answer: The table is covered with a tablecloth.

Caption: # catches a touchdown reception against army during the first half of a college football game .

QA #1: What is the player wearing?
Answer: The player is wearing a football uniform, which typically includes a jersey, pants, cleats, and a helmet.

QA #2: What is the main action in the image?
Answer: The main action in the image is a football player catching a football during a game.

QA #3: Who is the player in the image?
Answer: The player in the image is a football player, specifically a wide receiver.

Figure 0.0.15: **Παραδείγματα του επαυξημένου συνόλου δεδομένων.** Στο πάνω μέρος υπάρχουν εικόνες και κάτω από αυτές οι λεζάντες από το σύνολο δεδομένων CC3M και τα αντίστοιχα ζεύγη ερώτησης-απάντησης που δημιουργήθηκαν από το LLaVA.

Τελική Απώλεια Εκπαίδευσης Στην περίπτωσή μας, στόχος είναι να συνδυάσουμε την διαδικασια μάθησης με αντιδιαστολή και την ελαχιστοποίηση της απόστασης Μέγιστης Μέσης Απόκλισης όπως περιγράψαμε παραπάνω. Η τελική απώλεια διατυπώνεται ως εξής:

$$\begin{aligned}\mathcal{L}_{contrastive} = &-\frac{a}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top T(z_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top T(z_i)/\tau)} \\ &-\frac{(1-a)}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top T(y_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top T(y_i)/\tau)},\ a\in[0,1]\end{aligned}$$

$$\mathcal{L}_{MMD} = \left\| \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi(I(x_j)^\top T(y_i)/\tau) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi(I(x_j)^\top T(z_i)/\tau)\right\|_{\mathcal{H}} \tag{0.0.4}$$

$$\mathcal{L}_{DRAFT} = \mathcal{L}_{contrastive} + \mathcal{L}_{MMD} \tag{0.0.5}$$

όπου (x_i, y_i, z_i) είναι ένα από τα δείγματά μας και I, T είναι οι κωδικοποιητές εικόνας και κειμένου αντίστοιχα. a είναι η παράμετρος ανάμειξης μεταξύ των δύο απωλειών αντιδιαστολής. Τέλος, η ϕ είναι μια αντιστοίχηση χαρακτηριστικών που, στην περίπτωσή μας, είναι μια Συνάρτηση Πυρήνα Ακτινικής Βάσης (RBF) [90]. Μια συνολική παρουσίαση της μεθόδου μας παρουσιάζεται στο Σχήμα 0.0.18.

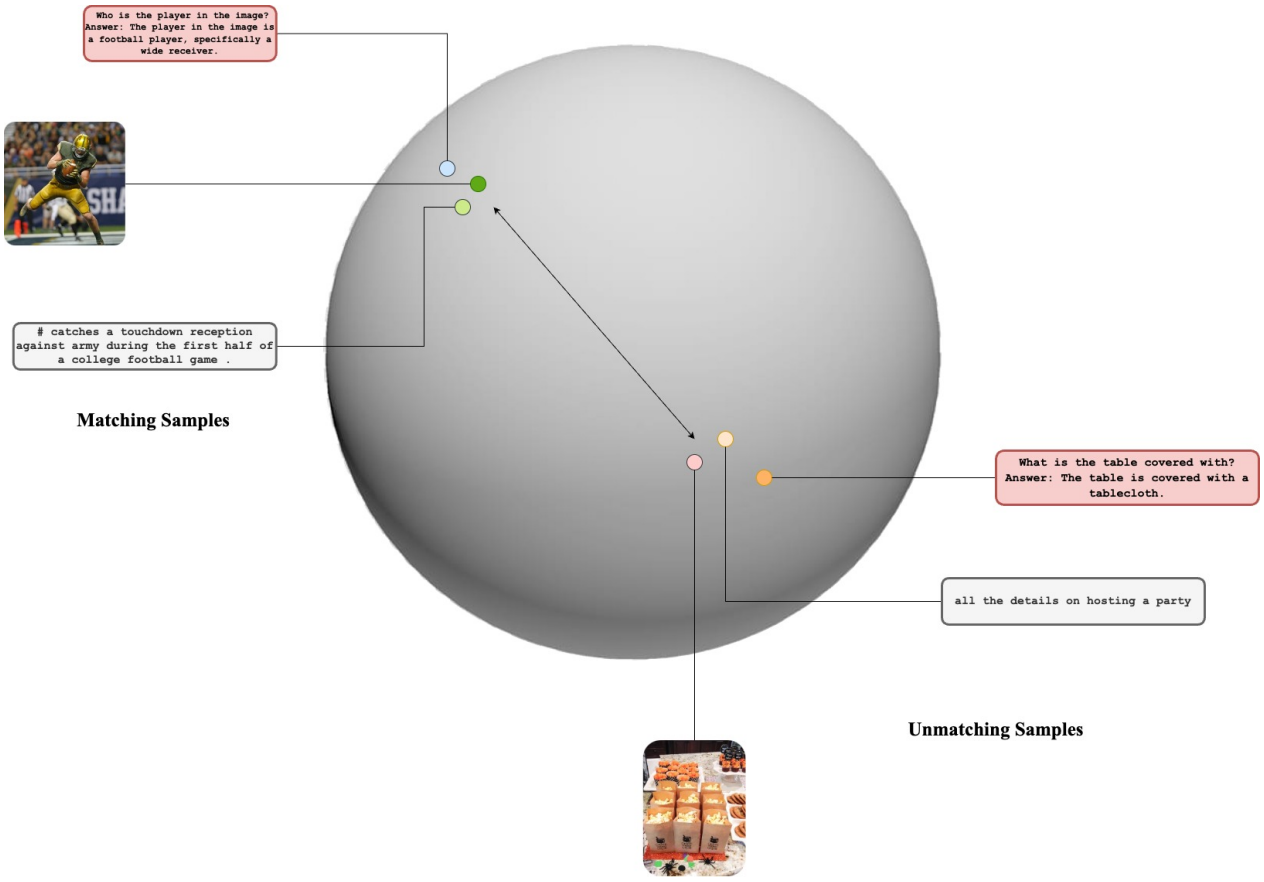


Figure 0.0.16: Οπτική αναπαράσταση της **διαδικασίας μαθησης με αντιδιαστολή**. Τα θετικά ζεύγη (πράσινο, μπλε) συγκεντρώνονται κοντά ενώ τα αρνητικά δείγματα (πορτοκαλί) απομακρύνονται. Στο πρόβλημά μας, οι θέσεις των αναπαραστάσεων εικόνας είναι σταθερές. Στη συνέχεια, προσπαθούμε να ευθυγραμμίσουμε τις αναπαραστάσεις λεζάντας και διαλόγου με αυτές ξεχωριστά.

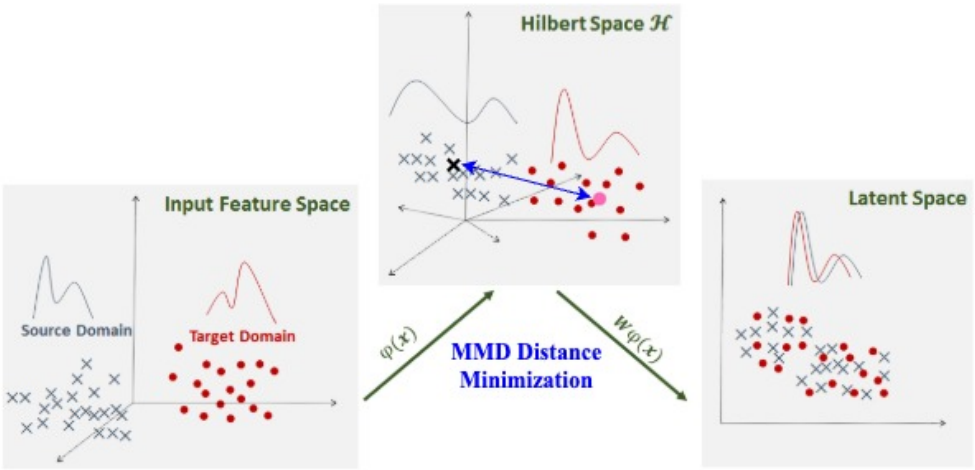


Figure 0.0.17: **Απώλεια Μέγιστης Μέσης Απόκλισης**. Δεδομένων δύο κατανομών, η MMD προσπαθεί να ευθυγραμμίσει τον μέσο κάθε κατανομής. Με αυτόν τον τρόπο, μπορούμε να ενθαρρύνουμε το μοντέλο μας να ευθυγραμμίσει την οριακή κατανομή ομοιότητας εικόνας-λεζάντας με την οριακή κατανομή ομοιότητας εικόνας-διαλόγου.

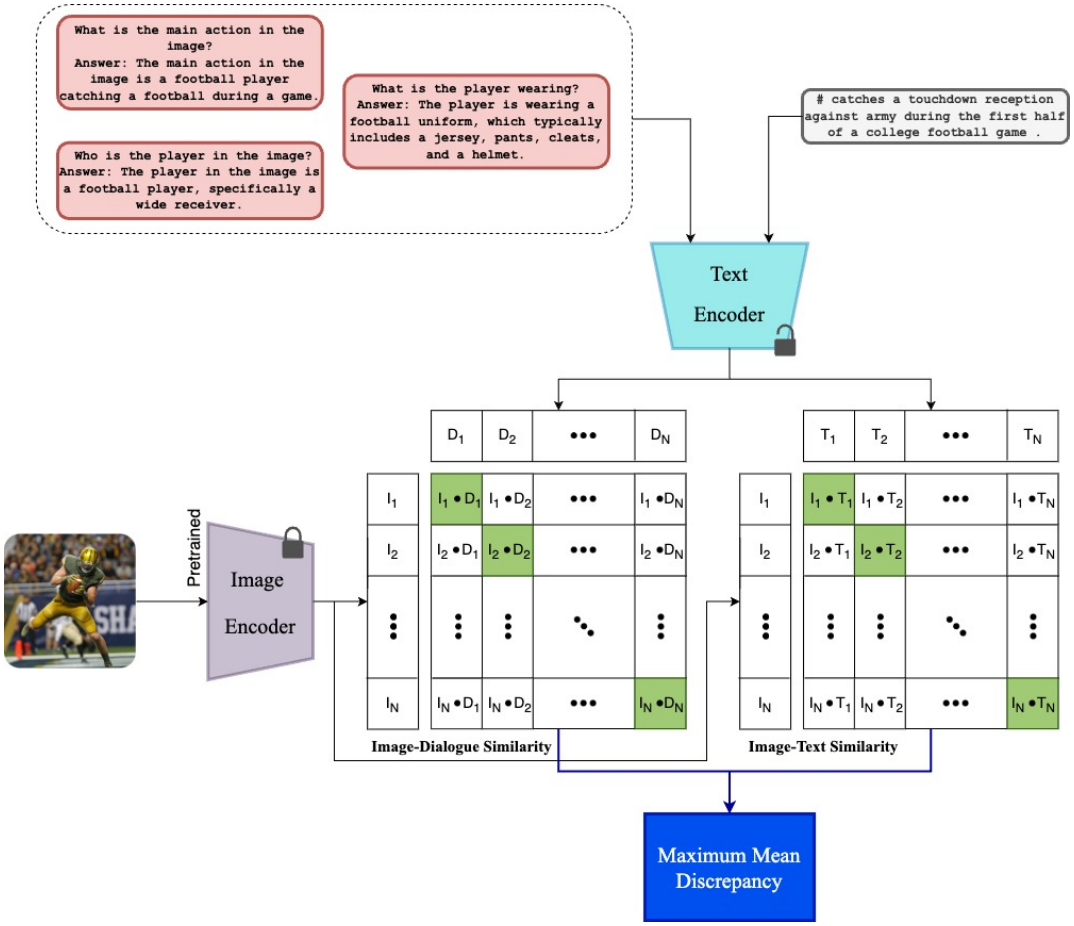


Figure 0.0.18: **DRAFT: Προσαρμογή του CLIP σε εισόδους κειμενικής περιγραφής τύπου ερωτήσεων-απαντήσεων.** Προτείνουμε τη χρήση απώλειας αντιδιαστολής ξεχωριστά για τα ζεύγη των αναπαραστάσεων εικόνας και των δύο κειμένων παραλλαγών μαζί με μια απώλεια απόστασης (MMD) πάνω στις σημασιολογικές κατανομές ομοιότητας εικόνας-λεζάντας και εικόνας-διαλόγου.

Πειραματικά Αποτελέσματα

Σύνολα Δεδομένων

Σε αυτή την ενότητα, θα περιγράψουμε τα σύνολα δεδομένων στα οποία θα αξιολογήσουμε τις μεθόδους μας. Θα χρησιμοποιήσουμε το MSCOCO [53] για πειράματα ανάκτησης χωρίς προσαρμογή και το σύνολο δεδομένων VQAv1 [4] για την αξιολόγηση σε απάντηση οπτικών ερωτήσεων.

MSCOCO

Το Microsoft Common Objects in Context (MSCOCO) [53] dataset είναι ένα μεγάλης κλίμακας σύνολο δεδομένων σχεδιασμένο για εργασίες ανίχνευσης αντικειμένων, κατάτμισης και λεκτικής περιγραφής. Το dataset περιλαμβάνει πάνω από 200.000 εικόνες, περίπου 1,5 εκατομμύριο επισημειώσεις αντικειμένων και 80 κατηγορίες αντικειμένων όπως άνθρωποι, ζώα, οχήματα και καθημερινά αντικείμενα. Παρέχονται ακριβείς κατατμίσεις για κάθε αντικείμενο, μαζί με αναγνωριστικά σημεία για την εκτίμηση της ανθρώπινης πόζας. Κάθε εικόνα συνοδεύεται από πέντε λεζάντες, προσφέροντας ποικίλες περιγραφές.

Το MSCOCO υποστηρίζει αρκετές εργασίες, συμπεριλαμβανομένων της ανίχνευσης αντικειμένων, της κατάτμισης, της εντοπισμού χαρακτηριστικών σημείων, της λεκτικής περιγραφής εικόνων και της πανοπτικής κατάτμισης. Το σύνολο δεδομένων είναι ποικίλο, με εικόνες από μια ευρεία γκάμα καθημερινών σκηνών από όλο τον κόσμο, παρουσιάζοντας αντικείμενα σε φυσικά περιβάλλοντα αντί να είναι απομονωμένα. Αυτό παρέχει ένα ρεαλ-

ιστικό σύνολο δεδομένων για την εκπαίδευση μοντέλων. Οι αναφορές είναι υψηλής ποιότητας και επισημειωμένες από ανθρώπους, εξασφαλίζοντας έτσι ακριβή εκπαίδευση και αξιολόγηση. Η εικόνα 5.1.1 παρουσιάζει κάποια παραδείγματα του dataset MSCOCO.

Figure 0.0.19: **Παραδείγματα του συνόλου δεδομένων MSCOCO.**[53]

VQAv1

Το Visual Question Answering v1 (VQAv1) dataset [4] είναι ένα σύνολο δεδομένων σχεδιασμένο για έρευνα στον τομέα της απάντησης οπτικών ερωτήσεων. Αυτό το σύνολο δεδομένων συνδυάζει το οπτικό περιεχόμενο μιας εικόνας με την επεξεργασία φυσικής γλώσσας, καθιστώντας το ως ένα πολύτιμο εργαλείο για την ανάπτυξη μοντέλων τεχνητής νοημοσύνης που μπορούν να κατανοήσουν και να σκεφτούν πράγματα σχετικά με εικόνες.

Το σύνολο δεδομένων VQAv1 περιλαμβάνει δύο διαφορετικούς τύπους εικόνων: πραγματικές εικόνες από το σύνολο δεδομένων MSCOCO και αφηρημένες σκηνές. Και οι δύο τύποι εξυπηρετούν μοναδικούς σκοπούς στην εκπαίδευση και αξιολόγηση των μοντέλων VQA. Οι πραγματικές εικόνες είναι υψηλής ποιότητας, φυσικές εικόνες που καταγράφουν καθημερινές σκηνές με κοινά αντικείμενα. Είναι πλούσιες σε λεπτομέρειες , παρουσιάζοντας ποικιλία φωτισμού, γωνιών και φόντων. Οι ερωτήσεις για πραγματικές εικόνες απαιτούν από τα μοντέλα να κατανοήσουν περίπλοκες οπτικές πληροφορίες και σχέσεις. Αντίθετα, οι αφηρημένες σκηνές στο σύνολο δεδομένων VQAv1 είναι συνθετικές εικόνες καρτούν που δημιουργήθηκαν χρησιμοποιώντας clip art. Αυτές οι αφηρημένες σκηνές σχεδιάστηκαν για να ελέγχουν συγκεκριμένες πτυχές της οπτικής κατανόησης και σκέψης χωρίς την πολυπλοκότητα των εικόνων του πραγματικού κόσμου. Οι αφηρημένες σκηνές είναι λιγότερο οπτικά πολύπλοκες, με σαφή και διακριτά αντικείμενα και δράσεις. Αυτή η απλότητα επιτρέπει τον πιο εύκολο έλεγχο και την αλληλεπίδραση συγκεκριμένων στοιχείων μέσα στην σκηνή για την δημιουργία στοχευμένων ερωτήσεων. Οι αφηρημένες σκηνές επικεντρώνονται στο να απομονώσουν συγκεκριμένες οπτικές ή νοητικές εργασίες, όπως η κατανόηση των σχέσεων μεταξύ αντικειμένων ή η βασική αναγνώριση αντικειμένων. Παραδείγματα ερωτήσεων για αφηρημένες σκηνές είναι "Τι κάνει το παιδί;" ή "Τι κοιτάει ο άντρας;".

Οι απαντήσεις στο σύνολο δεδομένων VQAv1 ποικίλλουν σε μορφή και τύπο, καλύπτοντας ένα ευρύ φάσμα δυνητικών ερωτήσεων που μπορούν να γίνουν για μια εικόνα. Αυτές οι απαντήσεις χωρίζονται σε αρκετές κατηγορίες: *Απαντήσεις Ναι/Όχι* είναι απλές δυαδικές απαντήσεις σε ερωτήσεις σχετικά με την παρουσία, την κατάσταση ή την ενέργεια που απεικονίζεται στην εικόνα. Οι *Αριθμητικές Απαντήσεις* είναι απαντήσεις σε ερωτήσεις που περιλαμβάνουν την μέτρηση αντικειμένων ή την καθορισμό ποσοτήτων (π.χ. "Πόσα από τα ελάφια κοιμούνται;" και η απάντηση θα μπορούσε να είναι "Ένα"). Οι *Σύντομες Απαντήσεις* είναι σύντομες περιγραφικές απαντήσεις που παρέχουν πληροφορίες σχετικά με διάφορες ιδιότητες, ενέργειες ή αντικείμενα στην εικόνα (π.χ. "Προς πού δείχνει το παιδί;" ("Μαμά"), "Τι άθλημα κάνουν;" ("Μπέιζμπολ")). Οι *Απαντήσεις Ανοιχτού Τύπου* είναι πιο πολύπλοκες απαντήσεις που μπορεί να περιλαμβάνουν σκέψη ή ερμηνεία της σκηνής που απεικονίζεται στην εικόνα. Για παράδειγμα, μπορούμε να έχουμε μια ερώτηση του τύπου "Γιατί τρέχει ο άνθρωπος;" και η απάντηση να είναι "Για να προλάβει το λεωφορείο". Αυτά τα παραδείγματα φαίνονται επίσης στην εικόνα 0.0.20.

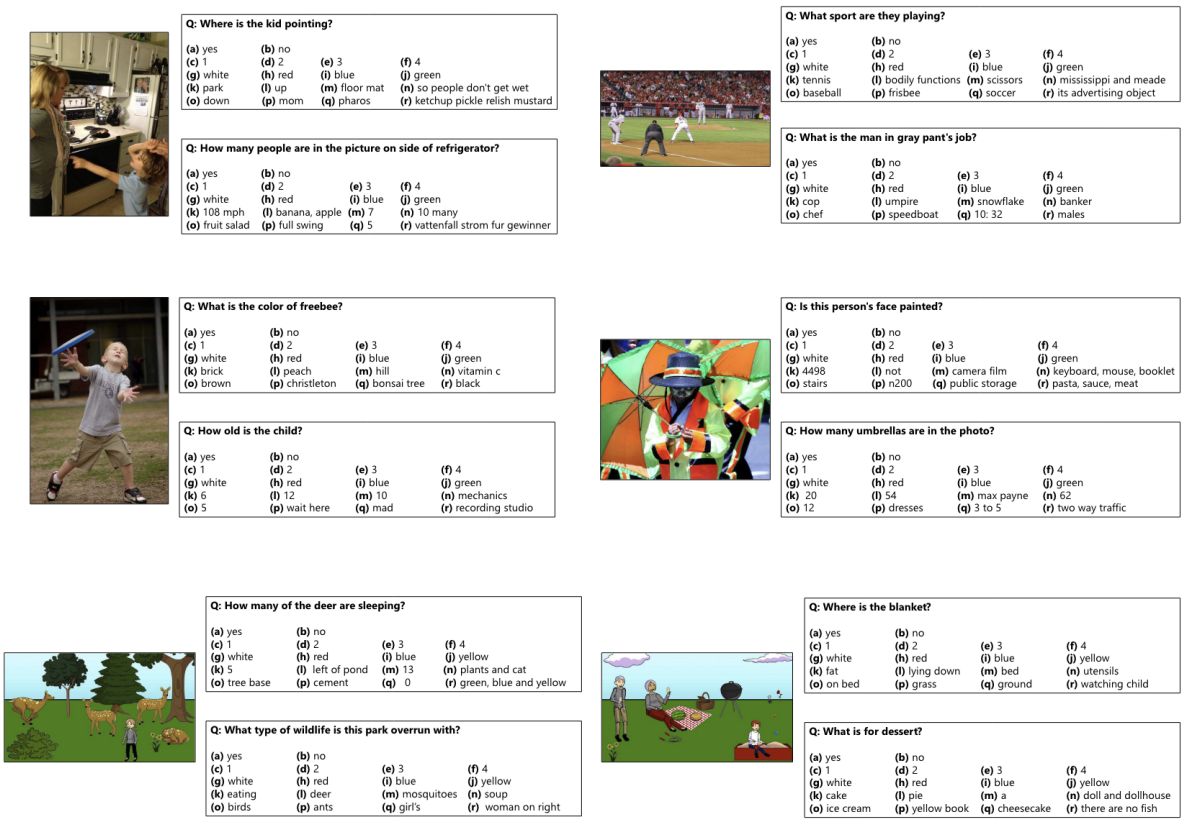


Figure 0.0.20: **Παραδείγματα του σύνολο δεδομένων VQAv1 .** Ερωτήσεις πολλαπλών επιλογών μαζί με τις πιθανές απαντήσεις τους για πραγματικές και αφηρημένες σκηνές.[4]

Αξιολόγηση του Τρίτου Κωδικοποιητή

Μέθοδος Αξιολόγησης

Για την αξιολόγηση του CLIP-3Modal επικεντρωθήκαμε σε εργασίες ανάκτησης εικόνας από κείμενο και το αντίστροφο χωρίς την έκθεση του μοντέλου σε δεδομένα του συνόλου αξιολόγησης. Χρησιμοποιήσαμε διάφορες αρχιτεκτονικές του CLIP που παρέχονται από το OpenCLIP ως βάση για την αξιολόγηση μας. Εκτελέσαμε τα πειράμτά μας σε αρχιτεκτονικές με ViT κωδικοποιητές εικόνας και BERT ή T5 ως κωδικοποιητές κειμένου. Αυτά τα προεκπαιδευμένα μοντέλα εκπαιδεύτηκαν σε 32 δισεκατομμύρια δείγματα του συνόλου δεδομένων LAION-2B [73]. Αυτό μας παρέχει ένα καλό σημείο εκκίνησης για τον τρίτο κωδικοποιητή μας. Για την παράμετρο βάρους a στη συνάρτηση απώλειάς μας, χρησιμοποιούμε $a = 0.65$. Παρατηρήσαμε ότι γενικά το μοντέλο μας πηγαίνει καλύτερα όταν το βάρος της απώλειας μεταξύ εικόνας και παραγόμενων λεκτικών περιγραφών είναι μεγαλύτερο από 0.5.

Μετά την εκπαίδευση του μοντέλου μας, συγχωνεύουμε τις ενσωματώσεις του κειμένου και των παραγόμενων λεκτικών περιγραφών, για να λάβουμε μια τελική ενσωμάτωση για το κείμενο. Για τη συγχώνευση των εξόδων παίρνουμε έναν άθροισμα των ενσωματώσεων που παρέχονται από τους κωδικοποιητές του κειμένου και των παραγόμενων λεκτικών περιγραφών:

$$X_{ensemble} = \beta \cdot T + (1 - \beta) \cdot G, \ \beta \in [0, 1] \tag{0.0.1}$$

όπου T και G είναι τα ενσωματώματα που παράγονται από τους κωδικοποιητές του αρχικού κειμένου και των παραγόμενων λεκτικών περιγραφών. Με την κατάλληλη επιλογή του βάρους, ο τρίτος κωδικοποιητής θα καταφέρει να ενσωματώσει επιπλέον πληροφορίες που ενισχύουν τις προβλέψεις που έχει ήδη μάθει το βασικό μοντέλο. Σε αυτήν την περίπτωση, χρησιμοποιούμε παράμετρο ανάμειξης $\beta = 0.9$. Η ιδέα πίσω από αυτήν την

υψηλή τιμή του β είναι ότι θέλουμε τα ενσωματώματα του τρίτου κωδικοποιητή να επηρεάσουν το αποτέλεσμα, αλλά παράλληλα οι αλλαγές που θα προκαλέσουν να είναι μικρές έστι ώστε να διατηρήσουμε αρκετό μέρος της αρχικής πληροφορίας. Μια μελέτη πάνω στην παράμετρο ανάμειξης παρουσιάζεται στον πίνακα 2.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32	39.7	65.4	75.6	56.3	79.8	87.1
C3M-ViT/B-32	**40.2**	**65.9**	**75.9**	**57.0**	**80.6**	**87.5**
CLIP-ViT/L-14	46.5	71.1	79.8	63.3	84.0	90.8
C3M-ViT/L-14	**46.8**	**71.2**	**80.0**	**63.6**	**84.1**	**91.0**
CLIP-ViT/H-14	**49.4**	**73.4**	**81.5**	**66.0**	86.1	**91.9**
C3M-ViT/H-14	49.1	73.3	**81.5**	**66.0**	**86.2**	**91.9**

Table 1: **Βασικά αποτελέσματα χρησιμοποιώντας το CLIP-3Modal μαζί με το BLIP-2 για την παραγωγή διαλόγου.** Το CLIP-3Modal βελτιώνει την ανάκληση σε σχεδόν κάθε περίπτωση. Η αξιολόγηση γίνεται στο MSCOCO και τόσο οι κωδικοποιητές εικόνας όσο και κειμένου έχουν προεκπαιδευτεί στο LAION-2B Dataset.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline ($\beta = 1$)	39.7	65.4	75.6	56.3	79.8	87.1
$\beta = 0.95$	39.9	65.7	75.8	**57.1**	80.4	87.3
$\beta = 0.90$	**40.2**	**65.9**	**75.9**	57.0	**80.6**	**87.5**
$\beta = 0.80$	40.0	65.6	75.8	56.3	79.6	86.2
$\beta = 0.60$	38.7	63.8	73.8	53.7	75.6	84.8

Table 2: **Οι υψηλές τιμές της παραμέτρου ανάμιξης βελτιώνουν την απόδοση, ενώ οι μικρότερες χειροτερεύουν την επίδοση του μοντέλου, ρίχνοντας τη κάτω από το βασικό μοντέλο**. Το β υποδηλώνει την παράμετρο βάρους ανάμιξης. Το μικρό βάρος στον κωδικοποιητή παραγωγής λεκτικών περιγραφών ωφελεί τις συγχωνευμένες ενσωματώσεις διατηρώντας τις αρχικές πληροφορίες απο το προεκπαιδευμένο μοντέλο και ενισχύοντάς τες με διαφορετικές ιδιαιτερότητες της εισόδου διαλόγου. Τα καλύτερα αποτελέσματα εμφανίστηκαν για $\beta = 0.9$.

Υλοποίηση

Εκπαιδεύσαμε τον τρίτο πύργο όπως περιγράφεται παραπάνω, χρησιμοποιώντας ως βάση διαφορετικά προεκπαιδευμένα μοντέλα CLIP (στο LAION-2B) βασισμένα σε ViTs, για τη κωδικοποίηση των εικόνων (παρέχονται από το OpenCLIP). Η εκπαίδευση του τρίτου κωδικοποιητή στο CLIP-3Modal-ViT/B-32 διαρκεί περίπου 1 ώρα ανά εποχή σε μία GPU, που είναι σημαντικά λιγότερος χρόνος από την εκπαίδευση του βασικού μοντέλου. Βασίσαμε την αξιολόγησή μας στο MSCOCO [53] μελετώντας την απόδοση ανάκτησης χωρίς προσαρμογή σε αυτό το συγκεκριμένο σύνολο δεδομένων, χρησιμοποιώντας την ίδια μετρική αξιολόγησης που χρησιμοποιήθηκε από το OpenCLIP. Καταφέραμε να ξεπεράσουμε το μοντέλο του OpenCLIP στην ανάκτηση τόσο εικόνας όσο και κειμένου χωρίς προσαρμογή, με περιθώριο από 0.3% έως 0.8%. Περισσότερες λεπτομέρειες παρουσιάζονται στον Πίνακα 1.

Αφαιρετικές Μελέτες

Χωρίς Ανάμειξη Προσθύτερα προτείναμε ένα σχήμα ανάμειξης για την αξιολόγηση του CLIP-3Modal στην ανάκτηση χωρίς προσαρμογή. Παρά τα υποσχόμενα αποτελέσματα, θέλουμε να δούμε αν ο προτεινόμενος τρίτος κωδικοποιητής μπορεί να σταθεί μόνος του χωρίς ανάμειξη. Όπως μπορούμε να δούμε στον Πίνακα 3, ο τρίτος κωδικοποιητής δεν αποδίδει πολύ καλά όταν αξιολογείται μόνος του. Αυτό είναι λογικό επειδή η προσαρμογή βλάπτει τη γενίκευση που είχε ο κωδικοποιητής από την προεκπαίδευση του. Για αυτόν τον λόγο, χρησιμοποιούμε τον τρίτο κωδικοποιητή σε συνδυασμό με τον αρχικό κωδικοποιητή κειμένου.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32*	34.2	60.0	70.6	52.4	76.3	84.3
C3M-ViT/B-32^{LLaVA}	**34.9**	**60.7**	**71.2**	**52.9**	**77.1**	**84.7**
C3M-ViT/B-32^{LLaVA} (only 3rd)	27.5	51.3	62.4	39.8	66.9	76.5
CLIP-ViT/L-14	46.5	71.1	79.8	63.3	84.0	90.8
C3M-ViT/L-14	**46.8**	**71.2**	**80.0**	**63.6**	**84.2**	**91.0**
C3M-ViT/L-14 (only 3rd)	35.8	61.4	71.8	44.3	69.8	80.1
CLIP-ViT/H-14	**49.4**	**73.4**	**81.5**	**66.0**	86.1	**91.9**
C3M-ViT/H-14	49.1	73.3	**81.5**	**66.0**	**86.2**	**91.9**
C3M-ViT/H-14 (only 3rd)	19.9	41.3	52.3	31.0	57.0	68.9

Table 3: **Ο τρίτος κωδικοποιητής ωφελείται μόνο όταν λειτουργεί προσθετικά.** Η προσαρμογή βλάπτει τη πληροφορία που έχει μάθει το μοντέλο από την προεκπαίδευση στον τρίτο κωδικοποιητή. Ωστόσο, παρέχει νέες πληροφορίες στον αρχικό κωδικοποιητή κειμένου που μπορούν να παρατηρηθούν μετά τη συγχώνευσή τους. * Προεκπαιδεύτηκε στο LAION-400m.

Γενετικό Μοντέλο Στην Ενότητα προτείναμε ένα νέο μοτίβο παραγωγής για τα συνθετικά δείγματα διαλόγου. Πιο συγκεκριμένα, αντί για το BLIP-2, χρησιμοποιήσαμε το LLaVA v1.5 για την παραγωγή διαλόγου. Επιπλέον, αλλάξαμε την αλληλεπίδραση με το μοντέλο. Αντί να κάνουμε δύο διαδοχικές ερωτήσεις σχετικά με την εικόνα και την μοναδικότητά της (BLIP-2), ζητάμε από το μοντέλο (LLaVA v1.5) να βρει τρεις ερωτήσεις που περιγράφουν την εικόνα και να τις απαντήσει.

Για την αξιολόγηση των νέων εισόδων, ενώνουμε όλα τα ζεύγη ερώτησης-απάντησης. Στη συνέχεια, τα παρέχουμε ως είσοδο στο μοντέλο CLIP-3Modal. Όπως μπορούμε να δούμε στον Πίνακα 4, τα συνθετικά δεδομένα που δημιουργήθηκαν με το LLaVA v1.5 και τη νέα αλληλεπίδραση με το μοντέλο βελτιώνουν την απόδοση του CLIP-3Modal σε κάθε περίπτωση. Στον Πίνακα 5

		Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	Generative Model	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32*	-	34.2	60.0	70.6	52.4	76.3	84.3
C3M-ViT/B-32	LLaVA	**34.9**	**60.7**	**71.2**	**52.9**	**77.1**	**84.7**
CLIP-ViT/H-14	-	49.4	73.4	81.5	66.0	86.1	91.9
C3M-ViT/H-14	BLIP-2	49.1	73.3	81.5	66.0	86.2	91.9
C3M-ViT/H-14	LLaVA	**49.6**	**73.5**	**81.7**	**66.6**	**86.5**	**92.1**

Table 4: **Η αντικατάσταση του BLIP-2 με το LLaVA βελτιώνει την απόδοση τόσο στις περιπτώσεις με όσο και χωρίς ανάμειξη.** * Χρησιμοποιήσαμε το προεκπαιδευμένο μοντέλο στο LAION-400m αντί του LAION-2B. Βλέπουμε ότι το LLaVA και το νέο μοτίβο παραγωγής προσφέρουν καλύτερα αποτελέσματα στην ανάκτηση. Ακόμα και το μοντέλο βασισμένο στο ViT-H-14 υπερτερεί του βασικού μοντέλου, κάτι που δεν επιτεύχθηκε με τα δεδομένα που παράγονται από το BLIP-2.

		Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	Generative Model	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/H-14	-	49.4	73.4	81.5	66.0	86.1	91.9
C3M-ViT/H-14	BLIP-2	19.9	41.3	52.3	31.0	57.0	68.9
C3M-ViT/H-14	LLaVA	30.4	55.5	66.5	46.0	71.7	80.3

Table 5: **Τα δεδομένα που παράγονται με το LLaVA βελτιώνουν επίσης την απόδοση του τρίτου κωδικοποιητή όταν λειτουργεί μόνος του.** Είναι εμφανές ότι τα δεδομένα που παράγονται από το LLaVA και το πρότυπο ερωτήσεων-απαντήσεων βοηθούν το προεκπαιδευμένο κωδικοποιητή κειμένου να διατηρήσει περισσότερες πληροφορίες από την προεκπαίδευσή του.

Αποτελέσματα

Μετά από προσεκτική εξέταση κάθε πειράματος που διεξήγαμε παρουσιάζουμε την καλύτερη καταγεγραμμένη απόδοση κάθε CLIP-3Modal μοντέλου στον Πίνακα 6. Βλέπουμε ότι η μέθοδός μας, ενισχυμένη με τα QA-ζευγάρια που παράγονται από το LLaVA παρέχει, τα καλύτερα αποτελέσματα και ξεπερνά το βασικό μοντέλο CLIP και την παραλλαγή του CLIP-3Modal που χρησιμοποιεί το μοντέλο BLIP-2.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32	39.7	65.4	75.6	56.3	79.8	87.1
C3M-ViT/B-32	**40.2**	**65.9**	**75.9**	**57.0**	**80.6**	**87.5**
CLIP-ViT/L-14	46.5	71.1	79.8	63.3	84.0	90.8
C3M-ViT/L-14	**46.8**	**71.2**	**80.0**	**63.6**	**84.1**	**91.0**
CLIP-ViT/H-14	49.4	73.4	81.5	66.0	86.1	91.9
C3M-ViT/H-14^{LLaVA}	**49.6**	**73.5**	**81.7**	**66.6**	**86.5**	**92.1**

Table 6: **Το CLIP-3Modal βελτιώνει την απόδοση σε κάθε τύπο ανάκτησης με μηδενική προσαρμογή.** Παρακάτω παρουσιάζονται τα μοντέλα με την καλύτερη απόδοση. Βλέπουμε ότι το μοντέλο μας υπερτερεί του βασικού OpenCLIP σε κάθε περίπτωση.

Αξιολόγηση DRAFT

Μέθοδος Αξιολόγησης

Στην Ενότητα , προτείνουμε μια εναλλακτική προσέγγιση που προσαρμόζει το CLIP σε κειμενικές εισόδους με μορφή διαλόγου. Για να παρατηρήσουμε την αποτελεσματικότητα της μεθόδου μας, την αξιολογήσαμε σε εργασίες Απάντησης Οπτικών Ερωτήσεων (Visual Question Answering - VQA) χρησιμοποιώντας το σύνολο δεδομένων VQAv1. Η αξιολόγηση των μοντέλων τύπου CLIP σε τέτοιου είδους εργασίες περιλαμβάνει τη μετατροπή του προβλήματος VQA σε πρόβλημα ανάκτησης με μηδενική προσαρμογή. Συγκεκριμένα, δεδομένης μιας εικόνας, μιας ερώτησης και όλων των N πιθανών απαντήσεων, συνενώνουμε κάθε απάντηση με την ερώτηση της για να δημιουργήσουμε N ζευγάρια ερώτησης-απάντησης. Στη συνέχεια προβλέπουμε το σωστό ζευγάρι εντοπίζοντας ποιο ζευγάρι έχει τη μεγαλύτερη ομοιότητα συνημιτόνου με την εικόνα. Αυτή η διαδικασία απεικονίζεται στο Σχήμα 0.0.21.

Ενώ κάποιοι προτείνουν τη λεπτομερή προσαρμογή του CLIP στο σύνολο εκπαίδευσης πριν από την αξιολόγηση, προτιμούμε να αξιολογούμε τα μοντέλα μας χωρίς καμία έκθεση σε δείγματα του συνόλου δεδομένων VQA. Αυτή η προσέγγιση αξιολογεί καλύτερα την προσαρμογή του μοντέλου χρησιμοποιώντας τα συνθετικά δεδομένα διαλόγου από το LLaVA. Επιπλέον, στο Σχήμα 0.0.21 φαίνεται ότι τα ζευγάρια ερώτησης-απάντησης στο σύνολο δεδομένων αξιολόγησης είναι διαφορετικά από τα εκτενή ζευγάρια ερωτήσεων-απαντήσεων που παράγονται κατά τη διάρκεια της εκπαίδευσης.

Αφού προσαρμόσουμε το μοντέλο μας για εισόδους Ερώτησης-Απάντησης, ακολουθούμε την παραπάνω διαδικασία χρησιμοποιώντας το προσαρμοσμένο κωδικοποιητή κειμένου για να ενσωματώσουμε τα συγχωνευμένα ζευγάρια ερώτησης-απάντησης. Επιπλέον, χρησιμοποιούμε την τεχνική ανάμειξης από το Αξιολόγηση του Τρίτου Κωδικοποιητή, η οποία συνδυάζει τον προσαρμοσμένο κωδικοποιητή κειμένου με τον αρχικό κωδικοποιητή από το βασικό CLIP. Πιστεύουμε ότι η προσέγγιση προσαρμογής τομέα μπορεί να βελτιώσει την απόδοση VQA όταν συνδυάζεται με το βασικό CLIP. Αξιολογήσαμε το μοντέλο σε πραγματικές εικόνες και αφηρημένες σκηνές από το VQAv1.

Αποτελέσματα

Εκπαιδεύσαμε το μοντέλο μας χρησιμοποιώντας μια αρχιτεκτονική CLIP βασισμένη στο ViT-B-32 που παρέχεται από το OpenCLIP. Το βασικό μας μοντέλο εκπαιδεύτηκε στο σύνολο δεδομένων LAION-400M. Οι πειραματισμοί μας διεξήχθησαν είτε με 2 ή με 4 GPU A100, με χρόνους εκπαίδευσης περίπου 16 λεπτά και 9 λεπτά ανά εποχή, αντίστοιχα. Η απόδοση της μεθόδου μας σε αφηρημένες σκηνές VQA φαίνεται στον Πίνακα 7, ενώ τα αποτελέσματα για VQA σε πραγματικές εικόνες παρουσιάζονται στον Πίνακα 8.

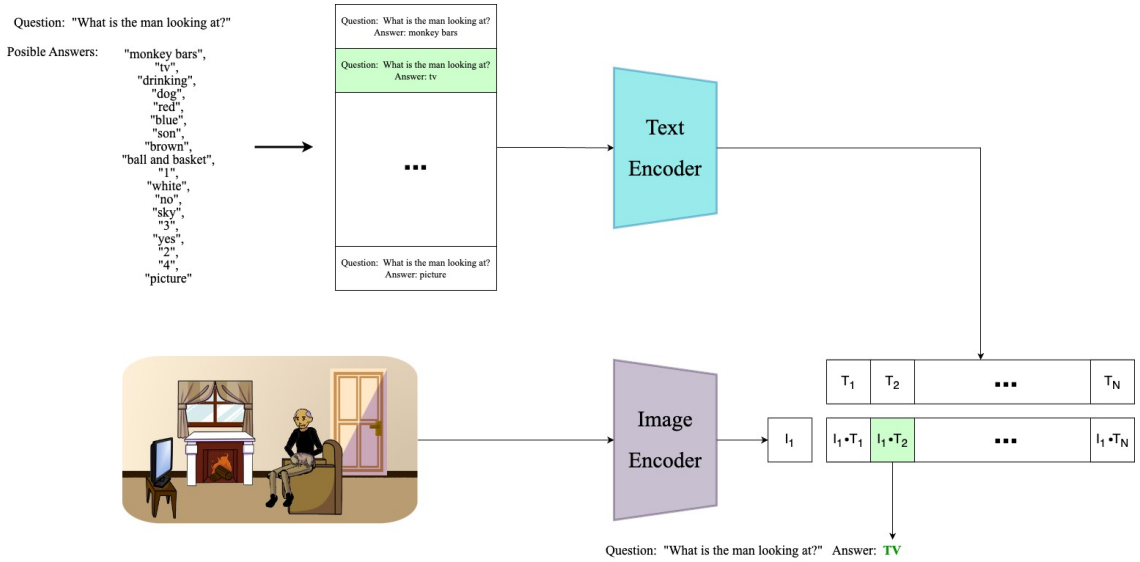


Figure 0.0.21: **Μέθοδος αξιολόγησης VQA για μοντέλα τύπου CLIP**. Συνενώνουμε κάθε πιθανή απάντηση με την ερώτηση και χρησιμοποιούμε την εικόνα για να βρούμε το ζευγάρι ερώτησης-απάντησης πο έχει τη μεγαλύτερη ομοιότητα συνημιτόνου με την εικόμα.

Όπως φαίνεται, ο προσαρμοσμένος κωδικοποιητής κειμένου βελτιώνει σημαντικά τη δυνατότητα του μοντέλου να απαντάει ερωτήσεις βασισμένες σε εικόνες. Αυτά τα αποτελέσματα αντιπροσωπεύουν ένα βήμα προόδου στην ανεπίβλεπτη προσαρμογή των μοντέλων CLIP σε νέες γλωσσικές υποδείξεις. Το μοντέλο μας υπερέχει του αρχικού CLIP σε εργασίες VQA με σημαντικό περιθώριο, χωρίς να έχει δει δείγματα από το σύνολο δεδομένων VQA. Αυτή η βελτίωση τονίζει την συνεισφορά της μεθόδου προσαρμογής μας και των παραγώμενων δεδομένων στην αύξηση της ακρίβειας του CLIP στην απάντηση ερωτήσεων.

Visual Question Answering (VQAv1) *Abstract Scenes*		
Model	Epochs	Accuracy
CLIP-ViT/B-32	-	9.88%
Ours-ViT/B-32	6	<u>10.68%</u>
Ours-ViT/B-32+blend	12	**11.54%**

Table 7: **Η μέθοδός μας κερδίζει το βασικό CLIP στις αφηρημένες σκηνές του VQAv1, με και χωρίς ανάμειξη.** Το βασικό μοντέλο CLIP, που παρέχεται από το OpenCLIP και εκπαιδεύτηκε στο σύνολο δεδομένων LAION-400M, επιτυγχάνει ακρίβεια 10.68%. Το δικό μας μοντέλο αυξάνει την ακρίβεια κατά περίπου 1.7% με την ανάμειξη του προσαρμοσμένου και του αρχικου κωδικοποιητη κειμένου (με $\beta = 0.9$). Χρησιμοποιώντας τη μέθοδο προσαρμογής μας, το μοντέλο φαίνεταινα έχει μια σημαντική αύξηση στην απόδοσή του στο VQA πρόβλημα.

Αφαιρετικές Μελέτες

Απώλεια MMD Για να αξιολογήσουμε την επίδραση της Απώλειας Μέγιστης Μέσης Απόκλισης (MMD), πραγματοποιήσαμε πειράματα αφαιρέσεων εκπαιδεύοντας το μοντέλο μας μόνο με τον αντιθετικό στόχο. Ο πίνακας 9 αναδεικνύει τον κρίσιμο ρόλο της απώλειας MMD στη βελτίωση της απόδοσης του CLIP σε VQA προβλήματα. Τα ευρήματά μας υπογραμμίζουν τη σημασία της ενσωμάτωσης της απώλειας MMD στη διαδικασία προσαρμογής.

Visual Question Answering (VQAv1) *Real Images*		
Model	Epochs	Accuracy
CLIP-ViT/B-32	-	26.2%
Ours-ViT/B-32	2	23.6%
Ours-ViT/B-32+blend	3	**28.3%**

Table 8: **Η μέθοδός μας βελτιώνει την ακρίβεια στις πραγματικές εικόνες του VQAv1 μέσω της ανάμειξης των κωδικοποιητών κειμένου.** Το μοντέλο βασικής σύγκρισης CLIP, που παρέχεται από το OpenCLIP και εκπαιδεύτηκε στο σύνολο δεδομένων LAION-400M, είχε αρχικά ακρίβεια 23.6%. Ωστόσο, με την ανάμειξη του προσαρμοσμένου κωδικοποιητή κειμένου με τον αρχικό κωδικοποιητή κειμένου (με $\beta = 0.9$), βελτιώσαμε σημαντικά την ακρίβεια, στο 28.3%, που υπερτερεί του CLIP κατά 2.1%.

Visual Question Answering (VQAv1) *Abstract Scenes*			
Model	Epochs	Batch Size	Accuracy
CLIP-ViT/B-32	-	-	9.88%
Ours-ViT/B-32	6	1536	10.68%
Ours-ViT/B-32+blend	12	1536	**11.54%**
Ours-ViT/B-32$^{no\ MMD}$	6	6144	9.46%
Ours-ViT/B-32+blend$^{no\ MMD}$	12	6144	10.24%

Table 9: **Η επίδραση της Απώλειας MMD στην απόδοση του προσαρμοσμένου κωδικοποιητή κειμένου.** Η ενσωμάτωση της απώλειας MMD στην εκπαίδευση επιτρέπει στο μοντέλο μας να ξεπεράσει το βασικό μοντέλο CLIP-ViT/B-32 στις αφηρημένες σκηνές του VQAv1. Τα μοντέλα που εκπαιδεύτηκαν χωρίς την απώλεια MMD, παρά τη χρήση μεγαλύτερου μεγέθους παρτίδας για βελτιωμένη μάθηση με αντιδιαστολή, έδειξαν χαμηλότερα ποσοστά ακρίβειας σε σύγκριση με τα αντίστοιχα μοντέλα που εκπαιδεύτηκαν με την απώλεια MMD. Ενώ η ανάμειξη (blending) έφτιαξε την απόδοση, δεν αντιστάθμισε πλήρως την απουσία της απώλειας MMD, δείχνοντας έτσι τον κρίσιμο ρόλο της στη βελτίωση της απόδοσης στο VQA πρόβλημα.

Ανάκτηση χωρίς προσαρμογή

Για να δείξουμε ότι η μέθοδός μας δεν βλάπτει τη γενίκευση του μοντέλου CLIP, την αξιολογήσαμε σε εργασίες ανάκτησης χωρίς προσαρμογή και συγκρίναμε τα αποτελέσματα με το CLIP και την βελτιωμένη μας μέθοδο, το CLIP-3Modal. Η σύγκριση φαίνεται στον πίνακα 10. Το μοντέλο μας βελτίωσε τις βαθμολογίες ανάκλησης του βασικού CLIP σε όλες τις εργασίες ανάκτησης και ελαφρώς υπερτερεί του CLIP-3Modal σε ορισμένες εργασίες ανάκτησης κειμένου.

Συμπεράσματα και Μελλοντικές Προεκτάσεις

Συμπεράσματα

Μπορούμε να δούμε εδώ ότι η χρήση μιας επιπλέον τροπικότητας είναι ένας εφικτός τρόπος βελτίωσης των μοντέλων εικόνας και κειμένου. Ο παραγωγικός διάλογος, που είναι πλούσιος σε επιπρόσθετες πληροφορίες και συμφραζόμενα, μας ενέπνευσε να τον ενσωματώσουμε στις αρχιτεκτονικές εικόνας-κειμένου. Αρχικά ενσωματώσαμε έναν επιπλέον κωδικοποιητή στην αρχιτεκτονική του CLIP μοντέλου για να συμπεριλάβουμε νέες τροπικότητες κατά τη διάρκεια της διαδικασίας εκπαίδευσης, επεκτείνοντας έτσι το υπάρχων μοντέλο εικόνας-κειμένου. Επιπλέον, παρουσιάσαμε τη μέθοδο **DRAFT**, μια καινοτόμο μέθοδο για την προσαρμογή μοντέλων τύπου CLIP σε διάφορα κειμενικά στυλ (π.χ., διάλογος) χρησιμοποιώντας μάθηση με αντιδιαστολή και ευθυγράμμιση κατανομών. Μιμούμενοι την προσέγγιση εκπαίδευσης του CLIP και ευθυγραμμίζοντας τις κατανομές ομοιότητας των διαφορετικών κειμενικών εισόδων με τις αντίστοιχες εικόνες, επιδείξαμε βελτιωμένη απόδοση σε εργασίες απάντησης οπτικών ερωτήσεων ενώ ταυτόχρονα διατηρήσαμε την ικανότητα γενίκευσης του μοντέλου.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32	34.2	60.0	70.6	52.4	76.3	84.3
C3M-ViT/B-32^{LLaVA}	**34.9**	**60.7**	**71.2**	52.9	**77.1**	84.7
Ours-ViT/B-32+blending	34.8	60.4	71.0	**53.1**	76.6	**84.8**
C3M-ViT/B-32^{LLaVA} (only 3rd)	27.5	51.3	62.4	39.8	66.9	76.5
Ours-ViT/B-32	25.4	49.0	60.6	36.7	64.3	75.5

Table 10: **Σύγκριση απόδοσης σε εργασίες ανάκτησης χωρίς προσαρμογη.** Η προσέγγισή μας υπερτερεί του βασικού CLIP-ViT/B-32 σε όλες τις μετρικές ανάκτησης. Με τη διατήρηση της απώλειας CLIP κατά τη διάρκεια της εκπαίδευσης, η μέθοδός μας διατηρεί την ικανότητά της γενίκευσης ενώ βελτιώνει την απόδοση μέσω της ανάμειξης των ενσωματώσεων.

Επιπλέον, διαπιστώσαμε ότι η απώλεια Maximum Mean Discrepancy (MMD) είναι ιδιαίτερα αποτελεσματική στην προσαρμογή των κατανομών ομοιότητας, κάνοντάς την κρίσιμο στοιχείο της μεθόδου προσαρμογής μας. Τέλος, δείξαμε ότι η ανάμειξη των παραμέτρων του βασικού μοντέλου με αυτές του προσαρμοσμένου κωδικοποιητη κειμένου βελτιώνει περαιτέρω την απόδοση.

Μελλοντικές Προεκτάσεις

Στο μέλλον, στοχεύουμε να αναλύσουμε περαιτέρω την ενσωμάτωση του τρίτου κωδικοποιητή στα μοντέλα τύπου CLIP και να εξετάσουμε εναλλακτικές επιλογές για τη τρίτη τροπικότητα. Και για τις δύο προσεγγίσεις, σχεδιάζουμε να πειραματιστούμε με διάφορα γενετικά μοντέλα και διάφορα στυλ για τη δημιουργία των παραγόμενων κειμένων. Επιπλέον, θα διεξάγουμε περαιτέρω πειράματα για τη μέθοδο DRAFT, συμπεριλαμβανομένης της αρχικής προσαρμογής του μοντέλου πριν από την αξιολόγηση σε εργασίες απάντησης οπτικών ερωτήσεων, ακολουθώντας το προτεινόμενο πρωτόκολλο αξιολόγησης. Ένα πολλά υποσχόμενο επόμενο έργο περιλαμβάνει την εφαρμογή της μεθόδου DRAFT σε μοντέλα κωδικοποιητή-αποκωδικοποιητή αντί για απλά μοντέλα κωδικοποιητή όπως το CLIP. Τέλος, η έννοια της ανάμειξης παραμέτρων προσφέρει ένα πολύ ενδιαφέρον πεδίο έρευνας λόγω των σημαντικών οφελών της για τις μεθόδους εκπαίδευσης μας. Για να βελτιώσουμε την ανάμειξη, σκεφτόμαστε να πειραματιστούμε με πιο δυναμικές προσεγγίσεις, όπως η κινούμενη μέση τιμή, αντί για την απλή πρόσθεση που προτάθηκε σε αυτήν την εργασία.

Chapter 1

Introduction

1.1 Self-supervised Learning

1.1.1 Overview

Self-Supervised Learning aims to provide rich representations and deep feature learning avoiding the use of annotated data as in Supervised Learning which defines the difficulty of practical deployment of Deep Learning nowadays. These methods made rapid advancements in recent years leading them to be comparably efficient to fully supervised pre-training methods and sometimes even surpassing their efficiency [42, 24].

Research on self-supervision motivated by the cost of manual annotation on datasets. With self-supervised methods we can design pretext tasks in order to obtain free labels to use on supervision for training a discriminative deep model. Specifically, we can train deep feature representations without annotation in order to train a deep neural network to solve a downstream task using comparatively little task specific annotated data compared to supervised learning.

As we mentioned above with self-supervised learning we are learning tasks that require to predict one part of the input or derive a label given another part of the input. We are able to see that self-supervision is in contrast with supervised learning. With supervised learning we try to predict a manually provided target output. In other words we train a generative model to estimate the density of the input data or learn a generator of them. Each Self-supervised method differs on the strategy for deriving labels to predict. This choice of the pretext task determines how effective the obtained representations will be in different downstream tasks. Some results suggest that besides the pretext task, the representation quality is also a logarithmic function of the amount of the unlabeled data. If this suggestion holds then we can achieve better performance just by the use of larger pre-training sets that would be allowed by the improvements in data collection and computing power [29].

These methods became popular across a variety of modalities such as image, text, speech, video and graphs. They managed to improve the sample efficiency of learning across the above modalities and also, boost diverse downstream tasks including: simple recognition, detection & localization [29], dense prediction [29], anomaly detection [25]. Furthermore, they lead to improvements on data efficiency of transfer learning [27, 28], semi-supervised learning [87] and active & meta learning [18].

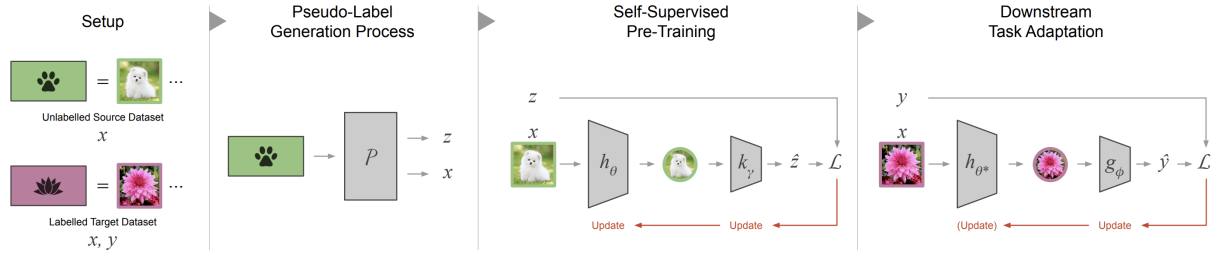


Figure 1.1.1: **Self-supervised learning pipeline.** The self-supervised workflow starts with an unlabelled source dataset and a labelled target dataset. As defined by the pretext task, pseudo-labels are programmatically generated from the unlabelled set. The resulting inputs, x and pseudo-labels z are used to pre-train the model $k_\gamma(h_\theta(\cdot))$ – composed of feature extractor h_θ and output k_γ modules – to solve the pretext task. After pre-training is complete, the learned weights θ^* of the feature extractor h_{θ^*} are transferred, and used together with a new output module g_φ to solve the downstream target task.[24]

1.1.2 Definition

At first we have to define the problem that self-supervised learning tries to solve and contrast it to the already known learning methods such as supervised and unsupervised learning [24].

Definition 1.1.1: Supervised Learning

Let a labeled dataset $D_i = \{x_i^{(t)}, y_i^{(t)}\}_{i=1}^N$ where D_i where N is the number of samples. Assume we have a problem, which we manage to solve in D_t by building a predictive model, $\hat{y} = f(x)$ that estimates the label y. For deep learning application as ours, the predictor is composed of a representation extractor function h_θ and a classifier or regression function g_φ:

$$f(x) = g_\varphi(h_\theta(x)) \tag{1.1.1}$$

So, we train our model to minimize a loss function $\mathcal{L}$:

$$arg\min_{\theta,\varphi} \sum_{(x_i^{(t)}, y_i^{(t)}) \in D_i} \mathcal{L}(g_\varphi(h_\theta(x_i^{(t)}), y_i^{(t)})) \tag{1.1.2}$$

The main problem on this approach is that h_θ may have hundreds of millions trainable parameters to fit which requires millions of labeled samples on dataset D_t.

Definition 1.1.2: Unsupervised Learning

In contrast to its supervised counterpart, unsupervised learning tries to learn from unlabeled data by estimating a density function over the input data or building a generative model. These techniques vary from Gaussian Mixture Models [33] to Generative Adversial Networks and Variational Autoencoders [27]. Other approaches focusing on learning latent representation such as clustering and autoencoders [27]. Following our previous notation, assume we have an autoencoder, our objective is to minimize a loss of reconstruction:

$$arg\min_{\theta,\varphi} \sum_{(x_i^{(t)}) \in D_i} \mathcal{L}(g_\varphi(h_\theta(x_i^{(t)}), x_i^{(t)})) \tag{1.1.3}$$

where h_θ extracts a feature representation and g_φ reconstructs the input x given the representation $h_\theta(x)$.

Now, we can see self-supervised learning as a special case of unsupervised learning where we don't have to reconstruct the input or estimate a density probability. Instead we build a pretext task $\mathcal{P}$ that exploits knowledge about the data.

Definition 1.1.3: Self-supervised Learning

Let $D_s = \{x_i^{(s)}\}_{i=1}^M$ be an unlabeled *source* dataset, where $M \gg N$ (unlabeled dataset is significant larger than the annotated), our objective is to make use of D_t and D_s s together to learn a predictive model $f(x) = g_\varphi(h_\theta(x))$. We create process $\mathcal{P}$ as our pretext task in order to generate pseudo-labels and an objective to guide learning. Given the source dataset D_s we generate pseudo-labels z and data points $\mathcal{P}(D_s) = \{x_i, z_i\}_{i=1}^M$.
Let $\bar{D}_s = \mathcal{P}(D_s) = \{x_i, z_i\}_{i=1}^M$ be the new pseudo-labeled dataset and k_γ be the pretext model. We try to optimize the self-supervised objective on the new dataset $\bar{D}_s$:

$$\theta^* = arg\min_{\theta,\gamma} \sum_{(x_i,z_i) \in \mathcal{P}(D_s)} \mathcal{L}(k_\gamma(h_\theta(x_i), z_i)) \tag{1.1.4}$$

Finally we discard the pretext function k_γ and keep only the optimized representation function h_{θ^*}. We use h_{θ^*} as a partial solution to solve the target problem using model $g_\varphi(h_{\theta^*}(\cdot))$.

Because parameters θ^* are well fitted, thus we have to learn only a minority of parameters in order to solve the downstream task. There are two common ways to solve the above problem using θ^*, ***fine-tuning*** and ***linear-readout***.

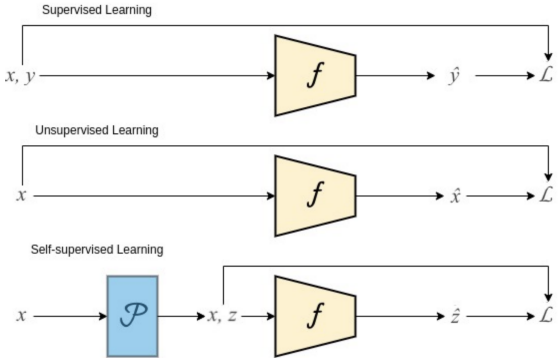


Figure 1.1.2: **An illustrative comparison between Self-supervised Learning, Supervised Learning and Unsupervised Learning.**

1.1.3 Pretext Tasks

As we noted before, the pretext task determines the nature of the self-supervised method. We can divide the literature in four broad families of tasks: *masked prediction, transformation prediction, instance discrimination, clustering.*

Masked Prediction

For masked prediction tasks the process $\mathcal{P}$ removes a part of the input data and we train our model to fill in the missing data. [46] proved that masked prediction is not restricted only in natural language application. In particular, they took a source instance $x_i^{(s)}$ and produced two new views x_i and z_i containing parts of the source. They showed that if there is a conditional independence between x_i and z_i given the downstream label, then by estimating that label we can predict z_i from the source $x_i^{(s)}$. There are various examples of tasks across all modalities. We can hide words for language modeling as in BERT [20], hide time-slots in speech [7] or hide a region of an image for inpainting [61].

Transformation Prediction

Assume the input has a canonical view and we applied a certain transformation to change it. We have to train the model to predict which transformation applied on the input. In computer vision, we can apply rotations to the raw images and requiring the network to predict the rotation angle [26]. In temporal data, such as videos, speech or other time-series we can shuffle the temporal order of the signal and train the network to predict the original order [72, 95].

Instance Discrimination

In this family, each input on the dataset is treated as its own class. The model is trained to discriminate between different instances. There are a few variations of this technique. The most straightforward way to manage the problem is to one-hot encode each instance on the dataset and with a categorical cross-entropy predict the correct instances [21]. However, the use of categorical cross-entropy loss become intractable for large datasets. Researchers suggested a more efficient way to approximate this loss by using contrastive methods. Specifically, they inspired by metric learning and contrastive estimations [32] and the core idea is to predict if a pair of inputs belongs to the same class instead of predicting the exact class. There are a lot established self-supervised frameworks in this family. For vision applications there are MoCo [31] and SimCLR [13], for speech we have CPC [59] and for multimodal applications, such as visuo-linguistic, there is CLIP [64].

Clustering

The clustering based methods traditionally focused on dividing the input data into a number of groups with high intra-group and low inter-group similarity (e.g. K-Means). However, in self-supervised learning applications, they aim to learn a good feature extractor instead of clustering assignments. Major examples in this category include ODC [12] for vision and XDC [2] for multimodal.

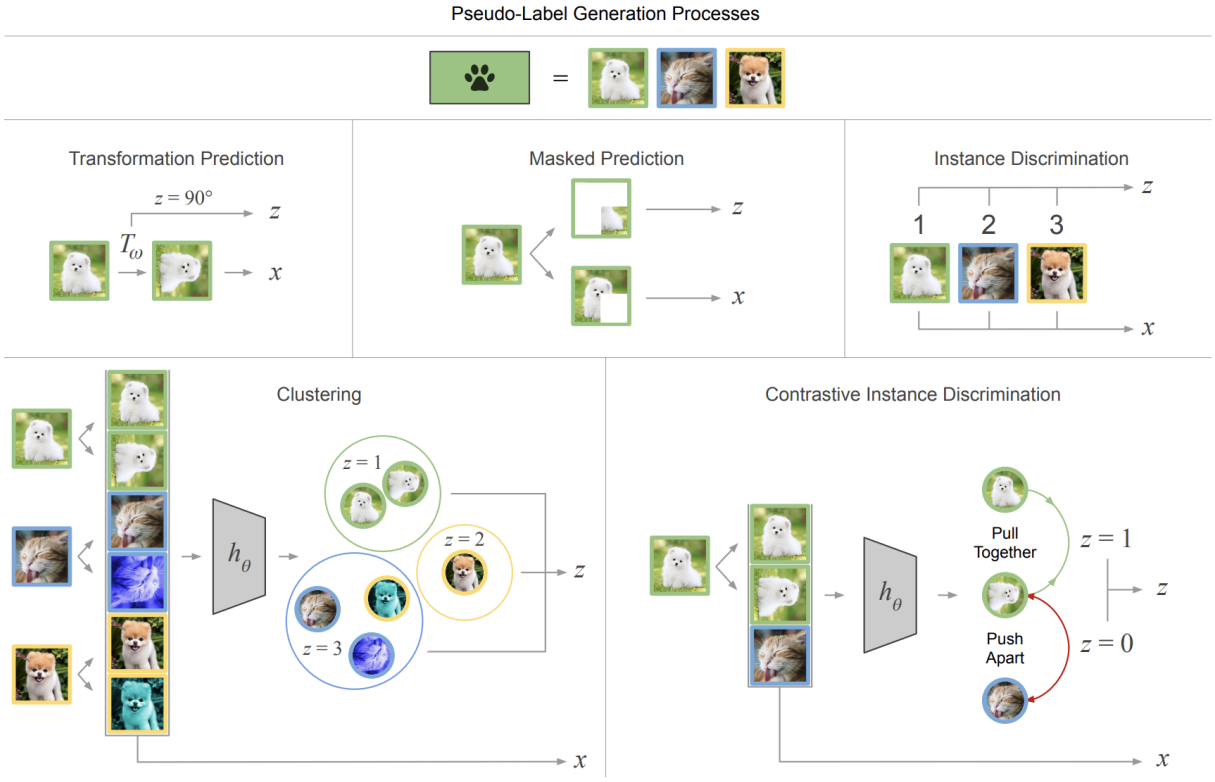


Figure 1.1.3: **Pseudo-label Generation Processes.** Illustrative examples of the way pseudo-labels are generated in the four families of pretext tasks of our taxonomy: transformation prediction, masked prediction, instance discrimination and clustering. An additional depiction is included of the popular version of instance discrimination using contrastive losses. Squares represent inputs x while circles portray the feature vectors of those inputs, $h_\theta(x)$.[24]

Chapter 2

Theoretical Background

2.1 Contrastive Learning

2.1.1 Overview

Contrastive learning, as a subset of self supervision, has been a major technique in recent state-of-the-art machine learning models, such as CLIP [64]. It relies on the use of multiple semantically related samples whose representations are then made as similar as possible. These are called positive samples in this context, and are often derived from the same underlying sample, under different transformations. Aside from views of the same sample, contrastive methods also make use of negative samples, which are considered to be semantically different from the original one. These are often obtained from random samples within the same batch [13] or from a past history of samples from the model [36]. By using these negative samples, the model is able to improve its representations, by learning to distinguish between samples which are different to each other. This form of representation learning has become immensely popular in the context of image-text models, being an integral element of several state-of-the-art works in this setting [64, 97].

2.1.2 Definition

Contrastive learning has become a popular technique for unsupervised learning in the past few years. The term 'contrastive learning' originally introduced by Arora et al. [6] as algorithms that remind of the well-known ***word2vec*** [58] that advantages the form of positive pairs containing semantically similar data points and negative samples. In a more formal way, we assume the existence of a positive pair $(x, x^+) \sim D_{sim}$ and k i.i.d. negative samples $x_1^-, x_2^-, ..., x_k^- \sim D_{neg}$ which are presumably unrelated to x. The positive pairs can obtained by taking two independently random augmented views of the same sample e.g. two crops of the same image. The main objective of contrastive learning is to maximize the semantic similarity of the positive pair and minimize it for all other pairs formed by the negative samples. *Cosine similarity* became the most common similarity function for these tasks as a result of its simplicity.

$$CosSim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \cdot \mathbf{y}}{||\mathbf{x}|| \cdot ||\mathbf{y}||} \tag{2.1.1}$$

Another perspective of contrastive learning given again by Arora et al. [6] in which we form latent classes and each augmented view belongs to the class the original data point defines. Then the objective for this task is to minimize a cross-entropy loss across all classes.

2.1.3 Contrastive Loss

In the same way, we define the contrastive loss as a variation of the popular cross-entropy loss. A widely used loss for contrastive learning which used in previous works [13, 79, 92, 59] is the *NT-Xent* (Normalized Temperature-Cross Entropy Loss). Suppose we have a mini-batch of size N and a positive pair (x_i, x_j) and every other pair counted as negative then the loss defined as:

$$l_{i,j}^{NT-Xent} = -\log \frac{\exp(sim(x_i, x_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} \exp(sim(x_i, x_k)/\tau)} \tag{2.1.2}$$

where τ is a temperature parameter. The above loss can be seen as a variation of another popular loss for contrastive learning objectives, the *InfoNCE* which proposed by [59] and applied by [100] for contrastive representation learning of *(image,text)* pairs on medical imaging. The *InfoNCE* loss defined as:

$$l_{i,j}^{InfoNCE} = -\log \frac{\exp(sim(x_i, x_j)/\tau)}{\sum_{k=1}^{N} \exp(sim(x_i, x_k)/\tau)} \tag{2.1.3}$$

Again we suppose we have a mini-batch of size N and (x_i, x_j) is a positive pair while all the other pairs assumed negative.

2.1.4 Alignment & Uniformity

In their work, Wang and Isola [91] introduced two key properties of contrastive loss: 1) ***alignment***, 2) ***uniformity***. These two are quantified measures of the representation quality. In particular, *alignment* assures that the representations of both samples of the positive pair will map to nearby features and *uniformity* secures that feature vectors are preserving as much information of data as possible because they should be roughly uniformly distributed on the unit hypersphere.

Definition 2.1.1: Perfect Alignment

We can say that an encoder f is ***perfectly aligned*** if $f(x) = f(y)$ a.s. over $(x, y) \sim D_{pos}$.

Definition 2.1.2: Perfect Uniformity

We can say that an encoder f is ***perfectly uniform*** if the distribution of $f(x)$ for $x \sim D_{pos}$ is the uniform distribution σ_{m-1} on $\mathcal{S}^{m-1}$, where $\mathcal{S}^{m-1}$ is a Borel space.

We can see that we can't have perfect alignment and perfect uniformity at the same time, that scenario implies that each augmented view of a data point should have the same feature vector to have perfect alignment. However this does not forms a uniform distribution, thus we not satisfy the uniformity limitation. Quantifying the above properties we can obtain the following two metrics:

Definition 2.1.3: Alignment Loss

Alignment loss is defined as the expected distance between positive pairs (x, x^+):

$$l_{align}(f; a) \triangleq \mathop{\mathbb{E}}_{(x,x^+)\sim D_{pos}} [||f(x) - f(x^+)||_2^a], \quad a > 0 \tag{2.1.4}$$

where f is an encoder and D_{pos} the distribution over the positive pairs.

Definition 2.1.4: Uniformity Loss

Uniformity loss is defined as the logarithm of the average pairwise Gaussian potential:

$$l_{uniform}(f; t) \triangleq \log \mathop{\mathbb{E}}_{\substack{i.i.d. \\ (x,y)\sim D_{data}}} \left[e^{-t||f(x)-f(y)||_2^2}\right], \quad t > 0 \tag{2.1.5}$$

where f is an encoder and D_{data} the distribution over the data and Gaussian kernel (Radial Basis Function (RBF) kernel) is defined as $G_t : \mathcal{S}^d \times \mathcal{S}^d \to \mathbb{R}_+$ [17, 9]:

$$G_t(u, v) \triangleq e^{-t|u-v||_2^2} = e^{2t \cdot u^\top v - 2t} \tag{2.1.6}$$

They proved that contrastive loss optimizes for both alignment and uniformity asymptotically and found strong agreement between both metrics and downstream performance. Let $\mathcal{L}_{contrastive}$ be the contrastive loss, τ a temperature parameter, f an encoder and inner product $(\cdot)$ the similarity function. Then, asymptotically for M negative samples we have:

$$\lim_{M\to\infty} \mathcal{L}_{contrastive} - \log M = -\frac{1}{\tau} \mathop{\mathbb{E}}_{(x,x^+)\sim D_{pos}} \left[f(x)^\top f(x^+)\right] + \mathop{\mathbb{E}}_{x\sim D_{data}} \left[\log \mathop{\mathbb{E}}_{x\sim D_{data}} e^{f(x^-)^\top f(x)/\tau}\right] \tag{2.1.7}$$

where,

$$\mathcal{L}_{contrastive}(f; \tau, M) = \mathop{\mathbb{E}}_{\substack{(x,x^+)\sim D_{pos} \\ \{x_i^-\}_{i=1}^M \sim^{i.i.d.} D_{data}}} \left[-\log \frac{\exp\left(f(x)^\top f(x^+)/\tau\right)}{\exp\left(f(x)^\top f(x^+)/\tau\right) - \sum_i \exp\left(f(x_i^-)^\top f(x_i^+)/\tau\right)}\right] \tag{2.1.8}$$

We can see that the first term is minimized if and only if f is perfectly aligned and the second term if and only if f is a perfectly uniform encoder.

2.1.5 Applications in Computer Vision

The most popular frameworks for contrastive learning are focusing on visual representations. Two of the most well-known frameworks of the field are SimCLR [13] and CLIP [64]. Another interesting work on the field of computer vision is Contrastive Multiview Coding (CMC) [83]. CLIP is analysed in the next chapter as one of the fundamental components of our work.

CLIP

Radford et al.[64] made a large step forward in multimodal learning for image-text data with their CLIP (Contrastive Language–Image Pre-training) model. Their work proposed a contrastive learning scheme to embed both the image and text in a shared multimodal representation space. CLIP made major progress in zero-shot image classification and multiple challenging distribution shifts. A more extensive analysis on CLIP is available on Section 3.1.

SimCLR

SimCLR showed that the composition of multiple data augmentation operations is crucial in order to yield effective image representations via a predictive task. In this specific work proposed a series of transformations which achieve a good performance. In particular, they crop the input image and resize it back to the original size, then they used random color distortions and finally a random Gaussian blur. After the augmentation stage they used ResNet [34] as a base encoder for feature extraction. For the prediction task they formed a mini-batch of size N and augmented every sample inside it, which formed $2N$ data points. They did not use negative sampling as in other works, but they simply treated the augmented pairs as positive and every other pair as negative, resulting to N positive pairs and $2(N-1)$ negative pairs. As mentioned before, they use cosine similarity to measure the semantic similarity between two data points. Finally, for the prediction task, they used the *NT-Xent* contrastive loss across all positive pairs in the mini-batch (both (x_i, x_j) and (x_j, x_i)).

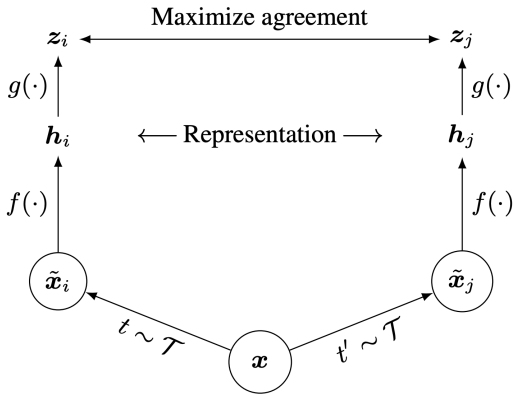


Figure 2.1.1: **SimCLR** [13]

2.1.6 Applications in NLP

Most recent works on NLP approaches of contrastive learning focused on sentence or phrase representations and text generation . In particular, we have to mention CoNT (Contrastive Neural Text Generation) [3] and SimCTG [80] for text generation, Quick-Thoughts [57] which proposed an efficient way to yield sentence representations and UCTopic [49] which focused on phrase representation and phrase mining. Notably for the text generation methods, CoNT samples examples from its own predictions which exposes the model on its own mistakes, using an N-pairs contrastive loss and incorporate the learned sequence similarity score directly

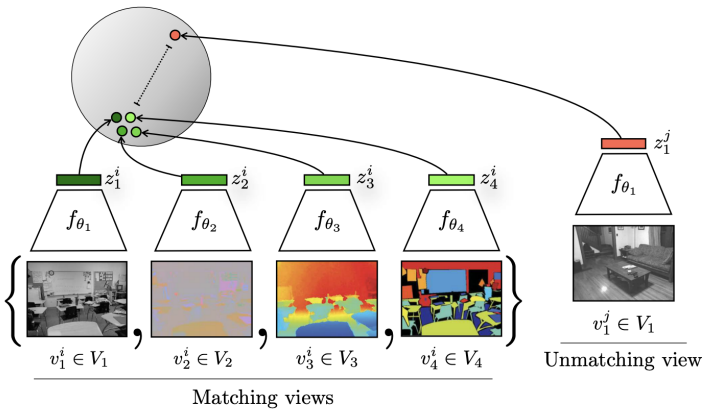


Figure 2.1.2: **Contrastive Multiview Coding framework** [83]

to inference stage. SimCTG encourages the model to learn discriminative and isotropic token representations and introduces contrastive search as a decoding method.

2.2 Multimodal Representation Learning

2.2.1 Overview

Multimodal research has been on the spot for over 50 years trying to design agents with intelligent capabilities such as understanding, reasoning and learning in the same way a human does. Multimodal learning is a very active field of research in many disciplines. Deep Learning took the world by storm the past decade which popularized again the research on the multimodal domain giving the tools to unlock new capabilities and perspectives of elements used in machine learning. The main modalities used for machine learning research are: language, vision, acoustics, touch, physiological signals and mobility. Various applications such as self-driving cars [93], text-to-speech technologies [5], video understanding [81], image and text generation[67, 77], embodied agents [10], multisensor fusion [47] are bringing us closer to our goal for intelligent systems in domains such as robotics, healthcare and multimedia affective computing and human-computer interaction.

Each modality has 3 key principles: *heterogeneity*, *connections* and *interactions* [52] which have driven subsequent innovations. Modalities are heterogeneous due to the diversity of the information's qualities and structure, for example, vision often captures as images while language as a text created by a set of characters. Connected are the modalities that are often related and share commonalities, e.g. acoustics and language are connected when we have a speech sample and its transcripts. Finally, modalities interact to yield new information for task inference. These principles lead researchers to a taxonomy [52] of the field containing 6 challenges in recent multimodal learning : *representation*, *alignment*, *reasoning*, *generation*, *transference* and *quantification*.

2.2.2 Representation Learning

Representations of raw data in a format that a computational model can work with has always been a challenge in the machine learning community. Multimodal representation learning studies the ways to represent and summarize multimodal information to reflect the heterogeneity and the interconnections between individual modalities. However this challenge shows many difficulties such us how to combine different modalities, how to deal with noise or how to deal with missing data. The ability to represent data in a meaningful way that contains crucial information about the nature of the entity forms the backbone of every multimodal learning model. In order to cover the above challenge we divide the field in 3 sub-challenges [52]: *representation fusion* where the number of modalities is greater than the number of separate representations, *representation coordination* where we keeping the same number of representations but we encourage the cross-modal interaction and *representation fission* where the number of representations is greater than the given modalities

and tries to capture structural information of the data. In Figure 2.2.1 we present schematically the above sub-challenges.

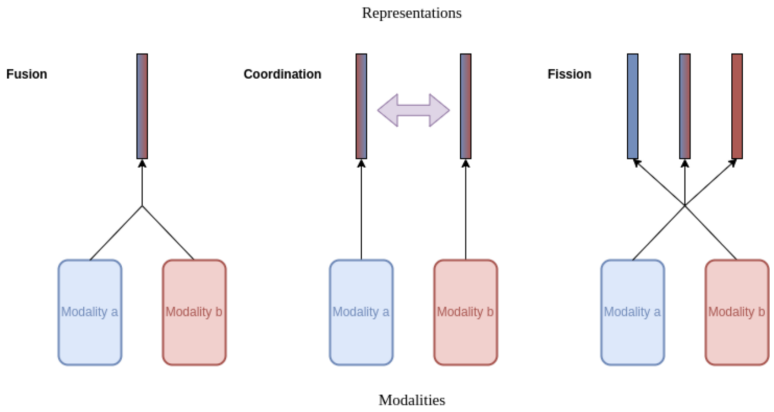


Figure 2.2.1: **Multimodal Representation Learning:** Fusion, Coordination, Fission

2.2.3 Fusion

Representation fusion aims to learn a joint representation space that enhance the cross-modal interactions between different modalities and at the same time effectively reducing the number of separate representations. We can divide approaches by the stage in which the fusion takes place. As a result we have early fusion and late fusion. In early fusion we fuse representation at very early stages with minimal pre-processing. Sometimes, early fusion even involves raw modalities themselves. The most common early fusion technique is a simple concatenation of the feature vectors. On the other hand in late fusion suitable unimodal encoders are applied to capture separate representations of every individual modality followed by several building blocks for fusion in order to learn the joint representation space.

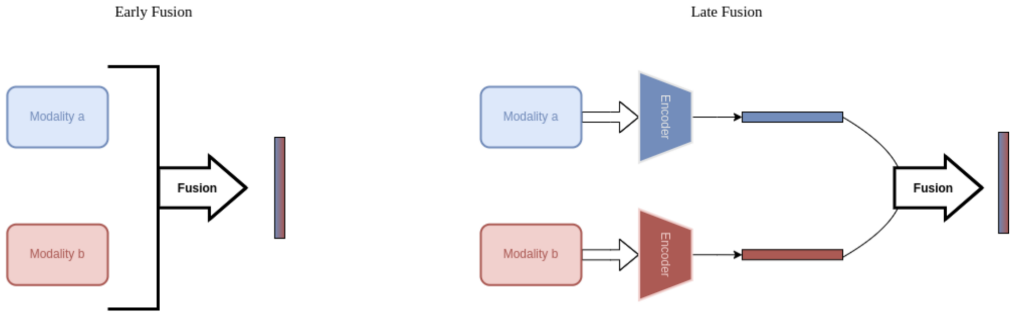


Figure 2.2.2: **Fusion Stages:** Early Fusion and Late Fusion

2.2.4 Coordination

Representation coordination aims to learn contextualized multimodal representations that are coordinated through their interconnections. Coordination tries to keep the number of representations the same as the number of modalities in a way that enhances the ability to capture multimodal context. At first we have strong coordination that enforces the equivalence between the modalities. In this category of coordination contrastive learning techniques are the main subject of study. For example CLIP [64] tries to encourage the representations of a picture of a dog and the word "dog" be close (semantically correlated) and in the meanwhile the pair between the picture of the dog and every other word except "dog" (semantically uncorrelated) should be far apart. Other works showed that contrastive learning provably captures redundant

information across two views but fails to capture non-redundant [83]. In addition to contrastive learning, other approaches focusing on learning a mapping of corresponding data from one modality to another in order to learn the coordinated space [78, 23]. If we want to capture more general connections between modalities we have partial correlation. Now our approaches do not try to make the two views equivalent, but to coordinate them in matter of order, hierarchy or other relationships beyond similarity. Works on partial correlation used CCA (canonical correlation analysis) [82], ordered or hierarchical spaces [89] and semantic relationships between modalities [98].

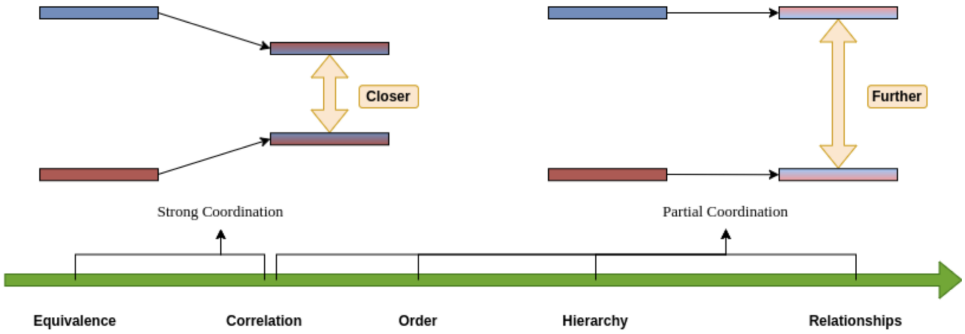


Figure 2.2.3: **Representation coordination:** Strong & Partial Coordination

2.2.5 Fission

Finally, representation fission aims to create new decoupled representations, more than the initial modalities, that exploit internal structure knowledge such us clusters or independent factors of variation. In comparison with joint and coordinated representation, fission allows a careful interpretation of the multimodal element and robust controllability of its properties. Depending the degree of the detail represented on the decoupled vectors, fission methods can be divided into modality level fission and fine-grained fission [52]. In modality level fission, at first we factorize into independent modality specific information and after into multimodal information redundant in both modalities. We can see fission as a problem of disentanglement representation learning, that tries to learn a variation of the data given independent latent variables, where given the modality specific vectors and the redundant multimodal vector encourages the independence between them [38, 8]. A suitable technique when its not easy retrain the disentangled model, for example when we have a large pretrained model, is post hoc representation disentanglement. Works lying in this category using EMAP (empirical multimodal additive function) [37] in order to retrieve the effects of unimodal interactions given the cross-modal interactions. Now, fine-grained fission attempts to break deeper into modalities by retrieving more information like clustering. Recent works suggested clustering approaches like, k-means algorithm combined with self-supervised contrastive learning in videos [14] or application of k-means algorithm in representations yielded by unsupervised audiovisual learning [40].

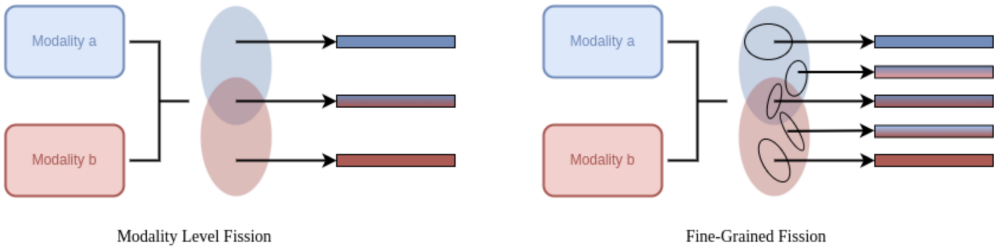


Figure 2.2.4: **Representation Fission categorization:** Modality Level Fission & Fine Grained Fission

Chapter 3

Related Work

3.1 CLIP

3.1.1 Overview

Recent works have proven that deep learning has revolutionized computer vision. However, there are still major problems as these approaches require enormous manually annotated datasets and the typical vision models are good at one task only and require significant effort to adapt to a new one. To address these problems Radford et al. introduced CLIP (Contrastive Language-Image Pretraining) [64]. The above neural network is trained on a wide variety of images with natural language supervision available from the web. Specifically, in order to learn from raw text its pre-training task is to predict which caption goes with which image. It proved that the above proposal is an efficient and scalable way to learn SOTA image representation. After the pre-training, natural language is used to reference different visual concepts, enabling zero-shot transfer of the model on the downstream task. CLIP was evaluated on over 30 computer vision datasets and a variety of tasks such as OCR (object character recognition), action recognition in videos. geolocalization and many types of fine-grained classification.

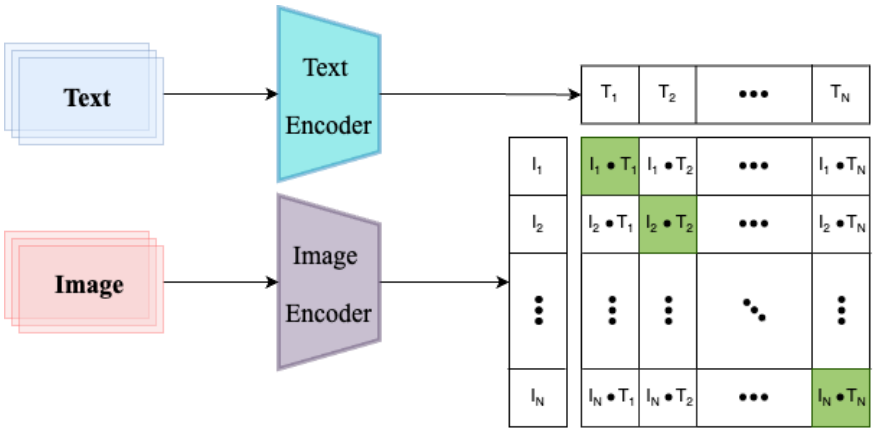


Figure 3.1.1: **CLIP:** Contrastive Language-Image Pre-training

3.1.2 Natural Language Supervision

Pre-training from raw text revolutionized NLP the past few years. Various task-agnostic objectives, such as masked prediction and autoregressive language modeling, became more scalable. Furthermore, task - agnostic architectures removed the need for specialized output heads after the development of "text-to-text" as an input-output interface. GPT-3 [11] for example became a widely popular system proving that these systems are now competitive across many tasks requiring little to zero task specific data for training. In other domains such as computer vision it is still a common practice to train models with crowd labeled datasets. The proposed idea was if computer vision systems can learn directly from web text the same way as in NLP? There are encouraging works and notable strengths of this perception. At first, is easier to scale natural language supervision compared to crowd-sourced labels for image classification since, natural language does not require annotation. In addition, supervision with natural language approaches yields better representations than supervised and self-supervised methods for the reason that connects the representation to language and enables flexible zero-shot transfer.

3.1.3 Method

CLIP follows the work of CMC [83] in which it was shown that a contrastive objective yields better representations than the predictive objective. Specifically, CLIP uses contrastive learning and tries to predict which text "as a whole" is paired with image instead of predicting the "exact" words, which is a difficult task. Given a batch of N pairs of image-text , CLIP tries to predict which of the $N \times N$ pairings occurred. In order to do this, learns a multimodal embedding space by jointly train an image encoder and a text encoder to maximize the cosine similarity of the image and text embeddings of the N true pairs, while minimizing the cosine similarity of the $N(2-N)$ incorrect pairs. As a task it has to optimize a symmetric cross-entropy

loss over these similarity scores similar to the InfoNCE loss' [59] adaption for (image,text) pairs for medical imaging in ConVIRT [100]. Specifically, for a batch of N image-text pairs $(v_i^T, v_i^I), i = 1, ..., N$, we calculate CLIP loss as it shows below:

$$l_i^{(I \to T)} = -\log \frac{\exp\left(v_i^I \cdot v_i^T / \tau\right)}{\sum_{k=1}^{N} \exp\left(v_i^I \cdot v_k^T / \tau\right)} \tag{3.1.1}$$

$$l_i^{(T \to I)} = -\log \frac{\exp\left(v_i^T \cdot v_i^I / \tau\right)}{\sum_{k=1}^{N} \exp\left(v_i^T \cdot v_k^I / \tau\right)} \tag{3.1.2}$$

$$\mathcal{L}_{CLIP} = \frac{1}{2N} \sum_{i=1}^{N} \left(l_i^{(I \to T)} + l_i^{(T \to I)}\right) \tag{3.1.3}$$

During training time, the only data augmentation applied is a random square crop from resized images. At evaluation time, the trained text encoder synthesizes a zero-shot linear classifier by embedding the descriptions of the target image classes. In Figures 4.3.1a and 3.1.2 there are schematical representations of the training and evaluation of CLIP respectively.

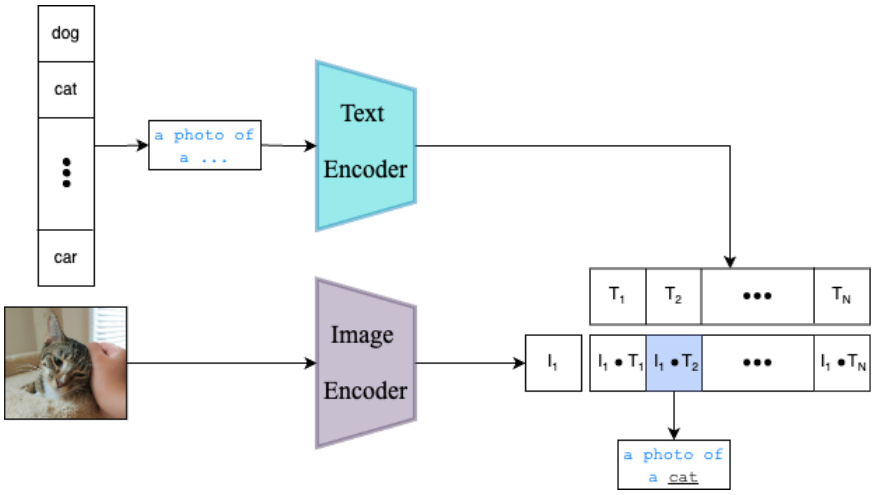


Figure 3.1.2: **CLIP's zero-shot prediction evaluation.** We convert all of a dataset's classes into captions such as "a photo of a cat" and predict the class of the caption CLIP estimates best pairs with a given image.

3.1.4 Model Selection

For the image encoder two architectures taken into consideration. The first is a modified version of ResNet-50 [34] as the baseline image encoder due to its proven performance and capabilities. Specifically, the authors used the ResNet-D improvements [36] and the antialiased rect-2 blur pooling from Zhang's work [99]. They also replaced the global average pooling layer with an attention pooling mechanism. The alternative architecture proposed used a ViT (Vision Transformer) [22] as the image encoder. For the text encoding they used a transformer architecture with modifications proposed by Radford et al. in [63].

3.1.5 Limitations

As the authors observed [64], there are still many limitations to CLIP. While CLIP's performance on common object recognition is well, it struggles on more abstract objects or more systematic tasks. In particular, counting the number of objects or predicting the depth of an object in the image are tasks that CLIP underperforms. On these tasks, zero-shot CLIP slightly outperforms random guessing. Furthermore, it

struggles compared to task-specific models on fine-grained classification tasks such as telling the difference between car model, dog breeds or flower species. CLIP also has poor generalization to images that are not covered in its pre-training dataset. For example, while zero-shot CLIP is evaluated on MNIST it only achieves 88% accuracy, well below 99.75% of humans. Finally, it was observed that CLIP's zero-shot classifier is sensitive to words and phrases and requires trials and error prompt engineering in order to perform well.

3.1.6 OpenCLIP

In our work, we use the pre-trained CLIP model provided by the OpenCLIP codebase [41]. OpenCLIP is an open source implementation of OpenAI's CLIP which enabled the training of CLIP models on a variety of data and compute budgets.

3.2 Transformers

3.2.1 Overview

In this chapter we will have a look on Transformer models, which have took the world by storm the last 7 years since their original proposal by Vaswani [88]. At this time Transformers have state-of-the-art performance across may tasks and datasets in a broad range of domains such as Natural Language processing, Computer Vision, Audio Processing. Their advantage over traditional methods is that Transformers can be trained using vast unlabeled data. As a result, very large Transformer models and improvements in the usage of unlabeled have led to applications that supersede their supervised counterparts and give valuable solutions in real-world problems.

Transformers are models that are based only in self-attention mechanisms [88], without using any CNNs or RNNs. We can classify them in 3 basic categories *1) autoencoding Transformers* which are a stack of encoders, *2) auto-regressive Transformers* that are a stack of decoders and *3) sequence-to-sequence Transformers* which is a stack of encoders followed by a stack of decoders. The input of a Transformer is a word embedding, however in the seq-to-seq case we need a positional embedding as well. Encoders consist of a self-attention layer and a neural network. Each encoder passes its output as input to the next encoder in the stack. Decoders on the other hand, contain a self-attention layer, followed by an encoder-decoder attention layer and a neural network. This architecture results in SOTA applications and faster training times than the traditional CNN and RNN architectures. BERT [20] (Bidirectional Encoder Representations from Transformers) should be considered one of the most characteristic Transformer architectures in literature. BERT is a general purpose transformer-based architecture that achieves SOTA results in various Natural Language Processing tasks and datasets. Furthermore, we can consider BERT as a deep bidirectional autoencoding Transformer. More specifically, each input word represented as a token embedding, a segment embedding and a positional embedding. In training time, BERT randomly masks a small segment of input embeddings and the goal is to complete the masked words by training the self-attention encoders. Finally, the output representation consists of the hidden state of the classification token which serves as an input in the classification head that is fine-tuned on top of BERT.

3.2.2 Attention

Self-Attention

Transformer models perform computations in parallel using self-attention [88] blocks, unlike the traditional Sequence Models that perform all computations sequentially. In this part, we have to formally define how the Self-Attention mechanism works. Let us have a sentence containing n words. We start by representing the n words in the sentence by using word embeddings so the result should be a d-dimensional vector $x_i \in \mathbb{R}^d$. So, our sentence should be the $n \times d$ matrix $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{(n \times d)}$. Because this representation does not take into account the surroundings of each word in the sentence, self-attention calculates in parallel n self-attention representations $A_1, \ldots, A_n$ for the n words. For each embedded word $x_i \in \mathbb{R}^d$ we calculate a query $q_i \in \mathbb{R}^{d_q}$, a key $k_i \in \mathbb{R}^{d_k}$ and a value $v_i \in \mathbb{R}^{d_v}$ represented as row vectors of the matrices X, Q, K and V, by linear projecting the X into the d-dimensional space of the queries, keys and values as shown below:

$$Q = XW_Q,\ K = XW_K,\ V = XW_V \tag{3.2.1}$$

and

$$q_i = x_i W_Q,\ h_i = x_i W_K,\ v_i = x_i W_V \tag{3.2.2}$$

where W_Q, W_K and W_V are learned $d \times d$ matrices and q_i, k_i and v_i are the i-th rows of the Q, K and V $n \times d$ matrices. Then we compute the self-attention representation A_i as the softmax of the inner product between q_i and k_j (both of dimension d_k) for $j = 1, ..., n$ multiplied by the values v_i (of dimension d_v).

$$A_i(q_i, K, V) = \sum_{i=1}^{n} \frac{\exp(q_i k_i)}{\sum_j \exp(q_i k_j)} v_i \tag{3.2.3}$$

if we calculate the sum for all the n words, we have a scaled dot-product attention with scaling factor $\frac{1}{\sqrt{d_k}}$ as it shows below:

$$A(X) = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{3.2.4}$$

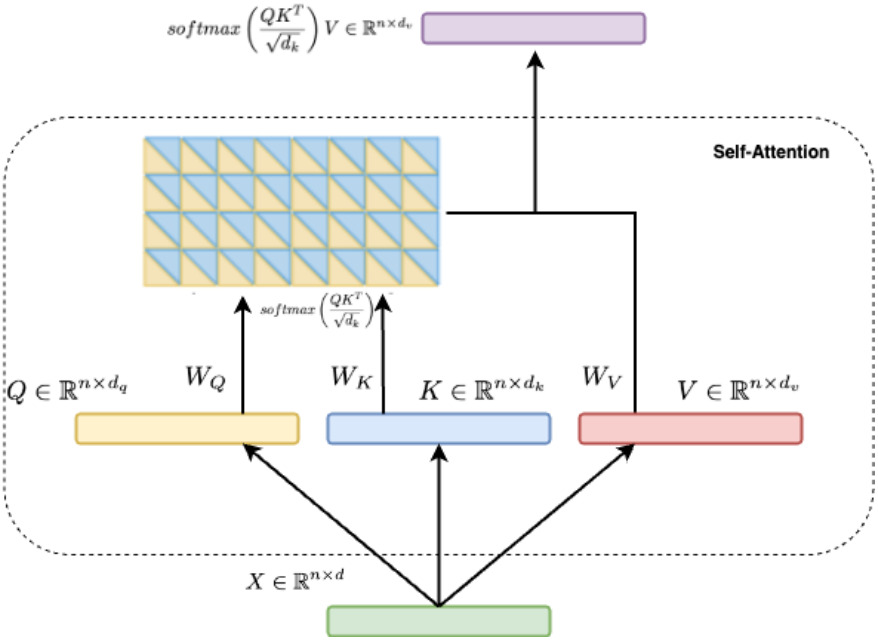


Figure 3.2.1: **Self Attention Mechanism**

Multi-head Attention

Multi-head attention performs self-attention multiple times. Instead of inner products, the query, key and value vectors are multiplied by matrices W_h^Q, W_h^K, W_h^V for each head $h = 1, ..., m$ and the multi-head attention representation are:

$$A_h(X) = \text{Attention}_h(Q, K, V) = \text{Attention}_h(W_h^Q Q, W_h^K K, W_h^V V) = \text{softmax}\left(\frac{W_h^Q Q W_h^K K^\top}{\sqrt{d_k}}\right) V \tag{3.2.5}$$

for each embedded word i. The m multi-head attention representations $A_h(X)$ for $h = 1, ..., m$ are concatenated and multiplied:

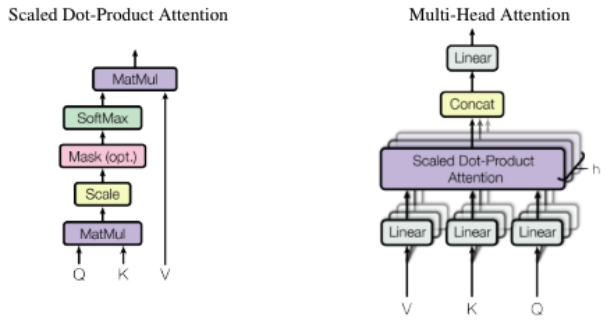


Figure 3.2.2: **Attention vs Multihead Attention.** Left: The Scaled Dot-Product Attention. Right: The Multi-Head Attention module. [88]

$$\text{MultiHead}(X) = \text{Concat}\left(A_1(X), ..., A_m(X)\right) W^O \tag{3.2.6}$$

where W^O is a learnable $md \times d$ matrix. In order to compute the multi-head attention, we first calculate the sum of the embedded sequence X of $n \times d$ dimensions and the position encoding P as input. Next, for each attention head, we compute queries Q , keys K and values V represented by $d \times d$ matrices, which are passed to the multi-head attention layer whose output is of dimension $n \times d$.

3.2.3 Architecture

Given sentences of embedded words, a Transformer may be used for diverse tasks such natural language understanding, text generation and translation of a sentence from one language to another. A Transformer consists of encoder or decoder blocks or both.

Positional Encoding

The position of words is required for the computation of the attention score. The positions of the words in the sentence are encoded as a position embedding by sine and cosine functions and added to X [88]. More specifically, position embedding matrix P is a $n \times d$ matrix and defined by:

$$P_{pos,2i} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right), \; P_{pos,2i+1} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{3.2.7}$$

where pos is the position of the word in the sentence, d is the dimension of the word embeddings and $i = 1, ..., d$ is the dimension index. By adding the positional embedding P to X allows the model to learn to attend to relative positions.

Encoder

The encoder block takes as input a matrix X of embedded words. The position encoding is added to the embedded words to form the input $X + P$. We then compute the queries Q, keys K and values V and pass the through a multi-head attention layer whose output is fed to a feed-forward neural network. In order to create an encoder, a block consisting of multi-head attention and a feed-forward neural network is repeated multiple times.

Decoder

The output of the encoder is fed to the decoder bock, which predicts the translated sentence. The decoder also consists of of multiple blocks of multi-head attention, which are fed into feed-forward neural networks and add positional embeddings to the inputs. Both the encoder and the decoder may consist of residual connections

between blocks and add normalization layers before the feed-forward neural networks. The output of the decoder is fed through a linear layer followed by a softmax layer. During generation, the decoder predicts new words, whereas during training , the decoder predict masked words from the input. The encoder-decoder architecture is shown in Figure 3.2.3

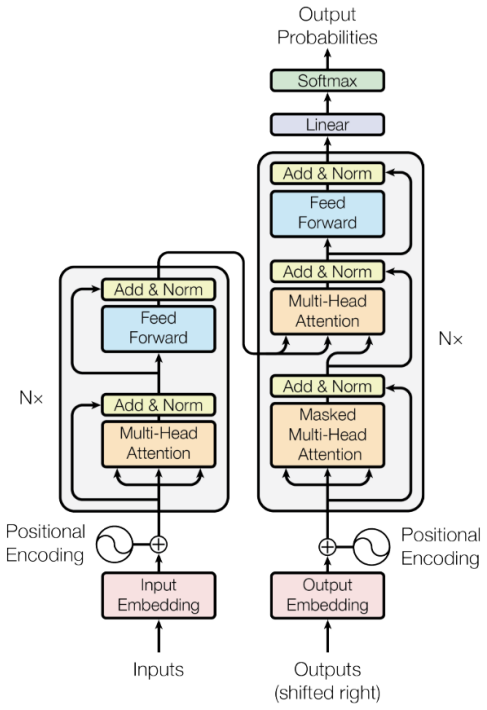


Figure 3.2.3: **The Transformer-model architecture** [88]

Training

Pre-training a Transformer is computationally expensive and most often involves vast amounts of unlabeled data. The most common optimization objectives for pre-training language models are *1)* masked word prediction, which is predicting a random deleted word in a sentence or predicting the next word; and *2)* classifying whether two sentences follow each other or not. This computationally expensive step is usually done once , followed by a relatively fast fine-tuning step. In fine-tuning, the pre-trained model is tuned on a specific dataset and task. Fine-tuning may be performed on a relatively small dataset very efficiently for specific usage. Pre-training followed by fine-tuning is referred to as transfer learning.

Transformer models

Transformers may be roughly split into three classes. At first we have autoencoding transformer models that only use an encoder and they are suitable for natural language understanding. These encoder transformers perform well in tasks such as question answering, sentence classification and other tasks that require to understand the whole sentence. In this category we include BERT [20] and its improvements such as RoBERTa [56]. BERT is based on the Transformer architecture and uses a mask token for training but not for testing. BERT predicts multiple mask tokens in parallel without modeling direct dependencies between different predictions. RoBERTa improves on BERT's training and results by fine-tuning hyperparameters. Next, we have auto-regressive transformers that use only a decoder. These architectures are suitable for text generation. In this category we can fine models such as GPT [62], GPT-2 [63] and GPT-3 [11]. Finally, architectures that have both an encoder and a decoder called sequence-to-sequence transformers. In this category we can find models like BART [48] and T5 [65]. Such models are suitable for translation, summarization, paraphrasing and question answering by generation.

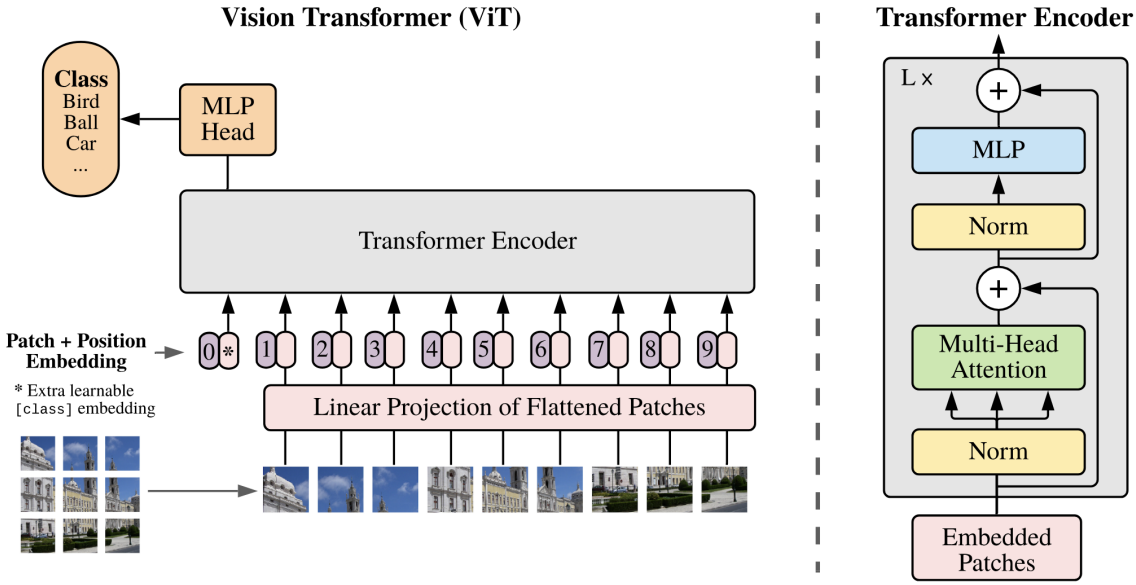


Figure 3.3.1: **ViT.** Model overview [22]

3.3 Vision Transformer

3.3.1 Overview

Vision Transformers [22] (ViTs) represent a transformative advancement in the field of computer vision, applying the principles of transformer architectures from natural language processing (NLP) to image data. By leveraging self-attention mechanisms, ViTs have redefined the capabilities and performance of image recognition models, achieving state-of-the-art results across various benchmarks and applications.

The journey of transformers began in NLP with models like BERT [20] (Bidirectional Encoder Representations from Transformers) and GPT [11] (Generative Pre-trained Transformer). These models revolutionized language processing by effectively capturing contextual relationships within text data through self-attention mechanisms. Inspired by this success, researchers adapted the transformer architecture to handle visual data, leading to the development of Vision Transformers. This adaptation involves processing images in a manner that enables the model to learn and infer complex visual patterns and relationships.

3.3.2 Architecture

The core innovation of Vision Transformers is the self-attention mechanism, which enables the model to weigh the importance of different parts of the input data dynamically. For an image divided into patches, the self-attention mechanism calculates the relationships between patches to understand the image as a whole. In Figure 3.3.1 is presented the process that described above.

Given an input image $x \in \mathbb{R}^{H \times W \times C}$, the image is divided into $N = \frac{HW}{P^2}$ patches of size $P \times P$. Each patch x_p is then flattened into a vector $x_p \in \mathbb{R}^{P^2 \cdot C}$. These flattened patches are linearly transformed into embeddings

$$z_0^i = W_p x_p^i + b_p, \tag{3.3.1}$$

where $W_p \in \mathbb{R}^{D \times P^2 C}$ and $b_p \in \mathbb{R}^D$. A classification token z_0^0 is added to the sequence of patch embeddings, and positional embeddings $E_{pos} \in \mathbb{R}^{(N+1) \times D}$ are added to form

$$z_0 = [z_0^0; z_0^1; \ldots; z_0^{N-1}] + E_{pos} \tag{3.3.2}$$

The sequence z_0 is processed by L layers of a Transformer encoder. Each layer consists of Layer Normalization, Multi-Head Self-Attention (MHSA), and a Feed Forward Network (FFN). Specifically, the processing within each layer is defined as:

$$hatz_{\ell-1} = \text{LayerNorm}(z_{\ell-1}), \tag{3.3.3}$$

$$z'_\ell = \hat{z}_{\ell-1} + \text{MHSA}(\hat{z}_{\ell-1}), \tag{3.3.4}$$

and

$$z_\ell = z'_\ell + \text{FFN}(\text{LayerNorm}(z'_\ell)) \tag{3.3.5}$$

After L Transformer encoder layers, the classification token z_0^0 is used for the final classification. The output is obtained by applying Layer Normalization followed by a Multi-Layer Perceptron (MLP) head:

$$\hat{z}_L = \text{LayerNorm}(z_L), \tag{3.3.6}$$

and

$$y = \text{MLP}(\hat{z}_L^0), \tag{3.3.7}$$

where

$$\text{MLP}(x) = xW_{head} + b_{head} \tag{3.3.8}$$

In summary, the Vision Transformer can be formulated as

$$y = \text{MLP}(\text{LayerNorm}(z_L^0)) \tag{3.3.9}$$

where z_L is the output after L Transformer encoder layers applied to the sequence of image patches with added positional embeddings.

3.3.3 Applications of Vision Transformers

Vision Transformers have achieved state-of-the-art performance in image classification tasks, surpassing traditional CNN-based models on benchmark datasets like ImageNet [19]. By leveraging their ability to capture global context, ViTs can accurately classify images across diverse categories. In object detection and segmentation, Vision Transformers excel by attending to relevant regions of an image and understanding their spatial relationships. This ability leads to precise localization and segmentation of objects, even in cluttered or complex scenes. ViTs are also used in image generation and understanding tasks, such as image synthesis, image captioning, and visual question answering. By modeling the joint distribution of images and textual descriptions, ViTs facilitate multimodal interaction and content generation.

3.3.4 Challenges

Regardless of their impressive capabilities, Vision Transformers face several challenges. The quadratic complexity of the self-attention mechanism can be computationally expensive, especially for high-resolution images. Techniques like sparse attention and efficient transformers are being explored to mitigate this issue. Furthermore, Vision Transformers typically require large amounts of training data to achieve optimal performance. Enhancements in data augmentation and semi-supervised learning can help reduce this dependency. Finally, understanding the decision-making process of Vision Transformers can be challenging due to their complex architecture. Research in explainable AI aims to make these models more transparent and interpretable.

3.4 T5

3.4.1 Overview

The T5 model [65], or "Text-To-Text Transfer Transformer," is a transformative neural network model developed by researchers at Google AI. Introduced in 2019, T5 is designed to handle a wide variety of natural language processing (NLP) tasks within a unified text-to-text framework. This groundbreaking approach involves converting all NLP tasks into a text generation problem, allowing T5 to utilize a single model architecture to address tasks such as translation, summarization, question answering, and classification.

3.4.2 The Text-to-Text Transfer Transformer Framework

The primary innovation of T5 is its text-to-text framework (Figure 3.4.1). Traditional NLP models often require task-specific architectures and loss functions. For instance, classification tasks typically use models that output a probability distribution over classes, while translation tasks generate sequences of words. T5 simplifies this by converting all tasks into a text generation problem, creating a more flexible and cohesive approach to NLP.

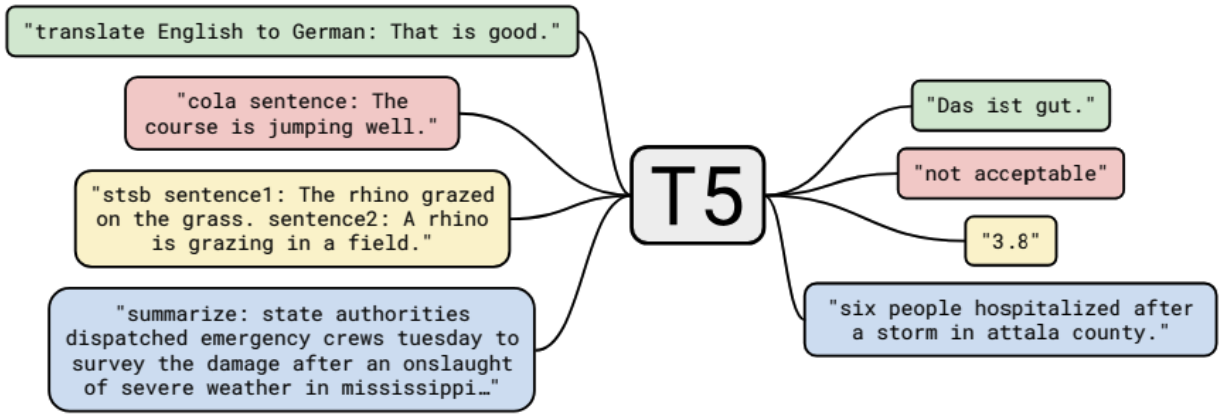


Figure 3.4.1: **Diagram of the T5 framework** [65]

In T5, both inputs and outputs are sequences of text. This means that the model can be trained on a single objective and can handle various tasks by interpreting and generating text strings. For example, for a translation task:

> Input: "translate English to French: How are you?"
>
> Output: "Comment ça va ?"

For a summarization task, in the same way we have:

> Input: "summarize: The quick brown fox jumps over the lazy dog"
>
> Output: "The fox jumps over the dog"

Finally in a question answering task, we could ask the model a question as it is shown below:

> Input: "question: Who wrote 'Pride and Prejudice'? context: 'Pride and Prejudice' is a novel by Jane Austen."
>
> Output: "Jane Austen".

This text-to-text paradigm simplifies the process of training and fine-tuning the model across different tasks, leveraging a consistent training objective. Moreover, it enables the use of transfer learning, where the knowledge gained from one task can be beneficial to another, improving overall performance.

3.4.3 Architecture and Training

T5 is based on the Transformer architecture, specifically employing the encoder-decoder structure originally introduced by Vaswani et al. [88]. The Transformer architecture is known for its self-attention mechanisms, which allow the model to weigh the relevance of different parts of the input data dynamically.

The T5 model as a Sequence-to-Sequence Transformer pertain to Encoder-Decoder architecture category, as we discussed in Section 3.2.3. An example of the model stracture is given in Figure 3.4.3. The encoder processes the input text sequence and transforms it into a set of context-rich representations, consisting of multiple layers of self-attention and feed-forward neural networks. The decoder generates the output text sequence based on the encoded representations and the previously generated tokens in the output sequence. It

also consists of multiple layers of self-attention and feed-forward neural networks and attends to the encoder's output representations as a sequence-to-sequence transformer.

T5 was trained on a massive corpus derived from the C4 dataset (Colossal Clean Crawled Corpus) [65], which is a cleaned and filtered version of web-crawled text data. The training process involved multiple phases. First, the model was pre-trained. T5 uses a denoising autoencoder objective where spans of text are masked, and the model is tasked with reconstructing them (Figure 3.4.2). This encourages the model to understand the context and generate coherent sequences. After that, during fine-tuning and generation, T5 uses a technique called teacher forcing. In this approach, the model receives the correct sequence of previous tokens as input during training, which helps stabilize and speed up the training process. For text generation tasks, T5 often employs beam search, a decoding strategy that maintains multiple hypotheses at each time step and selects the sequence with the highest overall probability. This approach helps produce more accurate and coherent outputs.

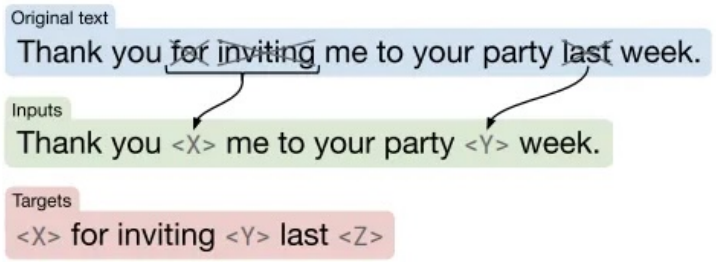


Figure 3.4.2: **T5 pre-training unsupervised objective.** In the original text, some words are dropped out with a unique sentinel token. Words are dropped out independently uniformly at random. The model is trained to predict basically sentinel tokens to delineate the dropped out text. [65]

Implementation Details

T5 is available in multiple sizes, ranging from small models like T5-Small with 60 million parameters to large models like T5-11B with 11 billion parameters. This range allows users to balance performance with computational resources. Key hyperparameters include the number of layers, the size of hidden states, the number of attention heads, and the learning rate are fine-tuned to optimize performance across various tasks. Furthermore, there is available a multilingual variation of the T5 model, called mT5 [96].

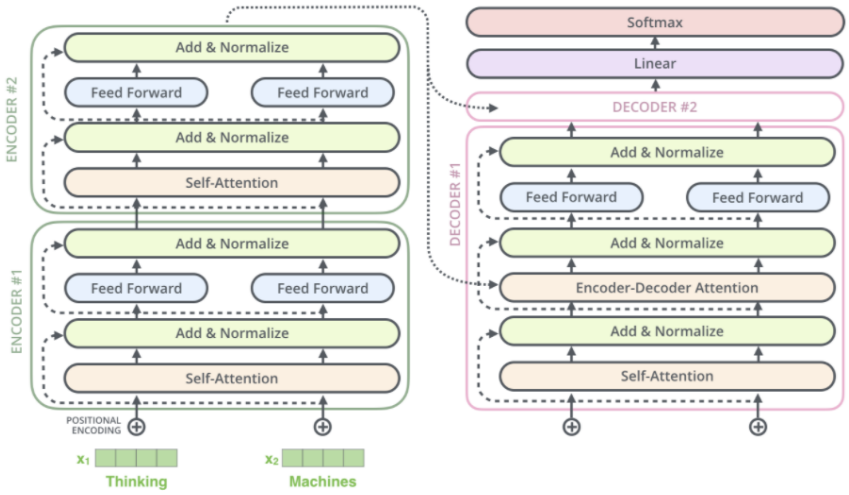


Figure 3.4.3: **T5 Architecture.** The T5 structure is just a standard vanilla encoder-decoder transformer

3.4.4 Applications

T5 has demonstrated state-of-the-art performance across a wide range of benchmarks and tasks. Some notable achievements and applications include language translation, summarization, question answering, and text classification.

In machine translation, T5 can translate text between numerous language pairs with high accuracy, often matching or surpassing the performance of specialized translation models. In summarization, it produces concise and coherent summaries of long texts, making it valuable for applications like news aggregation and report generation. For question answering, T5 excels in extracting and generating accurate answers from given contexts, performing well on benchmarks like SQuAD (Stanford Question Answering Dataset) [66]. In text classification, by converting classification tasks into text generation problems, T5 can label text data effectively, such as identifying sentiment or topic categories. Additionally, T5's versatility allows it to handle tasks like text completion, paraphrasing, and dialogue generation, making it a highly flexible tool for various NLP applications.

T5's versatility makes it suitable for various real-world applications. In healthcare, T5 can summarize lengthy medical reports, making it easier for healthcare professionals to review patient information quickly. Additionally, T5 can provide accurate answers to medical questions based on large medical text corpora, aiding in patient care and research. In the legal industry, T5 can analyze and summarize legal documents, helping lawyers and legal professionals save time and reduce manual effort. It can also provide answers to legal questions by referencing relevant statutes and case law, enhancing legal research. In customer service, T5 can be used to generate responses to customer queries, improving the efficiency and accuracy of customer service operations. By classifying the sentiment of customer feedback, T5 can help organizations better understand and respond to customer needs.

3.4.5 Challenges

Despite its versatility and performance, T5, like other large language models, faces several challenges. First, training and fine-tuning T5 require substantial computational power and memory, which can be a barrier for many researchers and organizations. The large model size and the need for high-performance hardware make it less accessible. Second, T5 can inherit biases present in the training data, leading to biased or unfair outputs. Continuous efforts are necessary to identify and mitigate these biases, especially when deploying the model in sensitive applications. Third, understanding the decision-making process of T5 remains a complex task, making it difficult to explain why the model generates certain outputs. This lack of interpretability can be problematic in applications where transparency is crucial. Finally, the ability of T5 to generate human-like text raises concerns about its potential misuse, such as generating misinformation or deepfake text.

3.5 Multimodal Transformer (MulT)

3.5.1 Overview

Instead of modeling language and vision independently, multi-modal transformers introduced techniques to use these modalities together with various applications in multimodal sentiment analysis, search and generation. For example, DALL-E [67] is a generative model that is trained jointly in text and images. Then the model receives a text description of an image as its input and generates an image that matches that description. Another example of multi-modal transformer is MulT [85]. A sequence-to-sequence transformer used for modeling human language time series. MulT fuses the two modalities using a cross-modal attention unit. The cross-modal transformer serves to repeatedly reinforce a *target* modality with low-level features from another *source* modality. One of the advantages of MulT is that it performs exceptionally for unaligned sequences while at the same time it captures long-range dependencies across deifferent modalities.

3.5.2 Core Idea

More specifically, MulT merges unaligned multimodal time-series via feed-forward cross-modal attention. Let two modalities α and β with two unaligned sequences from each denoted as $X_\alpha \in \mathbb{R}^{L_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{L_\beta \times d_\beta}$ respectively. $L_{(.)}$ represents the lenght of the sequence and $d_{(.)}$ the feature dimension. In the original paper,

the authors inspired by the decoder architecture introduced by Vaswani [88] for Neural Machine Translation, which translates one language to another. In the same way the proposed cross-modal attention on MulT tries to fuse cross-modal information by latent adapting the one modality to the other. The overall architecture is shown on Figure 3.5.1 taken from the original paper [85].

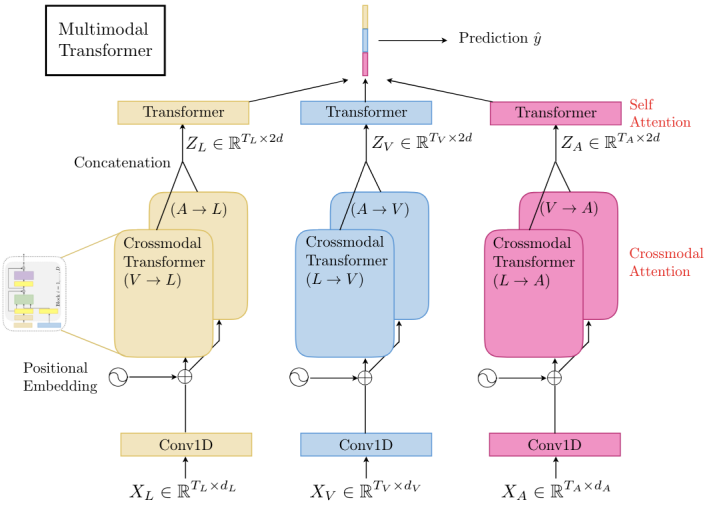


Figure 3.5.1: **Overall MulT architecture on modalities (L, V, A).** The crossmodal transformers, which suggest latent crossmodal adaptations, are the core components of MulT for multimodal fusion. [85]

3.5.3 Cross-modal Attention

For the cross-modal attention unit we define $Queries$, $Keys$ and $Values$ as $Q_\alpha = X_\alpha W_{Q_\alpha}$, $K_\beta = X_\beta W_{K_\beta}$ and $V_\beta = X_\beta W_{V_\beta}$ respectively. Where $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_K}$, $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_K}$ and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_V}$. Cross-modal attention is the latent adaption from modality β to α presented as $CM_{\beta \to \alpha}(X_\alpha, X_\beta) \in \mathbb{R}^{L_\alpha \times d_V}$. From now on, we define $Y_\alpha \coloneqq CM_{\beta \to \alpha}(X_\alpha, X_\beta)$ for easiest notation. It is noticeable that Y_α has the same length as Q_α, however it is presented in the vector space of V_β. Because cross-modal attention is basically a cross-attention mechanism we can easily derive that:

$$Y_\alpha = softmax\left(\frac{Q_\alpha K_\beta^\intercal}{\sqrt{d_K}}\right) \cdot V_\beta = softmax\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^\intercal X_\beta^\intercal}{\sqrt{d_K}}\right) \cdot X_\beta W_{V_\beta}$$

Now is more clear that Y_α computes a score matrix with the $softmax(\cdot) \in \mathbb{R}^{L_\alpha \times L_\beta}$ function which in the (i, j) position contains the attention given between the $i-th$ timestep of modality α to the $j-th$ timestep of modality β. Finally, the cross-modal transformer can be a deep stacking of several cross-attention blocks.

3.6 Image Captioning

3.6.1 Overview

Image captioning is among the key uses of foundation models that employ both image and text models.

Definition 3.6.1: Image Captioning

> *Image Captioning* is the task where we describe a visual content of an image in natural language, by using a visual understanding system along with a language model that can produce syntactically correct and meaningful sentences about the given concept.

The main goal of this field is to find the most effective pipeline that can understand the visual information the image contains, represent it and transform it in a text sequence which captures the connections between the visual and textual concepts.

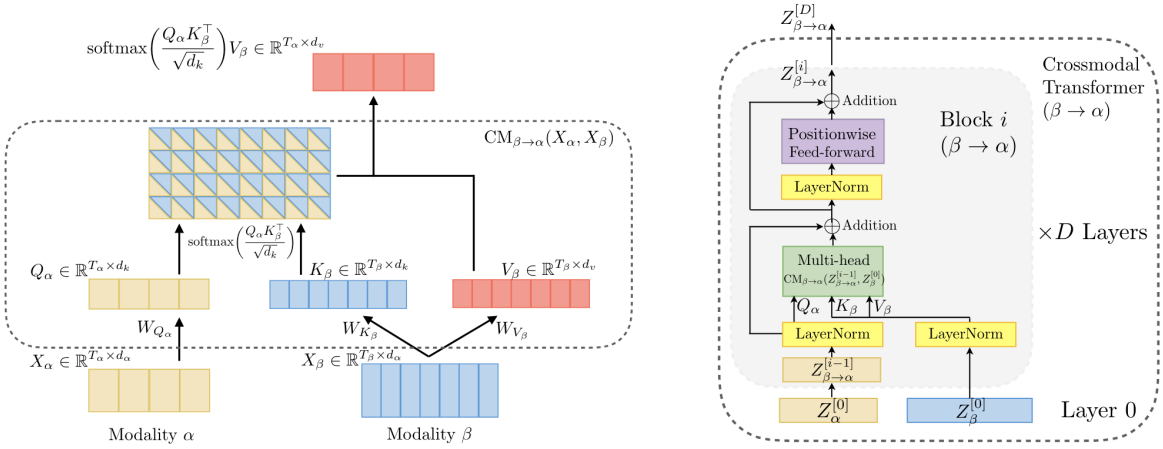


Figure 3.5.2: The **left** picture illustrates the **crossmodal attention** $CM_{\beta\to\alpha}(X_\alpha, X_\beta)$ between sequences X_α, X_β from modalities α, β. The **right** picture depicts the deep-stacking of crossmodal attention blocks that form the **multimodal transformer**. [85]

In recent years, two models have been prominent in the context of image captioning. The first is Flamingo [1], a model that operates by interleaving image and text tokens. This allows the model to perform captioning with a guided prompt, by combining the image tokens with those of the prompt into a single sequence. The second is BLIP-2 [50], which uses an extra transformer module (called a querying transformer) to combine the image and text modalities of the underlying models, allowing image captioning by passing information from the visual module to the language model. This follows the paradigm of foundation models, in that parts of the large pretrained models can be used as modules in architectures for a variety of tasks, with only a small part added to enable interaction between the two. Moreover, we shall also notice LLaVA [55], which connects an vision encoder and a large language model and trains them on produced data while uses instruction tuning. LLaVA is used for general purpose language-vision understanding and presents outstanding multimodal conversational abilities.

3.6.2 BLIP-2 (Bootstrapping Language-Image Pre-training)

BLIP-2 [50] represents a significant advancement in the field of vision-language pre-training. It introduces an efficient and scalable method to combine visual and textual information using pre-trained models. The core innovation lies in its ability to leverage off-the-shelf frozen image encoders and large language models (LLMs) to achieve state-of-the-art performance in various vision-language tasks while significantly reducing computational costs.

Architecture

The central module of BLIP-2 is the **Querying Transformer (Q-Former)**, which serves as a bridge between the frozen image encoder and the LLM. The Q-Former comprises two transformer submodules: one for visual feature extraction from the image encoder and another that functions as both a text encoder and decoder. This architecture allows for the extraction of fixed output features from images regardless of their resolution, facilitating more efficient and focused learning. This can also be seen in Figure 3.6.1.

Two-Stage Pre-Training Process

BLIP-2 employs a two-stage pre-training process. In the first stage, the Q-Former is connected to a frozen image encoder and trained using image-text pairs. This stage employs three main pre-training objectives: *Image-Text Contrastive Learning (ITC)*, which aligns image and text representations by maximizing mutual information; *Image-grounded Text Generation (ITG)*, which trains the model to generate text based on visual inputs, forcing the Q-Former to capture essential visual features; and *Image-Text Matching (ITM)*, which

enhances fine-grained alignment between image and text representations through a binary classification task. This process is illustrated in Figure 3.6.2

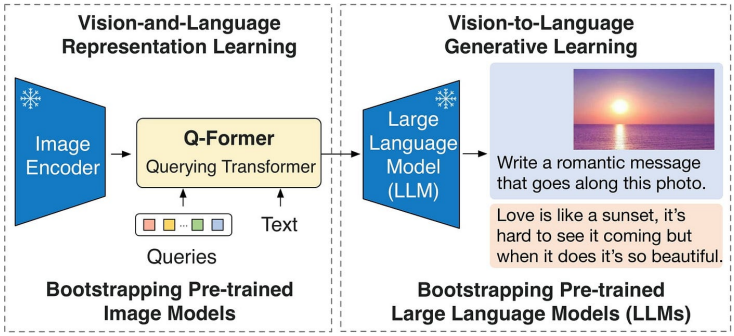


Figure 3.6.1: **BLIP-2 overall architecture.** It uses the lightweight Q-Former to bridge the gap between image and text modalities.[50]

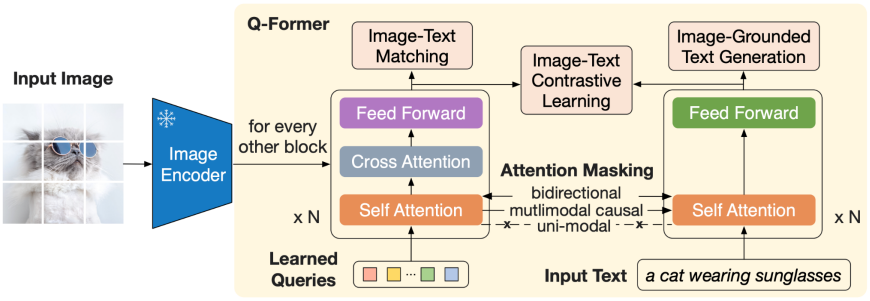


Figure 3.6.2: **First stage of BLIP's pre-training.** Model architecture of Q-Former and BLIP-2's first-stage vision-language representation learning objectives. BLIP-2 jointly optimizes 3 objectives that encourage the Q-Former's queries extract image representation most relevant to the text.[50]

In the second stage, the Q-Former, along with the frozen image encoder, is connected to a frozen LLM. This setup utilizes the generative capabilities of the LLM, with the Q-Former providing crucial visual context through projected query embeddings. This stage is crucial for tasks that require generating text based on visual inputs, such as image captioning and visual question answering. Figure 3.6.3 provides a visualisation of the second stage of BLIP's pre-training.

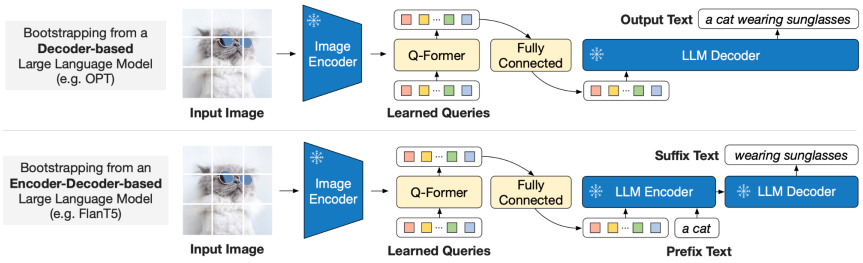


Figure 3.6.3: **Second stage of BLIP's pre-training.** We can see how BLIP-2 bootstrapping a decoder-based LLM on top while an encoder-decoder-based LLM on bottom. The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.[50]

BLIP's contribution

BLIP-2's architecture and training approach make it highly versatile for a range of applications, including image captioning, visual question answering, and image-text retrieval. Its novel approach of bootstrapping

from frozen pre-trained models sets a new standard for vision-language pre-training. By reducing the need for extensive computational resources and focusing on efficient feature extraction and alignment, BLIP-2 paves the way for more accessible and powerful vision-language applications.

3.6.3 LLaVA (Large Language and Vision Assistant)

LLaVA (Large Language and Vision Assistant) [55] represents a significant advancement in the fusion of language and vision models. It builds upon the foundational architecture of Vicuna [15], a large language model known for its robust instruction-following capabilities, and integrates it with a visual encoder from CLIP [64].

Visual Instruction Tuning

The training process of LLaVA involves a novel approach called *Visual Instruction Tuning*. This method combines language and vision by leveraging pre-existing, text-only large language models like GPT-4 [60] or ChatGPT to generate instruction-following data based on visual content. The visual data is encoded using symbolic representations such as captions and bounding boxes, which describe the visual scene and localize objects within it. This enables the creation of a rich dataset consisting of three types of instruction-following samples: conversations, detailed descriptions, and complex reasoning tasks.

For instance, the conversation samples involve designing a dialogue where an assistant answers questions about an image, as if it is interpreting the visual content in real time. Detailed descriptions provide comprehensive narratives about the image, while complex reasoning tasks require deeper logical inference based on the visual content. This method has resulted in a collection of 158,000 unique language-image instruction-following samples, enhancing the model's ability to understand and respond to multimodal queries

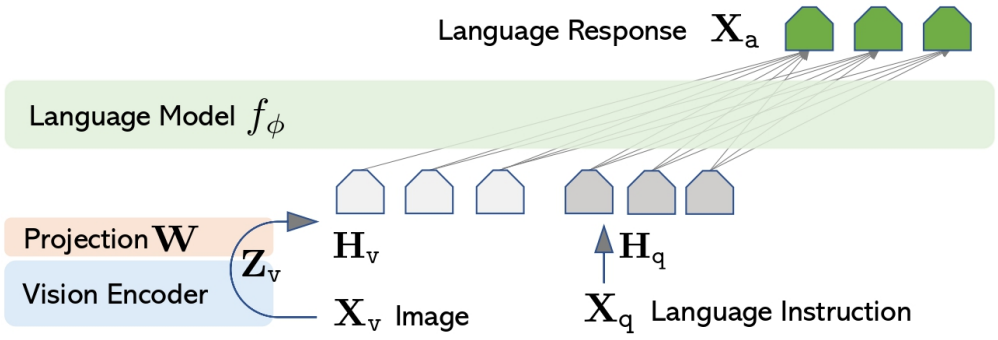


Figure 3.6.4: **LLaVA architecture.** For an input image X_v, the pre-trained CLIP visual encoder provides the visual feature $Z_v = g(X_v)$. Then, we apply a trainable projection matrix W to convert Z_v into language embedding tokens H_q, which have the same dimensionality of the word embedding space in the language model. Thus, we have a sequence of visual tokens H_v along with the instruction embeddings H_q to feed in the LLM.

Evolution: LLaVA v1.5

LLaVA has undergone several iterations, with LLaVA v1.5 [54] marking a significant milestone. This version improves upon the original by incorporating a more sophisticated vision-language connector, transitioning from a simple linear projection to a multi-layer perceptron (MLP). This change has enhanced the model's ability to represent and process visual information. LLaVA v1.5 also benefits from an expanded dataset, including additional academic-task-oriented VQA datasets, which further boost its performance across various benchmarks.

3.7 Visual Question Answering (VQA)

3.7.1 Overview

Visual Question Answering (VQA) is an interdisciplinary field that combines computer vision and natural language processing (NLP) to create systems capable of answering questions about images. The goal of VQA is to enable machines to understand and reason about visual content in a manner that is similar to human comprehension. This involves not only recognizing objects and scenes within an image but also interpreting and contextualizing this visual information to provide accurate answers to textual questions.

Definition 3.7.1: Visual Question Answering (VQA)

> *Visual Question Answering* is the task of developing AI systems that can provide accurate and relevant answers to questions posed about images.

The development of VQA systems has been fueled by advances in deep learning, particularly in convolutional neural networks (CNNs) for image processing and recurrent neural networks (RNNs) or transformers for handling text. Early VQA models primarily relied on extracting features from images and questions independently before merging this information to generate answers. Over time, more sophisticated architectures have been developed that utilize attention mechanisms, multimodal embeddings, and joint training strategies to better capture the complex relationships between visual and textual data.

3.7.2 VQA System Architecture

A VQA system comprises several stages: **image feature extraction**, **question encoding**, **multimodal fusion** and **answer prediction**. The first step in a VQA system involves extracting meaningful features from the input image. This is commonly achieved using pretrained image encoders, such as CNNs (ResNet[34], VGG [76]) Transformers (ViT [22]) which produce a rich representation of the image in the form of feature maps. Furthermore, the textual question is processed to create a feature representation that captures its semantic meaning. Many works using RNNs (like LSTMs [39], GRUs [16]) or transformer-based models (like BERT [20]), which can encode the question into a fixed-length vector or a sequence of embeddings. Next, the core challenge in VQA is to effectively combine the visual and textual features to generate a coherent representation that can be used to infer the answer. Various fusion techniques have been proposed, including:

- **Concatenation**: Directly combining the feature vectors from the image and question.
- **Element-wise Addition/Multiplication**: Performing element-wise operations to merge the features.
- **Attention Mechanisms**: Applying attention to focus on relevant parts of the image based on the question, enhancing the interaction between modalities.
- **Multimodal Transformers**: Utilizing transformer architectures that simultaneously process and integrate visual and textual information.

Finally, The fused multimodal representation is passed through one or more fully connected layers to predict the answer. The output layer typically uses a softmax activation function to produce a probability distribution over a predefined set of possible answers.

3.7.3 Datasets and Benchmarks

Several datasets and benchmarks have been created to evaluate the performance of VQA systems.

VQA Dataset

The VQA dataset [4] is one of the largest and most widely used benchmarks for evaluating VQA models. It contains real-world images sourced from the COCO dataset [53], accompanied by human-annotated questions and answers. The dataset covers a wide range of question types, including object recognition, counting, and common-sense reasoning. Some examples of VQAv1 dataset are shown in Figure 3.7.2a.

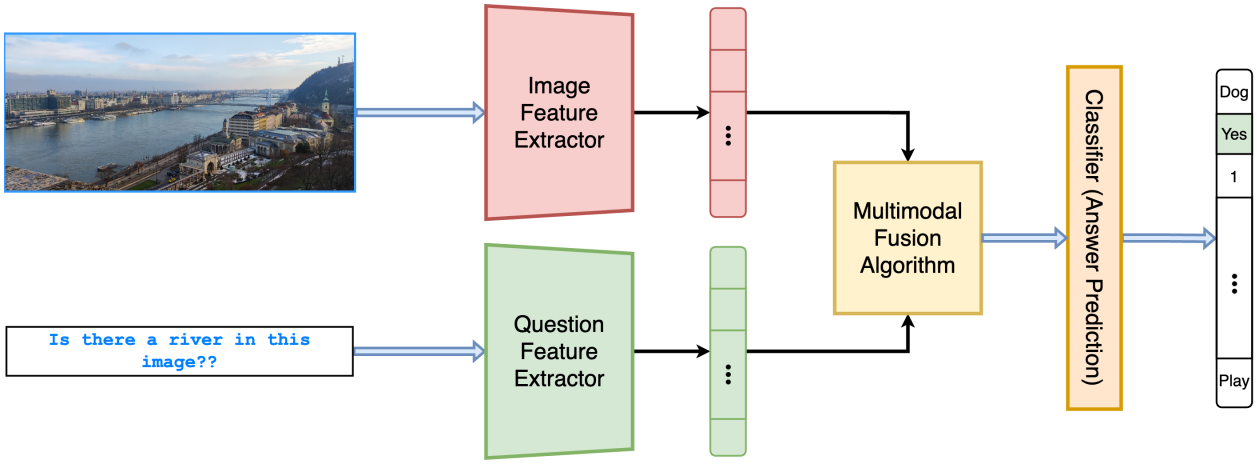


Figure 3.7.1: A typical **VQA architecture**

Visual Genome

The Visual Genome dataset [45] provides detailed annotations of objects, attributes, and relationships within images, making it a valuable resource for training and evaluating VQA models that require a deeper understanding of scene context and interactions.

CLEVR

CLEVR dataset [43] is designed to test the compositional reasoning abilities of VQA models. It consists of synthetic images and questions that require understanding of spatial relationships, counting, and logical operations.

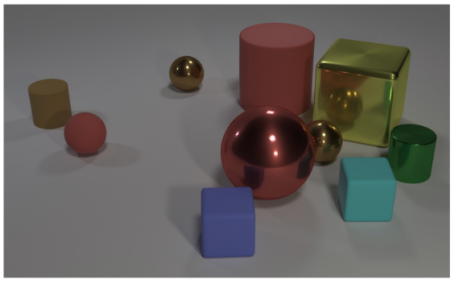
3.7.4 Challenges

Despite significant progress, VQA systems still face several challenges. Current VQA models often struggle with questions that require *compositional reasoning*, such as those involving multiple objects and their relationships. Improving the ability of models to understand and reason about complex scenes remains a key area of research. Furthermore, VQA models are *susceptible to biases* present in the training data, leading to poor generalization to unseen questions or images. Techniques such as adversarial training and balanced datasets are being explored to mitigate these biases. Finally, understanding the decision-making process of VQA models is crucial for building trust and reliability. At this time VQA systems lack *explainability*. Research in explainable AI aims to develop methods for visualizing and interpreting the reasoning paths taken by the VQA systems.

3.7.5 Considerations

Visual Question Answering represents a significant step towards creating AI systems that can understand and interact with the world in a human-like manner. By combining advances in computer vision and natural language processing, VQA systems have achieved impressive results across a variety of tasks and applications. However, challenges remain in areas such as compositional reasoning, bias mitigation, and explainability. As research in VQA continues to advance, it holds the potential to transform how machines perceive and understand visual information, paving the way for more intelligent and interactive AI systems. As the field evolves, VQA stands out as a promising and rapidly advancing area of artificial intelligence that bridges the gap between vision and language.

(a) VQAv1 Dataset [4]



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

(b) CLEVR [43]

(c) Visual Genome [45]

Figure 3.7.2: **VQA dataset examples.**

Chapter 4

The Proposed Methods

4.1 Introduction

Image-text models have become fundamental in machine learning, giving rise to several state-of-the-art architectures, such as CLIP, DALL-E and Stable Diffusion, among others [64, 67, 68, 71]. These foundational models can be used in a variety of tasks, usually different than the one they were trained on. This is because they use the each modality to infer knowledge for the former, therefore allowing them to operate without having seen data for the specific task at hand. At the same time, these architectures are also very costly to train. Training is often done on millions of image-text pairs. As such, using these large foundational models often becomes a task of **finetuning** them, rather than training them from scratch. These models can also be used simply for inference, and use them as the starting point around which a larger pipeline can be constructed.

4.2 Augmentation via MulT

In our first approach to enhance CLIP, we use MulT [85] to extract a new multimodal embeddings based on the image-text pair along with the two vectors the original modalities yield. At first, in order to create the fused multimodal embedding vectors we have to split the image into patches which are suitable as an input for MulT. MulT uses 1D positional embeddings, so as a way to not interfere with MulT's positional embedding we flatten our image in different axes in order to create additional vectors. We choose to use the same 1D positional encoding since there are not significant performance gains from using 2D-aware positional embeddings, as stated in ViT paper [22].

By creating different views of the same image-text pair we can enhance our data with augmentations of the same object. We project the new vector(s) on the joint multimodal embedding space along with the image and text embeddings. We follow CLIP's training pipeline by jointly training the MulT, image and text encoders to maximize the cosine similarity of the N true pairs and minimize the cosine similarity for the rest. Our new task is to optimize a weighted sum of each pair's symmetric cross-entropy loss score over these similarities. The overall architecture is shown on Figure 4.2.1.

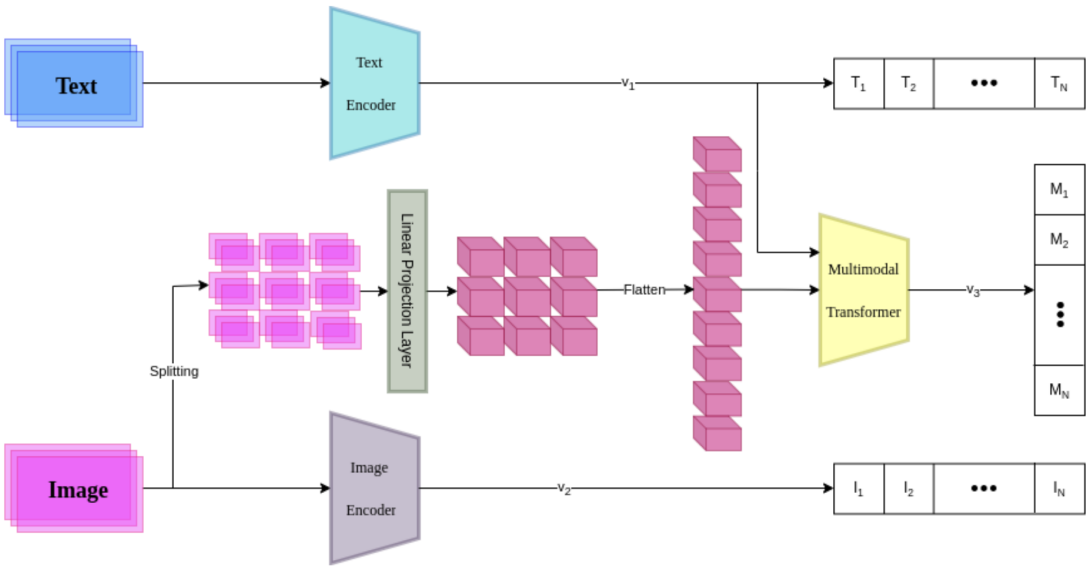


Figure 4.2.1: **Augmentation via MulT.** Overall architecture of the augmented CLIP model via MulT.

4.2.1 Multimodal Fusion

For a deeper understanding of the method's mechanics, assume an image $x_I \in \mathbb{R}^{H \times W \times C}$ and a text sequence $x_T \in \mathbb{R}^{\mathrm{L}}$,where (H, W) is the resolution of the image, C is the number of channels and L is the length of

the sentence. As in CLIP, we create minibatches $X_I \in \mathbb{R}^{N\times H\times W\times C}$and $X_T \in \mathbb{R}^{N\times L}$ of aligned images and texts. These minibatches serve as input for the image encoder and text encoder respectively. We project the output of each encoder on a joint multimodal space. Assume we have $z_I \in \mathbb{R}^{N\times D_I}$ and $z_T \in \mathbb{R}^{N\times D_T}$ on the output of each encoder and $W_I \in \mathbb{R}^{D_I\times D_E}$ and $W_T \in \mathbb{R}^{D_T\times D_E}$ two learned projections of each modality to the joint embedding space. D_T is the output dimension of the text representations, D_I is the dimension of the image representations and D_E is the dimension of the joint embedding space. So we obtain two vectors $v_1 \coloneqq z_T W_T = T \in \mathbb{R}^{N\times D_E}$ and $v_2 \coloneqq z_I W_I = I \in \mathbb{R}^{N\times D_E}$.

Because of the Multimodal Transformer's need to receive a sequence of each modality we reshape our initial image x_I on a sequence of 2D flattened patches$\hat{x_I} \in \mathbb{R}^{\mathrm{M}\times(P^2\cdot C)}$, where P is the size of the rectangle patches and $\mathrm{M} = \frac{HW}{P^2}$ is the length of the sequence of patches. For the text, we simply use our initial sequence of token embeddings x_T as an input on MulT. Let $z_M \in \mathbb{R}^{N\times D_M}$be the output of the Multimodal Transformer and $W_M \in \mathbb{R}^{D_M\times D_E}$a learned projection to the joint multimodal space. The fused multimodal vector occurs to be the $v_3 \coloneqq z_M\, W_M = M \in \mathbb{R}^{N\times D_E}$.

4.2.2 Training MulT

As we showed before, we yield three vectors for each image-text pair. These vectors serve as views of the pair. Specifically we have I: image embeddings, T: text embeddings and M: MulT embeddings. Similarly to ConVIRT [100] at training time we sample a minibatch of N input triplets (v^T, v^I, v^M) where (v_i^T, v_i^I, v_i^M) is the i-th triplet on the batch. The loss between text and image is the same as CLIP loss, which is a symmetric contrastive loss for image-to-text and text-to-image. We use the same formulation as in Equation 3.1.3 for $CE(I,T) = \mathcal{L}_{CLIP}$. In the same way we can derive $CE(T,M)$ and $CE(M,I)$. As our training loss we propose a weighted sum of the three symmetric cross-modal losses.

$$\mathcal{L}(I,T,M) = a\cdot CE(I,T) + b\cdot CE(T,M) + (1-a-b)\cdot CE(M,I) \tag{4.2.1}$$

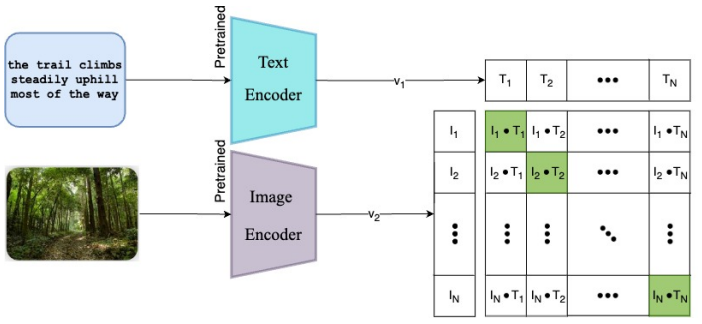
Parameters a, b must be $0 < a, b < 1$ and $a + b < 1$. $CE(\cdot)$ stands for the symmetric cross-entropy loss score.

After further experimentation, this method did not yield significant results. Despite its lack of effectiveness, it provided us with the foundational insights for the subsequent works presented in the following sections.

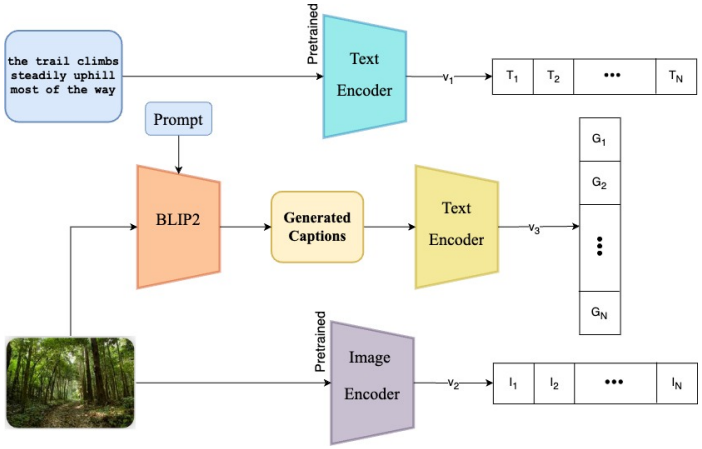
4.3 Enhancing CLIP with Generative Dialogue: Third Tower Approach

The motivation behind our work is derived from computer vision and the use of additional views [51, 94] of the same object in order to obtain more information about its characteristics. This idea was adapted on contrastive learning by CMC [84], where additional sensory views of the same image where used in order to enhance contrastive training. In the same way, in the case of image-text pairs we find an additional view in dialogues generated by generative models. This extra synthetic data will still be text but of a different nature compared to the captions that were used for the text encoder. We thus consider auxiliary textual datasets like metadata, dialogues about an image or product details obtained from a database to be additional modalities.

Within this context, we examine the use of a **third tower** in these image-text architectures. The addition of the third tower can be seen in Figure 4.3.1. We focus our study on a CLIP model trained by OpenCLIP [41], which we augment via an additional encoder, resulting in our **CLIP-3Modal** [86] architecture. This encoder serves to add additional modalities to the input, in a way that is consistent with the paradigm of reusing elements from foundation models. In contrast to previous works that examine the use of a third tower in the architecture [44], we explicitly consider the third tower as operating on a different modality, aside from the usual image and text ones. In our setting, we consider the additional modality to be the dialogue of a user with an image-captioning model such as BLIP-2 [50]. With this, we aim to augment the information from our input data with the outputs of a foundation model. We evaluate the performance of this addition to the CLIP architecture via retrieval performance on an image-text retrieval task.



(a) CLIP



(b) CLIP-3Modal

Figure 4.3.1: **Our proposed CLIP-3Modal architecture.** We propose incorporating a third tower in the CLIP model, which extends the existing image and text ones. This extra tower can be used during the evaluation of the model, along with the existing ones.

4.3.1 Using BLIP-2 for Captions

To incorporate the third tower into our architecture, we need to include a third modality in our input data. To do this, we use BLIP-2 [50] to expand an image text dataset with an additional modality. We choose CC3M [74] as the dataset that we enrich with BLIP-2 generated captions. For each image, we provide it as input to the BLIP-2 model. We then provide the following two questions, in sequence, as our prompt:

> "What do you see in this image?"
>
> "What makes this image unique?"

This can also be seen in Figure 4.3.2. This pair of questions provides a basic form of dialogue between the model and a user. Moreover, when providing the second question to the captioning model, we also give it the response to the first question as input. This allows for the response to the second question to be contextualized by the model by its own response to the first one.

We note here that the model that generates the third modality does not take the text from the image-text pairs as input. This means that, despite the third modality being provided to the training process in text format, the actual content is not directly dependent on the existing caption of the image. This allows the BLIP-2 model to provide new information about this particular sample.

4.3.2 Training the Third Tower

We now aim to train a CLIP architecture that incorporates a third tower in its construction, using BLIP-2 captions as input. To do this, we start from a pretrained CLIP model provided by OpenCLIP. The image and text towers from this model become the image and text towers for our architecture as well. To construct

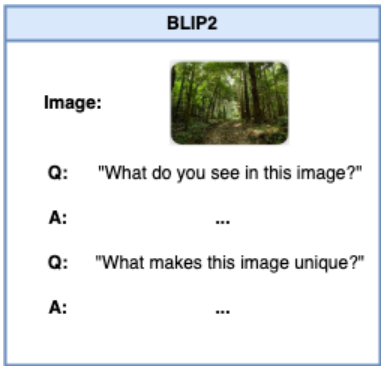


Figure 4.3.2: **Use of BLIP-2 model.** The questions with which the model is prompted provide a base form of dialogue for use as our third modality.

the third tower of our architecture, we start by making a copy of the pretrained text tower. We then freeze the original image and text towers, and train only the third tower on our extended CC3M dataset. Our loss function is similar to the one used in regular CLIP training:

$$\begin{aligned}\mathcal{L} = &-\frac{a}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top G(z_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top G(z_i)/\tau)} \\ &-\frac{(1-a)}{N}\sum_{i=1}^{N}\log\frac{\exp(T(y_i)^\top G(z_i)/\tau)}{\sum_{j=1}^{N}\exp(T(y_j)^\top G(z_i)/\tau)},\ a\in[0,1]\end{aligned} \tag{4.3.1}$$

where (x_i, y_i, z_i) is one of our samples and I, T, G are our image, text, and generated dialogue towers respectively. In the above, a is a blending hyperparameter between the two losses. Intuitively, we want a to be high enough to encourage proper behavior of the generated caption representations with respect to the corresponding image representation. At the same time, we don't want too high value of a, since this will just lead to simply replacing the pre-trained text encoder with the third tower. Careful assignment of parameter a can lead to the third tower taking into account both original modalities.

4.4 Enhancing CLIP with Generative Dialogue: Domain Adaptation Approach

CLIP-3Modal [86] provides very promising results where it improved the zero-shot retrieval scores of its base CLIP model (check Results). Despite its success to enhance CLIP with more knowledge from the new generated textual cues, CLIP-3Modal showed very poor performance on Visual Question Answering tasks. We believe that CC3M captions and the generated dialogue have a significant distribution shift and the third tower architecture along with the 3-way contrastive loss is unable to adapt to it. In this context, we give up the third tower approach and we keep CLIP architecture as it is. Our main task is to adapt CLIP's text tower to dialogue (without harming the generalization). We propose **DRAFT (Dual Representation Adaptive Fine-Tuning)** as a new adaptation technique for CLIP-like architectures on dialogue textual inputs.

Everyone should agree that *Text* and *Dialogue* are both part of Natural Language Processing, but they are different. Specifically, dialogue is more dynamic than text. The time sequencing in dialogue is critical and very important, but not in text which is static and not so sequential. Furthermore, dialogue has interaction. This interaction imposes targets and sequencing, however none of these exists in text. Finally, dialogue may be limited in content but is richer in context than a simple caption. These properties can be visible on the examples presented in Figure 4.4.2

We consider this problem a domain adaptation problem where the *source* domain is the *Captions* and the *target* domain is the *Dialogue*. Most of the works that focus on domain adaptation on CLIP are focusing on distribution shifts on images, which makes our problem significantly unique. For example, CLIPood [75] tries to generalize to out-of-distribution images by exploiting semantic relations between classes through

text. Furthermore, without the use of the third tower we focus our work on fine-tuning the existing text encoder on both captions and dialogue inputs. With this, we try to adapt the text encoder on the dialogue data without harming the in-distribution performance of the pre-trained model. Again, works like FLYP (Fine-tune Like You Pre-train)[30] proposes that by mimicking the contrastive training during fine-tuning (on visual distribution shifts) can preserve the generalization and achieve better results on out-of-distribution datasets, but it also aims at image distribution shifts.

4.4.1 Using LLaVA for generating Questions

Following our work on CLIP-3Modal, we need to include a third modality in our input data. However, instead of using BLIP-2, we utilize LLaVA v1.5 [55, 54] to expand the the existing CC3M [74] dataset by generating multiple questions along with their answers for each input. Specifically, for each image, we provide it as input to the LLaVA model and then ask the model to generate 3 fundamental questions about this image as it shows in Figure 4.4.1. These dialogue samples seem to provide better descriptions about the given image and add new information to the existing caption. We have a 1:3 caption per dialogue ratio, so with our augmentation the training dataset was tripled in size. Some examples are presented in Figure 4.4.2.

Again, we note that the captioning model does not take the original caption as input when it is generating the above questions

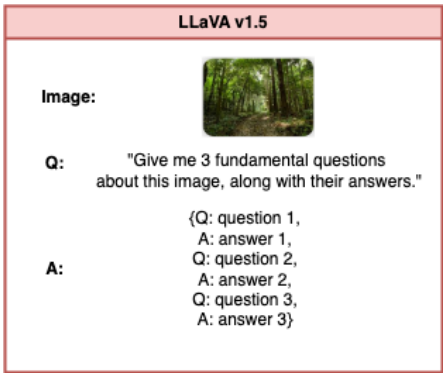


Figure 4.4.1: **Use of LLaVA-1.5 model**. The question-answer pairs the model generates provide textual cues for use as our third modality on the DA approach

4.4.2 Training for Domain Adaptation with DRAFT

Contrastive Objective

Now, we want to adapt a pre-trained CLIP model on the Question-Answer style textual descriptions of images. As in the Third Tower Approach, we start from a pretrained CLIP model provided by OpenCLIP. Then, we freeze the weights of the image tower while we keeping the text tower completely trainable. We train only the text tower on our LLaVA-extended dataset by giving both captions and QA-pairs as input on the text tower and the image as input on the frozen image tower. First, we constructing a contrastive objective, similar to CLIP, with loss function:

$$\begin{aligned}\mathcal{L}_{contrastive} = &-\frac{a}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top T(z_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top T(z_i)/\tau)} \\ &-\frac{(1-a)}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top T(y_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top T(y_i)/\tau)},\ a\in[0,1]\end{aligned} \tag{4.4.1}$$

where (x_i, y_i, z_i) is one of our samples and I, T are our image and text towers, respectively. The first part is the contrastive loss between the image and the dialogue sample, while the second part is the contrastive loss

Caption: all the details on hosting a party

QA #1: What is the main focus of the image?
Answer: The main focus of the image is a table filled with various snacks, including bags of popcorn and cookies.

QA #2: What is the table setting for?
Answer: The table setting appears to be for a Halloween party, as indicated by the presence of Halloween decorations and the overall festive atmosphere.

QA #3: What is the table covered with?
Answer: The table is covered with a tablecloth.

Caption: # catches a touchdown reception against army during the first half of a college football game .

QA #1: What is the player wearing?
Answer: The player is wearing a football uniform, which typically includes a jersey, pants, cleats, and a helmet.

QA #2: What is the main action in the image?
Answer: The main action in the image is a football player catching a football during a game.

QA #3: Who is the player in the image?
Answer: The player in the image is a football player, specifically a wide receiver.

Figure 4.4.2: **Examples of the augmented dataset.** On top there are images and below are the captions from the CC3M dataset and the corresponding question-answer pairs generated by LLaVA.

between image and text (CLIP loss). Furthermore, we use a blending parameter a to balance the two losses. Naturally, we want a to be high enough in order to encourage the model to yield meaningful representations for the dialogue in correspondence with the respective image representations. With the second loss, we try to do not let the text encoder forget its pre-training and harm its generalization abilities. With careful assignment of the hyperparameter a we can adapt the text tower to dialogue inputs and preserve generalizaton of the CLIP model.

Maximum Mean Discrepancy (MMD)

It is proven that contrastive loss assures that the distributions of the 2 modalities that take part in the loss (image-caption and image-dialogue) will be aligned (see Alignment & Uniformity). However, this property does not guarantee that the distributions of the captions and the dialogue are aligned too. We avoid to use a contrastive loss between them. Our intuition suggests that an extra contrastive loss will harm the representation space because the caption does not necessarily share the same semantic relation. In this context, our aim is to align the image-dialogue similarity distribution with the image-text similarity distribution by minimizing a distance over these distributions. Specifically, we try to minimize the **Maximum Mean Discrepancy (MMD)** which computes the deviation of their specific means and aligns the marginal distributions of the two input domains. MMD distance is formulated as:

$$\mathrm{MMD}^2(\mathcal{X}_s, \mathcal{X}_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(x_i^s) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(x_j^t) \right\|_{\mathcal{H}}^2 \tag{4.4.2}$$

where $\mathcal{X}_s$ and $\mathcal{X}_t$ are the source and target domain samples, respectively. n_s and n_t are the number of samples in the source and target domains and ϕ is the feature map that embeds the samples into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$.

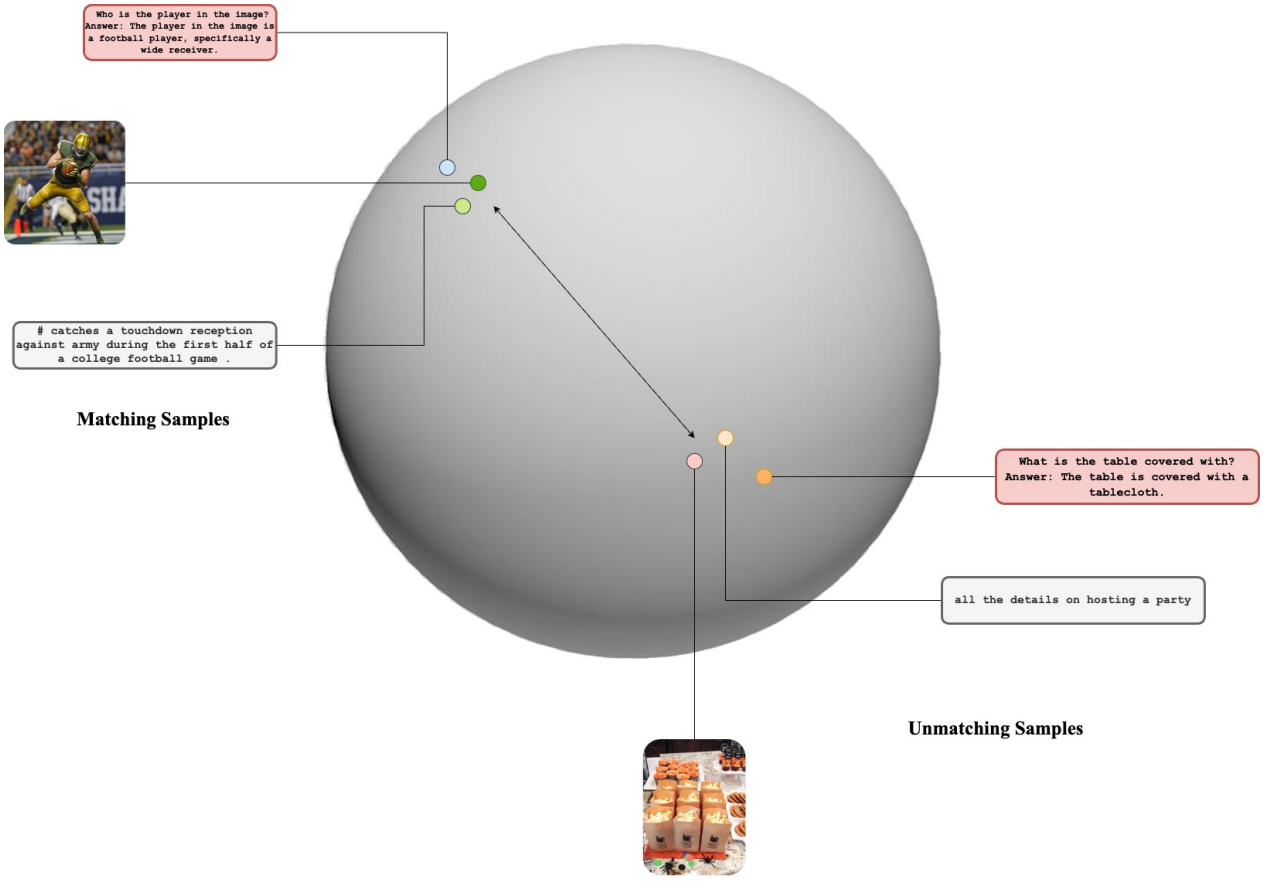


Figure 4.4.3: Visual representation of the **contrastive objective**. The positive pairs (green,blue) are concentrated together while the negative samples (orange) are pushed away. In our problem the positions of the image representations are fixed. Then, we try to align the caption and the dialogue representations with them separately.

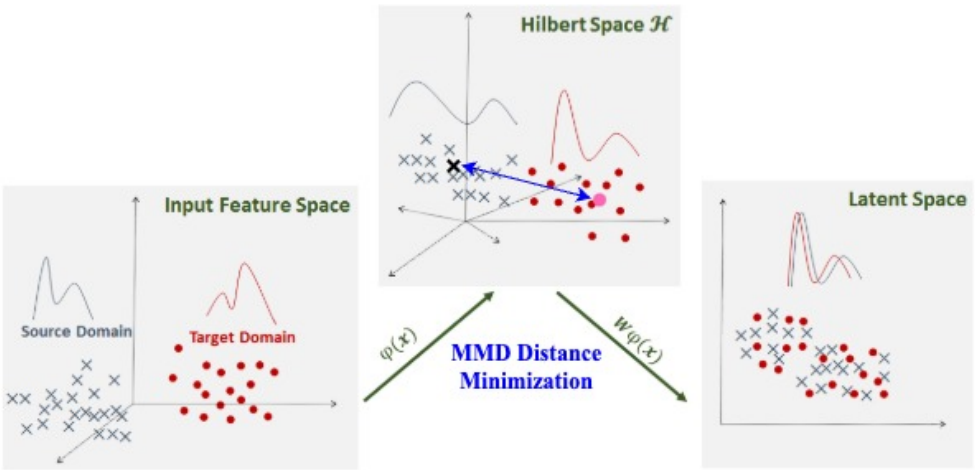


Figure 4.4.4: **Maximum Mean Discrepancy Loss**. Given two distributions, MMD tries to align the average embedding of each distribution. In this way, we can encourage our model to align the marginal image-caption similarity distribution with the marginal image-dialogue similarity distribution.

Final Training Loss

In our case we aim to combine the contrastive objective and the minimization of the Maximum Mean Discrepancy distance as we described above. Our final loss is formulated as:

$$\mathcal{L}_{contrastive} = -\frac{a}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top T(z_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top T(z_i)/\tau)} - \frac{(1-a)}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^\top T(y_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^\top T(y_i)/\tau)},\ a\in[0,1]$$

$$\mathcal{L}_{MMD} = \left\| \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi(I(x_j)^\top T(y_i)/\tau) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi(I(x_j)^\top T(z_i)/\tau) \right\|_{\mathcal{H}} \tag{4.4.3}$$

$$\mathcal{L}_{DRAFT} = \mathcal{L}_{contrastive} + \mathcal{L}_{MMD} \tag{4.4.4}$$

where (x_i, y_i, z_i) is one of our samples and I, T are our image, and text towers respectively. a is the blending hyperparameter between the two contrastive losses. Finally, ϕ is a feature mapping that, in our case, is a Radial Basis Function Kernel (RBF) [90]. An overall presentation of our method is presented in Figure 4.4.5.

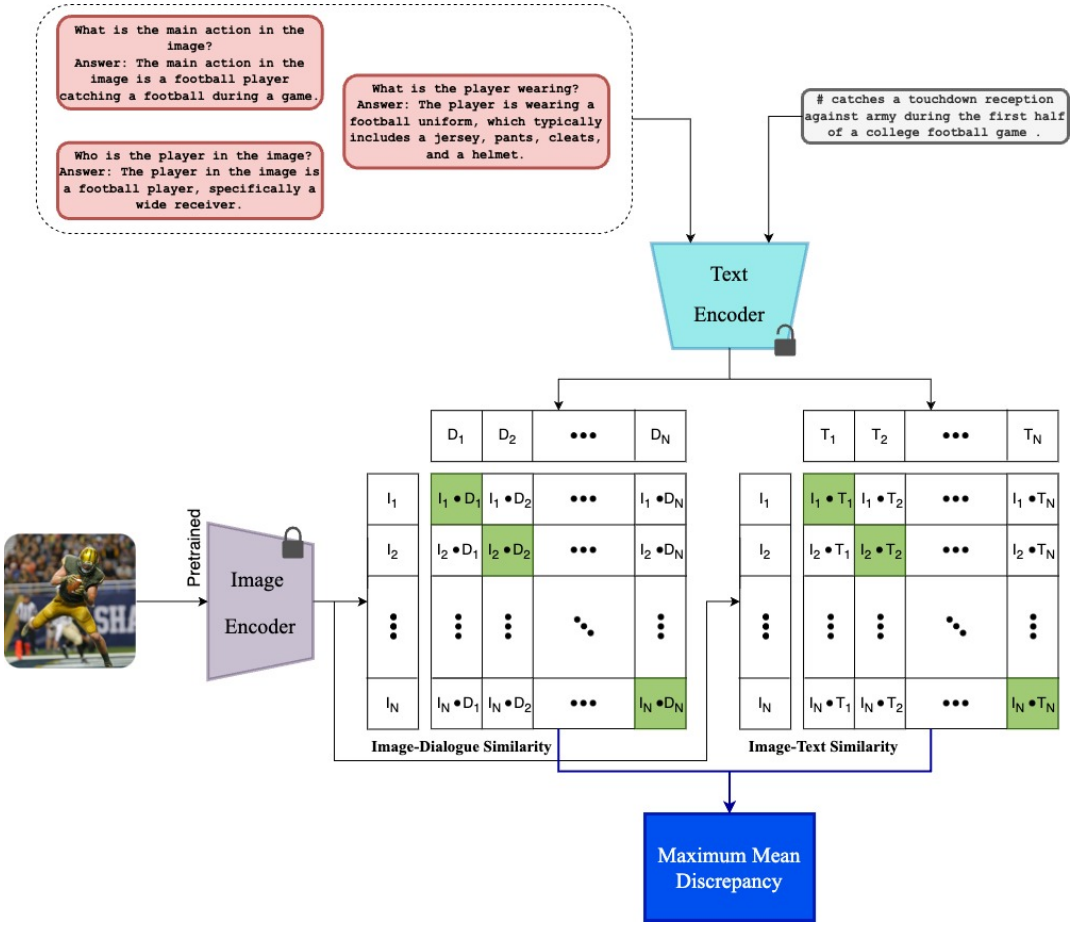


Figure 4.4.5: **DRAFT: Adapting CLIP on Question-Answer style text inputs.** We propose the use of contrastive loss between image representations and textual variations separately along with a distribution distance loss (MMD) over the semantic similarity distributions of images-captions and images-dialogue.

Chapter 5

Experimental Results

5.1 Datasets

In this section, we will describe the datasets that we will evaluate our methods. We will use MSCOCO [53] for zero-shot retrieval experiments and VQAv1 [4] dataset for the evaluation on Visual Question Answering tasks.

5.1.1 MSCOCO

The Microsoft Common Objects in Context (MSCOCO) [53] dataset is a large-scale, richly annotated image dataset designed for object detection, segmentation, and captioning tasks. The dataset contains over 200,000 images, around 1.5 million annotated object instances, and includes 80 object categories such as people, animals, vehicles, and everyday objects. Precise instance segmentations are provided for each object, along with keypoint annotations for human pose estimation. Each image is also paired with five captions, offering diverse descriptions.

MSCOCO supports several tasks including object detection, segmentation, keypoint detection, image captioning and panoptic segmentation. The dataset is diverse, with images taken from a wide range of everyday scenes around the world, presenting objects in natural contexts rather than isolated. This provides a realistic dataset for training models. The annotations are high-quality and manually annotated, ensuring accurate training and evaluation. Figure 5.1.1 shows some examples of the MSCOCO dataset.

Figure 5.1.1: **MSCOCO dataset examples.**[53]

The dataset is split into a training set with over 118,000 images, a validation set with about 5,000 images, and multiple test sets with tens of thousands of images used for competitions and evaluation purposes. It is widely used in computer vision research for developing and benchmarking new algorithms and training deep learning models for tasks such as object detection (e.g., Faster R-CNN [70], YOLO [69]), segmentation (e.g., Mask R-CNN [35]) and captioning.

5.1.2 VQAv1

The Visual Question Answering v1 (VQAv1) dataset [4] is a comprehensive dataset designed for research in the field of visual question answering. This dataset combines visual content with natural language processing, making it a challenging and valuable resource for developing AI models that can understand and reason about images.

VQAv1 dataset includes two distinct types of images: real images sourced from the MSCOCO dataset and abstract scenes. Both types serve unique purposes in training and evaluating VQA models. The real images are high-quality, natural images capturing everyday scenes with common objects. They are rich in detail and context, featuring varied lighting, angles, and backgrounds. Questions about real images require models to understand complex visual information and contextual relationships. For example "What is the person holding?" or "Is it raining in the image?". On the other hand, the abstract scenes in the VQAv1 dataset are synthetic, cartoon-like scenes created using clip art. These abstract scenes are designed to test specific aspects of visual understanding and reasoning without the complexity of real-world images. Abstract scenes are less complex visually, with clear and distinguishable objects and actions. This simplicity allows for easier control and manipulation of specific elements within the scene to create targeted questions. Abstract scenes

focus on isolating particular visual or reasoning tasks, such as understanding relationships between objects or basic object recognition. Examples of questions for abstract scenes may be "What is the child doing?" or "What the man is looking at?".

The answers in the VQAv1 dataset vary in format and type, covering a broad spectrum of potential queries that can be asked about an image. These answers fall into several categories: *Yes/No Answers* are simple binary responses to questions about the presence, state, or action depicted in the image. *Number Answers* are numerical responses to questions that involve counting objects or specifying quantities (e.g. "How many of the deer are sleeping?" and the answer could be "One"). *Short Answers* are short descriptive responses that provide information about various attributes, actions, or objects in the image (e.g. "Where is the kid pointing?" ("Mom"), "What sport are they playing?" ("Baseball")). *Open-ended Answers* are more complex responses that may involve reasoning or interpretation of the scene depicted in the image. For example we can have a question about "Why is the person running?" and the answer to be "To catch the bus". These examples can also be seen in Figure 5.1.2.

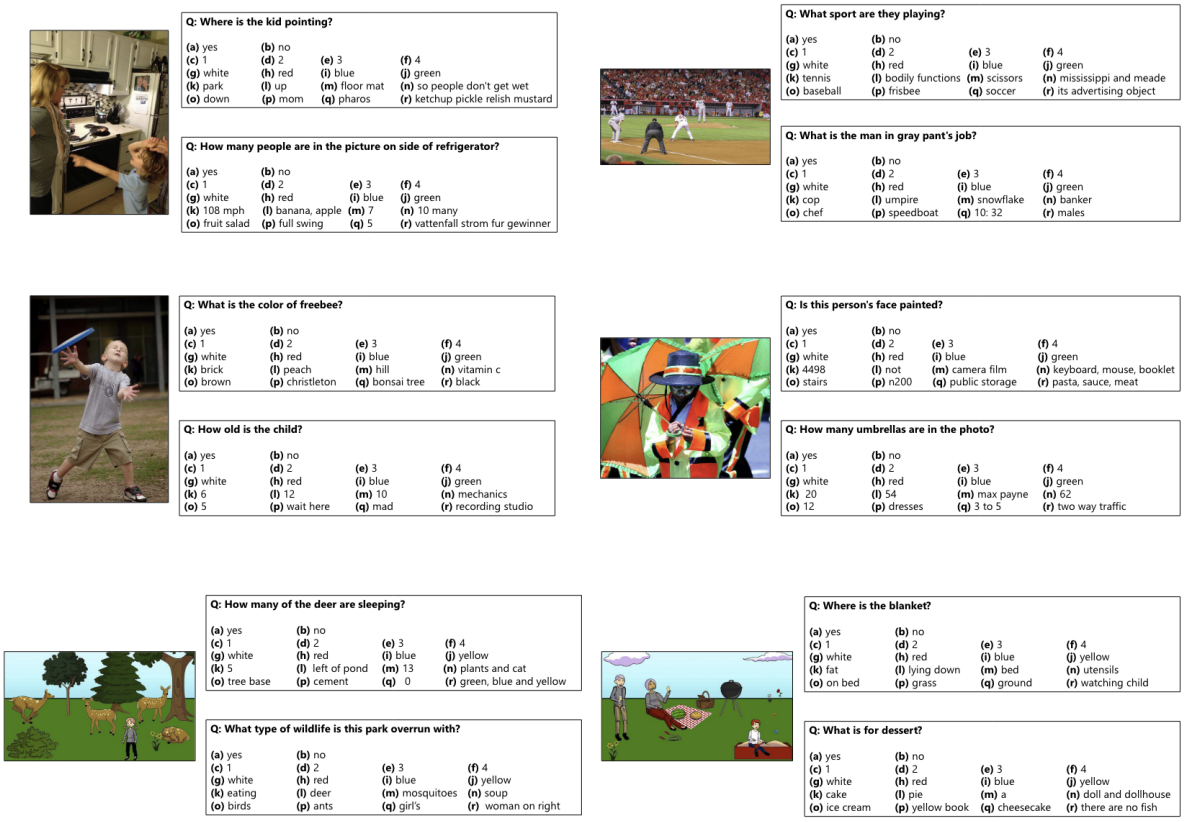


Figure 5.1.2: **Examples of the VQAv1 dataset.** Multiple-choice questions along with their answers for real and abstract scenes.[4]

5.2 Third Tower Evaluation

5.2.1 Evaluation Method

For the evaluation of CLIP-3Modal we focused on zero-shot retrieval tasks between image and text due to CLIP. We used multiple architecture of CLIP provided by OpenCLIP as the baseline for the evaluation. We run our experiments on architectures with ViT image encoders and BERT or T5 as text encoder. These are the same models used as the foundation of CLIP-3Modal. The aforementioned pre-trained models were

trained on 32 billion samples of the LAION-2B dataset [73]. This provides us with a good initial point for our third encoder. For the weighting hyperparameter a in our loss function, we use $a = 0.65$. We observed that overall our model performs the best when the weight of the loss between image and generated captions is higher than 0.5.

After training our model, we fuse the output embeddings of the text and the generated captions encoders, to obtain a final embedding for the text. For the fusion of the outputs we take a weighted sum of the embeddings provided by the text and the generated captions towers:

$$X_{ensemble} = \beta \cdot T + (1 - \beta) \cdot G, \ \beta \in [0, 1] \tag{5.2.1}$$

where T and G are the output embeddings of the original text and the generated captions tower. By weighting the two outputs appropriately, the third tower will incorporate additional information that enhances the predictions learned by the original model. In this case, we use a blending parameter of $\beta = 0.9$. The insight for this high value of β is that we want the embeddings from the third tower to influence the output, but we still want the changes to be small so as to not lose their initial valuable information. An ablation study on the blending parameter is presented in Table 5.2.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32	39.7	65.4	75.6	56.3	79.8	87.1
C3M-ViT/B-32	**40.2**	**65.9**	**75.9**	**57.0**	**80.6**	**87.5**
CLIP-ViT/L-14	46.5	71.1	79.8	63.3	84.0	90.8
C3M-ViT/L-14	**46.8**	**71.2**	**80.0**	**63.6**	**84.1**	**91.0**
CLIP-ViT/H-14	**49.4**	**73.4**	**81.5**	**66.0**	86.1	**91.9**
C3M-ViT/H-14	49.1	73.3	**81.5**	**66.0**	**86.2**	**91.9**

Table 5.1: **Base results using CLIP-3Modal along with BLIP-2 for captioning.** CLIP-3Modal improves recall on almost every zero-shot retrieval task. Evaluation is done on MSCOCO and both image and text encoders were pre-trained on the LAION-2B Dataset.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline ($\beta = 1$)	39.7	65.4	75.6	56.3	79.8	87.1
$\beta = 0.95$	39.9	65.7	75.8	**57.1**	80.4	87.3
$\beta = 0.90$	**40.2**	**65.9**	**75.9**	57.0	**80.6**	**87.5**
$\beta = 0.80$	40.0	65.6	75.8	56.3	79.6	86.2
$\beta = 0.60$	38.7	63.8	73.8	53.7	75.6	84.8

Table 5.2: **High values of the blending parameter improve performance, while smaller drop the evaluation scores lower than the baseline**. In this figure β denotes the blending weight hyperparameter. Smaller weight on the generated captions encoder benefits the fused embeddings by preserving their initial information and enhancing them with different aspects of the input. The best results were occurred for $\beta = 0.9$.

5.2.2 Implementation

We trained our third tower as described above, using different ViT based CLIP models pre-trained on LAION-2B (provided by OpenCLIP) as our foundation. For our training, on ViT-B-32 based architecture, we used batch size 1024, learning rate 10^{-5}, weight decay 0.1 and trained our model for 2 epochs on our custom dataset.For the small CLIP models (e.g. ViT-B-32) we train all the third tower while for the bigger we only train the last few layers (2 or 3 last layers). The training of the third tower on CLIP-3Modal-ViT/B-32 takes approximately 1 hour per epoch on a single GPU for, which is significantly less than the training time of the foundation model. We based our evaluation on MSCOCO [53] by studying the zero-shot retrieval

performance on this specific dataset, using the same evaluation metric provided by OpenCLIP. We managed to outperform the OpenCLIP's model in both image and text zero-shot retrieval, with a margin of 0.3% to 0.8%. More details are presented in Table 5.1. For the bigger ViT based architectures we used different batch size due to their larger size. More specifically, for both ViT-L-14 and ViT-H-14 we had batch size 768 and learning rate 10^{-3}.

5.2.3 Ablation Studies

No Blending

Above we proposed a blending scheme for the evaluation of CLIP-3Modal on zero-shot retrieval. Despite its promising results we want to see if the proposed third tower can stand alone without blending. As we can see in Table 5.3 the third tower is not performing very well when it is evaluated alone. That's makes sense because the fine-tuning harms the generalization that the tower had from its pre-training. For this reason, we use the third tower in addition with the original text tower.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32*	34.2	60.0	70.6	52.4	76.3	84.3
C3M-ViT/B-32^{LLaVA}	**34.9**	**60.7**	**71.2**	**52.9**	**77.1**	**84.7**
C3M-ViT/B-32^{LLaVA} (only 3rd)	27.5	51.3	62.4	39.8	66.9	76.5
CLIP-ViT/L-14	46.5	71.1	79.8	63.3	84.0	90.8
C3M-ViT/L-14	**46.8**	**71.2**	**80.0**	**63.6**	**84.2**	**91.0**
C3M-ViT/L-14 (only 3rd)	35.8	61.4	71.8	44.3	69.8	80.1
CLIP-ViT/H-14	**49.4**	**73.4**	**81.5**	**66.0**	86.1	**91.9**
C3M-ViT/H-14	49.1	73.3	**81.5**	**66.0**	**86.2**	**91.9**
C3M-ViT/H-14 (only 3rd)	19.9	41.3	52.3	31.0	57.0	68.9

Table 5.3: **The third benefits only when it operates additively.** The fine-tuning harms the information learned from pre-training on the third-tower. However, it provides new information to the original text encoder which can be seen after their fusion. * Pre-trained on LAION-400m.

Generative Model

In Section 4.4 we proposed a new generation pattern for the synthetic dialogue samples. More specifically, instead of BLIP-2 we used LLaVA v1.5 for dialogue generation. In addition to that we changed the interaction with the model. Instead of asking two sequantial questions about the image and its uniqueness (BLIP-2) we ask the model (LLaVA v1.5) to find three questions that describe the image and answer them.

In order to evaluate the new inputs, we concatenate all three question answer pairs. Then we provide it as input to the CLIP-3Modal model. As we cam see in Table 5.4 synthetic data generated with LLaVA v1.5 and the new prompt improve the performance of CLIP-3Modal in every case. In Table 5.5

		Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	Generative Model	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32*	-	34.2	60.0	70.6	52.4	76.3	84.3
C3M-ViT/B-32	LLaVA	**34.9**	**60.7**	**71.2**	**52.9**	**77.1**	**84.7**
CLIP-ViT/H-14	-	49.4	73.4	81.5	66.0	86.1	91.9
C3M-ViT/H-14	BLIP-2	49.1	73.3	81.5	66.0	86.2	91.9
C3M-ViT/H-14	LLaVA	**49.6**	**73.5**	**81.7**	**66.6**	**86.5**	**92.1**

Table 5.4: **Replacing BLIP-2 with LLaVA improves the performance for both blending and not blending cases.** * We used the pre-trained model on LAION-400m instead of LAION-2B. We can see that LLaVA and the new generation pattern provides better results for retrieval. Even the ViT-H-14 based model out-performed the baseline model, which was not achieved with the BLIP-2 genearated data.

		Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	Generative Model	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/H-14	-	49.4	73.4	81.5	66.0	86.1	91.9
C3M-ViT/H-14	BLIP-2	19.9	41.3	52.3	31.0	57.0	68.9
C3M-ViT/H-14	LLaVA	30.4	55.5	66.5	46.0	71.7	80.3

Table 5.5: **Generated data with LLaVA also improve the performance of the third tower alone.** It is visible that LLaVA generated data and the QA-pair pattern help the pre-trained text encoder to maintain more information from its pre-training.

5.2.4 Results

After careful consideration of every experiment we performed we present the best performance of every CLIP-3Modal in Table 5.6. We can see that our method enhanced with the LLaVA generated QA-pairs gives the better results and outperform the baseline CLIP model and its variant that utilizes the BLIP-2 model.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32	39.7	65.4	75.6	56.3	79.8	87.1
C3M-ViT/B-32	**40.2**	**65.9**	**75.9**	**57.0**	**80.6**	**87.5**
CLIP-ViT/L-14	46.5	71.1	79.8	63.3	84.0	90.8
C3M-ViT/L-14	**46.8**	**71.2**	**80.0**	**63.6**	**84.1**	**91.0**
CLIP-ViT/H-14	49.4	73.4	81.5	66.0	86.1	91.9
C3M-ViT/H-14^{LLaVA}	**49.6**	**73.5**	**81.7**	**66.6**	**86.5**	**92.1**

Table 5.6: **CLIP-3Modal improves recall on every zero-shot retrieval task.** Here are the models with the best performance. We see that our model outperforms the baseline OpenCLIP on every case.

5.3 DRAFT Evaluation

5.3.1 Evaluation Method

In Section 4.4, we proposed an alternative approach to CLIP-3Modal that adapts CLIP for dialogue textual inputs. To assess the effectiveness of our method, we evaluated it on Visual Question Answering (VQA) tasks using the VQAv1 dataset. Evaluating CLIP-like models on such tasks involves converting the VQA problem into a zero-shot retrieval problem. Specifically, given an image, a question, and all possible N answers, we concatenate each answer with the question to create N question-answer pairs. We then predict the correct pair by identifying which pair's embeddings have the highest cosine similarity with the image. This process is illustrated in Figure 5.3.1.

While some suggest fine-tuning CLIP on the training set before evaluation, we prefer to evaluate our models without any exposure to VQA dataset samples. This approach better evaluates the model's adaptation using generative dialogue data from LLaVA. Additionally, Figure 5.3.1 shows that the QA pairs in the test dataset are distinct from the extensive question-answer pairs generated during training.

After adapting our model for Question-Answer style inputs, we follow the above process using the adapted text encoder to embed the concatenated QA pairs. Moreover, we employ the blending technique from Third Tower Evaluation, which blends the adapted text encoder with the original encoder from the baseline CLIP. We believe that the domain adaptation approach can enhance VQA performance when it is combined with the base CLIP. We evaluated the model on both real images and abstract scenes from VQAv1.

5.3.2 Results

We trained our model using a ViT-B-32 based CLIP architecture provided by OpenCLIP. The baseline model was pre-trained on the LAION-400M dataset. Our experiments were conducted on either 2 or 4 A100 GPUs,

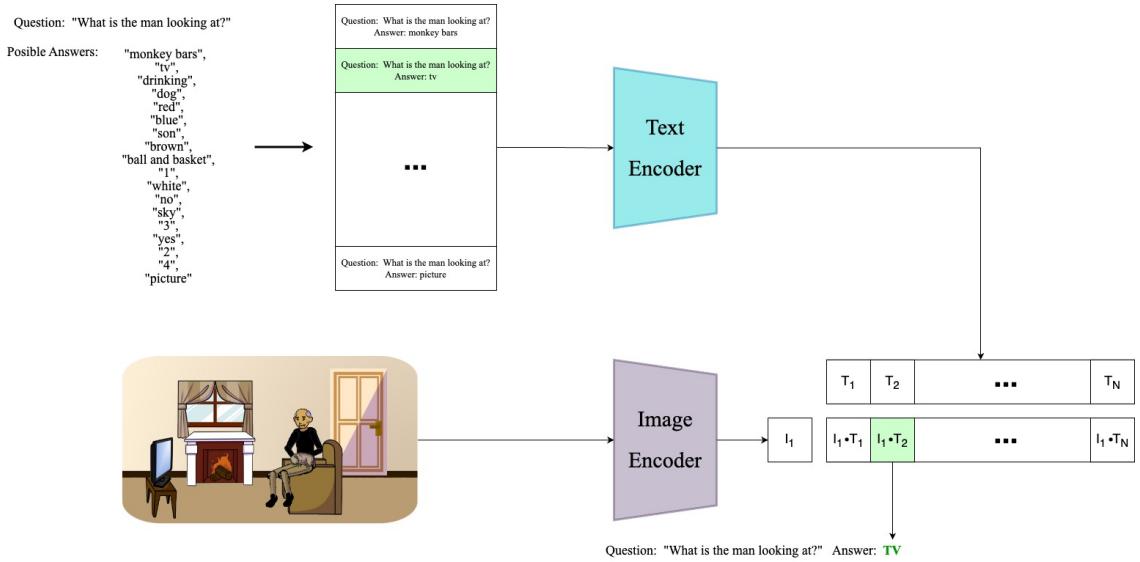


Figure 5.3.1: **VQA evaluation method for CLIP-like models**. We concatenate each possible answer with the question and use the image to find the QA pair with the highest cosine similarity.

with training times of approximately 16 minutes and 9 minutes per epoch, respectively. The performance of our method on abstract VQA scenes is shown in Table 5.7 , while the results for VQA on real images are presented in Table 5.8.

As demonstrated, the adapted text encoder significantly enhances the model's ability to answer questions based on specific image inputs. These results represent a step forward in the unsupervised adaptation of CLIP models to new textual cues. Our model outperforms the original CLIP on VQA tasks by a substantial margin without having seen samples from the VQA dataset. This improvement highlights the effectiveness of our adaptation method and the generated data in boosting CLIP's accuracy in answering questions.

Visual Question Answering (VQAv1) *Abstract Scenes*		
Model	Epochs	Accuracy
CLIP-ViT/B-32	-	9.88%
Ours-ViT/B-32	6	10.68%
Ours-ViT/B-32+blend	12	**11.54%**

Table 5.7: **Our method surpasses the baseline CLIP on VQAv1 abstract scenes, with and without blending.** The baseline CLIP model, provided by OpenCLIP and pre-trained on the LAION-400M dataset Our model, trained with a batch size of 1536, increased the accuracy to 10.68% . By blending the adapted and original text encoders (with $\beta = 0.9$), we further improved accuracy of the baseline model by almost 1.7%. This demonstrates a significant performance boost in the VQA task using our adaptation method.

5.3.3 Ablation Study

MMD Loss

To evaluate the impact of the Maximum Mean Discrepancy (MMD) Loss, we conducted ablation experiments by training our model solely on the contrastive objective. Table 5.9 highlights the crucial role of the MMD loss in enhancing CLIP's VQA performance. Our findings underscore the significance of incorporating MMD loss in the adaptation process, as evidenced by the results.

Visual Question Answering (VQAv1) *Real Images*		
Model	Epochs	Accuracy
CLIP-ViT/B-32	-	26.2%
Ours-ViT/B-32	2	23.6%
Ours-ViT/B-32+blend	3	**28.3%**

Table 5.8: **Our method enhances accuracy on VQAv1 real images through blended text encoders.** The baseline CLIP model provided by OpenCLIP and pre-trained on the LAION-400M dataset. Our model, trained with a batch size of 1536 for only 2 epochs, initially reached an accuracy of 23.6% . However, by blending the adapted text encoder with the original text encoder (using $\beta = 0.9$), we significantly improved the accuracy to 28.3%, which out-performed CLIP by 2.1% .

Visual Question Answering (VQAv1) *Abstract Scenes*			
Model	Epochs	Batch Size	Accuracy
CLIP-ViT/B-32	-	-	9.88%
Ours-ViT/B-32	6	1536	10.68%
Ours-ViT/B-32+blend	12	1536	**11.54%**
Ours-ViT/B-32$^{no\ MMD}$	6	6144	9.46%
Ours-ViT/B-32+blend$^{no\ MMD}$	12	6144	10.24%

Table 5.9: **Impact of MMD Loss on the performance of the adapted text encoder.** The inclusion of MMD loss in training enables our model to surpass the baseline CLIP-ViT/B-32 in VQAv1 abstract scenes. Models trained without MMD loss, despite using a larger batch size for enhanced contrastive learning, showed lower accuracies compared to their counterparts trained with MMD loss. While blending mitigated some performance reduction, it didn't fully compensate for the absence of MMD loss, highlighting its crucial role in improving VQA performance.

5.3.4 Zero-shot Retrieval

To demonstrate that our method does not harm the generalization of the CLIP model, we evaluated it on zero-shot retrieval tasks and compared the results with CLIP and our improved method, CLIP-3Modal. The comparison is shown in Table 5.10. Our model improved the recall scores of baseline CLIP across all zero-shot retrieval tasks and slightly outperformed CLIP-3Modal on some text retrieval tasks. This suggests that the generated data enhance CLIP by providing more information about images, as evidenced by the improved retrieval performance.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT/B-32	34.2	60.0	70.6	52.4	76.3	84.3
C3M-ViT/B-32^{LLaVA}	**34.9**	**60.7**	**71.2**	52.9	**77.1**	84.7
Ours-ViT/B-32+blending	34.8	60.4	71.0	**53.1**	76.6	**84.8**
C3M-ViT/B-32^{LLaVA} (only 3rd)	27.5	51.3	62.4	39.8	66.9	76.5
Ours-ViT/B-32	25.4	49.0	60.6	36.7	64.3	75.5

Table 5.10: **Performance comparison in zero-shot retrieval tasks.** Our approach surpasses the baseline CLIP-ViT/B-32 across all retrieval metrics. By keeping the CLIP loss during training, our method preserves its generalization capability while enhancing performance through embedding blending.

Chapter 6

Conclusion & Future Work

6.1 Conclusion

We can see here that the use of an extra modality is a viable way to improve upon an image and text model. Generative dialogue, rich in supplementary information and context, inspired us to creatively integrate it into image-text architectures. We initially incorporated an additional tower to the model's architecture to include new modalities during the training process, thereby extending existing image-text models. Moreover, we introduced **DRAFT**, a novel method for adapting CLIP-like models to various textual styles (e.g., dialogue) using contrastive learning and distribution alignment. By mimicking CLIP's training approach and aligning the similarity distributions of the different textual inputs with the corresponding images, we demonstrated improved performance on visual question answering tasks while maintaining the model's generalization ability. Additionally, we found that the Maximum Mean Discrepancy (MMD) loss is highly effective in adapting the similarity distributions, making it a crucial component of our adaptation method. Finally, we showed that blending the parameters of the base model with those of the fine-tuned text encoder further enhances performance.

6.2 Future Works

In the future, we aim to further analyze the integration of a third tower in CLIP-style models and explore alternative choices for the third modality. For both approaches, we plan to experiment with different generation models and various styles for the generated textual cues. Additionally, we will conduct further experiments on the DRAFT method, including initial fine-tuning of the model before evaluation in visual question answering (VQA), following the suggested evaluation pipeline. An exciting follow-up project involves applying the DRAFT method to encoder-decoder models instead of just encoder models like CLIP. Finally, the concept of parameter blending presents a very interesting area for research due to its significant benefits for our training methods. To enhance blending, we are considering experimenting with more dynamic approaches, such as moving average, instead of the simple summation proposed in this thesis.

Appendix A

Bibliography

[1] Alayrac, J.-B. et al. "Flamingo: a visual language model for few-shot learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 23716–23736.

[2] Alwassel, H. et al. "Self-supervised learning by cross-modal audio-video clustering". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9758–9770.

[3] An, C. et al. "Cont: Contrastive neural text generation". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 2197–2210.

[4] Antol, S. et al. "Vqa: Visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

[5] Arık, S. Ö. et al. "Deep voice: Real-time neural text-to-speech". In: *International conference on machine learning*. PMLR. 2017, pp. 195–204.

[6] Arora, S. et al. "A theoretical analysis of contrastive unsupervised representation learning". In: *arXiv preprint arXiv:1902.09229* (2019).

[7] Baevski, A. et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in neural information processing systems* 33 (2020), pp. 12449–12460.

[8] Bengio, Y., Courville, A., and Vincent, P. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[9] Borodachov, S. V., Hardin, D. P., and Saff, E. B. *Discrete energy on rectifiable sets*. Springer, 2019.

[10] Brodeur, S. et al. "Home: A household multimodal environment". In: *arXiv preprint arXiv:1711.11017* (2017).

[11] Brown, T. et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[12] Caron, M. et al. "Deep clustering for unsupervised learning of visual features". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149.

[13] Chen, T. et al. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[14] Chen, X. and He, K. "Exploring simple siamese representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15750–15758.

[15] Chiang, W.-L. et al. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. 2023. URL:

[16] Cho, K. et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[17] Cohn, H. and Kumar, A. "Universally optimal distribution of points on spheres". In: *Journal of the American Mathematical Society* 20.1 (2007), pp. 99–148.

[18] Collins, L. et al. "Maml and anil provably learn representations". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 4238–4310.

[19] Deng, J. et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[20] Devlin, J. et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[21] Dosovitskiy, A. et al. “Discriminative unsupervised feature learning with convolutional neural networks”. In: *Advances in neural information processing systems* 27 (2014).
[22] Dosovitskiy, A. et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
[23] Dyer, C. “Notes on noise contrastive estimation and negative sampling”. In: *arXiv preprint arXiv:1410.8251* (2014).
[24] Ericsson, L. et al. “Self-supervised representation learning: Introduction, advances, and challenges”. In: *IEEE Signal Processing Magazine* 39.3 (2022), pp. 42–62.
[25] Georgescu, M.-I. et al. “Anomaly detection in video via self-supervised and multi-task learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12742–12752.
[26] Gidaris, S., Singh, P., and Komodakis, N. “Unsupervised representation learning by predicting image rotations”. In: *arXiv preprint arXiv:1803.07728* (2018).
[27] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.
[28] Gouk, H., Hospedales, T. M., and Pontil, M. “Distance-based regularisation of deep networks for fine-tuning”. In: *arXiv preprint arXiv:2002.08253* (2020).
[29] Goyal, P. et al. “Scaling and benchmarking self-supervised visual representation learning”. In: *Proceedings of the ieee/cvf International Conference on computer vision*. 2019, pp. 6391–6400.
[30] Goyal, S. et al. “Finetune like you pretrain: Improved finetuning of zero-shot vision models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19338–19347.
[31] Grill, J.-B. et al. “Bootstrap your own latent-a new approach to self-supervised learning”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21271–21284.
[32] Gutmann, M. and Hyvärinen, A. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 297–304.
[33] Hastie, T. et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.
[34] He, K. et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
[35] He, K. et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.
[36] He, K. et al. “Momentum contrast for unsupervised visual representation learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.
[37] Hessel, J. and Lee, L. “Does my multimodal model learn cross-modal interactions? it’s harder to tell than you might think!” In: *arXiv preprint arXiv:2010.06572* (2020).
[38] Higgins, I. et al. “beta-vae: Learning basic visual concepts with a constrained variational framework”. In: *International conference on learning representations*. 2016.
[39] Hochreiter, S. and Schmidhuber, J. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.
[40] Hu, D., Nie, F., and Li, X. “Deep multimodal clustering for unsupervised audiovisual learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9248–9257.
[41] Ilharco, G. et al. *OpenCLIP*. July 2021. URL:
[42] Jing, L. and Tian, Y. “Self-supervised visual feature learning with deep neural networks: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 4037–4058.
[43] Johnson, J. et al. “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2901–2910.
[44] Kossen, J. et al. “Three Towers: Flexible Contrastive Learning with Pretrained Image Models”. In: *arXiv preprint arXiv:2305.16999* (2023).
[45] Krishna, R. et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations”. In: *International journal of computer vision* 123 (2017), pp. 32–73.
[46] Lee, J. D. et al. “Predicting what you already know helps: Provable self-supervised learning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 309–323.

[47] Lee, M. A. et al. "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 8943–8950.
[48] Lewis, M. et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: 1910.13461 [cs.CL].
[49] Li, J., Shang, J., and McAuley, J. "Uctopic: Unsupervised contrastive learning for phrase representations and topic mining". In: *arXiv preprint arXiv:2202.13469* (2022).
[50] Li, J. et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *International Conference on Machine Learning*. 2023.
[51] Li, Y., Yang, M., and Zhang, Z. "A Survey of Multi-View Representation Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 31.10 (2019), pp. 1863–1883. DOI: 10.1109/tkde.2018.2872063. URL:
[52] Liang, P. P., Zadeh, A., and Morency, L.-P. "Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions". In: *arXiv preprint arXiv:2209.03430* (2022).
[53] Lin, T.-Y. et al. "Microsoft coco: Common objects in context". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.
[54] Liu, H. et al. "Improved baselines with visual instruction tuning". In: *arXiv preprint arXiv:2310.03744* (2023).
[55] Liu, H. et al. "Visual instruction tuning". In: *arXiv preprint arXiv:2304.08485* (2023).
[56] Liu, Y. et al. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).
[57] Logeswaran, L. and Lee, H. "An efficient framework for learning sentence representations". In: *arXiv preprint arXiv:1803.02893* (2018).
[58] Mikolov, T. et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).
[59] Oord, A. v. d., Li, Y., and Vinyals, O. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).
[60] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774.
[61] Pathak, D. et al. "Context encoders: Feature learning by inpainting". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544.
[62] Radford, A. et al. "Improving language understanding with unsupervised learning". In: (2018).
[63] Radford, A. et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.
[64] Radford, A. et al. "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763.
[65] Raffel, C. et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2023. arXiv: 1910.10683 [cs.LG].
[66] Rajpurkar, P. et al. "Squad: 100,000+ questions for machine comprehension of text". In: *arXiv preprint arXiv:1606.05250* (2016).
[67] Ramesh, A. et al. "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.
[68] Ramesh, A. et al. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* (2022).
[69] Redmon, J. et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.
[70] Ren, S. et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).
[71] Rombach, R. et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: 2112.10752 [cs.CV].
[72] Sarkar, P. and Etemad, A. "Self-supervised learning for ecg-based emotion recognition". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 3217–3221.
[73] Schuhmann, C. et al. "Laion-5b: An open large-scale dataset for training next generation image-text models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 25278–25294.

[74] Sharma, P. et al. "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning". In: *Proceedings of ACL*. 2018.
[75] Shu, Y. et al. "Clipood: Generalizing clip to out-of-distributions". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 31716–31731.
[76] Simonyan, K. and Zisserman, A. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).
[77] Singer, U. et al. "Make-a-video: Text-to-video generation without text-video data". In: *arXiv preprint arXiv:2209.14792* (2022).
[78] Socher, R. et al. "Zero-shot learning through cross-modal transfer". In: *Advances in neural information processing systems* 26 (2013).
[79] Sohn, K. "Improved deep metric learning with multi-class n-pair loss objective". In: *Advances in neural information processing systems* 29 (2016).
[80] Su, Y. et al. "A contrastive framework for neural text generation". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 21548–21561.
[81] Sun, C. et al. "Videobert: A joint model for video and language representation learning". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 7464–7473.
[82] Thompson, B. "Canonical correlation analysis." In: (2000).
[83] Tian, Y., Krishnan, D., and Isola, P. "Contrastive multiview coding". In: *European conference on computer vision*. Springer. 2020, pp. 776–794.
[84] Tian, Y., Krishnan, D., and Isola, P. "Contrastive Multiview Coding". In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 776–794. ISBN: 978-3-030-58621-8.
[85] Tsai, Y.-H. H. et al. "Multimodal transformer for unaligned multimodal language sequences". In: *Proceedings of the conference. Association for computational linguistics. Meeting*. Vol. 2019. NIH Public Access. 2019, p. 6558.
[86] Tsaprazlis, E. et al. "Enhancing CLIP with a Third Modality". In: *NeurIPS 2023 Workshop: Self-Supervised Learning - Theory and Practice* (2023).
[87] Van Engelen, J. E. and Hoos, H. H. "A survey on semi-supervised learning". In: *Machine learning* 109.2 (2020), pp. 373–440.
[88] Vaswani, A. et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).
[89] Vendrov, I. et al. "Order-embeddings of images and language". In: *arXiv preprint arXiv:1511.06361* (2015).
[90] Vert, J.-P., Tsuda, K., and Schölkopf, B. "A primer on kernel methods". In: (2004).
[91] Wang, T. and Isola, P. "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere". In: *International conference on machine learning*. PMLR. 2020.
[92] Wu, Z. et al. "Unsupervised feature learning via non-parametric instance-level discrimination". In: *arXiv preprint arXiv:1805.01978* (2018).
[93] Xiao, Y. et al. "Multimodal end-to-end autonomous driving". In: *IEEE Transactions on Intelligent Transportation Systems* 23.1 (2020), pp. 537–547.
[94] Xu, C., Tao, D., and Xu, C. *A Survey on Multi-view Learning*. 2013. arXiv: 1304.5634 [cs.LG].
[95] Xu, D. et al. "Self-supervised spatiotemporal learning via video clip order prediction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10334–10343.
[96] Xue, L. et al. "mT5: A massively multilingual pre-trained text-to-text transformer". In: *arXiv preprint arXiv:2010.11934* (2020).
[97] Yu, J. et al. "CoCa: Contrastive Captioners are Image-Text Foundation Models". In: *Transactions on Machine Learning Research* (2022). ISSN: 2835-8856. URL:
[98] Zhang, H. et al. "Learning concept taxonomies from multi-modal data". In: *arXiv preprint arXiv:1606.09239* (2016).
[99] Zhang, R. "Making convolutional networks shift-invariant again". In: *International conference on machine learning*. PMLR. 2019, pp. 7324–7334.
[100] Zhang, Y. et al. "Contrastive learning of medical visual representations from paired images and text". In: *Machine Learning for Healthcare Conference*. PMLR. 2022, pp. 2–25.