# Μηχανισμοί προσοχής στα νευρωνικά δίκτυα και πιθανές εφαρμογές

Ομάδα 4: Γεωργιάδης Στέφανος, Κύρου Θεολόγος Μέρος Α

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών ΔΠΜΣ Προηγμένα Συστήματα Υπολογιστών και Επικοινωνιών Υπολογιστική Νοημοσύνη - Συστήματα εμπνευσμένα από τη βιολογία Επιβλέπων Καθηγητές: Πετρίδης Βασίλειος, Διαμαντόπουλος Θεμιστοκλής

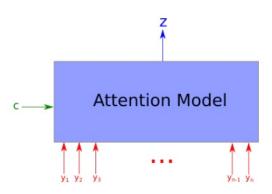
Περίληψη Σχοπός της παρούσας εργασίας, είναι να μελετήσει το θέμα της προσοχής στα νευρωνικά δίκτυα. Αρχικά, γίνεται περιληπτική θεωρητική εισαγωγή του μηχανισμού προσοχής και στη συνέχεια παρουσιάζονται οι κατηγορίες και οι τεχνικές/προσεγγίσεις που έχουν εφαρμοστεί εως τώρα σε μοντέλα νευρωνικών δικτύων. Τέλος, αναφέρονται ερευνητικά πεδία όπου εφαρμόζονται μηχανισμοί προσοχής βελτιώνοντας αισθητά την αποδοτικότητα του μοντέλου, καθώς και μελλοντικές επιστημονικές προκλήσεις και πιθανές εφαρμογές των μοντέλων προσοχής.

# 1 Θεωρητική εισαγωγή

Η βασιχή ιδέα πίσω από τους μηχανισμούς προσοχής (Attention Model) στα Νευρωνικά Δίκτυα πηγάζει από την γενικότερη έννοια του μηχανισμού προσοχής των ανθρώπων. Για παράδειγμα, το οπτικό μας σύστημα επεξεργασίας τείνει να εστιάζει επιλεκτικά σε ορισμένα μέρη της εικόνας, ενώ αγνοεί άλλες μη σχετικές πληροφορίες με τρόπο που μπορεί να βοηθήσει στην ικανότητα αντίληψης [Xu et al. 2015]. Ομοίως, σε αρκετά προβλήματα που αφορούν τη γλώσσα, την ομιλία ή την όραση, ορισμένα μέρη της εισόδου είναι πιο σημαντικά από άλλα. Για παράδειγμα, σε εργασίες μηχανικής μετάφρασης και περίληψης, μόνο ορισμένες λέξεις της ακολουθίας εισόδου μπορεί να είναι σχετικές για την πρόβλεψη της επόμενης λέξης. Η μοντελοποίηση με εισαγωγή μηχανισμού προσοχής είναι σχετικά πρόσφατη και γνωρίζει ραγδαία ανάπτυξη τα τελευταία χρόνια. Πέραν της χρησιμότητάς τους στην βελτίωση της απόδοσης των μοντέλων, έχουν χρησιμοποιηθεί εκτενώς για τη βελτίωση της ερμηνευσιμότητας των νευρωνικών δικτύων, τα οποία διαφορετικά θεωρούνται μοντέλα μαύρου κουτιού. Επίσης, βοηθούν στην υπέρβαση ορισμένων εως τώρα προκλήσεων με τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN).

Σε πρακτικό επίπεδο, η γενική ιδέα ενός μηχανισμού προσοχής είναι ότι παράλληλα με την εκπαίδευση του μοντέλου από τις εισόδους του δικτύου, δημιουργείται και ένα διάνυσμα προσοχής (attention vector) το οποίο δίνει τη δυνατότητα στο δίκτυο να εστιάσει σε συγκεκριμένες περιοχές της εισόδου, συνεισφέροντας έτσι στην ταχύτερη και πιο αποτελεσματική σύγκλιση στο επιθυμητό αποτέλεσμα. Αυτές οι περιοχές εστίασης προσαρμόζονται στις εισόδους, έτσι ώστε να υπάρχει

μεγαλύτερη σύνδεση με την έξοδο του μοντέλου. Ο τρόπος με τον οποίο ο μηχανισμός προσοχής επιχεντρώνεται/ξεχωρίζει στις περιοχές εστίασης γίνεται με τη χρήση βαρών (attention weights) επί των προηγούμενων χρυφών χαταστάσεων (source hidden states)  $\overrightarrow{h}_s$  συγχρατώντας έτσι σχετιχές πληροφορίες από την πλευρά της εισόδου για να βοηθήσει στην πρόβλεψη. Συνδυάζοντας λοιπόν τις πληροφορίες από την πλευρά της εισόδου με αυτές της τρέχουσας χατάστασης στόχου (target state)  $\overrightarrow{h}_t$  πετυχαίνουμε χαλύτερα ποσοστά πρόβλεψης.

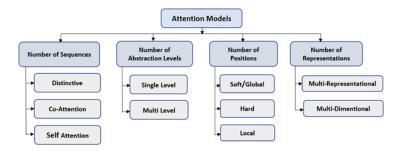


Σχήμα 1. Συνοπτική παρουσίαση της δομής ενός μηχανσιμου προοχής.

# 2 Κατηγορίες και τύποι μηχανισμών προσοχής

Υπάρχουν τέσσερις ευρύτερες κατηγορίες μοντέλων προσοχής. Κάθε κατηγορία εμπεριέχει διαφορετικούς τύπους μοντέλων προσοχής. Αν και οι κατηγορίες αυτές είναι αμοιβαία αποκλειόμενες, μπορούμε να εφαρμόσουμε μοντέλα προσοχής χρησιμοποιώντας έναν συνδυασμό διαφορετικών κατηγοριών. Ως εκ τούτου, οι κατηγορίες αυτές μπορούν επίσης να θεωρηθούν ως διαστάσεις.

Κάθε τύπος μηχανισμού προσοχής ορίζεται σαν μια συνάρτηση προσοχής η οποία, μπορεί να περιγραφεί ως μια αντιστοίχιση ενός ερωτήματος (query) και ενός συνόλου από ζεύγη κλειδιού-τιμής (key - value) σε μια έξοδο, όπου το ερώτημα, τα κλειδιά, οι τιμές και η έξοδος είναι όλα διανύσματα. Η έξοδος υπολογίζεται ως σταθμισμένο άθροισμα των τιμών, όπου το βάρος που αντιστοιχείται σε κάθε τιμή υπολογίζεται από μια συνάρτηση συμβατότητας του ερωτήματος με το αντίστοιχο κλειδί.



Σχήμα 2. Κατηγορίες και τύποι μοντέλων προσοχής.

#### 2.1 Αριθμός ακολουθιών

Σε αυτήν την κατηγορία ανήκουν τρεις τύποι προσοχής και καθορίζονται με βάση τον αριθμό των ακολουθιών που δέχονται ως είσοδο. Ο τύπος distinctive, χρησιμοποιείται όταν το κλειδί (key) και το ερώτημα (query) ανήκουν σε δύο διαφορετικές ακολουθίες εισόδου και εξόδου αντίστοιχα. Τα περισσότερα μοντέλα προσοχής που χρησιμοποιούνται για μετάφραση  $[Bahdanau\ et\ al.\ 2015]$ , λεζάντα εικόνας  $[Xu\ et\ al.\ 2015]$  και την αναγνώριση ομιλίας  $[Chan\ et\ al.\ 2016]$  εμπίπτουν στον συγκεκριμένο τύπο προσοχής

Ένα μοντέλο co-attention λειτουργεί σε πολλαπλές αχολουθίες εισόδου ταυτόχρονα και μαθαίνει από κοινού τα βάρη προσοχής των εισόδων (attention weights) , με σκοπό να καταγράψει τις αλληλεπιδράσεις μεταξύ αυτών των εισόδων. Οι εργασίες των [Lu et al. 2016] και [Yu et al. 2019] χρησιμοποιούν το εν λόγω μοντέλο σε VQA προβλήματα. Τέλος, στην περίπτωση που η είσοδος είναι μια αχολουθία αλλά η έξοδος δεν είναι μια αχολουθία, όπως ένα πρόβλημα ταξινόμησης ή ενα σύστημα συστάσεων (recommendation system) , χρησιμοποιούμε μηχανισμό αυτο-προσοχής (self attention).

#### 2.2 Αριθμός επιπέδων αφαίρεσης δεδομένων

Στην κατηγορία αυτή συναντούμμε δύο διαφορετικούς τύπους προσοχής. Όταν τα βάρη προσοχής υπολογίζονται μόνο για την αρχική ακολουθία εισόδου, ορίζεται ο μηχανισμός προσοχής ενός μόνο επιπέδου (single – level) Αντιθέτως, όταν ο μηχανισμός προσοχής εφαρμόζεται σε πολλαπλά επίπεδα αφαίρεσης της ακολουθίας εισόδου με διαδοχικό τρόπο, ορίζεται ο τύπος προσοχής πολλαπλών επιπέδων (multi – level). Τότε, η έξοδος (context vector) του χαμηλότερου επιπέδου αφαίρεσης γίνεται η κατάσταση ερωτήματος για το υψηλότερο επίπεδο αφαίρεσης. Επιπλέον, τα μοντέλα που χρησιμοποιούν προσοχή πολλαπλών επιπέδων μπορούν να ταξινομηθούν περαιτέρω με βάση την διαδικασία μέσω της οποίας μαθαίνονται

τα βάρη. Αν τα βάρη μαθαίνονται από πάνω προς τα κάτω [Zhao and Zhang 2018] (από το υψηλότερο επίπεδο αφαίρεσης στο χαμηλότερο επίπεδο) ή από κάτω προς τα πάνω [Yang et al. 2016].

# 2.3 Αριθμός των αρχικών τοποθετήσεων, στις ακολουθίες εισόδου

Στην τρίτη κατηγορία, οι διαφορετικοί τύποι πηγάζουν από τις αρχικές θέσεις της ακολουθίας εισόδου όπου υπολογίζεται η συνάρτηση προσοχής. Η προσοχή που εισάγεται από τους [Bahdanau et al. 2015] είναι γνωστή ως ήπια προσοχή. Η ήπια προσοχή χρησιμοποιεί έναν σταθμισμένο μέσο όρο όλων των κρυφών καταστάσεων της ακολουθίας εισόδου για τη δημιουργία του διανύσματος πλαισίου. Η χρήση της μεθόδου μαλακής στάθμισης καθιστά το νευρωνικό δίκτυο προσιτό σε αποδοτική μάθηση μέσω οπισθοδιάδοσης, αλλά έχει επίσης ως αποτέλεσμα τετραγωνικό υπολογιστικό κόστος. Για την αποφυγή μεγάλου υπολογιστικού κόστους η εργασία των [Xu et al. 2015] πρότεινει ένα μοντέλο σκληρής/αυστηρής προσοχής. Το διάνυσμα πλαισίου υπολογίζεται από στοχαστικά δειγματοληπτικές κρυφές καταστάσεις στην ακολουθία εισόδου. Αυτό επιτυγχάνεται με τη χρήση μίας 'multinoulli' κατανομής που παραμετροποιείται από τα βάρη της προσοχής. Το μοντέλο σκληρής προσοχής είναι επωφελές λόγω μειωμένου υπολογιστικού κόστους, αλλά η λήψη μιας σκληρής απόφασης σε κάθε θέση της εισόδου καθιστά το μοντέλο μη διαφοροποιήσιμο και δύσκολα βελτιστοποιήσιμο.

Στο πλαίσιο της εργασίας αυτόματης μετάφρασης, οι [Luong et al. 2015b] πρότειναν δύο άλλα μοντέλα προσοχής, τοπικό και παγκόσμιο. Το μοντέλο παγκόσμιας προσοχής είναι παρόμοιο με το μοντέλο ήπιας προσοχής. Το τοπικό μοντέλο προσοχής, από την άλλη πλευρά, είναι ένα υβρίδιο μεταξύ της ήπιας και της σκληρής προσοχής. Η βασική ιδέα είναι να εντοπίσουμε πρώτα ένα σημείο ή μια θέση προσοχής μέσα στην ακολουθία εισόδου και να επιλέξουμε ένα παράθυρο γύρω από αυτή τη θέση για τη δημιουργία ενός τοπικού μοντέλου ήπιας προσοχής. Το κύριο πλεονέκτημα της τοπικής προσοχής είναι ότι παρέχει μια παραμετρική αντιστάθμιση μεταξύ μαλακής και σκληρής προσοχής, υπολογιστικής απόδοσης και διαφοροποιησιμότητας εντός του παραθύρου.

#### 2.4 Αριθμός αναπαραστάσεων

Γενικά, στις περισσότερες εφαρμογές χρησιμοποιείται μια ενιαία αναπαράσταση χαρακτηριστικών της ακολουθίας εισόδου. Ωστόσο, σε μερικές περιπτώσεις, η χρήση μιας ενιαίας αναπαράστασης χαρακτηριστικών της εισόδου μπορεί να μην επαρκεί. Σε αυτές τις περιπτώσεις, μια προσέγγιση είναι η καταγραφή διαφορετικών πτυχών της εισόδου μέσω πολλαπλών αναπαραστάσεων χαρακτηριστικών. Η προσοχή μπορεί να χρησιμοποιηθεί για την ανάθεση βαρών σπουδαιότητας σε αυτές τις διαφορετικές αναπαραστάσεις, οι οποίες μπορούν να καθορίσουν τις πιο σχετικές πτυχές, αγνοώντας τον θόρυβο και περιττές πληροφορίες της εισόδου. Το μοντέλο που περιγράψαμε αναφέρεται ως πολυαναπαραστατική προσοχή(multi –

representational). Στη βάση της ίδιας λογικής, ορίζεται η πολυδιάστατη προσοχή. Τα βάρη εισάγονται για τον προσδιορισμό της συνάφειας των διάστασεων του διανύσματος ενσωμάτωσης. Ο υπολογισμός μιας βαθμολογίας για κάθε χαρακτηριστικό του διανύσματος μπορεί να επιλέξει τα χαρακτηριστικά που μπορούν να περιγράψουν καλύτερα τη συμβολική σημασία σε κάθε συγκεκριμένο νόημα. Αυτό είναι ιδιαίτερα χρήσιμο για εφαρμογές φυσικής γλώσσας όπου διάφορες λέξεις μπορούν να έχουν παρπαπάνω από μία έννοια ακόμα και στην ίδια πρόταση.

## 3 Εφαρμογές του μηχανισμού προσοχής και μελλοντικές προεκτάσεις

Τα μοντέλα προσοχής αποτελούν ένα σχετικά καινούριο πεδίο έρευνας που αναπτύσεται ραγδαία. Τα τελευταία χρόνια έχουν δημοσιευθεί δεκάδες ερευνητικές αναφορές, σχετικά με το θέμα της προσοχής και των εφαρμογών τους. Όπως αναφέραμε, η συνεισφορά των μηχανισμών προσοχής δεν αφορά μόνο την βελτίωση ήδη γνωστών νευρωνικών μοντέλων αλλά και την ικανότητα επεξηγηματικότητας περίπλοκων μοντέλων. Δεδομένου ότι οι μηχανισμοι προσοχής βρίσκουν εφαρμογή σε ένα μεγάλο έυρος πεδίων, θα περιγράψουμε τα μοντέλα προσοχής σε πέντε βασικούς τομείς, με ενδεικτικά παραδείγματα.

#### 3.1 Εφαρμογές σε γνωστά μοντέλα

Στον τομέα του NLP, η προσοχή εστιάζει στα πιο συσχετιζόμενα μέρη της αχολουθίας εισόδου, στην στοίχιση των αχολουθιών εισόδου και εξόδου και στη καταγραφή εξαρτήσεων μεγάλου εύρους για μεγαλύτερες αχολουθίες. Στην περίπτωση της μηχανικής μετάφρασης ο AM συνισφέρει στην υπέρβαση προβλημάτων που δημιουργούνται από την εν γενει διαφοροποίηση μεταξύ των γλώσσεων. Βοηθά, επίσης στην αποτελεσματικότερη στοίχηση υποχειμένου-ρήματος-ουσιαστικού που επίσης διαφέρει ανάλογα με την γλώσσα χαθώς και στην μετάφραση μεγάλων προτάσεων [Bahdanau et al. 2015]. Επιπλεόν, μηχανισμοί προσοχής χρησιμοποιούνται σε προβλήματα ερωτήσεων-απαντήσεων και σε εργασίες ανάλυσης συναισθήματος, δίνοντας βαρύτητα στις λέξεις που αφορούν συγχεχριμένα συναισθήματα.

Στον τομέα της οπτικής ικανότητας αντίληψης η οπτική προσοχή χρησιμοποιείται ώστε να επικεντρώνει σε ΄σημαντικές΄ περιοχές εντός της κύριας εικόνας καθώς και για να συλλάβει δομικές μακροσκελείς εξαρτήσεις μεταξύ κομματιών της εικόνας. Ο όρος της οπτικής προσοχής επινοήθηκε από τους  $[Mnih\ et\ al.\ 2014]$  για προβλήματα κατηγοριοποίησης εικόνων. Στη σχετική εργασία, η οπτική προσοχή χρησιμοποιείται για την ελαχιστοποίηση της υπολογιστικής πολυπλοκότητας του προτεινόμενου μοντέλου ανεξάρτητα από το μέγεθος της εικόνας εισόδου, σε σύγκριση με τα CNNs. Αξιοσημείωτη είναι επίσης η συνεισφορά μηχανισμών προσοχής σε προβλήματα ανίχνευσης αντικειμένων  $[Ba\ et\ al.\ 2014]$ , όπου γίνεται επεξεργασία εικόνας με διαδοχικό τρόπο, ώστε το μοντέλο να ανιχνέυει ένα αντικείμενο κάθε φορά και σε εργασίες δημιουργίας εικόνων με προσέγγιση βήμα προς βήμα σε αντίθεση με προηγούμενες αντίστοιχες μοντελοποιήσεις (μία σάρωση).

Η προσοχή έχει χρησιμοποιηθεί εκτενώς σε  $multi-modal\ applications$ , καθώς βοηθά στην κατανόηση μεταξύ διαφορετικών λειτουργειών. Ο σχολιασμός μίας ειχόνας ή ενός βίντεο ανήχει σε αυτήν την κατηγορία. Παρόμοια με την QA, εδώ η προσοχή εκτελεί τη λειτουργία της εύρεσης σχετικών τμημάτων της ειχόνας εισόδου [Xu et al. 2015] για την πρόβλεψη της επόμενης λέξης στη λεζάντα ή την εστίαση σε μιχρότερο υποσύνολο των χαρέ για την περιγραφή βίντεο [Yao et al. 2015]. Δύο πολύ σημαντικές λειτουργίες σε αυτήν την κατηγορία ειναι η ανγνώριση ομιλίας και η κατανόηση της ανθρώπινης δια ζώσης επικοινωνίας σε πραγματικό χρόνο [Zadeh et al. 2018]. Στη πρώτη περίπτωση η προσοχή αποτρέπει την υπεραιχπέδευση του μοντέλου [Chan et al. 2016]. Η αναγνώριση ομιλίας δεν έχει σαφή δομή, έχει πολλαπλά παρόμοια τμήματα ομιλίας και η είσοδος της είναι πολύ θορυβώδεις [Chorowski et al. 2015]. Η προσαρμογή ενός μηχανισμού προσοχής δρα αποτελεσματικά στα τρία παραπάνω προβλήματα. Στη δεύτερη περίπτωση, ένα δύσχολο πρόβλημα που περιλαμβάνει ταυτόχρονα γλωσσιχές, οπτικές και ομιλητικές λειτουργίες η προσοχή χρησιμοποιείται ανιχνέυοντας αλληλεπιδράσεις και συσχετίσεις μεταξύ των διαφορετικών λειτουργιών σε κάθε χρονικό βήμα (cross - view dynamics).

Στα συστήματα συστάσεων η προσοχή χρησιμέυει στη δημιουργία προφιλ/χρήστη. Η κεντρική ιδέα, είναι ότι ο ΜΑ κάνοντας αναδρομή στο ιστορικό του χρήστη καταγράφοντας τις αλληλεπιδράσεις του με τα άντικείμενα΄, τοποθετεί βάρη προσοχής στα αντίστοιχα αντικείμενα ώστε να καταγράψει τα ενδιαφέροντα του χρήστη για διαφορετικές χρονικές περιόδους καθώς και μεταβολές στις προτιμήσεις του. Τέλος, τα πραγματικά δεδομένα τις περισσότερες φορές περιέχουν πληροφορία σε μορφή γραφημάτων, διαγραμμάτων και δικτύων. Αυτά τα δεδομένα πηγάζουν από τα κοινωνικά δίκτυα, διαγράμματα γνώσεων και τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων. Η προσοχή έχει χρησιμοποιήθηκε για την ανάδειξη στοιχείων του γράφου (κόμβοι, ακμές, υπογράφοι) τα οποία είναι πιο σημαντικά για την την κύρια εργασία [Lee et al. 2019].

#### 3.2 Μελλοντικές πιθανές εφαρμογές

Ολοχληρώνοντας την παρούσα μελέτη, αναφέρουμε σύντομα ορισμένες εφαρμογές μηχανισμών προσοχής σε μοντέλα που βρίσκονται σε εξέλιξη.

Ένα μελλοντικό πεδίο εφαρμογής μηχανισμών προσοχής προτάθηκε από τους [Chiu and Raffel 2017] για να εφαρμοστεί σε ορισμένες εφαρμογές πραγματικού χρόνου, όπως ζωντανές λεζάντες βίντεο ή συνομιλίες μεταξύ ανθρώπων που μιλούν διαφορετικές γλώσσες. Σε αυτά τα μοντέλα χρειάζεται η διαδικασία της μετάφρασης να ξεκινήσει προτού ολοκληρώθεί η ανάγνωση ολόκληρης της αρχικής πρότασης. Η διαδικασία που ακολουθείται χωρίζει προσαρμοστικά την ακολουθία εισόδου σε μικρότερα κομμάτια στα οποία υπολογίζεται η ήπια προσοχή, επιτρέποντας έτσι την online και σε γραμμικό χρόνο αποκωδικοποίηση.

Ένα ακόμα ερώτημα προς διερέυνση είναι αν η προσοχή μπορεί να αποτελέσει ένα αυτοτελές εργαλείο για οπτικά μοντέλα προσοχής αντί να έχει συμπληρωματική λειτουργία στην αρχιτεκτονική του μοντέλου. Οι [Ramachandran et al. 2019] διερεύνησαν την αυτόνομη αυτοπροσοχή σε μοντέλα όρασης αντικαθιστώντας όλες

τις περιπτώσεις των χωριχών συνελίξεων με αυτοπροσοχή πάνω σε τοπιχές περιοχές. Μία αχόμη μελέτη που βρίσχεται σε εξέλιξη είναι η διερεύνηση της σχέσης μεταξύ των βαρών προσοχής και της ερμηνευσιμότητας του μοντέλου. Η μελλοντιχή έρευνα μπορεί να διερευνήσει τις χατανομές προσοχής των τρεχόντων μοντέλων και πώς μπορούν να τροποποιηθούν ώστε να προσφέρουν εύλογες αιτιολογήσεις της πρόβλεψης του μοντέλου.

Ο αυτοματοποιημένος σχεδιασμός αρχιτεχτονικών νευρωνικών διχτύων με τη χρήση της αναζήτησης νευρωνικής αρχιτεχτονικής (NAS) έχει ξεπεράσει τα ανθρώπινα σχέδια σε διάφορες εργασίες. Ένα ανοιχτό ερώτημα είναι αν μπορούμε να χρησιμοποιήσουμε τη NAS για την αναζήτηση της βέλτιστης αρχιτεχτονική της μονάδας προσοχής υψηλής τάξης. Η εργασία των  $[Ma\ et\ al.\ 2020]$  είναι η πρώτη προσπάθεια επέχτασης του NAS για την αναζήτηση plug-and-play μονάδων προσοχής πέρα από την αρχιτεχτονική χορμού. Τέλος, σε πρωτόλιο στάδιο βρίσχονται εργασίες που αφορούν προσοχή πολλαπλών περιπτώσεων, την χατανόηση χαι την μοντελοποίηση της συμπεριφοράς συστημάτων σε πολλές εφαρμογές του πραγματιχού χόσμου, όπως αυτόνομα οχήματα χαι θεωρία παιγνίων. Ένα σημαντιχό θέμα της έρευνας είναι η μείωση της τετραγωνιχής πολυπλοχότητας χρόνου χαι χώρου των μετασχηματιστών σε γραμμιχή, χωρίς απώλεια στην απόδοση των μοντέλων αυτών. Οι  $[Sukhbaatar\ et\ al.\ 2019b]$  προτείνουν μια προσέγγιση που χρησιμοποιεί μια μαθησιαχή συνάρτηση χάλυψης για τον δυναμιχό έλεγχο του εύρους προσοχής σε μηχανισμό αυτοπροσοχής.

## 4 Βιβλιογραφία

- 1.  $[Xu\ et\ al.\ 2015]$
- 2. [Bahdanau et al. 2015]
- 3. [Chan et al. 2016]
- 4. [Zhao and Zhang 2018]
- 5. [Yang et al. 2016]
- 6. [Luong et al. 2015b]
- 7. [Ramachandran et al. 2019]