National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Speech and Language Processing Group

Adapted Multimodal BERT with Layer-wise Fusion for Sentiment Analysis

Diploma Thesis

of

ODYSSEAS SPYRIDON CHLAPANIS

Supervisor: Alexandros Potamianos
Associate Professor

Athens, November 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Speech and Language Processing Group

Adapted Multimodal BERT with Layer-wise Fusion for Sentiment Analysis

Diploma Thesis

of

ODYSSEAS SPYRIDON CHLAPANIS

Supervisor: Alexandros Potamianos
Associate Professor

Approved by the examination committee on 2nd November 2022.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Alexandros Potamianos	Constantinos Tzafestas	Stefanos Kollias
Associate Professor	Associate Professor	Professor

Athens, November 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Speech and Language Processing Group

(Signature)

.....
Odysseas Spyridon Chlapanis

Electrical & Computer Engineer Graduate, NTUA

To my parents,
George and Athina
and in memory
of my grandmother,
Marianthi

Περίληψη

Τα τελευταία χρόνια η πληθώρα πολυμέσων και οι εξελίξεις στον τομέα της μηχανικής μάθησης έχει οδηγήσει στην εξάπλωση της πολυτροπικής μάθησης ως ένα από τα σημαντικότερα πεδία ερευνών εφαρμοσμένης τεχνητής νοημοσύνης. Η εκτεταμένη χρήση των μέσων κοινωνικής δικτύωσης έχει επιτρέψει την συλλογή τεράστιων συνόλων δεδομένων. Επιπρόσθετα, η πρόσφατη επιτυχία των Προεκπαιδευμένων Γλωσσικών Μοντέλων (ΠΓΜ) έχει οδηγήσει στην δημιουργία εκπληκτικών νέων εφαρμογών. Παρ' όλα αυτά η προ-εκπαίδευση νευρωνικών δικτύων μεγάλης κλίμακας σε πολλά στάδια που απαιτείται επιβάλλει ένα δυσθεώρητο κόστος παραμέτρων.

Στην παρούσα εργασία, προτείνεται το Προσαρμοσμένο Πολυτροπικό BERT (AMB), ένα μοντέλο βασισμένο στο γλωσσικό μοντέλο BERT το οποίο επεκτείνεται για πολυτροπική ανάλυση διάθεσης με ένα συνδυασμό από αντάπτορες (ή αλλιώς προσαρμογείς) και ενδιάμεσα επίπεδα συγχώνευσης. Το BERT είναι ένα προεκπαιδευμένο βαθύ νευρωνικό δίκτυο το οποίο είχε αρχικά σχεδιαστεί για την επεξεργασία γλωσσικής πληροφορίας και αποτελείται από πολλά επίπεδα του μοντέλου transformer. Ο αντάπτορας είναι ένα τμήμα της αρχιτεκτονικής το οποίο τοποθετείται ανάμεσα στα επίπεδα του BERT με σκοπό να προσαρμόσει το προεκπαιδευμένο γλωσσικό μοντέλο για το οποιοδήποτε πρόβλημα. Αυτή η διαδικασία ονομάζεται μεταφορά μάθησης, αλλά σε αντίθεση με την κλασική μέθοδο που ονομάζεται fine-tuning, οι αντάπτορες είναι πιο οικονομικοί ως προς τις παραμέτρους. Τα επίπεδα συγχώνευσης αποτελούνται από μία πιο απλή αρχιτεκτονική γνωστή ως feedforward network. Στοχεύουν στην συγχώνευση της οπτικοακουστικής πληροφορίας με τις αναπαραστάσεις κειμένου του BERT. Κατά τη διαδικασία της προσαρμογής, τα βάρη του προεκπαιδευμένου μοντέλου παραμένουν "παγωμένα", επιτρέποντας γρήγορη και οικονομική εκπαίδευση.

Με την διεξαγωγή εκτεταμένης αφαιρετικής μελέτης αποδεικνύεται πειραματικά ότι οι αντάπτορες βοηθούν την επίδοση αν και χρησιμοποιούν πολύ λιγότερες παραμέτρους, επειδή αποφεύγουν κάποια από τα προβλήματων των κλασικών τεχνικών μεταφοράς μάθησης. Επίσης, η προτεινόμενη λύση δείχνει σημάδια ευρωστίας σε θόρυβο εισόδου, το οποίο είναι θεμελιώδες για αληθινές εφαρμογές. Τα πειράματα στο πρόβλημα της ανάλυσης διάθεσης με το CMU-MOSEI αποκαλύπτουν ότι το AMB ξεπερνά σε όλες τις μετρικές το καλύτερο μοντέλο με 3.4% σχετική μείωση στο σφάλμα και 2.1% σχετική βελτίωση στην ακρίβεια 7 κλάσεων.

Λέξεις Κλειδιά

βαθιά νευρωνικά δίκτυα, μεταφορά μάθησης, ρύθμιση βαρών, βοηθητικά συμφραζόμενα, αντάπτορες, BERT, πολυτροπικά δεδομένα, συγχώνευση, ευρωστία, ανάλυση διάθεσης

Abstract

Over the past few years, the abundance of multimedia data and progress in core machine learning algorithms has set the scene for multimodal machine learning as one of the frontiers of applied AI research. The usage of social networks has exploded leading to massive amounts of data available. In addition, the recent success of the so-called Pretrained Language Models (PLMs) has encouraged the creation of many fascinating new applications. However, training these deep networks in multiple stages, as this trend suggests, comes at the cost of increased model parameters.

In this work, we propose Adapted Multimodal BERT (AMB), a BERT-based architecture for multimodal tasks that uses a combination of adapter modules and intermediate fusion layers. Specifically, the task that is going to be tackled is sentiment analysis on videos with text, visual and acoustic data. BERT is a deep pretrained neural network architecture that was originally used for processing language information and consists of multiple neural network layers, which are called transformer layers. The adapter is a neural module that is interleaved in between the layers of BERT in order to adjust the pretrained language model for the task at hand. This allows for transfer learning to the new task, but in contrast with fine-tuning which is the prevalent method, adapters are parameter-efficient. The fusion layers are composed of a simpler feedforward neural network aiming to perform task-specific, layer-wise fusion of audio-visual information with textual BERT representations. During the adaptation process the pretrained language model parameters remain frozen, allowing for fast, parameter-efficient training.

Extensive ablation studies are performed which reveal that this approach leads to an efficient model. Adapters prove empirically to help with performance although they train much less parameters, because they avoid some of the issues with standard approaches of transfer learning. They can outperform these costly approaches which consist of the aforementioned fine-tuning that refines the weights of the model to adapt it to the new task. Also, the proposed model shows signs of robustness to input noise, which is fundamental for real-life applications. The experiments on sentiment analysis with CMU-MOSEI reveal that AMB outperforms the current state-of-the-art across metrics, with 3.4% relative reduction in the resulting error and 2.1% relative improvement in 7−class classification accuracy.

Keywords

AI, deep neural network, transformer, tranfer learning, fine-tuning, prompt-tuning, adapters, BERT, multimodal, fusion, robustness, sentiment analysis

Ευχαριστίες

Η παρούσα διπλωματική εργασία αποτελεί ένα προσωπικό πόνημα στην διαμόρφωση του οποίου συνετέλεσαν πολλοί. Θα ήθελα να ευχαριστήσω πρωτίστως τον επιβλέποντα καθηγητή της εργασίας, τον καθηγητή Αλέξανδρο Ποταμιάνο για την ενασχόλησή του με την ερευνητική μου προσπάθεια, τις καίριες συμβουλές του αλλά και για τις πρωτοποριακές διαλέξεις που με ενέπνευσαν να ασχοληθώ με την μηχανική μάθηση και μου έδωσαν κουράγιο εν καιρώ πανδημίας, σε μία περίοδο δύσκολη για όλους μας. Ένα μεγάλο ευχαριστώ οφείλω στον Γιώργο Παρασκευόπουλο για την πολύτιμη βοήθεια του και τις καθοριστικές του ιδέες.
Θα ήθελα επίσης να ευχαριστήσω όλους όσους μου στάθηκαν στις δύσκολες στιγμές αυτής της προσπάθειας, που δεν ήταν λίγες, και με βοήθησαν να τις ξεπεράσω. Τέλος, το μεγαλύτερο ευχαριστώ το οφείλω στους γονείς μου για όλα όσα μου προσφέρουν καθημερινά.

Table of Contents

List of Figures

List of Tables

Εκτεταμένη Περίληψη στα Ελληνικά

Στο κεφάλαιο αυτό παρουσιάζεται μία εκτεταμένη περίληψη της εργασίας αυτής στα ελληνικά. Με συνοπτικό τρόπο θα διατυπωθούν οι κεντρικές ιδέες από κάθε ενότητα.

0.1 Εισαγωγή

Τα τελευταία χρόνια, ο τομέας των πολυτροπικών εφαρμογών αναπτύσσεται ραγδαία λόγω της πληθώρας πολυμεσικών δεδομένων και της ανάπτυξης των αλγορίθμων μηχανικής μάθησης. Η εκτεταμένη χρήση των μέσων κοινωνικής δικτύωσης έχει επιτρέψει την συλλογή τεράστιων συνόλων δεδομένων. Όμως, δεν είναι δυνατό να γίνει επεξεργασία τέτοιας κλίμακας δεδομένων με χειροκίνητο τρόπο. Για την επίλυση αυτής της δοκιμασίας απαιτείται η ανάπτυξη αποτελεσματικών συστημάτων τεχνητής νοημοσύνης.

0.1.1 Μηχανική Μάθηση

Η μηχανική μάθηση στηρίζεται στην σχεδίαση ενός μοντέλου το οποίο προσαρμόζει τις παραμέτρους του με τη βοήθεια αλγορίθμων βελτιστοποίησης πάνω σε μία πληθώρα στατιστικών δεδομένων. Ένα ιδιαίτερο είδος μοντέλο είναι το γνωστό και ως νευρωνικό δίκτυο. Τα νευρωνικά δίκτυα, και ιδιαίτερα τα βαθιά νευρωνικά δίκτυα, τα τελευταία χρόνια έχουν χρησιμοποιηθεί σε πολλές επαναστατικές εφαρμογές και έτσι βρίσκονται στο επίκεντρο της έρευνας στην τεχνητή νοημοσύνη.

Τα ίδια τα νευρωνικά δίκτυα χωρίζονται σε επιμέρους κατηγορίες μοντέλων. Το πιο απλό εξ αυτών είναι το πλήρως συνδεδεμένο δίκτυο, το οποίο αποτελείται από πολλά επίπεδα νευρώνων τα οποία συνδέουν κάθε νευρωνα της εισόδου με ένα πολλαπλασιαστικό βάρος για να παράξουν τις έξοδους τους. Ένα δεύτερο σημαντικό μοντέλο είναι το συνελικτικό δίκτυο. Αποτελεί τροποποίηση του προηγούμενου, με την προσθήκη συνελίξεων σε μία γειτονά η οποία ονομάζεται πυρήνας. Τα βάρη του πυρήνα είναι κοινά για κάθε νευρώνα. Τα συνελικτικά δίκτυα επεξεργάζονται συνήθως εικόνες. Για την περίπτωση των μονοδιάστατων ακολουθιών, όπως είναι η γλώσσα, ένα άλλο μοντέλο προτιμάται το αναδρομικό νευρωνικό δίκτυο. Το αναδρομικό νευρωνικό δίκτυο επεξεργάζεται κάθε σύμβολο της ακολουθίας με το ίδιο νευρωνικό δίκτυο, αποθηκεύοντας την πληροφορία από τα προηγούμενα σύμβολα. Συχνά τα αναδρομικά δίκτυα εμφανίζονται σε μορφή κωδικοποιητή-αποκωδικοποιητή και σε αυτή την περίπτωση είναι χρήσιμη η εφαρμογή μίας διαδικασίας που ονομάζεται προσοχή. Η προσοχή δίνει ένα βάρος σε κάθε τμήμα της ακολουθίας το οποίο καθορίζει την συνεισφορά του στο τελικό αποτέλεσμα. Με αφορμή την επιτυχία της προσοχής, κάποιοι επιστήμονες επινόησαν ένα εξαιρετικά πετυχημένο μοντέλο που ονομάζεται transformer και χρησιμοποιεί πολλά επίπεδα προσοχής και πλήρως συνδεδεμένου δικτύου για να επεξεργαστεί δεδομένα.

0.1.2 Μεταφορά Μάθησης σε Προεκπαιδευμένα Μοντέλα Transformer

Η πρώτη εργασία που εισήγαγε τη μέθοδο fine-tuning σε μεγάλη κλίμακα ήταν το GPT [6], με την ακόλουθη διατύπωση:

- Ένα βαθύ μοντέλο transformer προεκπαιδεύεται σε ένα μεγάλο σύνολο κειμένων με στόχο να αποκτήσει χρήσιμες γενικές γνώσεις χωρίς επίβλεψη
- Το ίδιο προεκπαιδευμένο μοντέλο στη συνέχεια προσαρμόζεται (fine-tuned) πάνω σε ένα συγκεκριμένο πρόβλημα με λιγότερα δείγματα

Μία άλλη σημαντική εργασία ήταν το BERT [7], το οποίο είναι ένα μοντέλο transformer κωδικοποιητή.

Η μέθοδος fine-tuning παρουσιάζει σημαντικά προβλήματα. Πρώτον, η ρύθμιση των βαρών προκαλεί απώλεια πληροφορίας η οποία ονομάζεται "αμνησία" (catastrophic-forgetting) επειδή τα βάρη περιέχουν χρήσιμη πληροφορία η οποία προέρχεται από το στάδιο της προεκπαίδευσης η οποία θα πρέπει να αλλοιωθεί για την προσαρμογή του μοντέλου στις νέες συνθήκες. Επίσης, το fine-tuning είναι μία κοστοβόρα διαδικασία η οποία απαιτεί πολλές παραμέτρους και ένα αντίγραφο ολόκληρου του μοντέλου για κάθε διαφορετικό πρόβλημα.

Την λύση σε αυτά τα προβλήματα προσφέρουν οι προσαρμογείς ή αλλιώς αντάπτορες [8]. Οι αντάπτορες είναι δίκτυα μικρής κλίμακας τα οποία τοποθετούνται πάνω από κάθε επίπεδο του transformer με σκοπό να το προσαρμόσουν στο νέο πρόβλημα. Όμως, έτσι, τα ίδια τα επίπεδα του transformer παραμένουν αμετάβλητα ή αλλιώς "παγωμένα".

0.1.3 Πολυτροπική Μάθηση και Ανάλυση Διάθεσης

Τα πολυτροπικά δεδομένα είναι δεδομένα που προέρχονται από διαφορετικά αισθητήρια (συχνά κείμενο, εικόνα και ήχος) και έχουν διαφορετικό τρόπο κωδικοποίησης σε ένα νευρωνικό δίκτυο. Αυτή η διαφορά στην αναπαράσταση προκαλεί σημαντικές δυσκολίες στην επίλυση πολυτροπικών προβλημάτων επειδή δυσκολεύει την συγχώνευση της πληροφορίας. Για την αποτελεσματική συγχώνευση μία μέθοδος [9] προτείνει την προβολή των δεδομένων σε έναν κοινό χώρο στον οποίο η πληροφορία από κάθε αισθητήριο έχει συμβατή αναπαράσταση για συγχώνευση. Το μοντέλο MISA [9] που ακολουθεί αυτή την μέθοδο πετυχαίνει αποτελέσματα αιχμής. Μία άλλη δημοφιλής μέθοδος είναι η ανάπτυξη περίπλοκων δικτύων προσοχής σε πολλά επίπεδα για την ευθυγράμμιση των αισθητηρίων.

Ένα άλλο πολύ ενδιαφέρον μοντέλο είναι το MAGMA [10], το οποίο χρησιμοποιεί ένα παγωμένο προεκπαιδευμένο γλωσσικό μοντέλο σε συνδυασμό με προσαρμογείς για να πραγματοποιήσει παραγωγή κειμένου με οπτικά συμφραζόμενα. Η χρήση προσαρμογέων ανάμεσα στα επίπεδα του γλωσσικού μοντέλου φαίνεται να λειτουργούν καλύτερα από την κλασική μέθοδο fine-tuning, σύμφωνα με τους συγγραφείς.

Στην περιοχή της πολυτροπικής ανάλυσης διάθεσης, έχουν προταθεί πολλές ενδιαφέρουσες ιδέες. Η χρήση της προσοχής επέτρεψε την ανάπτυξη πιο εξελιγμένων μορφών συγχώνευσης με πολλές παραλλαγές: ιεραρχική προσοχή [11], transformer με προσοχή για ευθυγράμμιση ανάμεσα σε διαφορετικά αισθητήρια [12], προσοχή για την ανανέωση των αναπαραστάσεων του BERT (μέθοδος shifting) [13], ακόμα και προσοχή για την επιλογή κατάλληλων χαρακτηριστικών [14].

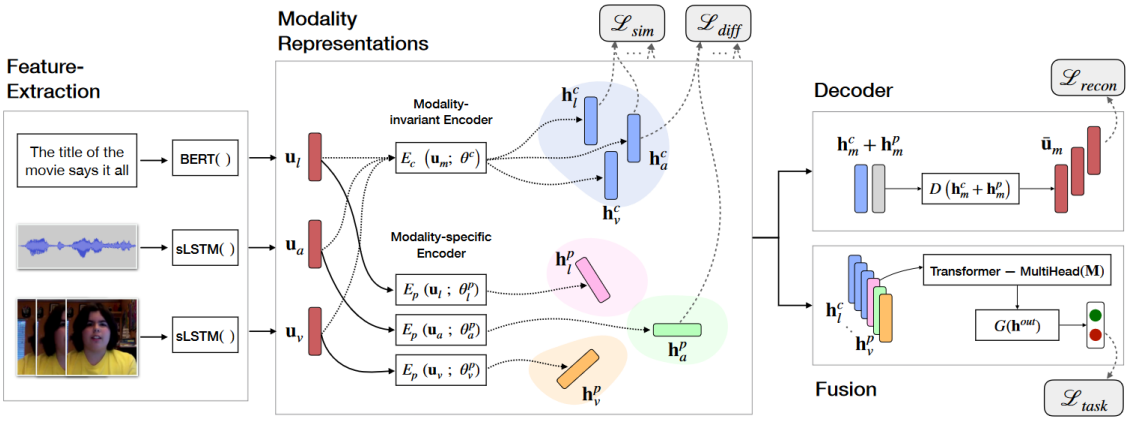


Figure 1. *Αρχιτεκτονική του MISA*

Figure 2. *Η αρχιτεκτονική MAGMA παρουσιάζει ικανότητες απάντησης σε ερωτήσεις με οπτικά συμφραζόμενα*

0.2 Προσαρμοσμένο Πολυτροπικό BERT με επίπεδα συγχώνευσης για ανάλυση διάθεσης

Η βασική συνεισφορά της εργασίας συνοψίζεται στα παρακάτω:

- Παρουσιάζεται το νέο μοντέλο: Προσαρμοσμένο Πολυτροπικό BERT (AMB). Αποτελεί επέκταση του δημοφιλούς BERT [7] και ρυθμίζεται με προσαρμογείς και ενδιάμεσα επίπεδα feedforward δικτύου τα οποία συγχωνεύουν πληροφορία από κείμενο, εικόνα και ήχο για να εκτελέσουν ανάλυση διάθεσης.

- Η αξιολόγηση του μοντέλου στο δημοφιλές σύνολο δεδομένων CMU-MOSEI δίνει αποτελέσματα αιχμής (state-of-the-art).

- Η χρήση αντάπτορα επιτρέπει οικονομική εκπαίδευση με λιγότερο από το ένα πέμπτο των παραμέτρων του προηγουμένου μοντέλου αιχμής, χωρίς να θυσιάζει τις επιδόσεις.

- Πραγματοποιείται αφαιρετική μελέτη η οποία αποδεικνύει πειραματικά ότι η κλασική μέθοδος “fine-tuning” η οποία προτιμάται στη βιβλιογραφία παρουσιάζει χαμηλότερες επιδόσεις αν και απαιτεί πολύ περισσότερες εκπαιδευόμενες παραμέτρους.

- Διεξάγεται μελέτη ευρωστίας η οποία δείχνει ότι το προτεινόμενο μοντέλο είναι εύρωστο στην παρουσία θορύβου στα αισθητήρια.

0.2.1 Αρχιτεκτονική

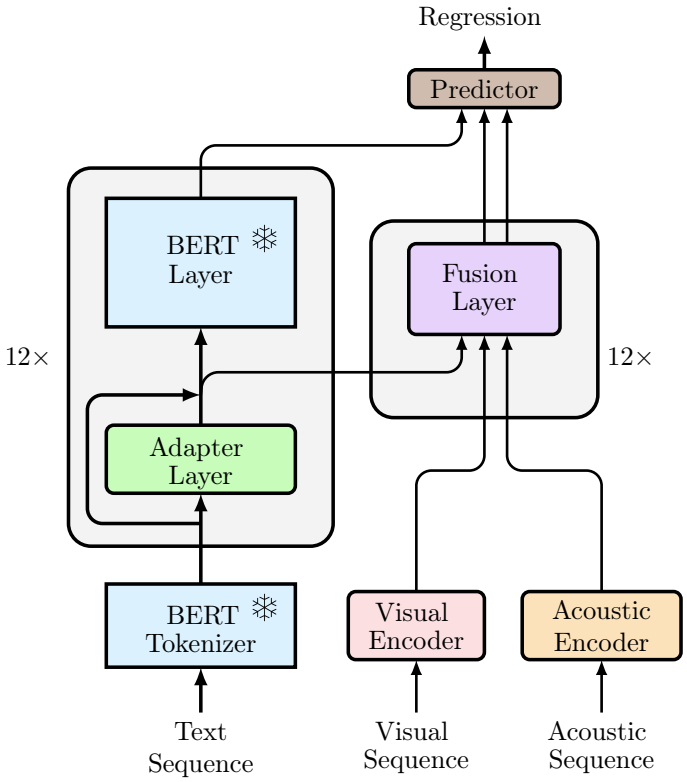


Figure 3. *Architecture of Adapted Multimodal BERT (AMB)*

Το σχήμα 3 παρουσιάζει την αρχιτεκτονική του συστήματος. Αρχικά, η ακολουθία του κειμένου της εισόδου περνάει από το παγωμένο στάδιο προεπεξεργασίας του BERT (BERT tokenizer) για να μετατραπεί σε ακολουθία από διανύσματα-λέξεις(tokens) του BERT. Παράλληλα, η ακουλουθίες με τα οπτικά και ακουστικά χαρακτηριστικά περνάνε από εκπαιδευόμενους κωδικοποιητές προκειμένου να μεταφραστούν σε ένα ειδικό διάνυσμα-λέξη συμβατό με την αναπαράσταση του BERT. Ο κορμός της αρχιτεκτονικής αποτελείται από ένα παγωμένο προεκπαιδευμένο μοντέλο BERT το οποίο ρυθμίζεται από επίπεδα αντάπτορα, χωρίς πρόσβαση στα άλλα αισθητήρια. Οι αναπαραστάσεις του BERT συνδυάζονται σε κάθε επίπεδο με οπτικο-ακουστική πληροφορία σε ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο (feed-forward network - FFN) για την επίτευξη πολυτροπικής συγχώνευσης. Η διαδικασία αυτή επαναλαμβάνεται σε 12 επίπεδα και οι τελευταίες αναπαραστάσης δίνονται σε ένα FFN για να προβλέωει το σκορ της διάθεσης.

Παγωμένα επίπεδα BERT

Το παγωμένο μοντέλο BERT αποτελεί τον κορμό της αρχιτεκτονικής για να δωθεί έμφαση στην σημασία της γλώσσας. Και το στάδιο προεπεξεργασίας του BERT (BERT tokenizer),

αλλά και τα 12 επίπεδα του BERT, παραμένουν αμετάβλητα κατά την εκπαίδευση, μειώνοντας έτσι τις επιπτώσεις της αμνησίας που μπορεί να προκύψουν από τη ρύθμιση των βαρών.

Επίπεδα Προσαρμογέων

Στο μοντέλο χρησιμοποιείται το πρωτότυπο είδος προσαρμογέα τύπου "bottleneck" [8]. Κάθε επίπεδο αποτελείται από μία γραμμική προβολή σε χαμηλότερη διάσταση ακολουθούμενη από μία μη-γραμμικότητα τύπου "ReLU" και τέλος μία γραμμική προβολή για επαναφορά στις αρχικές διαστάσεις. Χρησιμοποιούνται συνδέσεις υπολοίπου (residual) ανάμεσα στην είσοδο και την έξοδο του προσαρμογέα. Αντί για την προσθήκη προσαρμογέα και στην προσοχή και στο FFN, ακολουθώντας το [15], προστίθενται μόνο ένα επίπεδο προσαρμογέα, μετά από το FFN. Έτσι, μειώνεται ο αριθμός των παραμέτρων στο μισό. Οι προσαρμογείς είναι τοποθετημένοι με τέτοιο τρόπο ώστε είναι υπεύθυνοι για την προσαρμογή μόνο της πληροφορίας από το κείμενο και όχι από τα άλλα αισθητήρια.

Οπτικοί και Ακουστικοί Κωδικοποιητές

Οι οπτικοί και ακουστικοί κωδικοποιητές αποτελούνται από επίπεδα κωδικοποιητών transformer τα οποία επιδρούν σε κάθε αισθητήριο ξεχωριστά για την εξαγωγή πληροφορίας από μία ακολουθία αυθαίρετου μήκους και την συμπίεση της σε ένα συμπαγές οπτικοακουστικό διάνυσμα-λέξη. Αυτό το διάνυσμα προετοιμάζεται για το επόμενο στάδιο στο οποίο θα συντελεστεί συγχώνευση με την πληροφορία του κειμένου. Η σχεδίαση των κωδικοποιητών αντλεί έμπνευση από τις εργασίες των [16, 10, 17], με την προσθήκη του ακουστικού αισθητηρίου.

Επίπεδα Συγχώνευσης

Για τα επίπεδα συγχώνευσης χρησιμοποιείται feedforward network. Το πρώτο στοιχείο του BERT, γνωστό ως "CLS token" χρησιμοποιείται σαν σύνοψη για την πληροφορία των κρυφών καταστάσεων ενός επιπέδου [18]. Αυτό το στοιχείο προβάλλεται αρχικά σε χαμηλότερα διάσταση και έπειτα παρατίθεται με το οπτικοακουστικό διάνυσμα-λέξη για να δωθεί σαν είσοδος στο επίπεδο συγχώνευσης. Αν και οι [13, 17] επίσης εφαρμόζουν συγχώνευση, και οι δύο χρησιμοποιούν το αποτέλεσμα για να μετατοπίσουν τις αναπαραστάσεις του BERT. Η δική μας εκδοχή της συγχώνευσης είναι απλούστερη και αποδεικνύεται επαρκής για εξαιρετικά αποτελέσματα.

Προβλέπτης

Η συνολική αναπαράσταση του τελευταίου επιπέδου δίνεται σε έναν προβλέπτη τύπου FFN που δίνει την έξοδο του συστήματος. Το σύστημα εκπαιδεύεται με στρατηγική end-to-end, δηλαδή σε ένα στάδιο (με εξαίρεση την προεκπαίδευση του BERT).

Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	Trainable Parameters
MMLatch (G) [14]	0.582	0.704	52.1	82.8	82.9	2.6
MulT (G) [12]	0.580	0.703	51.8	82.5	82.3	1.8
LMF (B) [19]	0.623	0.677	50.2	82.0	82.1	1.0
TFN (B) [20]	0.593	0.700	51.8	82.5	82.3	0.6
MFM (B) [21]	0.568	0.717	51.3	84.4	84.3	1.7
ICCN (B) [22]	0.565	0.713	51.6	84.2	84.2	–
MAG-BERT* (FT) [13]	0.614	0.763	50.9	84.3	84.2	110.8
MISA (FT) [9]	0.555	0.756	52.2	85.3	85.3	47.1
AMB (Ours)	**0.536**	**0.766**	**53.3**	**85.8**	**85.8**	8.6

Table 1. *Αποτελέσματα στο CMU-MOSEI. Μοντέλα με (G) χρησιμοποιούν εμφυτεύματα glove. Τα μοντέλα με (B) χρησιμοποιούν παγωμένα χαρακτηριστικά BERT και προέρχονται από τα πειράματα στο [22]. Τα μοντέλα MISA και MAG-BERT χρησιμοποιούν μέθοδο fine-tuning (FT) για το BERT. Τα πειράματα του MAG-BERT* έχουν παραχθεί για αυτήν την εργασία από τον δημοσιευμένο κώδικα. Ο αριθμός των εκπαιδευόμενων παραμέτρων (Trainable Parameters) είναι σε εκατομμύρια.*

0.2.2 Πειραματική Διαδικασία

Διάταξη

Η εμπειρική αξιολόγηση της επίδοσης του μοντέλου γίνεται με τη βοήθεια του συνόλου δεδομένων για ανάλυση διάθεσης CMU-MOSEI. Αποτελείται από 23.454 βίντεο από το YouTube τα οποία αφορούν κριτικές ταινιών και άλλα θέματα. Τα δεδομένα εισόδου αποτελούνται από ακολουθίες του κειμένου απομαγνητοφώνησης, καθώς και οπτικά και ακουστικά χαρακτηριστικά. Για την αξιολόγηση χρησιμοποιούνται κλασικές μετρικές παλινδρόμησης: Μέσο Απόλυτο Σφάλμα (MAE), συντελεστής συσχέτισης (Corr) και ταξινόμησης: ακρίβεια 7 κλάσεων (Acc-7), ακρίβεια 2 κλάσεων (Acc-2) και F1 σκορ (F1). Για το μοντέλο BERT χρησιμοποιείται η έκδοση `bert-base-uncased`. Η εκπαίδευση διαρκεί 20 λεπτά σε μία κάρτα γραφικών GTX 1080Ti NVIDIA.

Αποτελέσματα

Τα αποτελέσματα της ανάλυσης διάθεσης στο CMU-MOSEI φαίνονται στον πίνακα 1. Παρατηρείται σημαντική βελτίωση σε όλες τις μετρικές από το προτεινόμενο μοντέλο AMB. Παράλληλα, όπως φαίνεται από την εικόνα 4, οι υψηλές επιδόσεις συνδυάζονται με χαμηλό κόστος παραμέτρων λόγω της χρήσης αντάπτορα.

Αφαιρετική Μελέτη (Ablation)

Ο Πίνακας 2 παρουσιάζει τις επιπτώσεις της αφαίρεσης αισθητηρίων και τα αποτελέσματα της χρήσης αντάπτορα στη θέση του fine-tuning για την προσαρμογή του γλωσσικού μοντέλου. Αρχικά, η απόρριψη του κειμένου είναι καταστροφική για το μοντέλο "AMB no-text", φανερώνοντας την κυριαρχία του κειμένου επί των άλλων αισθητηρίων. Αλλά και η απόρριψη των άλλων αισθητηρίων οδηγεί σε κάποια μικρή μείωση της επίδοσης, οπότε όλα τα αισθητήρια είναι σημαντικά για την επίτευξη καλών επιδόσεων.

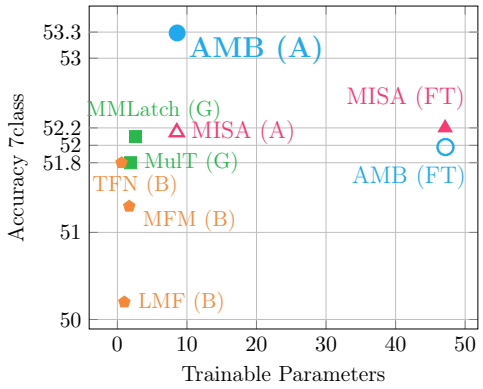


Figure 4. *Διάγραμμα ακρίβειας 7 κλάσεων - εκπαιδευόμενων παραμέτρων για τα καλύτερα μοντέλα από τη βιβλιογραφία. Το G συμβολίζει εμφυτεύματα GloVe, το A αντάπτορες, το B παγωμένα και το FT fine-tuned εμφυτεύματα BERT. Το προτεινόμενο AMB με αντάπτορες παρουσιάζει μία καλή ισορροπία ανάμεσα σε επίδοση και εκπαιδευόμενες παραμέτρους.*

Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	Trainable Parameters
AMB no-text	0.816	0.240	41.6	63.3	61.8	8.6
AMB text-only	0.541	0.760	52.8	85.7	85.7	8.6
MISA-Adapters	0.5480	0.758	52.1	85.8	85.8	8.5
MISA	0.555	0.756	52.2	85.3	85.3	47.1
AMB-FT	0.548	0.756	51.9	85.4	85.3	47.2
AMB	**0.536**	**0.766**	**53.3**	**85.8**	**85.8**	8.6

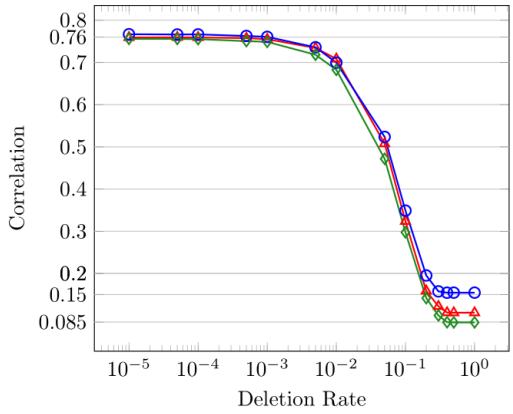
Table 2. *Αντάπτορες και fine-tuning. Παρατίθενται πειράματα με αφαίρεση είτε κειμένου (no-text) είτε οπτικοακουστικής πληροφορίας (text-only). Ο αριθμός των εκπαιδευόμενων παραμέτρων (Trainable Parameters) είναι σε εκατομμύρια.*

Για την σύγκριση του fine-tuning με τους προσαρμογείς, υλοποιούνται μία εκδοχή του MISA με προσαρμογείς ("MISA-Adapters") καθώς και μία εκδοχή του AMB με fine-tuning ("AMB-FT"). Στην περίπτωση του MISA, το fine-tuning αποδεικνύεται περιττό, ενώ στην περίπτωση του AMB είναι και επιβλαβές. Η μείωση στην επίδοση του AMB με fine-tuning δεν μπορεί παρά να οφείλεται στο φαινόμενο της αμνησίας.

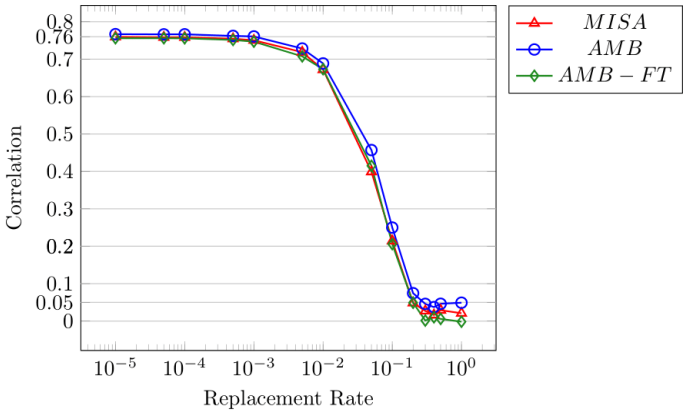
Μελέτη Ευρωστίας

Σε αυτήν την ενότητα ερευνάται η ευρωστία του AMB στην εισαγωγή θορύβου. Για τα οπτικά και τα ηχητικά χαρακτηριστικά ακολουθείται η μέθοδος των Hazarika et al. [23]. Για το κείμενο προτείνονται δύο νέες μέθοδοι: διαγραφή και αντικατάσταση λέξεων. Για το πείραμα της αντικατάστασης λέξεων ένα ποσοστό των λέξεων της εισόδου επιλέγεται τυχαία και αντικαθίστανται από μία λέξη του λεξικού, ενώ για την περίπτωση της διαγραφής αντικαθίστανται από το ειδικό στοιχείο της άγνωστης λέξης γνωστό ως [UNK].

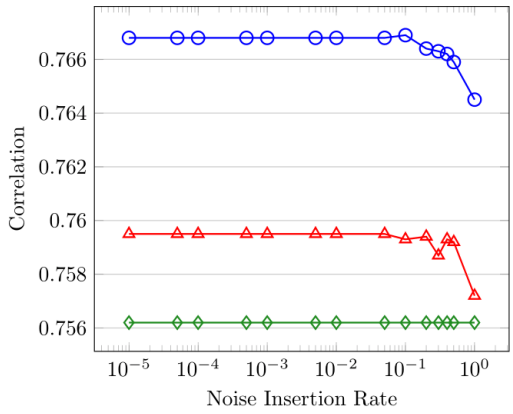
Η εικόνα 5 δείχνει τα αποτελέσματα της έρευνας. Στην περίπτωση της διαγραφής και της αντικατάστασης λέξεων παρατηρείται παρόμοια συμπεριφορά για τα τρία μοντέλα, αν και το AMB με προσαρμογείς φαίνεται πιο εύρωστο από το AMB με fine-tuning. Στην περίπτωση των οπτικών χαρακτηριστικών η πτώση της επίδοσης είναι φανερά μικρότερη για τα AMB και



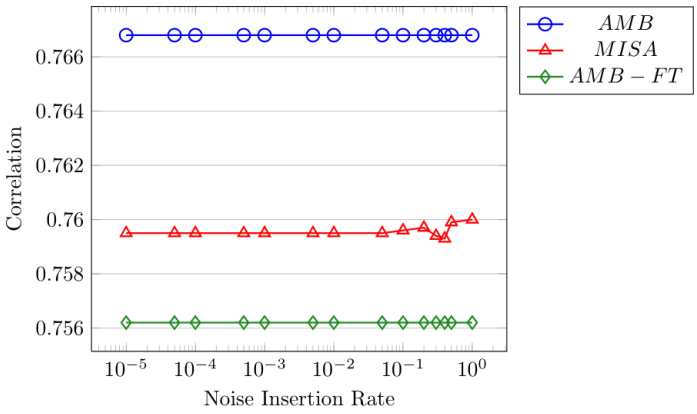
(a) *Τυχαία διαγραφή λέξεων*



(b) *Τυχαία αντικατάσταση λέξεων*



(c) *Τυχαία εισαγωγή θορύβου σε οπτικά χαρακτηριστικά*



(d) *Τυχαία εισαγωγή θορύβου σε ηχητικά χαρακτηριστικά.*
Τα αποτελέσματα είναι κάπως απρόσμενα.

Figure 5. *Ευρωστία του μοντέλου για διαφορετικά επίπεδα θορύβου. Τυχαία διαγραφή στοιχείων εισόδου (αριστερά), τυχαία αντικατάσταση στοιχείων εισόδου (μέση) και τυχαία εισαγωγή θορύβου στα οπτικά χαρακτηριστικά (δεξιά). Γαλάζιο °: AMB, Κόκκινο △: MISA, Πράσινο ◇: AMB-FT.*

MISA, όμως το AMB-FT παραμένει τελείως ανεπηρέαστο από την εισαγωγή θορύβου. Αυτό δείχνει ότι αυτό το μοντέλο επαφίεται αποκλειστικά στο κείμενο για να κάνει προβλέψεις. Άρα, η χρήση των επιπέδων προσαρμογέων φαίνεται να βοηθάει στην αξιοποίηση της πληροφορίας των λιγότερο κυριαρχικών αισθητηρίων. Αντίθετα, στην εισαγωγή θορύβου στα ηχητικά χαρακτηριστικά τα μοντέλα φαίνεται να συμπεριφέρονται κάπως απρόσμενα. Δηλαδή, αντί να μειώνεται η επίδοσή τους με την εισαγωγή όλο και περισσότερο θορύβου στο αισθητήριο, για τα μοντέλα AMB και AMB-FT παρατηρείται μηδενική μεταβολή. Άρα, αυτό σημαίνει ότι ούτε το AMB, αλλά ούτε και το AMB-FT επηρεάζονται από το αισθητήριο αυτό. Στην περίπτωση όμως του AMB, φάνηκε ότι τα οπτικά χαρακτηριστικά είναι σημαντικά για την λήψη αποφάσεων, άρα δεν μπορεί να φταίει ο μηχανισμός συγχώνευσης για το "απρόσμενο" αποτέλεσμα. Για το MISA συμβαίνει κάτι ακόμη πιο ύποπτο, αφού ενώ η εισαγωγή θορύβου στα οπτικά χαρακτηριστικά δυσκολεύει το MISA, στην περίπτωση των ακουστικών χαρακτηριστικών φαίνεται να υπάρχει ακόμη και βελτίωση. Κατά τη γνώμη μας, τα αποτελέσματα

δείχνουν μία ξεκάθαρη αδυναμία των ηχητικών χαρακτηριστικών, αλλά το ζήτημα απαιτεί περισσότερη διερεύνηση για να ληφθεί ένα τελικό συμπέρασμα.

0.3 Συμπεράσματα

Σε αυτήν την εργασία προτείνεται το μοντέλο AMB, ένα απλό αλλά ταυτόχρονα πρωτοποριακό μοντέλο το οποίο χτίζει πάνω στο ισχυρό προεκπαιδευμένο κωδικοποιητή BERT και αποφεύγει τους κινδύνους της κλασικής μεθόδου fine-tuning. Η χρήση του αντάπτορα επιτρέπει στο μοντέλο μας να μειώσουν το κόστος των εκπαιδευόμενων παραμέτρων χωρίς να θυσιάζει από την επίδοσή του, αφού καταφέρνει νέα επίδοση αιχμής στο CMU-MOSEI. Επιπρόσθετα, αποδείχτηκε πειραματικά ότι η χρήσιμη γνώση από το στάδιο προεκπαίδευσης συνδυάζεται αρμονικά με την οπτικοακουστική πληροφορία σε αυτό το μοντέλο, αποφεύγοντας έτσι τα δομικά ζητήματα της "αμνησίας" και της ανισορροπίας των αισθητηρίων. Τέλος, η μελέτη ευρωστίας έδειξε ότι το προτεινόμενο μοντέλο είναι αξιόπιστο και εύρωστο στην περίπτωση τυχαίας εισαγωγής θορύβου, το οποίο είναι ουσιαστικό ζήτημα για την ανάπτυξη εφαρμογών στον πραγματικό κόσμο.

Chapter 1

Introduction

1.1 Motivation

Over the past few years, the field of multimodal applications has witnessed impressive breakthroughs due to the abundance of multimedia data and progress in core machine learning algorithms. Social networks are more active than ever , as depicted in Fig. 1.1, uploading a massive amount of multimedia and other kinds of data online for anyone to access freely. It is not humanely possible to manually process and digest information from these extreme amounts of data. As E.O.Wilson puts it: "We are drowning in information, while starving for wisdom." Many digital applications, such as conversational virtual assistants, have emerged, aiming to help humans in this laborious task by processing their requests through their understanding of natural language. However, these approaches are limited by their inadequate abilities to leverage context from other modalities, such as visual and acoustic, that humans commonly use to enrich their social interactions. This has set the scene for multimodal machine learning as one of the frontiers of applied research in artificial intelligence (AI).

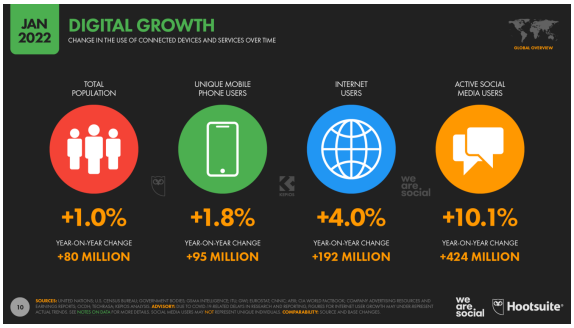


Figure 1.1. *The digital report was taken from [24].*

According to [25], "You can't learn language from the radio." The understanding of current methods in natural language understanding (NLU) is only based on statistical patterns that arise by studying words and their context in sentences. However, most of the knowledge that humans use to understand language is built on assumptions that have risen from sensory perception on the physical world. These assumptions, also referred to as "common sense", are only implied during conversation, never directly mentioned. For this reason, in order for an artificial system to acquire NLU abilities as close to human

understanding as possible, it is required to expand its perception to incorporate more modalities in order to "ground language".

Another important factor for real-world applications that humans might consider useful, especially in the context of AI assistants, is the illusion of empathy. Human behaviour is deeply connected with social interaction and non-verbal cues. Understanding the feelings of others is often demanded in order to understand completely what they mean. This is especially the case in the context of strong emotions that lead to humour or irony. In these cases, it is hard to infer accurate conclusions only with language, as the meaning is implied by visual cues, such as face expressions, and acoustic cues, that are known in the field of linguistics as prosody.

As mentioned, the rapid progression of machine learning research has recently enabled AI systems to tackle advanced multimodal problems. Innovative deep neural network models and specifically the transformer, have been instrumental to the success of large-scale language models that can understand natural language competently. This has inspired the research community to augment these models with other modalities in order to expand their abilities and increase their performance.

Real-world AI applications are extremely demanding. The industry's requirements on costs are strict and the computational power of portable devices is limited, imposing serious restrictions in the budget of trainable parameters. For this reason, models that strike the correct balance between performance and parameter efficiency should be developed in order to be flexible and modular. Flexible in the sense that they can be adapted easily on a new task if is necessary and modular in the sense that if the core system is updated, the internal system can remain exactly the same or adapt with a few minor modifications. Furthermore, when deploying software in the wild the importance of adapting to errors is priceless and for this reason it is crucial to study the robustness of these models to noise inserted in their input modalities. The popular techniques employed by deep learning researchers and engineers seem to be inadequate to the aforementioned requirements and this motivates the proposal of a new approach that overcomes these issues.

1.2 Contributions

The main contributions of this thesis are the following:

- We present Adapted Multimodal BERT or AMB for short, a multimodal extension of a popular deep learning model called BERT [7], which is tuned with adapter layers and intermediate feedforward network layers perform multimodal fusion with textual, visual and acoustic information in order to perform sentiment analysis.

- We test our model on a popular dataset called CMU-MOSEI and it achieves state-of-the-art performance.

- The use of adapters allows our model to be lightweight, training less than one fifth of the parameters of the previous state-of-the-art, without sacrificing performance.

- Our ablation study proves empirically that standard fine-tuning, which is the preferred method in the literature, falls behind adapter-tuning, even though it demands a much larger parameter budget.
- Our robustness study shows that our model is robust to random insertion of noise in its input modalities.

1.3 Outline

In chapter 2, Machine Learning, the preliminaries of machine learning are presented and the adoption of deep learning techniques is motivated. In addition, the most popular deep learning models are explained including the most fundamental for this and many other works, the transformer. In chapter 3, Transfer Learning, the standard approaches for successful pretraining techniques are discussed as well as alternative methods that avoid some of the issues that are present with the previous. Chapter 4, Multimodal Learning, is an introduction to the fascinating field of deep learnign with multiple sensory inputs. The most important techniques are analyzed and the operation of the advanced visual-language models is clarified. Multimodal Sentiment Analysis, which is the task that our method is evaluated, is the point of emphasis of chapter 5. The available datasets, the most influential methods as well as a method for diagnosing the robustness of the evaluated model are all discussed. In chapter 6, Adapted Multimodal BERT (AMB), which is our proposed method, is finally presented. The architecture is thoroughly explained and the experiments are presented in great detail with clear figures. In chapter 7, Conclusions, the results of our work are discussed and some ideas for future work are proposed together with some ethical considerations that every researcher should keep in mind.

Chapter 2

Machine Learning

2.1 Introduction

According to Mitchell [26] "machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks". Essentially, the programmer designs a machine learning algorithm which is then left to "learn" from data alone, only with the help of statistical models and optimization algorithms. The data used to train these models is called training data and the set of parameters that can be adjusted through this procedure, along with their underlying structure is often described as a machine learning model.

A very special kind of model is the so-called neural network. Neural networks, and especially deep neural networks, are at the forefront of machine learning research with fascinating results. In recent years they have been used plentifully in exciting applications and they are often speculated to possess "Artificial Intelligence" or in rare cases even consciousness [27]. In fact, more often than not, AI applications involve either neural networks or other machine learning techniques and for this reason all of these terms are used interchangeably.

Earlier efforts in artificial intelligence focused on expert knowledge systems which, based on logical inference rules, derived new fragments of knowledge or reasoned over statements. This is also referred to as symbolic AI and is characterized by serious limitations such as the difficulty of formally describing all possible knowledge based on a given task. Machine learning approaches such as neural networks were initially disregarded due to infeasibility concerns which, however, were refuted by the technological developments around storage and processing power, enabling the success of modern artificial intelligence with real life applications broadly used.

2.2 Machine Learning Concepts

2.2.1 Types of Learning

There are two main learning paradigms depending on the nature of the data samples and the type of feedback they can provide to the learning system. If the samples are organised in pairs of examples with their desired outputs, known as labels, then this paradigm is called supervised learning. In this case, models can be trained by minimizing an expression of the

error, named loss function, between predictions and labels. This encourages the model to ”learn” to predict the correct labels in a statistical manner. The other paradigm is called unsupervised learning and in this case only the samples are given, without any labels. The goal in this scenario is to apply a certain strategy or rule to guide the algorithm in order to explore its own view of the internal structure of the given samples. This can lead to grouping data in categories called clusters, or alternatively, projecting them to a latent space of lower dimensions, where each dimension has a specific intuitive intepretation. In this work we will focus on Supervised Learning.

2.2.2 Neural Networks

A neural network is the most popular machine learning model at the foundations of deep learning. Originally inspired by biological neurons in the human brain, it consists of nodes that interact with non-linear mappings. The nodes are organised in layers, typically the input and output layers, as well as a number of hidden layers that are meant for internal computations that will help the output layer predict the output successfully.

2.2.3 Training Machine Learning Models

Loss Function

The loss function is a metric to evaluate the distance between a prediction and the ground truth in the supervised learning setting. In other words, the loss function measures magnitude of the error of a prediction. The goal of training is to minimize the total loss as computed with this function.

Backpropagation

In order to train a model on a single sample-label pair, the first procedure is to obtain a prediction for this sample. This is called the forward pass and consists of calculating the output of each layer and feeding it as input to the next layer, until the output layer which produces the prediction. Now that the prediction is obtained, the loss function is used to evaluate the errors of this prediction compared to the given label. The key for training is not the value of the loss function but its gradient. After calculating the gradient of the output compared to the label, this gradient can be used with the help of the chain rule to compute the gradient of the last hidden layer before the output layer. This procedure can be iterated for all hidden layers, essentialy propagating the gradients back to the input layer, hence the name backpropagation. To complete the procedure, the gradients of each hidden layer are used to update the weights of each node in the layer, so that the total loss would decrease. Backpropagation is repeated on all the samples of a dataset once to complete an epoch and then multiple epochs take place until convergence of the model is succeded.

Generalization and Regularization

The ability of a machine learning model to achieve accurate predictions for previously unseen data is called generalization. Generalization is the end goal of every machine learning model in the supervised setting, because the labels for the samples used in the training procedure are already known. In the case that there are too many parameters, such as in deep models, and not enough data the model can learn the exact features of all the possible samples. This model forms a trivial solution that has a zero total loss in training, but cannot generalize. This problem is called overfitting. To solve this problem, some procedures have been proposed under the umbrella term of regularization. Regularization methods in general consist of introducting constraints in the parameters of the network in an effort to encourage it to learn meaningful relationships that lead to generalization.

2.3 Deep Learning

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." John von Neumann

2.3.1 Why depth?

A great explanation of the intuition behind the success of deep neural networks ironically comes from work on compressing neural networks. The concept is called "Lottery Ticket Hypothesis" [28] and it can be stated quite simply: "A randomly-initialized, dense neural network contains a subnetwork that is initialized such that —when trained in isolation— it can match the test accuracy of the original network after training for at most the same number of iterations."

This subnetwork is called a winning ticket. With the lottery ticket hypothesis in mind, the intuition of why depth works is obvious. Assuming perfect balance between depth and regularization, an assumption rather not trivial, deeper models will certainly provide more combinatorial possibilities and so more winning tickets. If combined in a complementary manner, like an ensemble of machine learning models, these tickets will display excellent performance.

2.3.2 Architecture

But if the success of deep networks is only based on randomness why do we need more than one basic model?
Not so fast. There are many more aspects to deep learning than just having a lot of random layers stuck one on top of the other. The most important is to design structure that will aid the emergence of winning tickets. This is done by ensuring the aforementioned regularization and also creating suitable conditions for winning tickets to occur. The introduction of structure in a deep neural network, aiming to find winning tickets more reliably and efficiently, is called architecture.

2.3.3 Feedforward Neural Networks

A Feedforward Neural Network (FFN) is a neural network that its nodes do not form recurrent connections. The simplest FFN only has a single node and is called the perceptron. The input node calculates the sum of the input features and applies a non-linear mapping, which is called the activation function, to obtain the output features. In a Multi-layer Perceptron (MLP) many nodes are combined to form a layer. There are three types of layers: an input layer, a number of hidden layers and an output layer. Multiple layer are stacked one on top of the other to form a chain of hidden layers. The number of hidden layers in the chain is called depth and the name "deep learning" arose from this terminology. They are the basis for most deep neural architectures. The only structural points introduced are in the choice of the depth and size of the hidden layer chain. The rest is left to be decided by the learning algorithm. This unstructured approach gives total freedom to the network to take its own decisions, resulting in a highly non-interpretable and unintuitive final model. This is a disadvantage in the general case, but if prior information about the structure of the input data is lacking but there are enough of them, then, it is a viable solution.

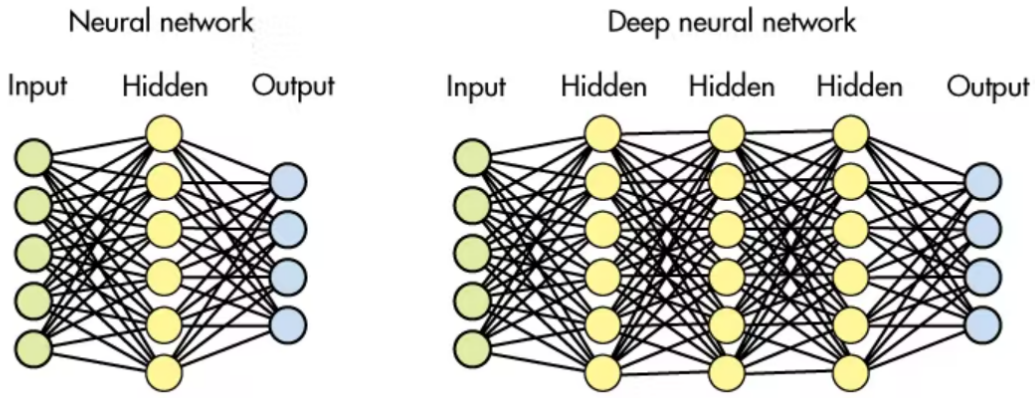


Figure 2.1. *Multilayer Perceptron (MLP) and Feedforward Network (FFN). A deep FFN is simply an extension of the MLP to include more hidden layers. Taken from [1]*

2.3.4 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are restricted versions of FFNs. The restriction is imposed by convolutional layers instead of fully connected layers. They are organised with many of these layers stacked one on top of the other to form a deep neural architecture. Convolution is achieved by connecting each neuron with its neighborhood of neurons. The weights of these connections are computed by a convolutional kernel depending on the relative positions of the two neurons. This kernel is then shared for the whole layer. The output of a convolutional layer is called feature map.

The mathematical equation for the operation of convolution is the following:

$$(I * K)(i, j) = \sum_m \sum_n I(i + m, j + n)K(m, n)$$

The goal of this design is to take advantage of the grid-structure of data (organised in pixels usually) and introduce an inductive bias. The introduced bias is called translation equivariance and it means that an object can be translated in any position of the image and the network will interact with it in the exact same way. This is true for any object in the 3-dimensional world. Unfortunately, this inductive bias does not include rotations, so the model has to learn how objects rotate on its own, but this does not seem to limit CNNs success.

In between convolutions, most CNN architectures interleave pooling layers. These layers lower the dimensions of data by picking a representative value for a cluster of neurons as the output of this cluster. This will be fed to the next layer of convolutions. Commonly, the representative value is simply the maximum output of the cluster and in this case this process is called max-pooling.

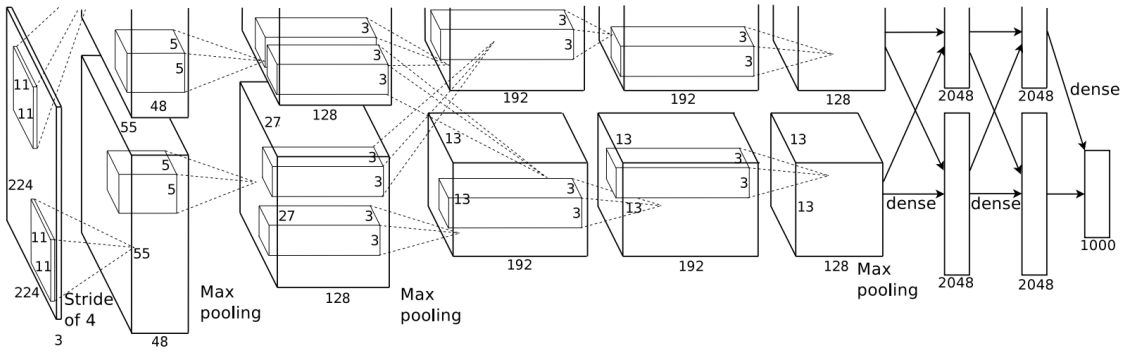


Figure 2.2. *AlexNet [2] Architecture*

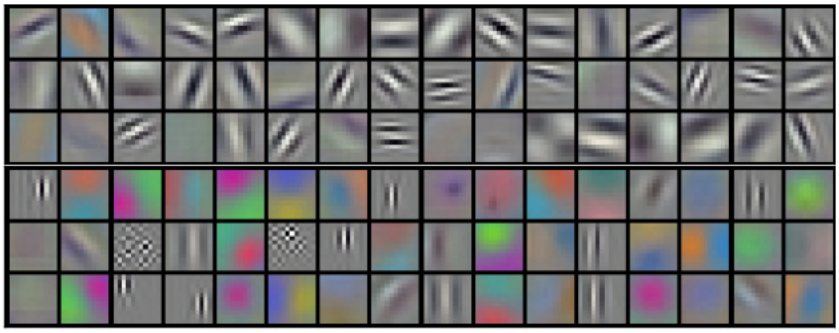


Figure 2.3. *Feature map of 96 layers learnt from the first convolutional layer of AlexNet [2]*

2.3.5 Recurrent Neural Networks

Convolutional networks are ideal for processing a grid of features, but they cannot be leveraged as successfully for symbolic sequences such as language. For this purpose recurrent neural networks (RNNs) are preferred. Their accepted input is a sequence of elements that could be for example parts of words. These elements are called tokens. As is the case with CNNs, imposing constraints is key for RNNs too. The primary characteristic

of a RNN is that the same weights that are used by the RNN function to process one token are used for all tokens; they are shared. This forces the model to learn a universal way to process all tokens and during this process the network learns a lot about the tokens' properties.

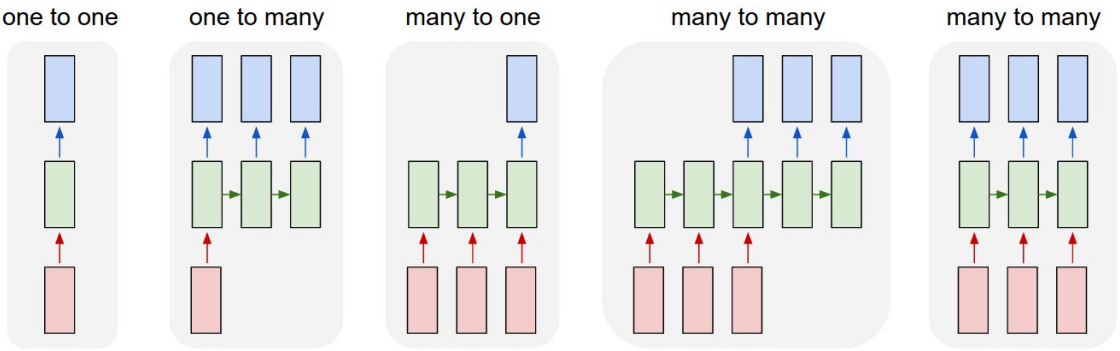


Figure 2.4. *Various alternative sequential arrangements. Image taken from [3].*

[29] proposed an innovative design for RNNs called sequence to sequence that was originally proposed for translation and is still used in various applications today. The concept of this design is simple: an encoder reads the input and the output of this encoder is passed to a decoder that has to produce the final output. The encoder reads the whole sentence and transforms it to a compact semantic hidden state and the decoder is responsible for using this latent code to produce natural output in the target-language. Instead of solving the problem all at once, the task is simplified and split to two subtasks that are solved by each module and this allows for effective end-to-end training. For an in-depth review of CNNs and RNNs we refer to Goodfellow et al. [30].

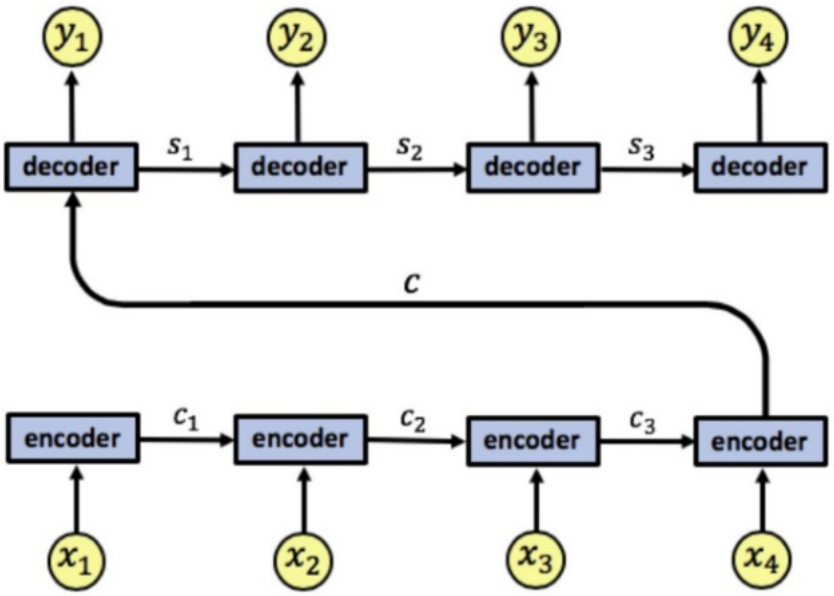


Figure 2.5. *RNN encoder-decoder architecture [4]*

2.3.6 Attention

The development of encoder-decoder architectures with RNNs gave rise to an old issue that was seemingly solved by LSTMs: long-term dependencies. LSTMs helped overcome some issues with the introduction of gates, but, still, very long-dependencies remained undetected. This became especially evident in the task of neural machine translation, where the goal is to translate a sentence from one language to another and exact alignment is demanded. The system pays more attention to the last parts of the sequence, because information from the earlier hidden states has been overwritten multiple times. There is no way to factor in earlier positions, or emphasize some input words compared to others while translating the sentence. A solution to this problem was provided with the introduction of the attention function. After the encoder, an attention layer is used to collect information from all hidden states, i.e. temporal attention. The collected information, which is a semantic summary of the input, is then passed as input to the decoder to make predictions. In this way long-range dependencies are taken into account by adjusting the attention weights.

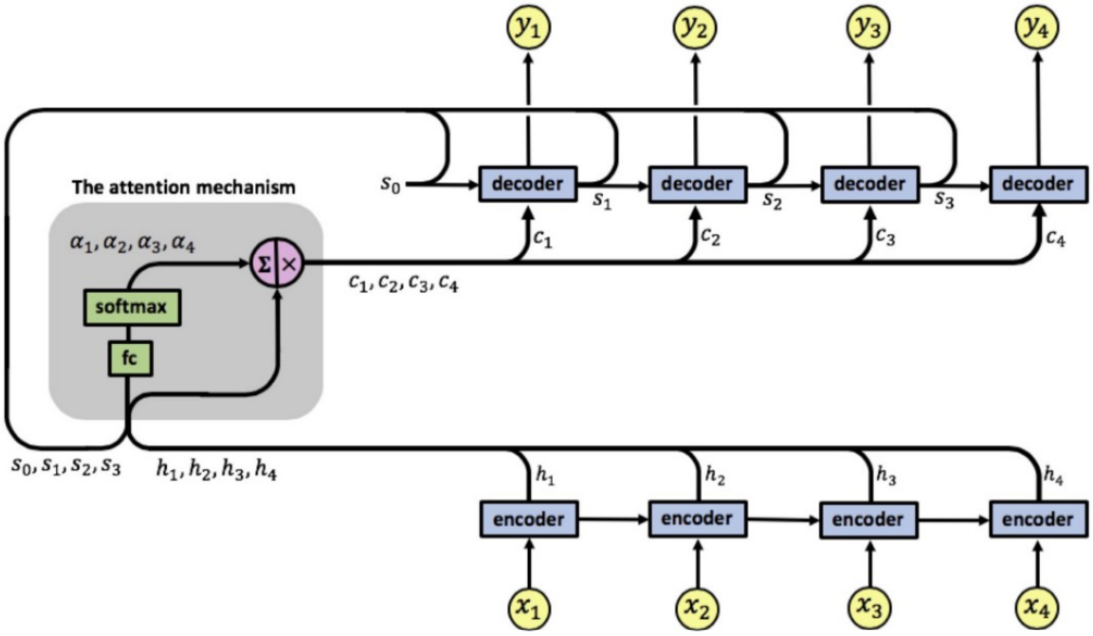


Figure 2.6. *RNN encoder-decoder with attention*

The attention function in principle maps a query to a weighted sum of values. Each value corresponds to a key to form a key-value pair. The weight of each value in the sum is computed according to the compatibility function between the query and the paired key. So if a key matches well with the query the corresponding paired value will have a high weight in the sum and so on.

This procedure is analogous to the mechanism of retrieval systems. For example in a search engine system the user inputs a string as a query and the system matches this string with a set of keys using a compatibility function. The values with the highest scoring keys will be returned as the best candidate results.

2.3.7 Transformers

The transformer model was introduced by [5]. It consists of a series of attention and FFN layers. In between the attention and the FFN a normalization function called "Add & Norm" is used. The specific attention is called self-attention because it does not require another sequence; it uses only the input sequence. The FFN is called position-wise because it is applied to each token of the sequence independently.

Self-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. The difference with vanilla attention is that instead of interacting temporally, self-attention interacts spatially. The Scaled Dot-Product attention of the transformer model is defined as the alignment scoring (via dot-product) on a series of keys (K) by a series of queries (Q), followed by softmax and application of the resulting weights on a series of values (V) to compute the output context representation. These keys, queries, and values are learnable linear transformations of the input vectors X.

Instead of computing a single self-attention of a sequence, the authors suggest multi-head attention with 8 attention layers. This gives freedom to each attention head to perform an independent computation on the sequence. The result is concatenated and projected with a FFN as the output of the multi-head attention layer.

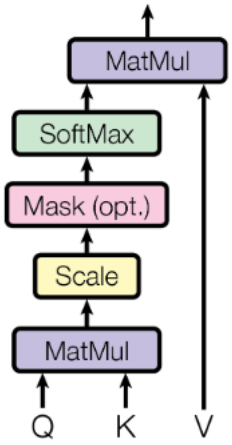


Figure 2.7. *Scaled Dot-Product Self Attention*

The original transformer architecture [5] follows a similar approach to the encoder-decoder design of a sequence-to-sequence RNN. However, other researchers experimented with decoder-only and encoder-only models. Decoder models are better suited for text generation tasks, while encoders are more suitable for classification tasks.

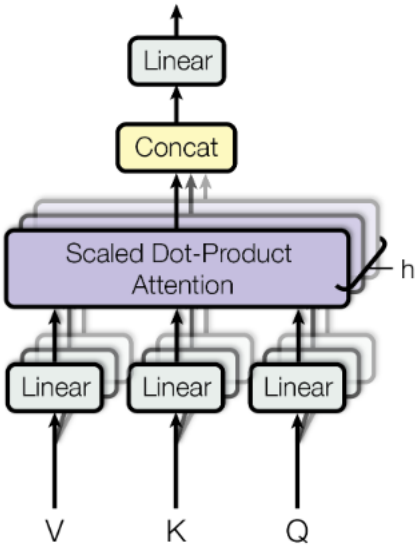


Figure 2.8. *Multi-Head Self Attention*

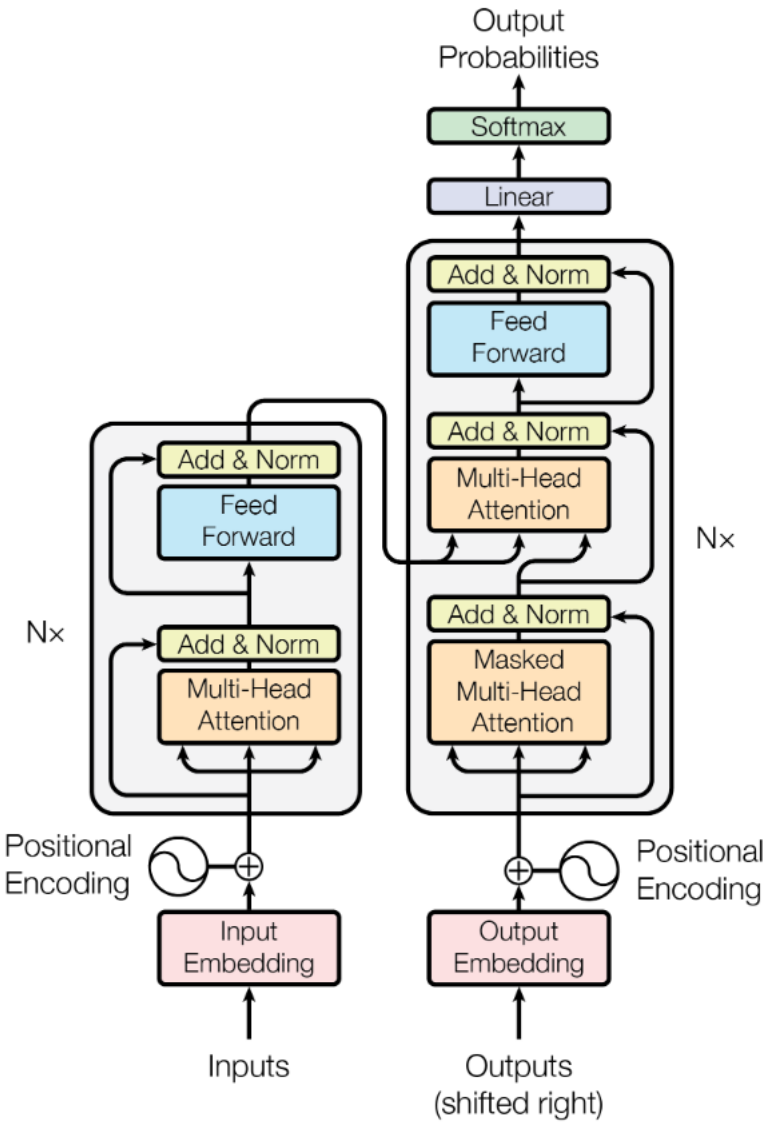


Figure 2.9. *Architecture of the Transformer introduced by Vaswani et al. in [5]*

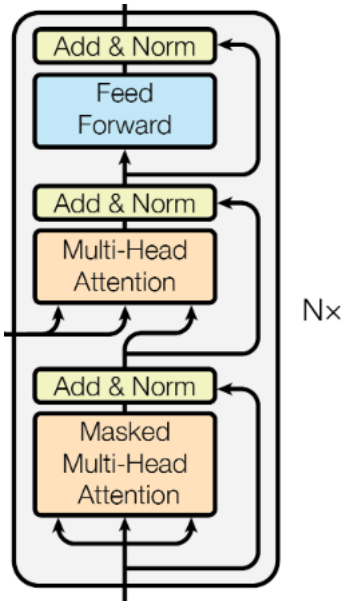


Figure 2.10. *Transformer Decoder Layer*

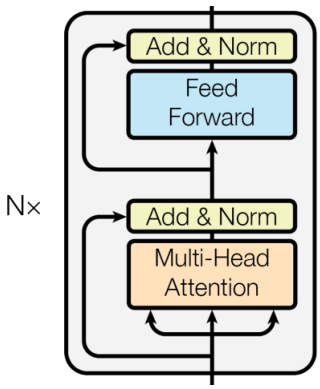


Figure 2.11. *Transformer Encoder Layer*

Chapter 3

Pretrained Language Models and Transfer Learning Methods

3.1 Transfer Learning

Domain: A domain D is defined by the pair of a feature space and a marginal distribution on this feature space. By feature space we mean the space of possible values for the features and by marginal distribution we simply refer to the distribution of dataset instances on this domain.

Task: Given a specific domain D, a task T is defined in an analogous way, as the pair of a label space and a predictive function f. Label space consists of the possible values for the labels and the predictive function is implicitly defined as the correct mapping from a sample x with features in D, to its label y.

We write: $f(x) = y$, where y is the correct label of x, if x is a sample in D. The goal of a learning algorithm is to create a model M that is a good approximation of f, meaning that $M(x) = f(x)$, for most instances of x in D.

Transfer Learning: Let T_s be a source task defined on a source domain D_s and a similar but not identical target task T_t, possibly defined on a different, but related, domain D_t. Also, let M be a model trained on T_s, meaning that we expect M to be a good approximation of f_s. We expect that the use of M as an initial approximation of f_t will help our learning algorithm to learn f_t. In other words, knowledge from the source task helps improve learning on the target task. This is the essence of the concept of transfer learning. For more a more formal definition of transfer learning, we refer to [31].

3.2 Pretrained Language Models

3.2.1 Large-scale Language Models

Advancements in computing power and the need to constantly exceed previous performance brought explosive growth to the depth of deep learning models. This trend was soon adopted for language models and especially the transformer model. However, growth in trainable parameters comes at a cost: these models required a large amount of data to converge to useful solutions. Datasets for supervised Natural Language Processing (NLP) tasks could not keep up with the requirements of these data-hungry models. To solve these

issues, the paradigm of pretraining was proposed. Huge unlabeled corpora were created semi-automatically from the unlimited resources of the internet. These datasets of raw text were leveraged with a variety of self-supervised objectives that were based on hiding information from the model and asking for predictions from the given context. The first and most popular models that adopted these techniques are GPT [6] and BERT [7].

3.2.2 GPT

GPT [6] is an autoregressive decoder transformer model that generates text. For the pretraining process, a simple unsupervised auxiliary task is used: predict the next token of the input sequence. Autoregressive means that during inference the prediction is fed as input to the model to predict the next output, until it decides to stop. In order to ensure that the model can execute the training process in parallel, the authors employed masked attention which blocks the view of the future tokens that the model should predict. This way training costs were reduced drastically.

3.2.3 BERT

Another landmark work that shaped research on PLMs was BERT [7]. BERT and GPT are in a way complementary. While GPT is best suited for autoregressive generation, BERT, which is an encoder model, was originally designed to perform classification and regression tasks. The first key feature that allows BERT to excel is that it is bidirectional. In this way, it avoids the use of masked-attention and allows the model to access the context of the whole input sentence. The drawback of this approach is that it cannot use the simple next-token prediction pretraining task. For this reason, the authors invented two novel pretraining auxilliary tasks. One of them is called Masked Language Modeling (MLM) and the other one Next Sentence Prediction (NSP).

3.3 Fine Tuning

3.3.1 Early Methods

Hinton and Salakhutdinov [32] were the first to pretrain a model (an autoencoder model, we refer to [30] for more details) on one task and then save the weights of this model to initialize another model. Inspired by this approach, [33] applied a similar technique in the field of computer vision. They trained a generic unsupervised CNN in layer-wise stages, using the output from the previous layer at each stage. After the pretraining stages they updated the whole hierarchy all together, in the same fashion as modern fine-tuning. Finally, RCNN (Regions with CNN features) [34] optimised this method and established it for computer vision. They leveraged pretraining for an auxiliary task followed by domain-specific fine-tuning on object detection in order to balance the limited amount of training data for this task.

3.3.2 Fine Tuning Paradigm for Transformers

Soon enough, these early "transfer learning" methods were "transferred" to the field of Natural Language Processing (NLP). The first work to introduce this paradigm to NLP in large scale with enormous success was GPT [6], with the following formulation:

- A high capacity, deep transformer model is pretrained on a large corpus of text, in order to collect useful general knowledge in an unsupervised manner
- This pretrained model is then refined (fine-tuned) on a specific task (downstream task), using much fewer labeled examples

BERT [7] expanded this approach for bidirectional encoder models by introducing two novel pretraining approaches: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, some of the words of the input sentence are randomly replaced by a mask-token ([MASK]) and the model has to predict the original token by understanding the context of the sentence. In the same spirit, NSP tests the model by choosing either two sentences that were found the one after the other in the original text or they were at random positions and essentially unconnected. To distinguish the sentences in NSP, two special tokens are added: [CLS] and [SEP]. The first one, called classification token, is very important because it is often used as a semantic summary [18] for the representation of the whole sentence.

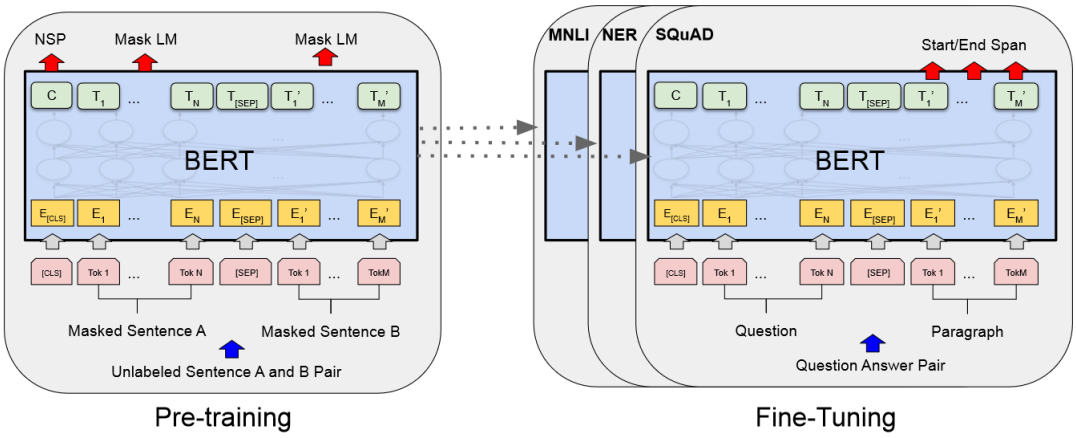


Figure 3.1. *BERT's fine-tuning strategy. A copy of the pretrained model is created for each specific task.*

3.3.3 Drawbacks of fine tuning

Fine-tuning consists of updating the pretrained weights, that contain general knowledge, to adapt them to a new specific task. The issue with this process is that the pretrained weights are not saved; they are overwritten. Clearly, if the original pretraining domain and the domain of the specific task are similar this is not a huge problem. But in some cases, such as multimodal learning, these two domains are completely distinct and the adaptation is not as smooth. In those cases, some of the useful knowledge that was stored in the

pretrained weights is lost in order to perform domain-adaptation. This is an unreasonable practice, as domain adaptation should be viewed as a completely separate process than pretraining and there is little connection between them. The only reasonable explanation to actually perform this practice is in some rare cases were the pretraining and the specific data come in similar quantities. In that case, maybe pretraining has inserted some biases and fine-tuning can be a way to eliminate them, however that is almost never the case with multimodal data.

Another important issue with fine-tuning is storage and training requirements. An extensible model [8] is a model that can be used as it is to solve a variety of tasks without forgetting previous ones. A fine-tuned model is not extensible as it is initially an exact copy of a pretrained model that has modified all of its weights slightly and so it has specialized in this and only this task. In addition, a lot of similar copies that serve more or less the same purpose are copied for each task. For all of these reasons, some alternative methods to fine-tuning have been proposed in the literature.

```
A "whatpu" is a small, furry animal native to Tanzania.  An example of a sentence that uses
the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

---

To do a "farduddle" means to jump up and down really fast.  An example of a sentence that uses
the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she
started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin.  An example of a sentence
that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden
there.  It was delicious.
```

Figure 3.2. *GPT3 performs few-shot learning with prompt tuning. Text in bold is generated by the model.*

3.4 Lightweight Tuning

3.4.1 Prompt tuning variations

Inspired by humans' ability to follow simple instructions, [35] instruct their model, called GPT-3, via text interaction, to perform a language task it has never seen before, without any gradient updates. These interactions, which are called "textual prompts", are designed by researchers, specifically for each task and they are prepended to the input. This method is often called "prompt tuning". Later, [36] proposed to search for prompts instead of designing them manually. [37] extended this idea, by learning prompt-embeddings with traditional backpropagation and showing that scaling is very beneficial for soft prompts. These "soft prompts" are not easily interpretable, but perform better and require less human intervention. [38] take the ideas of soft prompts yet another step further, by inserting a trainable prefix in the attention module of each layer of the transformer. Prefixes are very loosely related to prompts, but they achieve impressive results, displaying much greater flexibility than prompts, with an extremely low parameter budget. Their biggest

advantage is that they work equally well with encoder models, such as BERT. This method is called prefix tuning.

3.4.2 Adapters

The original adapters, called bottleneck adapters, were introduced by Houlsby et al. [8]. They consist of a linear down-projection, a non-linear activation and finally a linear up-projection to return to the original dimensions. The original input in this module is then added to the output to form the so called residual connection. Note that the weights of the adapter are initialized to have a very low norm. The residual connections together with the small norm initialization, ensure that they only intervene as much as they are needed to tweak the representations to be compatible with BERT.

This module is inserted in between BERT layers, in order to prepare the representation for the next layer. The original formulation from Houlsby suggested to insert adapter layers both after the attention layer and after the feedforward layer. However, Pfeiffer et al. [15] later proposed to insert an independent adapter module after the layernorm of the feedforward layer, which proved equally efficient with half additional parameters.

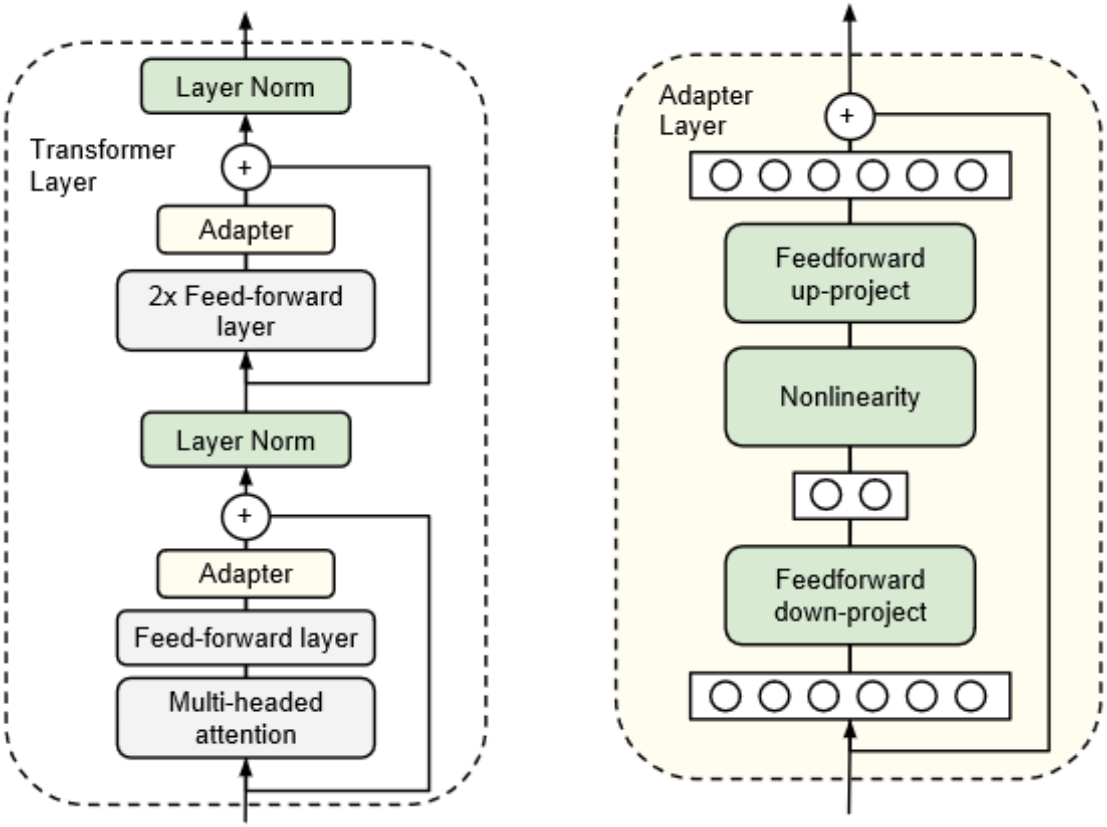


Figure 3.3. *Adapter Architecture*

The role of the adapters is to try to query the original BERT layer, essentially treating it as a blackbox algorithm. By separating the parameters each training stage has access to, in effect they are assigned a different task. AdapterFusion [15] first showed that this makes adapters much more than just a method for light tuning pretrained models. They are a

very useful tool that can be leveraged to avoid some of the dangers of naive fine tuning by providing the ability to disentangle training stages. Information learnt at a specific training stage is saved at a specific location for each layer and then it cannot change (we say it is frozen), thus mitigating the problems caused by catastrophic forgetting. In addition to that, it has proven to be a very data efficient process, partly because no knowledge is lost.

Chapter 4

Multimodal Learning

4.1 Introduction

4.1.1 What is a modality?

According to [39] a modality is: "The way in which something happens or is experienced." Jaimes et al. [40] clarify this as: "By modality we mean mode of communication according to human senses and computer input devices activated by humans or measuring human qualities." These two represent the human-centered approach, but there has also been proposed a machine perspective from [41] that is stated as: "A particular way or mechanism of encoding information." The final and most fulfilling approach is focused on the specific task. In their paper: "What is Multimodality?", Letitia et al. [42] propose that two streams of information belong to the same modality only and only if there exists a lossless 1-1 mapping between their domains in the preprocessing stage. These definitions suggest that the heterogeneous representations of modalities, also called modality gaps, is the bottleneck in a multimodal problem and our primary goal should be to bridge this gap and bring these representations closer together. When this happens, many techniques have been proposed in order to use these representations to draw accurate conclusions. This process is called multimodal fusion.

4.2 Multimodal Deep Learning Techniques

4.2.1 Learning Representations for Multimodal Fusion

There are many ways to produce successful multimodal representations, but in this section only deep neural methods are discussed. For different approaches we refer to [43]. According to Srivastava and Salakhhutidinov [44] a successful multimodal representation should possess three desirable properties: firstly, it should incorporate a kind of distance metric that reveals the amount of connection of the underlying concepts; secondly, it should be easy to obtain even in the absence of some modalities and finally, it should be possible to infer some elements of the missing modalities given the observed ones.

The goal of modern multimodal learning is to design a deep neural network that with its structure and some additional external regularization, it manages to project information from all the modalities to a common latent manifold that has the desired properties. The

hope is that in such a manifold, the heterogeneity between the modalities, also called the modality gap [9], will be bridged. In that case, the model can easily process this information and infer useful conclusions that lead to predictions. In other words, perform multimodal fusion.

4.2.2 Fusion Methods

In recent years, there have been numerous proposed methods for performing multimodal fusion effectively. For an in-depth review we suggest [43]. The focus of this study will be on deep learning methods, which are also the most successful up to date. It is important to note, that according to research [45], intermediate fusion applied to all of the layers of a deep neural network has great advantages because it each layer has its own properties. Specifically, the deeper the layer the more abstract the representation it processes. Two broad categories of fusion, introduced by [45] will be shortly reviewed:

- simple concatenation and feedforward network (FFN) fusion
- attention-based fusion

Concatenation and FFN

Concatenation is the most naive method of combining different features. The application of a feedforward network layer after concatenation is the most unstructured method of performing fusion. This is not necessarily a bad thing, as it means that total freedom is given to the model. In other words, no inductive bias is inserted in the architecture. If the designers have no access to such priors, or they have no clue how to establish them, then simple concatenation and FFN is the best and easiest solution they can hope for.

Attention-based

Attention mechanisms are widely used for fusion [45] [Mult, Flamingo, ViLBERT, Multimodal Intelligence]. They are usually preferred because they have a more intuitive interpretation compared to a simple feedforward network with no structure. The most popular variation of attention used for this purpose is symmetric cross-attention which is essentially standard attention with the keys and values set equal to one modality and the queries set equal to the other modality. Cross-attention allows each token of one modality to interact directly with each token from the other modality. This way, each token can ask questions and attend to specific features that are aligned with its own features. This alignment procedure has proven very efficient for multimodal fusion and this explains further why it is so popular.

4.3 Vision and Language

4.3.1 Visual BERT family

After the rapid success of BERT, many researchers [18, 46, 47, 48] implemented a multimodal version augmented with images. The general architecture is similar for all of these methods so only VilBERT [18] will be discussed here.

ViLBERT is composed of two parallel streams of transformer layers that interact in between each layer with a variation of the symmetric cross-attention fusion module, which the authors call co-attention. The first stream is a standard BERT that can process text and the other stream is a variation that can process visual features that correspond to image regions. The co-attention layers perform intermediate layer-wise fusion. ViLBERT also follows a complicated pretraining procedure in 3 stages, which is inspired by BERT's methods, that involve masking image-caption pairs. It is important to note that this procedure does not scale well if another modality is to be added.

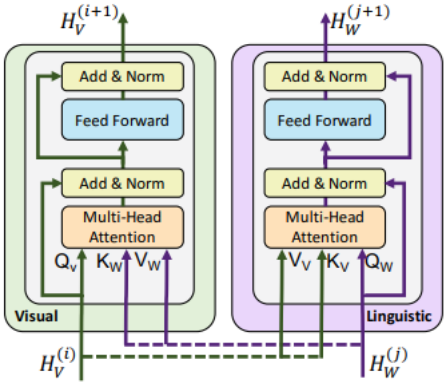


Figure 4.1. *ViLBERT architecture*

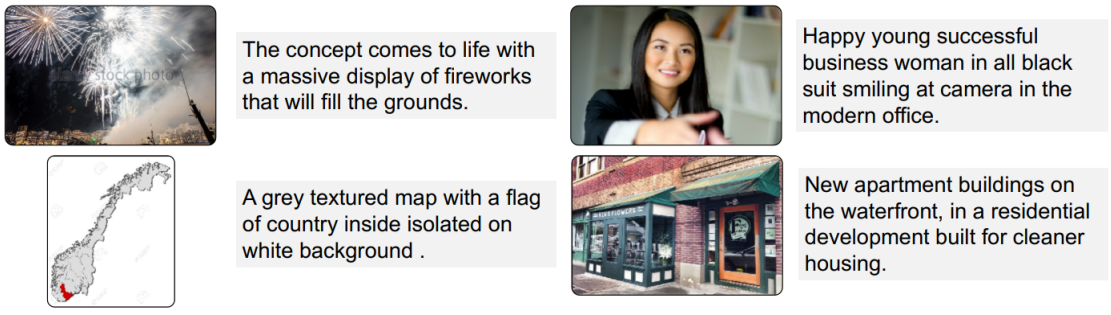


Figure 4.2. *A pretrained, but not fine-tuned, ViLBERT exhibits impressive performance*

4.3.2 Visual Language Models

Another important line of work on visual-language was initiated by a simple concept: a large-scale Pretrained Language Model (PLM) is augmented with visual information in

order to generate text in a rich multimodal context. Essentially, this method avoids to pretrain the model with expensive multimodal data and instead hopes that a powerful unimodal language model can be extended to understand visual information. In addition to that, fine-tuning is avoided as it causes catastrophic forgetting and alternative methods of lightweight tuning are leveraged [10] instead.

Frozen

Tsimboukelli et al. [16] proposed to augment a PLM with visual information, without changing its weights. The core concept was to translate image features into language-like embeddings which can be interpreted by the language transformer as "visual prompts" without the need to retrain the latter from scratch. A simple trainable visual encoder was responsible for preparing an input image to resemble language embeddings which are prepended to the input text of the frozen language model. This idea is similar to the light-tuning technique named "soft prompting", which, as described in chapter 3, trains a continuous embedding to interact with the language model. However, instead of training a soft prompt to effectively "ask questions" to the model, the visual soft prompt introduced by Frozen, provides the model with visual information.

The end goal of this model was to be able to complete the input sequence of text, in the context of an image, in the most reasonable way. To do a specific task the model was not tuned, i.e. no gradient updates were performed. It utilized the well-known property of the frozen PLM [35] to perform few-shot learning, revealing that it retained such abilities even in a visual context. A very important result of Frozen was that standard fine-tuning the language model in this scenario actually hurts performance because of the well studied catastrophic forgetting issues. The authors argued that much less paired image-caption data was available than the amount of text-only data used to pretrain the frozen model, so it was essential to keep as much knowledge from this training stage as possible. In addition, a very important attribute of this work was that it was very flexible because it is modular, meaning that it could use the best available language model off the shelf, just by re-training only the visual encoder with the visual-language pairs.

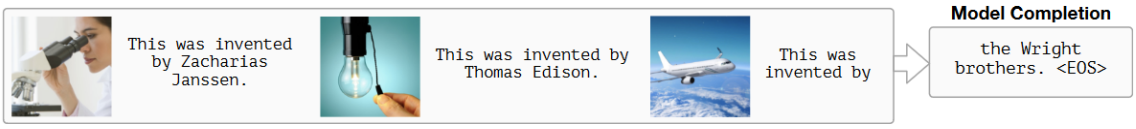


Figure 4.3. *Frozen understands images interleaved with text.*

MAGMA

MAGMA [10] extended this approach by adding adapter layers [8] in between the frozen layers. This proved empirically to boost performance as it outperformed Frozen by a significant margin. Adapter layers essentially allowed the PLM to adapt to a multimodal domain without losing any useful knowledge acquired in the pretraining stage. This is a key result that will turn out to be the foundation for the approach of this thesis to solve a different but related multimodal problem.

Figure 4.4. *MAGMA shows visual question-answering abilities*

Flamingo

Flamingo [17] scaled up and optimised this concept by introducing large scale multimodal pretraining and foundational innovations. They collected and curated a massive image-text dataset from the internet aiming to achieve higher quality multimodal few-shot learning abilities. The first innovation was a flexible visual encoder which can turn arbitrary sequences of images or even video frames to a fixed number of visual prompts. Moreover, instead of bottleneck adapters [8] they preferred cross-attention adapters that take fusion tokens and text tokens as input and output text token updates that are added with a residual gated connection to the previous text tokens, a technique known as shifting [49].

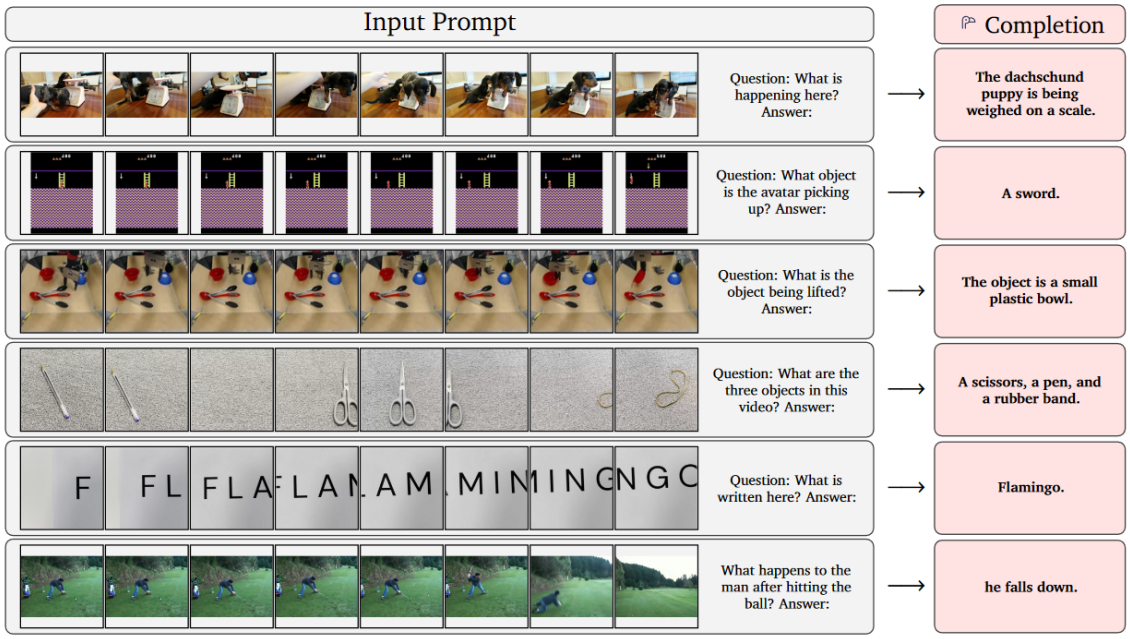


Figure 4.5. *Flamingo example question answering with video frames*

Chapter 5

Multimodal Sentiment Analysis

5.1 Introduction

One of the most distinguishable human traits is the feeling of empathy. Empathy leads to compassion and is crucial for the well-being of our society. Of course current AI systems are far from "feeling" empathy, but the illusion of empathy and sympathy is fascinating for users of AI applications because they encourage them to bond and feel "friendly" with the device. A great method to help achieve these traits is to evaluate the systems on sentiment analysis. Obviously, the first step in order to act in a compassionate manner as an AI system is to recognise the emotional state of the user. For these reasons, is is of immense importance to develop the field of sentiment analysis.

5.2 Datasets

There is a plethora of datasets for multimodal sentiment analysis (MSA) with text, visual and acoustic features as input. CMU-MOSI [50] contains 2199 opinion videos manually annotated with sentiment ranging from -3 to 3. ICT-MMMO [51], YouTube [52] and MOUD [53] all contain product reviews and opinion videos.

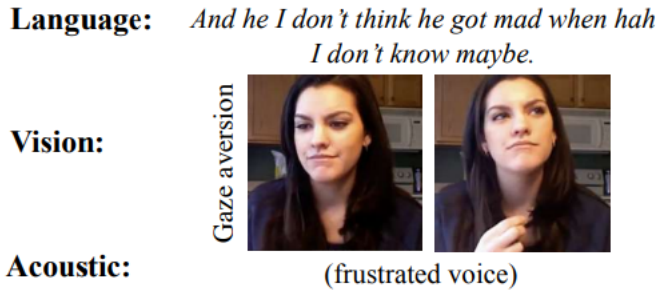


Figure 5.1. *An example from CMU-MOSEI*

CMU Multimodal Opinion Sentiment and Emotion Intensity or CMU-MOSEI [54] for short is a multimodal dataset which is the second and updated version of CMU-MOSI. The creators emphasized the importance of diversity in training samples by incorporating over 23,453 video segments with 250 topics and 1000 distinct speakers. Each sample consists of manual transcription aligned with acoustic and visual features. The videos are collected from YouTube and other online sources.

(a) *Word cloud of topics.*

Total number of sentences	23453
Total number of videos	3228
Total number of distinct speakers	1000
Total number of distinct topics	250
Average number of sentences in a video	7.3
Average length of sentences in seconds	7.28
Total number of words in sentences	447143
Total of unique words in sentences	23026
Total number of words appearing at least 10 times in the dataset	3413
Total number of words appearing at least 20 times in the dataset	1971
Total number of words appearing at least 50 times in the dataset	888

(b) *Summary of the statistics.*

Figure 5.2. *Word cloud and statistics for CMU-MOSEI*

Data collection for this dataset was done in an automatic way, but a curation process followed. Specifically, videos were analyzed to ensure that the camera was stationary and there was only one speaker. The diversity of the topics is remarkable, as there are about 250 of them, as can be seen in Fig. 5.2b. The most frequent topic is of course "reviews". Expert judges checked the dataset for mistakes. All videos have manual transcriptions and visual and acoustic features are aligned at word-level.

According to [12] for evaluation, although sentiment analysis is technically a regression task, both regression and accuracy metrics are leveraged. Specifically, Mean Absolute Error (MAE) and Pearson Correlation Coefficient are the preferred regression metrics. In order to use classification metrics, scores are split either in 7 classes to get 7-class Accuracy, or in 2 classes to get 2-class Accuracy and F1 score.

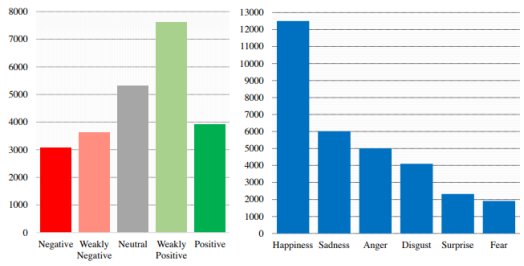


Figure 5.3. *Sentiment distribution shows a slight shift for positive sentiments*

5.3 Prior Work and State of the art

Early works on multimodal sentiment analysis used standard machine learning methods such as ensembles of Support Vector Machines [55]. Advanced neural network models such as Recurrent Neural Networks (RNNs) [56, 57, 58] Convolutional Neural Networks (CNNs) [59] and later Transformers surpassed them by taking advantage of context in the input sequences. Attention was used heavily as one of the most important tool for creating sophisticated modules that interact with different modalities. The variations that were created with this in mind are numerous: attention for fusion [11], cross-attention

transformer pairs for alignment [12], cross-attention for shifting [13], attention for creating multimodal prefix [10, 16], even attention for creating masks [14].
The following are the most successful models in MSA:

Mult [12] translates one modality to another using 3 pairs of cross-attention transformers that perform multimodal fusion. Impressively, this method of fusion allows it to align the input sequences even if they are not aligned in the preprocessing stage. Additionaly, an extensive study is performed showing attention activations visually. It uses GloVe embeddings [60] as language features and this hinders its performance.

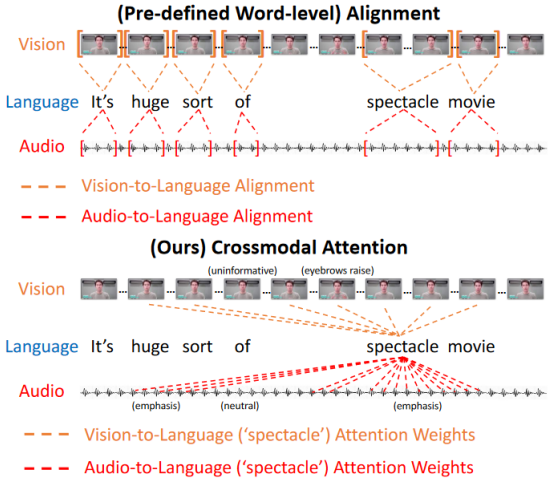


Figure 5.4. *MulT aligns modalities*

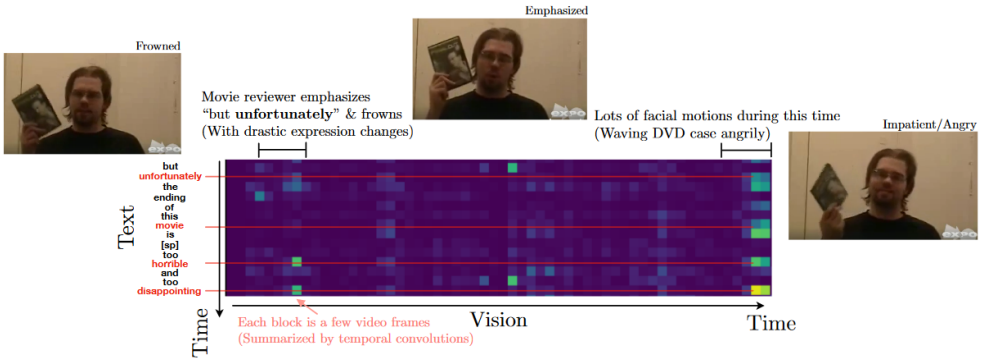


Figure 5.5. *MulT's attention activations*

MMLatch [14] uses high level representations to mask low level input features in a top-down manner. It is an innovative idea loosely inspired by congitive models of human perception. A feedback mechanism is used to capture top-down cross-modal interactions and update the input features. They achieve the best results among models which use GloVe embeddings.

MAG-BERT [13] presents a simple way to perform shifting of BERT representations, by adding a single multimodal adaptation gate to only one of BERT's layers.

In **ICCN** [22] audio-visual features are first extracted and then fused with text embeddings to get two outer-products, text-acoustic and text-visual. A Canonical Correlation

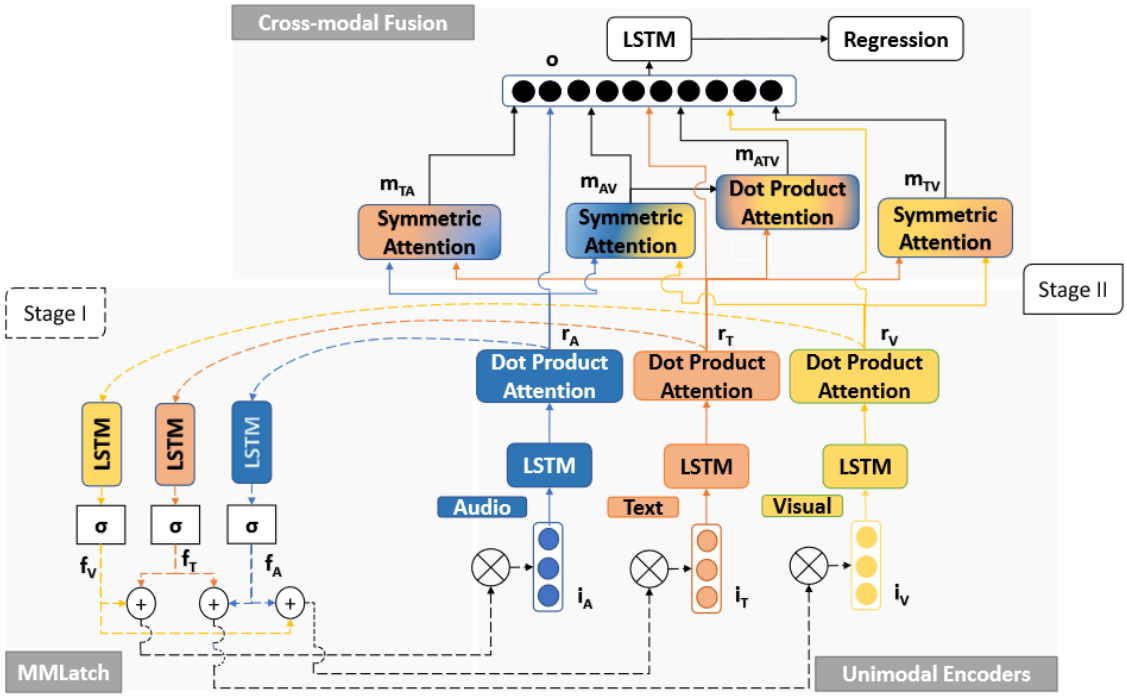


Figure 5.6. *MMLatch architecture*

Analysis (CCA) network takes these features fuses them to create a meaningful representation for prediction.For text embeddings, frozen BERT output is used as text features.

More recently, many researchers turned their efforts towards intricate multimodal pretraining strategies, such as [61, 62]. Such methods are model-agnostic and should be studied separately for a fair comparison.

MISA [9] is the current state-of-the-art in MSA. Instead of developing another intricate fusion approach, the authors emphasized the importance of building a reliable representation learning manifold. In this way, even a simple technique for fusion will be very effective. Specifically, they projected the input modalities in 4 subspaces: one modality-invariant and three modality-specific for each modality. To achieve this, they leveraged multiple loss functions to impose constraints for regularization. Namely, they used the original loss function of the task combined with a similarity loss, a differential loss and a reconstruction loss. Similarity is used in order to create the modality-invariant subspace and differential is used to ensure the modality-specific representation. The reconstruction loss is only there to avoid trivial solutions. Although it requires hand-crafting all of these losses, the rest of their architecture is quite simple and it surpasses the previous models by a large margin.

5.4 Robustness in MSA

Robustness is crucial for practical applications where unexpected errors are common. Data imbalances are often the cause of such errors and for this reason a clean and curated dataset is the foundation of a successful model. It has been observed [9, 12] that even

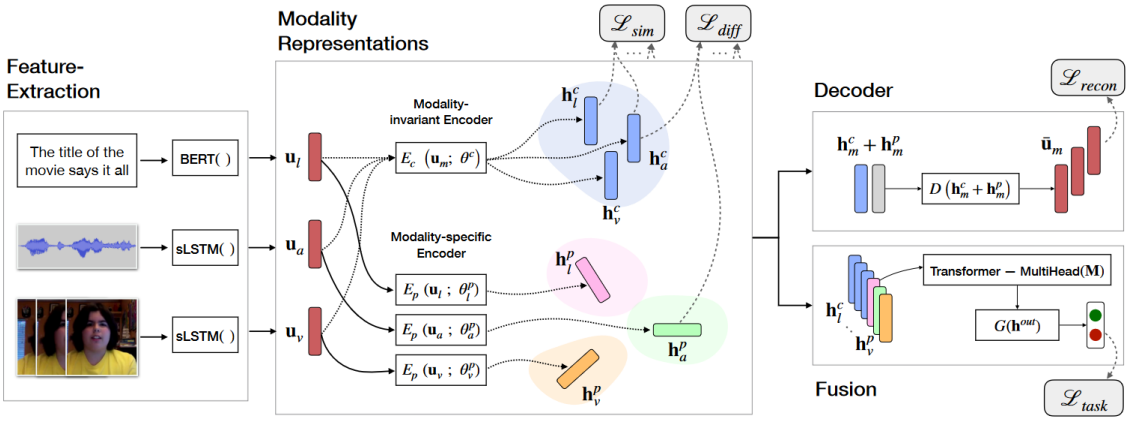


Figure 5.7. *MISA Architecture*

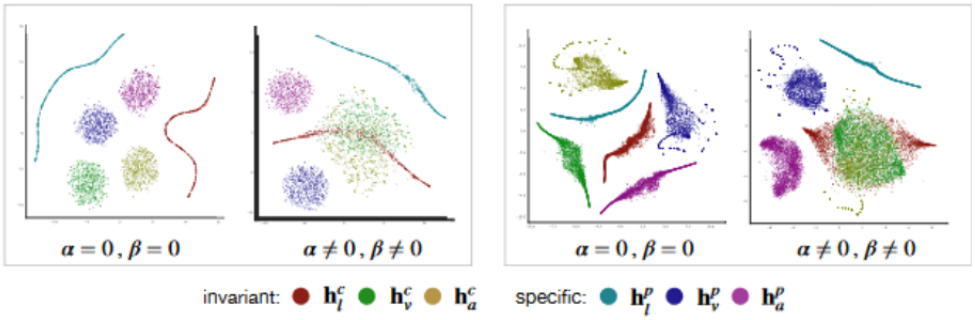


Figure 5.8. *MISA creates representation learning manifolds with modality-invariant and modality-specific subspaces*

for the highest performing models in MSA datasets language is by far dominating the other modalities. The same researchers performed simple ablation studies of removing one modality and letting the model use the rest of them only to find out that their models almost solely relied on language to make predictions. In order to study this phenomenon further Hazarika et al. [63] performed an extensive robustness study on state-of-the-art models for MSA.

Their method constitutes of two diagnostic checks at test time: removing a percentage of the modalities and inserting gaussian noise to a percentage of the modalities. The tests showed that even the state-of-the-art MISA is not as robust as expected to these tests. As a solution to the issues they proposed a training method that inserts noise in the training stage to better prepare the models for the noisy diagnostic tests. It is notable that with their method performance does not drop for the standard evaluation tests and at the same time perfomance on the diagnostic robustness tests rises significantly.

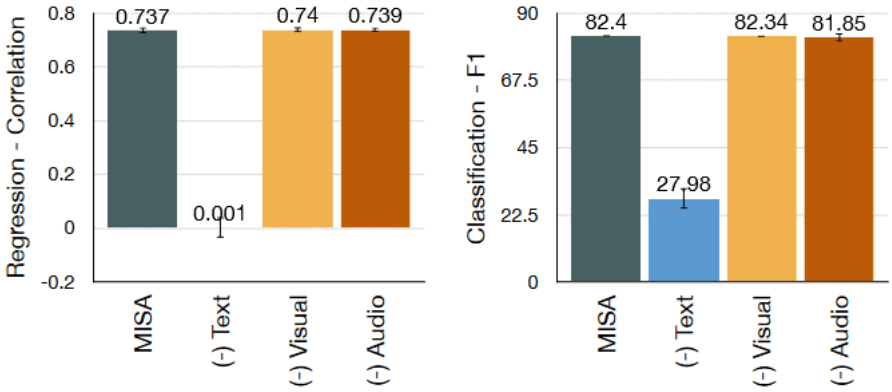


Figure 5.9. *The effect of missing a modality for MISA. Language modality dominates.*

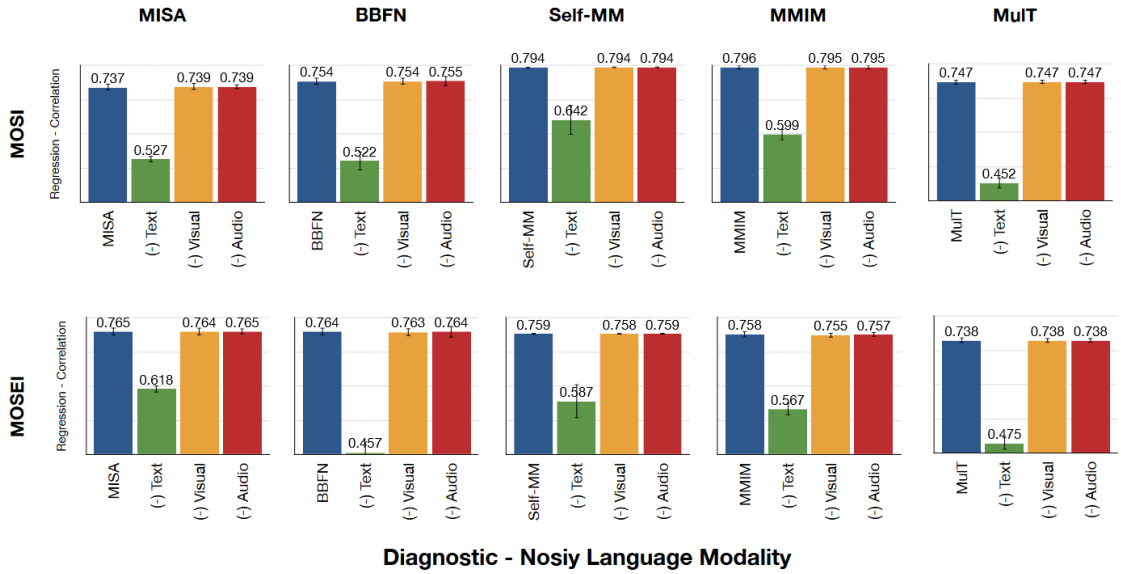


Figure 5.10. *Noise insertion as a diagnostic check for robustness.*

Chapter 6

Adapted Multimodal BERT (AMB)

6.1 Introduction

As discussed in the two previous chapters, there are some useful insights that prior work has established for multimodal learning that can be applied to MSA. First of all, the use of BERT seems to be a prerequisite for top-notch performance, as language has proven to be the most dominant modality. In addition, visual-language models have presented interesting novel ideas in the effective augmenting of PLMs, such as BERT, with the visual modality. It is obvious that such methods can be modified to include the acoustic modality as well, however the way in which to achieve this is not so obvious. Of course, methods such as ViLBERT[18] are not easily accessible because they do not scale well.

For these reasons, in this method we are inspired from visual-language models [16, 10, 17] to present an adapted BERT model for multimodal sentiment analysis. Our model is called Adapted Multimodal BERT (AMB). It leverages a novel but simple layer-wise feed-forward network interleaved with adapter and BERT layers in order to perform multimodal fusion. The details of our method's architecture are explained thoroughly, as well as the experimental procedure that leads to state-of-the-art results and important conclusions on the effect of adapters and the insertion of noise.

6.2 Architecture

6.2.1 Overview

Fig. 6.1 illustrates an overview of the system architecture. First of all, the input text sequence is fed to the frozen BERT preprocessing tokenizer to output a sequence of tokens. At the same time, the visual and audio sequences pass through their own trainable encoders in order to be translated to a single audio-visual token that is compatible with BERT representations. The core component is a frozen pretrained BERT model, which is tuned by adapter layers, without access to any other modalities. These BERT representations are combined with audio-visual information in a feedforward network (FFN) in order to perform layer-wise multimodal fusion. This process is repeated for 12 layers and the last representations are provided to a FFN to predict the sentiment score.

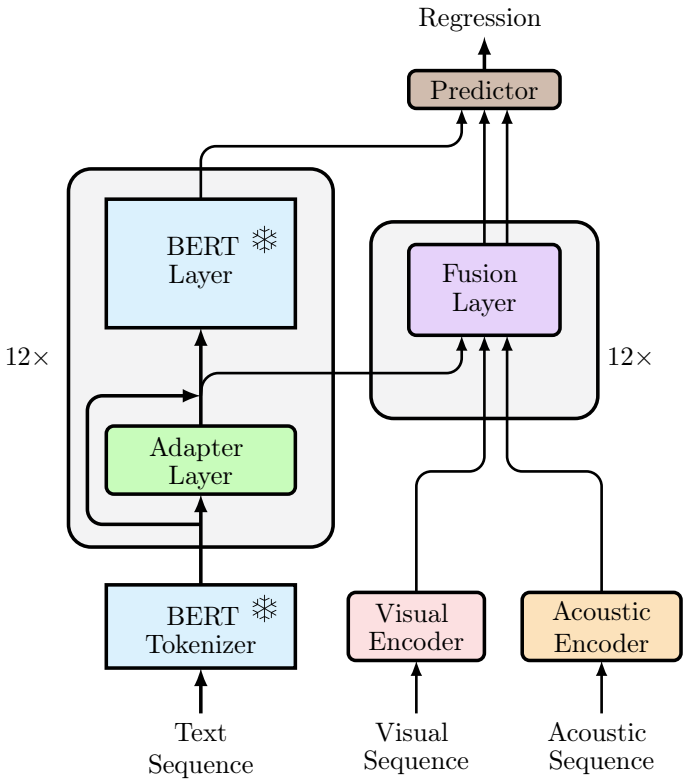


Figure 6.1. *Architecture of Adapted Multimodal BERT (AMB)*

6.2.2 Frozen BERT layers

The frozen BERT model is at the core of the architecture to emphasize the importance of language. Both BERT tokenizer and the 12 BERT layers are kept intact during training, limiting the effects of catastrophic forgetting that can incur during fine-tuning.

6.2.3 Adapter Layers

We use the original bottleneck adapters, introduced by Houlsby et al. [8]. Each adapter layer is composed of a linear down-projection followed by a ReLU non-linearity and then a linear up-projection to restore the original input dimensions. Residual connections are used between the input and output of each adapter layer. Instead of inserting an adapter layer both between the attention and the feedforward module, we follow [15] and only insert them after the feedforward layernorm layers, thus cutting the number of additional parameters in half. Our adapter layers are only responsible for adapting to the textual inputs.

6.2.4 Visual and Audio Encoders

Visual and audio encoders consist of transformer encoder layers that act on each modality separately to extract information from an arbitrary sequence of features and compress it in a concatenated visual-acoustic token. This token is then prepared for the next stage of layer-wise multimodal fusion. Our encoders are closely related to the approach of [16, 10, 17], with the addition of audio.

6.2.5 Fusion Layers

For multimodal fusion FeedForward Network Fusion (FFN-Fusion) is used in a layer-wise manner, between each BERT layer. The first BERT token (known as CLS token), which is commonly used to store a semantic summary of BERT's hidden states [18], is projected to a lower dimension and then concatenated with the modality tokens produced by the visual and audio encoders. This tensor is then fed into FFN-Fusion to output the fused representations. Although [13] and [17] also perform layer-wise multimodal fusion, both use the result to shift BERT representations in order to generate output text. We adopt a simpler approach without shifting.

6.2.6 Predictor

The fused representation of the last BERT and fusion layers are concatenated and fed into a classification head, consisting of a single Feedforward layer. Minimum Absolute Error loss is used for end-to-end training of the network.

6.3 Experiments

6.3.1 Setup

Data: The proposed model is evaluated for sentiment analysis on CMU-MOSEI [54]. It contains 23, 454 YouTube video clips of reviews on movies or other topics, where each sample is manually annotated with a sentiment score, ranging from -3 (strongly negative) to $+3$ (strongly positive). Text transcriptions are segmented into words, while visual FACET and acoustic COVAREP features are collected and aligned on these words. Standard train, development and test splits are provided. For evaluation, mean absolute error (MAE) and Pearson Correlation (Corr) between model and human predictions are used for regression, while seven-class accuracy (Acc-7), binary accuracy (Acc-2) and F1-score (F1) are used for classification.

Implementation Details: The `bert-base-uncased` version of BERT [7] is used for all experiments. It contains 12 transformer layers, where each token of the sentence has hidden size of 768 dimensions. The tokens are prepared for BERT with the standard tokenization procedure, while the two special tokens, [CLS] and [SEP], are added at the start and in the end of each sentence respectively. The encoders used for visual and acoustic modalities are randomly initialized transformer encoder modules with 2 layers and 1 attention head. We find that prepending a learnable [CLS] token and collecting this as a semantic summary works best. After a short hyper-parameter search in the range [128, 768] for the hidden size of the adapter layers, 384 is chosen as the optimal value. Similarly, for fusion layers 220 is chosen from [160, 820] as the hidden size.

For optimization, the Adam optimizer [64] is used with learning rate $5 * 10^{-5}$. Early stopping is used with patience set to 10 epochs and dropout is set to 0.2. Training takes 20 minutes on a single GTX 1080Ti NVIDIA GPU.

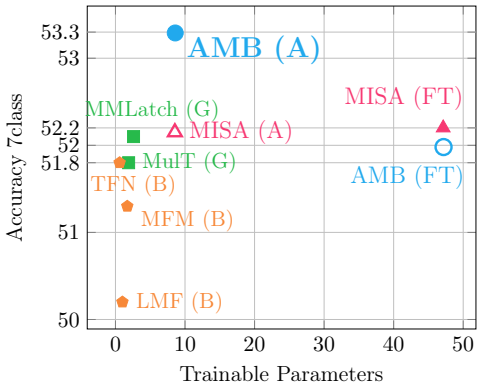


Figure 6.2. *7-class accuracy with respect to number of trainable parameters for the best performing models in the literature. G stands for GloVe embeddings, A for adapters, B for frozen and FT for fine-tuned BERT embeddings. The proposed AMB with adapters achieves a good balance between trainable parameters and performance.*

6.3.2 Results

The results for multimodal sentiment analysis on CMU-MOSEI are presented in Table 6.1. For fair comparison we only compare with methods in the literature that train in one stage, without leveraging their own, separate, pretraining stage on multimodal data. We observe that the proposed model outperforms all other methods by a significant margin across all metrics. As shown in Fig. 6.2, models that utilize Glove embeddings (G) [60], or frozen BERT embeddings (B), have fewer trainable parameters, sacrificing overall performance. Models that rely on fine-tuning of BERT have a significantly larger amount of trainable parameters. AMB with adapters surpasses fine-tuning based approaches on a small parameter budget.

Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	Trainable Parameters
MMLatch (G) [14]	0.582	0.704	52.1	82.8	82.9	2.6
MulT (G) [12]	0.580	0.703	51.8	82.5	82.3	1.8
LMF (B) [19]	0.623	0.677	50.2	82.0	82.1	1.0
TFN (B) [20]	0.593	0.700	51.8	82.5	82.3	0.6
MFM (B) [21]	0.568	0.717	51.3	84.4	84.3	1.7
ICCN (B) [22]	0.565	0.713	51.6	84.2	84.2	–
MAG-BERT* (FT) [13]	0.614	0.763	50.9	84.3	84.2	110.8
MISA (FT) [9]	0.555	0.756	52.2	85.3	85.3	47.1
AMB (Ours)	**0.536**	**0.766**	**53.3**	**85.8**	**85.8**	8.6

Table 6.1. *Results on CMU-MOSEI. Models indicated with (G) use glove embeddings. Models indicated with (B) use frozen BERT embeddings, and are taken from [22]. MISA and MAG-BERT use a fine-tuned (FT) BERT for feature extraction from language. MAG-BERT* is reproduced for CMU-MOSEI by the authors of this thesis. Trainable parameter are in millions.*

In our opinion, the interpretation of these results is clear. Methods that leverage a frozen BERT and the ones that use GloVe embeddings for feature extraction perform adequately, with very high parameter-efficiency. On the other hand, MISA fine-tunes

BERT to increase performance considerably, but it increases trainable parameters by a large margin in exchange. Finally, the proposed AMB with adapters manages to surpass the performance of MISA, by leveraging intermediate BERT representations instead of only the last layer BERT features. At the same time, adapters achieve a great parameter discount, making AMB competitive to frozen BERT and GloVe embeddings at parameter-efficiency. The following section aims to shed light to the comparison of adapters and fine-tuning even further.

6.3.3 Ablation Study

Table 6.2 shows an ablation study on the effect of the exclusion of modalities and the effect of using adapters versus finetuning for the adaptation of the language model. Firstly, the exclusion of the textual modality significantly degrades performance for the "AMB no-text" model, which demonstrates that text is the dominant modality for this task. With the exclusion of audio-visual information in "AMB text-only" performance still declines, though to a lesser degree, indicating that the use of multimodal information is beneficial.

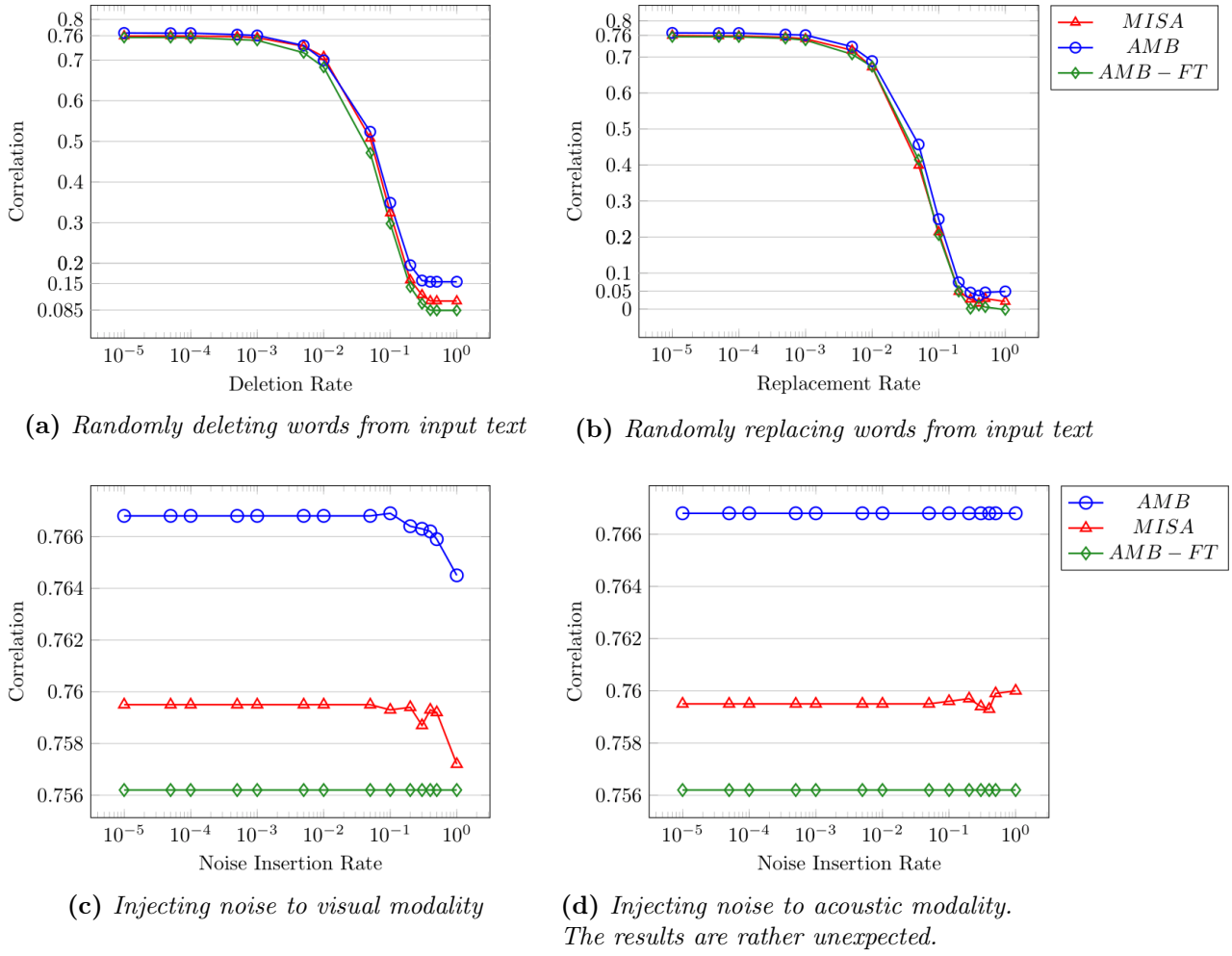
For the adapters versus fine-tuning experiments, an adapter based version of MISA ("MISA-Adapters") and a fine-tuned version of AMB ("AMB-FT") are implemented. We observe that fine-tuning is either unnecessary as in the case of MISA or even decreases model performance as in the case of AMB, revealing that some catastrophic forgetting occurs when performing fine-tuning on the text modality in this multimodal setting.

Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	Trainable Parameters
AMB no-text	0.816	0.240	41.6	63.3	61.8	8.6
AMB text-only	0.541	0.760	52.8	85.7	85.7	8.6
MISA-Adapters	0.5480	0.758	52.1	85.8	85.8	8.5
MISA	0.555	0.756	52.2	85.3	85.3	47.1
AMB-FT	0.548	0.756	51.9	85.4	85.3	47.2
AMB	**0.536**	**0.766**	**53.3**	**85.8**	**85.8**	8.6

Table 6.2. *Multimodal adapters vs fine-tuning. We include experiments, where the text, or the audio-visual modalities are missing. Trainable parameters are in millions.*

6.3.4 Robustness Study

In this section the robustness of AMB with respect to noise insertion is evaluated. For visual and acoustic robustness tests the work of Hazarika et al. [23] is followed. They propose the insertion of multiplicative Gaussian noise to a randomly selected set of input sequence elements for a given modality. For the text modality a different approach is employed that more closely simulates real-world errors, i.e. deleting and replacing input tokens. In the token replacement experiment a percentage of input tokens is selected randomly and replaced with random tokens from the vocabulary, while for the token deletion experiment they are instead replaced with the [UNK] token. The best checkpoint of each model is selected and the average correlation over three independent runs is reported, following [23].



(a) *Randomly deleting words from input text*

(b) *Randomly replacing words from input text*

(c) *Injecting noise to visual modality*

(d) *Injecting noise to acoustic modality. The results are rather unexpected.*

Figure 6.3. *Model robustness for varying levels of noise. Blue ∘: AMB, Red △: MISA, Green ◇: AMB-FT*

Fig. 6.3 displays the results of the robustness tests for varying levels of input noise. The deletion, replacement and noise insertion rate refer to the probability of corrupting each element in the input sequence. When corrupting textual inputs by deleting or replacing tokens we observe that performance starts to degrade after corrupting each token with 5% probability. Steeper performance degradation occurs in the case of replacement than in the case of deletion. This sensitivity to noise is expected, as text is the dominant modality. We observe similar robustness characteristics for AMB, MISA and AMB-FT, though adapter-based AMB appears to be somewhat more robust than its fine-tuned counterpart. In the extreme case from 50% probability and beyond AMB's lowest point is significantly higher than the rest, verifying that it considers all modalities to make predictions. In the case of noise injection to the visual modality performance drops off for AMB and MISA at 10% noise insertion rate. We observe that noise insertion in the visual modality affects both models less than noise insertion in text. Interestingly, the AMB-FT model is not affected by visual noise, revealing that this model relies completely on text, ignoring visual cues. These results highlight that, favoring adapter-based approaches over fine-tuning when using large pretrained language models for multimodal tasks may lead to improved model robustness

and better utilization of information from less dominant modalities (that contribute less to overall performance).

The results on acoustic modality were expected to be similar to the visual ones but they are qualitatively different. It is revealed that neither method seems to benefit from a clean acoustic modality. Not only that, but MISA surprisingly performs even better with a completely noisy stream than a clean acoustic stream of information. These results are in our opinion inconclusive and further investigation should be performed to get a clear insight. However, it seems that the acoustic modality of the dataset is inadequate to keep up with the others indicating a large imbalance and also MISA seems to interact with the acoustic modality although it is proven destructive, probably due to the additional regularization constraints that MISA uses.

Chapter 7

Conclusions

7.1 Discussion

In this work, AMB is proposed, a simple yet innovative model that builds on a powerful pretrained BERT transformer encoder and avoids the pitfalls of standard fine-tuning approaches for transfer learning. The use of adapters allows our model to lower the cost of trainable parameters without sacrificing performance, as it achieves new state-of-the-art on CMU-MOSEI. The most effective previous method, MISA [9], leverages fine-tuning which leads to five times more trainable parameters than our model. Other methods either use a frozen BERT for feature extraction or GloVe embeddings leading to parameter-efficient models that sacrifice performance. With this in mind, it is clear to us that adapters provide an excellent performance-parameter trade-off.

The ablation study showed that fine-tuning our model actually hinders its performance. In addition, it was proven empirically that useful knowledge from pretraining and non-dominant modalities is only leveraged effectively in the adapter-version. It seems that modality imbalances challenged the effectiveness of updating the weights of the model accurately. At the same time, a version of the previous state-of-the-art, MISA, that uses adapters achieves similar performance to the original that leverages fine-tuning, revealing that the same trends are followed by other models, although our architecture was optimised for this purpose. All of these results are a strong indication that fine-tuning suffers from the undesirable effects of catastrophic forgetting and adapter-based methods should be preferred instead.

Finally, our robustness study showed that the improvements introduced lead to reliable models that display robustness to various types of noise, a crucial trait for deploying applications in the wild. Although language modality is the foundation of our model, the study shows that a baseline performance is retained even with more than 50% deletion rate of input words by leveraging useful information from the other modalities. This is not observed for the fine-tuned version of our model and it is observed in a lesser degree for MISA. On the other hand, when visual features are pertrubed with noise, an expected drop in performance is observed, indicating the usefulness of these features for prediction. Surprisingly, this does not apply to acoustic features which is concerning for the quality of this stream of information. The latter result is in our opinion inconclusive and requires further investigation.

7.2 Future Work

Our experiments could be extended in many intriguing directions in the future in order to get clear insights on the degree of success of our proposed method. Firstly, it could be easily expanded in various ways in order to cement our results with further experimental evidence and at the same time explore the limits of this approach. One direction that our model could be expanded would be to incorporate more tasks, such as text generation from input prompts enriched with images in order to challenge AMB in a more demanding enviroment that requires generation. Also, the choice of the frozen PLM is not investigated enough in the literature, although it is the foundation of such methods. [13] experiment with BERT [7] and XL-net [65], while [16, 10] prefer GPT [6] for language generation tasks, but all of these works are incomparable and their results are inconclusive.

Moreover, exploring more sophisticated fusion methods compatible with our approach might be beneficial. For example, it was shown in chapter 5 that many successful multimodal approaches that require a similar intermediate fusion like AMB, they preferred a more complicated fusion method that leveraged cross-attention between modalities. It would be interesting to explore the effects of such a module on our design and our extensible architecture allows this change without severe modifications. The only part of the architecture that would have to change would be the feedforward network in the fusion layer. Ideally, an extensive ablation study on the effects of different modules in the fusion layer would shed light on the most appropriate method of fusing multimodal information in this scenario.

Another aspect to be considered is the effect of updating BERT representations. Our proposed approach extracts information from BERT to incorporate it in intermediate fusion modules, but BERT does not receive any feedback from the fusion process. In other words, BERT remains modality agnostic. As mentioned in chapters 4 and 5, there have been proposed approaches [49, 13, 17] that use feedback from the fusion layers in order to contextualize the next layer of BERT representations with modality information. This method is called shifting as feedback is used to modify slightly the multi-dimensional word hidden states with the hope of bringing them closer to the other modalities.

7.3 Ethical Considerations

Unfortunately, nowadays the costs of deep learning research are known to be prohibitive for low-budget institutions. It is our hope that our approach will be viewed as the blueprint for designing multimodal models based on pretrained unimodal encoders in a flexible and effective manner, that will be accesible to anyone independently of their budget. The readers should be already familiar by now with the fact that successful multimodal learning with AI systems requires massive amounts of cleaned, carefully curated data which are expensive to create manually and difficult to find reliably online in an automatic manner. To handle all of these data a deep learning system should use them as efficiently as possible without spending huge amounts of precious computational power, as standard methods are prone to do. In addition, if a system does not use all of the modalities to make decisions,

although it might exhibit excellent performance in the lab, it is vulnerable to the noisy real-life enviroment. It is clear for us that the proposed paradigm is a step forward in the right direction to achieve these goals and provide accessible AI research for everyone.

Bibliography

[1] *MLP and FFN*. https://www.electronicdesign.com/markets/automotive/article/21804976/whats-the-difference-between-machine-learning-techniques. Ημερομηνία πρόσβασης: 1-11-2022.

[2] Alex Krizhevsky, Ilya Sutskever και Geoffrey E Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. *Advances in Neural Information Processing Systems*F. Pereira, C.J. Burges, L. Bottou και K.Q. Weinberger, επιμελητές, τόμος 25. Curran Associates, Inc., 2012.

[3] *RNN Effectiveness*. http://karpathy.github.io/2015/05/21/rnn-effectiveness/. Ημερομηνία πρόσβασης: 1-11-2022.

[4] *RNNs Encoder-Decoder*. https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/. Ημερομηνία πρόσβασης: 1-11-2022.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser και Illia Polosukhin. *Attention is All you Need*. *Advances in Neural Information Processing Systems*I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan και R. Garnett, επιμελητές, τόμος 30. Curran Associates, Inc., 2017.

[6] Radfordet al. *Improving language understanding by generative pre-training*. 2018.

[7] Jacob Devlin, Ming Wei Chang, Kenton Lee και Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, σελίδες 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[8] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan και Sylvain Gelly. *Parameter-Efficient Transfer Learning for NLP*. *Proceedings of the 36th International Conference on Machine Learning*Kamalika Chaudhuri και Ruslan Salakhutdinov, επιμελητές, τόμος 97 στο *Proceedings of Machine Learning Research*, σελίδες 2790–2799. PMLR, 2019.

[9] Devamanyu Hazarika, Roger Zimmermann και Soujanya Poria. *MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, σελίδα 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery.

[10] Constantin Eichenberg, Sid Black, Samuel Weinbach, Letitia Parcalabescu και Anette Frank. *MAGMA - Multimodal Augmentation of Generative Models through Adapter-based Finetuning. ArXiv*, abs/2112.05253, 2021.

[11] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li και Ivan Marsic. *Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, σελίδες 2225–2235, Melbourne, Australia, 2018. Association for Computational Linguistics.

[12] Yao Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis Philippe Morency και Ruslan Salakhutdinov. *Multimodal Transformer for Unaligned Multimodal Language Sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, σελίδες 6558–6569, Florence, Italy, 2019. Association for Computational Linguistics.

[13] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis Philippe Morency και Ehsan Hoque. *Integrating Multimodal Information in Large Pretrained Transformers. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, σελίδες 2359–2369, Online, 2020. Association for Computational Linguistics.

[14] Georgios Paraskevopoulos, Efthymios Georgiou και Alexandros Potamianos. *Mmlatch: Bottom-Up Top-Down Fusion For Multimodal Sentiment Analysis. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, σελίδες 4573–4577. IEEE, 2022.

[15] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho και Iryna Gurevych. *AdapterFusion: Non-Destructive Task Composition for Transfer Learning. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, σελίδες 487–503, Online, 2021. Association for Computational Linguistics.

[16] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals και Felix Hill. *Multimodal Few-Shot Learning with Frozen Language Models. Advances in Neural Information Processing Systems*M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang και J. Wortman Vaughan, επιμελητές, τόμος 34, σελίδες 200–212. Curran Associates, Inc., 2021.

[17] Jean Baptiste Alayracet al. *Flamingo: a Visual Language Model for Few-Shot Learning.* 2022.

[18] Jiasen Lu, Dhruv Batra, Devi Parikh και Stefan Lee. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. Advances in Neural Information Processing Systems*H. Wallach, H. Larochelle, A. Beygelzimer, F.d' Alché-Buc, E. Fox και R. Garnett, επιμελητές, τόμος 32. Curran Associates, Inc., 2019.

[19] Z. Liu και others. *Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. Proc. 56th ACL*, σελίδες 2247–2256. ACL, 2018.

[20] A. Zadeh, M. Chen και others. *Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP*, 2017.

[21] Y.-H. H. Tsai, P. Liang και others. *Learning Factorized Multimodal Representations. ICLR*, 2019.

[22] Galen Andrew, Raman Arora, Jeff Bilmes και Karen Livescu. *Deep Canonical Correlation Analysis. Proceedings of the 30th International Conference on Machine Learning*Sanjoy Dasgupta και David McAllester, επιμελητές, τόμος 28 στο *Proceedings of Machine Learning Research*, σελίδες 1247–1255, Atlanta, Georgia, USA, 2013. PMLR.

[23] D. Hazarika, Y. Li και others. *Analyzing Modality Robustness in Multimodal Sentiment Analysis. NAACL*, σελίδες 685–696. ACL, 2022.

[24] *Digital Report.* `https://datareportal.com/reports/digital-2022-global-overview-report`. Ημερομηνία πρόσβασης: 1-11-2022.

[25] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Yue Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto και Joseph P. Turian. *Experience Grounds Language. EMNLP*, 2020.

[26] Tom M Mitchell. *Machine learning*, τόμος 1. McGraw-hill New York, 1997.

[27] *Lambda concioussness.* `https://theconversation.com/is-googles-lamda-conscious-a-philosophers-view-184987/`. Ημερομηνία πρόσβασης: 1-11-2022.

[28] Jonathan Frankle και Michael Carbin. *The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. International Conference on Learning Representations*, 2019.

[29] Ilya Sutskever, Oriol Vinyals και Quoc V. Le. *Sequence to Sequence Learning with Neural Networks. Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, σελίδα 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[30] Ian Goodfellow, Yoshua Bengio και Aaron Courville. *Deep Learning.* MIT Press, 2016. `http://www.deeplearningbook.org`.

[31] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong και Qing He. *A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE*, 109:43–76, 2021.

[32] G. E. Hinton και R. R. Salakhutdinov. *Reducing the Dimensionality of Data with Neural Networks. Science*, 313(5786):504–507, 2006.

[33] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala και Yann Lecun. *Pedestrian Detection with Unsupervised Multi-Stage Feature Learning. Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, σελίδα 3626–3633, USA, 2013. IEEE Computer Society.

[34] Ross Girshick, Jeff Donahue, Trevor Darrell και Jitendra Malik. *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[35] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever και Dario Amodei. *Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan και H. Lin, επιμελητές, τόμος 33, σελίδες 1877–1901. Curran Associates, Inc., 2020.

[36] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace και Sameer Singh. *AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, σελίδες 4222–4235, Online, 2020. Association for Computational Linguistics.

[37] Brian Lester, Rami Al-Rfou και Noah Constant. *The Power of Scale for Parameter-Efficient Prompt Tuning. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, σελίδες 3045–3059, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[38] Xiang Lisa Li και Percy Liang. *Prefix-Tuning: Optimizing Continuous Prompts for Generation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, σελίδες 4582–4597, Online, 2021. Association for Computational Linguistics.

[39] Tadas Baltrušaitis, Chaitanya Ahuja και Louis Philippe Morency. *Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[40] Alejandro Jaimes και Nicu Sebe. *Multimodal human–computer interaction: A survey. Computer Vision and Image Understanding*, 108(1):116–134, 2007.

[41] Wenzhong Guo, Jianwen Wang και Shiping Wang. *Deep Multimodal Representation Learning: A Survey*. *IEEE Access*, 7:63373–63394, 2019.

[42] Letitia Parcalabescu, Nils Trost και Anette Frank. *What is Multimodality? Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, σελίδες 1–10, Groningen, Netherlands (Online), 2021. Association for Computational Linguistics.

[43] Tadas Baltruvsaitis, Chaitanya Ahuja και Louis Philippe Morency. *Multimodal Machine Learning: A Survey and Taxonomy*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:423–443, 2019.

[44] Nitish Srivastava και Russ R Salakhutdinov. *Multimodal Learning with Deep Boltzmann Machines*. *Advances in Neural Information Processing Systems*F. Pereira, C.J. Burges, L. Bottou και K.Q. Weinberger, επιμελητές, τόμος 25. Curran Associates, Inc., 2012.

[45] Chao Zhang, Zichao Yang, Xiaodong He και Li Deng. *Multimodal Intelligence: Representation Learning, Information Fusion, and Applications*. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[46] Hao Tan και Mohit Bansal. *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, σελίδες 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics.

[47] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei και Jifeng Dai. *VL-BERT: Pre-training of Generic Visual-Linguistic Representations*. *International Conference on Learning Representations*, 2020.

[48] Liunian Harold Li, Mark Yatskar, Da Yin, Cho Jui Hsieh και Kai Wei Chang. *VisualBERT: A Simple and Performant Baseline for Vision and Language*. *ArXiv*, abs/1908.03557, 2019.

[49] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh και Louis Philippe Morency. *Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors*. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 33 1:7216–7223, 2019.

[50] Amir Zadeh, Rowan Zellers, Eli Pincus και Louis Philippe Morency. *Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages*. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[51] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae και Louis Philippe Morency. *YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context*. *IEEE Intelligent Systems*, 28(3):46–53, 2013.

[52] Louis Philippe Morency, Rada Mihalcea και Payal Doshi. *Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web*. *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, σελίδα 169–176, New York, NY, USA, 2011. Association for Computing Machinery.

[53] Anik Dey, Sebastian Krause, Ivelina Nikolova, Eva Vecchi, Steven Bethard, Preslav I. Nakov και Feiyu Xu, επιμελητές. *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[54] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria και Louis Philippe Morency. *Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph*. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, σελίδες 2236–2246, Melbourne, Australia, 2018. Association for Computational Linguistics.

[55] Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar και Rohit Prasad. *Ensemble of SVM trees for multimodal emotion recognition*. *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, σελίδες 1–4, 2012.

[56] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller και Shrikanth Narayanan. *Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification*. *IEEE Transactions on Affective Computing*, 3(2):184–198, 2012.

[57] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller και Gerhard Rigoll. *LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework*. *Image and Vision Computing*, 31(2):153–163, 2013. Affect Analysis In Continuous Input.

[58] Aman Shenoy και Ashish Sardana. *Multilogue-Net: A Context-Aware RNN for Multimodal Emotion Detection and Sentiment Analysis in Conversation*. *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, σελίδες 19–28, Seattle, USA, 2020. Association for Computational Linguistics.

[59] Soujanya Poria, Iti Chaturvedi, Erik Cambria και Amir Hussain. *Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis*. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, σελίδες 439–448, 2016.

[60] Jeffrey Pennington, Richard Socher και Christopher Manning. *GloVe: Global Vectors for Word Representation*. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, σελίδες 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics.

[61] W. Yu και others. *Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis*. *Proc.AAAI*. arXiv, 2021.

[62] J. Kim και J. Kim. *CMSBERT-CLR: Context-driven Modality Shifting BERT with Contrastive Learning for linguistic, visual, acoustic Representations*, 2022. arXiv:2209.07424.

[63] Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann και Soujanya Poria. *Analyzing Modality Robustness in Multimodal Sentiment Analysis. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, σελίδες 685–696, Seattle, United States, 2022. Association for Computational Linguistics.

[64] Diederik P. Kingma και Jimmy Ba. *Adam: A Method for Stochastic Optimization. CoRR*, abs/1412.6980, 2015.

[65] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov και Quoc V Le. *XLNet: Generalized Autoregressive Pretraining for Language Understanding. Advances in Neural Information Processing Systems*H. Wallach, H. Larochelle, A. Beygelzimer, F.d' Alché-Buc, E. Fox και R. Garnett, επιμελητές, τόμος 32. Curran Associates, Inc., 2019.

List of Abbreviations

AI	Artificial Intelligence
NLU	Natural Language Understanding
PLM	Pre-trained Language Model
MSA	Multimodal Sentiment Analysis
BERT	Bidirectional Encoder Representations from Transformers
AMB	Adapted Multimodal BERT
ICCN	Interaction Canonical Correlation Network
CMU-MOSEI	CMU Multimodal Opinion Sentiment and Emotion Intensity
MAE	Mean Absolute Error
MSA	Multimodal Sentiment Analysis
ML	Machine Learning
DL	Deep Learning
GPU	Graphics Processing Unit
CV	Computer Vision
NLP	Natural Language Processing
NMT	Neural Machine Translation
LM	Language Modelling
MLM	Masked Language Modelling
SVM	Support Vector Machine
MLP	Multi Layer Perceptron
FFN	Feed Forward Network
SGD	Stochastic Gradient Descent
ReLU	Rectified Linear Unit
CNN	Convolutional Neural Network
R-CNN	Region-based Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory network
ViLBERT	Vision and Language BERT
GPT	Generative Pre-trained Transformer