ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Cognitive methods for image captioning

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημήτριος Σωτηρίου

Επιβλέπων : Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Cognitive methods for image captioning

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημήτριος Σωτηρίου

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26η Ιουλίου 2021.

Αλέξανδρος Ποταμιάνος Αν. Καθηγητής Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.	Ανδρέας Σταφυλοπάτης Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

Δημήτριος Σωτηρίου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Δημήτριος Σωτηρίου, 2021.
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσοβίου Πολυτεχνείου.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή μου κύριο Αλέξανδρο Ποταμιάνο για την πολύ τιμητική για μένα εμπιστοσύνη που μου έδειξε, δίνοντάς μου την ευκαιρία να εκπονήσω τη διπλωματική μου εργασία στο Εργαστήριο Επεξεργασίας Φωνής και Φυσικής Γλώσσας, καθώς και για την συστηματική, ανελλιπή και πολύτιμη καθοδήγησή του. Είναι βεβαίως αυτονόητο ότι οποιεσδήποτε αδυναμίες στην παρούσα εργασία οφείλονται αποκλειστικά σε εμένα. Ευχαριστίες οφείλω επίσης στους Καθηγητές κυρίους Στέφανο Κόλλια και Ανδρέα Σταφυλοπάτη οι οποίοι ανταποκρίθηκαν με προθυμία στην παράκλησή μου για συμμετοχή στην τριμελή επιτροπή και με τίμησαν με την παρουσία τους. Εγκάρδια ευχαριστώ τους υποψήφιους διδάκτορες Γιώργο Παρασκευόπουλο και Ευθύμη Γεωργίου για την διαρκή υποστήριξη, τις συμβουλές τους και τον χρόνο που αφιέρωσαν στην επίλυση των αποριών μου σε όλη τη διάρκεια εκπόνησης της παρούσας εργασίας. Επίσης ευχαριστώ τους συναδέλφους και φίλους Νίκο και Νικήτα για τις εποικοδομητικές συζητήσεις και την ανταλλαγή απόψεων. Τέλος, ευχαριστώ την οικογένειά μου για τη στήριξη και συμπαράσταση σε όλη τη διάρκεια των σπουδών μου.

Δημήτριος Σωτηρίου
Ιούλιος 2021

National Technical University of Athens, School of Electrical and Computer Engineering
Division of Signals, Control and Robotics, Speech and Language Processing (SLP) laboratory.

Περίληψη

Γνωσιακές μέθοδοι για δημιουργία περιγραφών εικόνας

Παρόλο που το εγχείρημα της δημιουργίας λεζάντας σε μια εικόνα είναι δύσκολο για τους υπολογιστές, οι άνθρωποι μπορούν εύκολα να το φέρουν σε πέρας χάρη σε εγγενείς δυνατότητες του εγκεφάλου τους. Με βάση σχετικές έρευνες, συνάγεται ότι οι ενεργοποιήσεις του ανθρώπινου εγκεφάλου κωδικοποιούν σημασιολογικές πληροφορίες για το τι βλέπουμε και σκεπτόμαστε. Στο πεδίο της νευροεπιστήμης, πραγματοποιήθηκαν αρκετές μελέτες με στόχο την εξαγωγή πληροφοριών αυτού του τύπου από τις εγκεφαλικές ενεργοποιήσεις. Σε αυτή την εργασία, προτείνονται διάφορες τεχνικές ενσωμάτωσης των εγκεφαλικών ενεργοποιήσεων fMRI σε ένα μοντέλο δημιουργίας λεζάντας για εικόνα, που βασίζεται στην αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή των μετασχηματιστών. Ειδικότερα, εξετάζουμε την προσθήκη πληροφορίας fMRI στον κωδικοποιητή, την συνθηκοθέτηση του μηχανισμού προσοχής στον αποκωδικοποιητή και άλλες τεχνικές με χρήση ξεχωριστού κωδικοποιητή για τις εγκεφαλικές ενεργοποιήσεις. Επιπροσθέτως, διερευνούμε περισσότερο προσαρμοστικές εκδοχές των προαναφερθεισών τεχνικών ενσωμάτωσης, προκειμένου να διασφαλίσουμε την χρήση της αδύναμης τροπικότητας των εγκεφαλικών ενεργοποιήσεων και να επιτρέψουμε την αξιοποίησή τους μόνο στις περιπτώσεις όπου ενδέχεται να συνεισφέρουν σημαντικές πληροφορίες στο μοντέλο. Λόγω του γεγονότος ότι τα δεδομένα fMRI είναι περιορισμένα, εκτελείται με διαφορετικούς τρόπους ένα βήμα «λεξικής επέκτασης», όπου προβλέπονται οι εγκεφαλικές ενεργοποιήσεις για καινούρια οπτικά ερεθίσματα, τα οποία δεν χρησιμοποιήθηκαν κατά το πείραμα fMRI. Τα αποτελέσματα που προέκυψαν δείχνουν κατ' αρχάς ότι η κύρια διαδικασία αξιολόγησης που προτείνεται στη βιβλιογραφία δεν εγγυάται την ποιότητα της «λεξικής επέκτασης», καθώς μέσα από άλλες διαδικασίες αξιολόγησης φαίνεται ότι αυτή η αντιστοίχιση δεν είναι εύρωστη και δυνητικά εισάγει επιπρόσθετο θόρυβο στις προβλεφθείσες ενεργοποιήσεις. Δεύτερον, ότι το περιθώριο βελτίωσης με εγκεφαλικές ενεργοποιήσεις του μοντέλου δημιουργίας λεζάντας φαίνεται εδώ αρκετά περιορισμένο, καθώς σε όλα τα πειράματά μας παρατηρούνται μόνο μικρές αποκλίσεις σε σχέση με το βασικό μοντέλο. Αυτό υποδηλώνει ότι το μοντέλο δεν καταφέρνει να εξαγάγει χρήσιμη πληροφορία από την αδύναμη τροπικότητα των εγκεφαλικών ενεργοποιήσεων. Τέλος, συμπεραίνεται ότι απαιτείται περαιτέρω έρευνα, προκειμένου να δια-

πιστωθεί η αποτελεσματικότητα των εγκεφαλικών ενεργοποιήσεων στο εν λόγω πρόβλημα.

Λέξεις κλειδιά: μηχανική μάθηση, βαθιά μάθηση, νευρωνικά δίκτυα, μετασχηματιστές, γνωσιακή νευροεπιστήμη, λειτουργική μαγνητική τομογραφία, δημιουργία λεζάντας εικόνας

Abstract

Cognitive methods for image captioning

Even though image captioning is a difficult task for computers, humans can easily describe images through inherent capabilities of their brains with little effort. Recent research has shown that brain activations encode semantic information about what people see and think. In the domain of neuroscience, several studies have attempted to extract this information from brain activations. In this work, we propose several techniques of incorporating fMRI brain activations to an image captioning model that is based on the transformer encoder-decoder architecture. Specifically, we consider fusion at the encoder, attention conditioning on the decoder and other techniques with a separate transformer encoder for the brain activations. In addition, more adaptive variants of the aforementioned fusion techniques are explored in order to enforce the usage of the weak modality of brain activations or to enable the usage of the brain activations only when they are likely to contribute significant information to the model. Due to the fact that fMRI data are limited, a "lexical expansion" step is performed in various different ways, where brain activations are predicted for novel visual stimuli, that were not used in the fMRI experiment. Our results indicate that the quality of the "lexical expansion" is not guaranteed by the main evaluation process proposed in the literature, as other evaluation procedures indicate that this mapping is not very robust, potentially introducing additional noise to the predicted activations. Therefore, the scope for improvement of the model via brain activations seems to be quite limited and only minor deviations from the baseline are observed in all our experiments, suggesting that the model fails to extract meaningful information from the weak modality of brain activations. Finally, we conclude that additional research is needed in order to establish the usefulness of brain activations in complex computational tasks such as image captioning.

Keywords: machine learning, deep learning, neural networks, transformers, cognitive neuroscience, functional MRI, image captioning

Contents

List of Figures

List of Tables

Εκτεταμένη Περίληψη

Εισαγωγή

Το πρωτοποριακό έργο των Mitchell *et al.* [1] έδειξε ότι τα σήματα fMRI κωδικοποιούν σημαντικές σημασιολογικές πληροφορίες για συγκεκριμένα ουσιαστικά, οι οποίες μπορούν να είναι αποτελεσματικές, εάν χρησιμοποιηθούν για την αντιστοίχιση μεταξύ κατανεμημένων σημασιολογικών αναπαραστάσεων και ενεργοποιήσεων fMRI. Αυτό ήταν το πρώτο υπολογιστικό μοντέλο για την πρόβλεψη των εγκεφαλικών ενεργοποιήσεων που σχετίζονται με άγνωστες λέξεις. Από τότε πολλοί άλλοι προσπάθησαν να επεκτείνουν αυτή την πρώτη εργασία. [2–5] Πάντως, η χρήση γνωσιακών δεδομένων σε υπολογιστικά μοντέλα παραμένει ανοιχτό πεδίο για έρευνα. Χάρη στη διαθεσιμότητα συνόλων δεδομένων, όπου αξιοποιούνται εικόνες ως ερεθίσματα [6, 7], μπορούν να χρησιμοποιηθούν γνωσιακά δεδομένα ως επιπλέον είσοδος σε μοντέλα που λειτουργούν με εικόνες, όπως τα μοντέλα που δημιουργούν λεζάντες εικόνων. Στις ενεργοποιήσεις fMRI, λόγω των εγγενών σημασιολογικών πληροφοριών, οι παραγόμενες λεζάντες ενδέχεται να είναι περισσότερο εύλογες γνωσιακά, με αποτέλεσμα τη συνακόλουθη βελτίωση της απόδοσης αυτών των μοντέλων.

Σε αυτή την εργασία, προτείνουμε διάφορες τεχνικές ενσωμάτωσης των εγκεφαλικών ενεργοποιήσεων fMRI σε ένα μοντέλο που δημιουργεί λεζάντες για εικόνες, βασισμένο στην αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή των μετασχηματιστών. Ειδικότερα, εξετάζουμε την προσθήκη πληροφορίας στον κωδικοποιητή, την συνθηκοθέτηση του μηχανισμού προσοχής στον αποκωδικοποιητή και άλλες τεχνικές με χρήση ξεχωριστού κωδικοποιητή για τις εγκεφαλικές ενεργοποιήσεις. Λόγω του γεγονότος ότι τα δεδομένα fMRI είναι περιορισμένα, εκτελείται ένα βήμα «λεξικής επέκτασης», όπου προβλέπονται οι εγκεφαλικές ενεργοποιήσεις για καινούρια οπτικά ερεθίσματα, τα οποία δεν χρησιμοποιήθηκαν κατά το πείραμα fMRI.

Συνεισφορές

Οι κύριες συνεισφορές αυτής της εργασίας είναι οι εξής:

- Αξιολόγηση της αναπαραστατικής ομοιότητας μεταξύ οπτικών ή λεκτικών αναπαραστάσεων και ενεργοποιήσεων fMRI.
- Αξιολόγηση διαφορετικών μεθόδων για την κωδικοποίηση οπτικών ερεθισμάτων σε ενεργοποιήσεις fMRI.
- Προτάσεις διαφόρων αρχιτεκτονικών για την ενσωμάτωση των ενεργοποιήσεων fMRI σε μοντέλο δημιουργίας λεζάντας εικόνας με βάση την αρχιτεκτονική των μετασχηματιστών.
- Χρήση τεχνικών που επιτρέπουν περισσότερο προσαρμοστική ενσωμάτωση της ασθενούς τροπικότητας των ενεργοποιήσεων fMRI.

Σχετική Έρευνα

Τα τελευταία χρόνια έχει γίνει προσπάθεια σε αρκετές ερευνητικές εργασίες να αναλυθούν ποσοτικά οι σημασιολογικές αναπαραστάσεις του ανθρώπινου εγκεφάλου με τη βοήθεια ενεργοποιήσεων fMRI που προκλήθηκαν από οπτικά ερεθίσματα, όπως εικόνες ή κινηματογραφικές ταινίες [1, 2, 8–10]. Στα άρθρα [10, 11] αποκαλύφθηκε μια σχέση μεταξύ οπτικών ερεθισμάτων και εγκεφαλικών ενεργοποιήσεων με χρήση των σημασιολογικών κατηγοριών της Λεξικής Βάσης Δεδομένων (WordNet). Επίσης, κατασκευάστηκε ένας χάρτης σημασιολογικών αναπαραστάσεων του εγκεφαλικού φλοιού και αποδείχθηκε ότι οι σημασιολογικές πληροφορίες υπάρχουν σε ευρείες περιοχές μέσα στον εγκεφαλικό φλοιό. Στο άρθρο [12], υλοποιήθηκε ένα μοντέλο που ταξινομεί τις εγκεφαλικές ενεργοποιήσεις σε σημασιολογικές κατηγορίες και βρέθηκαν περιοχές του εγκεφάλου που ανταποκρίνονται σε συγκεκριμένες σημασιολογικές κατηγορίες. Στο [13], προτάθηκε ότι η οπτική προσοχή μεταβάλλει τις οπτικές αναπαραστάσεις στον εγκέφαλο για την βελτιστοποίηση της επεξεργασίας των σχετικών αντικειμένων στην περίπτωση της φυσικής όρασης. Στο [14] βρέθηκε μια συσχέτιση μεταξύ ενδιάμεσων αναπαραστάσεων στρωμάτων ενός Βαθέος Νευρωνικού Δικτύου και περιοχών της ραχιαίας οδού όταν το Βαθύ Νευρωνικό Δίκτυο (Deep Neural Network) που εκπαιδεύτηκε για αναγνώριση δράσεων χρησιμοποιήθηκε για την πρόβλεψη voxels της ραχιαίας οδού,

όπου ως ερεθίσματα χρησιμοποιήθηκαν κινηματογραφικές ταινίες. Στο [15] αποδείχθηκε ότι οι κατανεμημένες σημασιολογικές αναπαραστάσεις με βάση ένα μοντέλο skip-gram εκπαιδευμένο στην ιαπωνική Wikipedia είχαν συσχέτιση με ενεργοποιήσεις fMRI. Για να προκύψει μια πιο γνωσιακά αληθοφανής θεώρηση των κατανεμημένων σημασιολογικών μοντέλων, στο [16] ελέγχεται αν τα μοντέλα που βασίζονται σε εικόνα συλλαμβάνουν τα σημασιολογικά μοτίβα που προκύπτουν από ενεργοποιήσεις fMRI. Το συμπέρασμα είναι ότι τα μοντέλα που βασίζονται σε εικόνα βελτιώνουν αυτά που βασίζονται σε κείμενο. Στο [17] χρησιμοποιήθηκε ένα γλωσσικό μοντέλο με βάση το word2vec μαζί με ένα οπτικά βασισμένο μοντέλο για την αποκωδικοποίηση συγκεκριμένων και αφηρημένων ουσιαστικών από ενεργοποιήσεις του εγκεφάλου. Ωστόσο, δεν παρατηρήθηκε σημαντική βελτίωση για πολυτροπικά μοντέλα ακόμη και στην περίπτωση των συγκεκριμένων ουσιαστικών και το γλωσσικό μοντέλο ήταν ανώτερο στην περίπτωση των αφηρημένων ουσιαστικών. Στο [18] παρουσιάζεται ένα μοντέλο βαθέος αυτο-κωδικοποιητή που αποτελείται από CNN με LSTM, το οποίο εκπαιδεύεται σε ακολουθίες τομών fMRI και προβλέπει ολόκληρο τον όγκο του εγκεφάλου χρησιμοποιώντας πολυτροπικά ερεθίσματα ως είσοδο. Στο [4] τα ερεθίσματα από ουσιαστικά επεκτείνονται σε εικόνες και ουσιαστικά για να απεικονίσουν τον ισχυρό συσχετισμό μεταξύ γλωσσικών και οπτικών αναπαραστάσεων στον ανθρώπινο εγκέφαλο. Στο [19] κατασκευάζονται πολυτροπικά μοντέλα που αντιδιαστέλλουν μεταξύ των εσωτερικών οπτικών ιδιοτήτων των αντικειμένων και του εξωτερικού οπτικού πλαισίου. Αυτά τα μοντέλα, όταν αξιολογούνται κατά την διαδικασία αποκωδικοποίησης της εγκεφαλικής δραστηριότητας, αποδίδουν καλύτερα από εκείνα που βασίζονται σε πλήρεις εικόνες. Στο [20], αξιολογούνται αρκετά μοντέλα ενσωμάτωσης λέξεων μαζί με τους συνδυασμούς τους για αποκωδικοποίηση δεδομένων fMRI που σχετίζονται με τρεις κατηγορίες λέξεων και τρεις τρόπους εισαγωγής ερεθισμάτων. Τα πολυτροπικά και μετα-λεκτικά μοντέλα ενσωμάτωσης επιτυγχάνουν καλύτερη απόδοση από ό, τι τα συστατικά τους μοντέλα ενσωμάτωσης, δεδομένου ότι οι οπτικές πληροφορίες είναι σημαντικές για τη διάκριση μεταξύ λέξεων λόγω του μεγάλου τμήματος των πληροφοριακών voxel στα οπτικά δίκτυα.

Επισκόπηση Συνόλων Δεδομένων

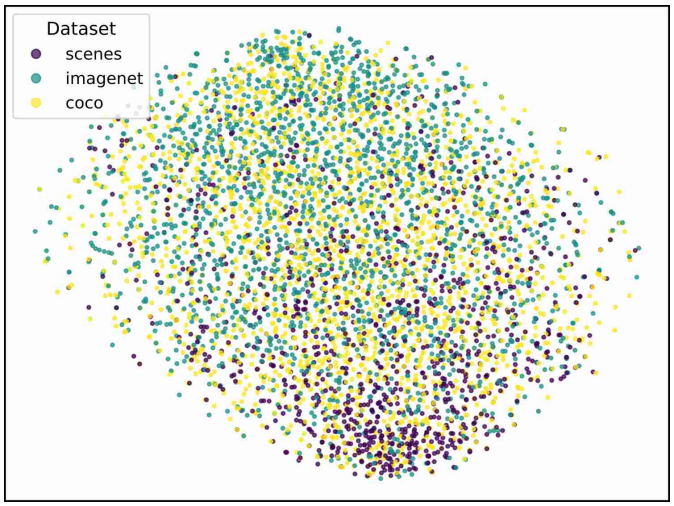
BOLD5000

Το BOLD5000 [21] είναι ένα σύνολο δεδομένων fMRI που βασίζεται σε αποκρίσεις από σχεδόν 5000 διαφορετικές εικόνες πραγματικού κόσμου που αλληλεπικαλύπτονται με τυπικά σύνολα δεδομένων όρασης υπολογιστών (SUN, COCO, ImageNet). Είναι μια τάξη μεγέθους μεγαλύτερο από προηγούμενα παρόμοια σύνολα δεδομένων και επιτρέπει πιο λεπτομερή έρευνα στη νευρωνική αναπαράσταση οπτικών χαρακτηριστικών, κατηγοριών και σημασιολογίας. Συγκεκριμένα, συλλέχθηκαν δεδομένα fMRI από τέσσερις συμμετέχοντες (16 συνεδρίες) σε μια αργή διαδικασία που σχετίζεται με συμβάντα και είχε ως στόχο την αποσύνδεση όποιων ερεθισμάτων προκαλούσαν συγκεκριμένες νευρικές αποκρίσεις. Οι εικόνες αποτελούνται από 1000 σκηνές εσωτερικών και εξωτερικών χώρων 250 κατηγοριών (SUN), 2000 αντικείμενα ενσωματωμένα σε ρεαλιστικό πλαίσιο (COCO) και 1916 εικόνες κυρίως μεμονωμένων αντικειμένων (ImageNet). Επιπλέον, υποδειγματοληπτούνται σε 375×375 εικονοστοιχεία και για να εξασφαλιστεί ομοιόμορφη φωτεινότητα, γίνεται κανονικοποίηση στο γκρι. Εκτός από τα ανεπεξέργαστα δεδομένα που παρέχονται για κάθε περίοδο / συνεδρία, παρέχονται επεξεργασμένα χαρακτηριστικά των voxel για δέκα περιοχές ενδιαφέροντος (ROIs)—πέντε περιοχές (PPA, RSC, OPA, LOC, EV) για κάθε εγκεφαλικό ημισφαίριο (LH, RH)—και για πέντε χρονικά διαστήματα δύο δευτερολέπτων (TR [1–5]). Σύμφωνα με λειτουργικούς εντοπιστές σε σχέση με το ερέθισμα από σκηνές, από αντικείμενα και ανακατεμένες εικόνες, τρεις ROI είναι επιλεκτικές σκηνής (PPA, RSC, OPA), η μία είναι επιλεκτική αντικειμένων (LOC) και μία αντιστοιχεί σε πρώιμο οπτικό σύστημα (EV). Για να εξαχθούν οι ενεργοποιήσεις voxel για κάθε περιοχή ενδιαφέροντος, πραγματοποιήθηκε χρήση κατωφλίου σε κάθε μία χρησιμοποιώντας ομαδική διόρθωση σφάλματος $p < 0.0001, k = 30$. Μια επισκόπηση του συνόλου δεδομένων παρουσιάζεται στον πίνακα 1.

Συμμετέχοντες	4
Συνεδρίες fMRI ανά συμμετέχοντα	16
Συνολικές διαδρομές σκηνής fMRI ανά συμμετέχοντα	142
Συνολικές λειτουργίες Localizer που εκτελούνται ανά συμμετέχοντα	8
Συνολικές δοκιμές σκηνής ανά συμμετέχοντα	5,254
Μοναδικά ερεθίσματα σκηνής ανά συμμετέχοντα	4,916

Πίνακας 1: Επισκόπηση συνόλου δεδομένων BOLD5000. Προσαρμογή του πίνακα από το [21].

Πραγματοποιήθηκε μια οπτικοποίηση t-SNE για τις ενεργοποιήσεις voxel μιας περιοχής επιλεκτικής για σκηνές (RH-RSC) για να προσδιοριστεί αν υπάρχει διαφορετική απόκριση στα σύνολα δεδομένων σε σχέση με τα ερεθίσματα. Από το σχήμα 1 παρατηρούμε ότι οι ενεργοποιήσεις που αντιστοιχούν στο σύνολο δεδομένων σκηνών SCENES σχηματίζουν συστάδα, πράγμα το οποίο συμφωνεί με τη διαίσθησή μας.



Σχήμα 1: t-SNE για τις ενεργοποιήσεις voxel της επιλεκτικής για σκηνές RH-RSC περιοχής ενδιαφέροντος για το χρονικό διάστημα 6–8 sec (TR4).

MS-COCO

Το MS-COCO [22] είναι ένα από τα πιο συχνά χρησιμοποιούμενα σύνολα δεδομένων αξιολόγησης μοντέλων για δημιουργία λεζάντας εικόνας [23–27]. Περιλαμβάνει 82.783 εικόνες για εκπαίδευση και 40.504 εικόνες για επικύρωση, με πέντε σχόλια από άνθρωπο ανά εικόνα. Επειδή τα σχόλια για το επίσημο σύνολο δοκιμών δεν είναι διαθέσιμα δημόσια, υιοθετούμε τον ευρέως χρησιμοποιούμενο διαχωρισμό “Karpathy” [28] και παίρνουμε 113.287 εικόνες για εκπαίδευση, 5.000 για επικύρωση και 5.000 για δοκιμές. Προκειμένου να αξιολογηθεί η ποιότητα για τις λεζάντες που δημιουργήθηκαν, χρησιμοποιούνται τυπικές μετρικές αυτόματης αξιολόγησης, και πιο συγκεκριμένα οι BLEU [29], METEOR [30], ROUGE [31], CIDEr [32] και SPICE [33].

Μεθοδολογία

Βασική Αρχιτεκτονική

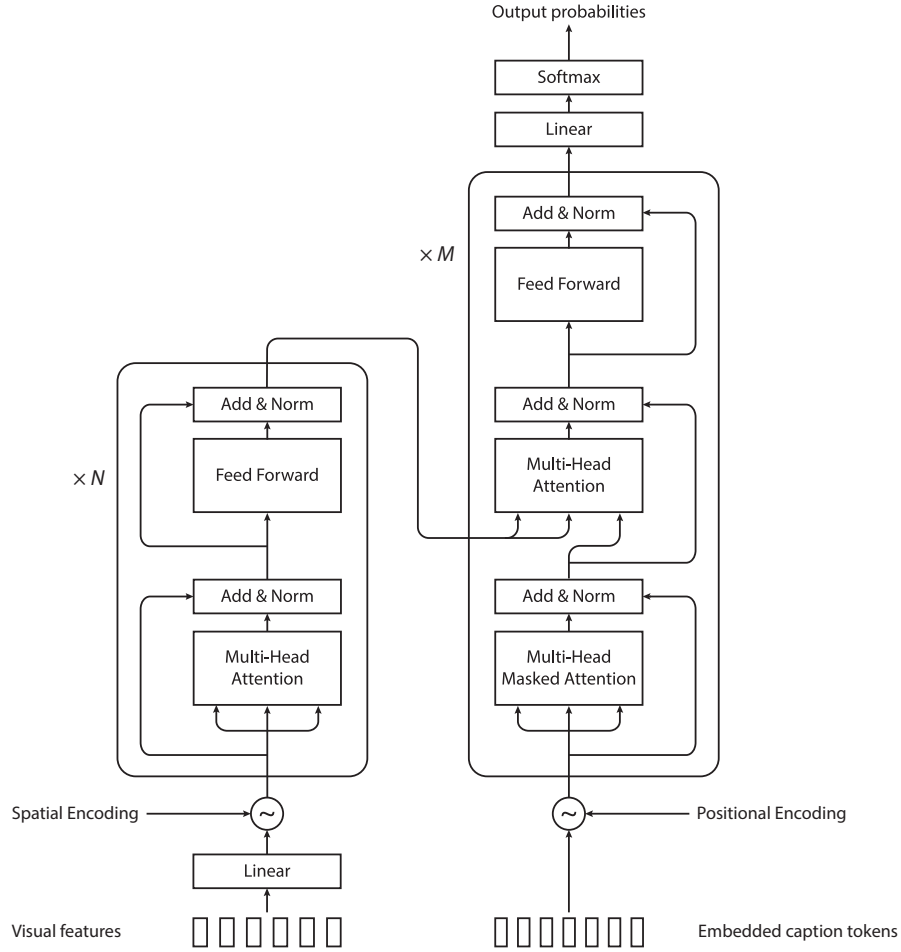
Η βασική αρχιτεκτονική δημιουργίας λεζάντων για εικόνες βασίζεται στην αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή των μετασχηματιστών [34]. Συγκεκριμένα, ο κωδικοποιητής εκτελεί αυτο-προσοχή σε οπτικά χαρακτηριστικά που έχουν εξαχθεί μέσω Ταχύτερου R-CNN, όπως περιγράφεται στο [23]. Ο αποκωδικοποιητής εκτελεί καλυμμένη αυτο-προσοχή σε διακριτικά λέξεων σε λεζάντες και οπτικογλωσσική προσοχή που συνθοκοθετεί την οπτική ροή στη ροή της γλώσσας. Τα οπτικά χαρακτηριστικά προβάλλονται μέσω γραμμικού στρώματος στον χαμηλότερο-διαστατικό χώρο αναπαράστασης του κωδικοποιητή. Η συνολική βασική αρχιτεκτονική απεικονίζεται στο σχήμα 2 και η απόδοσή της στο σύνολο δεδομένων MS-COCO φαίνεται στον πίνακα 2.

Σκορ	Βασικό μοντέλο
Bleu_1	0.740
Bleu_2	0.578
Bleu_3	0.445
Bleu_4	0.343
METEOR	0.279
ROUGE_L	0.558
CIDEr	1.117
SPICE	0.208

Πίνακας 2: Απόδοση της βασικής αρχιτεκτονικής για το σύνολο δεδομένων MS-COCO.

Representational Similarity Analysis

Η Ανάλυση Αναπαραστατικής Ομοιότητας (Representational Similarity Analysis) [35] χρησιμοποιήθηκε αρχικά για να συσχετιστούν οι αναπαραστάσεις οπτικών αντικειμένων στον εγκέφαλο με αναπαραστάσεις υπολογιστικών μοντέλων. Προκειμένου να αποκτήσουμε μια καλύτερη διαίσθηση αναφορικά με τις εγκεφαλικές αναπαραστάσεις, πριν πραγματοποιήσουμε «λεξική επέκταση», εκτελούμε μια προκαταρκτική RSA για τον προσδιορισμό της ομοιότητας των αναπαραστάσεων του εγκεφάλου μεταξύ υποκειμένων της μελέτης σε σχέση με τις διαφορετικές περιοχές ενδιαφέροντος. Έστω ότι $\boldsymbol{v}_i^{(r)} \in \mathbb{R}^d, d \in (100, 200)$ είναι οι ενεργοποιήσεις voxel για την περιοχή ενδιαφέροντος

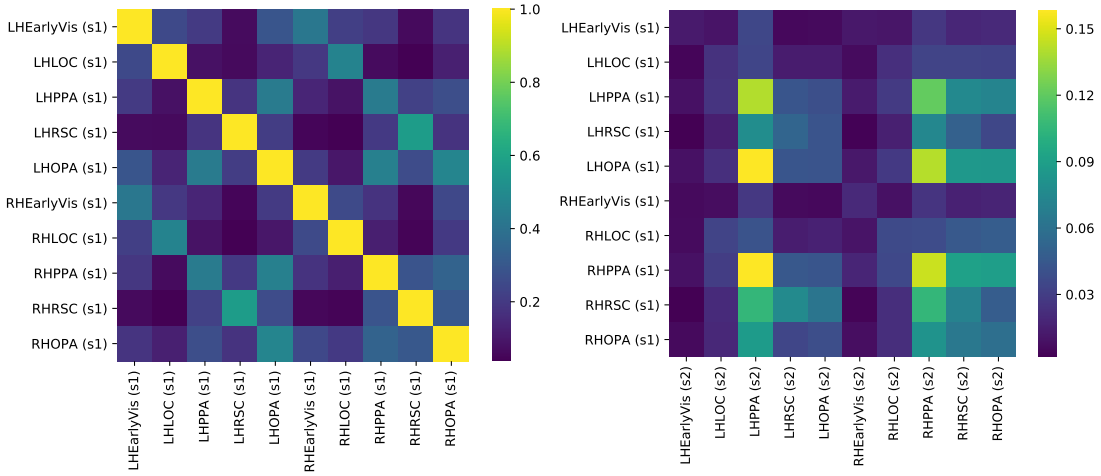


Σχήμα 2: Βασική αρχιτεκτονική δημιουργίας λεζάντας εικόνας που βασίζεται σε μετασχηματιστές.

r για το ερέθισμα εικόνας $i \in S$, όπου $S = \{1, 2, \ldots s\}$. Για κάθε περιοχή ενδιαφέροντος, υπολογίζουμε:

$$\boldsymbol{d}^{(r)} = (d_{ij}^{(r)}), \qquad d_{ij}^{(r)} = \cos(v_i^{(r)}, v_j^{(r)}) \quad i \in [1..s],\ j \in [i+1..s] \tag{1}$$

Κατόπιν υπολογίζουμε τη συσχέτιση Spearman $s(\boldsymbol{d}^{(r)}, \boldsymbol{d}^{(r')})$ για τις περιοχές ενδιαφέροντος r και r', οι οποίες μπορούν να επιλεγούν είτε από το ίδιο είτε από διαφορετικά υποκείμενα. Αυτό απεικονίζεται στο σχήμα 3, όπου στον αριστερό θερμικό χάρτη οι r και r' αντιστοιχούν στο υποκείμενο 1 και στον δεξιό θερμικό χάρτη η r αντιστοιχεί στο υποκείμενο 1 και η r' στο υποκείμενο 2, και κάθε καταχώριση του θερμικού χάρτη είναι η αντίστοιχη συσχέτιση Spearman.



Σχήμα 3: Ανάλυση Αναπαραστατικής Ομοιότητας για τις περιοχές ενδιαφέροντος του συμμετέχοντα 1 (αριστερά) και για τις περιοχές ενδιαφέροντος των συμμετεχόντων 1 και 2 (δεξιά).

Είναι ενδιαφέρον να σημειωθεί ότι όταν οι r και r' είναι από τον ίδιο συμμετέχοντα, μπορεί να παρατηρηθεί μια συσχέτιση μεταξύ αριστερού και δεξιού ημισφαιρίου από τα διαγώνια μοτίβα στο σχήμα 3 (αριστερά). Όταν οι r και r' προέρχονται από διαφορετικούς συμμετέχοντες, παρατηρούνται αδύναμες συσχετίσεις μόνο μεταξύ περιοχών ενδιαφέροντος υψηλότερου επιπέδου.

Μια παρόμοια ανάλυση πραγματοποιείται για να διαπιστωθεί η ομοιότητα των οπτικών αναπαραστάσεων, που εξάγονται από ένα προεκπαιδευμένο δίκτυο VGG16, 300-διαστάσεων αναπαραστάσεων GloVe840B [36] και ενεργοποιήσεων voxel. Από τον πίνακα 3 μπορούμε να δούμε ότι οι ενεργοποιήσεις voxel είναι περισσότερο συσχετισμένες με τις οπτικές αναπαραστάσεις παρά με τις αναπαραστάσεις λέξεων, με τις μεγαλύτερες συσχετίσεις να παρατηρούνται για περιοχές ενδιαφέροντος υψηλότερου επιπέδου. Αυτή η παρατήρηση υποδηλώνει ότι το μοντέλο κωδικοποίησης που προβλέπει ενεργοποιήσεις voxel πρέπει να βασίζεται σε οπτικές αναπαραστάσεις και όχι σε αναπαραστάσεις λέξεων.

		P1	P2	P3	P4
LHEarlyVis	visual	0.051	0.029	0.012	0.031
LHEarlyVis	GloVe	−0.017	−0.014	−0.007	−0.025
LHLOC	visual	−0.008	−0.005	0.007	0.010
LHLOC	GloVe	0.000	−0.013	−0.006	−0.037
LHPPA	visual	0.145	0.172	0.126	0.159
LHPPA	GloVe	0.061	0.088	0.116	0.085
LHRSC	visual	0.077	0.048	0.034	0.036
LHRSC	GloVe	0.049	0.026	0.029	0.019
LHOPA	visual	0.164	0.045	0.068	0.102
LHOPA	GloVe	0.100	0.027	0.061	0.048
RHEarlyVis	visual	0.054	0.037	−0.000	0.030
RHEarlyVis	GloVe	0.007	−0.013	−0.013	−0.032
RHLOC	visual	0.005	0.014	−0.010	0.012
RHLOC	GloVe	−0.024	−0.010	−0.010	−0.015
RHPPA	visual	0.202	0.172	0.118	0.165
RHPPA	GloVe	0.084	0.111	0.084	0.056
RHRSC	visual	0.148	0.123	0.101	0.062
RHRSC	GloVe	0.096	0.097	0.092	0.036
RHOPA	visual	0.134	0.110	0.057	0.128
RHOPA	GloVe	0.104	0.086	0.031	0.051
visual	GloVe	0.265	0.266	0.268	0.262

Πίνακας 3: Ανάλυση Αναπαραστατικής Ομοιότητας μεταξύ περιοχών ενδιαφέροντος, οπτικών χαρακτηριστικών (VGG16) και αναπαραστάσεων λέξεων (300-διαστάσεων GloVe840B) για όλους τους συμμετέχοντες P[1-4]. Τα «LH» και «RH» δηλώνουν το αριστερό και το δεξιό ημισφαίριο του εγκεφάλου, αντίστοιχα.

Λεξική Επέκταση

Λόγω των περιορισμένων δεδομένων fMRI είναι απαραίτητο να επεκταθεί το «λεξικό» των διαθέσιμων fMRI με τη δημιουργία ενός μοντέλου κωδικοποίησης που προβλέπει ενεργοποιήσεις voxel για εικόνες που δεν χρησιμοποιήθηκαν ως ερεθίσματα στο σύνολο δεδομένων fMRI. Αυτό μας επιτρέπει να λάβουμε ενεργοποιήσεις voxel για οποιαδήποτε είσοδο εικόνας στο μοντέλο που παράγει λεζάντες. Για το σκοπό αυτό, εκπαιδεύουμε ένα μοντέλο περίπου 5000 ζευγών εικόνων και αντίστοιχων fMRI, το οποίο μαθαίνει μια αντιστοίχιση από οπτικά χαρακτηριστικά σε fMRI. Τα οπτικά χαρακτηριστικά $\boldsymbol{v}_i \in \mathbb{R}^{1\times 512}$ λαμβάνονται από τον μέσο όρο συγκέντρωσης (pooling) στο τελευταίο στρώμα ενός προεκπαιδευμένου δικτύου VGG-16. Για τα fMRI παραλείπουμε τις περιοχές ενδιαφέροντος του αρχικού οπτικού συστήματος, ούτως ώστε $\boldsymbol{f}_i \in \mathbb{R}^{1\times 1100}$. Στις ακόλουθες παραγράφους περιγράφουμε τρεις διαφορετικές προσεγγίσεις που χρησιμοποιήθηκαν για αυτή τη διαδικασία κωδικοποίησης.

Παλινδρόμηση Κορυφογραμμής

Η παλινδρόμηση κορυφογραμμής (ridge regression) είναι η επικρατούσα τεχνική που χρησιμοποιείται για αντιστοίχιση σε ενεργοποιήσεις voxel [2, 6, 37, 38] λόγω της απλότητας και της αρκετά καλής απόδοσής της. Θεωρώντας ότι ο πίνακας $\boldsymbol{F} \in \mathbb{R}^{5000\times1100}$ αντιπροσωπεύει τα χαρακτηριστικά fMRI για όλα τα ερεθίσματα και ότι ο $\boldsymbol{V} \in \mathbb{R}^{5000\times512}$ αντιπροσωπεύει τα αντίστοιχα οπτικά χαρακτηριστικά, τα χαρακτηριστικά fMRI νέων εικόνων $\boldsymbol{V}_{\text{new}}$ που δεν χρησιμοποιούνται ως ερεθίσματα λαμβάνονται από

$$\hat{\boldsymbol{F}} = \boldsymbol{V}_{\text{new}}\hat{\boldsymbol{W}} \tag{2}$$

όπου $\hat{\boldsymbol{W}} \in \mathbb{R}^{512\times1100}$ είναι ένας πίνακας που έχει εκμαθηθεί και του οποίου η i-οστή στήλη $\hat{\boldsymbol{W}}_{:,i}$ δίνεται από

$$\hat{\boldsymbol{W}}_{:,i} = \underset{\boldsymbol{W}_{:,i}\in\mathbb{R}^{512\times1}}{\arg\min} ||\boldsymbol{F}_{:,i} - \boldsymbol{V}\boldsymbol{W}_{:,i}||_2^2 + \lambda_i||\boldsymbol{W}_{:,i}||_2^2 \tag{3}$$

και $\lambda_i \geq 0$ είναι μια παράμετρος κανονικοποίησης που λαμβάνεται μέσω διασταυρούμενης επικύρωσης.

Εκμάθηση Αραιών Λεξικών

Μια λιγότερο συχνά χρησιμοποιούμενη τεχνική είναι η Εκμάθηση Αραιών Λεξικών (Sparse Dictionary Learning) όπου τα χαρακτηριστικά fMRI $\boldsymbol{F} \in \mathbb{R}^{5000\times1100}$ αποσυντίθενται έτσι ώστε

$$\boldsymbol{F} = \boldsymbol{C}\boldsymbol{D} \tag{4}$$

όπου $\boldsymbol{C} \in \mathbb{R}^{5000\times a}$ είναι ένας αραιός πίνακας συντελεστών λεξικού και ο $\boldsymbol{D} \in \mathbb{R}^{a\times1100}$ περιέχει τα διανύσματα βάσης του λεξικού ως γραμμές. Αυτή η αποσύνθεση στοχεύει ουσιαστικά στην εξομάλυνση των αναπαραστάσεων fMRI, εκφράζοντας τις ως γραμμικό συνδυασμό ατόμων του λεξικού που θεωρητικά δεν συλλαμβάνουν πλήρως τον θόρυβο που υπάρχει στο σήμα των fMRI. Κατόπιν, εκμαθαίνεται ένας πίνακας $\hat{\boldsymbol{W}}$ που αντιστοιχίζει οπτικά χαρακτηριστικά σε συντελεστές λεξικού μέσω παλινδρόμησης κορυφογραμμής. Τα προβλεπόμενα χαρακτηριστικά fMRI λαμβάνονται σε δύο βήματα

από τις

$$\hat{\boldsymbol{C}} = \boldsymbol{V}_{\text{new}}\hat{\boldsymbol{W}} \tag{5}$$

$$\hat{\boldsymbol{F}} = \hat{\boldsymbol{C}}\boldsymbol{D} \tag{6}$$

Ομοιότητα

Μια άλλη τεχνική που περιγράφεται στο [5] εκφράζει τις προβλεπόμενες ενεργοποιήσεις fMRI ως γραμμικό συνδυασμό γνωστών ενεργοποιήσεων fMRI. Τυπικά, οι προβλεπόμενες ενεργοποιήσεις fMRI $\hat{\boldsymbol{f}}_{\text{new}} \in \mathbb{R}^{1\times 1100}$ για μια νέα εικόνα $\hat{\boldsymbol{v}}_{\text{new}} \in \mathbb{R}^{1\times 512}$ δίνονται από την

$$\hat{\boldsymbol{f}}_{\text{new}} = \sum_{i=1}^{5000} g(\boldsymbol{v}_{\text{new}}, \boldsymbol{v}_i)\boldsymbol{f}_i \tag{7}$$

όπου $g(\boldsymbol{v}_i, \boldsymbol{v}_j)$ είναι μια συνάρτηση που εκφράζει την ομοιότητα μεταξύ των εικόνων $\boldsymbol{v}_i$ και $\boldsymbol{v}_j$.

Αξιολόγηση της Λεξικής Επέκτασης

Για να εκτιμηθεί η απόδοση των διαφορετικών μεθόδων που προτάθηκαν για λεξική επέκταση, χρησιμοποιούμε την τυπική διαδικασία αξιολόγησης από τη βιβλιογραφία [38]. Πιο συγκεκριμένα, εκπαιδεύουμε επανειλημμένα το μοντέλο με διαφορετικά υποσύνολα $m-2$ ζεύγων εικόνων–fMRI και εκτελούμε την ακόλουθη αξιολόγηση στα δύο υπόλοιπα ζεύγη, όπου το m είναι ο συνολικός αριθμός ζευγών εικόνων–fMRI. Μια πρόβλεψη θεωρείται επιτυχής εάν

$$g(\boldsymbol{f}_1, \hat{\boldsymbol{f}}_1) + g(\boldsymbol{f}_2, \hat{\boldsymbol{f}}_2) > g(\boldsymbol{f}_1, \hat{\boldsymbol{f}}_2) + g(\boldsymbol{f}_2, \hat{\boldsymbol{f}}_1) \tag{8}$$

όπου τα πραγματικά fMRI είναι $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, τα προβλεπόμενα είναι $\hat{\boldsymbol{f}}_1$, $\hat{\boldsymbol{f}}_2$ και g είναι μια συνάρτηση ομοιότητας π.χ. συσχέτιση Spearman. Ο πίνακας 4 εμφανίζει τα αποτελέσματα εκτέλεσης αυτής της διαδικασίας για $n = 1000$ επαναλήψεις.

	Παλινδρόμηση	Εκμάθηση Λεξικού	Ομοιότητα
υποκείμενο 1	0.8775	0.925	0.886
υποκείμενο 2	0.8575	0.925	0.865
υποκείμενο 3	0.8370	0.845	0.873
υποκείμενο 4	0.8500	0.905	0.861

Πίνακας 4: Αξιολόγηση της λεξικής επέκτασης από οπτικά χαρακτηριστικά (VGG16) σε ενεργοποιήσεις fMRI. Ο αριθμός των επαναλήψεων της εξίσωσης 8 είναι $n = 1000$. Για την μέθοδο «Ομοιότητα», ως συνάρτηση ομοιότητας χρησιμοποιήθηκε το εσωτερικό γινόμενο.

Και οι τρεις μέθοδοι επιτυγχάνουν γενικά καλά αποτελέσματα με την εκμάθηση αραιών λεξικών να έχει καλύτερες βαθμολογίες για τρεις από τους τέσσερις συμμετέχοντες. Η μέθοδος που βασίζεται στην ομοιότητα, με το εσωτερικό γινόμενο ως συνάρτηση ομοιότητας g στην εξίσωση 7 παρήγαγε καλές βαθμολογίες αξιολόγησης ακόμη και αν οι προβλεπόμενες ενεργοποιήσεις voxel είχαν εξωπραγματικά υψηλές τιμές. Αυτό το γεγονός δείχνει ότι η διαδικασία αξιολόγησης που χρησιμοποιείται στη βιβλιογραφία δεν παρέχει ισχυρές εγγυήσεις ποιότητας των προβλεπόμενων χαρακτηριστικών fMRI.

Ένας άλλος τρόπος αξιολόγησης των προβλεπόμενων χαρακτηριστικών fMRI, προκειμένου να διασφαλιστεί ότι περιέχουν χρήσιμες πληροφορίες, είναι η χρήση τους αντί των οπτικών χαρακτηριστικών. Η είσοδος στον κωδικοποιητή γίνεται

$$\hat{\boldsymbol{F}}\boldsymbol{W}_{\text{fmri}} + \boldsymbol{P}\boldsymbol{W}_{\text{pos}} \tag{9}$$

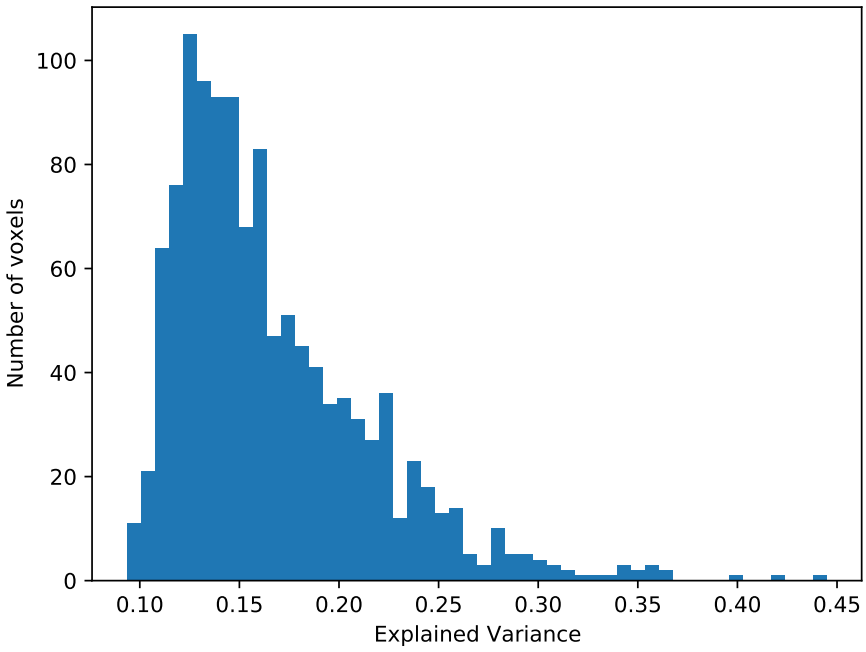
όπου $\hat{\boldsymbol{F}} \in \mathbb{R}^{\text{boxes}\times 1100}$ περιέχει τα προβλεπόμενα χαρακτηριστικά fMRI για κάθε πλαίσιο οριοθέτησης, $\boldsymbol{P} \in \mathbb{R}^{\text{boxes}\times 5}$ κωδικοποιεί τις πληροφορίες θέσης των πλαισίων οριοθέτησης και $\boldsymbol{W}_{\text{fmri}} \in \mathbb{R}^{1100\times 512}$, $\boldsymbol{W}_{\text{pos}} \in \mathbb{R}^{5\times 512}$ είναι εκμαθήσιμοι πίνακες. Ο πίνακας 5 παρουσιάζει τα αποτελέσματα που προέκυψαν για όλες τις μεθόδους λεξικής επέκτασης. Βλέπουμε ότι το μοντέλο επιτυγχάνει ικανοποιητικά αποτελέσματα ακόμα και δίχως τα οπτικά χαρακτηριστικά. Το καλύτερο μοντέλο βασίζεται στην Παλινδρόμηση Κορυφογραμμής, μολονότι η Εκμάθηση Λεξικού είχε λίγο καλύτερη απόδοση στην προηγούμενη μέθοδο αξιολόγησης. Αυτό μπορεί να σημαίνει ότι η Παλινδρόμηση Κορυφογραμμής διατηρεί καλύτερα κάποιες ιδιότητες από τον χώρο των οπτικών χαρακτηριστικών.

Σκορ	Βασικό μοντέλο	Παλινδρόμηση	Εκμάθηση Λεξικού	Ομοιότητα
Bleu_1	0.740	0.693	0.672	0.635
Bleu_2	0.578	0.523	0.498	0.453
Bleu_3	0.445	0.393	0.369	0.326
Bleu_4	0.343	0.298	0.276	0.241
METEOR	0.279	0.262	0.249	0.227
ROUGE_L	0.558	0.530	0.516	0.483
CIDEr	1.117	0.979	0.893	0.751
SPICE	0.208	0.189	0.174	0.154

Πίνακας 5: Απόδοση στην δημιουργία λεζάντας εικόνας για το σύνολο δεδομένων MS-COCO κατά την εκπαίδευση με προβλεπόμενα fMRI αντί για οπτικά χαρακτηριστικά. Τα fMRI έχουν προβλεφθεί χρησιμοποιώντας οπτικά χαρακτηριστικά (VGG-16) που εξάγονται από πλαίσια οριοθέτησης.

Ερμηνευόμενη Διακύμανση για Αντιστοίχιση σε fMRI

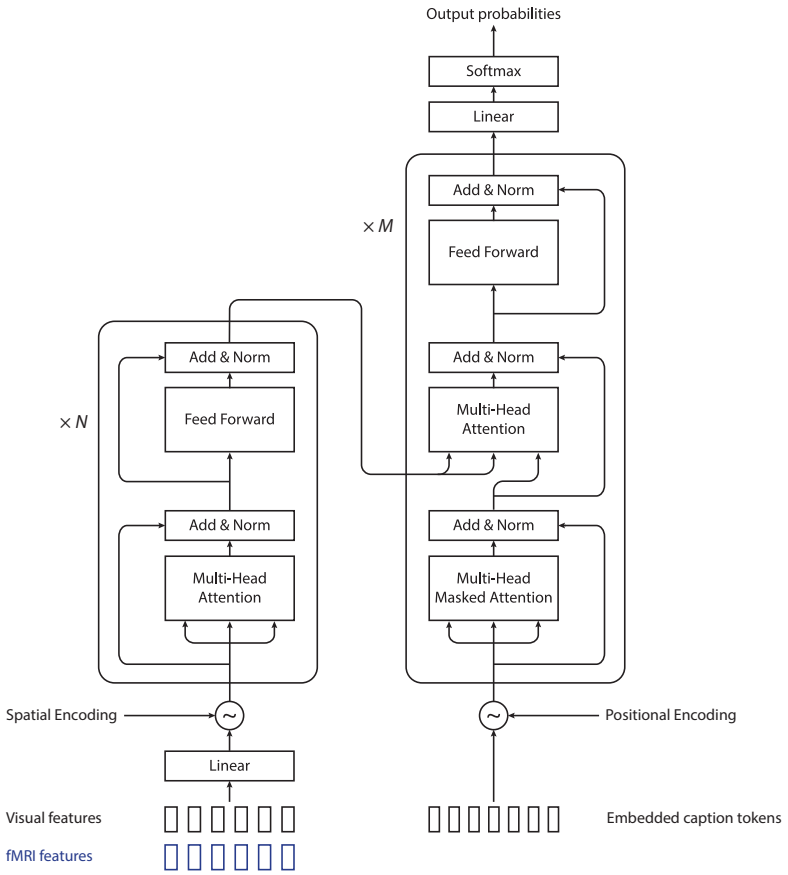
Στην περίπτωση της Παλινδρόμησης Κορυφογραμμής μπορούμε να αξιολογήσουμε επίσης την αντιστοίχιση από οπτικά χαρακτηριστικά $\boldsymbol{V} \in \mathbb{R}^{5000\times512}$ στα προβλεπόμενα fMRI $\hat{\boldsymbol{F}} \in \mathbb{R}^{5000\times1100}$ μέσω της ερμηνευόμενης διακύμανσης για όλα τα voxel. Το σχήμα 4 δείχνει το ιστόγραμμα που προκύπτει. Η αντιστοίχιση δεν είναι ικανοποιητική, καθώς για τα περισσότερα voxel προβλέπεται η μέση απόκρισή τους. Αυτό μπορεί να αποδοθεί στο γεγονός ότι ένα γραμμικό μοντέλο δεν είναι αρκετά ισχυρό για αυτή την αντιστοίχιση, στον εγγενή θόρυβο του σήματος fMRI ή στο γεγονός ότι τα voxel περιέχουν πληροφορίες που δεν μπορούν να προβλεφθούν αποκλειστικά από οπτικά χαρακτηριστικά.



Σχήμα 4: Ιστόγραμμα Ερμηνευόμενης Διακύμανσης ανά voxel στην περίπτωση της Παλινδρόμησης Κορυφογραμμής. Όταν η ΕΔ είναι πλησίον στο 0, τότε προβλέπεται η μέση απόκριση του voxel.

Ενσωμάτωση στον Κωδικοποιητή

Η απλούστερη μέθοδος μέσω της οποίας μπορούν να ενσωματωθούν τα χαρακτηριστικά fMRI στο μοντέλο δημιουργίας λεζάντας είναι μέσω ενσωμάτωσης στον κωδικοποιητή. Αυτή η περίπτωση απεικονίζεται στο σχήμα 5. Διάφορες παραλλαγές περιγράφονται στις ακόλουθες παραγράφους, όπου τα χαρακτηριστικά fMRI είτε προστίθενται είτε συνενώνονται με τα οπτικά χαρακτηριστικά που λειτουργούν ως είσοδος στον κωδικοποιητή.



Σχήμα 5: Ενσωμάτωση στον κωδικοποιητή. Τα χαρακτηριστικά fMRI προβλέπονται για τα οπτικά χαρακτηριστικά κάθε πλαισίου οριοθέτησης και προστίθενται ή συνενώνονται στα αντίστοιχα οπτικά χαρακτηριστικά.

Πλήρωση, πρόσθεση Τα προβλεπόμενα fMRI προστίθενται στα οπτικά χαρακτηριστικά κάθε πλαισίου οριοθέτησης με κατάλληλη πλήρωση μηδενικών. Τα τελικά χαρακτηριστικά εισόδου για το i-οστό πλαίσιο οριοθέτησης δίνονται από

$$\boldsymbol{X}_{i,:} = (\boldsymbol{V}_{i,:} + \text{pad}(\hat{\boldsymbol{F}}_{i,:}))\boldsymbol{W}_{\text{visual}} + \boldsymbol{P}_{i,:}\boldsymbol{W}_{\text{pos}} \tag{10}$$

όπου ο $\boldsymbol{V} \in \mathbb{R}^{\text{boxes}\times 2048}$ περιέχει τα οπτικά χαρακτηριστικά για κάθε πλαίσιο οριοθέτησης τα οποία έχουν εξαχθεί μέσω Faster-RCNN και ο $\hat{\boldsymbol{F}} \in \mathbb{R}^{\text{boxes}\times 1100}$ περιέχει τα προβλεπόμενα χαρακτηριστικά fMRI που έχουν αντιστοίχιση ενός προς ένα με τα οπτικά χαρακτηριστικά των οριοθετημένων πλαισίων. Επίσης, ο $\boldsymbol{P} \in \mathbb{R}^{\text{boxes}\times 5}$ κωδικοποιεί τις πληροφορίες θέσης των πλαισίων οριοθέτησης.

Προβολή, πρόσθεση Αυτή η περίπτωση είναι παρόμοια με την προηγούμενη, με τη διαφορά ότι τα προβλεπόμενα fMRI προβάλλονται πρώτα μέσω γραμμικού στρώματος στην κατώτερη διάσταση του χώρου ενσωμάτωσης του κωδικοποιητή και στη συνέχεια προστίθενται στα οπτικά χαρακτηριστικά. Τα τελικά χαρακτηριστικά εισόδου για το i-οστό πλαίσιο οριοθέτησης δίνονται από

$$\boldsymbol{X}_{i,:} = \boldsymbol{V}_{i,:}\boldsymbol{W}_{\text{visual}} + \hat{\boldsymbol{F}}_{i,:}\boldsymbol{W}_{\text{fmri}} + \boldsymbol{P}_{i,:}\boldsymbol{W}_{\text{pos}} \tag{11}$$

Συνένωση Μια άλλη εναλλακτική είναι η συνένωση των προβλεπόμενων fMRI με τα οπτικά χαρακτηριστικά όπου

$$\boldsymbol{X}_{i,:} = (\boldsymbol{V}_{i,:}||\hat{\boldsymbol{F}}_{i,:})\boldsymbol{W} + \boldsymbol{P}_{i,:}\boldsymbol{W}_{\text{pos}} \tag{12}$$

και ο $\boldsymbol{W} \in \mathbb{R}^{(2048+1100)\times 512}$ προκύπτει μέσω μάθησης. Ο πίνακας 6 εμφανίζει τα αποτελέσματα των παραπάνω μεθόδων, με τα χαρακτηριστικά fMRI να προβλέπονται μόνο μέσω Παλινδρόμησης Κορυφογραμμής, καθώς οι άλλες μέθοδοι είχαν παρόμοια ή χειρότερη απόδοση. Είναι προφανές ότι η συνένωση χειροτερεύει την απόδοση, ενώ η προσθήκη δεν επηρεάζει το μοντέλο με σημαντικό τρόπο.

Σκορ	Βασικό μοντέλο	Παλινδρόμηση (pad, add)	Παλινδρόμηση (project, add)	Παλινδρόμηση (concat)	Παλινδρόμηση (concat, ℓ_2-norm)
Bleu_1	0.740	0.736	0.735	0.730	0.727
Bleu_2	0.578	0.574	0.573	0.566	0.563
Bleu_3	0.445	0.441	0.441	0.433	0.430
Bleu_4	0.343	0.341	0.340	0.332	0.330
METEOR	0.279	0.283	0.283	0.280	0.278
ROUGE_L	0.558	0.561	0.560	0.557	0.553
CIDEr	1.117	1.124	1.124	1.105	1.092
SPICE	0.208	0.212	0.211	0.208	0.206

Πίνακας 6: Απόδοση ενσωμάτωσης fMRI στον κωδικοποιητή για τη δημιουργία λεζάντας εικόνας στο σύνολο δεδομένων MS-COCO.

Drop-net

Μια πιθανή ανεπάρκεια της προηγούμενης προσέγγισης είναι ότι το δίκτυο ενδέχεται να βασίζεται κυρίως στην ισχυρή οπτική τροπικότητα και να αντιμετωπίζει τα πρόσθετα χαρακτηριστικά fMRI ως θόρυβο. Εμπνευσμένοι από το τέχνασμα Drop-net που προτείνεται στο [39], μπορούμε να μετριάσουμε αυτό πρόβλημα αναγκάζοντας το μοντέλο

να χρησιμοποιεί μόνο την οπτική ροή είτε μόνο τη ροή fMRI είτε και τις δύο ταυτόχρονα, με βάση κάποια ομοιόμορφα κατανεμημένη τυχαία μεταβλητή U^l από το διάστημα $[0, 1]$. Για την περίπτωση που τα χαρακτηριστικά fMRI υπερτίθενται στα οπτικά χαρακτηριστικά, η είσοδος στον κωδικοποιητή γίνεται

$$\begin{aligned}\boldsymbol{X}_{i,:} = &\mathbb{I}\left(U^l < \frac{p_{\text{net}}}{2}\right)\boldsymbol{V}_{i,:}\boldsymbol{W}_{\text{visual}} + \mathbb{I}\left(U^l > 1 - \frac{p_{\text{net}}}{2}\right)\hat{\boldsymbol{F}}_{i,:}\boldsymbol{W}_{\text{fmri}} \\ &+ \frac{1}{2}\mathbb{I}\left(\frac{p_{\text{net}}}{2} \leq U^l \leq 1 - \frac{p_{\text{net}}}{2}\right)\left(\boldsymbol{V}_{i,:}\boldsymbol{W}_{\text{visual}} + \hat{\boldsymbol{F}}_{i,:}\boldsymbol{W}_{\text{fmri}}\right) + \boldsymbol{P}_{i,:}\boldsymbol{W}_{\text{pos}}\end{aligned} \tag{13}$$

όπου $\mathbb{I}(\cdot)$ είναι η δείκτρια συνάρτηση και $p_{\text{net}} \in [0, 1]$ είναι ο ρυθμός drop-net. Τα αποτελέσματα εμφανίζονται στον πίνακα 7 για $p_{\text{net}} = 0.8$ και δεν είναι ικανοποιητικά. Αυτό μπορεί να αποδοθεί στο γεγονός ότι η ισχυρή οπτική τροπικότητα δεν χρησιμοποιείται συνεχώς και η αδύναμη τροπικότητα των fMRI δημιουργεί προβλήματα στη διαδικασία εκπαίδευσης.

Σκορ	Βασικό μοντέλο	Παλινδρόμηση (pad, add)	Παλινδρόμηση (project, add)
Bleu_1	0.740	0.632	0.629
Bleu_2	0.578	0.460	0.451
Bleu_3	0.445	0.340	0.336
Bleu_4	0.343	0.256	0.252
METEOR	0.279	0.226	0.219
ROUGE_L	0.558	0.486	0.480
CIDEr	1.117	0.782	0.776
SPICE	0.208	0.154	0.146

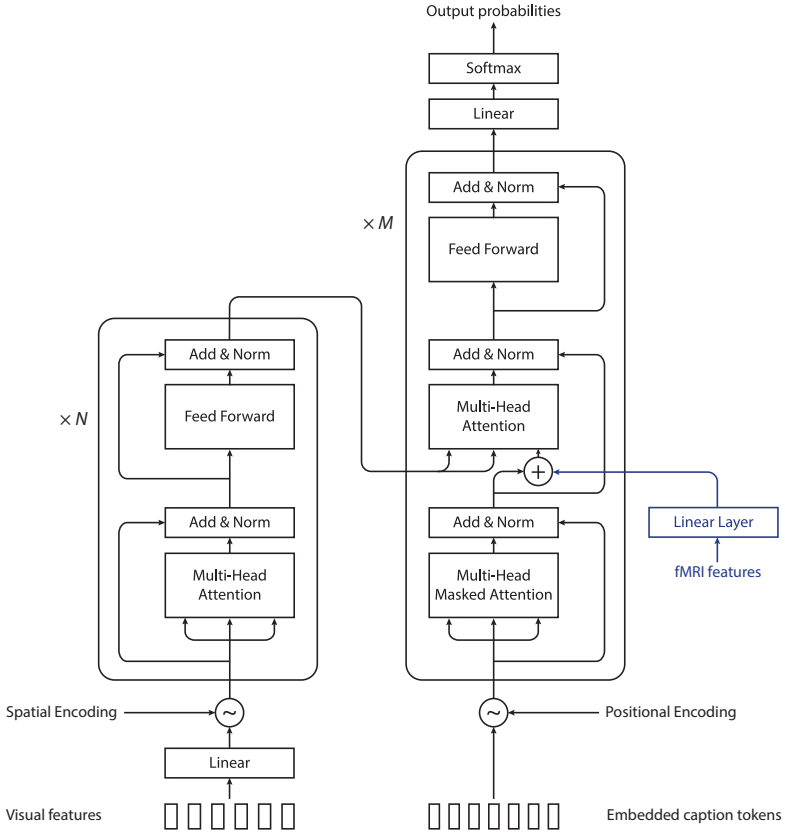
Πίνακας 7: Απόδοση της ενσωμάτωσης fMRI με drop-net ($p_{\text{net}} = 0.8$) στον κωδικοποιητή για δημιουργία λεζάντας εικόνας στο σύνολο δεδομένων MS-COCO.

Συνθηκοθέτηση Προσοχής στον Αποκωδικοποιητή

Μια διαφορετική προσέγγιση για την ενσωμάτωση των χαρακτηριστικών fMRI είναι μέσω συνθηκοθέτησης προσοχής στον αποκωδικοποιητή. Σε αυτή την περίπτωση δεν χρησιμοποιούμε πλέον τα οπτικά χαρακτηριστικά από τα πλαίσια οριοθέτησης, αλλά από ολόκληρη την εικόνα εισόδου. Έστω ότι $\hat{\boldsymbol{f}} \in \mathbb{R}^{1\times 1100}$ είναι τα προβλεπόμενα χαρακτηριστικά fMRI, τότε τα διανύσματα ερωτημάτων $\boldsymbol{Q}_{i,:} \in \mathbb{R}^{1\times 512}$ του αποκωδικοποιητή γίνονται

$$\boldsymbol{Q}'_{i,:} = \boldsymbol{Q}_{i,:} + \hat{\boldsymbol{f}}\boldsymbol{W}_{\text{fmri}} \tag{14}$$

όπου $\boldsymbol{W}_{\text{fmri}} \in \mathbb{R}^{1100\times512}$ είναι ένας εκμαθήσιμος πίνακας που προβάλλει τα fMRI διανύσματα σε έναν χώρο 512 διαστάσεων. Η τροποποιημένη αρχιτεκτονική εμφανίζεται στο σχήμα 6. Για να προσδιορίσουμε εάν τα χαρακτηριστικά fMRI αντιμετωπίζονται ως θόρυβος από το μοντέλο, εκτελέστηκε ένα επιπλέον πείραμα, όπου προστέθηκε τυχαίος θόρυβος στα ερωτήματα. Πράγματι, τα αποτελέσματα που φαίνονται στον πίνακα 8 επιβεβαιώνουν την προηγούμενη υπόθεση.

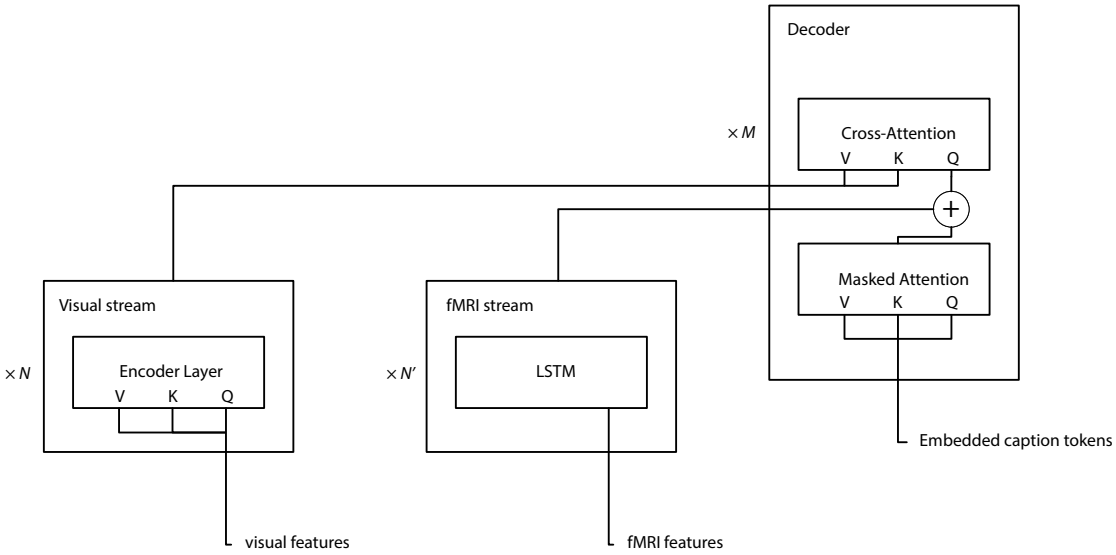


Σχήμα 6: Το προβλεπόμενο διάνυσμα χαρακτηριστικών fMRI $\hat{\boldsymbol{f}}$ που προκύπτει από τα οπτικά χαρακτηριστικά ολόκληρης της εικόνας προβάλλεται γραμμικά σε χώρο 512 διαστάσεων και προστίθεται σε κάθε διάνυσμα ερωτήματος $\boldsymbol{Q}_{i,:}$.

Σκορ	Βασικό μοντέλο	Παλινδρόμηση	Εκμάθηση Λεξικού	Τυχαίο
Bleu_1	0.740	0.737	0.735	0.736
Bleu_2	0.578	0.575	0.574	0.576
Bleu_3	0.445	0.443	0.442	0.443
Bleu_4	0.343	0.341	0.342	0.342
METEOR	0.279	0.283	0.283	0.282
ROUGE_L	0.558	0.562	0.560	0.562
CIDEr	1.117	1.129	1.127	1.127
SPICE	0.208	0.212	0.212	0.210

Πίνακας 8: Απόδοση της δημιουργίας λεζάντας εικόνας για το σύνολο δεδομένων MS-COCO με συνθηκοθέτηση προσοχής στον αποκωδικοποιητή. Σε κάθε διάνυσμα ερωτήματος προστίθεται μια γραμμική προβολή του ίδιου διανύσματος fMRI. Τα fMRI έχουν προβλεφθεί με Παλινδρόμηση Κορυφογραμμής και Εκμάθηση Λεξικού, χρησιμοποιώντας οπτικά χαρακτηριστικά (VGG-16) που εξάγονται από ολόκληρη την εικόνα. Η τελευταία στήλη προέκυψε μετά από πρόσθεση τυχαίου θορύβου αντί χαρακτηριστικών fMRI.

Αντί να χρησιμοποιούμε χαρακτηριστικά fMRI που προβλέπονται από ολόκληρη την εικόνα, μπορούμε να συνδυάσουμε τα χαρακτηριστικά fMRI που προβλέπονται από τα οπτικά χαρακτηριστικά των πλαισίων οριοθέτησης μέσω ενός LSTM. Όπως και προηγουμένως, το αποτέλεσμα αυτής της διαδικασίας προστίθεται στα ερωτήματα του αποκωδικοποιητή. Η τροποποιημένη αρχιτεκτονική εμφανίζεται στο σχήμα 7. Τα αποτελέσματα, που φαίνονται στον πίνακα 9, είναι παρόμοια με αυτά του πίνακα 8, πράγμα που υποδηλώνει ότι τα χαρακτηριστικά fMRI αντιμετωπίζονται ως θόρυβος από το μοντέλο.



Σχήμα 7: LSTM για τη συγκέντρωση των fMRI που προβλέπονται από τα πλαίσια οριοθέτησης και κατόπιν συγχώνευση στα ερωτήματα του αποκωδικοποιητή.

Σκορ	Βασικό μοντέλο	Παλινδρόμηση + LSTM	Λεξικό + LSTM
Bleu_1	0.740	0.735	0.733
Bleu_2	0.578	0.574	0.573
Bleu_3	0.445	0.441	0.440
Bleu_4	0.343	0.341	0.339
METEOR	0.279	0.283	0.282
ROUGE_L	0.558	0.560	0.556
CIDEr	1.117	1.126	1.125
SPICE	0.208	0.212	0.210

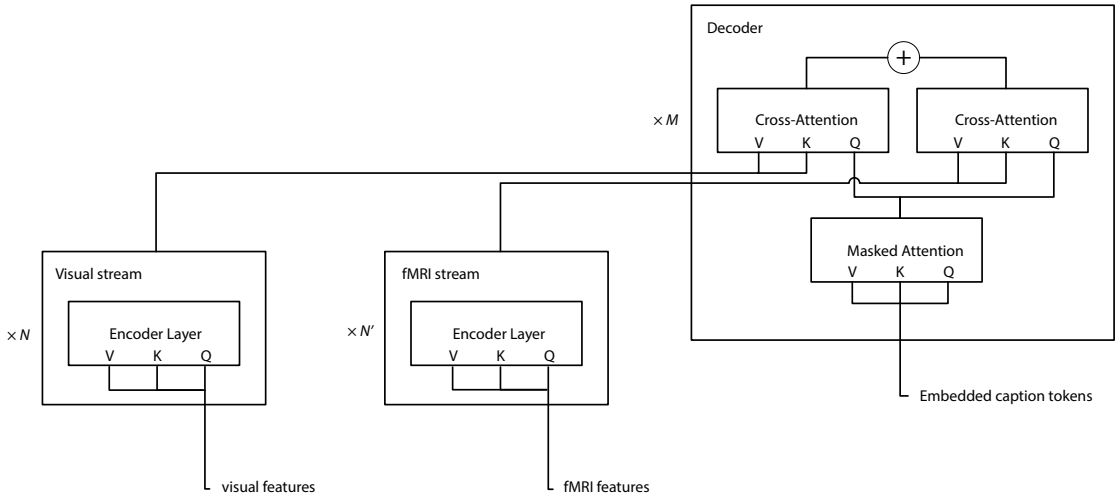
Πίνακας 9: Απόδοση της αρχιτεκτονικής LSTM. Χρησιμοποιήθηκε LSTM τριών στρωμάτων, Παλινδρόμηση Κορυφογραμμής και Εκμάθηση Λεξικού για την λεξική επέκταση από τα οπτικά χαρακτηριστικά που εξήχθησαν από τα πλαίσια οριοθέτησης.

Δύο ξεχωριστοί κωδικοποιητές

Με βάση τις στρατηγικές ενσωμάτωσης πληροφορίας για αρχιτεκτονικές μετασχηματιστών που προτάθηκαν στο [40] προσθέτουμε έναν ξεχωριστό κωδικοποιητή και ένα επιπλέον μπλοκ προσοχής στον αποκωδικοποιητή για τη ροή των fMRI. Η προκύπτουσα αρχιτεκτονική φαίνεται στο σχήμα 8. Η συνολική προσοχή δίνεται από

$$\mathcal{A}_{\text{tot}} = \mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}_{\text{vis}}, \boldsymbol{V}_{\text{vis}}) + \mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}_{\text{fmri}}, \boldsymbol{V}_{\text{fmri}}) \tag{15}$$

όπου $\boldsymbol{K}_{\text{vis}} \equiv \boldsymbol{V}_{\text{vis}}$ είναι η έξοδος του οπτικού κωδικοποιητή και $\boldsymbol{K}_{\text{fmri}} \equiv \boldsymbol{V}_{\text{fmri}}$ είναι η έξοδος του κωδικοποιητή των fMRI.



Σχήμα 8: Δύο ξεχωριστοί κωδικοποιητές με «παράλληλη» ενσωμάτωση στην πλευρά του αποκωδικοποιητή.

Μηδενική είσοδος για fMRI

Σε αυτήν την ενότητα παρουσιάζουμε μια παραλλαγή της αρχιτεκτονικής που συζητήθηκε προηγουμένως η οποία λαμβάνει υπόψη ότι η ροή fMRI ενδέχεται να μην είναι χρήσιμη σε όλες τις περιπτώσεις. Βασίστηκε στην ιδέα μηδενικής εισόδου που προτάθηκε στο [41]. Θεωρώντας b τον αριθμό των πλαισίων οριοθέτησης μιας εικόνας και $\boldsymbol{F} \in \mathbb{R}^{b\times 512}$ την έξοδο του κωδικοποιητή fMRI, προσαρτούμε μια «μηδενική είσοδο» (τυχαίο θόρυβο) $\boldsymbol{n}_\circ$ στον ακόλουθο πίνακα έτσι ώστε

$$\boldsymbol{F}^+ = \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{n}_\circ \end{bmatrix} \tag{16}$$

Στη διασταυρούμενη προσοχή για τα fMRI στον αποκωδικοποιητή έχουμε ότι $\boldsymbol{K} \equiv \boldsymbol{F}^+$. Επομένως

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^\mathsf{T} = \begin{bmatrix} \boldsymbol{Q}\boldsymbol{F}^\mathsf{T} & \boldsymbol{Q}\boldsymbol{n}_\circ^\mathsf{T} \end{bmatrix} \tag{17}$$

Το άθροισμα των σκορ προσοχής για κάθε κλειδί $\boldsymbol{K}_{i,:}$ δίνεται από

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \boldsymbol{A} \tag{18}$$

όπου $\boldsymbol{S} \in \mathbb{R}^{1\times(b+1)}$. Στη συνέχεια, εξετάζουμε τις ακόλουθες δύο περιπτώσεις:

1. Εάν ο μέσος όρος των σκορ που αντιστοιχούν στα κωδικοποιημένα διανύσματα fMRI είναι μεγαλύτερος από το σκορ για την μηδενική είσοδο, η συνολική προσοχή περιλαμβάνει την έξοδο της διασταυρούμενης προσοχής για τα fMRI.

2. Αλλιώς, λαμβάνεται υπόψη μόνο η έξοδος της οπτικής διασταυρούμενης προσοχής.

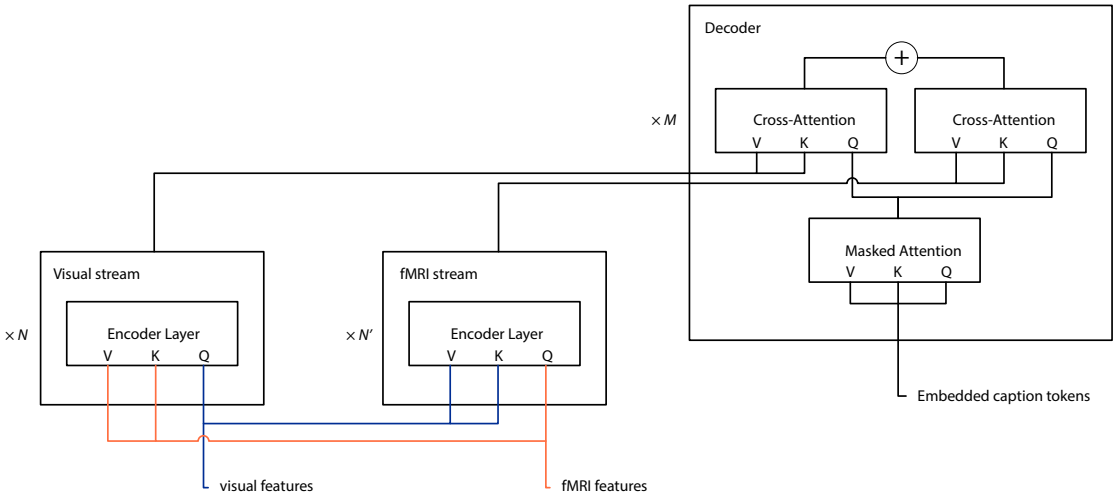
Η συνολική προσοχή $\mathcal{A}_{\text{tot}}$ δίνεται από

$$\mathcal{A}_{\text{tot}} = \begin{cases} \alpha\mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}_{\text{vis}}, \boldsymbol{V}_{\text{vis}}) + (1-\alpha)\mathcal{A}(\boldsymbol{Q}, \boldsymbol{F}^{+}, \boldsymbol{F}^{+}) & \text{if } \frac{1}{b}\sum_{i=1}^{b} \boldsymbol{S}_{1,i} > \boldsymbol{S}_{1,b+1} \\ \mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}_{\text{vis}}, \boldsymbol{V}_{\text{vis}}) & \text{otherwise} \end{cases} \tag{19}$$

όπου $K_{\text{vis}} \equiv V_{\text{vis}}$ είναι η έξοδος του οπτικού κωδικοποιητή και $\alpha \in (0, 1)$.

Δύο κωδικοποιητές με συγχώνευση διασταυρούμενης τροπικότητας

Μια διαφορετική παραλλαγή της αρχιτεκτονικής με τους δύο κωδικοποιητές βασίζεται στο άρθρο [42]. Συγκεκριμένα, ο οπτικός κωδικοποιητής και ο κωδικοποιητής των fMRI επικοινωνούν μέσω στρωμάτων διασταυρούμενης προσοχής, όπως φαίνεται στο σχήμα 9. Τα συγκεντρωτικά αποτελέσματα για τις τρεις παραλλαγές των προηγούμενων ενοτήτων παρουσιάζονται στον πίνακα 10. Για την λεκτική επέκταση χρησιμοποιήθηκε Παλινδρόμηση Κορυφογραμμής, καθώς άλλες μέθοδοι είχαν παρόμοια απόδοση. Στην περίπτωση της συγχώνευσης διασταυρούμενης τροπικότητας τα σκορ είναι ελαφρώς χειρότερα, καθώς η ροή fMRI ενδέχεται να επηρεάζει αρνητικά την περισσότερο σημαντική οπτική ροή. Συνολικά, κανένα από τα μοντέλα δεν παρουσιάζει σημαντικές βελτιώσεις ως προς το βασικό μοντέλο. Κάποιες μικρές αποκλίσεις θα μπορούσαν πιθανώς να αποδοθούν στα fMRI τα οποία ενεργούν ως θόρυβος κανονικοποίησης.



Σχήμα 9: Δύο κωδικοποιητές με συγχώνευση διασταυρούμενης τροπικότητας και «παράλληλη» συγχώνευση στην πλευρά του αποκωδικοποιητή.

Σκορ	Βασικό μοντέλο	Δυο κωδ.	Δυο κωδ. + μηδ. είσοδος	Δυο κωδ. + διασταυρ.
Bleu_1	0.740	0.736	**0.741**	0.736
Bleu_2	**0.578**	0.574	**0.578**	0.572
Bleu_3	**0.445**	0.441	0.444	0.440
Bleu_4	**0.343**	0.341	0.342	0.340
METEOR	0.279	**0.283**	0.281	0.279
ROUGE_L	0.558	**0.561**	0.559	0.556
CIDEr	1.117	**1.124**	1.123	1.122
SPICE	0.208	**0.212**	0.209	0.210

Πίνακας 10: Απόδοση της αρχιτεκτονικής δύο κωδικοποιητών (με και χωρίς την τροποποίηση μηδενικής εισόδου στον αποκωδικοποιητή) και της αρχιτεκτονικής των δύο κωδικοποιητών με συγχώνευση διασταυρούμενης τροπικότητας. Για την λεξική επέκταση από τα οπτικά χαρακτηριστικά που εξάγονται από τα πλαίσια οριοθέτησης, χρησιμοποιείται Παλινδρόμηση Κορυφογραμμής.

Oracle

Όλα τα προηγούμενα πειράματα δεν έδωσαν σημαντικές βελτιώσεις με τη χρήση δεδομένων fMRI. Προκειμένου να εκτιμηθεί εάν υπάρχουν πληροφορίες στα fMRI που είναι δύσκολο να ενσωματωθούν στο μοντέλο δημιουργίας λεζάντας, πραγματοποιούμε ένα πείραμα oracle υπολογίζοντας την καλύτερη δυνατή βαθμολογία CIDEr για κάθε εικόνα σε σχέση με το βασικό μοντέλο και με ένα άλλο από τα προτεινόμενα μοντέλα. Το σκορ

της i-οστής εικόνας δίνεται από

$$s_i = \max(s_i^{(b)}, s_i^{(t)}) \tag{20}$$

όπου ο δείκτης (b) δηλώνει το βασικό μοντέλο και ο εκθέτης (t) δηλώνει το προτεινόμενο μοντέλο. Τα αποτελέσματα παρουσιάζονται στον πίνακα 11. Έχει ενδιαφέρον να σημειωθεί ότι το μοντέλο που χρησιμοποιεί δύο ξεχωριστούς κωδικοποιητές επιτυγχάνει σχεδόν τα ίδια αποτελέσματα όταν εκπαιδεύεται με χρήση χαρακτηριστικών fMRI και όταν εκπαιδεύεται με χρήση οπτικών χαρακτηριστικών και στους δύο κωδικοποιητές. Αυτό υποδηλώνει ότι η βελτίωση που παρατηρείται στην περίπτωση των χαρακτηριστικών fMRI μπορεί να αποδοθεί στα οπτικά χαρακτηριστικά από τα οποία προήλθαν. Το απλό μοντέλο συνθηκοθέτησης προσοχής στον αποκωδικοποίητη, με ένδειξη "fmri add", αποδίδει παρόμοια με το αντίστοιχο "random add", όπου έχει χρησιμοποιηθεί τυχαίος θόρυβος αντί για fMRI, υποδεικνύοντας ότι είναι αναποτελεσματικό στην εκμετάλλευση των πληροφοριών που ενδεχομένως παρέχονται από τα fMRI.

Σκορ	Βασικό μοντέλο	Βασικό + Δύο κωδ. (fMRI)	Βασικό + Δύο κωδ. (visual)	Βασικό + fMRI add	Βασικό + random add
Bleu_1	0.740	0.777	0.777	0.772	0.768
Bleu_2	0.578	0.629	0.629	0.623	0.617
Bleu_3	0.445	0.500	0.500	0.492	0.488
Bleu_4	0.343	0.396	0.396	0.388	0.385
METEOR	0.279	0.303	0.303	0.298	0.297
ROUGE_L	0.558	0.595	0.594	0.587	0.587
CIDEr	1.117	1.290	1.285	1.254	1.246
SPICE	0.208	0.228	0.227	0.223	0.222

Πίνακας 11: Οι καλύτερες δυνατές βαθμολογίες που προκύπτουν συνδυάζοντας το αρχικό μοντέλο (πρώτη στήλη) με άλλα μοντέλα: Δύο κωδικοποιητές (μοντέλο από το σχήμα 8), "fmri add" (προσθήκη ενός καθολικού διανύσματος fMRI στα ερωτήματα του αποκωδικοποιητή), "random add" (προσθήκη ενός τυχαίου διανύσματος στα ερωτήματα του αποκωδικοποιητή).

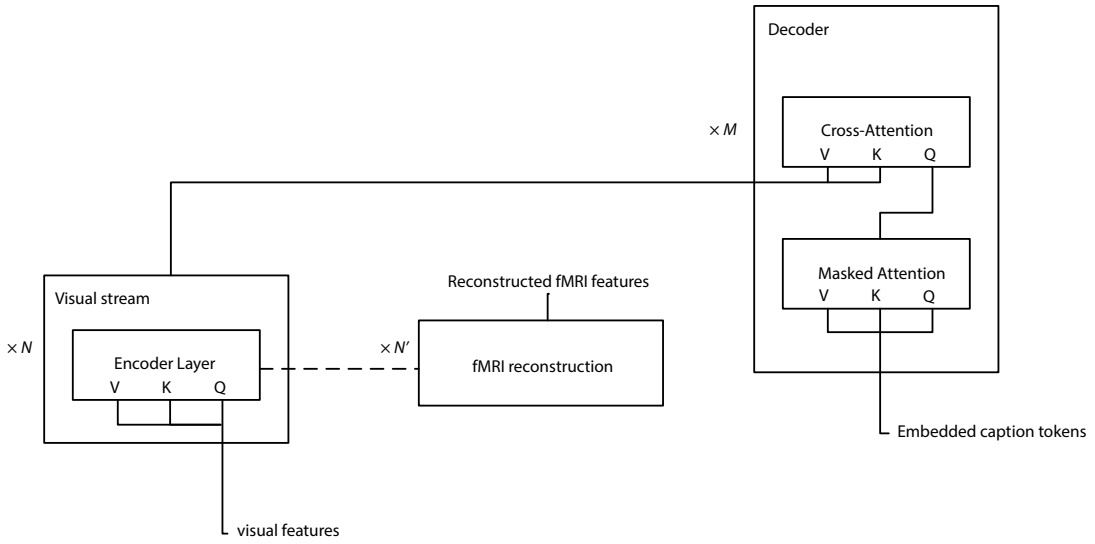
Ανακατασκευή fMRI

Τέλος, μια άλλη ιδέα που αξίζει να ερευνηθεί είναι η χρήση βοηθητικής διαδικασίας πρόβλεψης χαρακτηριστικών fMRI αξιοποιώντας τις αναπαραστάσεις ενός αρχικού στρώ-

ματος του κωδικοποιητή. Η συνολική απώλεια του μοντέλου γίνεται

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \alpha \mathcal{L}_{\text{fmri}} \tag{21}$$

όπου $\mathcal{L}_{\text{cap}}$ είναι η εξομαλυμένη διασταυρούμενη εντροπία που σχετίζεται με την δημιουργία λεζάντας, $\mathcal{L}_{\text{fmri}}$ είναι η απώλεια παλινδρόμησης που σχετίζεται με την ανακατασκευή των χαρακτηριστικών fMRI και το $\alpha > 0$ είναι ρυθμιστική παράμετρος. Το σχήμα 10 απεικονίζει αυτή την αρχιτεκτονική και ο πίνακας 12 δείχνει τα αντίστοιχα αποτελέσματα. Παρατηρούμε ότι η απόδοση είναι παρόμοια με το βασικό μοντέλο. Αυτό μπορεί να αποδοθεί είτε στο γεγονός ότι η εργασία ανακατασκευής fMRI υπερπροσαρμόζεται και γενικεύεται με διαφορετικό ρυθμό [43] είτε στις αδύναμες αναπαραστάσεις των fMRI.



Σχήμα 10: Ανακατασκευή fMRI

Σκορ	Βασικό μοντέλο	Παλινδρόμηση Κορυφογραμμής	Εκμάθηση Λεξικού	Ομοιότητα
Bleu_1	0.740	0.734	0.735	0.737
Bleu_2	0.578	0.572	0.572	0.576
Bleu_3	0.445	0.440	0.441	0.443
Bleu_4	0.343	0.339	0.340	0.342
METEOR	0.279	0.283	0.282	0.279
ROUGE_L	0.558	0.559	0.559	0.558
CIDEr	1.117	1.120	1.120	1.121
SPICE	0.208	0.211	0.210	0.208

Πίνακας 12: Απόδοση για την δημιουργία λεζάντας εικόνας στο σύνολο δεδομένων MS-COCO, όπου πραγματοποιείται ανακατασκευή fMRI από τις αναπαραστάσεις του πρώτου στρώματος του κωδικοποιητή. Τα fMRI έχουν προβλεφθεί με Παλινδρόμηση Κορυφογραμμής, με Εκμάθηση Λεξικού και με Ομοιότητα, χρησιμοποιώντας οπτικά χαρακτηριστικά που έχουν εξαχθεί από τα πλαίσια οριοθέτησης της εικόνας.

Συμπεράσματα

Παρόλο που υπάρχει συσχέτιση μεταξύ ενεργοποιήσεων fMRI και οπτικών αναπαραστάσεων και, σε μικρότερο βαθμό, συσχέτιση με αναπαραστάσεις λέξεων, το σήμα fMRI περιέχει αρκετό θόρυβο και η ενσωμάτωσή του σε υπολογιστικά μοντέλα είναι δύσκολη διαδικασία, δεδομένου ότι τις περισσότερες φορές δεν βελτιώνεται η απόδοση. Μεταξύ των μέθοδων που δοκιμάστηκαν και αξιολογήθηκαν για την κωδικοποίηση οπτικών αναπαραστάσεων σε ενεργοποιήσεις fMRI, η εκμάθηση αραιού λεξικού φαίνεται να υπερτερεί της παλινδρόμησης κορυφογραμμής και της μεθόδου που βασίζεται στην ομοιότητα. Ωστόσο, η μέθοδος αξιολόγησης που χρησιμοποιείται στην βιβλιογραφία είναι αρκετά απλή και δεν είναι εγγυάται απαραίτητα ότι οι προκύπτουσες προβλεπόμενες ενεργοποιήσεις fMRI είναι υψηλής ποιότητας. Εκτός αυτού, τα μοντέλα που ενσωματώνουν fMRI μέσω της παλινδρόμησης της κορυφογραμμής φαίνεται να έχουν ελαφρώς καλύτερη απόδοση, κάτι που μπορεί να υποδηλώνει ότι ο προκύπτων χώρος fMRI μοιράζεται περισσότερες ιδιότητες με τον οπτικό χώρο από τον οποίο προήλθαν, συγκριτικά με τις άλλες μεθόδους. Όσον αφορά τις τεχνικές που χρησιμοποιήθηκαν για την ενσωμάτωση ενεργοποιήσεων fMRI, τόσο η ενσωμάτωση στον κωδικοποιητή, όσο και η συνθηκοθέτηση προσοχής στον αποκωδικοποιητή δίνουν παρόμοια αποτελέσματα, υποδηλώνοντας ότι τα fMRI ενεργούν κυρίως ως θόρυβος κανονικοποίησης. Όταν προστίθεται ξεχωριστός κωδικοποιητής μετασχηματιστή για τα fMRI, ο οποίος επικοινωνεί με τον οπτικό κωδικοποιητή μέσω προσοχής διασταυρούμενης τροπικότη-

τας, η απόδοση υποβαθμίζεται καθώς η ροή fMRI ενδέχεται να επηρεάζει αρνητικά την οπτική ροή. Οι προσαρμοστικές μέθοδοι, όπως το DropNet, που δεν βασίζονται πάντα στην οπτική ροή, αλλά επιτρέπουν την αποκλειστική χρήση της ροής fMRI, έτσι ώστε να μην αγνοείται από το μοντέλο, δεν έχουν καλή απόδοση. Έτσι, είναι προφανές ότι η οπτική ροή πρέπει να χρησιμοποιείται πάντοτε και η ροή fMRI μόνο όταν είναι είναι πιθανό ότι θα συνεισφέρει χρήσιμες πληροφορίες. Αυτό επιχειρήθηκε στο πείραμα της «μηδενικής εισόδου», το οποίο ωστόσο δεν επέφερε κάποια σημαντική βελτίωση της απόδοσης. Μια άλλη ενδιαφέρουσα παρατήρηση προέκυψε στο πείραμα oracle, που έδειξε ότι οι ενεργοποιήσεις fMRI δεν δίνουν κάποια βελτίωση σε σχέση με τις οπτικές αναπαραστάσεις από τις οποίες προήλθαν. Αυτό υποδηλώνει ότι η διαδικασία κωδικοποίησης fMRI παρουσιάζει προβλήματα και ιδανικά θα πρέπει να βελτιωθεί ή να παρακαμφθεί μέσω μεγαλύτερων συνόλων δεδομένων fMRI.

Μελλοντικές επεκτάσεις

Λόγω του θορύβου που υπάρχει στο σήμα fMRI και της μεταβλητότητας της εγκεφαλικής ανατομίας και της λειτουργικής απόκρισης σε διαφορετικά υποκείμενα, είναι δύσκολο να εξαχθούν ισχυρές αναπαραστάσεις fMRI για χρήση σε υπολογιστικά μοντέλα. Ένας τρόπος για να μετριαστεί αυτό το πρόβλημα είναι να αναπαρασταθούν δεδομένα fMRI από πολλά υποκείμενα σε έναν κοινό χώρο [44–46]. Επιπλέον, θα μπορούσαν να ερευνηθούν διαφορετικές τεχνικές επιλογής voxel [38, 47–49], προκειμένου να επιλεγούν πιο αντιπροσωπευτικά voxel που να έχουν ως αποτέλεσμα καλύτερες γνωσιακές αναπαραστάσεις για υπολογιστικές εργασίες σε επόμενα στάδια. Επιπλέον, διαφορετικά σύνολα δεδομένων θα μπορούσαν να συγκεντρωθούν μαζί για να επιτρέψουν προεκπαίδευση σε δεδομένα fMRI με αρχιτεκτονικές δικτύων που χρησιμοποιούνται στην προεκπαίδευση μοντέλων όρασης και γλώσσας [50, 51]. Αυτή η διαδικασία θα μπορούσε να οδηγήσει σε βελτιωμένες αναπαραστάσεις όρασης και γλώσσας που είναι γνωσιακά βασισμένες και που μπορούν να χρησιμοποιηθούν σε άλλες εργασίες, όπως για παράδειγμα στην δημιουργία περιγραφών εικόνων. Επιπλέον, άλλες προσαρμοστικές τεχνικές ενσωμάτωσης μπορούν να εφαρμοστούν βάσει της παραδοχής ότι οι εγκεφαλικές ενεργοποιήσεις μπορούν να οδηγήσουν σε βελτίωση μόνο σε συγκεκριμένες περιπτώσεις και όχι καθολικά. Υπό αυτό το πρίσμα, θα ήταν ενδιαφέρον να ερμηνευθεί το είδος των πληροφοριών που παρέχουν οι ενεργοποιήσεις του εγκεφάλου. Αυτό θα μπορούσε ενδεχομένως να βοηθήσει στην βελτίωση της διαδικασίας ενσωμάτωσης.

CHAPTER 1

Introduction

1.1 Motivation

The seminal work of Mitchell *et al.* [1] demonstrated that fMRI signals encode meaningful semantic information for concrete nouns, which can be effectively used to map between distributed semantic representations and fMRI activations. This was the first computational model to predict brain patterns associated with unknown words. Many others have since attempted to extend this initial work [2–5], and the use of cognitive data in computational models remains an open field of research. With the availability of datasets where images were used as stimuli [6, 7], cognitive data can be used as an extra input modality to models that operate on images, such as image captioning models. Due to the semantic information inherent in fMRI activations, it may be possible that generated captions will be more cognitively plausible, resulting in a subsequent performance improvement.

1.2 Contributions

The main contributions of this thesis are the following:

- Assessment of the representational similarity between visual or word representations and fMRI activations.
- Evaluation of different methods for mapping visual stimuli to fMRI activations.
- Implementation of several architectures for the incorporation of fMRI activations to a transformer based image captioning model.
- Usage of techniques that allow a more adaptive fusion of the weak modality of fMRI activations.

1.3 Thesis organization

Chapter 2 and chapter 3 provide background material in machine learning and computational neuroscience respectively, that serves as a basis for subsequent sections. Specifically, chapter 2 introduces techniques that are used to map visual representations to fMRI activations, and deep learning models such as transformers and Faster-RCNN that form the foundation of the image captioning model. Chapter 3 describes the fMRI signal in more detail, its preprocessing steps, voxel selection methods, as well as encoding models that are commonly used to map e.g. word or image representations to fMRI activations and decoding models that perform the opposite process. In chapter 4 different methods for encoding to fMRI are evaluated and several different architectures are proposed for the fusion of fMRI activations to the image captioning model. Finally, chapter 5 provides discussion of the results, as well as future directions.

CHAPTER 2

Machine Learning Background

2.1 Introduction

Machine Learning (ML) is a subset of Artificial Intelligence that studies algorithms which improve automatically by using data [52]. Essentially, these algorithms build a model from the training data and are able to make decisions or predictions without explicit programming [53]. Various approaches are employed in order to develop ML computer algorithms that are able to accomplish tasks where there are no satisfactory alternatives. One approach involves labelling training data, which are subsequently used by the algorithm to improve the underlying model. For example, a system that automatically recognizes hand-written digits would be trained on the MNIST dataset which has labelled examples of digits [54]. Machine Learning is used in a variety of domains, such as speech recognition, computer vision, natural language processing and robotics, where the development of equivalent conventional algorithms would be often difficult or even unfeasible [34, 55–58].

Depending on the nature of the input available to a learning system, ML algorithms are divided into four broad categories. In *supervised learning*, the goal is to learn a mapping from inputs to outputs by training on labelled examples where the inputs and the desired outputs are provided. On the contrary, in *unsupervised learning*, no labels are provided and the goal of the learning algorithm is to uncover hidden patterns on the input. *Semi-supervised learning* is at the intersection of supervised and unsupervised learning, where unlabeled data are used in combination with a small amount of labeled data. In *reinforcement learning*, a learning agent interacts with its environment, in which it must accomplish a certain goal e.g. play a game against an opponent or drive a vehicle. The agent essentially tries to maximize the rewards while it navigates the problem space [59].

2.2 Regression

Linear regression is a linear method that models the relationship between a scalar response (dependent) variable and one or more explanatory (independent) variables. When there is one explanatory variable, simple linear regression is performed, whereas for more than one, multiple linear regression is performed [60]. In multivariate linear regression, multiple correlated response variables are predicted instead of a single scalar variable [61]. Relationships between variables are modeled using linear models whose unknown parameters are estimated from the data. Moreover, the conditional mean of the response variable given the explanatory variables is usually assumed to be an affine function of these variables.

Given a dataset $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n examples, a linear regression model makes the assumption that the relationship between the dependent variable y_i and the vector of regressors $\boldsymbol{x_i}$ is linear, given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.1}$$

where ε_i is an error term. In a more compact notation, these n equations are often written as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

where

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \tag{2.3}$$

and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$. In practice, a constant term is included as one of the regressors e.g. $\mathbf{x}_{i0} = 1$ for $i = 1, \ldots, n$, and the corresponding element of $\boldsymbol{\beta}$ is the intercept. To fit a linear model to a given dataset, the regression coefficients $\boldsymbol{\beta}$ need to be estimated such that the error term $\boldsymbol{\varepsilon} = \mathbf{y} - X\boldsymbol{\beta}$ is minimized. As a measure of minimization,

usually the sum of squared errors $\|\varepsilon\|_2^2$ is used, so that

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 \tag{2.4}$$

2.2.1 Ridge Regression

Ridge regression is often used in models with many parameters in order to moderate the problem of multicollinearity where independent variables are highly correlated [62]. It is very similar to linear regression, except that the measure of minimization includes a regularization term [63] and the $\boldsymbol{\beta}$ coefficients are given by

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{2.5}$$

where $\lambda \geq 0$ is a tuning parameter that is determined separately. The term $\lambda\|\boldsymbol{\beta}\|_2^2$ is called a *shrinkage penalty* that has the effect of shrinking the elements of $\boldsymbol{\beta}$ towards zero. The parameter λ controls the relative impact of the regularization. When $\lambda = 0$, ridge regression produces least squares estimates that may lead to overfitting and as $\lambda \to \infty$ the shrinkage penalty grows and ridge regression gives coefficient estimates close to zero, leading to underfitting. For each value of λ, a different estimate $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ is produced. Thus, selecting an appropriate value for λ is important and often cross-validation is used.

2.3 Sparse Dictionary Learning

Sparse Dictionary Learning, also called sparse coding, is a method that aims to find a sparse representation of the input data as a linear combination of basis elements (*atoms*) which compose a *dictionary*. These atoms can be an overcomplete spanning set and they do not need to be orthogonal, allowing for more flexible dictionaries and better data representations. Dictionary Learning is useful for signal denoising since usually the interesting part of an input signal can be represented in a sparse way, in contrast to the noise that has a much less sparse representation [64].

For a dataset $X = \{x_1, \ldots, x_K\}$, $x_i \in \mathbb{R}^d$ we need to find a dictionary $\mathbf{D} = \{d_1, \ldots, d_n\}$, $d_i \in \mathbb{R}^d$ and a representation $R = \{r_1, \ldots, r_K\}, r_i \in \mathbb{R}^n$ where the norm $\|X - \mathbf{D}R\|_F^2$ is minimized and the representations r_i are sparse. Mathematically,

this can be written as

$$\underset{\mathbf{D}\in\mathcal{C},r_i\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{K} \|x_i - \mathbf{D}r_i\|_2^2 + \lambda \|r_i\|_0 \tag{2.6}$$

where

$$\mathcal{C} \equiv \left\{\mathbf{D} \in \mathbb{R}^{d\times n} : \|d_i\|_2 \leq 1 \,\forall i = 1,\ldots,n\right\}, \lambda > 0. \tag{2.7}$$

The constraint $\mathbf{D} \in \mathcal{C}$ prevents the atoms from reaching arbitrarily high values, allowing arbitrarily low values for the represenations r_i. The tuning parameter λ compromises between the sparsity condition and the minimization of error. Due to the ℓ_0-norm, the problem is not convex and belongs to the NP-hard class [65]. If the ℓ_1-norm is used instead to ensure sparsity [66], the problem becomes convex with respect to $\mathbf{D}$ when R is fixed and vice versa, but not jointly convex in $(\mathbf{D}, R)$.

A dictionary $\mathbf{D}$ is said to be undercomplete when $n < d$ and overcomplete when $n > d$ which is the typical case. The case $n = d$ is not considered, since the resulting dictionary does not have any representational benefit. Undercomplete dictionaries are related to dimensionality reduction, while overcomplete dictionaries allow for richer data representations. The dictionary itself can be a transform matrix e.g. wavelets transform, or it can be learned so that the input is sparsely represented in an optimal way.

2.4 Principal Component Analysis

Principal Component Analysis (PCA) is an orthogonal linear transformation which transforms data to a new set of variables (principal components) that are uncorrelated and ordered in such a way so as to retain as much as possible the variation present in the dataset [67]. The largest variance through a scalar projection of the data corresponds to the first principal component, the second largest to the second component, et cetera.

Given a dataset of n examples $\mathbf{X} \in \mathbb{R}^{n\times p}$ with zero mean column-wise, the transformation is defined by a set of l vectors $\mathbf{w}_{(k)} = (w_1,\ldots,w_p)_{(k)}$, $\mathbf{w}_{(k)} \in \mathbb{R}^p$ that map each row vector $\mathbf{x}_{(i)}$ of $\mathbf{X}$ to a vector of principal component scores $\mathbf{t}_{(i)} = (t_1,\ldots,t_l)_{(i)}$ where $t_{k(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)}$ for $i = 1,\ldots,n$ and $k = 1,\ldots,l$, so that the individual elements $t_1,\ldots,t_l$ successively inherit the maximum possible variance from $\mathbf{X}$, where each $\mathbf{w}$ is

constrained to be a unit vector. The first weight vector $\mathbf{w}_{(1)}$ needs to satisfy

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w} \right)^2 \right\} \tag{2.8}$$

in order to maximize variance. Then, the first principal component is given by $\left(\mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}\right) \mathbf{w}_{(1)}$. The remaining weight vectors are given by

$$\mathbf{w}_{(k)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \left\| \hat{\mathbf{X}}_k \mathbf{w} \right\|^2 \right\} = \arg\max \left\{ \frac{\mathbf{w}^\top \hat{\mathbf{x}}_k^\top \hat{\mathbf{x}}_k \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\} \tag{2.9}$$

where

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X} \mathbf{w}_{(s)} \mathbf{w}_{(s)}^\top \tag{2.10}$$

The full decomposition of $\mathbf{X}$ can be written as $\mathbf{T} = \mathbf{XW}$, where the columns of $\mathbf{W}$ are the weight vectors and map each vector $\mathbf{x}_{(i)}$ from an original space of p variables to a new space of p variables that are uncorrelated over the dataset. An example for a synthetic dataset is shown in figure 2.1. By keeping only the first L principal components we can achieve *dimensionality reduction* and get a truncated transformation, where vectors in the transformed space have only L variables.

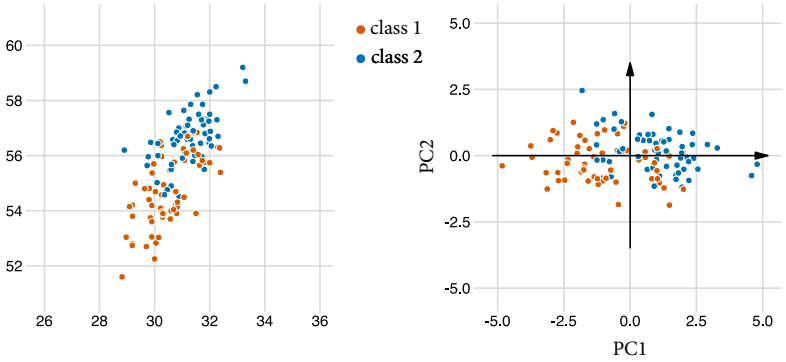


Figure 2.1: Principal Component Analysis. Synthetic dataset (left) and transformed version on the coordinate system defined by the two principal components.

2.5 Deep Learning

2.5.1 Feedforward Neural Networks

Feedforward Neural Networks, also often called Multilayer Perceptrons (MLPs), are the cornerstone of deep learning models. Their goal is to approximate some function f^* by defining a mapping from the input $\boldsymbol{x}$ to the output $\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{\theta})$ and learning parameters $\boldsymbol{\theta}$ which result in the best approximation [68]. Information flows only in the forward direction, through intermediate computation layers that define the output $\boldsymbol{y}$ and there are no feedback loops. The network can be modeled through a directed acyclic graph that describes how functions are chained together to form the output $f(\boldsymbol{x}) = (f^{(n)} \circ f^{(n-1)} \cdots \circ f^{(1)})(\boldsymbol{x})$, where $f^{(i)}$ represents the i-th layer of the network. The length of this chain is the depth of the network.

The goal of training is to drive $f(\boldsymbol{x})$ close to $f^*(\boldsymbol{x})$ by using approximate examples of $f^*(\boldsymbol{x})$ that are evaluated at different training examples $\boldsymbol{x}$. Given these training examples, the output layer of the network, which is the final one, must produce a value that is near to $\boldsymbol{y} \approx f^*(\boldsymbol{x})$. The behavior of the previous layers is not governed directly by the training data, but by the learning algorithm which decides how to use them in order to implement an approximation of $f^*(\boldsymbol{x})$ in the best possible way. These layers are called *hidden* since they are not directly observable.

Each hidden layer can be seen as a vector whose elements represent a neuron. These neurons act in parallel, performing a vector-to-scalar mapping, where they receive input from neurons of the previous layer and compute their own activation value. This concept draws inspiration from neuroscience and the functions $f^{(i)}(\boldsymbol{x})$ of each layer are based upon neuroscientific observations regarding the computation that actual biological neurons perform. However, the network does not attempt to be a model of the brain but rather it draws several insights from it. A typical model of an artifical neuron is illustrated in figure 2.2 and a feedforward network with two hidden layers is shown in figure 2.3.

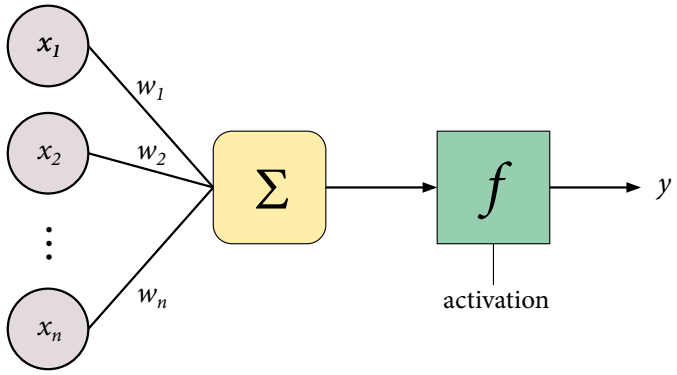


Figure 2.2: Schematic diagram of an artificial neuron. Each input element x_i is multiplied by a weight coefficient w_i. The resulting terms are added together and an activation function is applied to the sum.

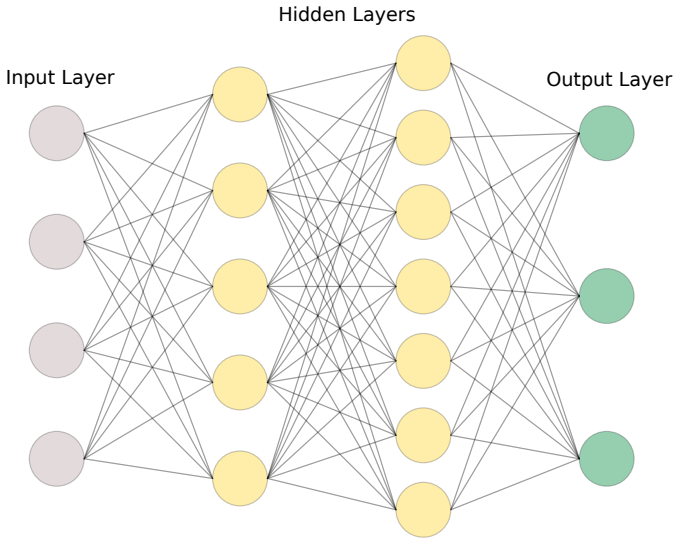


Figure 2.3: Feedforward Neural Network with two hidden layers.

To overcome the limited capacity of linear models that cannot capture interactions between input variables, FNNs essentially transform the input $\boldsymbol{x}$ via a non-linear learnable mapping ϕ, so that $f(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{w}) = \phi(\boldsymbol{x}; \boldsymbol{\theta})^\top \boldsymbol{w}$ in the case of one hidden layer. The parameter vector $\boldsymbol{\theta}$ is used to learn ϕ from a broad family of functions and $\boldsymbol{w}$ maps the transformed input linearly to the target output. The training problem is not convex anymore, but the optimization algorithm can usually find the $\boldsymbol{\theta}$ that results in a good representation. In addition, engineers can encode knowledge by designing function families

$\phi(\boldsymbol{x}; \boldsymbol{\theta})$ that are expected to perform well in certain domains without the need to choose exactly the right function.

Feedforward Neural Networks have been applied successfully in many applications and apart from good experimental results, they also have important theoretical properties. The universal approximation theorem for arbitrary width and bounded depth neural networks states that every well-behaved function that maps intervals of real numbers to some output interval of real numbers can be approximated with arbitrary precision by a feedforward network with only one hidden layer and a sufficient number of neurons. Moreover, a suitable activation function is required and one of the first proofs [69] uses a sigmoid function.

2.5.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [70, 71] are regularized versions of multilayer perceptrons which are not as prone to overfitting and their training is computationally less demanding. They are cognitively motivated by orientation-selective and locally sensitive neurons found in the visual cortex of animals [72] and each neuron in CNNs receives input from a corresponding neighborhood of the previous layer, as shown in figure 2.4. Unlike feed forward networks that rely on matrix multiplication, CNNs are based on convolution, which is defined as

$$(I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n) \tag{2.11}$$

where I is a 2-D matrix that represents the input and K is a weighting 2-D matrix that represents the kernel (filter) whose dimensions are much smaller than the input. The resulting output is commonly referred to as a feature map. The definition in equation (2.11) is cross-correlation strictly speaking. However, it is equivalent to a convolution due to the fact that the kernel is learnable. Intuitively, the kernel can be seen as a detector that finds patterns at different parts of the input and these patterns are assembled hierarchically to form more complex patterns at subsequent layers.

Convolutional layers perform convolution between their input and a set of kernels and the resulting feature maps are passed from a non-linear activation function. One commonly used such function is a Rectified Linear Unit (ReLU) [73]. Then, a pooling operation is performed where the feature maps are downsampled by summarizing

nearby features. This process is similar to a convolution in that it operates in local neighborhoods but with a hard-coded function instead of a learnable filter. Pooling is essential as it allows subsequent layers to process representations which capture increasing fractions of the input. It also decreases the size of the feature maps, thus reducing the total parameters of the model. A commonly used pooling operation is max-pooling which selects the maximum value within a fixed-size neighborhood [74].

Another property that arises from pooling operations is invariance to small-scale transformations of the input. For example, the pooled representations are translation invariant since they aggregate information over significant regions of the input. This is particularly useful in cases where we are interested in the presence or absence of a pattern and not in its exact location. By pooling between different feature maps that have been created using different kernels (filters), max-pooling can result in invariance to rotations [68].

Convolutional Neural Networks exhibit three important properties that set them apart from feedforward networks: *parameter sharing*, *sparse connectivity* and *equivariance*. Parameter sharing relies on the assumption that a pattern should be able to be detected at different positions by reusing the same parameters. This is enforced by the convolution operation, where the kernel K can be seen as a filter that slides across the input. In this way, the number of parameters is greatly diminished, reducing overfitting and computational costs. Sparse connectivity means that each neuron is connected to a local region of neurons that has a fixed size and is referred to as the receptive field of that neuron. This prevents the number of parameters from increasing substantially for high dimensional inputs. Equivariance is granted by the convolution operation and it implies that translating the input leads to a translation of the output feature map. This allows kernels to detect patterns across the input.

Due to the reasons described, CNNs have become very popular especially for computer vision tasks [56, 75]. The large reduction of training parameters allowed for very deep models that when trained on large-scale datasets surpassed conventional computer vision methods by a large margin.

2.5.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a family of neural networks that process sequential input data. They are different from feedforward neural networks in that they do not

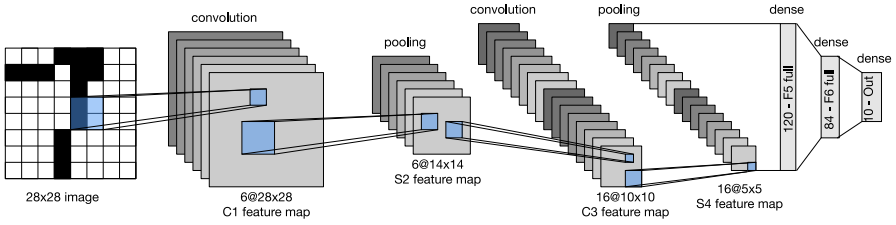


Figure 2.4: Convolutional Neural Network architecture. Figure reproduced from [76].

operate under the assumption that all training instances are independent and in that they produce their output taking into consideration accumulated contextual information from previous instances [68]. This is achieved through a feedback loop that allows information to flow from one timestep to the next, as shown in figure 2.5. For an in-

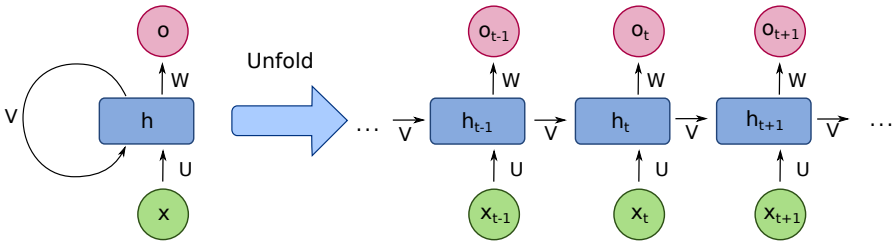


Figure 2.5: Recurrent Neural Network unit. Compact diagram (left) and unfolded version (right). Figure reproduced from [77].

put sequence x_1, x_2,..., x_t, the hidden state h_t and the output o_t at time t of an Elman RNN [78], which is one of the simplest variations, is given by

$$\begin{aligned} h_t &= \sigma_h \left(U x_t + V h_{t-1} + b_h \right) \\ o_t &= \sigma_y \left(W h_t + b_y \right) \end{aligned} \tag{2.12}$$

where σ_h, σ_y are activation functions, U, V, W are weight matrices and b_y, b_h are bias vectors. Since the same weight matrices and bias vectors are used for each time-step, the number of parameters of the model does not depend on the sequence length. Moreover, this allows information to be extracted from different positions of the input sequence, thus improving generalization.

Despite the above advantages of RNNs, in practice it is very difficult for gradient based learning algorithms to optimize them as the duration of dependencies that have to be captured increases [79]. The gradients need to be propagated across the depth of the model through the unfolded computation graph and this can result in either "explod-

ing" or "vanishing" gradients. A simple heuristic that can be used to solve the problem of "exploding" gradients is *gradient clipping* where gradients are clipped when they exceed a certain threshold [80]. To deal with the problem of "vanishing" gradients, a soft constraint can be imposed through a regularization term so that there is a preference for parameter values such that gradients neither increase or decrease through backpropagation [80]. In addition, other variants of RNNs can be used, such as Long Short-Term Memories (LSTMs) which specifically avoid the aforementioned problem [81].

Long Short-Term Memories (LSTMs). Long Short-Term Memories (LSTMs) are a special type of RNN that was introduced in order to cope with "vanishing" gradients during the training process [81]. This allowed learning long-term dependencies and consequently their successful application in a wide variety of tasks, such as speech recognition, image captioning and text classification [82–84]. An LSTM cell, which is the building block of LSTM networks, is shown in figure 2.6. It comprises three gating units that control the flow of information, either by preserving it to the internal states or by "forgetting" it. The *forget gate* f_t decides what information should be kept or forgotten according to

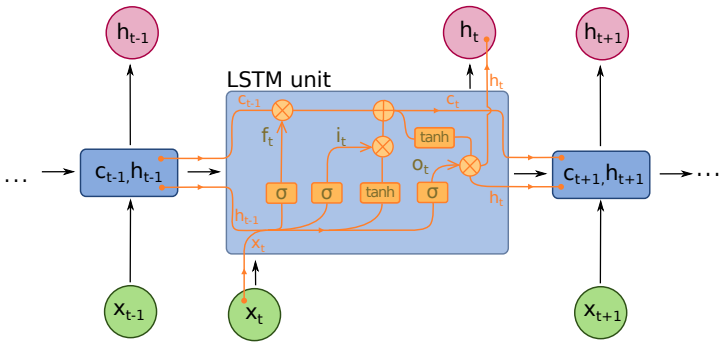


Figure 2.6: Schematic for an LSTM cell. Figure adapted from [77].

the previous hidden state h_{t-1} and the current input x_t. Values of f_t closer to zero imply "forgetting" while values closer to one imply keeping information to the cell state. Similarly, the *input gate* i_t determines what input values will be propagated to the new cell state. The *output gate* o_t controls the exposure of the cell state to the hidden state based on the current input and the previous hidden state. The *cell state* c_t contains internal information and is updated based on past information filtered via the forget gate and on current information filtered via the input gate. Finally, the *hidden state* h_t encodes the

input sequence until time-step t. Formally, the equations that define the operation of an LSTM cell are

$$\begin{aligned} f_t &= \sigma_g \left(W_f x_t + U_f h_{t-1} + b_f \right) \\ i_t &= \sigma_g \left(W_i x_t + U_i h_{t-1} + b_i \right) \\ o_t &= \sigma_g \left(W_o x_t + U_o h_{t-1} + b_o \right) \\ \tilde{c}_t &= \sigma_c \left(W_c x_t + U_c h_{t-1} + b_c \right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot \sigma_h \left(c_t \right) \end{aligned} \tag{2.13}$$

where σ_g is the sigmoid function, σ_c is the hyperbolic tangent function, and σ_h can be either the hyperbolic tangent function or the identity function [85]. The operator $\odot$ denotes element-wise multiplication (Hadamard product), W_*, U_* are weight matrices and b_* are bias vectors that are learned during training.

2.5.4 Transformers

Transformers is a deep learning encoder-decoder architecture for sequence transduction that is based on an attention mechanism instead of recurrences [34]. It achieved state-of-the-art performance on the WMT 2014 English-to-French translation task and surpassed the best performing models on the WMT 2014 English-to-German translation task. Unlike recurrent models that process input sequentially by generating each hidden state as a function of the current input and the previous hidden state, transformers process the input in parallel due to the attention mechanism, requiring less training time. The overall architecture is displayed on figure 2.7. The encoder, shown on the left, consists of N layers and each layer includes a multi-head self attention block and a position-wise fully connected feed-forward network (FFN). The decoder, on the right, has N layers and each of these comprises a masked multi-head attention block that operates on the output embeddings, a multi-head self attention block and a position-wise fully connected FFN. The masking on the first block is necessary in order to prevent attending to future positions. In addition, the output embeddings are shifted right by one position so that the predictions for a specific position are conditioned only on the known outputs for previous positions.

The attention mechanism that is used is called “scaled dot product” attention and is

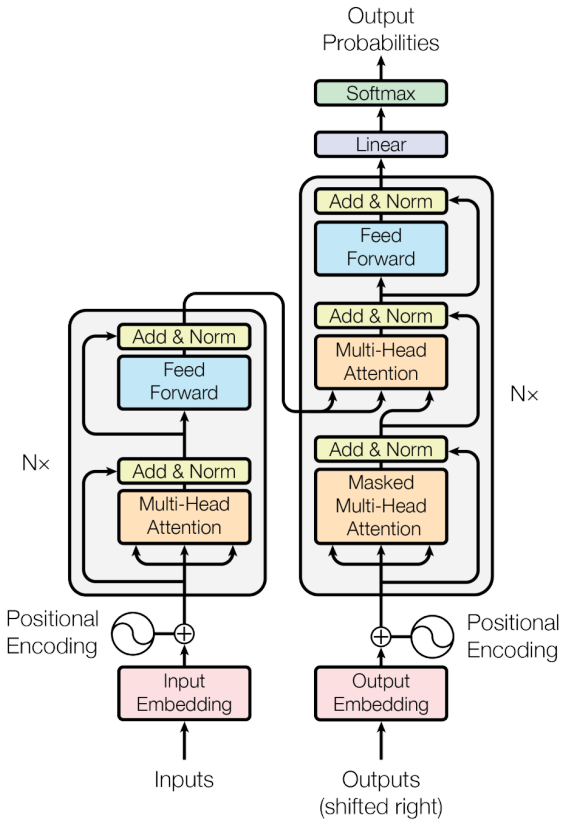


Figure 2.7: Transformer architecture. Figure reproduced from [34].

a modified version of dot product attention. Letting Q, K, V be the query, key and value matrices respectively, the attention matrix is given by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.14}$$

where d_k is the dimension of the queries and keys. The scaling factor $1/\sqrt{d_k}$ helps to avoid the problem of vanishing gradients by preventing the saturation of the softmax function.

In practice, though, Multi-Head attention is used which allows for better parallelization and enables the model to attend to information from different representation subspaces at different positions. The modified formula for h attention heads is

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O \\ \text{where head}_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{2.15}$$

where $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ are learnable parameter matrices. Figure 2.8 illustrates simple scaled dot-product attention along with multi-head attention, which comprises many scaled dot-product attention heads.

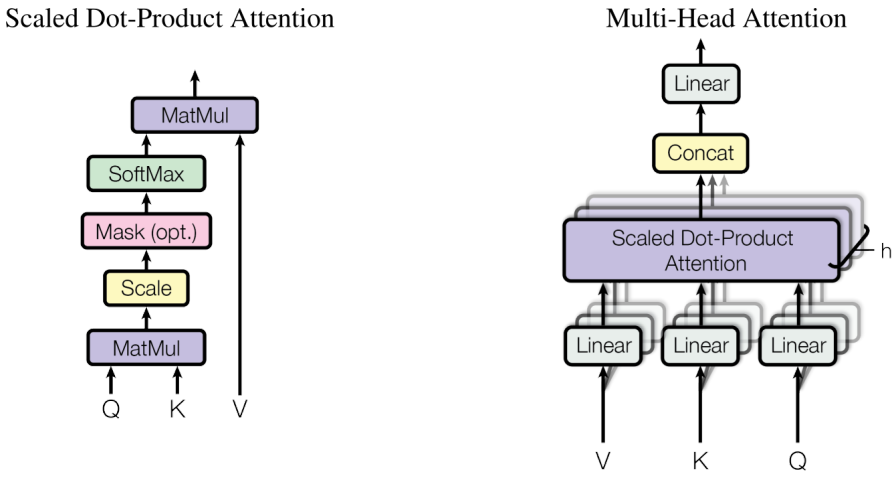


Figure 2.8: Scaled Dot-Product Attention (left) and Multi-Head Attention (right). Figure reproduced from [34].

Positional encoding is employed so that the model can utilize the ordering of the input sequence. This is necessary since no recurrence or convolution is involved—the input is treated uniformly. The encoding involves injecting information about the absolute or relative position of the tokens in the input sequence. The positional encodings have the same dimension d_{model} as the embeddings and are superimposed on them. One choice for positional encodings is to use the sine and cosine functions at different frequencies:

$$\begin{aligned} \text{PE}_{(pos,2i)} &= \sin\left(pos/10000^{2i/d_{\text{model}}}\right) \\ \text{PE}_{(pos,2i+1)} &= \cos\left(pos/10000^{2i/d_{\text{model}}}\right) \end{aligned} \tag{2.16}$$

where i is the dimension and pos the position. With this formulation of positional encodings, the model is able to extrapolate to longer sequence lengths than those encountered during training.

2.5.5 Faster R-CNN

Traditional object detection techniques that used swallow architectures on handcrafted features have been surpassed by deep neural network architectures, most of which are

improvements on R-CNN [86]. R-CNN uses a selective search algorithm to generate region proposals which are passed on to a convolutional neural network for feature extraction. The extracted features are fed to an SVM that predicts the presence of an object inside the region proposal and four offset values that increase the precision within that region [87]. However, this model is computationally slow due to the large number of region proposals that have to be classified per image and the fixed selective search algorithm may generate poor candidate regions.

Faster R-CNN [88] overcomes these problems by letting the network learn the region proposals through a Region Proposal Network (RPN) that shares convolutional features with the object detection network.

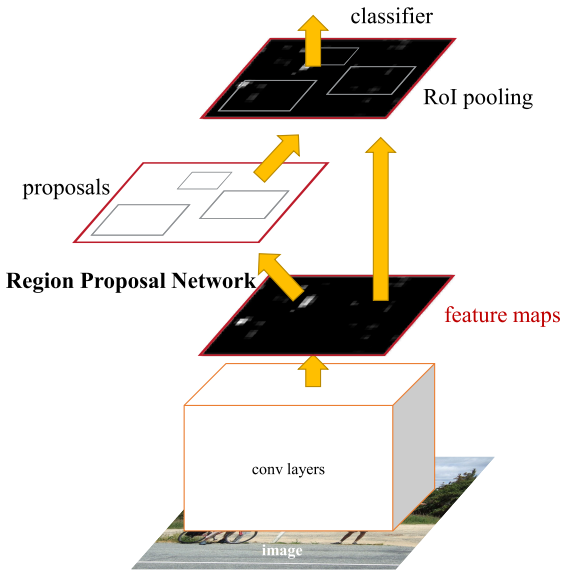


Figure 2.9: Faster R-CNN overall architecture. Figure reproduced from [88].

The input image is passed through a convolutional network that produces a feature map. A Region Proposal Network (RPN) predicts region proposals from this feature map by computing a score that corresponds to the likelihood of existence of an object for each region. Since the proposed regions' feature maps have different sizes, a Region of Interest (RoI) pooling layer is used in order to obtain fixed size representations for all regions, where each feature map is split to a fixed number of regions and the maximum value is kept in every region. Finally, the proposed region features are passed to a classifier that predicts the label of the image within the proposed region as well as the offset values that refine the bounding boxes.

2.6 Captioning Evaluation Metrics

2.6.1 BLEU

Bilingual Evaluation Understudy (BLUE) is a quick, inexpensive and language independent evaluation method for automatic translation [29]. It is defined as the geometric mean of the modified precision scores of the test corpus multiplied by an exponential brevity penalty factor. The geometric average of the modified n-gram precisions p_n is computed, using n-grams with maximum length N and positive weights w_n that sum to one. Then, the brevity penalty is calculated as follows

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{2.17}$$

where c is the length of the candidate translation and r is the effective reference corpus length. In compact notation, the BLEU score is given by

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right). \tag{2.18}$$

It ranges from 0 to 1 and indicates how close the generated text is to the reference texts, where values closer to 1 represent higher similarity. A score equal to 1 implies that the generated text is identical to the reference one. In addition, the more reference texts there are, the higher the score becomes. Thus, comparisons of different evaluations are only meaningful when the number of reference texts is constant.

2.6.2 METEOR

Metric for Evaluation of Translation with Explicit Ordering (METEOR) is an automatic metric used for evaluation of machine translation that addresses several shortcomings found in the BLEU metric [30]. Specifically, BLEU does not directly account for *recall*, which is the proportion of matched n-grams out of the total number of n-grams in the reference text, and the use of higher order n-grams does not necessarily capture correct grammar. In addition, n-grams do not provide explicit word-matching between the generated text and the reference. It should be noted that geometric averaging can be problematic when an n-gram has a zero score.

To overcome these issues, METEOR is based upon the harmonic mean of unigram precision and recall, where recall is weighted higher than precision. Moreover, besides the exact word matching, stemming and synonymy matching is used. At first, an alignment is iteratively created between the generated and the reference text, which is a set of unigram mappings. Then, unigram precision P and unigram recall R are given by

$$P = \frac{m}{w_t}, \; R = \frac{m}{w_r} \tag{2.19}$$

where m is the number of unigrams of the generated text that also exist in the reference text, w_t is the number of unigrams in the generated text and w_r is the number of unigrams in the reference text. Precision and recall are combined through an harmonic mean

$$F_{\text{mean}} = \frac{10PR}{R + 9P} \tag{2.20}$$

where recall is weighted 9 times more. In order to account for congruity with respect to larger segments that appear both in the reference and generated text, a penalty p is computed which is high when there are many non-adjacent mappings in the reference and generated text. Unigrams are grouped to the minimum possible number c of chunks, where a chunk is a set of adjacent unigrams in the generated and reference text. Mathematically, we have that

$$p = 0.5 \left(\frac{c}{u_m} \right)^3 \tag{2.21}$$

where u_m is the number of mapped unigrams. Finally, the METEOR score is given by

$$M = F_{\text{mean}} \, (1 - p). \tag{2.22}$$

The penalty effectively reduces F_{mean} if there are no bigram or longer matches. To calculate METEOR against multiple reference texts, the generated text is compared with each of these and the highest score is selected.

2.6.3 ROUGE-L

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) comprises several measures that determine automatically the quality of a summary by comparing it to reference summaries created by humans [31]. ROUGE-L is based on the longest common subse-

quence (LCS) of the generated and the reference summaries. The longer the LCS of these two summary sentences, the more similar they are. Letting X be the reference summary of length m and Y the generated summary of length n, ROUGE-L for sentence-level LCS is the LCS-based F-measure F_{lcs} given by

$$\begin{aligned} R_{lcs} &= \frac{LCS(X,Y)}{m} \\ P_{lcs} &= \frac{LCS(X,Y)}{n} \\ F_{lcs} &= \frac{(1+\beta^2)\, R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \end{aligned} \tag{2.23}$$

where $LCS(X,Y)$ is the length of the longest common subsequence of X and Y and $\beta = P_{lcs}/R_{lcs}$. ROUGE-L is equal to 1 when $X = Y$ and equal to 0 when $LCS(X,Y) = 0$, *i.e.* nothing is common between X and Y. By using the LCS criterion, there is not a necessity for a predefined n-gram length, as the longest in-sequence common n-grams are automatically included. Moreover, consecutive matches are not required, but rather in-sequence matches that reflect word order at the sentence level. However, one disadvantage of LCS is that alternative LCSes of shorter sequences are not taken into account in the final score.

2.6.4 CIDEr

Consensus-based Image Description Evaluation (CIDEr) is a new paradigm for evaluation of image captions that is based on human consensus [32]. It measures the similarity of a generated caption against a set of ground truth captions written by humans and it inherently captures sentence similarity, grammaticality, importance, saliency and accuracy. To evaluate how well a generated caption c_i matches the consensus of a set of captions $S_i = \{s_{i1}, \ldots, s_{im}\}$, all words are first mapped to their stem forms and each caption is represented using the set of $n-$grams that are present in it, where an n-gram ω_k is a set of one or more ordered words. Then, a Term Frequency Inverse Document Frequency (TF-IDF) weighting is performed for each n-gram to encode how often n-grams in the generated caption are present in the reference ones, and how often n-grams not present in the reference captions are not in the generated captions. Additionally, frequent n-grams across all images are given low weight. The TF-IDF $g_k(s_{ij})$ for each

n-gram ω_k is

$$g_k\left(s_{ij}\right) = \frac{h_k\left(s_{ij}\right)}{\sum_{\omega_l \in \Omega} h_l\left(s_{ij}\right)} \log\left(\frac{|I|}{\sum_{I_p \in I} \min\left(1, \sum_q h_k\left(s_{pq}\right)\right)}\right) \tag{2.24}$$

where $h_k\left(c\right)$ is the frequency that an n-gram ω_k occurs in the caption c, Ω is the vocabulary of n-grams and I is the set of all images. The CIDEr_n score for n-length n-grams is the average cosine similarity betwen the generated caption and the reference captions and is given by

$$\text{CIDEr}_n\left(c_i, S_i\right) = \frac{1}{m} \sum_j \frac{\boldsymbol{g}^n\left(c_i\right) \cdot \boldsymbol{g}^n\left(s_{ij}\right)}{\left\|\boldsymbol{g}^n\left(c_i\right)\right\| \left\|\boldsymbol{g}^n\left(s_{ij}\right)\right\|} \tag{2.25}$$

where $\boldsymbol{g^n}\left(c_i\right)$ is a vector with elements $g_k\left(c_i\right)$ that correspond to all n-length n-grams. The final CIDEr score combines the scores of variable length n-grams as follows

$$\text{CIDEr}\left(c_i, S_i\right) = \sum_{n=1}^{N} w_n \text{CIDEr}_n\left(c_i, S_i\right) \tag{2.26}$$

and uniform weights $w_n = 1/N$ typically work best.

2.6.5 SPICE

Semantic Propositional Image Caption Evaluation (SPICE) is an automated evaluation metric for captioning that is defined over scene graphs [33] and compares semantic propositional content. Moreover, it better captures human judgments over model generated captions than other automated metrics. At first, the generated caption c and the reference captions $S = \{s_1, \ldots, s_m\}$ are transformed to the scene graphs $G(c)$ and $G(S)$ respectively, where $G(S)$ is the union of scene graphs $G(s_i)$ for $s_i \in S$. The semantic relations in a scene graph are considered to be a conjuction of logical propositions or tuples and the function T returns these tuples from a scene graph as

$$T(G(c)) \triangleq O(c) \cup E(c) \cup K(c), \tag{2.27}$$

where $O(c)$ is the set of object mentions in c, $E(c)$ is the set of hyper-edges that represent relations between objects and $K(c)$ is the set of attributes associated with objects. The

precision P, recall R and SPICE score are defined as

$$\begin{aligned} P(c, S) &= \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|} \\ R(c, S) &= \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|} \\ \text{SPICE}(c, S) &= F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)} \end{aligned} \tag{2.28}$$

respectively, where the binary matching operator $\otimes$ returns the matching tuples in two scene graphs. Since SPICE is an F-score, it is easily interpretable and its range is between 0 and 1. Moreover, it can be applied equally to both small and large datasets.

CHAPTER 3

Cognitive Background

3.1 Introduction

Neuroimaging, which is broadly defined as techniques that enable measuring brain activity, has increased the ability to study the neural basis of cognition in the recent years [89]. Since its inception in the early 1990s, Functional Magnetic Resonance Imaging (fMRI) has become one of the most commonly used methods for studying the functionality of the human brain. fMRI comprises several imaging methods that have been developed to determine regional and time-varying signals that can be attributed to task induced changes or to resting-state unregulated processes of the brain. Due to its widespread availability, relatively low cost, non-invasive nature and good spatial resolution, it has been used extensively in a large number of studies in cognitive neuroscience, psychology, neurolinguistics and related fields [90].

3.2 BOLD Signal

The most common approach of fMRI is based on the fact that activated brain neurons cause the amount of blood that flows through the corresponding brain area to increase. In other words, the surplus of blood flow that is induced by brain activity leads to a relative increase of blood oxygen. The signal that is measured through fMRI depends on this variation in oxygenation and it is referred to as Blood Oxygenation Level Dependent (BOLD) response and it is measured from small cubical brain regions that are called *voxels*. Each voxel includes hundreds of thousands neurons, thus the BOLD signal that is measured from a specific voxel indicates the group activity of the neurons that are located within it.

The BOLD signal varies over time as the neurons of voxels are activated due to the function of the brain. Hence, fMRI data are large and complex time series that are represented as a sequence of 3D volumes that are images of the brain, with time being a 4th dimension. In a typical fMRI experiment, each image contains approximately $100,000$ voxels and 3D volumes are collected continuously with a repetition time (TR) of 2–4 seconds. Tasks are designed so that measurable changes in the BOLD signal can be recorded in order to make inferences about task-induced brain activity, where signal changes can be mapped to brain function.

It is noteworthy that the BOLD signal does not change instantly and after the stimulus ends, it does not return immediatly to the baseline condition. Changes in blood flow are slow, taking several seconds, meaning that the signal is a distorted and delayed representation of the actual neural signal. Mathematically, the local change in the BOLD signal with respect to a stimulus presentation is modelled by the haemodynamic response function (HRF), which represents an ideal response to an infinitely brief stimulus. After the first few seconds of the stimulus presentation, an initial dip occurs and then the blood flows with increasing volume. This increase continues gradually for approximately five seconds and a peak value is reached. Afterwards the blood flow decreases quickly and the BOLD signal goes below the baseline for a prolonged period of time, resulting in an undershoot. The hemodynamic response function is usually modeled through some fixed functions, such as the Gamma function, the Poisson function, the Gaussian function and the double Gamma function. A canonical HRF that corresponds to the double Gamma function is shown in figure 3.1.

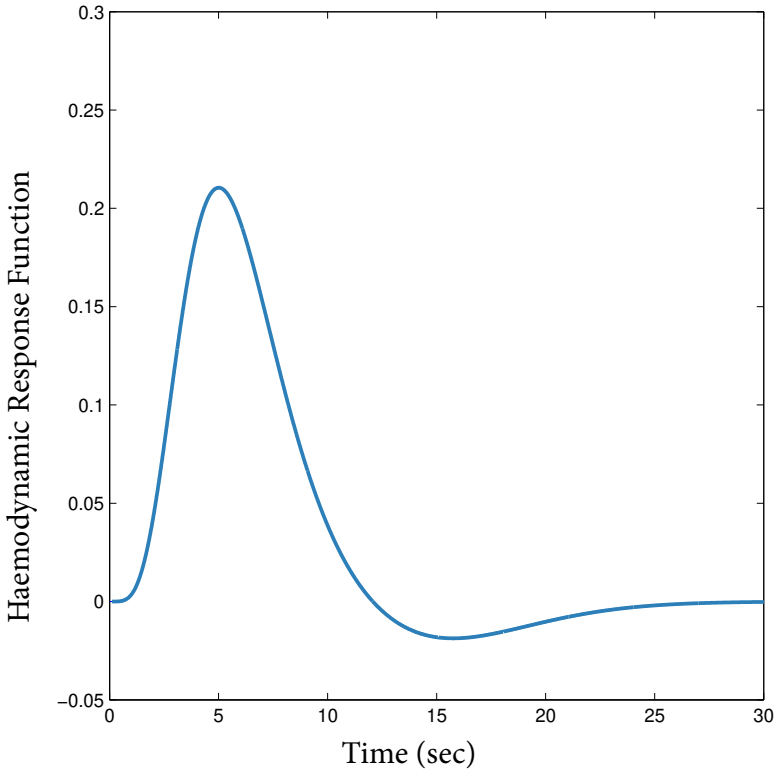


Figure 3.1: Haemodynamic Response Function.

3.3 Preprocessing

Even though fMRI is a powerful method for the detection of functional activations in the brain, the collected data are difficult to analyze since the signal is fairly weak and there are various sources of noise that must be carefully taken into account e.g. thermal noise, head motion artifacts, physiological noise caused from respiration and heartbeat, variations in the subject's cognition et cetera. Moreover, due to the constant operation of the brain, the detection of which activations are inducted by the task at interest is non-trivial, and the BOLD response represents only a small fraction of the variance of the signal. Thus, for the previous reasons, it is crucial to perform several preprocessing steps to mitigate these problems [91]. Several different preprocessing pipelines have been proposed [92–95], and in general there is no agreement on what is the optimal series of steps. However, most of these include *slice timing* and *head-motion correction*, *co-registration*, *normalization* and *spatial smoothing*.

Slice timing correction (STC) compensates for the fact that slices of the 3D volume of the brain are captured at different times and as a consequence they are not aligned properly. The time series for each slice are shifted with respect to a reference point in time in order to correct the slice dependent delays. Even though STC can interfere with other steps in the preprocessing pipeline, it is a required step and even more so in cases where the time to repetition (TR) is substantial and the hemodynamic response may change significantly among slices [96]. Head-motion correction accounts for the fact that the position of the head changes with respect to the scanner even for short times to repetition. This movement is responsible for a spatial misalignment of the voxels and head-motion correction ensures that each voxel consistently represents a specific part of the brain.

Co-registration aligns the obtained functional images with anatomical ones that have increased spatial resolution in order to facilitate further preprocessing steps and to enable detection of activations with a higher resolution anatomical reference frame. Normalization maps functional images into a normalized space that allows making comparisons between multiple subjects and enables generalization of results to a larger population. Finally, the purpose of spatial smoothing is to improve SNR and to make the distribution of data close to normal in order to facilitate further statistical analysis. A usual approach is to compare results prior and after spatial smoothing to better understand the effect it has on the results of the fMRI analysis [97].

3.4 Voxel selection

Voxel selection is an important process that has to be performed before fMRI data can be used for further analysis. This is because in a typical fMRI experiment, the whole brain can contain up to a hundred thousand voxels and this can lead to overfitting of machine learning models [48]. In addition, the fMRI signal is inherently noisy due to scanning conditions or other physical processes of the subject and consequently the activity measured in many of these voxels may be irrelevant to subsequent analysis. For example, an encoding or decoding model that utilises all voxel activations will likely not succeed as most voxels are not informative and the stimulus does not have a direct effect on them. Prior to voxel selection, a grey matter mask can be applied to select voxels that contain neuronal tissue [98, 99]. Due to the challenges posed in dealing with

the high-dimensional noisy fMRI signal, several different voxel selection strategies have been proposed in the literature.

In theory-driven analysis, the voxels are reduced by using previous knowledge regarding brain regions of interest (ROIs). In ROI-based methods [100] the analysis is restricted to a cortical area that is defined using separate anatomical or functional localizers under the assumption that these localizers produce indeed relevant cortical regions and the interesting voxels are in a homogenous region in the brain, having the same cognitive functionality. Theory-driven approaches have the benefit of bypassing challenges that occur when statistical analyses are performed at the whole-brain level.

In contrast to ROI-based methods, information-driven approaches such as searchlight analyses or whole-brain contrasts do not rely on pre-conceived notions about the location of brain responses. For example, in searchlight analyses, a spherical window is slided in the brain to select voxels by analyzing the predictive power of the signal within the region of the sphere [101]. However, this requires statistical corrections for multiple comparisons which may deteriorate brain responses, thus preventing the data analysis at a fine spatial scale. Moreover, searchlight methods implicitly require human biases about the spatial scale of observed effects and about the shape of regions of interest [47].

Another issue that needs to be considered is the variability of the signal quality over the whole brain. The fMRI signal varies significantly across the cortex and in several cases it contains severe noise [102]. In addition, the signal quality may be task-dependent. For example, in the case of an attentionally-demanding task, regions in the fronto-parietal attention network might respond more regularly than in tasks that are based on viewing images freely. Therefore, other approaches combine theory-based and information driven methods where a broad cortical region with high-quality voxels is isolated first, such as visually-responsive voxels in the occipito-temporal lobe. This procedure has the benefit that the cortical region under consideration is larger than a small region of interest while it is still relevant to the theoretical questions that are examined. However, the problem of choosing these high-quality voxels is non trivial and several different methods have been proposed.

A popular approach is to select the most active voxels with respect to the stimuli [35, 103, 104] by contrasting between resting state and stimulus conditions. Voxels that have a t-value greater than a predefined threshold are considered to be active. In this way, the regions that respond positively across stimulus conditions are isolated. This method is sensitive to the signal strength of voxels since the t-values can be low due to excessive

noise. However, it is not sensitive when systematic differences exist across conditions—voxels that respond equally across conditions will be considered to be active.

Other approaches select voxels that maximize the variance across conditions [105] or that can be predicted well by a model [11], where the latter requires setting a threshold that determines voxels that are well-fit. Similarly, voxels that are stable across conditions can be selected [1] using a cross-validation process that isolates voxels that respond consistenly across different folds. Moreover, activity-based methods can be combined with stability based ones [106]. In cases where trials are not present in the dataset and only one stimulus presentation exists per participant, a prediction driven method can be used to select informative voxels [38]. A separate encoding model can be fitted for each voxel and the model performance for a single voxel can be calculated as the Pearson correlation coefficient between real and predicted responses [107].

3.5 Encoding and Decoding Models

Due to the lack of large-scale fMRI datasets, the most commonly used method for mapping from and to fMRI data is linear and ridge regression [1, 2, 6, 38], with neural networks much less frequently used. A model that maps from the stimulus space to the voxel space is often referred to as an encoder, and conversely, a decoder operates in the opposite direction. In [1], the activation of a voxel v with respect to a word w is given by

$$y_v(w) = \sum_{i=1}^{m} c_{v,i} f_i(w), \quad \forall v = 1 \cdots V, \tag{3.1}$$

where $f_i(w)$ is a function that estimates the association between a seed word i and another word w, V is the total number of voxels and $c_{v,i}$ are weights that are learned via regression using the fMRI data for words that were used as stimuli in the fMRI experiment. For the representation of words, the co-occurrence similarity with 25 manually selected seed verbs was used. These verbs were chosen with respect to psycholinguistic criteria and their relatedness to basic sensory and motor activities. The same method is followed in [2], where the resulting cognitive embeddings are evaluated in NLP downstream tasks. Other works [108] report that equally good results can be achieved even with automatically choosing the seed words or with utilising WordNet based features [109, 110]. In [111], it is concluded that no input representation is better overall at predicting brain activations, although morphological and dependency based models seem to perform

better. In [5], a similarity-based approach is considered, where the voxel activations for an unknown word w are computed as a sum of activations of known words u_i, weighted by $f(u_i, w)$, with f being a similarity function. In [112] many different semantic models are reviewed for input representation, including dependency, association and image based ones and it is concluded that visual information is a stronger predictor of brain activity than linguistic information for concrete nouns. A generative model is used in [11] to map from low dimensional co-occurence word embeddings to fMRI data of cortical areas, where the semantic category clusters in the brain are modeled as a probability distribution and emission probabilities are modeled as Gaussians. In [6], ridge regression is used in order to predict GloVe embeddings from voxel activations. A decoder that is trained on isolated words, which may include abstract nouns, is able to accurately classify sentences through their corresponding fMRI data, with different levels of granularity. In [4, 113], conventional word embeddings are mapped to cognitive embeddings using a neural network with one hidden layer. Specifically, in [4] it is reported that through the use of neural networks, both encoding and decoding accuracy is improved, compared to a linear regression model on the same input. The low temporal resolution problem of fMRI activations is addressed in [114] by sliding a Gaussian window across tokens, which accounts for the Haemodynamic delay. The resulting representations are used in conjuction with a Hidden Markov Model in order to improve performance in part-of-speech induction. In [115, 116], it is found that LSTM-based sentence representations correlate well with brain activations. Moreover, to map sentence stimuli to fMRI activations, ridge regression is used on top of a pretrained LSTM for language modelling [107], where it is stated that LSTMs, which encode context, are better at predicting activations for individual words in a sentence.

CHAPTER 4

Image captioning with fMRI fusion

4.1 Introduction

Even though image captioning is a difficult task for computers, humans can easily describe images through inherent capabilities of their brains with little effort. It is evident that human brain activations encode semantic information about what people see and think. In the domain of neuroscience, several studies have attempted to extract semantic information about what people see or imagine from brain activations [117, 118]. In this work, we propose several techniques of incorporating fMRI brain activations to an image captioning model that is based on the transformer architecture. Due to the fact that fMRI data are limited, a "lexical expansion" step is performed, where brain activations are predicted for new visual stimuli.

4.2 Related Work

In recent years, several papers have tried to analyze quantitatively the semantic representations of the human brain through fMRI activations that were evoked from visual stimuli such as images or natural movies [1, 2, 8–10]. Using the semantic categories of WordNet, a relationship between visual stimuli and brain activations was revealed in [10, 11], where a map of semantic representations of the cerebral cortex was constructed, and it was shown that semantic information exists over broad areas within the cortex. In [12], a model that classifies brain activations to semantic categories was constructed and brain areas that respond to specific semantic categories were revealed. In [13], it was suggested that visual attention alters visual representations in the brain to optimize processing of relevant objects during natural vision. A correlation between intermediate representations of Deep Neural Network layers and areas of the dorsal stream was found in [14]

when a Deep Neural Network trained for action recognition was used to predict voxels of the dorsal stream where natural movies were used as stimuli. In [15], it was shown that distributed semantics based on a skip-gram model trained on the Japanese Wikipedia were correlated with fMRI activations. In order to provide a more cognitively plausible view of distributional semantic models, in [16] it is tested whether image-based models capture the semantic patterns that emerge from fMRI recordings, concluding that image-based models improve text-based ones. In [17] a linguistic word2vec based model is used along with a visually grounded one for decoding concrete and abstract nouns from brain activations. However, no significant improvement was observed for multimodal models even in the case of concrete nouns and the linguistic model was superior in the case of abstract nouns. In [18], the authors present a deep autoencoder model consisting of a CNN with an LSTM. The model is trained on fMRI slice sequences and predicts the entire brain volume using multimodal stimuli as input. In [4], the stimuli is extended from nouns to images and nouns in order to illustrate the strong correlation between linguistic and visual representations in the human brain. In [19], multimodal models that differentiate between internal visual properties of objects and their external visual context are constructed. These models, when evaluated on the task of decoding brain activity, perform better than those that are based on complete images. In [20], several word embedding models are evaluated along with their combinations for decoding fMRI data associated with three word classes and three stimuli input modalities. Multimodal and meta-word embedding models achieve better performance than their component embedding models since visual information is important for distinguishing words from each other due to the large portion of informative voxels in the visual networks.

4.3 Datasets overview

4.3.1 BOLD5000

BOLD5000 [21] is a functional MRI dataset that is based on responses from almost 5000 diverse real world images that overlap with typical computer vision datasets (SUN, COCO, ImageNet). It is an order of magnitude larger than previous similar datasets and it allows for more fine-grained research in the neural representation of visual features, categories and semantics. Specifically, fMRI data was collected from four participants (16 sessions) in a slow event-related process that aimed to disentangle which stimuli

caused specific neural responses. Images are comprised of 1000 indoor and outdoor scenes of 250 categories (SUN), 2000 objects embedded in realistic context (COCO) and 1916 images of mostly singular objects (ImageNet). Moreover, they are downsampled to 375×375 pixels and, in order to ensure uniform luminance, gray world normalization is performed. In addition to the raw data that are provided for each session, processed voxel features are provided for ten Regions of Interest (ROIs)—five regions (PPA, RSC, OPA, LOC, EV) for each brain hemisphere (LH, RH)—and for five two-second time intervals (TR[1–5]). According to functional localizers with respect to stimulus from scenes, objects and scrambled images, three ROIs are scene selective (PPA, RSC, OPA), one is object selective (LOC) and one corresponds to early visual (EV). In order to extract the voxel activations for each region of interest, thresholding was performed in each one using a family-wise error correction of $p < 0.0001$, $k = 30$. An overview of the dataset is presented in table 4.1

Participants	4
fMRI Sessions per Participants	16
Total fMRI Scene Runs Per Participant	142
Total Functional Localizer Runs Per Participant	8
Total Scene Trials Per Participant	5,254
Unique Scene Stimuli Per Participant	4,916

Table 4.1: BOLD5000 dataset overview. Table was adapted from [21].

A t-SNE visualization for voxel activations of a scene selective region (RH-RSC) was performed to determine whether it responds differently with respect to stimuli across datasets. From figure 4.1 we observe that activations that correspond to the SCENES dataset are clustered together, which agrees with our intuition.

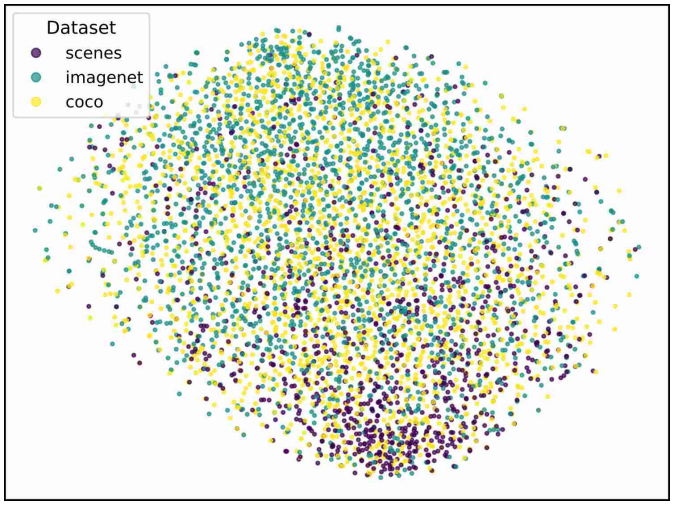


Figure 4.1: t-SNE for the voxel activations of scene selective region of interest RH-RSC for the time interval 6–8 sec (TR4).

4.3.2 MS-COCO

MS-COCO [22] is one of the most commonly used datasets for evaluating image captioning models [23–27]. It contains 82,783 images for training and 40,504 images for validation, with five human-annotated descriptions per image. Since the annotations for the official testing set are not available publicly, we adopt the widely used "Karpathy" split [28], and take 113,287 images for training, 5,000 for validation and 5,000 for testing. In order to evaluate the generated caption quality, the standard automatic evaluation metrics are used, namely BLEU [29], METEOR [30], ROUGE [31], CIDEr [32] and SPICE [33].

4.4 Methodology

4.4.1 Baseline Architecture

The baseline image captioning model is based on the Transformer encoder-decoder architecture [34]. Specifically, the encoder performs self-attention on visual features that have been extracted via Faster R-CNN as described in [23]. The decoder performs masked self-attention on caption tokens and visio-linguistic attention that conditions the visual stream on the language stream. The visual features are projected via a linear layer to the

lower-dimensional representation space of the encoder. The overall baseline architecture is illustrated in figure 4.2 and its performance on the MS-COCO dataset is shown on table 4.2.

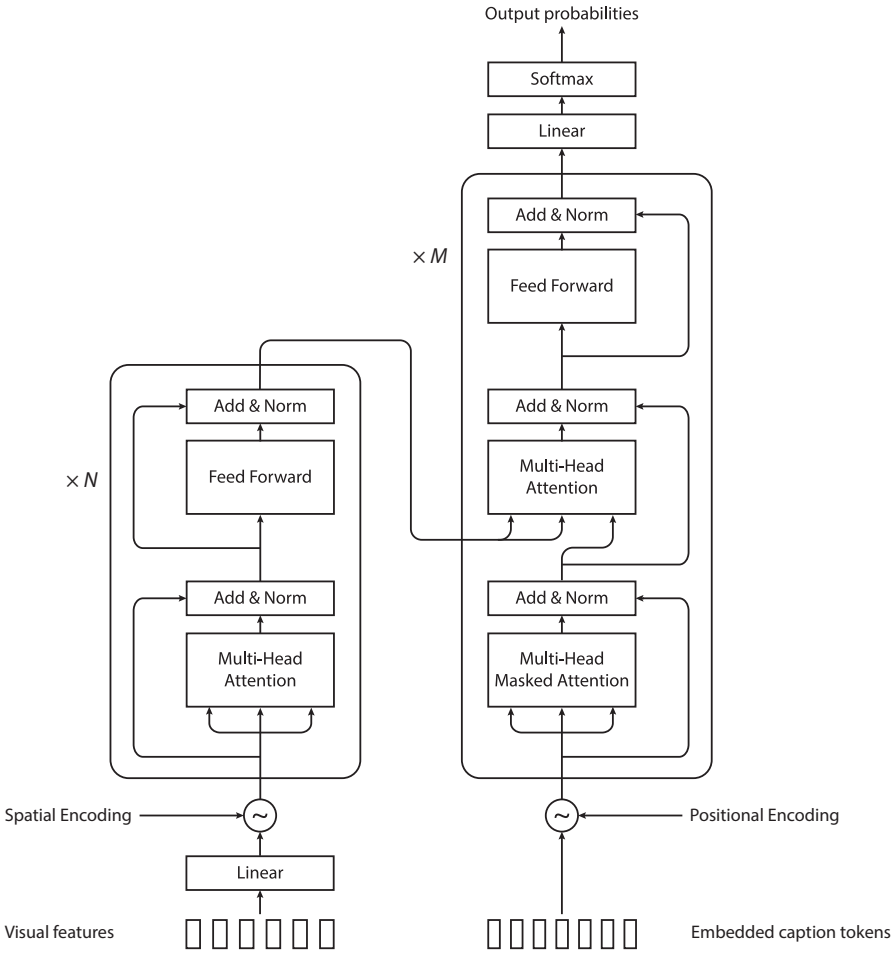


Figure 4.2: Baseline image captioning transformer-based architecture.

Scores	Baseline model
Bleu_1	0.740
Bleu_2	0.578
Bleu_3	0.445
Bleu_4	0.343
METEOR	0.279
ROUGE_L	0.558
CIDEr	1.117
SPICE	0.208

Table 4.2: Performance of the baseline architecture evaluated on the MS-COCO dataset.

4.4.2 Representational Similarity Analysis

Representational Similarity Analysis (RSA) [35] was initially used to relate brain representations of visual objects to representations of computational models. In order to gain some insight about the brain representations prior to performing lexical expansion (section 4.4.3), we perform a preliminary RSA to determine the similarity of brain representations across subjects with respect to the different regions of interest. Let $\boldsymbol{v}_i^{(r)} \in \mathbb{R}^d$, $d \in (100, 200)$, be the voxel activations for region of interest r for image stimulus $i \in S$, where $S = \{1, 2, \dots s\}$. For each region of interest, we compute:

$$\boldsymbol{d}^{(r)} = (d_{ij}^{(r)}), \qquad d_{ij}^{(r)} = \cos(v_i^{(r)}, v_j^{(r)}) \quad i \in [1..s],\ j \in [i+1..s]. \tag{4.1}$$

Then, we compute the Spearman correlation $s(\boldsymbol{d}^{(r)}, \boldsymbol{d}^{(r')})$ for regions of interest r and r', which can be chosen either from the same or from different subjects. This is displayed in figure 4.3, where in the left heatmap r and r' correspond to subject 1 and in the right heatmap r corresponds to subject 1 and r' to subject 2. Each entry of the heatmap is the respective Spearman correlation.

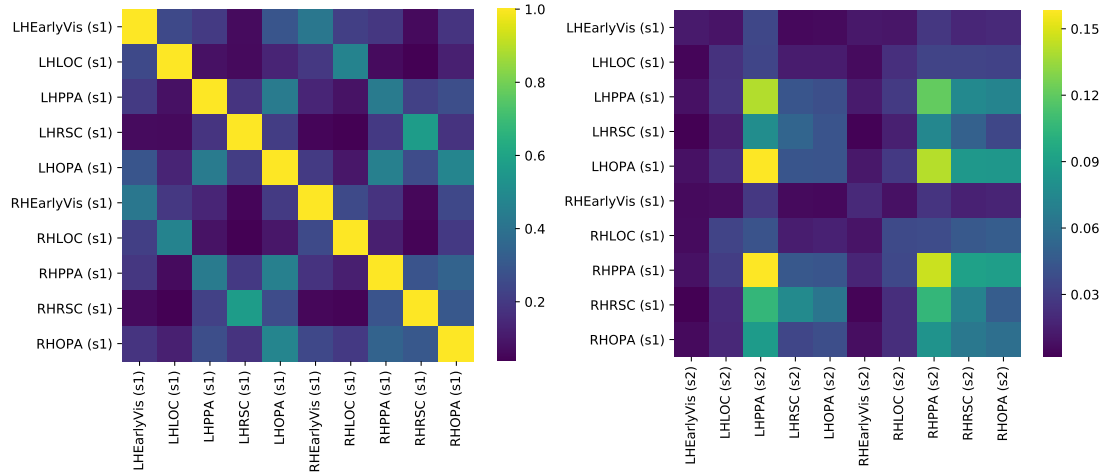


Figure 4.3: Representational Similarity Analysis for ROIs of participant 1 (left) and for ROIs of participants 1 and 2 (right).

It is interesting to note that when r and r' are from the same subject, a correlation between the left and right hemispheres can be observed from the diagonal patterns in figure 4.3 (left). When r and r' are from different subjects, weak correlations are observed only among higher level ROIs.

A similar analysis is performed to establish the similarity of visual embeddings, extracted from a pretrained VGG16 network, 300-dimensional GloVe840B [36] embeddings and voxel activations. From table 4.3 we can see that voxel activations are more correlated with visual embeddings than with word embeddings with greater correlations being observed for higher level regions of interest. This observation suggests that the encoding model that predicts voxel activations should be based on visual and not word embeddings.

		P1	P2	P3	P4
LHEarlyVis	visual	0.051	0.029	0.012	0.031
LHEarlyVis	GloVe	−0.017	−0.014	−0.007	−0.025
LHLOC	visual	−0.008	−0.005	0.007	0.010
LHLOC	GloVe	0.000	−0.013	−0.006	−0.037
LHPPA	visual	0.145	0.172	0.126	0.159
LHPPA	GloVe	0.061	0.088	0.116	0.085
LHRSC	visual	0.077	0.048	0.034	0.036
LHRSC	GloVe	0.049	0.026	0.029	0.019
LHOPA	visual	0.164	0.045	0.068	0.102
LHOPA	GloVe	0.100	0.027	0.061	0.048
RHEarlyVis	visual	0.054	0.037	−0.000	0.030
RHEarlyVis	GloVe	0.007	−0.013	−0.013	−0.032
RHLOC	visual	0.005	0.014	−0.010	0.012
RHLOC	GloVe	−0.024	−0.010	−0.010	−0.015
RHPPA	visual	0.202	0.172	0.118	0.165
RHPPA	GloVe	0.084	0.111	0.084	0.056
RHRSC	visual	0.148	0.123	0.101	0.062
RHRSC	GloVe	0.096	0.097	0.092	0.036
RHOPA	visual	0.134	0.110	0.057	0.128
RHOPA	GloVe	0.104	0.086	0.031	0.051
visual	GloVe	0.265	0.266	0.268	0.262

Table 4.3: Representational Similarity Analysis between ROIs, visual features (VGG16) and word embeddings (300-dimensional GloVe840B) for all participants P[1-4]. "LH" and "RH" denote left and right hemisphere respectively.

4.4.3 Lexical Expansion

Due to the scarcity of fMRI data, it is necessary to expand the "lexicon" of available fMRIs by building an encoding model that predicts voxel activations for images that were not used as stimuli in the fMRI experiment. This will enable us to obtain voxel activations for any input image of the captioning model. To this end, we train a model of approximately 5000 pairs of images and corresponding fMRIs to learn a mapping from visual features to fMRIs. The visual features $\boldsymbol{v}_i \in \mathbb{R}^{1\times 512}$ are obtained via average pooling on the last layer of a pretrained VGG-16 network and for the fMRIs, we omit the ROIs of the early visual system so that $\boldsymbol{f}_i \in \mathbb{R}^{1\times 1100}$. On the following paragraphs, we describe three different approaches that were used for this encoding process.

Ridge Regression

Ridge regression is the prevalent technique that is used for mapping to voxel activations [2, 6, 37, 38] due to its simplicity and satisfactory performance. Assuming that $\boldsymbol{F} \in \mathbb{R}^{5000\times1100}$ represents the fMRI features for all stimuli and $\boldsymbol{V} \in \mathbb{R}^{5000\times512}$ represents the corresponding visual features, the fMRI features of new images $\boldsymbol{V}_{\text{new}}$ that were not used as stimuli are obtained by

$$\hat{\boldsymbol{F}} = \boldsymbol{V}_{\text{new}}\hat{\boldsymbol{W}} \tag{4.2}$$

where $\hat{\boldsymbol{W}} \in \mathbb{R}^{512\times1100}$ is a learned matrix whose i-th column $\hat{\boldsymbol{W}}_{:,i}$ is given by

$$\hat{\boldsymbol{W}}_{:,i} = \underset{\boldsymbol{W}_{:,i}\in\mathbb{R}^{512\times1}}{\arg\min} ||\boldsymbol{F}_{:,i} - \boldsymbol{V}\boldsymbol{W}_{:,i}||_2^2 + \lambda_i||\boldsymbol{W}_{:,i}||_2^2 \tag{4.3}$$

and $\lambda_i \geq 0$ is a regularization parameter that is obtained through cross-validation.

Sparse Dictionary Learning

A less frequently used technique is *sparse dictionary learning* where the fMRI features $\boldsymbol{F} \in \mathbb{R}^{5000\times1100}$ are decomposed so that

$$\boldsymbol{F} = \boldsymbol{C}\boldsymbol{D}, \tag{4.4}$$

where $\boldsymbol{C} \in \mathbb{R}^{5000\times a}$ is a sparse matrix of dictionary coefficients and $\boldsymbol{D} \in \mathbb{R}^{a\times1100}$ contains the dictionary basis vectors as its rows. This decomposition essentially aims to smooth the fMRI representations by expressing them as a linear combination of dictionary atoms that in theory do not fully capture the noise present in the fMRI signal. Then a matrix $\hat{\boldsymbol{W}}$ that maps visual features to dictionary coefficients is learned via ridge regression. The predicted fMRI features are obtained in two steps from

$$\hat{\boldsymbol{C}} = \boldsymbol{V}_{\text{new}}\hat{\boldsymbol{W}} \tag{4.5}$$

$$\hat{\boldsymbol{F}} = \hat{\boldsymbol{C}}\boldsymbol{D} \tag{4.6}$$

Similarity

Another technique described in [5] expresses predicted fMRI activations as a linear combination of known fMRI activations. Formally, the predicted fMRI activations $\hat{\boldsymbol{f}}_{\text{new}} \in$

$\mathbb{R}^{1\times 1100}$ for a new image $\hat{\boldsymbol{v}}_{\text{new}} \in \mathbb{R}^{1\times 512}$ are given by

$$\hat{\boldsymbol{f}}_{\text{new}} = \sum_{i=1}^{5000} g(\boldsymbol{v}_{\text{new}}, \boldsymbol{v}_i)\boldsymbol{f}_i \tag{4.7}$$

where $g(\boldsymbol{v}_i, \boldsymbol{v}_j)$ is a function that measures the similarity of images $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$.

4.4.4 Evaluation of lexical expansion

In order to assess the performance of the different methods that were proposed for lexical expansion, we use the standard evaluation procedure from the literature [38]. In particular, we repeatedly train each model with different subsets of $m-2$ image–fMRI pairs and perform the following evaluation on the two remaining pairs, where m is the total number of image–fMRI pairs. A prediction is considered successful if

$$g(\boldsymbol{f}_1, \hat{\boldsymbol{f}}_1) + g(\boldsymbol{f}_2, \hat{\boldsymbol{f}}_2) > g(\boldsymbol{f}_1, \hat{\boldsymbol{f}}_2) + g(\boldsymbol{f}_2, \hat{\boldsymbol{f}}_1), \tag{4.8}$$

where the real fMRIs are $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, the predicted ones are $\hat{\boldsymbol{f}}_1$, $\hat{\boldsymbol{f}}_2$ and g is a similarity function e.g. Spearman correlation. Table 4.4 displays the results of performing this process for $n = 1000$ iterations.

	Ridge Regression	Dictionary Learning	Similarity
subject 1	0.8775	0.925	0.886
subject 2	0.8575	0.925	0.865
subject 3	0.8370	0.845	0.873
subject 4	0.8500	0.905	0.861

Table 4.4: Evaluation of lexical expansion from visual features (VGG16) to fMRI activations. The number of iterations of equation (4.8) is $n = 1000$. For the "Similarity" method, dot product was used as a similarity function.

All three methods achieve good results in general, with dictionary learning having the best scores for three out of four subjects. The similarity-based method, with dot product as a similarity function in equation (4.7), produced good evaluation scores even though the predicted voxel activations had unrealistically high values. This fact indicates that the evaluation procedure used in the literature does not give us strong guarantees of the quality of the predicted fMRI features.

Another way to evaluate the predicted fMRIs in order to assess whether they contain useful information is to use them instead of visual features in the baseline architecture. The input to the encoder becomes

$$\hat{\boldsymbol{F}}\boldsymbol{W}_{\text{fmri}} + \boldsymbol{P}\boldsymbol{W}_{\text{pos}}, \tag{4.9}$$

where $\hat{\boldsymbol{F}} \in \mathbb{R}^{\text{boxes}\times 1100}$ contains the predicted fMRI features that correspond to the visual features of each bounding box, $\boldsymbol{P} \in \mathbb{R}^{\text{boxes}\times 5}$ encodes the positional information of the bounding boxes and $\boldsymbol{W}_{\text{fmri}} \in \mathbb{R}^{1100\times 512}$, $\boldsymbol{W}_{\text{pos}} \in \mathbb{R}^{5\times 512}$ are learnable matrices. Table 4.5 displays the results that were obtained for all lexical expansion methods. We can see that the model achieves decent scores even in the absence of visual features. The best performing method is clearly Ridge Regression, even though Dictionary Learning had slightly better performance in the previous evaluation method. This may indicate that Ridge Regression better preserves some properties of the visual embedding space.

Scores	Baseline model	Ridge Regression	Dictionary Learning	Similarity
Bleu_1	0.740	0.693	0.672	0.635
Bleu_2	0.578	0.523	0.498	0.453
Bleu_3	0.445	0.393	0.369	0.326
Bleu_4	0.343	0.298	0.276	0.241
METEOR	0.279	0.262	0.249	0.227
ROUGE_L	0.558	0.530	0.516	0.483
CIDEr	1.117	0.979	0.893	0.751
SPICE	0.208	0.189	0.174	0.154

Table 4.5: Performance on image captioning for the MS-COCO dataset when training with predicted fMRI features instead of visual features. The fMRIs have been predicted using visual features (VGG-16) extracted from bounding boxes.

Explained Variance

In the case of Ridge Regression, we can additionally evaluate the linear mapping from visual features $\boldsymbol{V} \in \mathbb{R}^{5000\times 512}$ to predicted fMRIs $\hat{\boldsymbol{F}} \in \mathbb{R}^{5000\times 1100}$ via the explained variance for all voxels. Figure 4.4 illustrates the resulting histogram. The mapping is far from perfect, since for most voxels, the mean voxel response is predicted. This can be attributed to the fact that a linear model is not powerful enough for this mapping, to the inherent noise of the fMRI signal or to the fact that voxels contain information that cannot be derived purely from visual features.

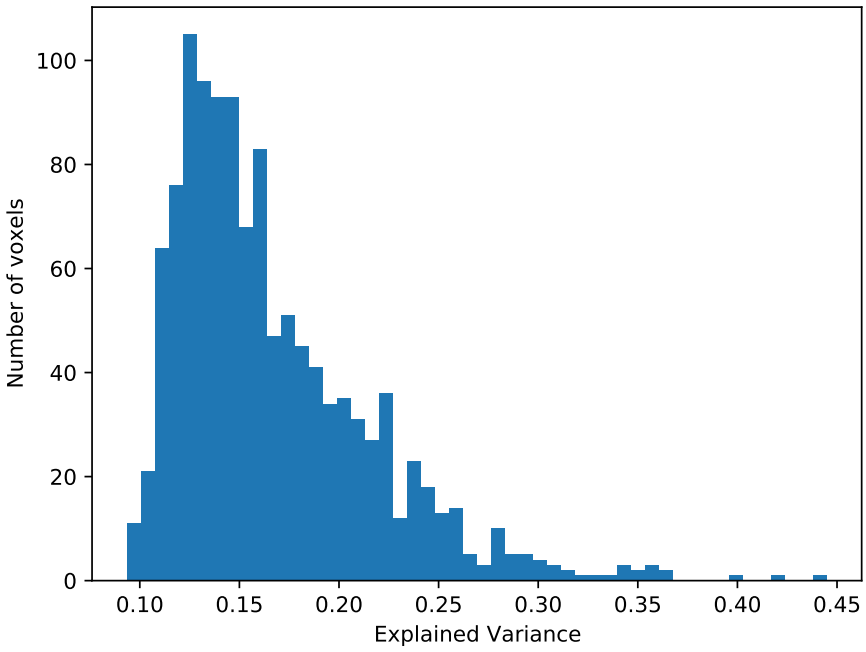


Figure 4.4: Histogram for Explained Variance (EV) per voxel in the case of Ridge Regression. When EV is close to 0, the mean voxel response is predicted.

4.4.5 Fusion at the Encoder

The simplest approach in which fMRI features can be incorporated to the captioning model is via fusion at the encoder, which is illustrated in figure 4.5. Several variants are described in the following paragraphs, where the fMRI features are either added or concatenated to the visual features that act as input to the encoder.

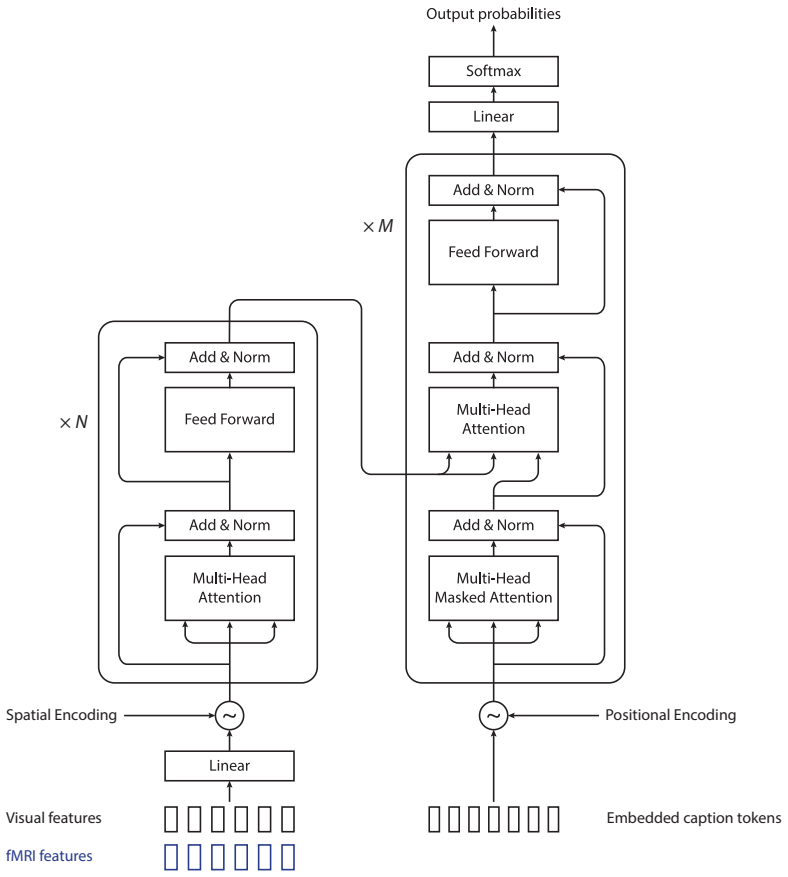


Figure 4.5: Fusion at the encoder. fMRI features are predicted using the visual features of each bounding box and are added or concatenated to the corresponding visual features.

Padding, addition The predicted fMRIs are added to the visual features of each bounding box with appropriate zero-padding. Formally, the final input features for the i-th bounding box are given by

$$\boldsymbol{X}_{i,:} = (\boldsymbol{V}_{i,:} + \text{pad}(\hat{\boldsymbol{F}}_{i,:}))\boldsymbol{W}_{\text{visual}} + \boldsymbol{P}_{i,:}\boldsymbol{W}_{\text{pos}} \tag{4.10}$$

where $\boldsymbol{V} \in \mathbb{R}^{\text{boxes}\times 2048}$ contains the visual features for each bounding box that have been extracted via Faster-RCNN and $\hat{\boldsymbol{F}} \in \mathbb{R}^{\text{boxes}\times 1100}$ contains the predicted fMRI features that have one-to-one correspondence to the visual features of the bounding boxes. Also, $\boldsymbol{P} \in \mathbb{R}^{\text{boxes}\times 5}$ encodes the positional information of the bounding boxes.

Projection, addition This case is similar to the previous one, with the difference that first the predicted fMRIs are linearly projected through a linear layer to the lower dimen-

sional embedding space of the encoder and then they are added to the visual features. Mathematically, the final input features for the i-th bounding box are given by

$$\boldsymbol{X}_{i,:} = \boldsymbol{V}_{i,:}\boldsymbol{W}_{\text{visual}} + \hat{\boldsymbol{F}}_{i,:}\boldsymbol{W}_{\text{fmri}} + \boldsymbol{P}_{i,:}\boldsymbol{W}_{\text{pos}} \tag{4.11}$$

Concatenation Another alternative is to concatenate the predicted fMRIs to the visual features where

$$\boldsymbol{X}_{i,:} = (\boldsymbol{V}_{i,:}||\hat{\boldsymbol{F}}_{i,:})\boldsymbol{W} + \boldsymbol{P}_{i,:}\boldsymbol{W}_{\text{pos}} \tag{4.12}$$

and $\boldsymbol{W} \in \mathbb{R}^{(2048+1100)\times 512}$ is a learnable matrix. Table 4.6 displays the results of the above methods, with fMRI features being predicted only through Ridge Regression, since other methods performed similarly or worse. It is obvious that concatenation deteriorates performance, while addition does not affect the model in any significant way.

Scores	Baseline model	Regression (pad, add)	Regression (project, add)	Regression (concat)	Regression (concat, ℓ_2-norm)
Bleu_1	0.740	0.736	0.735	0.730	0.727
Bleu_2	0.578	0.574	0.573	0.566	0.563
Bleu_3	0.445	0.441	0.441	0.433	0.430
Bleu_4	0.343	0.341	0.340	0.332	0.330
METEOR	0.279	0.283	0.283	0.280	0.278
ROUGE_L	0.558	0.561	0.560	0.557	0.553
CIDEr	1.117	1.124	1.124	1.105	1.092
SPICE	0.208	0.212	0.211	0.208	0.206

Table 4.6: Performance of fMRI fusion on the encoder for image captioning on the MS-COCO dataset.

4.4.6 Drop-net

One possible deficiency in the previous approach is that the network may rely mostly on the strong visual modality and treat the additional fMRI features as noise. Inspired by the Drop-net trick that is proposed in [39], we can mitigate this problem by forcing the model to either use only the visual modality, or the fMRI modality or both at the same time, based on a randomly sampled uniform variable U^l from the interval $[0, 1]$. For the case when fMRI features are superimposed to the visual features, the input at the

encoder becomes

$$\begin{aligned}\boldsymbol{X}_{i,:} = &\mathbb{I}\left(U^l < \frac{p_{\text{net}}}{2}\right)\boldsymbol{V}_{i,:}\boldsymbol{W}_{\text{visual}} + \mathbb{I}\left(U^l > 1 - \frac{p_{\text{net}}}{2}\right)\hat{\boldsymbol{F}}_{i,:}\boldsymbol{W}_{\text{fmri}} \\ &+ \frac{1}{2}\mathbb{I}\left(\frac{p_{\text{net}}}{2} \leq U^l \leq 1 - \frac{p_{\text{net}}}{2}\right)\left(\boldsymbol{V}_{i,:}\boldsymbol{W}_{\text{visual}} + \hat{\boldsymbol{F}}_{i,:}\boldsymbol{W}_{\text{fmri}}\right) + \boldsymbol{P}_{i,:}\boldsymbol{W}_{\text{pos}}\end{aligned} \tag{4.13}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $p_{\text{net}} \in [0, 1]$ is the drop-net rate. The results, shown in table 4.7 for $p_{\text{net}} = 0.8$, are really poor and this can be attributed to the fact that the strong visual modality is not used continuously and the weak fMRI modality disrupts the training process.

Scores	Baseline model	Regression (pad, add)	Regression (project, add)
Bleu_1	0.740	0.632	0.629
Bleu_2	0.578	0.460	0.451
Bleu_3	0.445	0.340	0.336
Bleu_4	0.343	0.256	0.252
METEOR	0.279	0.226	0.219
ROUGE_L	0.558	0.486	0.480
CIDEr	1.117	0.782	0.776
SPICE	0.208	0.154	0.146

Table 4.7: Performance of fMRI fusion with drop-net ($p_{\text{net}} = 0.8$) on the encoder for image captioning on the MS-COCO dataset.

4.4.7 Decoder Attention Conditioning

A different approach to incorporate the fMRI features is via attention conditioning on the decoder. In this case, we no longer utilize the visual features from the bounding boxes, but rather from the whole input image. Letting $\hat{\boldsymbol{f}} \in \mathbb{R}^{1\times 1100}$ the predicted fMRI features, the query vectors $\boldsymbol{Q}_{i,:} \in \mathbb{R}^{1\times 512}$ of the decoder become

$$\boldsymbol{Q}'_{i,:} = \boldsymbol{Q}_{i,:} + \hat{\boldsymbol{f}}\boldsymbol{W}_{\text{fmri}} \tag{4.14}$$

where $\boldsymbol{W}_{\text{fmri}} \in \mathbb{R}^{1100\times 512}$ is a learnable matrix that projects the fMRI vectors to a 512-dimensional embedding space. The modified architecture is shown on figure 4.6. To determine whether the fMRI features are treated as noise by the model, an additional experiment where random noise was added to the queries was performed. Indeed, the results, shown on table 4.8, validate the previous hypothesis.

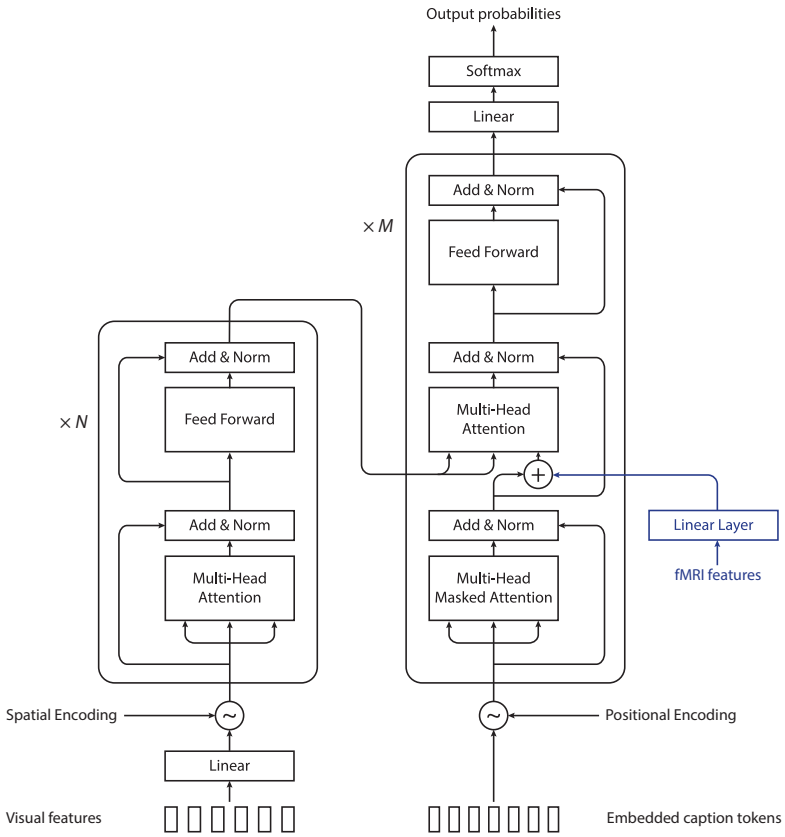


Figure 4.6: A predicted fMRI feature vector $\hat{\boldsymbol{f}}$, obtained from the visual features of the whole image, is linearly projected to a 512-dimensional space and added to each query vector $\boldsymbol{Q}_{i,:}$.

Scores	Baseline model	Regression	Dictionary Learning	Random
Bleu_1	0.740	0.737	0.735	0.736
Bleu_2	0.578	0.575	0.574	0.576
Bleu_3	0.445	0.443	0.442	0.443
Bleu_4	0.343	0.341	0.342	0.342
METEOR	0.279	0.283	0.283	0.282
ROUGE_L	0.558	0.562	0.560	0.562
CIDEr	1.117	1.129	1.127	1.127
SPICE	0.208	0.212	0.212	0.210

Table 4.8: Performance on image captioning for the MS-COCO dataset with attention conditioning on the decoder. At each query vector, a linear projection of the same fMRI vector is added. The fMRIs have been predicted with Ridge Regression and Dictionary learning, using visual features (VGG-16) extracted from the whole image. For the last column, random noise has been added instead of fMRI features.

Instead of using fMRI features predicted from the whole image, we can aggregate fMRI features predicted from visual features of the bounding boxes via an LSTM. Similarly as before, the result of this aggregation is added to the queries of the decoder. The modified architecture is shown on figure 4.7. Results, shown on table 4.9, are similar to those of table 4.8, implying again that the fMRI features are treated as noise by the model.

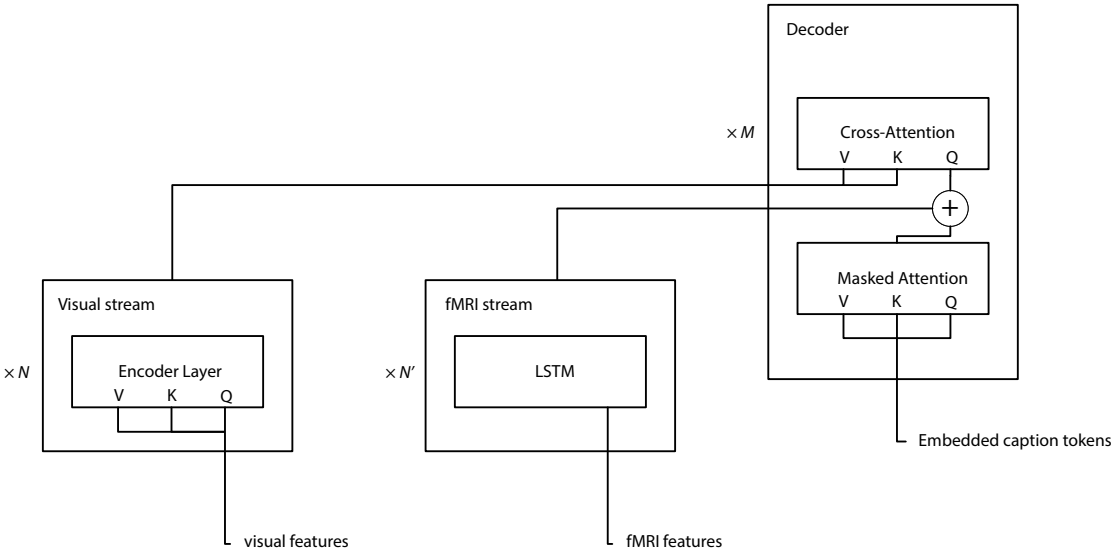


Figure 4.7: LSTM for aggregating fMRIs predicted from bounding boxes and fusion on the queries of the decoder.

Scores	Baseline model	Regression + LSTM	Dictionary + LSTM
Bleu_1	0.740	0.735	0.733
Bleu_2	0.578	0.574	0.573
Bleu_3	0.445	0.441	0.440
Bleu_4	0.343	0.341	0.339
METEOR	0.279	0.283	0.282
ROUGE_L	0.558	0.560	0.556
CIDEr	1.117	1.126	1.125
SPICE	0.208	0.212	0.210

Table 4.9: Performance of the LSTM architecture where a 3-layer LSTM was used. Ridge Regression and Dictionary Learning were used for lexical expansion from visual features extracted from bounding boxes.

4.4.8 Two separate encoders

Inspired by the fusion strategies for transformer architectures that were proposed in [40], we add a separate transformer encoder and an additional multi-head attention block at the decoder for the fMRI stream. The resulting architecture is illustrated on figure 4.8. Formally, the total attention is given by

$$\mathcal{A}_{\text{tot}} = \mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}_{\text{vis}}, \boldsymbol{V}_{\text{vis}}) + \mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}_{\text{fmri}}, \boldsymbol{V}_{\text{fmri}}) \tag{4.15}$$

where $\boldsymbol{K}_{\text{vis}} \equiv \boldsymbol{V}_{\text{vis}}$ is the output of the visual encoder and $\boldsymbol{K}_{\text{fmri}} \equiv \boldsymbol{V}_{\text{fmri}}$ is the output of the fMRI encoder.

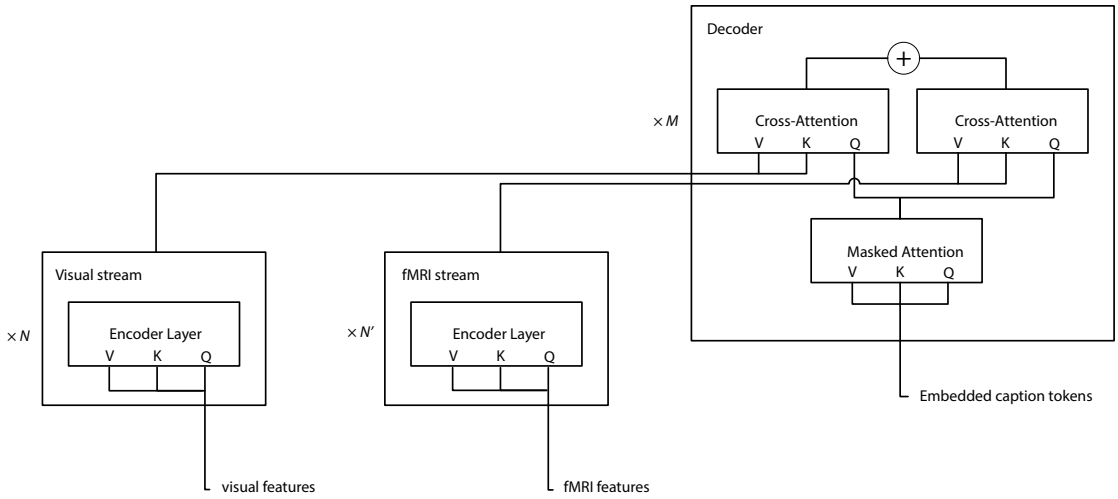


Figure 4.8: Two separate encoders with "parallel" fusion on the decoder side.

4.4.9 Null Input for fMRIs

In this section, we introduce a variation of the architecture discussed previously that compensates for the fact that the fMRI stream may not be useful in all cases. It is inspired from the null input idea that was proposed in [41]. Letting b the number of bounding boxes of an image and $\boldsymbol{F} \in \mathbb{R}^{b\times 512}$ the output of the fMRI encoder, we append a null input (random noise) $\boldsymbol{n}_\circ$ to this matrix so that

$$\boldsymbol{F}^+ = \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{n}_\circ \end{bmatrix} \tag{4.16}$$

At the fMRI cross-attention at the decoder, we have that $\boldsymbol{K} \equiv \boldsymbol{F}^{+}$, thus

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^{\mathsf{T}} = \begin{bmatrix} \boldsymbol{Q}\boldsymbol{F}^{\mathsf{T}} & \boldsymbol{Q}\boldsymbol{n}_{\circ}^{\mathsf{T}} \end{bmatrix}. \tag{4.17}$$

The sum of the attention scores for each key $\boldsymbol{K}_{i,:}$ is given by

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \boldsymbol{A} \tag{4.18}$$

where $\boldsymbol{S} \in \mathbb{R}^{1\times(b+1)}$. Then, we consider the following two cases:

1. If the average of the scores that correspond to the encoded fMRI vectors is greater than the score for the null input, the total attention includes the output of the fMRI cross-attention.

2. Else, only the output of the visual cross-attention is considered.

Formally, the total attention $\mathcal{A}_{\text{tot}}$ is given by

$$\mathcal{A}_{\text{tot}} = \begin{cases} \alpha\mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}_{\text{vis}}, \boldsymbol{V}_{\text{vis}}) + (1-\alpha)\mathcal{A}(\boldsymbol{Q}, \boldsymbol{F}^{+}, \boldsymbol{F}^{+}) & \text{if } \frac{1}{b}\sum_{i=1}^{b} \boldsymbol{S}_{1,i} > \boldsymbol{S}_{1,b+1} \\ \mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}_{\text{vis}}, \boldsymbol{V}_{\text{vis}}) & \text{otherwise} \end{cases} \tag{4.19}$$

where $K_{\text{vis}} \equiv V_{\text{vis}}$ is the output of the visual encoder and $\alpha \in (0, 1)$.

4.4.10 Two encoders with cross-modal fusion

A different variant of the architecture discussed in section 4.4.8 is inspired from [42]. Specifically, the visual and the fMRI encoders communicate via cross-attention layers, as shown on figure 4.9. Results for all three variations of the previous sections are presented in table 4.10 with only Ridge Regression being used for lexical expansion, as other methods performed similarly. In the case of cross-modal fusion, scores are slightly worse since the fMRI stream may negatively affect the more important visual stream. Overall, none of the models exhibit any significant improvements over the baseline and minor deviations can be probably attributed to the fMRIs acting as regularization noise.

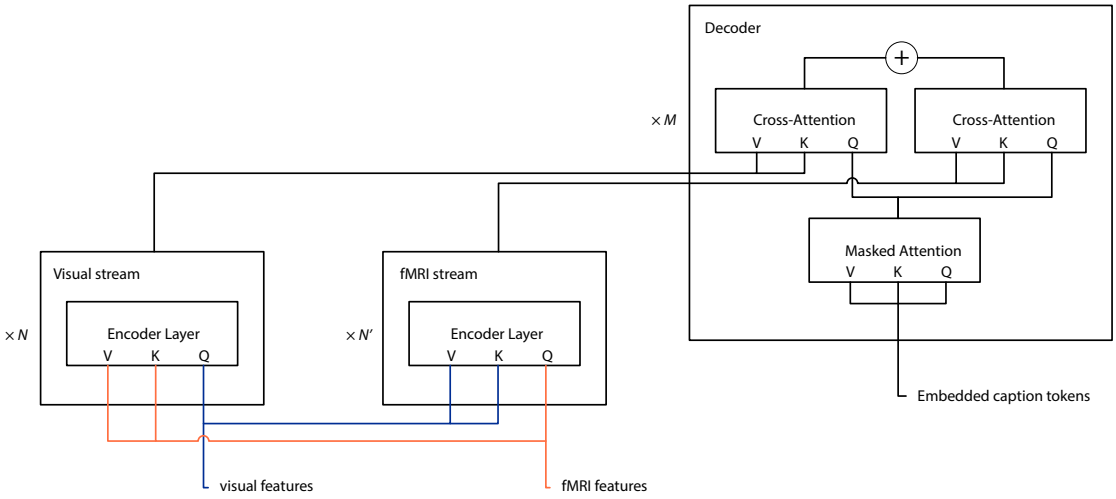


Figure 4.9: Two encoders with cross-modal fusion and "parallel" fusion on the decoder side.

Scores	Baseline model	Two encoders	Two enc. + null input	Two enc. + cross-modal
Bleu_1	0.740	0.736	**0.741**	0.736
Bleu_2	**0.578**	0.574	**0.578**	0.572
Bleu_3	**0.445**	0.441	0.444	0.440
Bleu_4	**0.343**	0.341	0.342	0.340
METEOR	0.279	**0.283**	0.281	0.279
ROUGE_L	0.558	**0.561**	0.559	0.556
CIDEr	1.117	**1.124**	1.123	1.122
SPICE	0.208	**0.212**	0.209	0.210

Table 4.10: Performance of the two-encoder architecture (with and without the "null" input modification on the decoder) and the two-encoder architecture with cross-modal fusion. Regression has been used for lexical expansion from visual features extracted from bounding boxes.

4.4.11 Oracle

All previous experiments did not hint to any significant improvements with the usage of fMRI data. In order to assess whether fMRIs have information which is difficult to be incorporated into the captioning model, we perform an oracle experiment by computing the best attainable CIDEr score for each image with respect to the baseline model and another proposed model. Formally, the score of the i-th image is given by

$$s_i = \max(s_i^{(b)}, s_i^{(t)}) \tag{4.20}$$

where the superscript (b) denotes the baseline model and the superscript (t) denotes a proposed model. Results are presented in table 4.11. It is interesting to note that the model using two separate encoders, described in section 4.4.8, achieves almost identical results when trained using fMRI features and when trained using only visual features on the second encoder. This suggests that the improvement observed in the case of the fMRI features can be attributed to the visual features from which they were derived. The model of section 4.4.7, labeled "fmri add" performs similarly to its counterpart "random add", where random noise has been used instead of fMRIs, indicating that it is ineffective at capturing the information potentially provided by the fMRIs.

Scores	Baseline	Baseline + Two encoders (fMRI)	Baseline + Two encoders (visual)	Baseline + fMRI add	Baseline + random add
Bleu_1	0.740	0.777	0.777	0.772	0.768
Bleu_2	0.578	0.629	0.629	0.623	0.617
Bleu_3	0.445	0.500	0.500	0.492	0.488
Bleu_4	0.343	0.396	0.396	0.388	0.385
METEOR	0.279	0.303	0.303	0.298	0.297
ROUGE_L	0.558	0.595	0.594	0.587	0.587
CIDEr	1.117	1.290	1.285	1.254	1.246
SPICE	0.208	0.228	0.227	0.223	0.222

Table 4.11: Best scores by combining the original model (first column) with other models: two encoders (model from figure 4.8), "fmri add" (addition of a global fMRI vector to the queries of the decoder), "random add" (addition of a random vector to the queries of the decoder).

4.4.12 fMRI Reconstruction

Finally, another idea worth exploring is to add an auxiliary task of predicting fMRI features using the representations of an early layer of the encoder. The total loss of the model becomes

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \alpha \mathcal{L}_{\text{fmri}} \tag{4.21}$$

where $\mathcal{L}_{\text{cap}}$ is the smoothed cross-entropy loss associated with the captioning task, $\mathcal{L}_{\text{fmri}}$ is the regression loss associated with the reconstruction of fMRI features and $\alpha > 0$ is a tunable parameter. Figure 4.10 illustrates this architecture and table 4.12 displays the results. We observe that performance is similar to the baseline model, which can be

attributed either to the fact that the fMRI reconstruction task overfits and generalizes at a different rate [43] or to the weak fMRI represenations.

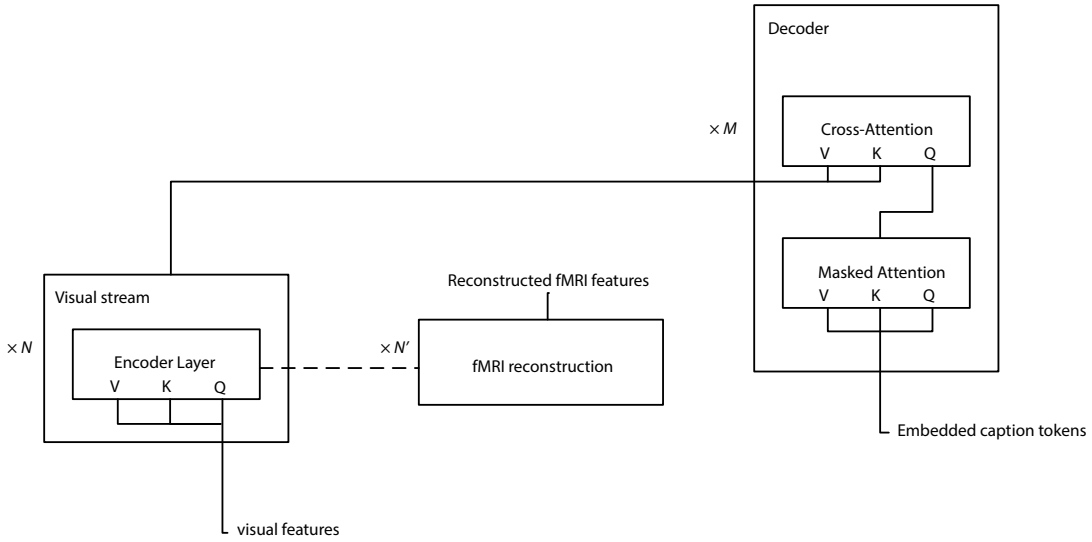


Figure 4.10: fMRI reconstruction task

Scores	Baseline model	Ridge Regression	Dictionary Learning	Similarity
Bleu_1	0.740	0.734	0.735	0.737
Bleu_2	0.578	0.572	0.572	0.576
Bleu_3	0.445	0.440	0.441	0.443
Bleu_4	0.343	0.339	0.340	0.342
METEOR	0.279	0.283	0.282	0.279
ROUGE_L	0.558	0.559	0.559	0.558
CIDEr	1.117	1.120	1.120	1.121
SPICE	0.208	0.211	0.210	0.208

Table 4.12: Performance on image captioning for the MS-COCO dataset while reconstructing fMRIs from the represenations of the first layer of the encoder. The fMRIs have been predicted with Ridge Regression, Dictionary learning and Similarity, using visual features extracted from the bounding boxes of the image.

CHAPTER 5

Conclusion

5.1 Discussion

Even though Representational Similarity Analysis demonstrated a correlation between fMRI activations and visual embeddings and to a lesser extent with word embeddings, the fMRI signal contains a lot of noise and its fusion to computational models is non trivial, since most of the time it does not improve performance. Among the evaluated methods for encoding visual embeddings to fMRI activations, sparse dictionary learning outperforms ridge regression and the similarity based method. However, the evaluation method that is used in the literature is fairly naive and does not necessarily guarantee that the resulting predicted fMRI activations are of high quality. In addition to this, models that incorporate fMRIs obtained through ridge regression seem to perform slightly better, which may indicate that the resulting fMRI space shares more properties with the visual space it is derived from. With regard to the techniques that were used to incorporate fMRI activations, fusion at the encoder or decoder attention conditioning give similar results, hinting that fMRI features act only as regularization or that the model is unable to extract useful information. When a separate transformer encoder is added for fMRIs, that communicates with the visual encoder via cross-modal attention, performance slightly degrades since the fMRI stream may negatively affect the visual stream. Adaptive methods, such as DropNet, that do not always rely on the visual stream, but allow the exclusive use of the fMRI stream so that it is not neglected by the model, perform poorly. Thus, it is evident that the visual stream should be always used and the fMRI stream only when it is likely that it will contribute valuable information. This is attempted at the "null-input" experiment which, however, does not yield any significant performance improvements. Another interesting observation was due to the oracle experiment, which showed that fMRI activations do not yield any improvement over the

visual representations that they were derived from, suggesting that the fMRI encoding process is far from perfect and ideally it should be improved or eliminated via richer fMRI datasets.

5.2 Future work

Due to the noise present in the fMRI signal and the variability of brain anatomy and functional response across different subjects, it is difficult to obtain robust fMRI representations for use in computational models. One way to mitigate this problem is to aggregate fMRI data from multiple subjects into a single shared response space [44–46]. Moreover, different voxel selection techniques could be explored [38, 47–49] to select more discriminative voxels that result in more informative cognitive representations for subsequent computational tasks. Furthermore, different datasets could be aggregated together to allow for pretraining on fMRI data with network architectures that are used on vision-language pretraining models [50, 51]. This process could lead to improved cognitively based vision-language representations that can be used on downstream tasks such as image captioning. In addition, other adaptive fusion techniques can be applied based on the premise that brain activations can lead to an improvement only in specific instances and not in all cases. From this perspective, it would be interesting to try and interpret what kind of information is contributed by brain activations. This could potentially help in improving the fusion process.

Bibliography

1. T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science* **320**, 1191–1195 (2008). https://doi.org/10.1126/science.1152876. Cited on pages 1, 2, 29, 56 & 59.

2. N. Athanasiou, E. Iosif, and A. Potamianos, "Neural activation semantic models: Computational lexical semantic models of localized neural activations," in *Proceedings of the 27th International Conference on Computational Linguistics*. Aug., 2018. Cited on pages 1, 2, 10, 29, 56, 59 & 67.

3. S. R. Oota, N. Manwani, and B. R. S, "fMRI Semantic Category Decoding using Linguistic Encoding of Word Embeddings," *arXiv:1806.05177 [cs, q-bio]* (2018). http://arxiv.org/abs/1806.05177. arXiv: 1806.05177. Cited on pages 1 & 29.

4. L. Cao and Y. Zhang, *Investigating Lexical and Semantic Cognition by Using Neural Network to Encode and Decode Brain Imaging*, pp. 84–100. 11, 2019. Cited on pages 1, 3, 29, 57 & 60.

5. A. J. Anderson, J. R. Binder, L. Fernandino, C. J. Humphries, L. L. Conant, M. Aguilar, X. Wang, D. Doko, and R. D. S. Raizada, "Predicting Neural Activity Patterns Associated with Sentences Using a Neurobiologically Motivated Model of Semantic Representation," *Cerebral Cortex* cercor;bhw240v1 (2016). http://cercor.oxfordjournals.org/cgi/doi/10.1093/cercor/bhw240. Cited on pages 1, 11, 29, 57 & 67.

6. F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, and E. Fedorenko, "Toward a universal decoder of linguistic meaning from brain activation,". https://doi.org/10.1038/s41467-018-03068-4. Cited on pages 1, 10, 29, 56, 57 & 67.

7. N. Chang, J. Pyles, A. Marcus, A. Gupta, M. Tarr, and E. Aminoff, "BOLD5000,". https://kilthub.cmu.edu/articles/BOLD5000/6459449. https://kilthub.cmu.edu/articles/BOLD5000/6459449. Cited on pages 1 & 29.

8. S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Current biology* **21**, 1641–1646 (2011). Cited on pages 2 & 59.

9. F. Pereira, M. Botvinick, and G. Detre, "Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments," *Artificial intelligence* **194**, 240–252 (2013). Cited on pages 2 & 59.

10. A. Huth, S. Nishimoto, A. Vu, and J. Gallant, "A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain," *Neuron* **76**, 1210–1224 (2012). https://linkinghub.elsevier.com/retrieve/pii/S0896627312009348. Cited on pages 2 & 59.

11. A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature* **532**, 453–458 (2016). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4852309/. Cited on pages 2, 56, 57 & 59.

12. D. E. Stansbury, T. Naselaris, and J. L. Gallant, "Natural scene statistics account for the representation of scene categories in human visual cortex," *Neuron* **79**, 1025–1034 (2013). Cited on pages 2 & 59.

13. T. Cukur, S. Nishimoto, A. G. Huth, and J. L. Gallant, "Attention during natural vision warps semantic representation across the human brain," *Nature neuroscience* **16**, 763–770 (2013). Cited on pages 2 & 59.

14. U. Güçlü and M. A. van Gerven, "Increasingly complex representations of natural movies across the dorsal stream are shared between subjects," *NeuroImage* **145**, 329–336 (2017). Cited on pages 2 & 59.

15. S. Nishida, A. G. Huth, J. L. Gallant, and S. Nishimoto, "Word statistics in large-scale texts explain the human cortical semantic representation of objects, actions, and impressions," in *Society for Neuroscience Annual Meeting*, vol. 333. 2015. Cited on pages 3 & 60.

16. A. J. Anderson, E. Bruni, U. Bordignon, M. Poesio, and M. Baroni, "Of Words, Eyes and Brains: Correlating Image-Based Distributional Semantic Models with Neural Representations of Concepts,". Cited on pages 3 & 60.

17. A. J. Anderson, D. Kiela, S. Clark, and M. Poesio, "Visually Grounded and Textual Semantic Models Differentially Decode Brain Activity Associated with Concrete and Abstract Nouns," *Transactions of the Association for Computational Linguistics* **5**, 17–30 (2017). https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00043. Cited on pages 3 & 60.

18. V. Rowtula, S. R. Oota, M. Gupta, and R. S. Bapi, "A Deep Autoencoder for Near-Perfect fMRI Encoding,". Cited on pages 3 & 60.

19. C. Davis, L. Bulat, A. L. Vero, and E. Shutova, "Deconstructing multimodality: visual properties and visual context in human semantic processing," in *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pp. 118–124. Association for Computational Linguistics, Minneapolis, Minnesota, 2019. https://www.aclweb.org/anthology/S19-1013. Cited on pages 3 & 60.

20. S. Wang, J. Zhang, H. Wang, N. Lin, and C. Zong, "Fine-grained neural decoding with distributed word representations," *Information Sciences* **507**, 256–272 (2020). https://linkinghub.elsevier.com/retrieve/pii/S0020025519307820. Cited on pages 3 & 60.

21. N. Chang, J. A. Pyles, A. Marcus, A. Gupta, M. J. Tarr, and E. M. Aminoff, "BOLD5000, a public fMRI dataset while viewing 5000 visual images," *Scientific Data* **6**, 49 (2019). https://www.nature.com/articles/s41597-019-0052-3. Cited on pages 4, 60 & 61.

22. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer. 2014. Cited on pages 5 & 62.

23. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *arXiv:1707.07998 [cs]* (2018). http://arxiv.org/abs/1707.07998. Cited on pages 5, 6 & 62.

24. M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," *arXiv:1912.08226 [cs]* (2020). http://arxiv.org/abs/1912.08226. Cited on pages 5 & 62.

25. Y. Pan, T. Yao, Y. Li, and T. Mei, "X-Linear Attention Networks for Image Captioning,". Cited on pages 5 & 62.

26. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical Sequence Training for Image Captioning," *arXiv:1612.00563 [cs]* (2017). http://arxiv.org/abs/1612.00563. Cited on pages 5 & 62.

27. D. Wang, D. Beck, and T. Cohn, "On the Role of Scene Graphs in Image Captioning," in *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pp. 29–34. Association for Computational Linguistics, Hong Kong, China, 2019. https://www.aclweb.org/anthology/D19-6405. Cited on pages 5 & 62.

28. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137. 2015. Cited on pages 5 & 62.

29. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318. 2002. Cited on pages 5, 46 & 62.

30. S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72. 2005. Cited on pages 5, 46 & 62.

31. C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81. 2004. Cited on pages 5, 47 & 62.

32. R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575. 2015. Cited on pages 5, 48 & 62.

33. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*, pp. 382–398, Springer. 2016. Cited on pages 5, 49 & 62.

34. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv:1706.03762 [cs]* (2017). http://arxiv.org/abs/1706.03762. arXiv: 1706.03762 version: 5. Cited on pages 6, 31, 42, 43, 44 & 62.

35. N. Kriegeskorte, M. Mur, and P. A. Bandettini, "Representational similarity analysis-connecting the branches of systems neuroscience," *Frontiers in systems neuroscience* **2**, 4 (2008). Cited on pages 6, 55 & 64.

36. J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar, 2014. http://aclweb.org/anthology/D14-1162. Cited on pages 8 & 65.

37. S. Takada, R. Togo, T. Ogawa, and M. Haseyama, "Estimating Viewed Images with Natural Language Question Answering from fMRI Data," in *2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)*, pp. 99–100. IEEE, Kyoto, Japan, Mar., 2020. https://ieeexplore.ieee.org/document/9080971/. Cited on pages 10 & 67.

38. L. Beinborn, S. Abnar, and R. Choenni, "Robust Evaluation of Language-Brain Encoding Experiments," *arXiv:1904.02547 [cs]* (2019), arXiv:1904.02547 [cs]. http://arxiv.org/abs/1904.02547. Cited on pages 10, 11, 27, 56, 67, 68 & 82.

39. J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu, "Incorporating BERT into Neural Machine Translation," *arXiv:2002.06823 [cs]* (2020). http://arxiv.org/abs/2002.06823. arXiv: 2002.06823. Cited on pages 16 & 72.

40. J. Libovický, J. Helcl, and D. Mareček, "Input Combination Strategies for Multi-Source Transformer Decoder," *arXiv:1811.04716 [cs]* (2018). http://arxiv.org/abs/1811.04716. Cited on pages 20 & 76.

41. A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf, "Recurrent Independent Mechanisms," *arXiv:1909.10893 [cs, stat]* (2020). http://arxiv.org/abs/1909.10893. Cited on pages 21 & 76.

42. J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," *arXiv:1908.02265 [cs]* (2019). http://arxiv.org/abs/1908.02265. Cited on pages 22 & 77.

43. W. Wang, D. Tran, and M. Feiszli, "What Makes Training Multi-Modal Classification Networks Hard?," *arXiv:1905.12681 [cs]* (2020). http://arxiv.org/abs/1905.12681. Cited on pages 25 & 80.

44. P.-H. Chen, X. Zhu, H. Zhang, J. S. Turek, J. Chen, T. L. Willke, U. Hasson, and P. J. Ramadge, "A Convolutional Autoencoder for Multi-Subject fMRI Data Aggregation," *arXiv:1608.04846 [cs, stat]* (2016). http://arxiv.org/abs/1608.04846. arXiv: 1608.04846. Cited on pages 27 & 82.

45. S. A. Nastase, Y.-F. Liu, H. Hillman, K. A. Norman, and U. Hasson, "Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space," *NeuroImage* **217**, 116865 (2020). https://linkinghub.elsevier.com/retrieve/pii/S1053811920303517. Cited on pages 27 & 82.

46. H. Wen, J. Shi, W. Chen, and Z. Liu, "Transferring and generalizing deep-learning-based neural encoding models across subjects," *NeuroImage* **176**, 152–163 (2018). https://linkinghub.elsevier.com/retrieve/pii/S105381191830363X. Cited on pages 27 & 82.

47. L. Tarhan and T. Konkle, "Reliability-based voxel selection," *NeuroImage* **207**, 116350 (2020). https://www.sciencedirect.com/science/article/pii/S1053811919309413. Cited on pages 27, 55 & 82.

48. O. Yamashita, M.-a. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns," *NeuroImage* **42**, 1414–1429 (2008). https://linkinghub.elsevier.com/retrieve/pii/S1053811908006940. Cited on pages 27, 54 & 82.

49. Y. Wang, Z. Li, Y. Wang, X. Wang, J. Zheng, X. Duan, and H. Chen, "A Novel Approach for Stable Selection of Informative Redundant Features from High Dimensional fMRI Data,". Cited on pages 27 & 82.

50. L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," *arXiv:1908.03557 [cs]* (2019). http://arxiv.org/abs/1908.03557. Cited on pages 27 & 82.

51. Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: Learning UNiversal Image-TExt Representations,". https://openreview.net/forum?id=S1eL4kBYwr. Cited on pages 27 & 82.

52. T. M. Mitchell *et al.*, "Machine learning,". Cited on page 31.

53. J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane, "Automated design of both the topology and sizing of analog electrical circuits using genetic programming," in *Artificial Intelligence in Design'96*, pp. 151–170. Springer, 1996. Cited on page 31.

54. E. Alpaydin, *Introduction to machine learning*. MIT press, 2020. Cited on page 31.

55. J. Hu, H. Niu, J. Carrasco, B. Lennox, and F. Arvin, "Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning," *IEEE Transactions on Vehicular Technology* **69**, 14413–14423 (2020). Cited on page 31.

56. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems* **25**, 1097–1105 (2012). Cited on pages 31 & 39.

57. D. Yu and L. Deng, *Automatic Speech Recognition*. Springer, 2016. Cited on page 31.

58. R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. Cited on page 31.

59. C. M. Bishop, "Pattern recognition," *Machine learning* **128**, (2006). Cited on page 31.

60. D. A. Freedman, *Statistical models: theory and practice*. cambridge university press, 2009. Cited on page 32.

61. A. C. Rencher and W. F. Christensen, "Chapter 10, multivariate regression–section 10.1, introduction," *Methods of multivariate analysis, Wiley Series in Probability and Statistics* **709**, 19 (2012). Cited on page 32.

62. D. E. Hilt and D. W. Seegrist, *Ridge, a computer program for calculating ridge regression estimates*, vol. 236. Department of Agriculture, Forest Service, Northeastern Forest Experiment …, 1977. Cited on page 33.

63. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013. Cited on page 33.

64. M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing* **54**, 4311–4322 (2006). Cited on page 33.

65. A. M. Tillmann, "On the computational intractability of exact and approximate dictionary learning," *IEEE Signal Processing Letters* **22**, 45–49 (2014). Cited on page 34.

66. D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* **59**, 797–829 (2006). Cited on page 34.

67. I. Jolliffe, "Principal component analysis," *Encyclopedia of statistics in behavioral science* (2005). Cited on page 34.

68. I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016. Cited on pages 36, 39 & 40.

69. G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems* **2**, 303–314 (1989). https://doi.org/10.1007/BF02551274. Cited on page 38.

70. K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics* **36**, 193–202 (1980). https://doi.org/10.1007/BF00344251. Cited on page 38.

71. Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object Recognition with Gradient-Based Learning," Lecture Notes in Computer Science, pp. 319–345. Springer, Berlin, Heidelberg, 1999. https://doi.org/10.1007/3-540-46805-6_19. Cited on page 38.

72. D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology* **160**, 106–154.2 (1962). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1359523/. Cited on page 38.

73. V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, p. 807–814. Omnipress, Madison, WI, USA, 2010. Cited on page 38.

74. J. Weng, N. Ahuja, and T. Huang, "Learning recognition and segmentation of 3-D objects from 2-D images," in *1993 (4th) International Conference on Computer Vision*, pp. 121–128. May, 1993. Cited on page 39.

75. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]* (2015). http://arxiv.org/abs/1512.03385. arXiv: 1512.03385. Cited on page 39.

76. A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. 2020. https://d2l.ai. Cited on page 40.

77. "Recurrent neural network." Wikipedia, June, 2021. https://en.wikipedia.org/w/index.php?title=Recurrent_neural_network&oldid=1027494214. Page Version ID: 1027494214. Cited on pages 40 & 41.

78. J. L. Elman, "Finding Structure in Time," *Cognitive Science* **14**, 179–211 (1990). https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1. Cited on page 40.

79. Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks* **5**, 157–166 (1994). Cited on page 40.

80. R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, pp. 1310–1318. PMLR, May, 2013. http://proceedings.mlr.press/v28/pascanu13.html. Cited on page 41.

81. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735. Cited on page 41.

82. X. Li and X. Wu, "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition," *arXiv:1410.4281 [cs]* (2015). http://arxiv.org/abs/1410.4281. arXiv: 1410.4281. Cited on page 41.

83. M. Chen, G. Ding, S. Zhao, H. Chen, Q. Liu, and J. Han, "Reference Based LSTM for Image Captioning," *Proceedings of the AAAI Conference on Artificial Intelligence* **31**, (2017). https://ojs.aaai.org/index.php/AAAI/article/view/11198. Cited on page 41.

84. C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A C-LSTM Neural Network for Text Classification," *arXiv:1511.08630 [cs]* (2015). http://arxiv.org/abs/1511.08630. arXiv: 1511.08630. Cited on page 41.

85. F. Gers and E. Schmidhuber, "LSTM recurrent networks learn simple context-free and context-sensitive languages," *IEEE Transactions on Neural Networks* **12**, 1333–1340 (2001). Cited on page 42.

86. Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *arXiv:1807.05511 [cs]* (2019). http://arxiv.org/abs/1807.05511. arXiv: 1807.05511. Cited on page 45.

87. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *arXiv:1311.2524 [cs]* (2014). http://arxiv.org/abs/1311.2524. arXiv: 1311.2524. Cited on page 45.

88. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *arXiv:1506.01497 [cs]* (2016). http://arxiv.org/abs/1506.01497. arXiv: 1506.01497. Cited on page 45.

89. M. D'Esposito, A. Kayser, and A. Chen, "fmri: Applications in cognitive neuroscience," in *fMRI techniques and protocols*, pp. 285–322. Springer, 2009. Cited on page 51.

90. S. A. Huettel, A. W. Song, G. McCarthy, *et al.*, *Functional magnetic resonance imaging*, vol. 1. Sinauer Associates Sunderland, MA, 2004. Cited on page 51.

91. S. C. Strother and N. Churchill, "Neuroimage preprocessing," *Handbook of neuroimaging data analysis* 264–308 (2017). Cited on page 53.

92. N. W. Churchill, A. Oder, H. Abdi, F. Tam, W. Lee, C. Thomas, J. E. Ween, S. J. Graham, and S. C. Strother, "Optimizing preprocessing and analysis pipelines for single-subject fmri. i. standard temporal motion and physiological noise correction methods," *Human brain mapping* **33**, 609–627 (2012). Cited on page 53.

93. N. W. Churchill, R. Spring, B. Afshin-Pour, F. Dong, and S. C. Strother, "An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional mri," *PloS one* **10**, e0131520 (2015). Cited on page 53.

94. J. Zhang, J. R. Anderson, L. Liang, S. K. Pulapura, L. Gatewood, D. A. Rottenberg, and S. C. Strother, "Evaluation and optimization of fmri single-subject processing pipelines with npairs and second-level cva," *Magnetic resonance imaging* **27**, 264–278 (2009). Cited on page 53.

95. O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, *et al.*, "fmriprep: a robust preprocessing pipeline for functional mri," *Nature methods* **16**, 111–116 (2019). Cited on page 53.

96. D. B. Parker and Q. R. Razlighi, "The benefit of slice timing correction in common fmri preprocessing pipelines," *Frontiers in neuroscience* **13**, 821 (2019). Cited on page 54.

97. N. Lazar, *The statistical analysis of functional MRI data*. Springer Science & Business Media, 2008. Cited on page 54.

98. D. Schwartz, M. Toneva, and L. Wehbe, "Inducing brain-relevant bias in natural language processing models," in *Advances in Neural Information Processing Systems 32*, p. 14123–14133. Curran Associates, Inc., 2019. http://papers.nips.cc/paper/9559-inducing-brain-relevant-bias-in-natural-language-processing-models.pdf. Cited on page 54.

99. A. J. Reddy and L. Wehbe, "Syntactic representations in the human brain: beyond effort-based metrics," preprint, June, 2020. http://biorxiv.org/lookup/doi/10.1101/2020.06.16.155499. Cited on page 54.

100. H. Johansen-Berg, T. E. J. Behrens, M. D. Robson, I. Drobnjak, M. F. S. Rushworth, J. M. Brady, S. M. Smith, D. J. Higham, and P. M. Matthews, "Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex," *Proceedings of the National Academy of Sciences* **101**, 13335–13340 (2004), https://www.pnas.org/content/101/36/13335.full.pdf. https://www.pnas.org/content/101/36/13335. Cited on page 55.

101. N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping,". https://doi.org/10.1073/pnas.0600244103. Cited on page 55.

102. A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates," *Proceedings of the National Academy of Sciences* **113**, 7900–7905 (2016), https://www.pnas.org/content/113/28/7900.full.pdf. https://www.pnas.org/content/113/28/7900. Cited on page 55.

103. M. Mur, M. Meys, J. Bodurka, R. Goebel, P. A. Bandettini, and N. Kriegeskorte, "Human object-similarity judgments reflect and transcend the primate-it object representation," *Frontiers in psychology* **4**, 128 (2013). Cited on page 55.

104. K. N. Kay and J. D. Yeatman, "Bottom-up and top-down computations in word- and face-selective cortex," *Elife* **6**, e22341 (2017). Cited on page 55.

105. F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fmri: a tutorial overview," *Neuroimage* **45**, S199–S209 (2009). Cited on page 56.

106. S. Norman-Haignere, N. G. Kanwisher, and J. H. McDermott, "Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition," *Neuron* **88**, 1281–1296 (2015). Cited on page 56.

107. S. Jain and A. G. Huth, "Incorporating Context into Language Encoding Models for fMRI," preprint, May, 2018. http://biorxiv.org/lookup/doi/10.1101/327601. Cited on pages 56 & 57.

108. B. Devereux, C. Kelly, and A. Korhonen, "Using fmri activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora," in *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, CN '10, p. 70–78. Association for Computational Linguistics, USA, 2010. Cited on page 56.

109. A. Jelodar, M. Alizadeh, and S. Khadivi, "Wordnet based features for predicting brain activity associated with meanings of nouns,". Cited on page 56.

110. J. António Rodrigues, R. Branco, J. Silva, C. Saedi, and A. Branco, "Predicting Brain Activation with WordNet Embeddings," in *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pp. 1–5. Association for Computational Linguistics, Melbourne, July, 2018. https://www.aclweb.org/anthology/W18-2801. Cited on page 56.

111. S. Abnar, R. Ahmed, M. Mijnheer, and W. Zuidema, "Experiential, Distributional and Dependency-based Word Embeddings have Complementary Roles in Decoding Brain Activity," in *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pp. 57–66. Association for Computational Linguistics, Salt Lake City, Utah, Jan., 2018. https://www.aclweb.org/anthology/W18-0107. Cited on page 56.

112. L. Bulat, S. Clark, and E. Shutova, "Speaking, Seeing, Understanding: Correlating semantic models with conceptual representation in the brain," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*,

pp. 1081–1091. Association for Computational Linguistics, Copenhagen, Denmark, Sept., 2017. https://www.aclweb.org/anthology/D17-1113. Cited on page 57.

113. N. Hollenstein, A. de la Torre, N. Langer, and C. Zhang, "CogniVal: A Framework for Cognitive Word Embedding Evaluation," *arXiv:1909.09001 [cs]* (2019). http://arxiv.org/abs/1909.09001. arXiv: 1909.09001. Cited on page 57.

114. J. Bingel, M. Barrett, and A. Søgaard, "Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 747–755. Association for Computational Linguistics, Berlin, Germany, Aug., 2016. https://www.aclweb.org/anthology/P16-1071. Cited on page 57.

115. S. Abnar, L. Beinborn, R. Choenni, and W. Zuidema, "Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains," *arXiv:1906.01539 [cs, q-bio]* (2019). http://arxiv.org/abs/1906.01539. arXiv: 1906.01539. Cited on page 57.

116. P. Qian, X. Qiu, and X. Huang, "Bridging LSTM Architecture and the Neural Dynamics during Reading," *arXiv:1604.06635 [cs]* (2016). http://arxiv.org/abs/1604.06635. arXiv: 1604.06635. Cited on page 57.

117. J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science* **293**, 2425–2430 (2001). Cited on page 59.

118. K. N. Kay, "Principles for models of neural information processing," *NeuroImage* **180**, 101–109 (2018). Cited on page 59.