ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Κυβερνοασφάλεια και Επιστήμη Δεδομένων»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Αυτόματη Περιγραφή Οπτικοακουστικών Σκηνών με χρήση Βαθέων Νευρωνικών Δικτύων.** **Automatic Description of Audiovisual Scenes using Deep Neural Networks.**
Ονοματεπώνυμο Φοιτητή	**Ασαντούρ Βαρτιάν**
Πατρώνυμο	**Χαρουτιούν**
Αριθμός Μητρώου	**ΜΠΚΕΔ 21003**
Επιβλέπων	**Άγγελος Πικράκης, Επίκουρος Καθηγητής**

Ημερομηνία Παράδοσης: **Δεκέμβριος 2023**

Τριμελής Εξεταστική Επιτροπή

Άγγελος Πικράκης Επίκουρος Καθηγητής	Μιχαήλ Ψαράκης Αναπληρωτής Καθηγητής	Διονύσιος Σωτηρόπουλος Επίκουρος Καθηγητής

Ευχαριστίες

Η παρούσα μεταπτυχιακή διατριβή ήταν μια πραγματική πρόκληση για εμένα, καθώς εμπεριείχε αρκετές δυσκολίες πλην όμως μέσα από τις οποίες ήρθα σε επαφή με ποικίλα και ενδιαφέροντα ερεθίσματα. Φυσικά, σε όλες τις προκλήσεις/δυσκολίες που κατά καιρούς αντιμετώπισα είχα την στήριξη του επιβλέποντος καθηγητή μου κ. Άγγελου Πικράκη, ο οποίος με τις πολύτιμες συμβουλές καθώς και την έμπειρη καθοδήγησή του συνέβαλε καθοριστικά στην αντιμετώπιση καθενός από τα εκάστοτε προβλήματα. Δεν θα μπορούσα να μην εκφράσω ένα μεγάλο ευχαριστώ για όλη τη βοήθειά του, το πλήθος συζητήσεων αλλά και τις σχολαστικές αναλύσεις που μου παρείχε όλους αυτούς τους μήνες.

Τέλος, το μεγαλύτερο ευχαριστώ αναλογεί στην οικογένειά μου. Γνωρίζοντας ότι η εκπόνηση της μεταπτυχιακής αυτής διατριβής αποτελούσε ένα στόχο ήδη από την εποχή που ολοκλήρωσα τις προπτυχιακές μου σπουδές, η στήριξη που μου προσέφεραν και που συνεχίζουν να μου προσφέρουν σε κάθε νέο μου βήμα ήταν και είναι αμέριστη.

Σύνοψη

Η Μηχανική Μάθηση είναι ένα ταχέως αναπτυσσόμενο πεδίο της πληροφορικής, ικανό να παρέχει λύσεις σε απαιτητικά προβλήματα αυξανόμενης πολυπλοκότητας. Στο πλαίσιο αυτό, ο στόχος αυτής της διατριβής είναι να οικοδομηθεί ένα σύστημα με σκοπό την αυτόματη περιγραφή της σκηνής βίντεο χρησιμοποιώντας μια αλληλουχία συνδυαστικών μοντέλων μηχανικής μάθησης. Για το σκοπό αυτό, ένα σήμα βίντεο αντιμετωπίζεται ως μια ακολουθία εικόνων και κάθε εικόνα τροφοδοτείται ως είσοδος σε μια αρχιτεκτονική CLIP που αυτή τελικά παράγει μια περιγραφή εικόνας. Το CLIP είναι ένα embedding ανοιχτού κώδικα εκπαιδευμένο να συσχετίζει την εικόνα με το κείμενο. Σε επόμενο βήμα, η ακολουθία των παραγόμενων περιγραφών συνενώνεται και δίνεται ως είσοδος σε ένα μοντέλο μετασχηματιστή που παράγει την τελική περιγραφή της σκηνής του βίντεο. Προκειμένου να επιτύχουμε καλύτερα αποτελέσματα σε αυτό το δεύτερο στάδιο επεξεργασίας, επανεκπαιδεύσαμε και βελτιστοποιήσαμε τα μοντέλα μετασχηματιστών τους BART και Pegasus χρησιμοποιώντας το Σύνολο Δεδομένων LSMDC (Large Scale Movie Description Challenge Dataset). Η απόδοση του προτεινόμενου συστήματος αξιολογήθηκε χρησιμοποιώντας διάφορες καθιερωμένες μετρικές.

Abstract

Machine Learning is a rapidly growing field of informatics, capable of providing solutions to demanding problems of increasing complexity. In that context, the goal of this thesis is to build a system for the purposed of automatic video scene description using machine learning pipeline. To this end, a video signal is treated as a sequence of images and each image is fed as input to a CLIP architecture which generated an image description. CLIP is an open-source embedding
trained to associate image with text. At a next step, the sequence of generated descriptions is concatenated and it is given as input to a transformer model which produces the final description of the video scene. In order to obtain better results at this second processing stage, we re-trained and fine-tuned the BART and Pegasus transformer models using the Large Scale Movie Description Challenge Dataset (LSMDC). The performance of the the proposed pipeline was assessed using various established metrics.

Πίνακας Περιεχομένων

1 Εισαγωγή

1.1 Ορισμός προβλήματος

Η αυτόματη περιγραφή βίντεο αποτελεί ένα ανοιχτό πρόβλημα στον τομέα της τεχνητής νοημοσύνης και της υπολογιστικής όρασης, όπου συνεχώς προτείνονται λύσεις και τεχνικές που θα μπορούσαν να εφαρμοστούν. Υπάρχουν διάφορες τεχνικές που έχουν προταθεί για αυτόματη περιγραφή (auto-description) των βίντεο. Ορισμένες από αυτές περιλαμβάνουν:

- Χρήση μοντέλων βαθιάς μάθησης (Deep Learning): Με την εμφάνιση των προηγμένων μοντέλων βαθιάς μάθησης, όπως τα μοντέλα μετάδοσης μάθησης (π.χ. από την προεκπαίδευση σε μεγάλα σύνολα εικόνων/κειμένου) και τα μοντέλα βασισμένα στους μετασχηματιστές (Transformers), έχουν επιτευχθεί εντυπωσιακά αποτελέσματα στο κομμάτι παραγωγής περιγραφής των βίντεο.
- Χρήση πολυμορφισμού (zero-shot learning): Με τη χρήση μοντέλων όπως το CLIP που είναι εκπαιδευμένα να κατανοούν και εικόνες και κείμενο, έχει δοκιμαστεί η δυνατότητα να περιγράφουν αυτόματα βίντεο ακολουθώντας φυσικά την απαραίτητη διαδικασία.
- Συνδυασμός πολλαπλών μοντέλων: Ορισμένες προσεγγίσεις συνδυάζουν πολλαπλά μοντέλα και τεχνικές για αυτόματη περιγραφή των βίντεο, πετυχαίνοντας έτσι καλύτερη ακρίβεια και απόδοση.

Παρόλο που υπάρχουν πολλές προσπάθειες και μέθοδοι που αναφέρονται παραπάνω, η διαδικασία παραγωγής περιγραφής των βίντεο εξακολουθεί να είναι μια πολύπλοκη και δύσκολη πρόκληση στον κλάδο της τεχνητής νοημοσύνης και απαιτεί συνεχή έρευνα και εύρεση καινοτόμων λύσεων προκειμένου να βελτιώνεται και να επεκτείνεται.

Όλη η υλοποίηση βασίζεται στην έννοια της διαδικασίας αυτόματης παραγωγής περιγραφών για εικόνες. Αυτή η τεχνική συνδυάζει την ανάλυση εικόνας και τη γλωσσική κατανόηση για να παράγει ακριβείς περιγραφές του περιεχομένου μιας εικόνας [11]. Οι χρήσεις της διαδικασίας παραγωγής λεζάντας εικόνων (image-captioning) περιλαμβάνουν:

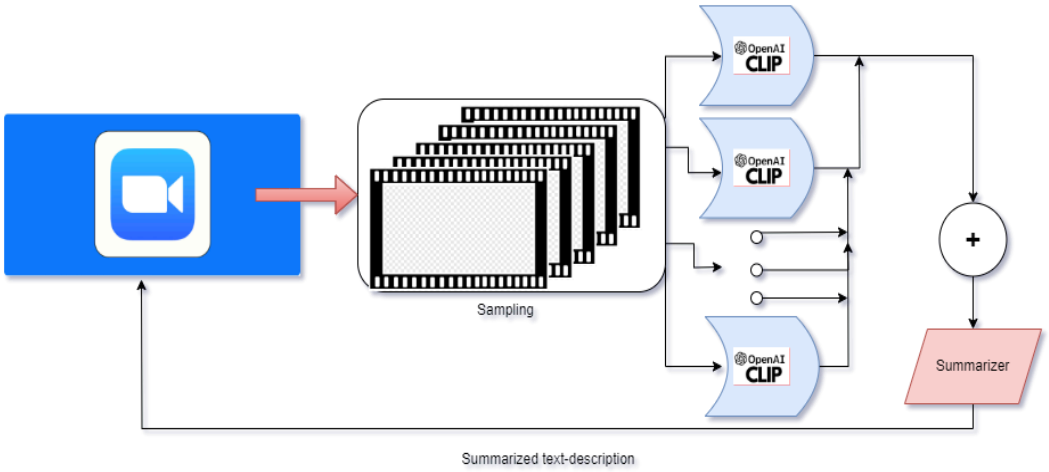
- Περιγραφή εικόνων: Αυτό μπορεί να είναι χρήσιμο για αυτοματοποιημένες εφαρμογές που απαιτούν περιγραφές εικόνων.
- Ανάκτηση πληροφοριών: Η παραγωγή περιγραφών για εικόνες μπορεί να βοηθήσει στην ανάκτηση πληροφοριών. Για παράδειγμα, σε έναν ψηφιακό φωτογραφικό κατάλογο, οι περιγραφές εικόνων μπορούν να χρησιμοποιηθούν για την αναζήτηση εικόνων βάσει περιεχομένου.
- Κοινωνικά μέσα και διαμοιρασμός εικόνων: Οι περιγραφές εικόνων μπορούν να χρησιμοποιηθούν για τη βελτίωση της διαδικασίας της αναζήτησης εικόνων και την παροχή πληροφοριών για εικόνες σε κοινωνικά μέσα και πλατφόρμες socializing.
- Βοήθημα για ανθρώπους με προβλήματα όρασης: Οι λεζάντες εικόνων που παράγονται μπορούν να χρησιμοποιηθούν ως βοήθημα για ανθρώπους με προβλήματα όρασης. Μπορεί να επιτρέψει σε αυτούς να αναγνωρίζουν και να κατανοούν το περιεχόμενο των εικόνων, φυσικά με χρήση text-to-voice τεχνικών.
- Εφαρμογές πραγματικού χρόνου: Η παραγωγή λεζάντας εικόνας μπορεί να χρησιμοποιηθεί σε εφαρμογές πραγματικού χρόνου, όπως ρομποτική ή επαυξημένη πραγματικότητα, για να παράγει αυτόματα περιγραφές για το περιβάλλον που παρουσιάζεται μπροστά στο σύστημα.

Αυτές είναι μερικές από τις κύριες χρήσεις παραγωγής λεζάντας. Η τεχνική αυτή έχει ευρεία χρήση και μπορεί να βοηθήσει σε πολλούς τομείς όπου απαιτείται η αντιστοίχιση εικόνων με τις περιγραφές τους, αλλά και το αντίστροφο.

Στην παρούσα εργασία λοιπόν κληθήκαμε να κατασκευάσουμε ένα σύστημα παραγωγής περιγραφής για οποιοδήποτε οπτικο-ακουστικό υλικό (βίντεο) με αυτόματο τρόπο με χρήση νευρωνικών δικτύων. Το αποτέλεσμα της δουλειάς αυτής μπορεί να γίνει απαραίτητη σε πληθώρα εφαρμογών, όπως για παράδειγμα ο θεατής μιας ταινίας/σειράς να μπορεί γρήγορα να μελετήσει την υπόθεση της ταινίας και να κρίνει εάν του αρέσει ή όχι, ακόμα και να καθοριστεί και το είδος της ταινίας βάσει της παραγόμενης περιγραφής, ώστε να καθορίζεται εύκολα η ηλικιακή ομάδα που μπορεί να παρακολουθήσει την εκάστοτε ταινία/σειρά, αλλά φυσικά κι άλλες πολλές χρήσεις.

Η γενική ιδέα της υλοποίησης του συστήματος αυτού είναι η εξής: Δειγματοληπτούμε το οπτικο-ακουστικό υλικό που έχουμε στην κατοχή μας. Στη συνέχεια, τα καρέ που λαμβάνουμε από την δειγματοληψία τα εισάγουμε ως είσοδο στο CLIP (της OpenAI), το οποίο μας παράγει την περιγραφή για κάθε καρέ. Το σύνολο των περιγραφών των καρέ συγκεντρώνονται και αποτελούν είσοδο πλέον σε έναν συνοψιστή (summarizer), ώστε να παραχθεί η τελική περίληψη όλων των περιγραφών που έχουμε συγκεντρώσει. Το τελικό αποτέλεσμα που προκύπτει αποτελεί πλέον την περιγραφή (description) του οπτικο-ακουστικού υλικού.

Ας περιγράψουμε τώρα πως δουλέψαμε για να φτάσουμε στο στάδιο να υλοποιηθεί το παραπάνω σύστημα που αναφέρουμε και η διαδικασία είναι η εξής: Αρχικά, συγκεντρώσαμε μερικά δοκιμαστικά βίντεο αρκετά μικρής διάρκειας, κάναμε δειγματοληψία για το κάθε ένα, εισάγαμε κάθε καρέ (frame) ως είσοδο στο CLIP, μας παρήγαγε μια περιγραφή για κάθε ένα και στην συνέχεια το σύνολο των περιγραφών τα εισάγαμε σε έναν συνοψιστή (summarizer) για να προκύψει η τελική περιγραφή του βίντεο.



Εικόνα 1 : Η αρχιτεκτονική του συστήματός μας

Στη συνέχεια όλη αυτή η διαδικασία ξεκίνησε να υλοποιείται και πιο επίσημα σε ένα πολύ μεγάλο σύνολο δεδομένων (Large Scale Movie Description Challenge Dataset), το οποίο περιλάμβανε έτοιμες περιγραφές για κάθε απόσπασμα βίντεο, αλλά και το ίδιο το απόσπασμα. Το συγκεκριμένο σύνολο δεδομένων ήταν χωρισμένο σε σύνολα train, validation και test. Για κάθε απόσπασμα που διαθέτει το κάθε σύνολο παρήξαμε νέες περιγραφές με τον ίδιο τρόπο που αναφέραμε λίγο παραπάνω. Στη συνέχεια εφαρμόσαμε κάποιες μετρικές για να κάνουμε μια συνολική ανάλυση των δεδομένων που είχαμε στην διάθεσή μας σε εκείνο το στάδιο. Κάπου εδώ να σημειώσουμε ότι η περιγραφή για κάθε απόσπασμα σε όλες τις περιπτώσεις προκύπτει μόνο από το οπτικό υλικό (frame) και όχι κι από το ακουστικό. Πιο συγκεκριμένα δε κάνουμε κάποια ανάλυση του ήχου (voice-to-text) για να εξάγουμε επιπλέον χαρακτηριστικά για το βίντεο. Στα επόμενα κεφάλαια που ακολουθούν εξηγούμε αναλυτικά ποια η λειτουργία των CLIP και των μετασχηματιστών, που είναι τα κύρια μέρη της εργασίας μας, όπως επίσης ποια είναι η αρχιτεκτονική του καθενός και οι χρήσεις τους σε γενικό πλαίσιο.

Τελικό στάδιο της εργασίας μας ήταν να δημιουργήσουμε και να εκπαιδεύσουμε - βελτιώσουμε (fine-tuning) έναν δικό μας συνοψιστή χρησιμοποιώντας ως σύνολο δεδομένων το ίδιο το Large Scale Movie Description Challenge Dataset (LSMDC), όπου αναφέρεται παραπάνω, το οποίο περιέχει δεδομένα για ένα μεγάλο πλήθος ταινιών. Αναλυτικότερα, πρόκειται για ένα μεγάλο σύνολο δεδομένων που έχει συγκεντρωθεί από διάφορες πηγές, συμπεριλαμβανομένων των ιστοσελίδων βαθμολογίας ταινιών, public βάσεων δεδομένων και άλλων διαθέσιμων πηγών.

Το εν λόγω σύνολο έχει χρησιμοποιηθεί ευρέως στον τομέα της επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP) και της ανάπτυξης συστημάτων συστάσεων για ταινίες. Οι ερευνητές και οι προγραμματιστές μπορούν να χρησιμοποιήσουν αυτό το σύνολο δεδομένων για να αναπτύξουν αλγορίθμους και μοντέλα που εξάγουν πληροφορίες από τις ταινίες, να πραγματοποιήσουν αναζητήσεις, να κατανοήσουν τις προτιμήσεις των χρηστών και να προτείνουν ταινίες που ενδέχεται να τους αρέσουν κ.ο.κ.

1.2 Σχετικές Εργασίες

Υπάρχουν πολλά ενδιαφέροντα έργα που έχουν υλοποιηθεί στον τομέα της περίληψης βίντεο, χρησιμοποιώντας διάφορες τεχνικές και προσεγγίσεις. Μερικά από αυτά περιλαμβάνουν:

1. YouTube Automatic Video Chapters [23]: Το YouTube εισήγαγε αυτόματα κεφάλαια σε ορισμένα βίντεο με βάση το περιεχόμενό τους, επιτρέποντας στους χρήστες να πλοηγούνται ευκολότερα μέσα στο βίντεο.

2. DeepMind's "Summarization TV": Ένα έργο της εταιρίας DeepMind που χρησιμοποιεί μηχανική μάθηση για τη δημιουργία περιλήψεων βίντεο, εστιάζοντας σε συγκεκριμένα σημεία ή γεγονότα.

3. News Video Summarization: Αρκετές εφαρμογές έχουν αναπτυχθεί για την αυτόματη περίληψη ειδησεογραφικών βίντεο, επιτρέποντας στους χρήστες να λάβουν γρήγορα ενημέρωση για ειδήσεις.

4. Real-Time Video Summarization: Κάποια έργα επικεντρώνονται στη δημιουργία περιλήψεων από ροή βίντεο σε πραγματικό χρόνο, επιτρέποντας την ταχεία εξαγωγή συνόψεων βίντεο.

Τα παραπάνω έργα αντιπροσωπεύουν μερικές από τις ποικίλες προσεγγίσεις και εφαρμογές της τεχνολογίας περίληψης βίντεο που έχουν αναπτυχθεί σε διάφορους τομείς.

2 CLIP

2.1 Τι είναι το CLIP

Όπως αναφέραμε και στην εισαγωγή, το CLIP (Contrastive Language-Image Pretraining) είναι ένα μοντέλο μηχανικής μάθησης που αναπτύχθηκε από την OpenAI [2]. Αυτό το μοντέλο συνδυάζει την επεξεργασία φυσικής γλώσσας (NLP) με την ανάλυση εικόνων, προκειμένου να αντιληφθεί και να αναπαραστήσει τη σχέση μεταξύ κειμένου και εικόνας [3].

Ο σκοπός του CLIP είναι να εκπαιδευθεί σε αναπαραστάσεις που συνδέουν την εννοιολογική σημασία του κειμένου με το περιεχόμενο της εικόνας [3]. Το μοντέλο εκπαιδεύεται πάνω σε ένα τεράστιο σύνολο δεδομένων, που περιέχει ζεύγη κειμένων και εικόνων από τον ιστό. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο "κωδικοποιεί" την πληροφορία από το κείμενο και την εικόνα σε έναν κοινό χώρο αναπαραστάσεων.

Η συμπεριφορά του CLIP βασίζεται στην αρχή της αντίθεσης (contrastive learning), όπου το μοντέλο εκπαιδεύεται να αναγνωρίζει τις θετικές συσχετίσεις (το πώς το κείμενο συσχετίζεται με την

εικόνα του) και να απομονώνει τις αρνητικές συσχετίσεις (το πώς διαφέρει το κείμενο από την εικόνα) [1]. Αυτή η αντίθεση βοηθά το μοντέλο να αναπαραστήσει σημασιολογικές συνδέσεις μεταξύ διαφορετικών κειμένων και εικόνων.

Μετά την εκπαίδευση, το CLIP μπορεί να χρησιμοποιηθεί για πολλές εφαρμογές, όπως είναι η αναζήτηση εικόνων βάσει κειμένου, η κατάταξη εικόνων με βάση το περιεχόμενό τους, η περιγραφή εικόνων, η αναζήτηση κειμένων βάσει εικόνων και άλλες προβλέψεις που συνδέουν κείμενο και εικόνες.

Το CLIP αντιπροσωπεύει ένα σημαντικό βήμα στην ανάπτυξη της αντίληψης των υπολογιστών για τον τρόπο με τον οποίο αλληλεπιδρούν οι άνθρωποι με τον κόσμο τους, συνδυάζοντας κείμενο και εικόνες.

2.2 Χρήσεις του CLIP

Το CLIP (Contrastive Language-Image Pretraining) της OpenAI είναι ένα μοντέλο μηχανικής μάθησης που εκπαιδεύεται να κατανοεί τόσο τις εικόνες όσο και το κείμενο. Έχει πολλές χρήσεις και μπορεί να χρησιμοποιηθεί σε διάφορους τομείς. Ορισμένες από τις κύριες χρήσεις του CLIP είναι οι εξής:

1. Αναζήτηση και ταξινόμηση εικόνων: Το CLIP μπορεί να χρησιμοποιηθεί για να αναζητήσει και να ταξινομήσει εικόνες βάσει του περιεχομένου τους. Δηλαδή εισάγοντας μια περιγραφή στο CLIP, τότε αυτό έχει την δυνατότητα να μας επιστρέψει μία ή και πολλές εικόνες βάσει της περιγραφής αυτής.

2. Σχεδιασμός εικονικής πραγματικότητας και παιχνιδιών: Το CLIP μπορεί να χρησιμοποιηθεί για την ανάπτυξη εργαλείων σχεδίασης σε εικονική πραγματικότητα και παιχνίδια. Με την ικανότητα του να κατανοεί κείμενο και εικόνες, το CLIP μπορεί να διευκολύνει τη δημιουργία αντικειμένων και σκηνών μέσα στα παιχνίδια.

3. Ανίχνευση και ταξινόμηση περιεχομένου σε κοινωνικά μέσα και διαδικτυακά φόρουμ. Αυτό μπορεί να βοηθήσει στην αυτόματη ανίχνευση παραβατικών ή ανεπιθύμητων περιεχομένων και στην εφαρμογή μέτρων ασφαλείας στον κυβερνοχώρο.

4. Ανάλυση κοινωνικής συμπεριφοράς: Το CLIP μπορεί να χρησιμοποιηθεί για την ανάλυση της κοινωνικής συμπεριφοράς σε εικόνες και κείμενο. Με τη χρήση μεγάλου όγκου δεδομένων, το μοντέλο μπορεί να εντοπίσει μοτίβα και τάσεις στις αλληλεπιδράσεις και τις εκφράσεις των ανθρώπων.

5. Δημιουργία εικόνων από κείμενο: Το CLIP μπορεί επίσης να χρησιμοποιηθεί για τη δημιουργία εικόνων από κείμενο περιγραφής. Αυτή η δυνατότητα επιτρέπει στο μοντέλο να παράγει εικόνες που αντιστοιχούν σε περιγραφές κειμένου, αναπαριστώντας τη σημασιολογική κατανόηση του περιεχομένου.

2.3 Η αρχιτεκτονική του CLIP

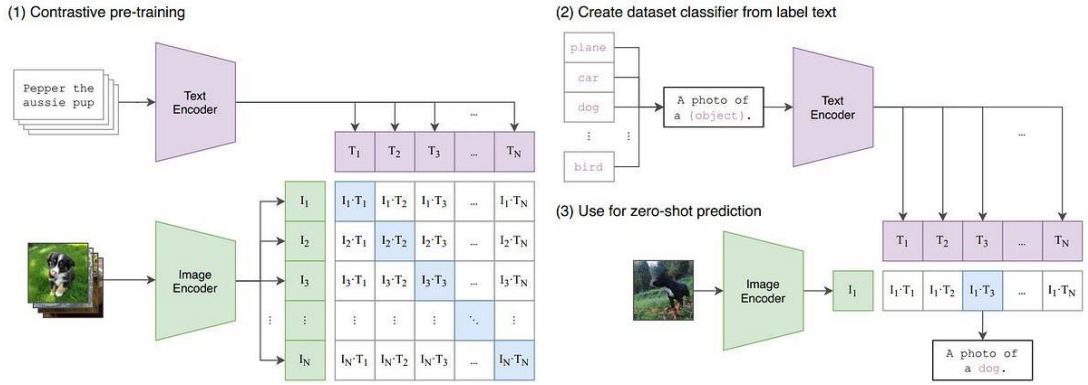
Το CLIP (Contrastive Language-Image Pretraining) αποτελείται από μια σύνθετη αρχιτεκτονική που συνδυάζει μοντέλα μηχανικής μάθησης για την κατανόηση τόσο των εικόνων όσο και του κειμένου. Συγκεκριμένα, το CLIP αποτελείται από δύο κύρια μοντέλα μηχανικής μάθησης: τον Vision Encoder και τον Text Encoder.

1. Vision Encoder: Το μέρος του Vision Encoder αναλαμβάνει την επεξεργασία των εικόνων. Αρχικά, η εικόνα υποβάλλεται σε ένα προεπεξεργαστή (preprocessor), ο οποίος μετατρέπει την εικόνα σε μια αναπαράσταση διανύσματος χαρακτηριστικών (feature vector) [8]. Στη συνέχεια, το διάνυσμα χαρακτηριστικών περνάει μέσα από ένα συνελικτικό νευρωνικό δίκτυο (convolutional neural network) που εξάγει πιο αφηρημένα χαρακτηριστικά μέσα από την εικόνα [6],[7]. Το τελικό αποτέλεσμα είναι ένα διάνυσμα χαρακτηριστικών που αναπαριστά την εικόνα.

2. Text Encoder: Το μέρος του κωδικοποιητή κειμένου (Text Encoder) ασχολείται με την επεξεργασία του κειμένου. Αρχικά, το κείμενο παρέχεται σε έναν κωδικοποιητή (encoder), όπως ένα αναδραστικό νευρωνικό δίκτυο (recurrent neural network)[12] ή έναν αυτοσχέδιο κωδικοποιητή που χρησιμοποιεί μηχανισμούς προσοχής (attention mechanisms) [10]. Αυτός ο κωδικοποιητής αντιστοιχεί το κείμενο σε ένα διάνυσμα χαρακτηριστικών, γνωστό και ως κωδικοποιημένο σημασιολογικό χώρο (encoded semantic space).

Η αρχιτεκτονική του CLIP περιλαμβάνει επίσης μια διαδικασία αντίθεσης (contrastive learning), κατά την οποία το μοντέλο εκπαιδεύεται να ενώνει εικόνες και κείμενο με έναν κοινό χώρο αναπαράστασης. Κατά τη διάρκεια της εκπαίδευσης, το CLIP εκπαιδεύεται να μάθει να αντιστοιχίζει τα ζεύγη εικόνας-κειμένου που σχετίζονται και να απομονώνει τα ζεύγη που δεν σχετίζονται.

Συνολικά, η αρχιτεκτονική του CLIP ενσωματώνει μηχανισμούς επεξεργασίας εικόνων και κειμένου, επιτρέποντας στο μοντέλο να κατανοήσει και να συνδέσει τη σημασία των δύο μορφών πληροφορίας. Αυτή η ολοκληρωμένη προσέγγιση επιτρέπει στο CLIP να εκτελέσει πολλές χρήσιμες λειτουργίες, όπως αναζήτηση και ταξινόμηση εικόνων βάσει κειμένου, ανάλυση σημασιολογικής συμπεριφοράς και πολλές άλλες εφαρμογές.



Εικόνα 2 : Η αρχιτεκτονική του CLIP [2]

2.4 Ανάλυση αρχιτεκτονικής του CLIP

2.4.1 Image Encoder

Ένας κωδικοποιητής εικόνας (Image Encoder) είναι ένα μοντέλο μηχανικής μάθησης που χρησιμοποιείται για τη μετατροπή μιας εικόνας σε μια αναπαράσταση χαρακτηριστικών (feature representation). Η αναπαράσταση αυτή περιέχει πληροφορίες που αντιπροσωπεύουν τα χαρακτηριστικά, τη δομή και το περιεχόμενο της εικόνας.

Ο κωδικοποιητής εικόνας συνήθως βασίζεται σε συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNNs) που έχουν αποδειχθεί αποτελεσματικά στην επεξεργασία εικόνων [6]. Τα CNNs αναγνωρίζουν χαρακτηριστικά σε διάφορα επίπεδα και εφαρμόζουν φίλτρα για την εξαγωγή χαρακτηριστικών από την εικόνα. Μετά την επεξεργασία αυτή, ο συγκεκριμένος κωδικοποιητής παράγει ένα διάνυσμα χαρακτηριστικών που αναπαριστά την εικόνα.

Οι κωδικοποιητές εικόνων χρησιμοποιούνται σε πολλές εφαρμογές, όπως αναγνώριση αντικειμένων, αναζήτηση εικόνων βάσει περιεχομένου, ταξινόμηση εικόνων και πολλές άλλες. Επίσης, συχνά χρησιμοποιούνται ως μέρος μεγαλύτερων μοντέλων όπως το CLIP της OpenAI, όπου η αναπαράσταση της εικόνας συνδυάζεται με αναπαραστάσεις κειμένου, αλλά και το αντίστροφο.

Μερικοί δημοφιλείς κωδικοποιητές είναι το VGGNet, το ResNet, το Inception, το DenseNet και πολλά άλλα. Αυτά τα μοντέλα έχουν εκπαιδευτεί σε μεγάλα σύνολα δεδομένων, όπως το ImageNet, για την εξαγωγή υψηλής ποιότητας αναπαραστάσεων εικόνας.

Γενικά, ο κωδικοποιητής εικόνας αναλαμβάνει τη μετατροπή της εικόνας σε έναν αριθμητικό πίνακα χαρακτηριστικών που μπορεί να χρησιμοποιηθεί για πολλές επεξεργασίες και αναλύσεις εικόνας.

2.4.2 Text Encoder

Ένας κωδικοποιητής κειμένου (Text Encoder) είναι ένα μέρος μιας αρχιτεκτονικής μηχανικής μάθησης που αναλαμβάνει την επεξεργασία του κειμένου. Ο βασικός σκοπός του είναι να μετατρέψει ένα κείμενο σε μια αναπαράσταση διανύσματος ή σε μια αναπαράσταση χαρακτηριστικών που αποτυπώνει τη σημασία και τα χαρακτηριστικά του.

Ο συγκεκριμένος κωδικοποιητής εκπαιδεύεται να κατανοεί το κείμενο και να εξάγει πληροφορίες από αυτό με τρόπο που να μπορεί να αντιληφθεί τις σημασιολογικές συνδέσεις, τις δομικές ιδιότητες και τα συναισθήματα που περιέχει το κείμενο. Αυτό επιτυγχάνεται μέσω της χρήσης συνελικτικών νευρωνικών δικτύων (CNNs), αναδραστικών νευρωνικών δικτύων (RNNs) ή ακόμη και μετασχηματισμούς όπως ο μετασχηματισμός του BERT [5].

Οι κωδικοποιητές κειμένων χρησιμοποιούνται σε πολλές εφαρμογές μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας. Μπορούν να χρησιμοποιηθούν για τον σχολιασμό και την ανάλυση κειμένου, την κατηγοριοποίηση κειμένου, τη μετάφραση, την ανάλυση συναισθήματος μέσα από κείμενο και πολλές άλλες εφαρμογές.

Οι πηγές κειμένου που χρησιμοποιούνται για την εκπαίδευση ενός τέτοιου κωδικοποιητή μπορούν να περιλαμβάνουν δημόσια κείμενα όπως ιστοσελίδες, άρθρα, βιβλία, εφημερίδες, αλλά και ιδιωτικά δεδομένα όπως κείμενα από κοινωνικά δίκτυα, ηλεκτρονικά μηνύματα κ.ά. Οι ακριβείς λεπτομέρειες και οι αρχιτεκτονικές τους μπορούν να διαφέρουν ανάλογα με το συγκεκριμένο μοντέλο και την αρχιτεκτονική που χρησιμοποιείται.

2.4.3 Zero-Shot Learning

Η μηδενική μάθηση (Zero-Shot Learning) είναι ένα πεδίο της Τεχνητής Νοημοσύνης που ασχολείται με την ανάπτυξη αλγορίθμων μηχανικής μάθησης που μπορούν να αναγνωρίζουν και να κατηγοριοποιούν νέα αντικείμενα ή έννοιες για τις οποίες δεν έχουν λάβει καμία πρότερη εκπαίδευση.

Στην παραδοσιακή μηχανική μάθηση, ένα μοντέλο εκπαιδεύεται σε ένα σύνολο δεδομένων που περιλαμβάνει παραδείγματα από κάθε κατηγορία που θέλουμε να αναγνωρίσει. Ωστόσο, στην περίπτωση της μηδενικής μάθησης, δεν υπάρχουν δεδομένα εκπαίδευσης για τις νέες κατηγορίες που θέλουμε να αναγνωρίσει.

Ο κεντρικός στόχος της μηδενικής μάθησης είναι να επιτρέψει σε ένα μοντέλο μηχανής να αντιστοιχίσει σωστά νέα δείγματα σε κατηγορίες με βάση μια περιγραφή ή έναν περιορισμένο αριθμό προηγούμενων δειγμάτων. Αυτό επιτυγχάνεται συνήθως με τη χρήση εξωτερικών πηγών πληροφορίας όπως γλωσσικά μοντέλα και οντολογίες.

Οι τεχνικές που χρησιμοποιούνται για την υλοποίηση της μηδενικής μάθησης περιλαμβάνουν:

1. Γλωσσικές εμφωλευμένες αναπαραστάσεις: Αυτή η τεχνική χρησιμοποιεί γλωσσικά μοντέλα για να μετατρέψει τις περιγραφές των νέων κατηγοριών σε αριθμητικές αναπαραστάσεις, που στη συνέχεια μπορούν να χρησιμοποιηθούν για την αναγνώριση κατηγοριών.
2. Εξαγωγή χαρακτηριστικών με βάση την οντολογία: Σε αυτήν την προσέγγιση, χρησιμοποιείται μια οντολογία ή μια γραφική αναπαράσταση των κατηγοριών για να εξάγουμε χαρακτηριστικά για τα νέα δείγματα και να τα κατηγοριοποιήσουμε.
3. Πολυτροπική προσέγγιση: Αυτή η προσέγγιση επιχειρεί να εκμεταλλευτεί τις συσχετίσεις μεταξύ των υπάρχουσων κατηγοριών για να αναγνωρίσει νέες κατηγορίες. Συχνά χρησιμοποιούνται μοντέλα εκπαίδευσης με ενίσχυση (reinforcement learning) για να επιτευχθεί αυτό.

Το Zero-Shot Learning έχει πολλές εφαρμογές στην αναγνώριση εικόνων, την επεξεργασία φυσικής γλώσσας και τη γενικότερη ταξινόμηση αντικειμένων σε νέες κατηγορίες που δεν ήταν γνωστές κατά τη διάρκεια της εκπαίδευσης του μοντέλου.

2.5 Εφαρμογή του CLIP

Όπως αναφέραμε και στα προηγούμενα υποκεφάλαια, το CLIP είναι ένα πολύ ισχυρό νευρωνικό μοντέλο, όπου η εκπαίδευση του στηρίζεται σε ένα ευρύ και ογκώδες σύνολο δεδομένων. Ως εκ τούτου, η χρήση αυτού στο δικό μας πείραμα ήταν μονόδρομος, καθώς υπερέχει εν σχέσει με κάθε ένα από τα μοντέλα που μέχρι σήμερα κυκλοφορούν ή και να υπάρχουν υπάρχουν αρκετά αρνητικά ως προς την χρήση τους.

Προτού προχωρήσουμε σε οποιαδήποτε υλοποίηση, πρωταρχικός μας στόχος ήταν να επικεντρωθούμε στην όσο το δυνατόν καλύτερη και πληρέστερη κατανόηση του συγκεκριμένου μοντέλου, κυρίως όσον αφορά στον τρόπο χρήσης αυτού αλλά και τις δυνατότητες που αυτό προσφέρει. Για να το επιτύχουμε αυτό, εκμεταλλευτήκαμε μία online πλατφόρμα στην οποία ναι μεν παρέχεται δωρεάν ένα άλλο μοντέλο, δεν παύει όμως να πρόκειται για ένα μοντέλο παρόμοιας λογικής και δυνατοτήτων (https://replicate.com/rmokady/clip_prefix_caption) -συνεπώς και εξίσου χρήσιμο. Στα πρώτα στάδια του πειραματισμού μας με το μοντέλο της πλατφόρμας, ήταν απαραίτητο να πραγματοποιήσουμε διάφορες δοκιμές με τις αντίστοιχες παραμέτρους ούτως ώστε να επιτύχουμε την αποτελεσματική παραγωγή caption για κάποια τυχούσα εικόνα. Παραδείγματος χάριν, στην παρακάτω εικόνα (εικόνα 3) δίνεται μία τυπική εικόνα από αυτές που χρησιμοποιήθηκαν μαζί με το αντίστοιχο caption που παρήγαγε το online εργαλείο.

Εικόνα 3 : Εικόνα δοκιμής για το σύστημά μας [19]

Λεζάντα online εργαλείου: **A bunch of bananas sitting on top of a table.**

Στη συνέχεια, αφού στήσαμε το δικό μας σύστημα [18], εισάγαμε την ίδια εικόνα και παρατηρήσαμε πως προέκυψε παρόμοια περιγραφή με αυτήν της έτοιμης πλατφόρμας η οποία είναι η παρακάτω:

Λεζάντα του δικού μας συστήματος: **a basket of bananas at a market.**

Παρατηρούμε ότι τα αποτελέσματα είναι αρκετά κοντά μεταξύ τους, κάτι που σημαίνει ότι οι παράμετροι που θέσαμε στο δικό μας σύστημα είναι σωστές. Να αναφέρουμε στο σημείο αυτό, ότι παρατηρήθηκαν πολλά προβλήματα στα τελικά αποτελέσματα σε περιπτώσεις όπου θέταμε άστοχες παραμέτρους.

Αφότου εξασφαλίσαμε ότι το σύστημα που δομήσαμε για παραγωγή λεζαντών είναι τόσο αποτελεσματικό όσο και σταθερό, προχωρήσαμε στο επόμενο στάδιο ανάλυσης κατά το οποίο εξερευνούμε την εγκυρότητα του μοντέλου σε εικόνες με όσο το δυνατόν πλουσιότερο περιεχόμενο ή εικόνες των οποίων η παραγόμενη περιγραφή πιθανότατα θα είναι αρκετά εμπλουτισμένη. Ορισμένα παραδείγματα τέτοιων εικόνων δίνονται παρακάτω (εικόνες 4-5).

Εικόνα 4 :
Λεζάντα: biological genus is a large carnivorous lizard that lives in the temperate rainforests [21]

Εικόνα 5 : Λεζάντα Εικόνας: pink peonies in a vase [20]

Η παραπάνω διαδικασία παραγωγής των λεζαντών που εφαρμόστηκε αρχικά σε δεδομένα εικόνων, άμεσα επεκτάθηκε και στη χρήση αυτής για την παραγωγή λεζάντας σε δεδομένα οπτικο-ακουστικού υλικού που επίσης είχαμε στη διάθεσή μας. Η διαδικασία που ακολουθήσαμε συνοψίζεται στα εξής:

Ξεκινώντας με την επιλογή του επιθυμητού βίντεο, αρχικά εφαρμόζουμε κατάλληλη διαδικασία δειγματοληψίας, μετατρέποντας με τον τρόπο αυτό το βίντεο σε τίποτα άλλο παρά από ένα σύνολο εικόνων - καρέ (frames). Ακολούθως, εφαρμόζοντας την τυπική διαδικασία παραγωγής λεζάντας εικόνας όπως αναλύθηκε νωρίτερα, καταφέρνουμε να αποκτήσουμε ένα caption για κάθε ένα από τα frames που έχουμε στην κατοχή μας. Τέλος, το σύνολο των περιγραφών (captions) που αποθηκεύεται, εισάγεται σε έναν συνοψιστή (summarizer). Ο τρόπος λειτουργίας αλλά και η αρχιτεκτονική του που χρησιμοποιήθηκε θα αναλυθεί λεπτομερώς σε επόμενα κεφάλαια.

Τέλος, φυσικά να αναφέρουμε ότι το CLIP συνοδεύεται και από σφάλματα σε κάποιον βαθμό, κατά την προσπάθειά του να παράξει περιγραφή για μια εικόνα, όπως παρακάτω (εικόνα 6) όπου ναι μεν έχει αναγνωρίσει σωστά τα χαρακτηριστικά του άντρα που εμφανίζεται στην εικόνα, παράλληλα όμως αγνοεί πλήρως την ύπαρξη της γυναικείας παρουσίας στο βάθος.

Εικόνα 6 : Λεζάντα εικόνας: A man wearing glasses and a tie [22]

3 Μετρικές

3.1 Σκοπός χρήσης μετρικών

Οι μετρικές χρησιμοποιούνται στη μηχανική μάθηση για να μετρήσουν και να αξιολογήσουν την απόδοση ενός μοντέλου μηχανικής μάθησης για ένα συγκεκριμένο πρόβλημα. Οι μετρικές μας δίνουν μια ποσοτική αναπαράσταση της ποιότητας των προβλέψεων του μοντέλου μας.

Οι λόγοι για τους οποίους λαμβάνουμε μετρικές είναι οι εξής:

1. Αξιολόγηση της απόδοσης: Οι μετρικές μας επιτρέπουν να κατανοήσουμε πόσο καλά επιτυγχάνει το μοντέλο μας το στόχο του προβλήματος. Μπορούμε να συγκρίνουμε διάφορα μοντέλα και αλγορίθμους μηχανικής μάθησης με βάση τις μετρικές, προκειμένου να επιλέξουμε αυτό με την καλύτερη απόδοση.

2. Βελτιστοποίηση του μοντέλου: Οι μετρικές μας βοηθούν να κατανοήσουμε τα σημεία ισχύος και αδυναμίας του μοντέλου μας. Μπορούμε να χρησιμοποιήσουμε αυτές τις πληροφορίες για να βελτιώσουμε το μοντέλο μας, αναπτύσσοντας νέες τεχνικές ή προσαρμόζοντας τις υπάρχουσες.

3. Παρακολούθηση της απόδοσης στον χρόνο: Οι μετρικές μας μπορούν επίσης να χρησιμοποιηθούν για να παρακολουθούν την απόδοση του μοντέλου μας καθ' όλη τη διάρκεια της λειτουργίας του. Αυτό μας επιτρέπει να ανιχνεύσουμε πιθανά προβλήματα ή αλλαγές στην απόδοση και να προβούμε σε απαραίτητες προσαρμογές.

Οι μετρικές μπορούν να ποικίλουν ανάλογα με το πρόβλημα μηχανικής μάθησης που αντιμετωπίζουμε. Ορισμένες συνηθισμένες μετρικές περιλαμβάνουν την ακρίβεια (accuracy), την ακρίβεια ταξινόμησης (precision), την ανάκληση (recall), την F1-μέτρηση (F1-score), το απόλυτο σφάλμα (mean absolute error - MAE) και το τετραγωνικό σφάλμα (mean squared error - MSE).

Τώρα λοιπόν η διαδικασία σύνοψης (summarization) με χρήση των μετασχηματιστών βασίζεται συνήθως σε μερικές κύριες μετρικές για την αξιολόγηση της ποιότητας των παραγόμενων περιλήψεων. Οι συνηθέστερες μετρικές είναι: ο ROUGE, ο BLEU και ο METEOR. Αυτές οι μετρικές χρησιμοποιούνται για να αξιολογήσουν την ποιότητα των περιλήψεων που παράγονται από συστήματα σύνοψης βασισμένα στους μετασχηματιστές. Οι περισσότερες μέθοδοι σύνοψης χρησιμοποιούν μία ή περισσότερες από αυτές τις μετρικές για να αξιολογήσουν την ποιότητα των περιλήψεων και να τις βελτιστοποιήσουν.

3.2 Μετρική ROUGE

Το ROUGE είναι ένα σύνολο μετρικών που αξιολογούν την ομοιότητα μεταξύ της πραγματικής περίληψης και της παραγόμενης περίληψης [9]. Αυτές οι μετρικές μετρούν την ακρίβεια, την ανάκληση και το F1-score των κοινών λέξεων, των κοινών φράσεων ή των κοινών n-gram (συνεχόμενων λέξεων) μεταξύ της πραγματικής και της παραγόμενης περίληψης.

Ας υποθέσουμε ότι έχουμε το αρχικό κείμενο (reference) και την παραγόμενη αυτόματη περίληψη (system summary) για αυτό το κείμενο. Η μετρική ROUGE υπολογίζει τον βαθμό ομοιότητας μεταξύ αυτών των δύο κειμένων.

Ας πάρουμε ένα παράδειγμα:

Αρχικό κείμενο (reference): «Ο καιρός σήμερα ήταν υπέροχος. Ο ήλιος έλαμπε και οι θερμοκρασίες ήταν ιδανικές για μια βόλτα στο πάρκο. Οι άνθρωποι απόλαυσαν την ηλιοθεραπεία και το πικνίκ.»

Παραγόμενη αυτόματη περίληψη (system summary): «Ο καιρός ήταν καλός και οι άνθρωποι βγήκαν έξω για να απολαύσουν την ηλιοθεραπεία και το φαγητό.»

Για να υπολογίσουμε τη μετρική ROUGE, πρέπει να χωρίσουμε τα κείμενα σε λέξεις ή n-grams και να συγκρίνουμε τις συμβολοσειρές αυτές. Ας υποθέσουμε ότι η μετρική ROUGE χρησιμοποιείται για να υπολογίσει την ROUGE-1, η οποία μετρά την ακρίβεια των μονολεκτικών αντιστοιχιών μεταξύ των κειμένων.

Λέξεις στο αρχικό κείμενο (reference): Ο καιρός σήμερα ήταν υπέροχος. Ο ήλιος έλαμπε και οι θερμοκρασίες ήταν ιδανικές για μια βόλτα στο πάρκο. Οι άνθρωποι απόλαυσαν την ηλιοθεραπεία και το πικνίκ.

Λέξεις στην παραγόμενη περίληψη (system summary): Ο καιρός ήταν καλός και οι άνθρωποι βγήκαν έξω για να απολαύσουν την ηλιοθεραπεία και το φαγητό.

Αν κρατήσουμε μόνο τις μοναδικές λέξεις από αυτά τα κείμενα και τις αντιστοιχίσουμε μεταξύ τους, θα έχουμε το εξής:

Κοινές μοναδικές λέξεις: ο, καιρός, ήταν, καλός, άνθρωποι, ηλιοθεραπεία, και, το

Σύνολο μοναδικών λέξεων στο αρχικό κείμενο: 15

Σύνολο μοναδικών λέξεων στην παραγόμενη περίληψη: 8

Για να υπολογίσουμε την ROUGE-1, υπολογίζουμε το ποσοστό των κοινών μοναδικών λέξεων προς το σύνολο των μοναδικών λέξεων στο αρχικό κείμενο.

ROUGE-1 = (Κοινές μοναδικές λέξεις) / (Σύνολο μοναδικών λέξεων στο αρχικό κείμενο) = 7/15 ≈ 0.47

Στο παράδειγμά μας, η ROUGE-1 είναι περίπου 0.47, υποδεικνύοντας έναν βαθμό ομοιότητας μεταξύ της παραγόμενης περίληψης και του αρχικού κειμένου. Πειραματικά είναι γνωστό ότι ένα ποσοστό γύρω στο 50% ή 0.5 αντικατοπτρίζει ένα άριστο αποτέλεσμα κατά την περίληψη.

3.3 Μετρική BLEU

Το BLEU (Bilingual Evaluation Understudy) μετρά την ποιότητα μιας παραγόμενης περίληψης συγκρίνοντάς τη με μια ή περισσότερες αναφορές περιλήψεων. Χρησιμοποιείται για να μετρήσει την ακρίβεια των κοινών λέξεων ή φράσεων μεταξύ της παραγόμενης περίληψης και των αναφορών περιλήψεων.

Ας υποθέσουμε ότι έχουμε την εξής μετάφραση μηχανής:

Μετάφραση μηχανής: «Ο γάτος είναι στο σπίτι.»

Και μια αναφορά μετάφρασης ανθρώπου:

Αναφορά μετάφρασης: «Η γάτα είναι μέσα στο σπίτι.»

Για να υπολογίσουμε τη μετρική BLEU, ακολουθούμε τα παρακάτω βήματα:

1. Υπολογισμός των n-γραμμάτων για κάθε φράση (τα n-γράμματα είναι ακολουθίες από n λέξεις).

- Παράδειγμα: Για n=1, τα 1-γράμματα για τη μετάφραση μηχανής είναι: [«Ο», «γάτος», «είναι», «στο», «σπίτι»] και για την αναφορά μετάφρασης είναι: [«Η», «γάτα», «είναι», «μέσα», «στο», «σπίτι»].

- Για n=2, τα 2-γράμματα για τη μετάφραση μηχανής είναι: [«Ο γάτος», «γάτος είναι», «είναι στο», «στο σπίτι»] και για την αναφορά μετάφρασης είναι: [«Η γάτα», «γάτα είναι», «είναι μέσα», «μέσα στο», «στο σπίτι»].

2. Υπολογισμός της ακρίβειας (precision) για κάθε n-gram.

- Η ακρίβεια είναι ο λόγος του αριθμού των n-gram που εμφανίζονται τόσο στη μετάφραση μηχανής όσο και στην αναφορά μετάφρασης, προς τον αριθμό των n-gram που εμφανίζονται στη μετάφραση μηχανής.

- Παράδειγμα: Για n=1, η ακρίβεια είναι 4/5 = 0.8 (καθώς υπάρχουν 4 λέξεις που εμφανίζονται και στις δύο μεταφράσεις, από τις συνολικά 5 λέξεις της μετάφρασης μηχανής).

- Για n=2, η ακρίβεια είναι 2/4 = 0.5 (καθώς υπάρχουν 2 2-grams που εμφανίζονται και στις δύο μεταφράσεις, από τις συνολικά 4 2-grams της μετάφρασης μηχανής).

3. Υπολογισμός του βάρους (weight) για κάθε n-gram.

- Οι βαθμοί βάρους είναι μια σταθερά που αντιπροσωπεύει τη σημασία κάθε n-gram.

- Παράδειγμα: Για n=1, ο βαθμός βάρους είναι 1/1 = 1 (καθώς η ακρίβεια υπολογίστηκε για 1-gram).

- Για n=2, ο βαθμός βάρους είναι ½ = 0.5 (καθώς η ακρίβεια υπολογίστηκε για 2-grams).

4. Υπολογισμός του BLEU σκορ.

- Ο υπολογισμός του BLEU σκορ γίνεται με βάση τις ακρίβειες και τους βαθμούς βάρους για όλες τις n-grams.

- Παράδειγμα: Αν έχουμε n=1 και n=2, τότε το BLEU σκορ υπολογίζεται ως εξής:

BLEU = exp(0.5 * log(0.8) + 0.5 * log(0.5))

Συνοψίζοντας, ο υπολογισμός του BLEU σκορ συνδυάζει την ακρίβεια για διάφορες n-grams με τα αντίστοιχα βάροι, για να προκύψει ένα συνολικό μέτρο ποιότητας μετάφρασης μηχανής.

3.4 Μετρική METEOR

Το METEOR (Metric for Evaluation of Translation with Explicit Ordering) είναι μια μετρική που συνδυάζει την ακρίβεια και την πληρότητα. Αξιολογεί την ποιότητα της παραγόμενης περίληψης με βάση την ομοιότητα μεταξύ των λέξεων, των φράσεων και της σύνταξης μεταξύ της παραγόμενης και της πραγματικής περίληψης.

Η μετρική METEOR (Metric for Evaluation of Translation with Explicit Ordering) χρησιμοποιείται συνήθως για την αξιολόγηση μηχανικής μετάφρασης κειμένου. Ακολουθεί ένα παράδειγμα για να εξηγήσω πώς υπολογίζεται η μετρική METEOR.

Ας υποθέσουμε ότι έχουμε την πρόταση που θέλουμε να μεταφράσουμε:

Πρόταση στην πηγαία γλώσσα: “The cat is sitting on the mat.”

Η μεταφραστική πρόταση που παρήγαγε ένα μηχάνημα είναι:

Μεταφραστική πρόταση: «Η γάτα κάθεται στο χαλί.»

Για να υπολογίσουμε το METEOR, πρέπει να ακολουθήσουμε τα παρακάτω βήματα:

Βήμα 1: Προετοιμασία

Αρχικά, πρέπει να προετοιμάσουμε τις δύο προτάσεις για την αξιολόγηση. Αυτό περιλαμβάνει την αφαίρεση των σημείων στίξης και τη διαίρεση του κειμένου σε λέξεις.

Πρόταση 1: “The cat is sitting on the mat.”

Πρόταση 2: «Η γάτα κάθεται στο χαλί.»

Βήμα 2: Αντιστοίχιση λέξεων

Στο επόμενο βήμα, αντιστοιχούμε τις λέξεις της πρώτης πρότασης με τις λέξεις της δεύτερης πρότασης. Οι αντιστοιχίες μπορούν να είναι μονόπλευρες (μία λέξη της πρώτης πρότασης αντιστοιχεί σε μία λέξη της δεύτερης πρότασης) ή πολύπλευρες (πολλές λέξεις της πρώτης πρότασης αντιστοιχούν σε μία λέξη της δεύτερης πρότασης).

Αντιστοίχιση: «The» -> «Η», «cat» -> «γάτα», «is» -> «κάθεται», «sitting» -> «κάθεται», «on» -> «στο», «the» -> «το», «mat» -> «χαλί».

Βήμα 3: Υπολογισμός σκορ

Στο τελευταίο βήμα, υπολογίζουμε τα ποσοστά αντιστοίχισης και τα βάρη για κάθε αντιστοίχιση. Τα βάρη εξαρτώνται από την απόσταση των λέξεων και την ακρίβεια της αντιστοίχισης.

Σκορ αντιστοίχισης:

- «The» ➔ «Η»: ακρίβεια 1, απόσταση 0
- «cat» ➔ «γάτα»: ακρίβεια 1, απόσταση 0
- «is» ➔ «κάθεται»: ακρίβεια 1, απόσταση 0
- «sitting» ➔ «κάθεται»: ακρίβεια 1, απόσταση 0.5 (διπλή αντιστοίχιση)
- «on» ➔ «στο»: ακρίβεια 1, απόσταση 0
- «the» ➔ «το»: ακρίβεια 1, απόσταση 0
- «mat» ➔ «χαλί»: ακρίβεια 1, απόσταση 0

Σκορ ταιριάσματος: (1 + 1 + 1 + 1 + 1 + 1 + 1) / 7 = 1

Βάρη:

- Αντιστοίχιση με ακρίβεια: 7/7 = 1
- Αντιστοίχιση με απόσταση: 6/7

Σκορ METEOR: (1^1) * (1^(6/7)) ≈ 0.55

Έτσι, στο παράδειγμά μας, το σκορ METEOR για τη μηχανική μετάφραση της πρότασης είναι περίπου 0.55.

Οι παραπάνω μετρικές χρειάστηκαν στην παρούσα διαδικασία, ώστε να μεταφράσουμε κατά μία έννοια τα τελικά αποτελέσματα από τις περιλήψεις (summaries), αλλά και να τα αξιολογήσουμε φυσικά.

4 Transformers

4.1 Είδη Transformers

Οι μετασχηματιστές (Transformers) είναι ένας τύπος μοντέλων μηχανικής μάθησης που έχουν αποδειχθεί εξαιρετικά αποτελεσματικά σε πολλά προβλήματα επεξεργασίας φυσικής γλώσσας και αναγνώρισης ακολουθιών [4]. Αναλυτικά, υπάρχουν διάφορα είδη μετασχηματιστών που έχουν αναπτυχθεί και χρησιμοποιηθεί σε διάφορες εφαρμογές. Ορισμένα από αυτά περιλαμβάνουν:

1. Transformer: Το αρχικό μοντέλο μετασχηματιστή που παρουσιάστηκε από τους Vaswani και συνεργάτες το 2017. Αυτό το μοντέλο χρησιμοποιήθηκε αρχικά για μετάφραση και αποτελείται από έναν αριθμό επαναλαμβανόμενων επιπέδων (layers) με αυτο-προσοχή(self-attention) και πλήρως συνδεδεμένα επίπεδα.

2. BERT (Bidirectional Encoder Representations from Transformers): Το BERT είναι ένα μοντέλο Transformer που προτάθηκε από την εταιρεία Google το 2018. Αυτό το μοντέλο εκπαιδεύεται κυρίως για κατανόηση φυσικής γλώσσας, χρησιμοποιώντας ένα μεγάλο πλήθος δεδομένων. Ένα κοινό χαρακτηριστικό του BERT είναι η προ-εκπαίδευση (pre-training) σε μεγάλους όγκους κειμένων, ακολουθούμενη από την μετ-εκπαίδευση/βελτίωση (fine-tuning) σε συγκεκριμένες εφαρμογές.

3. GPT (Generative Pre-trained Transformer): Το GPT είναι μια σειρά από μοντέλα Transformers που αναπτύχθηκαν από την OpenAI. Το αρχικό μοντέλο GPT παρουσιάστηκε το 2018 και αποτελούνταν από ένα μεγάλο αριθμό επιπέδων Transformer. Το GPT και οι εκδόσεις του έχουν χρησιμοποιηθεί για διάφορες εργασίες, όπως αναγνώριση κειμένου, παραγωγή κειμένου κ.λπ.

4. XLNet: Το XLNet είναι ένα μοντέλο Transformer που προτάθηκε το 2019 και επεκτείνει τον αλγόριθμο αυτοπαλινδρομικής προσοχής (autoregressive attention) που χρησιμοποιείται στο GPT. Η αυτοπαλινδρομική προσοχή επιτρέπει στο μοντέλο να αποφύγει την περιορισμένη προσέγγιση των προβλέψεων και να λαμβάνει υπόψη τη συνοχή του κειμένου κατά την εκπαίδευση.

Αυτά είναι μερικά από τα κύρια είδη των μοντέλων των μετασχηματιστών που έχουν παρουσιαστεί και χρησιμοποιηθεί σε εφαρμογές επεξεργασίας φυσικής γλώσσας. Υπάρχουν επίσης πολλές παραλλαγές και βελτιωμένες εκδόσεις αυτών των μοντέλων, καθώς η έρευνα συνεχίζει να προχωρά σε αυτόν τον τομέα.

4.2 Χρήσεις των Transformers

Οι μετασχηματιστές (Transformers) είναι ένα είδος αρχιτεκτονικής νευρωνικών δικτύων που έχει αποδειχθεί ότι είναι εξαιρετικά αποτελεσματική σε πολλές εφαρμογές επεξεργασίας φυσικής γλώσσας. Ας εξετάσουμε μερικές από τις πολύ σημαντικές χρήσεις των μετασχηματιστών:

1. Μηχανές Μετάφρασης: Οι μετασχηματιστές είναι εξαιρετικά αποτελεσματικοί στη μετάφραση φυσικής γλώσσας από μία γλώσσα σε μία άλλη. Έχουν επιτευχθεί σημαντικές πρόοδοι στον τομέα της μετάφρασης χάρη στη χρήση των μετασχηματιστών, ειδικά με την εισαγωγή του μοντέλου «Transformer» στο περιβάλλον του «Google Translate».

2. Αναγνώριση Ονομάτων Προσώπων (NER): Οι μετασχηματιστές χρησιμοποιούνται για την ανίχνευση και την κατηγοριοποίηση ονομάτων προσώπων σε κείμενα. Αυτό είναι χρήσιμο σε πολλές εφαρμογές, όπως η εξαγωγή πληροφοριών από κείμενα, η ανάλυση κοινωνικών μέσων κ.λπ.

3. Προεπεξεργασία Κειμένου: Οι μετασχηματιστές χρησιμοποιούνται επίσης για την προεπεξεργασία κειμένου πριν από την είσοδο σε άλλα μοντέλα επεξεργασίας γλώσσας. Αυτό περιλαμβάνει την αποκωδικοποίηση, την αναπαράσταση λέξεων και την εξαγωγή χαρακτηριστικών για περαιτέρω επεξεργασία.

4. Chatting Συστήματα: Οι μετασχηματιστές χρησιμοποιούνται για την ανάπτυξη chatting συστημάτων, όπως οι εικονικοί βοηθοί και τα συστήματα επικοινωνίας ανθρώπου-μηχανής. Μπορούν να μάθουν να αντιλαμβάνονται την φύση εκάστοτε περίστασης αλλά και την παραγωγή φυσικής γλώσσας, καθιστώντας την επικοινωνία με τις μηχανές πιο φυσική και αποδοτική.

5. Ανάλυση Συναισθήματος: Οι μετασχηματιστές είναι επίσης χρήσιμοι στην ανάλυση συναισθημάτων από κείμενα. Μπορούν να αναγνωρίσουν τον τόνο, την απόχρωση και τη σημασιολογία που σχετίζεται με τα συναισθήματα που εκφράζονται σε ένα κείμενο.

6. Ανάλυση εικόνας και αναγνώριση αντικειμένων: Οι μετασχηματιστές χρησιμοποιούνται για την ανάλυση εικόνας, την αναγνώριση αντικειμένων και την επεξεργασία εικόνας. Αποτελούν το κύριο κομμάτι των συστημάτων αναγνώρισης προσώπων, των συστημάτων ανίχνευσης αντικειμένων σε εικόνες και βίντεο, και χρησιμοποιούνται σε εφαρμογές όπως αυτόματη ετικετοποίηση φωτογραφιών και αναγνώριση χαρακτηριστικών εικόνας.

7. Συστήματα συστάσεων και πρόβλεψης: Οι μετασχηματιστές χρησιμοποιούνται σε συστήματα συστάσεων για την πρόβλεψη των προτιμήσεων των χρηστών και τη σύσταση περιεχομένου. Χρησιμοποιούνται επίσης σε μοντέλα πρόβλεψης και προγνώσεων για πολλούς τομείς, όπως οικονομία, κλιματικές μεταβολές και χρηματοοικονομικές αγορές.

Αυτές είναι μερικές από τις κύριες χρήσεις των μετασχηματιστών στην καθημερινότητά μας. Ωστόσο, οι μετασχηματιστές έχουν ευρύτερη εφαρμογή και συνεχώς εξελίσσονται, επηρεάζοντας πολλούς τομείς όπως η ρομποτική, η ιατρική διάγνωση, η αυτόνομη οδήγηση και πολλοί άλλοι.

Ως γλωσσικό μοντέλο, οι μετασχηματιστές έχουν πολλά πλεονεκτήματα, αλλά υπάρχουν και μερικά αρνητικά που μπορούν να αναφερθούν. Ορισμένα από αυτά τα αρνητικά περιλαμβάνουν:

1. Απαιτητικά σε πόρους: Είναι αρκετά απαιτητικά σε πόρους, ειδικά όταν πρόκειται για μεγάλα μοντέλα όπως το GPT-3.5. Αυτό σημαίνει ότι απαιτούν πολλή υπολογιστική ισχύ και μνήμη για την εκπαίδευση και τη λειτουργία τους.

2. Απαιτητικά σε πολλά δεδομένα: Απαιτούν μεγάλα σύνολα δεδομένων και πολύ χρόνο εκπαίδευσης για να αποκτήσουν υψηλή απόδοση. Αυτό σημαίνει ότι η δημιουργία και η εκπαίδευση ενός μοντέλου μετασχηματιστή μπορεί να είναι πολύ χρονοβόρα και απαιτητική από άποψη υπολογιστικών πόρων και δεδομένων.

3. Αδυναμία κατανόησης του πλήρους περιεχομένου: Παρόλο που οι μετασχηματιστές έχουν εντυπωσιακή απόδοση στην παραγωγή κειμένου, μπορεί να είναι δύσκολο να κατανοήσουν το πλήρες περιεχόμενο ή να διακρίνουν τις αλλαγές στο νόημα κατά την παραγωγή πολυπλοκότερων κειμένων.

4. Έλλειψη γνώσεων πέρα από το εκπαιδευμένο σύνολο δεδομένων: Βασίζονται σε δεδομένα που έχουν χρησιμοποιηθεί για την εκπαίδευσή τους. Αυτό σημαίνει ότι δεν έχουν πρόσβαση σε πραγματικές γνώσεις που υπερβαίνουν το υλικό του συνόλου δεδομένων, και μπορεί να παρουσιάσουν περιορισμένη κατανόηση σε ζητήματα που απαιτούν γνώσεις εκτός του εκπαιδευμένου περιβάλλοντος.

Αυτά είναι μερικά από τα αρνητικά που μπορούν να συνδεθούν με τους μετασχηματιστές. Όμως είναι σημαντικό να σημειωθεί ότι η τεχνολογία τους βελτιώνεται συνεχώς και αυτά τα προβλήματα μπορεί να αντιμετωπιστούν ή να μειωθούν στο μέλλον.

4.3 Αρχιτεκτονική απλού Transformer

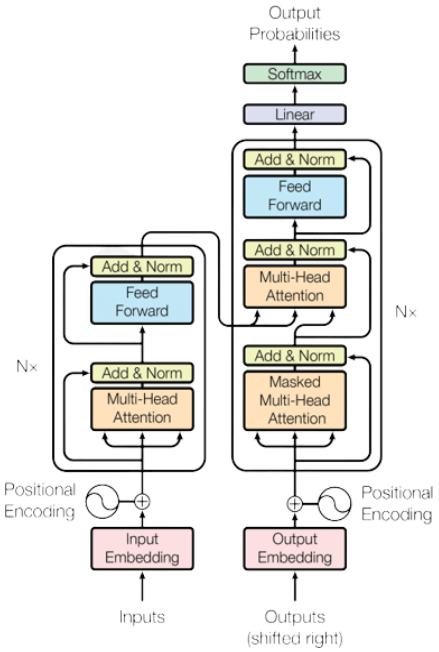
Ένας απλός μετασχηματιστής (Transformer) ακολουθεί την αρχιτεκτονική που παρουσιάστηκε στο αρχικό άρθρο «Attention is All You Need» των Vaswani και συναδέλφων το 2017. Αυτή η αρχιτεκτονική αποτελείται από δύο βασικά μέρη: τον encoder (κωδικοποιητή) και τον decoder (αποκωδικοποιητή).

Ο κωδικοποιητής αποτελείται από μια στοίβα από έναν ορισμένο αριθμό ίδιων επιπέδων (layers) κωδικοποίησης. Κάθε επίπεδο κωδικοποίησης περιλαμβάνει δύο σημαντικά μέρη: την πολλαπλή προσοχή στον εαυτό του (Multi-Head Self-Attention) και ένα νευρωνικό δίκτυο με προώθηση προς τα εμπρός (Feed-Forward Neural Network).

Ο αποκωδικοποιητής επίσης αποτελείται από μια στοίβα με τα αντίστοιχα επίπεδα (layers). Υπάρχει ένα επιπλέον επίπεδο κωδικοποίησης πριν από τον decoder, που ονομάζεται encoder-decoder attention (προσοχή ανάμεσα σε κωδικοποιητή και αποκωδικοποιητή). Αυτό το επίπεδο χρησιμοποιείται για να επιτρέψει στον αποκωδικοποιητή να έχει πρόσβαση στις πληροφορίες που έχουν κωδικοποιηθεί από τον κωδικοποιητή.

Και στα δύο μέρη, encoder και decoder, χρησιμοποιείται η ιδέα της αυτοπροσοχής (self-attention), όπου οι λέξεις σε μια πρόταση αλληλεπιδρούν και προσδίδουν βάρος στην κατανόηση του κειμένου.

Συνολικά, η αρχιτεκτονική του απλού μετασχηματιστής περιλαμβάνει πολλές στοίβες επιπέδων κωδικοποίησης και αποκωδικοποίησης, μεταξύ των οποίων χρησιμοποιείται το self-attention για την αλληλεπίδραση των λέξεων και την εξαγωγή σημαντικών χαρακτηριστικών του κειμένου.



Εικόνα 7 : Αρχιτεκτονική απλού Transformer [4]

4.4 Προηγμένοι μηχανισμοί μηχανικής μάθησης

4.4.1 Positional Encoding

Η κωδικοποίηση θέσης (positional encoding) είναι ένας σημαντικός μηχανισμός που εισήχθη στον αλγόριθμο των μετασχηματιστών για την αντιμετώπιση του προβλήματος της αντιστροφής της σειράς των λέξεων και της έλλειψης κατανόησης της θέσης των λέξεων στην είσοδο.

Οι αλγόριθμοι αυτοί, και ειδικότερα τα μοντέλα αναγνώρισης φυσικής γλώσσας (NLP), είναι αρχικά σχεδιασμένα για να χειριστούν αναπαραστάσεις λέξεων, οι οποίες προσδιορίζουν το νόημα των λέξεων ανεξαρτήτως της θέσης τους στην πρόταση. Ωστόσο, στις περισσότερες φυσικές γλώσσες, η θέση μιας λέξης είναι σημαντική και μπορεί να επηρεάσει το νόημα της πρότασης.

Η κωδικοποίηση θέσης προστίθεται στις ενσωματώσεις (embeddings) των λέξεων πριν αυτές εισέλθουν στα επίπεδα των μετασχηματιστών. Σκοπός της είναι να προσφέρει πληροφορία σχετικά με τη θέση των λέξεων στην είσοδο.

Ο τρόπος που επιτυγχάνεται κάτι τέτοιο είναι μέσω μιας συνάρτησης κωδικοποίησης που δίνει μια αναπαράσταση του πολυδιάστατου διανύσματος θέσης. Υπάρχουν διάφορες μέθοδοι για την κωδικοποίηση θέσης, αλλά η πιο κοινή προσέγγιση είναι η χρήση συναρτήσεων συνημιτόνου ή ημιτόνου.

Ένα παράδειγμα για κωδικοποίηση θέσης είναι:

Για μια λέξη στη θέση «pos» και μια συγκεκριμένη διάσταση «i» στο διάνυσμα, η κωδικοποίηση θέσης υπολογίζεται ως εξής:

```
PE(pos, 2i) = sin( pos / 10000 ^ (2*i / d_model) )
PE(pos, 2i + 1) = cos( pos / 10000 ^ (2*i / d_model) )
```

όπου «pos» είναι η θέση της λέξης, «i» είναι ο δείκτης της διάστασης, και «d_model» είναι η διάσταση του διανύσματος ενσωμάτωσης.

Με την προσθήκη της κωδικοποίησης θέσης, το μοντέλο μπορεί να διαχωρίσει λέξεις που εμφανίζονται στις ίδιες θέσεις σε διαφορετικές προτάσεις, επιτρέποντας έτσι την αντιμετώπιση προβλημάτων που προκύπτουν από την σειρά των λέξεων. Αυτό είναι κρίσιμο για τις εφαρμογές NLP που απαιτούν κατανόηση της σύνταξης και της σημασίας των λέξεων στο πλαίσιο του κειμένου.

4.4.2 Attention

Στους μηχανισμούς μετασχηματιστών, η έννοια της προσοχής (Attention) αναφέρεται ως μια βασική λειτουργία που επιτρέπει στο μοντέλο να εστιάσει περισσότερο σε συγκεκριμένα κομμάτια της εισόδου κατά την επεξεργασία της. Αυτό το χαρακτηριστικό είναι καθοριστικό για την απόδοση των μετασχηματιστών σε ποικίλες διαδικασίες επεξεργασίας φυσικής γλώσσας και έχει συμβάλει σημαντικά στην επίτευξη καταπληκτικών αποτελεσμάτων σε αυτόν τον τομέα.

Για να κατανοήσουμε τον μηχανισμό προσοχής, ας δούμε πώς λειτουργεί σε ένα μοντέλο μετασχηματιστή:

1. Είσοδος:

Η είσοδος σε έναν μετασχηματιστή αποτελείται από μια ακολουθία λέξεων ή διανυσμάτων που αναπαριστούν τα δεδομένα που θέλουμε να επεξεργαστούμε. Για παράδειγμα, αν έχουμε μια πρόταση «Ο γάτος κοιμάται,» κάθε λέξη («Ο», «γάτος», «κοιμάται») μετατρέπεται σε ένα διάνυσμα με διαστάσεις που αντιπροσωπεύουν τον σημασιολογικό χώρο των λέξεων.

2. Έρωτημα (Query), κλειδί (Key) και τιμή (Value):

Σε κάθε βήμα, δημιουργούνται τρεις σημασιολογικές προβολές για κάθε λέξη (ή διάνυσμα) της εισόδου:

- Ερώτημα: Χρησιμοποιείται για να ρωτήσουμε το μοντέλο ποια μέρη της εισόδου πρέπει να εστιάσει.
- Κλειδί: Χρησιμοποιείται για να προσδιορίσουμε τις περιοχές της εισόδου που αξίζει να ληφθούν υπόψη κατά την αναζήτηση πληροφοριών.
- Τιμή: Χρησιμοποιείται για να παρέχουμε τις πληροφορίες που αντιστοιχούν σε κάθε λέξη ή διάνυσμα εισόδου.

3. Υπολογισμός της προσοχής:

Με βάση την τριπλέτα ερώτημα-κλειδί-τιμή, υπολογίζεται ένα σκορ που μετράει το βάρος αξίας (ή «attention weight») που πρέπει να δοθεί σε κάθε λέξη της εισόδου. Αυτός ο υπολογισμός γίνεται μέσω διαφόρων μετρικών, συχνά με χρήση του εσωτερικού γινομένου ή του soft-max.

4. Εφαρμογή της προσοχής:

Το βάρος που υπολογίστηκε στο βήμα 3 χρησιμοποιείται για να δώσει περισσότερη ή λιγότερη έμφαση στις διάφορες λέξεις της εισόδου κατά την επεξεργασία. Οι λέξεις με μεγαλύτερο βάρος λαμβάνονται περισσότερο υπόψη κατά την ενσωμάτωση πληροφοριών από το μοντέλο, ενώ οι λιγότερο σημαντικές λέξεις αγνοούνται σχεδόν ή πλήρως.

Συνοψίζοντας, ο μηχανισμός της προσοχής (attention mechanism) επιτρέπει στο μοντέλο να επιλέγει δυναμικά ποια μέρη της εισόδου πρέπει να ληφθούν υπόψη κατά την επεξεργασία. Αυτό οδηγεί σε πιο αποτελεσματική αντιμετώπιση προβλημάτων και ικανότητα του μοντέλου να αντιληφθεί συγκεκριμένες συσχετίσεις και παραδοχές που είναι απαραίτητες για τις περισσότερες εφαρμογές επεξεργασίας φυσικής γλώσσας.

4.4.3 Self Attention

Ο μηχανισμός αυτοπροσοχής (self-attention mechanism) είναι μια βασική συνιστώσα που χρησιμοποιείται στους μετασχηματιστές, μια προηγμένη αρχιτεκτονική νευρωνικών δικτύων, που επιτρέπει την αναλλοίωτη αναζήτηση σχέσεων μεταξύ διαφορετικών λέξεων ή τμημάτων της εισόδου. Οι μετασχηματιστές αυτοί είναι εξαιρετικά δημοφιλείς στον τομέα της επεξεργασίας φυσικής γλώσσας και χρησιμοποιούνται σε πολλές εφαρμογές, όπως μετάφραση, αναγνώριση φωνής και ανάλυση κειμένου.

Ο μηχανισμός αυτοπροσοχής λειτουργεί ως μηχανισμός εξαγωγής σημαντικών σχέσεων μεταξύ λέξεων σε μια πρόταση ή άλλου τύπου ακολουθίας. Όταν το μοντέλο δέχεται μια ακολουθία λέξεων (ή διανύσματα αναπαράστασης), δημιουργεί τρεις σημαντικές οντότητες:

1. Query (Ερώτηση): Ποια λέξη χρειάζεται να εστιάσουμε ή να αναλύσουμε σε βάθος.

2. Key (Κλειδί): Οι λέξεις που βοηθούν στον προσδιορισμό της σχέσης μεταξύ της ερώτησης και των άλλων λέξεων στην ακολουθία.

3. Value (Τιμή): Οι τιμές που παρέχουν το περιεχόμενο για τον υπολογισμό του εξόδου του self-attention μηχανισμού.

Από τις παραπάνω πληροφορίες, υπολογίζονται τα βάρη προσέγγισης προσοχής μεταξύ των λέξεων της εισόδου. Ουσιαστικά, η έννοια της αυτοπροσοχής επιτρέπει στο μοντέλο να δώσει περισσότερη προσοχή σε σημαντικές λέξεις ή πτυχές της εισόδου ενώ αγνοεί τις άσχετες ή λιγότερο σημαντικές λέξεις.

Η αυτοπροσοχή χρησιμοποιείται κατά την επεξεργασία κάθε επιπέδου (layer) του μετασχηματιστή και επιτρέπει την αποτελεσματική αναπαράσταση του περιεχομένου και την αναγνώριση πολύπλοκων προτύπων στα δεδομένα εισόδου. Αυτό συμβάλλει σημαντικά στην απόδοση του μοντέλου σε διάφορες εφαρμογές επεξεργασίας φυσικής γλώσσας.

Οι μηχανισμοί προσοχής και αυτοπροσοχής αναφέρονται σε δύο διαφορετικούς τρόπους αναπαράστασης και χρήσης πληροφορίας στο μοντέλο των μετασχηματιστών. Και οι δύο τεχνικές είναι κρίσιμες για τη λειτουργία τους, αλλά χρησιμοποιούνται για διαφορετικούς σκοπούς.

1. Attention (Προσοχή):

Η έννοια της προσοχής είναι μια γενική ιδέα που χρησιμοποιείται στη μηχανική μάθηση και τα νευρωνικά δίκτυα για να δώσει μεγαλύτερη βαρύτητα σε συγκεκριμένα στοιχεία ενός συνόλου. Στο πλαίσιο των μετασχηματιστών, η έννοια της προσοχής (attention) χρησιμοποιείται για να επιτρέψει στο μοντέλο να επικεντρωθεί στα σημαντικά κομμάτια της εισόδου κατά την επεξεργασία της πληροφορίας. Συνεπώς, οι συνδέσεις (βάρη) μεταξύ των διάφορων στοιχείων ενός συνόλου αποτελούνται από την αναγνώριση των σημαντικών σχέσεων μεταξύ τους.

2. Self-Attention (Αυτοπροσοχή):

Η αυτοπροσοχή είναι μια συγκεκριμένη υλοποίηση του μηχανισμού της προσοχής που χρησιμοποιείται εντός των μετασχηματιστών. Αυτό σημαίνει ότι το μοντέλο δημιουργεί συνδέσεις μεταξύ των διάφορων λέξεων ή διανυσμάτων της εισόδου (συνήθως λέξεις σε μοντέλα γλώσσας) για να αντιληφθεί τις συσχετίσεις μεταξύ τους. Σε αντίθεση με τον κλασικό μηχανισμό προσοχής, όπου υπάρχουν δύο διαφορετικά σύνολα εισόδου (π.χ. ερώτηση και περίληψη κειμένου), ο μηχανισμός αυτοπροσοχής δημιουργεί συνδέσεις εντός του ίδιου συνόλου εισόδου.

Συνοψίζοντας, η διαδικασία της προσοχής είναι μια γενική έννοια που αναφέρεται στη βαρύτητα που δίνεται σε διαφορετικά στοιχεία, ενώ η αυτοπροσοχή είναι μια συγκεκριμένη εφαρμογή του

μηχανισμού της προσοχής στο εσωτερικό των μετασχηματιστών, όπου δημιουργούνται συνδέσεις μεταξύ των διαφορετικών λέξεων ή διανυσμάτων ενός κειμένου για την αναγνώριση συσχετίσεων. Η αυτο-προσοχή είναι καθοριστική για τη λειτουργία των μετασχηματιστών, καθώς τους επιτρέπει να αντιληφθούν τις μακροπρόθεσμες συσχετίσεις μεταξύ λέξεων και να επιτύχουν εξαιρετική απόδοση σε πολλές εφαρμογές που σχετίζονται με επεξεργασία φυσικής γλώσσας.

4.4.4 Multi-Head Attention

Με τον μηχανισμό προσοχής πολλών κεφαλών (multi-head attention) αναφερόμαστε στην χρήση πολλαπλών μηχανισμών αυτοπροσοχής και πολλές φορές εν παραλληλία. Αυτό γίνεται για να πετύχουμε καλύτερη και γρηγορότερη κατανόηση γραμματικών συνδέσεων/εννοιών μεταξύ διάφορων λέξεων μεταξύ των προτάσεων που εισάγονται στην είσοδο. Ήταν μια προσθήκη η οποία έφερε την επανάσταση σε έρευνες πάνω στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP).

Ο μηχανισμός αυτός που βρίσκεται στους μετασχηματιστές δίνει τη δυνατότητα στο μοντέλο να επικεντρώνεται σε διαφορετικά σημεία της εισόδου κατά την επεξεργασία της. Υπολογίζει τα σταθμισμένα βάρη αυτών βασισμένο σε σκορ ομοιότητας μεταξύ ενός ερωτήματος (query) και ζευγών κλειδιού-τιμής (key-value). Αυτό επιτρέπει στο μοντέλο να συσχετίζει σημαντικές πληροφορίες και να αγνοεί αδιάφορα σημεία της εισόδου.

Κάθε κεφαλή γνωρίζει ένα μοναδικό σύνολο γραμμικών μετασχηματισμών (πίνακες) για τον υπολογισμό ερωτημάτων (queries), κλειδιών (keys) και τιμών (values), παρέχοντας αποτελεσματικά πολλαπλές προοπτικές στα δεδομένα εισόδου. Η έξοδος κάθε κεφαλής προσοχής στη συνέχεια συνενώνεται με το γενικό σύνολο ή υπολογίζεται ο μέσος όρος για να παραχθεί η τελική αναπαράσταση. Αυτό επιτρέπει στο μοντέλο να αντιλαμβάνεται διαφορετικούς τύπους εξαρτήσεων και σχέσεων μεταξύ λέξεων ή των tokens στην ακολουθία εισόδου, παρέχοντας πιο πλούσιες και πιο αναλυτικές αναπαραστάσεις.

Η φόρμουλα για τον υπολογισμό προσοχής πολλών κεφαλών έχει ως εξής:

1) Χωρίζουμε σε πολλαπλές κεφαλές τα δεδομένα εισόδου: Τα Q(Queries), K(Keys) και V(Values) χωρίζονται σε πολλαπλά σύνολα πινάκων, ένα για κάθε κεφαλή προσοχής.

2) Υπολογίζονται τα βάρη προσοχής για κάθε κεφαλή. Για κάθε κεφαλή, υπολογίζεται το εσωτερικό γινόμενο των πινάκων ερωτήματος (Q) και κλειδιού (K), ακολουθεί αντίστοιχη κλιμάκωση τιμών, στη συνέχεια εφαρμόζεται μια συνάρτηση softmax για να ληφθούν τα βάρη προσοχής και πολλαπλασιάζονται με τον πίνακα τιμών (V).

3) Συνδυάζονται οι έξοδοι από όλες τις κεφαλές: Οι έξοδοι από όλες τις κεφαλές προσοχής συνενώνονται ή υπολογίζονται κατά μέσο όρο για να σχηματίσουν την τελική αναπαράσταση εξόδου.

Ο μηχανισμός προσοχής πολλαπλών κεφαλών έχει δείξει σημαντικές βελτιώσεις σε διάφορες εργασίες NLP, όπως μηχανική μετάφραση, μοντελοποίηση γλώσσας, δημιουργία κειμένου και ανάλυση συναισθήματος. Χρησιμοποιώντας πολλαπλές κεφαλές, το μοντέλο μπορεί να καταγράψει διαφορετικά μοτίβα και εξαρτήσεις μέσα στα δεδομένα, καθιστώντας το πιο ισχυρό για σύνθετες εργασίες κατανόησης γλώσσας.

5 Μέθοδος Περίληψης

5.1 Στόχος βελτιστοποίησης συνοψιστή

Η βελτιστοποίηση (fine-tuning) ενός συνοψιστή (summarizer) αναφέρεται στην πρακτική εκπαίδευσης ενός υπάρχοντος μοντέλου με νέα δεδομένα προκειμένου να βελτιωθεί η απόδοσή του για συγκεκριμένες εργασίες. Αυτή η διαδικασία έχει πολλά οφέλη, καθώς επιτρέπει στο μοντέλο να μάθει από τα νέα δεδομένα και να προσαρμοστεί σε συγκεκριμένες απαιτήσεις. Ανάμεσα στα κυριότερα οφέλη της βελτιστοποίησης ενός μοντέλου που υλοποιεί την διαδικασία της περίληψης περιλαμβάνονται τα εξής:

1. Προσαρμογή στον συγκεκριμένο στόχο: Η βελτιστοποίηση επιτρέπει στο υπάρχον μοντέλο να εξελιχθεί και να προσαρμοστεί στον συγκεκριμένο τύπο περίληψης που απαιτείται για μια συγκεκριμένη εφαρμογή ή περίπτωση χρήσης. Αυτό μπορεί να βελτιώσει σημαντικά την ποιότητα των περιλήψεων που παράγονται για αυτήν την ειδική χρήση.

2. Επίλυση προβλημάτων διαθέσιμων δεδομένων: Οι υπάρχοντες αλγόριθμοι summarizing μπορεί να μην είναι ικανοποιητικοί για όλα τα είδη δεδομένων. Με την διαδικασία της βελτιστοποίησης, μπορούμε να προσαρμόσουμε το μοντέλο για να ανταποκρίνεται καλύτερα στη δομή και το περιεχόμενο των διαθέσιμων δεδομένων, προκειμένου να παράγει ποιοτικές περιλήψεις.

3. Βελτίωση απόδοσης: Η βελτιστοποίηση μπορεί να βοηθήσει στην ενίσχυση της απόδοσης του μοντέλου σε περιπτώσεις που το αρχικό μοντέλο δεν είναι ιδιαίτερα καλό ή δεν έχει εκπαιδευτεί με τα συγκεκριμένα δεδομένα που χρησιμοποιούνται τώρα.

4. Αποφυγή εκ νέου εκπαίδευσης: Η εκπαίδευση μοντέλων βασισμένων σε μεγάλα μοντέλα όπως το GPT-3 είναι χρονοβόρα και απαιτεί πολλούς υπολογιστικούς πόρους. Μέσω της βελτιστοποίησης μας επιτρέπεται να επωφεληθούμε από την υπάρχουσα γνώση του μοντέλου και να το προσαρμόσουμε στις νέες απαιτήσεις, χωρίς να ξεκινήσουμε την εκπαίδευση από το μηδέν.

Συνοψίζοντας, η βελτιστοποίηση ενός μοντέλου περίληψης μπορεί να βελτιώσει την ποιότητα και την απόδοση του μοντέλου για συγκεκριμένες εργασίες και δεδομένα, και ταυτόχρονα μειώνει τον χρόνο και τους πόρους που απαιτούνται για την εκπαίδευση του μοντέλου.

5.2 Προεπεξεργασία δεδομένων

Ο στόχος μας σε αυτό το κεφάλαιο είναι να περιγράψουμε όλη την διαδικασία που ακολουθήσαμε εώς ότου φτάσουμε στο στάδιο να εκπαιδεύσουμε δικό μας συνοψιστή με αντίστοιχο σύνολο δεδομένων. Το συνόλο αυτό δεδομένων το οποίο είχαμε στη διάθεσή μας ήταν το Large Scale Movie Description Challenge Dataset, το οποίο περιέχει αποσπάσματα από ένα αρχικό βίντεο/ταινία, όπου για κάθε απόσπασμα εμπεριέχει αντίστοιχη περιγραφή σε αντίστοιχα csv αρχεία. Η δομή όλου του συνόλου δεδομένων είναι σύνθετη καθώς στοχεύει σε πολλών ειδών εφαρμογές, άρα και η δομή θα πρέπει να εξυπηρετεί κι άλλες περιπτώσεις χρήσης. Συγκεκριμένα να αναφέρουμε το σύνολο δεδομένων περιλαμβάνει ένα σύνολο εκπαίδευσης (περ. 100000 εγγραφές), ένα σύνολο τεστ (περ. 10000 εγγραφές) και τέλος ένα σύνολο validation (περ. 7500 εγγραφές). Για την δική μας περίπτωση αρκούσαν μόνο τα αρχεία δεδομένων που περιείχαν μόνο τον τίτλο του αποσπάσματος και την περιγραφή του ίδιου του αποσπάσματος. Έχοντας στην κατοχή μας τα δεδομένα αυτά, αρκούσαν να προχωρήσουμε στο επόμενο βήμα το οποίο ήταν να παράξουμε δικές μας περιγραφές μέσω του CLIP για όλες τις δειγματοληπτημένες εικόνες κάθε αποσπάσματος.

Ένα σημαντικό πρόβλημα που δημιουργήθηκε πειραματικά κατά την παραγωγή περιγραφών για ένα απόσπασμα με το CLIP, ήταν ότι μερικές περιγραφές ήταν πολύ παρόμοιες μεταξύ τους, εώς και πολλές φορές ίδιες. Ο λόγος ήταν ότι τα περισσότερα αποσπάσματα είναι μικρά σε διάρκεια (περ. 3 - 10 secs) και έτσι η πιθανότητα τα καρέ να μοιάζουν μεταξύ τους μεγάλη. Έτσι λοιπόν, το πρόβλημα αυτό το λύσαμε ή τουλάχιστον ελαχιστοποιήθηκε με τον εξής τρόπο:

Εφόσον είμαστε σε θέση να γνωρίζουμε το μέγιστο πλήθος καρέ κάθε αποσπάσματος, ορίζουμε ένα threshold (π.χ. 120), όπου δηλαδή θεωρούμε ότι αν το πλήθος των καρέ είναι πάνω από 120 τότε πρέπει να λάβουμε 4 frames συνολικά, σε αντίθετη περίπτωση μόνο 3 και θα εξηγήσουμε αμέσως την διαδικασία αυτή.

Αν το πλήθος των frames που θέλουμε να πάρουμε είναι 3, τότε χωρίζουμε σε 3 ίσα νοητά groups τα συνολικά frames. Δηλαδή αν συνολικά έχουμε 60 frames τότε έχουμε 3 ίσα σύνολα όπως παρακάτω:

- 1ο Group (1^{ο} Frame – 20^{ο} Frame)
- 2ο Group (21^{ο} Frame – 40^{ο} Frame)
- 3ο Group (41^{ο} Frame – 60^{ο} Frame)

Έτσι λοιπόν από κάθε διάστημα λαμβάνουμε ένα frame με τυχαίο τρόπο. Με αυτόν τον τρόπο καταφέρνουμε να λαμβάνουμε frames από την αρχή εώς το τέλος του αποσπάσματος και η τυχαιότητα να κάνει πιο αμερόληπτη την επιλογή μας. Το ίδιο φυσικά ισχύει και για την περίπτωση των 4 frames. Να σημειώσουμε εδώ ότι αν σε περίπτωση που το αποτέλεσμα της διαίρεσης (για εύρεση πλήθους frames ανά group) προκύψει ως δεκαδικό νούμερο, τότε γίνεται στρογγυλοποίηση προς τα κάτω, με αποτέλεσμα τα τελευταία 1-2 frames να χάνονται, γεγονός που δεν μας επηρεάζει φυσικά.

Διατηρώντας τα frames με την σειρά που τα λάβαμε από την ολοκλήρωση της δειγματοληψίας, παράγουμε για κάθε ένα από αυτά τα frames από μια περιγραφή και με κατάλληλη επεξεργασία τα ενώνουμε μεταξύ τους κι έτσι δημιουργείται ένα μικρό κειμενάκι που αναφέρεται εν συντομία στο εκάστοτε απόσπασμα. Σε δεύτερο χρόνο το κειμενάκι αυτό θα χρησιμοποιηθεί ως είσοδο (input) σε κάποιον έτοιμο συνοψιστή για να παραχθεί η περιληψή του και να γίνει ακόμα πιο μικρό, όμως αυτό θα το περιγράψουμε παρακάτω, ας επικεντρωθούμε τώρα στο πως καταφέραμε να παράξουμε περιγραφές για κάθε απόσπασμα καθώς όταν μιλάμε για σύνολα δεδομένων των 100000^{ων} εγγραφών, αναφερόμαστε σε μια πρόκληση. Μην ξεχνάμε πως το CLIP είναι ένα σύνθετο νευρωνικό δίκτυο και για να εκτελεστεί απαιτεί χρόνο. Πιο απλά, για κάθε απόσπασμα (των 3 ή 4 frames) θα εκτελεστεί 3 ή 4 φορές ανάλογα, το οποίο απαιτεί χρόνο και υπολογιστικούς πόρους φυσικά.

Η λύση μας σε αυτό το υπολογιστικό πρόβλημα ήταν να παραλληλοποιήσουμε (με χρήση threads - σε γλώσσα Python) το πρόβλημα μας και έτσι χωρίσαμε τα σύνολα δεδομένων μας σε υπο-ομάδες ίσου πλήθους και έτσι κάθε thread ανέλαβε κι από μια εκτελώντας την διαδικασία που προαναφέραμε. Δηλαδή την παραγωγή περιγραφών ανά απόσπασμα και τέλος κάθε thread τα εισάγει σε δικό του excel αρχείο, με συγκεκριμένο id κάθε αρχείο. Να αναφέρουμε εδώ ότι προτιμήσαμε excel αρχεία, καθώς είναι πιο εύχρηστα σε αντίθεση με τα csv αρχεία.

Τέλος, με αντίστοιχο κώδικα συνενώσαμε τα excel αρχεία που παρήγαγε το κάθε thread με την κατάλληλη ταυτοποίηση του κάθε αρχείου (βάσει id), ώστε να μπουν με την αρχική σειρά για να τηρηθεί η αρχική δομή.

Παρακάτω παρουσιάζουμε τους χρόνους (σε μέρες) που μας πήρε για να επεξεργαστούμε και να δημιουργήσουμε τα 3 σύνολα δεδομένων (train – test – validation).

Είδος συνόλου δεδομένων	Ημέρες
Train	4
Test	1.5
Validation	1

Πίνακας 1: Χρόνοι δημιουργίας (σε ημέρες) κολώνας ανά σύνολο δεδομένων με CLIP

Επόμενο μας βήμα αφού κατασκευάσαμε το νέο σύνολο δεδομένων μας με τις αντίστοιχες περιγραφές για κάθε απόσπασμα, ήταν να εξάγουμε την περίληψη για κάθε περιγραφή που προέκυψε. Η διαδικασία αυτή έγινε χωριστά από την διαδικασία εξαγωγής περιγραφής και όχι παράλληλα δηλαδή, καθώς και οι 2 διαδικασίες αυτές είναι αρκετά βαριές υπολογιστικά οπότε υπήρχε θέμα με την μνήμη RAM (έλλειψη πόρων). Έτσι η εκτέλεση αυτή έγινε σε άλλη φάση και η οποία αποδείχθηκε ακόμη πιο βαριά υπολογιστικά σε αυτήν την περίπτωση και δεν κατέστη δυνατή η παραλληλοποίηση του προβλήματος αυτού και έπρεπε κάθε περιγραφή να εκτελείται μόνη της χωρίς κάποια αντίστοιχη διαδικασία να εκτελείται παράλληλα (parallel threads). Κάθε δοκιμή που έγινε δεν απέφερε καρπούς, οπότε ακολουθήσαμε την κλασική οδό, αυτή της σειριακής εκτέλεσης της διαδικασίας που αναφέρουμε.

Παρακάτω βλέπουμε την παραγωγή περίληψης για κάθε κείμενο που αρχικά παρήχθη μέσω του CLIP εξ' ολοκλήρου, για όλα τα διαθέσιμα σύνολα δεδομένων (train – test – validation).

Είδος συνόλου δεδομένων	Ημέρες
Train	7
Test	3
Validation	2.5

Πίνακας 2: Χρόνοι δημιουργίας κολώνας ανά σύνολο δεδομένων με χρήση συνοψιστή.

Τελικά τα σύνολα δεδομένων που δημιουργούνται έχουν την εξής μορφή (ενδεικτική παρουσίαση για 1 γραμμή):

Όνομα αποσπάσματος	Δοσμένη περίληψη (Target)	CLIP **(Είσοδος 1)**	Περίληψη **(Είσοδος 2)**

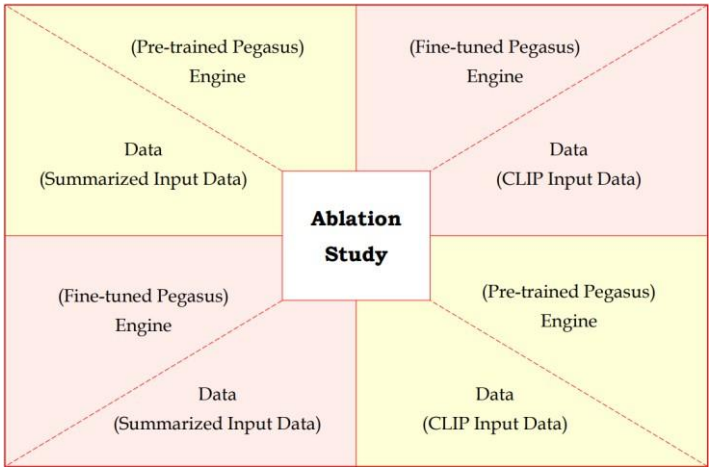
Πίνακας 3: Μορφή ενός συνόλου δεδομένων για κάθε γραμμή.

5.3 Βελτιστοποίηση μοντέλων

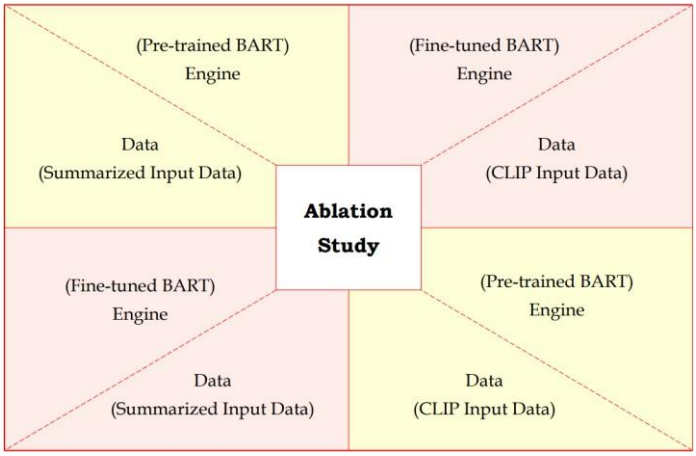
Αφού έχουμε ένα εκτενές σύνολο δεδομένων πλέον, είναι σημαντικό να προχωρήσουμε σε μια διερευνητική μελέτη (ablation study). Αυτό σημαίνει ότι εφόσον ο σκοπός μας είναι να φτιάξουμε ή να χρησιμοποιήσουμε έναν ή πολλούς αποδοτικούς συνοψιστές, καλούμαστε να ακολουθήσουμε δύο οδούς. Η πρώτη οδός είναι να ελέγξουμε ήδη εκπαιδευμένους συνοψιστές αν κάνουν καλή δουλειά, δηλαδή παράγουν καλά summaries. Τέλος, η δεύτερη οδός είναι να χτίσουμε δικό μας συνοψιστή, ή πιο σωστά να εκπαιδεύσουμε με δικά μας δεδόμενα έναν ή και παραπάνω συνοψιστές και να ελέγξουμε πάλι την αποδοτικότητά του καθενός που εκπαιδεύσαμε. Το αποτέλεσμα της διερευνητικής μελέτης (ablation study) θα είναι ο τελικός συνοψιστής είτε που θα χει προκύψει από υπάρχουσα εκπαίδευση είτε βελτιστοποιημένος, που θα καλύπτει όσο δυνατόν περισσότερο τις ανάγκες μας.

Έχοντας υπόψιν μας ότι σε ένα μοντέλο εξετάζουμε την ακρίβειά του με χρήση του συνόλου test εισάγοντας την αντίστοιχη είσοδο και αναμένοντας αντίστοιχη έξοδο, αλλά και το ότι κάθε σύνολο

έχει δύο ειδών εισόδους (CLIP, Περίληψη του CLIP), τότε έχουμε τα παρακάτω διαγράμματα να ισχύουν και θα εξηγήσουμε λίγο παρακάτω το πως ερμηνεύονται.



Εικόνα 8 : 1ο σκέλος ablation study



Εικόνα 9: 2ο σκέλος ablation study

Αρχικά στο πρώτο διάγραμμα [Εικόνα 8] βλέπουμε να γίνεται χρήση του Engine Pegasus (της Google) [16]. Επίσης τα δύο κίτρινα πλαίσια αφορούν στον προ-εκπαιδευμένο Pegasus με είσοδο-test κείμενο CLIP και συνοψισμένο κείμενο CLIP. Τέλος, τα πορτοκαλί πλαίσια αφορούν στον βελτιωμένο (fine-tuned) Pegasus, αυτόν δηλαδή που εκπαιδεύσαμε με δικά μας δεδομένα και πάλι γίνεται χρήση των ίδιων εισόδων-test αντίστοιχα για κάθε περίπτωση. Η ίδια λογική ακολουθείται και στο δεύτερο διάγραμμα [Εικόνα 9], απλώς πλέον έχουμε τον BART [5],[13],[15] ως Engine.

Με τα δεδομένα που έχουμε πλέον στην διάθεση μας ας προχωρήσουμε αρχικά σε ενδεικτικό υπολογισμό μετρικών (κύριες μετρικές: Rouge-1, Meteor).

Υποψήφια-πρόταση **CLIP** **Περίληψη**	**Μετρικές** **Rouge1, Meteor**	**Πρόταση-στόχος**
The library in the thriller film.	Rouge1: 0.45, meteor: 0.23	Now in a library
The library in a movie.	Rouge1: **0.60**, meteor: **0.59**	
A man is standing in a dark room and he is holding a cup of coffee. The toilet in the bathroom. The bathroom in the movie. A man walks through a dark room and looks at the camera.	Rouge1: **0.14**, meteor:**0.26**	Passing a row of toilet stalls
A man is standing in a dark room.	Rouge1: 0.14, meteor: 0.16	
Young women walking in the park. A group of teenage girls walk down a sidewalk.	Rouge1: 0.19, meteor: 0.14	The roommates walk through the campus
A group of girls walk down the sidewalk.	Rouge1: **0.28**, meteor: **0.16**	
A pilot in a cockpit of an aircraft. The view from the cockpit of a helicopter. A view from the window of a plane.	Rouge1: 0.07, meteor: 0.31	In the cockpit.
A pilot is in the cockpit.	Rouge1: **0.44**, meteor: **0.92**	
The movie opens with a scene of a massive explosion. The movie opens with a scene from the film.	Rouge1: 0, meteor: 0.060	It shows craters and burning wreckage.
The movie opens with a scene.	Rouge1: 0, meteor: **0.07**	

Πίνακας 4: Προβολή μετρικών για CLIP–στόχος και Περίληψη–στόχος

Παρατηρούμε βάσει του παραπάνω πίνακα πως η χρήση συνοψιστή ορισμένες φορές μπορεί να λειτουργήσει σωτήρια, ενώ άλλες ίσως δεν προσφέρει κάποια βελτίωση. Να σημειωθεί εδώ ότι μια τιμή πολύ κοντά στο 0.50 είναι ένδειξη μιας πολύ καλής προσέγγισης. Επίσης παρατηρούμε και περίπτωση στην οποία ούτε το CLIP, αλλά ούτε κι ο συνοψιστής κατάφεραν να προσεγγίσουν το πρόταση-στόχος (target), με αποτέλεσμα οι μετρικές να χουν τιμές μηδέν ή κοντά στο μηδέν.

Οι μετρικές είναι μια καλή πρακτική φυσικά για να αξιολογήσουμε τα text-based μοντέλα, όμως δεν είναι αντιπροσωπευτικές μερικές φορές.

Τώρα ενδεικτικά θα παρουσιάσουμε με εικόνες, το πως λειτουργήσαμε για την παραγωγή περιγραφών για κάθε frame και πως τελικά παρήχθη η περίληψη. Στο παράδειγμά μας θα χρησιμοποιήσουμε το απόσπασμα που αφορά την 4η γραμμή του παραπάνω πίνακα (Πίνακας 4).

1) CLIP: A pilot in a cockpit of an aircraft.

2) CLIP: The view from the cockpit of a helicopter.

3) CLIP: A view from the window of a plane.

Συνένωση περιγραφών από CLIP:

(A pilot in a cockpit of an aircraft.)

+ (The view from the cockpit of a helicopter.)

+ (A view from the window of a plane.)

Να αναφέρουμε ότι πριν την συνένωση όλων των περιγραφών ακολουθεί πάντα μια διόρθωση στις προτάσεις, ώστε το συνολικό αποτέλεσμα της συνένωσης να φαίνεται ως πραγματικό κείμενο. Π.χ. Μετατροπή σε κεφαλαίο γράμμα το πρώτο γράμμα της εκάστοτε πρότασης, αλλά και προσθήκη τελείας στο τέλος κάθε πρότασης.

Περίληψη περιγραφών από CLIP - βίντεο: A pilot is in the cockpit.

(Με πρόταση-στόχο : In the cockpit.)

Πράγματι είναι σκηνές που αφορούν ένα πιλοτήριο.

Ολοκληρώνοντας την ανάλυση των δεδομένων μας χρησιμοποιώντας κατάλληλες μετρικές, λαμβάνοντας υπόψιν κάποιες ειδικές περιπτώσεις κ.ο.κ., μπορούμε να προχωρήσουμε στην εφαρμογή της διερευνητικής μελέτης που αναφέραμε παραπάνω.

Έχουμε λοιπόν στην διάθεση μας δύο Engines: τον Pegasus, αλλά και τον BART [14], τα οποία και αυτά θα χρησιμοποιήσουμε στο πείραμά μας. Οι περιπτώσεις είναι οι εξής και ισχύουν για κάθε ένα Engine που προαναφέραμε:

1) Εκπαίδευση αντίστοιχου μοντέλου (Pegasus, BART) με είσοδο ένα από τα είδη εισόδων που έχουμε αναφέρει και έξοδο (target) τη δοσμένη περιγραφή από το αρχικό σύνολο δεδομένων. Τέλος, λήψη μετρικών και εξαγωγή συμπερασμάτων.
2) Χρήση αντίστοιχου μοντέλου (Pegasus, BART) με είσοδο ένα από τα είδη εισόδων που έχουμε αναφέρει και λήψη μετρικών ως προς τα αποτελέσματα που δίνει κάθε μοντέλο λαμβάνοντας υπόψιν την έξοδο που δίνει με την έξοδο-στόχος (output-target) [17].

Παρακάτω βλέπουμε στο πινακάκι τις τιμές της μετρικής Rouge-1 που προέκυψαν ανά περίπτωση (8 περιπτώσεις). Ας εξηγήσουμε τώρα πως ερμηνεύεται το πινακάκι αυτό. Η πρώτη (μεγάλη) στήλη αφορά το μοντέλο Pegasus, όπου όπως φαίνεται υπάρχουν δύο υπο-στήλες με τίτλους “Finetuned/Βελτιστοποιημένο” και “Non Finetuned/ Όχι βελτιστοποιημένο” αντίστοιχα. Σε κάθε μια περίπτωση έχουμε και δύο ειδών εισόδους (συνοψισμένο κείμενο-CLIP και κείμενο-CLIP) με τις αντίστοιχες τιμές στην μετρική Rouge-1.

Pegasus		*BART*	
Finetuned	**Non Finetuned**	**Finetuned**	**Non Finetuned**
INPUT	***INPUT***	***INPUT***	***INPUT***
Summarized CLIP Text: 0.1175	**Summarized CLIP Text: 0.046**	**Summarized CLIP Text: 0.15**	**Summarized CLIP Text: 0.0368**
CLIP Text: 0.129	**CLIP Text: 0.044**	**CLIP Text: 0.138**	**CLIP Text: 0.0359**

Εικόνα 10 : Πίνακας αποτελεσμάτων αξιολόγησης μοντέλων

Παρατηρούμε ότι και στις δύο περιπτώσεις μοντέλων η διαδικασία της βελτιστοποίησης απέφερε καρπούς, καθώς είναι σχετικά καλά τα αποτελέσματα ως προς την τιμή του Rouge, συγκριτικά με τις τιμές Rouge για τις περιπτώσεις που δεν γίνονται ενέργειες βελτίωσης των αντίστοιχων μοντέλων. Προχωρώντας και στις επιμέρους περιπτώσεις, βλέπουμε ότι ο BART έχει αρκετά καλύτερες τιμές σε αντίθεση με τον Pegasus (Finetuned). Τέλος, μια βασική παρατήρηση που θα κάναμε και θα ήταν καλό να το αναφέρουμε είναι πως στον βελτιστοποιημένο Pegasus, η εκπαίδευση με είσοδο “κείμενο CLIP” δίνει καλύτερα αποτελέσματα αν δει κανείς την τιμή μετρικής που λαμβάνεται έπειτα από εκπαίδευση με είσοδο “συνοψισμένο κείμενο CLIP”. Το αντίστροφο ισχύει για το μοντέλο του BART.

Για να ολοκληρωθεί το πείραμά μας και η τελική ανάλυση να γίνει όσο το δυνατόν πιο σωστά λαμβάνοντας όλα τα δυνατά εργαλεία που έχουμε στη διάθεσή μας (κυρίως αναφερόμαστε σε δωρεάν εργαλεία), ήταν σημαντικό να πειραματιστούμε και με το πιο εμπορικό εργαλείο που κανείς μπορεί να κάνει διάφορα ενδιαφέροντα πράματα, μέσα σε αυτά περιλαμβάνεται και η περίληψη κειμένου, δεν είναι άλλο από το Chat-GPT (έκδοση 3.5). Ας αναφέρουμε μερικά πράματα για το Chat-GPT τα οποία είναι τα εξής: Το Chat-GPT είναι ένα προηγμένο σύστημα τεχνητής νοημοσύνης που βασίζεται στην

αρχιτεκτονική GPT (Generative Pre-trained Transformer) που αναπτύχθηκε από την OpenAI. Αυτό το σύστημα είναι σχεδιασμένο να κατανοεί και να απαντά σε ανθρώπινη γλώσσα, παρέχοντας λύσεις και πληροφορίες σε ποικίλες εφαρμογές.

Το Chat-GPT λειτουργεί με τη χρήση ενός εκπαιδευμένου μοντέλου που έχει μάθει από τεράστια ποσότητα κειμένου από τον παγκόσμιο ιστό. Αυτή η εκπαίδευση του επιτρέπει να έχει γενικές γνώσεις, να αναγνωρίζει λεξιλογικές συνδέσεις και συντακτικούς κανόνες, επιτρέποντάς του να ανταποκρίνεται σε ερωτήσεις και αιτήματα.

Εν τέλει, το Chat-GPT είναι ένα ισχυρό εργαλείο που βοηθά στην αυτοματοποίηση εργασιών που σχετίζονται με τη φυσική γλώσσα και επιτρέπει τη δημιουργία και την πρόσβαση σε πληροφορίες με άνεση και αποτελεσματικότητα.

Ας έρθουμε τώρα στο δικό μας πείραμα και πως έγινε η χρήση του Chat-GPT. Για το πείραμά μας έγινε χρήση της δωρεάν έκδοσης μέσω του API (Application Programming Interface) που μας προσφέρει η OpenAI με έναν περιορισμό στην χρήση του φυσικά, οπότε έπρεπε να γίνει σύντομη και περιεκτική έρευνα για τα αποτελέσματα που δίνει όσον αφορά στο summarizing. Τα αποτελέσματα παρόλα αυτά δεν ήταν τόσο εντυπωσιακά, καθώς ήταν πολύ κοντά σε αυτά που δίνανε BART και Pegasus. Επιλέχθηκαν με τυχαίο τρόπο 50 περιγραφές προς περίληψη και με τον περιορισμό ότι δεν μπορούσαμε να κάνουμε κλήση στο API πολλές συνεχόμενες φορές, αλλά τοποθετώντας κατάλληλο timer καταφέραμε κατά κάποιον τρόπο να προσπεράσουμε το συγκεκριμένο πρόβλημα και να καταλήξουμε τελικά να εξάγουμε κάποια σημαντικά συμπεράσματα. Θα ήταν δύσκολο να χρησιμοποιηθεί το Chat – GPT καθόλη την διάρκεια της εκπόνησης της εργασίας μας για τον λόγο που αναφέραμε.

Σε ένα ερευνητικό θα λέγαμε πλαίσιο, καλό είναι να αναφέρουμε και τα χαρακτηριστικά του συστήματος που είχαμε στην διάθεση μας, για την εκπόνηση όλων των απαιτούμενων διαδικασιών, καθώς παίζουν πολύ σημαντικό ρόλο για τα τελικά αποτελέσματα, αλλά παράλληλα επεξηγούνται πολλά στα σημεία όπου παρουσιάζουμε χρόνους εκτέλεσης κ.λπ. Οπότε λοιπόν τα τεχνικά χαρακτηριστικά του συστήματός μας ήταν τα εξής:

CPU	Intel Core i7 , 12th Gen
RAM	16 Gb
HARD DISK TYPE	NVMe M.2
OPERATING SYSTEM	Windows 11

6 Συμπεράσματα – Περίληψη

Το NLP (Natural Language Processing) είναι ένα πεδίο της τεχνητής νοημοσύνης (ΤΝ) που ασχολείται με τον τρόπο με τον οποίο οι υπολογιστές αλληλεπιδρούν με την ανθρώπινη γλώσσα. Ακολούθως στοιχειοθετούνται τα βασικά χαρακτηριστικά του NLP:

1. Κατανόηση Γλώσσας: Το NLP ασχολείται με την κατανόηση της ανθρώπινης γλώσσας από τους υπολογιστές. Αυτό περιλαμβάνει την ανάλυση, την ερμηνεία και την εξαγωγή νοήματος από κείμενο ή ομιλία.

2. Παραγωγή Γλώσσας: Εκτός από την κατανόηση, το NLP επιδιώκει επίσης τη δημιουργία ανθρώπινης φυσικής γλώσσας από υπολογιστές, όπως η γραφή κειμένου ή η σύνθεση φωνητικής ομιλίας.

3. Εφαρμογές: Το NLP έχει εφαρμογές σε πολλούς τομείς, συμπεριλαμβανομένων των μηχανικών αναζήτησης, της ανάλυσης κοινωνικών μέσων, της αυτόματης μετάφρασης, της αναγνώρισης φωνής, των ειδήσεων και της ρομποτικής.

4. Εξέλιξη: Το NLP εξελίσσεται συνεχώς, και τα προηγμένα μοντέλα όπως το GPT (Generative Pre-trained Transformer) έχουν επιτύχει εντυπωσιακά αποτελέσματα σε πολλούς τομείς. Η έρευνα στο NLP συνεχίζεται για τη βελτίωση της κατανόησης και της παραγωγής γλώσσας.

Το NLP έχει ευρείες εφαρμογές και αναπτύσσεται συνεχώς με την ανάπτυξη νέων τεχνικών και μοντέλων, καθιστώντας το ένα από τα πλέον σημαντικά πεδία της τεχνητής νοημοσύνης.

Το παρόν έργο θίγει με πολύ πρακτικό τρόπο το επιστημονικό επίτευγμα που δίνει την δυνατότητα στον υπολογιστή να αναγνωρίσει, να κατανοήσει πολύπλευρες σημασίες της ανθρώπινης γλώσσας, αλλά και κατανόηση εικόνων -ακόμα και σύνθετων εικόνων πολλές φορές. Πλέον η δυνατότητα αυτή του υπολογιστή είναι γεγονός αφού απορρέει από την εξέλιξη του επιστημονικού πεδίου της τεχνητής νοημοσύνης και παρέχεται φυσικά σε ποικίλες εφαρμογές για την επίτευξη ενός ευρέος συνόλου στόχων.

Όπως είναι γνωστό όλοι οι επιστημονικοί τομείς έχουν το περιθώριο να βελτιωθούν και να εξελιχθούν, εξού και η έρευνα με την επιστήμη είναι δύο συνυφασμένες έννοιες και άρρηκτα συνδεδεμένες. Όμως, αυτό το γεγονός επηρεάζει φυσικά και τον τομέα του NLP, καθώς όλα τα μοντέλα που χτίζονται για αναγνώριση-κατανόηση γλώσσας κ.ο.κ. έχουν ακόμα σημαντικά προβλήματα, τα οποία αναμφίβολα με την πάροδο των ετών σημειώνουν πολύ σημαντική βελτίωση, γεγονός που καθιστά το αντικείμενο της τεχνητής νοημοσύνης ένα συναρπαστικό και ιδιαίτερα γοητευτικό αντικείμενο. Παρόλα αυτά πρέπει να εξετάσουμε το παρόν και να βρίσκουμε τρόπους να προσπερνάμε τα εμπόδια που δημιουργούνται με όλα τα δυνατά μέσα που κατέχουμε και έχουμε την δυνατότητα να χρησιμοποιήσουμε στα εκάστοτε υπολογιστικά συστήματα που μας παρέχονται, καθώς είναι γνωστό ότι πολλά μοντέλα έχουν αρκετά υψηλό υπολογιστικό κόστος και πολλές φορές καλούμαστε να επιλέγουμε ανάμεσα σε ποιότητα αποτελεσμάτων με αντίστοιχο βαρύ υπολογιστικό κόστος ή σε γρήγορα αποτελέσματα επαρκούς σχετικά ποιότητας.

Στην προκειμένη περίπτωση ήταν μια αρκετά σύνθετη εργασία, όμως αν μπορούσαμε να απαντήσουμε στην ερώτηση ποια είναι τα πιο ζωτικά σημεία της όλη διαδικασίας, η απάντηση θα ήταν χωρίς αμφιβολία η παραγωγή ερμηνείας μιας εικόνας μέσω του CLIP καθώς και το αντίστοιχο engine που παράγει την περίληψη. Τυπικά, η διαδικασία παραγωγής ερμηνείας εικόνας φυσικά προηγείται της διαδικασίας της παραγωγής περίληψης αλλά και από αυτήν εξαρτάται και η τελική περίληψη, οπότε θα λέγαμε πως είναι το πιο σημαντικό κομμάτι της όλης εργασίας. Έτσι λοιπόν, μια προφανής θα λέγαμε βελτίωση που ίσως θα μπορούσαμε να κάνουμε είναι να δοκιμάσουμε άλλα

μοντέλα εκτός του CLIP, όπως για παράδειγμα το BLIP (Bootstrapping Language - Image Pre-training for Unified Vision-Language Understanding and Generation- https://huggingface.co/docs/transformers/main/model_doc/blip), όπου πρακτικά έχουμε υψηλότερο υπολογιστικά κόστος συγκριτικά με το CLIP, όμως μεγαλύτερο accuracy, γεγονός που σημαίνει ότι ίσως θα είχαμε καλύτερα αποτελέσματα όσον αφορά στις ερμηνείες εικόνων, όμως θα υπήρχε μεγάλη αναμονή αποτελεσμάτων λόγω του ότι σαν μοντέλο απαιτεί πολλούς πόρους. Καταλήγουμε δηλαδή σε αυτό που αναφέραμε παραπάνω ότι ανάλογα το τι θέλουμε να επιτύχουμε καλούμαστε να επιλέξουμε το κατάλληλο «πακέτο» που μπορούμε να έχουμε στην κατοχή μας, αλλά και να μπορούμε να θέσουμε όλο αυτό το σύστημα να λειτουργεί αρμονικά, καθώς οι δυσκολίες είναι πολλές όσο τα μοντέλα που χρησιμοποιούνται είναι σύνθετα και υπολογιστικά βαριά.

Επίσης, δεν θα ήταν ασήμαντο να αναφέρουμε πως και ως προς το κομμάτι της δειγματοληψίας ενός βίντεο θα μπορούσαμε να κάνουμε κάποιες έξυπνες υλοποιήσεις, όπως για παράδειγμα οι εξής:

- Ορίζοντας ένα ποσοστό ομοιότητας που θα δρα ως κατώφλι (threshold), μπορούν να συγκρίνονται δύο δειγματοληπτημένες εικόνες και βάσει του ποσοστού να αποθηκεύεται μόνο μία από τις 2.
- Εικόνες που δεν προσφέρουν κάτι στο σύνολο για την παραγωγή περίληψης να αφαιρούνται (π.χ. τίτλοι τέλους μιας ταινίας).
- Χρήση έξυπνων φίλτρων σε κάθε εικόνα για βελτίωση της ποιότητας, ώστε να βοηθηθεί το μοντέλο στην καλύτερη κατανόηση της.

Αυτές είναι μερικές από τις τροποποιήσεις/προσθήκες που θα μπορούσαμε να κάνουμε στο κομμάτι δειγματοληψίας, όμως τα αναφέρουμε με μια μικρή επιφύλαξη καθώς δεν είναι βέβαιο πως θα βελτιωθούν οι ερμηνείες που παράγονται από το CLIP, κι οποιοδήποτε μοντέλο δηλαδή που έχουμε στην διάθεσή μας. Στον αντίποδα, όσον αφορά στους συνοψιστές δεν υπάρχει κάτι προφανές που θα μπορούσαμε να κάνουμε όταν πειραματικά αποδεικνύεται πως το βασικό πρόβλημα που δημιουργήθηκε στις περιλήψεις οφειλόταν κυρίως στα προϋπάρχοντα λάθη του CLIP. Επιπρόσθετα, η χρήση τριών διαφορετικών συνοψιστών (BART, Pegasus, ChatGPT) που πραγματοποιήθηκε για τις ανάγκες της εργασίας, περιορίζει σημαντικά τη λίστα με τους πιθανούς διαφορετικούς συνοψιστές που θα μπορούσαν να χρησιμοποιηθούν.

Ολοκληρώνοντας την παρουσίαση βελτιώσεων θα θέλαμε να αναφέρουμε και την εφαρμογή τεχνικών Speech – To – Text με σκοπό την εξαγωγή ομιλίας σε μορφή κειμένου μέσα από τις ταινίες/σειρές. Με αυτόν τον τρόπο θα μπορούσαμε να καλύψουμε πιθανά κενά που θα υπήρχαν ερμηνεύοντας μόνο την εικόνα. Μια τέτοια υλοποίηση φυσικά απαιτεί πολύ έξυπνους αλγορίθμους και τεχνικές, που όπως είναι αναμενόμενο υπερβαίνει κατά πολύ το φόρτο εργασίας και το συνολικό μέγεθος της.

Εν κατακλείδι, είναι απαραίτητο να υπογραμμιστεί ότι η παρούσα εργασία ήρθε αντιμέτωπη με πληθώρα προκλήσεων, όπως ο ρυθμός δειγματοληψίας, το περιεχόμενο των εικόνων που έπρεπε να περιοριστεί καθώς το CLIP δεν είναι εκπαιδευμένο σε εικόνες όπως εικόνες κινουμένων σχεδίων ή επιστημονικής φαντασίας, αλλά και σε εικόνες αρκετά σύνθετες με αφηρημένο περιεχόμενο, και τέλος το βασικό πρόβλημα που έπρεπε να αντιμετωπίσουμε ήταν το υπολογιστικό κόστος που υπήρχε στο γενικό πλαίσιο της εργασίας. Αντιμετωπίστηκαν με την κατάλληλη χρήση αλγορίθμων και διαμέρισης του προβλήματος σε μικρότερα και «άθροιση» τους στο τέλος (μια μορφή του αλγορίθμου διαίρει και βασίλευε – divide and conquer). Φυσικά με τις κατάλληλες βελτιώσεις και προσθήκες μπορεί να παρέχει εύρος δυνατοτήτων και λύσεων σε πολλούς τομείς, όπως αυτόματη παραγωγή περιλήψεων σε ταινίες/σειρές σε γνωστές πλατφόρμες, δυνατότητα παρακολούθησης μιας ταινίας από άτομα με προβλήματα όρασης εφαρμόζοντας φυσικά περαιτέρω αλλαγές και τροποποιήσεις, αλλά και χρήση σε περιπτώσεις όπου είναι ανάγκη η γρήγορη περίληψη βίντεο που αφορά κάποιο συνέδριο, κάποιο σεμινάριο ή κάποια παρουσίαση, σε περιπτώσεις όπου κάποιος θα ήθελε πολύ γρήγορα να καταλάβει τι πραγματεύεται το εκάστοτε υλικό.

7 Βιβλιογραφία

[1] Cherti M, Beaumont R, Wightman R, Wortsman M, Ilharco G, Gordon C, Schuhmann C, Schmidt L, Jitsev J. Reproducible scaling laws for contrastive language-image learning. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (pp. 2818-2829).

[2] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision" (Ημερομηνία δημοσίευσης: 15 Ιανουαρίου 2021).

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "CLIP: Connecting Text and Images through Contrastive Learning" (Ημερομηνία δημοσίευσης: 4 Φεβρουαρίου 2021).

[4] Vaswani et al., (2017). "Attention Is All You Need".

[5] Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension" (2019).

[6] "ImageNet Classification with Deep Convolutional Neural Networks".

[7] "Very Deep Convolutional Networks for Large-Scale Image Recognition".

[8] "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation".

[9] Chin-Yew Lin and Eduard Hovy. "ROUGE: A Package for Automatic Evaluation of Summaries (2004)".

[10] Yann LeCun, Léon Bottou, Yoshua Bengio και Patrick Haffner. "Gradient-Based Learning Applied to Document Recognition".

[11] Xu et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" (2015).

[12] Chung et al. "Attention and Augmented Recurrent Neural Networks" (2014).

[13] Yinhan Liu, Marjan Ghazvininejad, Omer Levy, James Y. Wu, Luke Zettlemoyer, και Veselin Stoyanov. "BART: Sequence-to-Sequence Generative Model with Denoising Autoencoder Objectives".

[14] Ziniu Hu, Lingfei Wu, P. M. Podsadni, και Carlotta Domeniconi. "BART Fine-Tuning: How to Derive a Contextualized Model for Tree-Based Algorithms?".

[15] Daya Guo, Yanan Wu, Yongkui Lai, Kang Liu, and Jun Zhao. "Meta-Learning with BART for Few-Shot Intent Detection and Slot Filling".

[16] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning* (pp. 11328-11339). PMLR.

[17] Shleifer, Sam, and Alexander M. Rush. "Pre-trained summarization distillation." *arXiv preprint arXiv:2010.13002* (2020).

[18] https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb

[19] https://replicate.com/rmokady/clip_prefix_caption

[20] https://wallpapers.com/wallpapers/pink-flowers-aesthetic-in-glass-vase-w8loqa8ztvk5ebk9.html

[21] https://www.wallpaperflare.com/reptiles-komodo-dragon-wildlife-wallpaper-gixno

[22] https://www.themoviedb.org/movie/34653-a-single-man

[23] https://www.tubics.com/blog/youtube-chapters