ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΤΥΠΟΠΟΙΗΣΗ σε ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ και στη ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ»

Σύστημα Ερωτοαπαντήσεων Βάσει Οπτικού Περιεχομένου με Χρήση Τεχνικών Βαθιάς Μηχανικής Μάθησης

Κακογεωργίου Ιωάννης
ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ: 09316012

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ, Ιούνιος 2019

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΤΥΠΟΠΟΙΗΣΗ σε ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ και στη ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ»

Σύστημα Ερωτοαπαντήσεων Βάσει Οπτικού Περιεχομένου με Χρήση Τεχνικών Βαθιάς Μηχανικής Μάθησης

Κακογεωργίου Ιωάννης

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.
ΣΥΝΕΠΙΒΛΕΠΩΝ: Γεώργιος Σιόλας
ΕΔΙΠ Ε.Μ.Π.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
……	……	……
Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Σταύρος Περαντώνης Διευθυντής Έρευνας στο ΕΚΕΦΕ "Δημόκριτος"	Κωνσταντίνος Καράντζαλος Αν. Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ, Ιούνιος 2019

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με το πρόβλημα της αυτόματης απάντησης σε ερώτηση φυσικής γλώσσας που αναφέρεται στο περιεχόμενο μιας εικόνας (Visual Question Answering ή VQA). Είναι ένα πρόβλημα που εντοπίζεται στην τομή των επιστημονικών πεδίων της Όρασης Υπολογιστών (CV) και της Επεξεργασίας Φυσικής Γλώσσας (NLP), το οποίο στα πλαίσια της εργασίας προσεγγίζεται με τη χρήση Βαθιών Νευρωνικών Δικτύων.

Στην εργασία αυτή, υλοποιείται και παρουσιάζεται ένα Σύστημα Ερωτήσεων - Απαντήσεων Βάσει Οπτικού Περιεχομένου το οποίο βασίζεται σε Συνελικτικά Δίκτυα (CNN) και σε Ανατροφοδοτούμενα Δίκτυα (RNN). Συγκεκριμένα για την αναπαράσταση των εικόνων γίνεται χρήση των Συνελικτικών Δικτύων VGGNet-19 και DenseNet-161, ενώ για την αναπαράσταση των ερωτήσεων γίνεται αρχικά χρήση μεθόδων Εμφύτευσης των Λέξεων μέσω Πίνακα Εμφύτευσης και του Γλωσσικού Μοντέλου ELMo, οι οποίες έπειτα τροφοδοτούνται σε ένα LSTM για τη δημιουργία της αναπαράστασης των ερωτήσεων. Οι δύο αυτές αναπαραστάσεις συνδυάζονται μέσω Επιπέδων Πολλαπλής Εστίασης (Stacked Attention Networks) το οποία εντοπίζουν περιοχές της εικόνας που σχετίζονται με την ερώτηση. Με βάση τις περιοχές αυτές εξάγονται τελικά χαρακτηριστικά τα οποία τροφοδοτούνται σε ένα Πλήρως Συνδεδεμένο Επίπεδο το οποίο παράγει την τελική απάντηση.

Τα παραπάνω μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν στο σύνολο δεδομένων VQA v.2 και τα αποτελέσματα έδειξαν ότι το βέλτιστο μοντέλο που αποτελείται από το συνδυασμό πέντε μεμονωμένων μοντέλων (Ensemble Model) επιτυγχάνει αρκετά υψηλή απόδοση.

Λέξεις-κλειδιά: Μοντέλο Ερωτήσεων - Απαντήσεων, Οπτικό περιεχόμενο, Εξόρυξη Γνώσης, Επεξεργασία Φυσικής Γλώσσας, Μηχανική Μάθηση, Τεχνητά Νευρωνικά Δίκτυα, Νευρώνες Μακράς και Βραχείας Μνήμης, Ανατροφοδοτούμενα Νευρωνικά Δίκτυα, Βαθιά Νευρωνικά Δίκτυα, Εμφύτευση Λέξεων, Γλωσσικό Μοντέλο, ELMo, Επίπεδα Πολλαπλής Εστίασης, VGGNet, DenseNet

Abstract

This thesis tackles the problem of Visual Question Answering (VQA) where an algorithm is given as input an image and a natural language question and generates a natural language answer as the output. VQA lies at the intersection of the fields of Computer Vision and Natural Language Processing and has been historically considered a very challenging problem.

In this work, we adopt Deep Neural Networks (DNN) to address this problem. Specifically, we evaluate the performance of various DNN pipelines consisting of different architectures of Convolutional Neural Networks (CNN) and word representations. With respect to the task of feature map extraction from images, we evaluate the VGGNet-19 and DenseNet-161 CNN architectures. With respect to word representation, we evaluate the performance of Embedding Matrix and Language Model ELMo methods, the output of which is fed to an LSTM RNN network to produce the final question embeddings. Both representations are combined with Stacked Attention Networks that focus on image regions related to the question. Features within these regions are extracted and fed to a Fully Connected Layer that produces the final answer.

The aforementioned models are trained and validated using the VQA v.2 dataset. Results indicate that using an Ensemble of five models outperforms all evaluated single models and offers additional accuracy.

Keywords: Visual Question Answering, Data Mining, Natural Language Processing (NLP), Computer Vision (CV), Machine Learning, Artificial Neural Networks, Long short-term memory (LSTM), Recurrent Neural Network (RNN), Deep Neural Networks, Word Embeddings, Language Model, biLM, biLSTM, ELMo, Stacked Attention Networks (SANS), VGGNet, DenseNet, Highway Networks, T-distributed Stochastic Neighbor Embedding (t-SNE)

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, για την ευκαιρία που μου έδωσε να εκπονήσω την παρούσα διπλωματική. Ακολούθως, θα ήθελα να ευχαριστήσω τους κ. Σταυρό Περαντώνη και κ. Κωνσταντίνο Καράντζαλο για τις χρήσιμες συμβουλές τους. Η συνεργασία που είχα και με τους τρεις ξεχωριστά κατά την διάρκεια των μεταπτυχιακών μου σπουδών αποτέλεσαν το έναυσμα για την ενασχόληση μου με τον τομέα της Μηχανικής Μάθησης. Επίσης, οφείλω να ευχαριστήσω ιδιαιτέρως τον κ. Γιώργο Σιόλα για τη συνεχή καθοδήγηση και συνεισφορά του στην εκπόνηση της παρούσας εργασίας. Η άμεση ανταπόκριση και το ενδιαφέρον του έπαιξαν σημαντικό ρόλο έτσι ώστε να εξαχθεί το βέλτιστο δυνατό αποτέλεσμα. Στη συνέχεια, θα ήθελα να ευχαριστήσω τον κ. Αριστείδη Δούμα για την καθοδήγηση και τις συμβουλές του τα τελευταία 12 χρόνια. Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους γονείς μου, στην Κατερίνα μου και τους φίλους μου για την έμπρακτη υποστήριξή τους όλα αυτά τα χρόνια.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1. Εισαγωγή

1.1 Σύστημα Ερωτήσεων - Απαντήσεων Βάσει Οπτικού Περιεχομένου – VQA

Ο επιστημονικός κλάδος που βρίσκεται στην τομή των πεδίων της Όρασης Υπολογιστών (CV), της Επεξεργασίας Φυσικής Γλώσσας (NLP) και της Αναπαράστασης Οντολογικής Γνώσης και Συλλογιστικής (KR) έχει αναπτυχθεί ταχύτατα τα τελευταία χρόνια. Η ανάπτυξη του συγκεκριμένου πεδίου έχει φέρει στο προσκήνιο ένα πολύ μεγάλο πλήθος εξαιρετικών εφαρμογών. Δημοφιλές παράδειγμα αποτελεί η αυτόματη περιγραφή του περιεχομένου μιας εικόνας (Image Captioning) το οποίο σχετίζεται με την ικανότητα ενός συστήματος να μπορεί, αυτόματα, να δημιουργεί περιγραφές του περιεχομένου των εικόνων γεννώντας σωστές συντακτικά και σημασιολογικά προτάσεις (Mao et al. 2014) (Kiros et al. 2014) (Fang et al., 2015) (Chen et al., 2015) (Donahue et al., 2015). Προβλήματα σαν και αυτό αποτελούν μία εξαιρετικά απαιτητική πρόκληση και ανοίγουν τον δρόμο προς πιο ολοκληρωμένες λύσεις τεχνικής νοημοσύνης.
Ένα άλλο πρόβλημα αρκετά συγγενικό με την αυτόματη περιγραφή του περιεχομένου μιας εικόνας είναι η αυτόματη απάντηση σε ερωτήσεις φυσικής γλώσσας που αναφέρεται στο περιεχόμενο μιας εικόνας (Visual Question Answering ή VQA). Στόχος ενός τέτοιου συστήματος το οποίο δέχεται μία εικόνα και μία οποιαδήποτε ερώτηση σε φυσική γλώσσα που σχετίζεται με την εικόνα, είναι να παράξει μία σωστή απάντηση σε φυσική γλώσσα σαν αποτέλεσμα. Η πρόοδος σε ένα τέτοιο πεδίο ενισχύει σημαντικά τη δυνατότητα δημιουργίας συστημάτων προς ολοκληρωμένες λύσεις τεχνητής νοημοσύνης. Συγκεκριμένες εφαρμογές επίσης μπορούν να προκύψουν σε περιπτώσεις όπου άτομα με προβλήματα όρασης εξάγουν μια οπτική πληροφορία αλληλεπιδρώντας με ένα τέτοιο σύστημα (Antol et al., 2015).

1.2 Σχετικές Προσεγγίσεις

Το πρόβλημα VQA όπως αναφέρθηκε σχετίζεται άμεσα με το πρόβλημα της αυτόματης περιγραφής του περιεχομένου μιας εικόνας (Image Captioning). Παράδειγμα μιας προσέγγισης του προβλήματος Image Captioning αποτελεί εκείνη των (Vinyals et al., 2014), όπου το σύστημα αρχικά εξάγει τα χαρακτηριστικά της εικόνας μέσω ενός Βαθιού Συνελικτικού Δικτύου GoogleNet, τα οποία έπειτα τροφοδοτούνται σε ένα δίκτυο Νευρώνων Μακράς-Βραχείας Μνήμης (LSTM) για την παραγωγή της περιγραφής. Άλλη προσέγγιση αποτελεί η μέθοδος που προτάθηκε από τους (Xu et al., 2015) και πήγε ένα βήμα παραπέρα χρησιμοποιώντας ένα μηχανισμό Εστίασης (Attention Layers) πάνω στις εικόνες κατά τη διαδικασία παραγωγής των περιγραφών.
Στην περίπτωση του VQA το πρόβλημα εστιάζεται στην κατανόηση της συσχέτισης μεταξύ της οπτικής πληροφορίας και της ερώτησης ώστε να παραχθεί η απάντηση. Αρκετά διαφορετικά

μοντέλα έχουν προταθεί για τη μοντελοποίηση και δημιουργία ενός αποτελεσματικού VQA συστήματος όπου τα περισσότερα βασίζονται σε νευρωνικά δίκτυα (Gao et al., 2015) (Ren et al., 2015) (Malinowski et al., 2015) (Antol et al., 2015). Τα περισσότερα από αυτά ακολουθούν παρόμοια λογική με εκείνα της αυτόματης περιγραφής του περιεχόμενου μιας εικόνας. Η πιο συνήθης πρακτική προσέγγιση του προβλήματος αποτελείται από την εξαγωγή χαρακτηριστικών στο σύνολο της εικόνας μέσω κάποιου Βαθιού Συνελικτικού Δικτύου (CNN), την αναπαράσταση της ερώτησης ως ένα διάνυσμα χαρακτηριστικών μέσω κάποιου δικτύου Νευρώνων Μακράς-Βραχείας Μνήμης (LSTM) και το συνδυασμό των δύο αναπαραστάσεων με σκοπό την τελική εκτίμηση της απάντησης. Οι (Gao et al., 2015) και (Malinowski et al., 2015) χρησιμοποίησαν τη λογική του Κωδικοποιητή - Αποκωδικοποιητή (*LSTM Encoder Decoder*) ώστε δοθείσας της ερώτησης και της εικόνας να παράξουν μία μεταβλητού μεγέθους απάντηση. Έτσι σε κάθε βήμα εκτέλεσης της Κωδικοποίησης (encoding) το LSTM δίκτυο τροφοδοτούταν με τα χαρακτηριστικά που έχουν εξαχθεί από την εικόνα μαζί με την αναπαράσταση της κάθε λέξης της ερώτησης. Οι (Ren et al., 2015) πρότειναν μια ποικιλία από νευρωνικά μοντέλα τα οποία χρησιμοποιούσαν είτε μονής είτε διπλής κατεύθυνσης LSTM (biLSTM). Ωστόσο, παρατήρησαν ότι η απλή σειριακή συνένωση των χαρακτηριστικών της εικόνας και της ερώτησης εμφάνισε τα καλύτερα αποτελέσματα. Μία παρόμοια προσέγγιση αποτελεί το μοντέλο που πρότειναν οι (Antol et al., 2015) όπου στην περίπτωση αυτή ο συνδυασμός των χαρακτηριστικών επιτυγχάνεται μέσω του κατά σημείο πολλαπλασιασμού. Πιο σύγχρονες τεχνικές εστίασαν την προσοχή τους στην εξαγωγή χαρακτηριστικών από περιοχές της εικόνας που σχετίζονται με την ερώτηση. Οι (Yang et al., 2016) χρησιμοποιούν τα χαρακτηριστικά της ερώτησης ώστε να παράξουν μια κατανομή πιθανοτήτων (Εστίαση) πάνω στην εικόνα μέσω Πολλαπλών Επιπέδων Εστίασης (Stacked Attention Networks). Μέσω της κατανομής αυτής εξάγονται έπειτα χαρακτηριστικά της εικόνας που σχετίζονται με τις περιοχές μεγάλης πιθανότητας. Μια πρόσφατη προσέγγιση είναι των (Lu et al., 2016) όπου πέραν της εστίασης στην εικόνα από τα χαρακτηριστικά της ερώτησης, προτείνουν και την αντίθετη κατεύθυνση, δηλαδή εστίαση σε λέξεις ή φράσεις της ερώτησης από τα χαρακτηριστικά της εικόνας και τη διαδικασία αυτή την ονομάζουν Co-Attention. Μία διαφορετική προσέγγιση για τις περιοχές εστίασης της εικόνας πρότειναν οι (Anderson et al. 2018) όπου αντί η εικόνα να χωρίζεται σε πλέγμα και πάνω σε αυτό να υπολογίζονται οι πιθανότητες εστίασης, χρησιμοποιείται αρχικά ένα Faster R-CNN (Ren et al., 2015) για την ανίχνευση αντικειμένων πάνω στην εικόνα και οι πιθανότητες εστίασης παράγονται πάνω στα χαρακτηριστικά των περιοχών που ορίζουν τα αντικείμενα αυτά. Άλλες τεχνικές όπως των (Fukui et al., 2016) εστιάζουν στον τρόπο συγχώνευσης των αναπαραστάσεων της ερώτησης και της εικόνας και προτείνουν αντί της απλής σειριακής επέκτασης ή της κατά σημείο πρόσθεσης/πολλαπλασιασμού να χρησιμοποιείται κάποιος Διγραμμικός μετασχηματισμός. Έτσι παρατήρησαν ότι μπορεί να επιτευχθεί μια πιο αντιπροσωπευτική αναπαράσταση της συγχώνευσης των δύο αναπαραστάσεων όπου δύναται να μοντελοποιήσει καλύτερα τις αλληλεπιδράσεις μεταξύ τους.

Για τη διερεύνηση της συσχέτισης μεταξύ της οπτικής πληροφορίας και της ερώτησης ώστε να παραχθεί η απάντηση και κατασκευή κατάλληλων μοντέλων, αρκετά σύνολα δεδομένων δημιουργήθηκαν είτε με αυτόματη παραγωγή ερωτήσεων μέσω περιγραφών εικόνας είτε μέσω

ανθρώπινης επισήμανσης (Maliniwski et al., 2014) (Ren et al., 2015) (Antol et al., 2015) (Gao et al., 2015) (Zhu et al., 2016) (Krishna et al., 2017). Ένα από τα πιο δημοφιλή σύνολα δεδομένων που σχετίζεται με το συγκεκριμένο πρόβλημα αποτελεί το VQA v1 και v2 το οποίο βασίζεται στο σύνολο εικόνων MS COCO (Antol et al., 2015) (Goyal et al., 2017). Οι δημιουργοί του συγκεκριμένου συνόλου δεδομένων διοργανώνουν κάθε χρόνο έναν διαγωνισμό που ονομάζεται VQA Challenge με σκοπό τη δημιουργία καλύτερων μοντέλων πάνω στο VQA v2 σύνολο δεδομένων. Σημειώνεται ότι το μοντέλο που προτείνεται από τους (Antol et al., 2015) αποτελεί για το διαγωνισμό VQA Challenge το μοντέλο σημείου αναφοράς (Baseline) για τη σύγκριση της απόδοσης του σε σχέση με άλλα.

1.3 Αντικείμενο Διπλωματικής

Στα περισσότερα VQA συστήματα οι προσεγγίζεις που προτείνονται ακολουθούν τη γενικότερη δομή:

- Αναπαράσταση της ερώτησης ως ένα διάνυσμα χαρακτηριστικών μέσω κάποιου δικτύου Νευρώνων Μακράς-Βραχείας Μνήμης (LSTM).
- Εξαγωγή χαρακτηριστικών στο σύνολο της εικόνας μέσω κάποιου Βαθιού Συνελικτικού Δικτύου (CNN).
- Συνδυασμό των δύο αναπαραστάσεων με σκοπό την τελική εκτίμηση της απάντησης.

Η προσοχή στις διαφορετικές προσεγγίσεις εστιάζεται στα τρία αυτά κομβικά σημεία. Δηλαδή, προτείνονται εναλλακτικές στον τρόπο αναπαράστασης της ερώτησης όπως με κάποιο CNN ή bi-LSTM αντί του πιο κλασικού LSTM, εναλλακτικές στον τρόπο εξαγωγής των χαρακτηριστικών όπως με τον εντοπισμό περιοχών εντός της εικόνας που σχετίζονται με την ερώτηση και εναλλακτικές στον τρόπο συγχώνευσης των δύο αυτών αναπαραστάσεων ώστε να μοντελοποιεί πιο αποτελεσματικά τις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών.
Κεντρικός στόχος της παρούσας εργασίας είναι η υλοποίηση ενός Συστήματος Ερωτήσεων - Απαντήσεων Βάσει Οπτικού Περιεχομένου όπως αυτά που προτείνονται στις σχετικές προσεγγίσεις. Η βασική δομή του συστήματος που ακολουθείται είναι αυτή που περιγράφεται στους (Antol et al., 2015) και αξιοποιείται το σύνολο δεδομένων VQA v2. Δηλαδή χρησιμοποιείται ένα Συνελικτικό Δίκτυο CNN για την εξαγωγή του χάρτη χαρακτηριστικών της εικόνας και ένα LSTM για την εξαγωγή των χαρακτηριστικών της ερώτησης.

Για την εξαγωγή των χαρτών χαρακτηριστικών των εικόνων γίνεται πειραματισμός με τέσσερις εναλλακτικές:

- VGGNet-19 με ανάλυση εικόνας εισόδου 224x224 (Simonyan et al., 2015)
- VGGNet-19 με ανάλυση εικόνας εισόδου 448x448 (Simonyan et al., 2015)
- DenseNet-161 με ανάλυση εικόνας εισόδου 224x224 (Huang et al., 2017)
- DenseNet-161 με ανάλυση εικόνας εισόδου 448x448 (Huang et al., 2017)

Για τη δημιουργία των αναπαραστάσεων των λέξεων των ερωτήσεων πριν αυτά τροφοδοτηθούν στο LSTM, χρησιμοποιήθηκαν οι προσεγγίσεις:

- Απλή εμφύτευση των one-hot διανυσμάτων λέξεων σε ένα χώρο μικρότερης διάστασης.
- Αξιοποίηση του υπερσύγχρονου Γλωσσικού Μοντέλου ELMo (Peters et al., 2018) το οποίο δημιουργεί αναπαραστάσεις των λέξεων κάνοντας χρήση ολόκληρης της ερώτησης. Στόχος του είναι να δημιουργηθεί μία αναπαράσταση που πέραν της ίδιας της λέξης να αξιοποιεί γενικότερα τον τρόπο που χρησιμοποιήθηκε στη συγκεκριμένη ερώτηση.

Για το συνδυασμό των δύο αναπαραστάσεων γίνεται χρήση των Πολλαπλών Επιπέδων Εστίασης (Stacked Attention Networks) όπως προτείνονται από τους (Yang et al., 2016), τα οποία εντοπίζουν περιοχές που συσχετίζονται με την ερώτηση πάνω στο χάρτη των χαρακτηριστικών της εικόνας. Η συγκεκριμένη προσέγγιση με Επίπεδα Εστίασης αποτελεί τη βασική αφετηρία όλων των σύγχρονων προσεγγίσεων που αφορούν το VQA πρόβλημα και για αυτό υιοθετείται και στην παρούσα εργασία. Η υλοποίηση των (Yang et al., 2016) δεν εμφάνισε στο διαγωνισμό VQA Challenge εξαιρετικά βελτιωμένα αποτελέσματα, όμως η υιοθέτηση της από τις μετέπειτα βέλτιστες προσεγγίσεις καθιστούν εξαιρετικά σημαντική την περαιτέρω ανάλυση και διερεύνηση της.
Για την εκτίμηση της τελικής απάντησης το πρόβλημα προσεγγίζεται σαν πρόβλημα Πολλαπλής Ταξινόμησης και όχι σαν Αποκωδικοποίηση (Decoding) όπου παράγεται μία μεταβλητού μεγέθους απάντηση. Η προσέγγιση αυτή αποτελεί την κύρια προσέγγιση για όλες τις υλοποιήσεις που κάνουν χρήση του συνόλου δεδομένων VQA v2. Ο βασικός λόγος σχετίζεται με το γεγονός ότι το 98% των απαντήσεων δεν υπερβαίνει τις τρεις λέξεις καθώς και ότι οι 1000 απαντήσεις με τη μεγαλύτερη συχνότητα στο σύνολο εκπαίδευσης εμφανίζονται σε ποσοστό 87,47% των ερωτήσεων.
Επίσης στην παρούσα εργασία εκτελούνται διαγνωστικοί έλεγχοι για την κατανόηση της λειτουργίας του συστήματος και γίνεται ανάλυση των σφαλμάτων. Στην προσπάθεια αυτή γίνεται ερμηνεία των αναπαραστάσεων των ερωτήσεων μέσω τεχνικών μείωσης της διάστασης των χαρακτηριστικών τους και παρουσιάζονται οι κατανομές πιθανοτήτων που δημιουργούνται από τα Επίπεδα Εστίασης πάνω στις εικόνες για διαφορετικές ερωτήσεις.
Ακόμα γίνεται έλεγχος της απόδοσης του βέλτιστου συστήματος που υλοποιήθηκε πάνω στο ‘κρυφό’ σύνολο ελέγχου του διαγωνισμού το οποίο αποτελεί τη βασική σύγκριση σε σχέση με άλλα μοντέλα. Η αξιολόγηση αυτή επιτυγχάνεται μέσω ενός εξυπηρετητή (Server) που παρέχουν οι δημιουργοί των δεδομένων.
Τέλος για την υλοποίηση αξιοποιήθηκε η πλατφόρμα Pytorch (Paszke et al. 2017) και από όσο γνωρίζουμε κατά τη διάρκεια της υλοποίησης δεν υπάρχει αντίστοιχη υλοποίηση του μοντέλου Πολλαπλών Επιπέδων Εστίασης σε Pytorch που να παρουσιάζει ισοδύναμη απόδοση με εκείνη που παρουσιάζεται στο (Yang et al., 2016).
Επίσης σημειώνεται ότι πριν την παρούσα εργασία δεν είχε διερευνηθεί η προσέγγιση της αναπαράστασης των λέξεων των ερωτήσεων μέσω του ELMo και το γεγονός αν ενισχύει την απόδοση ενός VQA συστήματος.

2. Θεωρητικό Υπόβαθρο

2.1 Νευρωνικά Δίκτυα

Στο κεφάλαιο αυτό παρουσιάζονται ορισμοί και περιγραφές των αλγορίθμων που χρησιμοποιήθηκαν στην παρούσα εργασία.

2.1.1 Τεχνητή Νοημοσύνη

Ο όρος τεχνητή νοημοσύνη αναφέρεται στον κλάδο της πληροφορικής ο οποίος ασχολείται με τη σχεδίαση και την υλοποίηση υπολογιστικών συστημάτων που μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς τα οποία υπονοούν έστω και στοιχειώδη ευφυΐα: μάθηση, προσαρμοστικότητα, εξαγωγή συμπερασμάτων, κατανόηση από συμφραζόμενα και επίλυση προβλημάτων. Η τεχνητή νοημοσύνη συστηματοποιώντας και αυτοματοποιώντας τις διανοητικές αυτές εργασίες μπορεί να έχει εφαρμογή σε ολόκληρο το φάσμα της ανθρώπινης διανοητικής δραστηριότητας.
Η σύγχρονη τεχνητή νοημοσύνη αποτελεί ένα από τα πλέον «μαθηματικοποιημένα» και ταχέως εξελισσόμενα πεδία της πληροφορικής. Σήμερα, ο τομέας αξιοποιεί περισσότερο εργαλεία που προέρχονται από τα εφαρμοσμένα μαθηματικά και τις επιστήμες μηχανικών και λιγότερο από τη θεωρητική πληροφορική και τη μαθηματική λογική όπως συνέβαινε πριν το 1990 (Russel et al., 1995).

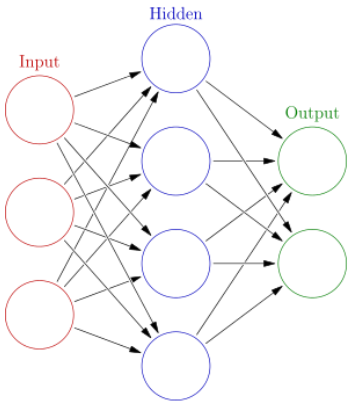
2.1.2 Μηχανική Μάθηση

Ένα υποπεδίο της τεχνητής νοημοσύνης που τα τελευταία χρόνια γνωρίζει μεγάλη ανάπτυξη είναι αυτό της μηχανικής μάθησης. Ως μηχανική μάθηση, ορίζεται το πεδίο εκείνο που παρέχει τη δυνατότητα στους υπολογιστές να “μαθαίνουν” κατασκευάζοντας μοντέλα από δεδομένα χωρίς να προγραμματίζονται ρητά. Πρόκειται για ένα επιστημονικό πεδίο προερχόμενο από το συνδυασμό της αναγνώρισης προτύπων και της υπολογιστικής μάθησης στην τεχνητή νοημοσύνη, που ερευνά τη μελέτη και τη δημιουργία αλγορίθμων, οι οποίοι είναι σε θέση να “μαθαίνουν” και να κάνουν προβλέψεις βάσει δεδομένων, ξεπερνώντας, ουσιαστικά, τον ρητό, στατικό προγραμματισμό (Simon 2013).

2.1.3 Μοντέλο του Νευρωνικού Δικτύου

Ένας αλγόριθμος μηχανικής μάθησης που εμπνέεται από τη δομή και τις λειτουργικές πτυχές των βιολογικών νευρωνικών δικτύων αποτελεί το τεχνητό Νευρωνικό Δίκτυο (ΝΝ) το οποίο είναι ένα κύκλωμα διασυνδεδεμένων νευρώνων.

Οι νευρώνες είναι τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος νευρώνας-κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές, επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου.



Σχήμα 2.1: Παράδειγμα Νευρωνικού Δικτύου [1].

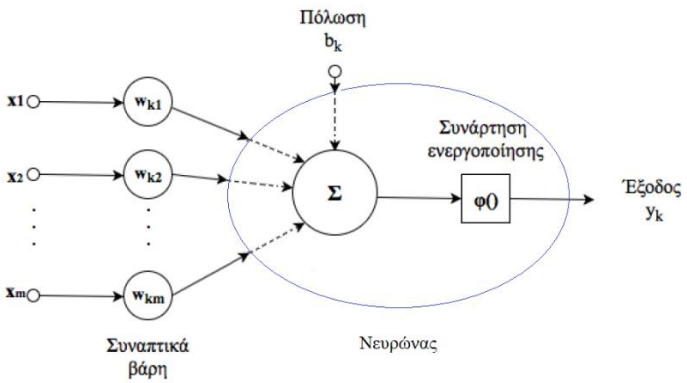
Ένα απλό νευρωνικό δίκτυο αποτελείται από τρία επίπεδα νευρώνων:

- το επίπεδο με τους νευρώνες εισόδου
- το επίπεδο με τους νευρώνες εξόδου
- το επίπεδο με τους υπολογιστικούς νευρώνες ή κρυμμένους νευρώνες

Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Συνεπώς σε ένα απλό νευρωνικό δίκτυο υπάρχουν δύο επίπεδα νευρώνων που συντελούν στις υπολογιστικές εργασίες. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό ονομάζεται Δυναμικό του νευρώνα. Το Δυναμικό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη.

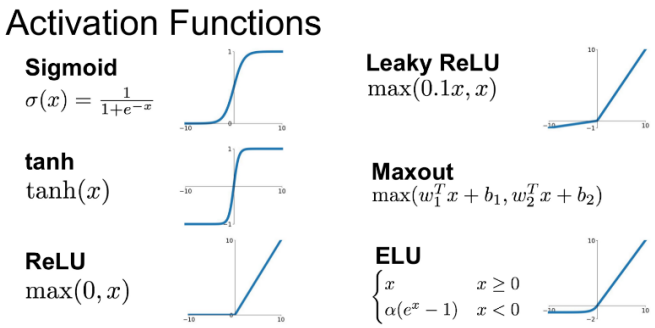
Συγκεκριμένα, εάν η $x_{ki} \in \mathbb{R}$ είναι η i-οστή είσοδος του k νευρώνα, $w_{ki} \in \mathbb{R}$ το i-οστό συναπτικό βάρος του k νευρώνα και Φ η συνάρτηση ενεργοποίησης του νευρωνικού δικτύου, τότε η έξοδος y_kτου k νευρώνα δίνεται από την εξίσωση:

$$y_k = \Phi\left(\sum_{i=0}^{N} x_{ki} w_{ki}\right) \in \mathbb{R} \qquad (2.1)$$



Σχήμα 2.2: Μοντέλο Νευρώνα. (Haykin, 2009 Μετά από επεξεργασία)

Στον k-οστό νευρώνα υπάρχει ένα συναπτικό βάρος w_{k0} με ιδιαίτερη σημασία, το οποίο καλείται **πόλωση** ή **κατώφλι** (*bias,threshold*). Η τιμή της εισόδου του είναι πάντα η μονάδα, $x_{k0} = 1$. Εάν το συνολικό άθροισμα από τις υπόλοιπες εισόδους του νευρώνα είναι μεγαλύτερο από την τιμή αυτή, τότε ο νευρώνας ενεργοποιείται. Εάν είναι μικρότερο, τότε ο νευρώνας παραμένει ανενεργός. Κάποιες από τις πιο συνηθισμένες συναρτήσεις ενεργοποίησης εμφανίζονται στο επόμενο Σχήμα 2.3.



Σχήμα 2.3: Παραδοσιακές Συναρτήσεις Ενεργοποίησης [2].

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η εγγενής ικανότητα μάθησης. Δηλαδή η ικανότητα του δικτύου να επιλύει κάποιο πρόβλημα (π.χ. η σταδιακή προσέγγιση μίας συνάρτησης). Η μάθηση επιτυγχάνεται μέσω της εκπαίδευσης, μίας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου (των βαρών και της πόλωσής του) σε τιμές κατάλληλες ώστε να επιλύεται με επαρκή επιτυχία το προς εξέταση πρόβλημα.
Μια διαφορετική αναπαράσταση ενός νευρωνικού δικτύου είναι μέσω της άλγεβρας πινάκων. Θέτοντας ως W∈ $\mathbb{R}^{NxM}$ για το σύνολο των βαρών των συνάψεων προς ένα επίπεδο νευρώνων,

όπου N είναι οι είσοδοι στο επίπεδο, M το πλήθος των νευρώνων και $b \in \mathbb{R}^M$ οι τιμές κατωφλίου του κάθε νευρώνα, τότε η έξοδος του επιπέδου μπορεί να αναπαρασταθεί και ως:

$$y = \Phi(x \cdot W^T + b) \in \mathbb{R}^M \quad (2.2)$$

Η αναπαράσταση αυτή ενός απλού επιπέδου νευρωνικού δικτύου ονομάζεται Πλήρως Συνδεδεμένο Επίπεδο (Fully-Connected Layer) και αποτελεί δομικό στοιχείο των περισσότερων νευρωνικών δικτύων. Τα επίπεδα αυτά πέραν της χρήσης τους στα απλά νευρωνικά δίκτυα που περιγράφηκαν παραπάνω, εφαρμόζονται και στα περισσότερα βαθιά νευρωνικά δίκτυα που θα παρουσιαστούν στη συνέχεια (Haykin 1999).

2.1.4 Βαθιά Νευρωνικά Δίκτυα

Ένα βαθύ νευρωνικό δίκτυο είναι ένα νευρωνικό δίκτυο με περισσότερα από δύο επίπεδα το οποίο χρησιμοποιεί αφηρημένα μαθηματικά μοντέλα. Αποτελεί κομμάτι μίας ευρύτερης οικογένειας μεθόδων μηχανικής μάθησης βασιζόμενων στην αναπαράσταση δεδομένων, σε αντίθεση με αλγορίθμους επικεντρωμένους σε υπολογιστικές εργασίες.
Τα βαθιά νευρωνικά δίκτυα απαιτούν πολλούς υπολογιστικούς πόρους κατά την εκπαίδευση τους. Για το λόγο αυτό για την εκπαίδευση τους αξιοποιείται η υπολογιστική δύναμη των καρτών γραφικών (GPU). Η εξέλιξη και βελτίωση των GPU οδήγησε στην περαιτέρω ανάπτυξη της Βαθιάς Μάθησης. Έτσι οι αρχιτεκτονικές βαθιάς μηχανικής μάθησης και νευρωνικών δικτύων έχουν πλέον εφαρμογή σε πολλά πεδία της τεχνητής νοημοσύνης, όπως η όραση των υπολογιστών, η αναγνώριση της φωνής και κατανόηση της φυσικής γλώσσας (Bengio 2009).

2.1.5 Επιβλεπόμενη Μάθηση Τεχνητών Νευρωνικών Δικτύων

Στην επιβλεπόμενη μάθηση (supervised learning), διατίθεται ένα σύνολο δεδομένων (dataset), το οποίο αποτελείται από παρατηρήσεις ή παραδείγματα (training examples), για κάθε ένα από τα οποία είναι γνωστό το σωστό αποτέλεσμα. Σε αυτή την περίπτωση λέγεται ότι το σύνολο δεδομένων είναι επισημασμένο (labeled). Στόχος αυτού του είδους μάθησης, είναι να ανακαλυφθεί μία σχέση ανάμεσα στις παρατηρήσεις και στα αποτελέσματα. Η σχέση η οποία ανακαλύπτεται είναι ένα μοντέλο τους προβλήματος. Το μοντέλο αυτό μπορεί να αποτελεί ένα νευρωνικό δίκτυο. Έτσι ένα εκπαιδευμένο νευρωνικό δίκτυο είναι το αποτέλεσμα της μάθησης και το χρησιμοποιούμε για να κάνουμε προβλέψεις σε νέες, άγνωστες παρατηρήσεις (Mohri 2012).
Ένα είδος επιβλεπόμενης μάθησης αποτελεί η Ταξινόμηση (Classification). Η ταξινόμηση αφορά τα προβλήματα στα οποία οι παρατηρήσεις κατηγοριοποιούνται σε ένα σύνολο προεπιλεγμένων διακριτών τιμών ή κλάσεων. Στόχος είναι να εκπαιδευτεί ένα σύστημα μηχανικής μάθησης, το οποίο παρατηρώντας τα επισημασμένα δεδομένα που του παρέχονται,

να ανακαλύψει μία σχέση, ανάμεσα στις παρατηρήσεις και τις κλάσεις στις οποίες ανήκουν, ώστε να είναι σε θέση να προβλέψει την κλάση νέων μελλοντικών παρατηρήσεων.
Συγκεκριμένα για ένα σύνολο από N δεδομένα που αποτελούνται από παρατηρήσεις $x_i \in \mathbb{R}^F$, και τις αντίστοιχες κλάσεις τους c_i , όπου $i = 1, 2, .., N$ και $c_i \in \{1,2, \dots k\}$ στόχος είναι η εύρεση μιας συνάρτησης f (ένα νευρωνικό δίκτυο) η οποία θα απεικονίζει παρατηρήσεις σε κλάσεις, δηλαδή:

$$f_W: \mathbb{R}^F \rightarrow \{1,2, \dots k\}$$

όπου W είναι τα βάρη ή οι παράμετροι της συνάρτησης (νευρωνικό δίκτυο) (Alpaydin 2010).
Μέσω της διαδικασίας της εκπαίδευσης, γίνεται προσαρμογή των παραμέτρων του νευρωνικού δικτύου ώστε να "ταιριάξει" στις παρατηρήσεις που διαθέτουμε. Για τη διαμόρφωση των παραμέτρων του δικτύου, απαιτούνται δύο διαδικασίες:

- <u>Συνάρτηση κόστους</u>
  Η συνάρτηση κόστους (loss function) ή αντικειμενική συνάρτηση (objective function) αξιολογεί το πόσο καλά το μοντέλο ταιριάζει στα δεδομένα. Στόχος είναι η εύρεση των παραμέτρων του μοντέλου που θα ελαχιστοποιήσουν τη συνάρτηση κόστους. Συνεπώς οι συναρτήσεις κόστους χρησιμοποιούνται ώστε η αποδοτική εκπαίδευση του νευρωνικού δικτύου να ανάγεται σε ένα πρόβλημα βελτιστοποίησης.
  Υπάρχουν πολλές συναρτήσεις κόστους που χρησιμοποιούνται στα προβλήματα βελτιστοποίησης. Σε προβλήματα κατηγοριοποίησης (classification) συνήθως χρησιμοποιείται η συνάρτηση διασχιζόμενης εντροπίας (cross entropy) (Goodfellow 2016).

$$J(W, b) = -\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij} \cdot \log\left(\widehat{y_{ij}}\right) \qquad (2.3)$$

  Όπου:
  - J(W, b): συνάρτηση κόστους
  - W: βάρη του νευρωνικού δικτύου
  - b: οι τιμές κατωφλίων του νευρωνικού δικτύου (biases)
  - N: πλήθος δεδομένων
  - M: πλήθος εξόδων του δικτύου (Διαφορετικές κλάσεις)
  - $\widehat{y_{ij}}$: έξοδος του νευρωνικού δικτύου
  - y: πραγματική έξοδος - στόχος (ground truth)

- <u>Ενημέρωση των παραμέτρων του δικτύου</u>
  Χρησιμοποιώντας τη συνάρτηση κόστους, εφαρμόζεται ένας μηχανισμός ο οποίος ενημερώνει το νευρωνικό δίκτυο, αυξομειώνοντας τα βάρη της f_W. Μια πιθανή προσέγγιση στο πρόβλημα αυτό θα ήταν να επιλέγουμε τυχαία *W* έως ότου βρούμε το W εκείνο για το οποίο ελαχιστοποιείται η J. Ωστόσο, ένα τυπικό νευρωνικό δίκτυο που θέλουμε να εκπαιδεύσουμε μπορεί να έχει διανύσματα παραμέτρων με εκατομμύρια ή δισεκατομμύρια παραμέτρους και έτσι μέθοδοι σαν και αυτήν γίνονται υπολογιστικά αδύνατες.

Μία δημοφιλής τεχνική είναι η Κατάβαση Κλίσης ή Gradient Descent (GD). Αυτή η τεχνική διορθώνει επαναληπτικά τα βάρη του δικτύου. Αρχικά υπολογίζονται οι μερικές παράγωγοι κάθε βάρους w_i ως προς τη συνάρτηση κόστους *J*. Ο υπολογισμός των μερικών παραγώγων γίνεται μέσω του αλγόριθμου οπισθοδιάδοσης (backpropagation algorithm), ο οποίος είναι μια αναδρομική εφαρμογή του κανόνα της αλυσίδας (Chain Rule) και περιγράφηκε για πρώτη φορά στο πλαίσιο των τεχνητών νευρωνικών δικτύων από τον Werbos το 1982. Ανάλογα λοιπόν με την κλίση ως προς τα w_i, δηλαδή $\nabla_{w_i} J(w_i)$, αυξομειώνονται οι τιμές των βαρών του δικτύου. Το μέγεθος της διόρθωσης (βήμα), ονομάζεται ρυθμός μάθησης και συμβολίζεται με η. Η ενημέρωση των βαρών γίνεται με την ακόλουθη σχέση:

$$w = w - \eta \cdot \nabla_w J(w) \tag{2.4}$$

Στην πράξη, χρησιμοποιείται ο αλγόριθμος στοχαστικής μείωσης κλίσης (Stochastic Gradient Decent) ο οποίος αποτελεί μία υποπερίπτωση του γενικού αλγορίθμου, όπου ο υπολογισμός της κλίσης και η ενημέρωση των βαρών γίνεται με χρήση ενός ή λίγων (mini batch SGD) δειγμάτων εκπαίδευσης (Robbins et al., 1951). Ο λόγος που είναι ιδιαίτερα διαδεδομένη η συγκεκριμένη τεχνική είναι γιατί μειώνει τις απαιτήσεις σε μνήμη και εμφανίζεται να επιταχύνει την εκπαίδευση του νευρωνικού δικτύου.
Εκτός από την απλή SGD, υπάρχουν και πιο εξεζητημένες παραλλαγές της, οι οποίες πετυχαίνουν ταχύτερη σύγκλιση, όπως με Nesterov Momentum (Nesterov 1983). Επιπλέον, τα τελευταία χρόνια χρησιμοποιούνται σε όλο και μεγαλύτερο βαθμό τεχνικές αυτόματης βελτιστοποίησης, στις οποίες γίνεται αυτόματη ρύθμιση του ρυθμού μάθησης, όπως Adagrad (Duchi et al., 2011), Adadelta (Zeiler 2012), RMSprop (Hinton et al., 2012) και Adam (Kingma et al., 2014), η οποία είναι αυτή τη στιγμή η πιο δημοφιλής τεχνική.

Αλγόριθμος: Οπισθοδιάδοση για ενημέρωση των βαρών

1. Αρχικοποίησε τα βάρη του δικτύου με τυχαίες τιμές.
2. Για κάθε παράδειγμα στα δεδομένα εκπαίδευσης κάνε τα ακόλουθα:
    a. Υπολόγισε την έξοδο του δικτύου και την συνάρτηση κόστους.
    b. Υπολόγισε τις παραγώγους της συνάρτησης κόστους ως προς τον κάθε νευρώνα του επιπέδου εξόδου.
    c. Ενημέρωσε τα βάρη που οδηγούν στο επίπεδο εξόδου.
    d. Για κάθε κρυφό επίπεδο ξεκινώντας από το επίπεδο εξόδου κάνε τα εξής:
        i. Υπολόγισε τις παραγώγους της συνάρτησης κόστους ως προς τον κάθε νευρώνα του επιπέδου αυτού κάνοντας χρήση του κανόνα της αλυσίδας.
        ii. Ενημέρωσε τα βάρη του επιπέδου.
3. Επανέλαβε το βήμα 2 μέχρι να συγκλίνει το δίκτυο.

Υπάρχουν δύο προβλήματα τα οποία μπορούν να προκύψουν αναφορικά με την ικανότητα του νευρωνικού δικτύου να μοντελοποιήσει το πρόβλημα. Είναι επιθυμητό για ένα δίκτυο να μάθει τις δομές ή τις αρχές που διέπουν το πρόβλημα και να αγνοήσει τον πιθανό θόρυβο. Τις περισσότερες φορές αυτό απαιτεί την εύρεση μίας λεπτής ισορροπίας, ανάμεσα σε δύο φαινόμενα:

- Υποπροσαρμογή
  Η υποπροσαρμογή (underfitting) προκύπτει, όταν το μοντέλο δεν είναι αρκετά σύνθετο ώστε να περιγράψει το πρόβλημα. Σε αυτή την περίπτωση σημαίνει ότι το πρόβλημα είναι πιο περίπλοκο από το ορισμένο νευρωνικό δίκτυο (Geman et al., 1992). Το πρόβλημα αυτό μπορεί να ξεπεραστεί με τους εξής τρόπους:
  - Αύξηση της πολυπλοκότητας του δικτύου. Αυτό μπορεί να επιτευχθεί με την αύξηση των παραμέτρων των επιπέδων, ή του βάθους του δικτύου.
  - Αύξηση της ευαισθησίας του μοντέλου. Αυτό επιτυγχάνεται με την επιλογή των κατάλληλων τιμών για ορισμένες από τις υπερπαραμέτρους των αλγορίθμων εκπαίδευσης.
  - Εξαγωγή περισσότερων χαρακτηριστικών. Μπορεί η αδυναμία του δικτύου να μοντελοποιήσει τα δεδομένα, να οφείλεται στο ότι δεν είναι αρκετά αντιπροσωπευτικός ο τρόπος με τον οποίο αναπαρίστανται. Με την εξαγωγή περισσότερων ή και πιο σύνθετων χαρακτηριστικών, το δίκτυο έχει τη δυνατότητα να ανακαλύψει πιο σύνθετες σχέσεις στα δεδομένα.
- Υπερπροσαρμογή
  Η υπερπροσαρμογή (overfitting) είναι πιο συνηθισμένο πρόβλημα και προκύπτει όταν το μοντέλο είναι πάρα πολύ σύνθετο και είναι ικανό να ταιριάξει τέλεια στα δεδομένα εκπαίδευσης. Αυτό σημαίνει ότι έχει μάθει ακόμα και το "θόρυβο" στα δεδομένα, δηλαδή ακόμη και τις μικρές ιδιαιτερότητές τους οι οποίες δεν αντιστοιχούν σε κάποια πραγματική δομή (δεν φέρουν χρήσιμη πληροφορία) (Geman et al., 1992).
  Το πρόβλημα τις υπερπροσαρμογής είναι πιο συχνό σήμερα, καθώς η διαθέσιμη υπολογιστική ισχύ επιτρέπει την εύκολη δημιουργία πολύ σύνθετων μοντέλων. Τα τεχνητά νευρωνικά δίκτυα είναι αρκετά επιρρεπή στο πρόβλημα αυτό καθώς συνήθως διαθέτουν εκατομμύρια παραμέτρους. Υπάρχουν διάφοροι τρόποι αντιμετώπισης ή περιορισμού του προβλήματος. Ορισμένες λύσεις είναι:
  - Αύξηση των δεδομένων εκπαίδευσης. Με αυτό τον τρόπο είναι πιο δύσκολο για το μοντέλο να ταιριάξει στο θόρυβο των δεδομένων, αλλά γίνονται και πιο ξεκάθαρες οι πραγματικές δομές στα δεδομένα. Είναι η καλύτερη λύση, αλλά και συχνά η πιο δύσκολη.
  - Ενίσχυση των δεδομένων εκπαίδευσης (data augmentation). Με αυτό τον τρόπο προστίθενται στο σύνολο δεδομένων, τεχνητές παρατηρήσεις. Οι παρατηρήσεις αυτές είτε δημιουργούνται από την αρχή είτε αποτελούν τροποποιημένες εκδόσεις των υπαρχόντων δεδομένων.
  - Περιορισμός μοντέλου. Με αυτό τον τρόπο μειώνονται οι παράμετροι του δικτύου, μειώνοντας κατά συνέπεια τους βαθμούς ελευθερίας του μοντέλου. Έτσι, ο αλγόριθμος δεν έχει την ευκαιρία να μάθει σχέσεις, πέρα από τις πραγματικές σχέσεις στα δεδομένα.

- Μείωση χαρακτηριστικών. Με αυτό τον τρόπο απλοποιείται η αναπαράσταση των παρατηρήσεων. Αυτή πολλές φορές δεν είναι καλή λύση. Ο λόγος είναι ότι αν οι παρατηρήσεις δεν περιέχουν περιττά χαρακτηριστικά, τότε απορρίπτεται χρήσιμη πληροφορία.
- Ένας άλλος τρόπος αντιμετώπισης του προβλήματος είναι με τεχνικές εξομάλυνσης (regularization). Μία συνηθισμένη λύση, η οποία χρησιμοποιείται σε διαδικασίες μάθησης νευρωνικών δικτύων, αφορά την αποθάρρυνση μεγάλων βαρών στο δίκτυο. Με τον τρόπο αυτό εμποδίζεται η υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης, καθώς δεν επιτρέπεται στο μοντέλο να δώσει υπερβολική σημασία σε συγκεκριμένα χαρακτηριστικά.
- Μία ακόμα συνήθης τεχνική είναι ο Περιορισμός Ενεργοποίησης (Dropout) που εφαρμόζεται μεταξύ δύο επιπέδων των νευρωνικών δικτύων, με βάση την οποία για κάθε επανάληψη του αλγορίθμου εκπαίδευσης επιλέγονται τυχαία κάποιες συνδέσεις νευρώνων οι οποίες μηδενίζονται. Για τον προσδιορισμό του ποσοστού των συνδέσεων που μηδενίζονται χρησιμοποιείται μία υπερπαράμετρος p που εκφράζει το ποσοστό αυτό.
- Μια άλλη τακτική που αντιμετωπίζει το πρόβλημα της υπερπροσαρμογής είναι αυτή του πρόωρου σταματήματος της εκπαίδευσης (Early Stopping). Με αυτόν τον τρόπο, διακόπτεται η διαδικασία της εκπαίδευσης μόλις παρατηρηθεί πτώση της απόδοσης του μοντέλου σε ένα σύνολο δεδομένων που χρησιμοποιείται μόνο για αξιολόγηση και δεν συμμετέχει ουσιαστικά για την εκπαίδευση του δικτύου. Ουσιαστικά καθώς το μοντέλο προσαρμόζεται στα διαθέσιμα δεδομένα εκπαίδευσης και μαθαίνει μοτίβα και συσχετίσεις, η απόδοση του στο σύνολο δεδομένων εκπαίδευσης και αξιολόγησης αυξάνεται. Ωστόσο, μετά από πολλά περάσματα του εκπαιδευτικού συνόλου δεδομένων, το μοντέλο τείνει να υπερπροσαρμόζεται και να μαθαίνει και τον θόρυβο που υπάρχει στο εκπαιδευτικό σύνολο δεδομένων. Σε αυτήν την περίπτωση, ενώ η απόδοση στο σύνολο εκπαίδευσης θα αυξάνεται, στο σύνολο αξιολόγησης θα μειώνεται. Το πρόωρο σταμάτημα έγκειται στην εύρεση αυτής της σωστής στιγμής με την μέγιστη απόδοση στο σύνολο αξιολόγησης.

2.2 Βαθιά Συνελικτικά Νευρωνικά Δίκτυα (CNN)

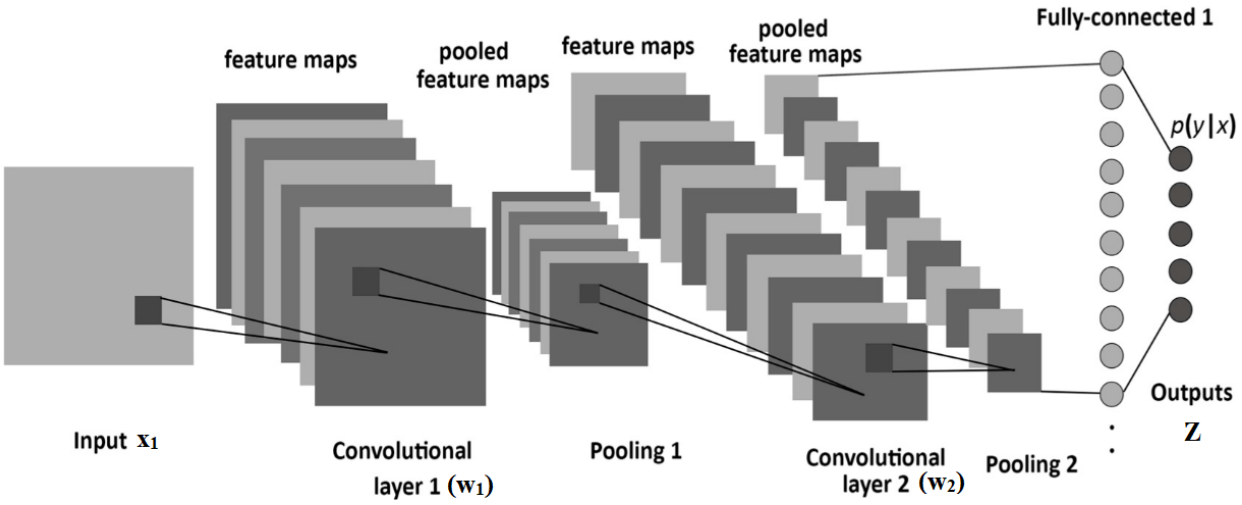
2.2.1 Ορισμός

Τα Βαθιά Συνελικτικά Νευρωνικά Δίκτυα (CNNs ή ConvNets) είναι νευρωνικές αρχιτεκτονικές δικτύων, ειδικά σχεδιασμένες για τη διαχείριση δεδομένων με κάποια χωρική τοπολογία (π.χ. εικόνες, βίντεο, φασματόγραμμα ήχου στην επεξεργασία φωνής, ακολουθίες χαρακτήρων σε κείμενο). Είναι μια υποκατηγορία των Τεχνητών Νευρωνικών Δικτύων τα οποία έχουν αποδειχθεί ότι είναι πολύ αποτελεσματικά σε πεδία όπως η αναγνώριση και ταξινόμηση εικόνων, με συγκεκριμένες επιτυχίες στην αναγνώριση προσώπων, αντικειμένων και φωτεινών σηματοδοτών, ενώ επίσης παρέχουν όραση σε ρομπότ και αυτοκινούμενα οχήματα. Τα δίκτυα αυτά αποτελούνται από νευρώνες οι οποίοι έχουν εκπαιδεύσιμα βάρη (weights) και κλίσεις (biases) παρόμοια με εκείνα των απλών Νευρωνικών Δικτύων που περιγράφηκαν αρχικά. Η διαφορά των συνελικτικών σε σχέση με των απλών δικτύων οφείλεται στην αρχιτεκτονική τους και στον τρόπο που προωθείται η πληροφορία εντός αυτών. Η αρχιτεκτονική τους είναι έτσι σχεδιασμένη ώστε πέραν την αξιοποίηση μεμονωμένων τιμών εισόδου να αξιοποιείται και η χωρική τους εξάρτηση (LeCun et al., 2015).

2.2.2 Επισκόπηση Αρχιτεκτονικής

Ένα Βαθύ Συνελικτικό Νευρωνικό Δίκτυο αποτελείται από έναν αριθμό από συνελικτικά (convolutional) και υποδειγματοληπτικά (subsampling) επίπεδα, τα οποία, προαιρετικώς, ακολουθούνται από πλήρως συνδεδεμένα επίπεδα (fully-connected layers). Η είσοδος σε ένα τέτοιο δίκτυο είναι, συνήθως, ένας πίνακας τριών διαστάσεων. Για παράδειγμα, μια εικόνα μπορεί να αναπαρασταθεί σε ένα πίνακα τριών διαστάσεων H x W x C, όπου το H (height) αντιστοιχεί στον αριθμό των pixel της εικόνας στον κάθετο άξονα, το W (width) αντιστοιχεί στον αριθμό των pixel της εικόνας στον οριζόντιο άξονα και το C (channels) αναφέρεται συνήθως σε φασματικά κανάλια όπου για συνήθεις εικόνες είναι τα χρωματικά κανάλια RGB (Red, Green, Blue) . Η είσοδος, στη συνέχεια, περνάει ακολουθιακά από μια σειρά επεξεργασιών.
Ένα βήμα επεξεργασίας, συνήθως, αποκαλείται επίπεδο, το οποίο θα μπορούσε να είναι ένα συνελικτικό επίπεδο (convolutional layer), ένα συγκεντρωτικό επίπεδο (pooling layer), ένα πλήρως συνδεδεμένο επίπεδο (fully connected layer) ή ένα επίπεδο απωλειών (loss layer). Τα επίπεδα αυτά περιγράφονται λεπτομερώς στη συνέχεια.
Προς το παρόν δίνεται μία αφηρημένη περιγραφή της δομής ενός βαθιού συνελικτικού δικτύου:

Έστω X_1 η είσοδος του δικτύου, για παράδειγμα μία εικόνα, η οποία περνά από διαδοχικά επίπεδα επεξεργασίας μέχρι να εξαχθεί η τελική έξοδος Z. Ως επίπεδο επεξεργασίας ορίζεται το w_i, όπου w είναι το διάνυσμα παραμέτρων του επιπέδου i. Κάθε επίπεδο επεξεργασίας i δέχεται μία είσοδο X_i, τη μετασχηματίζει και εξάγει μία έξοδο X_{i+1}, η οποία αποτελεί είσοδο του επόμενου επιπέδου i+1.



Σχήμα 2.4: Αρχιτεκτονική Συνελικτικού Δικτύου το οποίο αποτελείται από Συνελικτικά, Συγκεντρωτικά και Πλήρως Συνδεδεμένα Επίπεδα (Albelwi et al., 2017).

Η διαδικασία είναι σειριακή. Όταν ολοκληρωθεί το προωθητικό πέρασμα, ξεκινά μία επιπλέον διαδικασία, η οπισθοδιάδοση σφάλματος (backward error propagation), η οποία βασίζεται σε κάποια συνάρτηση κόστους και είναι απαραίτητη για την ενημέρωση και εκπαίδευση των παραμέτρων w του συνελικτικού νευρωνικού δικτύου. Υπάρχει μεγάλη ποικιλία στις αρχιτεκτονικές συνελικτικών νευρωνικών δικτύων, όμως ακολουθείται συνήθως μία συγκεκριμένη γενική μορφή (Σχήμα 2.4) (Schmidhuber 2014).

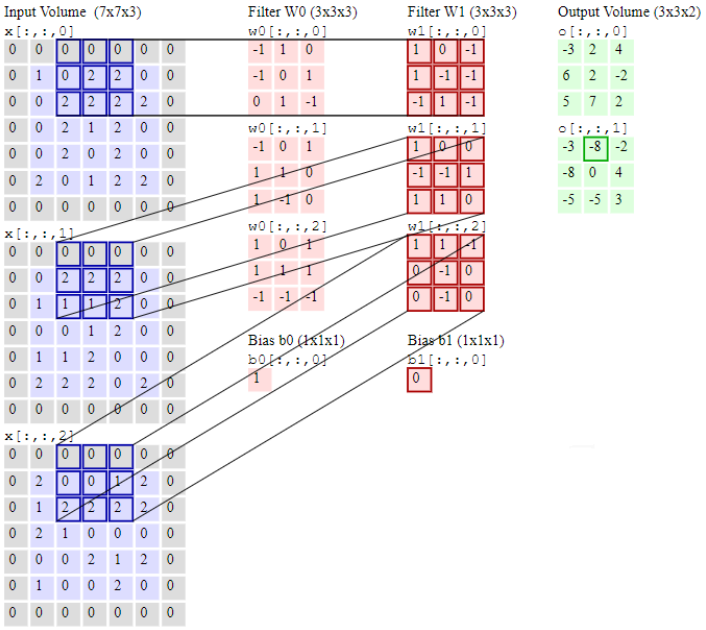
2.2.3 Συνελικτικο Επίπεδο (Convolutional Layer)

Το συνελικτικό επίπεδο είναι το πιο βασικό δομικό μέρος στην αρχιτεκτονική του συνελικτικού νευρωνικού δικτύου. Ένα συνελικτικό επίπεδο, όπως φανερώνει και η ονομασία του, συνελίσσει την είσοδο του με μία σειρά από φίλτρα ή πυρήνες (kernels), παράγοντας έτσι χάρτες χαρακτηριστικών. Στην περίπτωση της δισδιάστατης συνέλιξης, αν η είσοδος είναι μία εικόνα (RGB) ή ένας τρισδιάστατος τανυστής $input \in \mathbb{R}^{W \times H \times C_{in}}$, η μαθηματική περιγραφή της παραγόμενης εξόδου $Output \in \mathbb{R}^{W \times H \times C_{out}}$ ενός συνελικτικού επιπέδου για κάθε χάρτη χαρακτηριστικών C_{out} είναι:

$$Output(C_{out}) = f\left(bias(C_{out}) + \sum_{k=0}^{C_{in}} weight(C_{out}, k) \star input(k)\right) \qquad (2.5)\,,$$

όπου $\star$ συμβολίζει τον τελεστή διασυσχέτισης (cross-correlation), C_{in} δηλώνει τον αριθμό των καναλιών εισόδου, C_{out} τον αριθμό των παραγόμενων χαρτών χαρακτηριστικών, H το ύψος και W το πλάτος σε pixels της εισόδου, k ο δείκτης που αντιστοιχεί το φίλτρο και το κανάλι εισόδου και f μία μη γραμμική συνάρτηση ενεργοποίησης.
Για την μη γραμμική συνάρτηση ενεργοποίησης συνήθως επιλέγεται ReLU, όμως θα μπορούσε να τοποθετηθεί οποιαδήποτε άλλη μη γραμμική συνάρτηση ενεργοποίησης.

Ένα παράδειγμα εμφανίζεται στο επόμενο Σχήμα 2.5 όπου για μία εικόνα 7x7 με 3 κανάλια (Input Volume) παράγονται μέσω ενός συνελικτικού επιπέδου 2 χάρτες χαρακτηριστικών (Output Volume). Οι πίνακες με πορτοκαλί χρώμα αναφέρονται στα εκπαιδευόμενα βάρη. Το παράδειγμα αναφέρεται στον υπολογισμό της τιμής -8 του δεύτερου χάρτη χαρακτηριστικών. Κάθε μία περιοχή των δεδομένων εισόδου πολλαπλασιάζεται κατά σημείο με το αντίστοιχο φίλτρο και το τελικό αποτέλεσμα -8 είναι το συνολικό τους άθροισμα (Andrej Karpathy, CS231n).



Σχήμα 2.5: Παράδειγμα λειτουργίας ενός Συνελικτικού Επιπέδου [3].

2.2.4 Αριθμός Παραμέτρων ενός Συνελικτικού Επιπέδου

Έστω ένα συνελικτικό επίπεδο δισδιάστατης συνέλιξης που δέχεται σαν είσοδο έναν τρισδιάστατο τανυστή διαστάσεων $W_1 \times H_1 \times C_1$ (πλάτος, ύψος, κανάλια). Αν αυτό αποτελείται από φίλτρα διαστάσεων $M \times N \times C_1$ (πλάτος, ύψος, κανάλια) το καθένα, *Stride* είναι το βήμα που χρησιμοποιείται κατά τον υπολογισμό της συνέλιξης, και *Pad* το εύρος του zero padding στα άκρα της εισόδου, τότε η έξοδος του στρώματος είναι διαστάσεων $W_2 \times H_2 \times C_2$ όπου (Andrej Karpathy, CS231n):

$$W_2 = \frac{W_1 - M + 2P}{Stride} + 1, H_2 = \frac{H_1 - M + 2P}{Stride} + 1, C_2 = K \qquad (2.6)$$

Στο παράδειγμα του Σχήματος 2.5 το *Stride* είναι 1 και το Pad είναι 2.

Με βάση τα παραπάνω και λαμβάνοντας υπόψη τη διάσταση του διανύσματος πόλωσης, το πλήθος των υπό μάθηση παραμέτρων του συνελικτικού επιπέδου είναι:

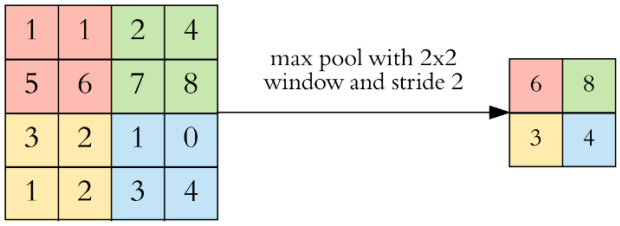
$$cart(P) = K \cdot (M \cdot N \cdot C_1) + K \tag{2.7}$$

2.2.5 Συγκεντρωτικό Επίπεδο (Pooling Layer)

Στην συντριπτική πλειοψηφία των μοντέλων, λόγω της ανάγκης μείωσης της διάστασης του μοντέλου αλλά και τη βελτίωση της απόδοσης του μοντέλου, χρησιμοποιείται μετά από κάποια εκ των επιπέδων συνέλιξης ένα Συγκεντρωτικό Επίπεδο. Σε ένα τέτοιο επίπεδο ενός συνελικτικού δικτύου, η έξοδος *Output* ενός προηγούμενου επιπέδου συνέλιξης υφίσταται τη δράση ενός τελεστή συγκέντρωσης. Συνηθίζεται ο τελεστής αυτός να είναι η συνάρτηση τοπικού μεγίστου που εφαρμόζεται σε μη επικαλυπτόμενες υποπεριοχές του τανυστή εισόδου με αποτέλεσμα αυτός να υποδειγματοληπτείται.
Σημειώνεται πως το επίπεδο συγκέντρωσης δεν προσθέτει υπό μάθηση παραμέτρους στο μοντέλο. Η μαθηματική περιγραφή της λειτουργίας του στρώματος συσσώρευσης μεγίστου (Max Pooling Layer) σε ένα δίκτυο δισδιάστατης ή τρισδιάστατης συνέλιξης περιγράφεται ως εξής: Σε κάθε μη επικαλυπτόμενη υποπεριοχή διάστασης M×N και για κάθε χάρτη χαρακτηριστικών (δηλαδή για κάθε κανάλι) $Output(C_{out})$ του προηγούμενου επιπέδου του δικτύου εφαρμόζεται η εξής σχέση:

$$Output_{C_{out}}^{Region} = \max_{\forall x \in Region}\{Output_{C_{out}}(x)\} \tag{2.8}$$

Ουσιαστικά, με βάση την παραπάνω σχέση διατηρείται μόνο η μέγιστη ενεργοποίηση εντός μίας υποπεριοχής του χάρτη χαρακτηριστικών. Επίσης όπως και στην περίπτωση του επιπέδου συνέλιξης έτσι και εδώ συνηθίζεται να χρησιμοποιείται μία παράμετρος βήματος Stride που καθορίζει αν η παραπάνω σχέση θα εφαρμοστεί σε κάθε υποπεριοχή της εισόδου ή αν αυτό θα συμβεί για λιγότερες από όλες τις δυνατές υποπεριοχές. Αν επιλεχθεί το δεύτερο, η είσοδος υποδειγματοληπτείται περαιτέρω (Andrej Karpathy, CS231n).



Σχήμα 2.6: Παράδειγμα λειτουργίας ενός Συγκεντρωτικού Επιπέδου [3].

Συνεπώς, για ένα δίκτυο δισδιάστατης συνέλιξης, αν η είσοδος είναι ένας τανυστής διαστάσεων $W_1 \times H_1 \times C_1$ (πλάτος,ύψος,κανάλια), Stride το βήμα και M×N η διάσταση των υποπεριοχών, τότε η έξοδος του στρώματος συσσώρευσης είναι ένας τανυστής διαστάσεων $W_2 \times H_2 \times C_2$ όπου:

$$W_2 = \frac{W_1 - M}{Stride} + 1, H_2 = \frac{H_1 - M}{Stride} + 1, C_2 = C_1 \quad (2.9)$$

2.2.6 Πλήρως Συνδεδεμένο Επίπεδο (Fully Connected Layer)

Οι έξοδοι των συνελικτικών και συγκεντρωτικών επιπέδων αναπαριστούν χαρακτηριστικά υψηλών στρωμάτων. Ο σκοπός του πλήρως συνδεδεμένου επιπέδου είναι να χρησιμοποιήσει αυτά τα χαρακτηριστικά προκειμένου να κατηγοριοποιήσει την εικόνα εισόδου σε διάφορες κλάσεις, βασιζόμενο στο σύνολο δεδομένων που χρησιμοποιήθηκαν για εκπαίδευση. Στο παράδειγμα που εμφανίζεται στο Σχήμα 2.4 παρατηρείται μετά το τέλος του 2ου συγκεντρωτικού επιπέδου όλοι οι χάρτες χαρακτηριστικών τροφοδοτούνται σε μορφή διανύσματος σε ένα πλήρως συνδεδεμένο επίπεδο το οποίο κατηγοριοποιεί τελικά την εικόνα εισόδου σε κλάσεις με τις αντίστοιχες πιθανότητες. Για την εκτίμηση των πιθανοτήτων χρησιμοποιείται σαν συνάρτηση ενεργοποίησης η Softmax $\sigma: \mathbb{R}^k \to \mathbb{R}^k$ η οποία χρησιμεύει για την εκτίμηση πιθανοτήτων στην περίπτωση πολλών διαφορετικών κλάσεων (Andrej Karpathy, CS231n).

$$\sigma(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}, \text{ όπου } i = 1,2, \dots, k\ , \boldsymbol{z} = (z_1, z_2, \dots z_k) \quad (2.10)$$

2.2.7 Το Πρόβλημα των Εξαφανιζόμενων Κλίσεων (Vanishing Gradient Problem)

Καθώς αυξάνεται το βάθος ενός δικτύου, αρχίζει να πάσχει από το πρόβλημα εξαφανιζόμενων κλίσεων. Το συγκεκριμένο πρόβλημα εμφανίζεται στο στάδιο εκπαίδευσης των βαθιών νευρωνικών δικτύων. Σε κάθε βήμα εκπαίδευσης και προσαρμογής του δικτύου, κάθε ένα από τα βάρη του λαμβάνει μια ενημέρωση ανάλογη της μερικής παραγώγου της συνάρτησης κόστους ως προς το βάρος αυτό. Σε κάποιες περιπτώσεις η ενημέρωση αυτή είναι εξαιρετικά μικρή εμποδίζοντας να μεταβληθούν ουσιαστικά οι τιμές των βαρών. Αυτό έχει σαν αποτέλεσμα να σταματήσει τελείως η περαιτέρω εκπαίδευση του δικτύου.
Συγκεκριμένα, ο αλγόριθμος οπισθοδιάδοσης του λαθους υπολογίζει τις μερικές παραγώγους μέσω του κανόνα της αλυσίδας. Συνεπώς για ένα δίκτυο με *n* επίπεδα ο υπολογισμός των μερικών παραγώγων της συνάρτησης κόστους ως προς τα βάρη του πρώτου επιπέδου προκύπτει από τον πολλαπλασιασμό των μερικών παραγώγων των *n* επόμενων επιπέδων μέσω του κανόνα της αλυσίδας. Έτσι για τις περιπτώσεις των περισσότερων παραδοσιακών συναρτήσεων ενεργοποίησης που οι παράγωγοι τους βρίσκονται στο εύρος (0,1) έχει σαν αποτέλεσμα οι κλίσεις ως προς τα βάρη των πρώτων επιπέδων να μειώνεται με εκθετικό ρυθμό καθώς αυξάνεται το βάθος του δικτύου (Hochreiter 1991).

Σημειώνεται ότι στην περίπτωση που οι παράγωγοι των συναρτήσεων ενεργοποίησης παίρνουν μεγαλύτερες τιμές τότε μπορεί να συμβεί το ισοδύναμο πρόβλημα που ονομάζεται Πρόβλημα Ανατινασσόμενων Κλίσεων (exploding gradient problem) το οποίο δημιουργεί εξίσου σημαντικά προβλήματα στην διαδικασία μάθησης.
Αρκετές λύσεις έχουν προταθεί στην προσπάθεια επίλυσης του προβλήματος των Εξαφανιζόμενων κλίσεων. Κάποιες από αυτές είναι:

- η σταδιακή εκπαίδευση των δικτύων με μη επιβλεπόμενο τρόπο με ένα επίπεδο την φορά, σταδιακή επέκταση τους και στο τέλος προσαρμογή με επιβλεπόμενη μάθηση μέσω του αλγορίθμου backpropagation (Schmidhuber 1992)
- η χρήση της συνάρτησης ReLU ως συνάρτησης ενεργοποίησης
- τεχνικές αρχικοποίησης των βαρών του δικτύου
- η βελτίωση της ταχύτητας των GPU (το οποίο δεν αντιμετωπίζει δομικά το πρόβλημα)
- οι παραλειπόμενες συνδέσεις οι οποίες αναλύονται σε επόμενη υποενότητα.

2.2.8 Γνωστά Συνελικτικά Νευρωνικά Δίκτυα

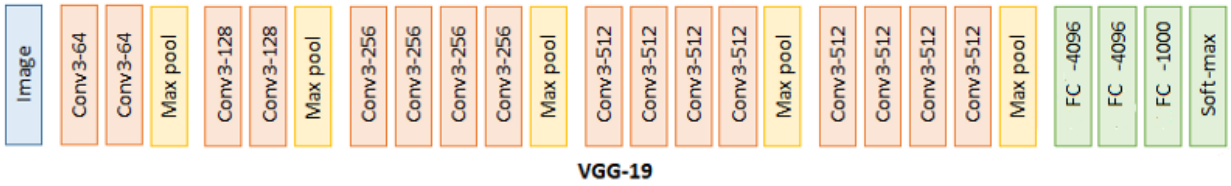
Στο σημείο αυτό παρουσιάζονται επιγραμματικά μερικά, ήδη υπάρχοντα, γνωστά συνελικτικά δίκτυα που χρησιμοποιούνται ευρέως στον χώρο της Βαθιάς Μηχανικής Μάθησης.

- *LeNet* (LeCun et al., 1998): Αποτελεί την πρώτη επιτυχημένη εφαρμογή Συνελικτικών Δικτύων που αναπτύχθηκε από τον Yann LeCun την δεκαετία του 1990. Η συγκεκριμένη αρχιτεκτονική χρησιμοποιήθηκε κυρίως για αναγνώριση κωδικών, ψηφίων κ.λ.π.
- *AlexNet* (Krizhevsky et al., 2012): Το συγκεκριμένο δίκτυο ήταν το πρώτο το οποίο έκανε τα Συνελικτικά Δίκτυα διάσημα στο χώρο της Όρασης Υπολογιστών. Το δίκτυο αυτό είχε μία πολύ παρόμοια αρχιτεκτονική με αυτή του LeNet, ωστόσο, ήταν βαθύτερο, μεγαλύτερο και είχε πολλά συνελικτικά επίπεδα, στοιβαγμένα το ένα πάνω στο άλλο, που είχε αποτελέσει μια πρωτοποριακή τεχνική.
- *ZF Net* (Zeiler et al., 2013): Το δίκτυο αυτό αποτελεί μια βελτίωση του AlexNet, διορθώνοντας κάποιες υπερπαραμέτρους της αρχιτεκτονικής. Πιο συγκεκριμένα, επεκτάθηκε το μέγεθος του μεσαίου συνελικτικού επιπέδου, ενώ παράλληλα έκαναν το βήμα και το μέγεθος φίλτρου του πρώτου επιπέδου μικρότερο. Το συγκεκριμένο δίκτυο απέσπασε την πρώτη θέση στον διαγωνισμό ILSVRC 2013.
- *GoogleNet* (Szegedy et al., 2015): Νικητής του διαγωνισμού ILSVRC 2014 αναδείχθηκε το συγκεκριμένο δίκτυο το οποίο αναπτύχθηκε από την Google. Το βασικό πλεονέκτημά του έναντι στα προηγούμενα μοντέλα, έγκειται στην δραματική μείωση των παραμέτρων του δικτύου. Πιο συγκεκριμένα, ο αριθμός των παραμέτρων μειώθηκε στα 4 εκατομμύρια σε σύγκριση με το AlexNet που είχε 60 εκατομμύρια παραμέτρους.

2.2.9 Δίκτυο VGGNet

Το VGGNet έχει αποσπάσει στον διαγωνισμό ILSVRC 2014 τη δεύτερη θέση στο πρόβλημα κατηγοριοποίησης εικόνων, καθώς και την πρώτη θέση στο πρόβλημα εντοπισμού αντικειμένων

μέσα στην εικόνα. Το συγκεκριμένο δίκτυο έχει αναπτυχθεί από την ομάδα Visual Geometry Group (VGG) του πανεπιστημίου της Οξφόρδης και είναι αρκετά διάσημο καθώς αποδίδει πάρα πολύ καλά αλλά σε πολλές εργασίας Όρασης Υπολογιστών. Επίσης η ομάδα που το δημιούργησε παραχωρεί ελεύθερα τα βάρη από το εκπαιδευμένο δίκτυο σε συμβατή μορφή με την βιβλιοθήκη Caffe. Η βασική συνεισφορά του είναι στο γεγονός ότι το βάθος ενός δικτύου είναι ένα σημαντικό συστατικό για την καλή απόδοση.
Το συγκεκριμένο δίκτυο χρησιμοποιεί το μικρότερο και απλούστερο 3x3 παράθυρο συνέλιξης στα συνελικτικά επίπεδα που περιέχει και εμφανίζει βελτιωμένες αποδόσεις καθώς αυξάνεται το βάθος του. Το VGGNet εμφανίζεται σε τέσσερις εκδόσεις ως προς το βάθος του. Τα VGG-11, VGG-13 , VGG-16 και VGG-19. Στη συνέχεια παρατίθενται η αρχιτεκτονική του δικτύου VGG-19 το οποίο εμφανίζει και την βέλτιστη απόδοση (Σχήμα 2.7). Το δίκτυο αποτελείται σχηματικά από δύο μέρη. Το πρώτο μέρος αποτελείται από διαδοχικά συνελικτικά επίπεδα (Convolutional layers) στα οποία παρεμβάλλονται κάποια συγκεντρωτικά επίπεδα (Max pooling layers) μέσω των οποίων γίνεται η εξαγωγή των χαρακτηριστικών της εικόνας. Το δεύτερο τμήμα αποτελείται από πλήρως συνδεδεμένα επίπεδα μέσω τον οποίων γίνεται η τελική ταξινόμηση. Ο συνολικός αριθμός των παραμέτρων του είναι 144 εκατομμύρια, το οποίο αποτελεί ένα μειονέκτημα όσον αφορά τη μνήμη που απαιτεί κατά την εκπαίδευση αλλά και κατά την αποθήκευσή του (Simonyan et al., 2015).



Σχήμα 2.7: Αρχιτεκτονική δικτύου VVGNet-19 [4] (Μετά από επεξεργασία).

2.2.10 Δίκτυα ResNet και DenseNet

Στις κλασικές αρχιτεκτονικές συνελικτικών δικτύων η πληροφορία περνάει διαδοχικά από το κάθε επίπεδο στο επόμενο. Ένα τέτοιο δίκτυο είναι και το VGGNet-19 όπως περιγράφηκε προηγουμένως. Όπως αναφέρθηκε, στα αρχικά στάδια αύξησης του βάθους ενός δικτύου υπάρχει η τάση για αύξηση της ακρίβειάς του. Παρόλα αυτά καθώς αυξάνεται το βάθος, το δίκτυο αρχίζει να πάσχει από το πρόβλημα των Εξαφανιζόμενων Κλίσεων όπως περιγράφηκε προηγουμένως. Ένα ακόμα πρόβλημα που έχει παρατηρηθεί είναι ότι καθώς προστίθενται νέα επίπεδα, το πλήθος των παραμέτρων αυξάνεται δραματικά το οποίο οδηγεί σε αύξηση της υπολογιστικής πολυπλοκότητας.

Οι He et al. (2015) για την επίλυση των προβλημάτων αυτών πρότειναν το δίκτυο ResNet το όποιο ήταν το νικητήριο για τον διαγωνισμό ILSVRC 2015. Το δίκτυο αυτό εισήγαγε για πρώτη φορά ως λογική στην εκπαίδευση ενός δικτύου την μάθηση των Συναρτήσεων Κατάλοιπων (Residual Functions) μέσω παραλειπόμενων συνδέσεων (skip connections). Ο βασικός στόχος είναι η αύξηση της απόδοσης καθώς αυξάνεται το βάθος του δικτύου. Επομένως θεωρείται ότι τα επιπρόσθετα συνελικτικά επίπεδα που δεν χρειάζονται θα πρέπει να προσεγγίζουν την ταυτοτική απεικόνιση και να μην επηρεάζουν την απόδοση του δικτύου, δηλαδή :

$$x \rightarrow H(x) = x, \tag{2.11}$$

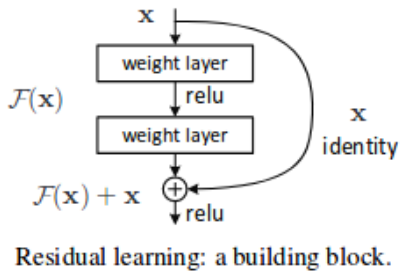
όπου x αποτελεί την είσοδο στο Συνελικτικό Επίπεδο H. Επίσης τα κατάλοιπα F ενός συνελικτικού επιπέδου (που διατηρεί την διάσταση της εισόδου του) μπορούν να οριστούν και ως:

$$F(x) = H(x) - x. \tag{2.12}$$

Με αποτέλεσμα όλα τα επίπεδα που δεν συνεισφέρουν στην επίλυση του προβλήματος να εμφανίζουν κατάλοιπα:

$$F(x) = 0 \Leftrightarrow H(x) = x \tag{2.13}$$

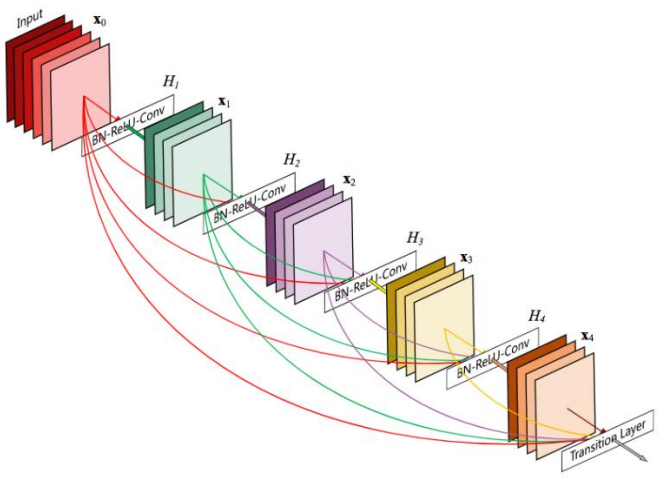
Οι δημιουργοί των δικτύων αυτών παρατήρησαν ότι η εκτίμηση της ταυτοτικής συνάρτησης μέσω ενός συνελικτικού επιπέδου που αποτελείται από μη γραμμικές συναρτήσεις είναι ένα πιο δύσκολο πρόβλημα από την εκτίμηση της συνάρτησης κατάλοιπων γύρω από το 0. Με απλά λόγια είναι πιο εύκολη η εύρεση μιας συνάρτησης F(x) = 0 έναντι της F(x) = x όταν εμπλέκονται μη γραμμικές συναρτήσεις.



Σχήμα 2.8: Δομικό στοιχείο ενός Επιπέδου Κατάλοιπων (He at al., 2015).

Σχηματικά λοιπόν στα Επίπεδα Κατάλοιπων (Residual Layers) η είσοδος προστίθεται κατά σημείο στην έξοδο μέσω παραλειπόμενων συνδέσεων με αποτέλεσμα το δίκτυο να προσπαθεί να μάθει τα κατάλοιπα (Σχημα 2.8). Η αρχιτεκτονική αυτή εκτός των άλλων ενισχύει και την οπισθοδιάδοση των κλίσεων (He et al., 2015).

Μία επέκταση των ResNet αποτελεί το δίκτυο DenseNet. Το DenseNet είναι αρκετά όμοιο με το ResNet ωστόσο διαφοροποιούνται σε δύο σημεία. Η πρώτη σημαντική διαφορά είναι ότι το DenseNet έχει κάθε επίπεδο απευθείας συνδεδεμένο με κάθε άλλο επίπεδο που έπεται μέσω παραλειπόμενων συνδέσεων. Η δεύτερη είναι ότι στην περίπτωση του DenseNet χρησιμοποιείται σειριακή επέκταση (concatenation) των χαρακτηριστικών έναντι της κατά σημείο πρόσθεσης που χρησιμοποιεί το ResNet. Συνεπώς στο DenseNet κάθε επίπεδο δέχεται μία συλλογή γνώσης από τα προηγούμενα επίπεδα. Αποτέλεσμα αυτού είναι ότι το δίκτυο μπορεί να είναι μικρότερο και να αξιοποιήσει καλύτερα την πληροφορία που δέχεται και λόγω αυτού να καταλαμβάνει λιγότερη μνήμη. Πιο συγκεκριμένα τα DenseNet αποτελούνται από Πυκνές Στοίβες (Dense Blocks) τα οποία αποτελούνται από συνελικτικά επίπεδα τα οποία είναι απευθείας συνδεδεμένα μεταξύ τους. Τα Dense blocks συνδέονται μεταξύ τους μέσω των επιπέδων μετάβασης (Transition layers). Εντός του Dense block κάθε επίπεδο λαμβάνει σαν είσοδο τις εξόδους όλων των προηγούμενων επιπέδων όπως αποτυπώνεται στο Σχήμα 2.9.



Σχήμα 2.9: Ένα Dense Block 5 επιπέδων με ρυθμό ανάπτυξης k=4. Κάθε επίπεδο δέχεται σαν είσοδο τους χάρτες χαρακτηριστικών όλων των προηγούμενων επιπέδων (Huang et al., 2017).

Το κάθε επίπεδο ενός Dense block αποτελείται από τα εξής επιμέρους επίπεδα:

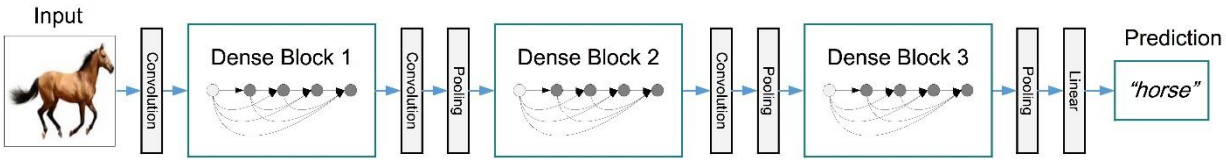
1. Κανονικοποίηση του πακέτου δεδομένων εισόδου (Batch Normalization)
2. Συνάρτηση ενεργοποίησης ReLU
3. Συνελικτικό επίπεδο με 3x3 πυρήνα

Οι δημιουργοί του δικτύου παρατήρησαν ότι η τοποθέτηση της συνάρτησης ενεργοποίησης πριν από το συνελικτικό επίπεδο εμφανίζει καλύτερα αποτελέσματα έναντι της πιο συνηθισμένης αντίστροφης διάταξης.

Ένα ακόμα χαρακτηριστικό των Dense blocks είναι ότι η είσοδος κάθε επιπέδου είναι η έξοδος όλων των προηγούμενων επιπέδων. Συνεπώς κάθε επίπεδο παράγει κάποια επιπλέον χαρακτηριστικά (feature maps) και ο αριθμός των επιπλέον χαρακτηριστικών ορίζεται ως ο ρυθμός ανάπτυξης (growth rate). Συνεπώς με ρυθμό ανάπτυξης 32, το εντέκατο επίπεδο ενός Dense block δέχεται σαν είσοδο (11-1)*32 = 320 χάρτες χαρακτηριστικών. Για το λόγο αυτό έχει προταθεί μία παραλλαγή των Dense blocks στην οποία πριν από κάθε επίπεδο ενός Dense Block όπως ορίστηκε προηγουμένως παρεμβάλονται και συσσωρευτικά επίπεδα (Bottleneck) τα οποία ουσιαστικά αποτελούν 1x1 συνελικτικά επίπεδα με πλήθος χαρακτηριστικών ίσο με 4 φορές τον ορισμένο ρυθμό ανάπτυξης. Η περίπτωση των Dense blocks με Bottleneck επίπεδα είναι:

1. Κανονικοποίηση του πακέτου δεδομένων εισόδου (Batch Normalization)
2. Συνάρτηση ενεργοποίησης ReLU
3. Συνελικτικό συσσωρευτικό επίπεδο με 1x1 πυρήνα που παράγει 4*ρυθμό ανάπτυξης χάρτες χαρακτηριστικών
4. Κανονικοποίηση του πακέτου δεδομένων εισόδου (Batch Normalization)
5. Συνάρτηση ενεργοποίησης ReLU
6. Συνελικτικό επίπεδο με 3x3 πυρήνα

Συνεπώς σχετικά με το προηγούμενο παράδειγμα, στην περίπτωση που περιλαμβάνονται συσσωρευτικά επίπεδα, θα δίνονται σαν είσοδο σε κάθε επίπεδο σταθερά 128 χάρτες χαρακτηριστικών.



Σχήμα 2.10: Ένα DenseNet με 3 Dense Blocks. Τα επίπεδα μεταξύ δύο γειτονικών Dense Block αναφέρονται ως Επίπεδα Μετάβασης τα οποία αλλάζουν το μέγεθος των χαρτών χαρακτηριστικών μέσω Συνελικτικών και Συγκεντρωτκών Επιπέδων (Huang et al., 2017).

Επίσης εντός των Dense blocks οι διαστάσεις των χαρακτηριστικών παραμένουν σταθερές καθώς σε άλλη περίπτωση δεν θα μπορούσε να επιτευχθεί σειριακή επέκταση των χαρτών χαρακτηριστικών. Ωστόσο επειδή η υποδειγματοληψία ή η σταδιακή μείωση της χωρικής διάστασης των χαρτών χαρακτηριστικών είναι απαραίτητη για την αρχιτεκτονική ενός συνελικτικού δικτύου, μεταξύ των Dense blocks εμφανίζονται επίπεδα μετάβασης (Transition layers) τα οποία επιτυγχάνουν αυτό τον σκοπό. Ένα τέτοιο επίπεδο περιλαμβάνει:

1. Κανονικοποίηση του πακέτου δεδομένων εισόδου (Batch Normalization - BN)
2. Συνελικτικό επίπεδο με 1x1 πυρήνα
3. Συγκεντρωτικό επίπεδο ως προς τον μέσο όρο (Average pooling)

Στην συνέχεια παρουσιάζεται το δίκτυο DenseNet-161.

Layers	Output Size	DenseNet-161($k = 48$)
Convolution	112×112	7×7 conv, stride 2
Pooling	56×56	3×3 max pool, stride 2
Dense Block (1)	56×56	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	56×56	1×1 conv
	28×28	2×2 average pool, stride 2
Dense Block (2)	28×28	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	28×28	1×1 conv
	14×14	2×2 average pool, stride 2
Dense Block (3)	14×14	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 36$
Transition Layer (3)	14×14	1×1 conv
	7×7	2×2 average pool, stride 2
Dense Block (4)	7×7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$
Classification Layer	1×1	7×7 global average pool
		1000D fully-connected, softmax

Σχήμα 2.11: Αρχιτεκτονική ενός DenseNet-161. Σημειώνεται ότι κάθε "conv" επίπεδο αποτελείται ουσιαστικά από την ακολουθία BN-ReLU-Conv (Huang et al., 2017).

Το DenseNet-161 παρόλο που αποτελείται από 161 επίπεδα περιέχει περίπου 29 εκατομμύρια παραμέτρους, δηλαδή σχεδόν 5 φορές λιγότερες από το VGGNet-19. Ο λόγος είναι ότι χρησιμοποιούν σε κάθε επίπεδο 48 χάρτες χαρακτηριστικών αντίθετα με εκείνους του VGGNet-19 όπου συνεχώς αυξάνονται (2*64, 2*128, 4*256, 8*512). Συνεπώς λόγω των πυκνών συνδέσεων που χρησιμοποιούνται μεταξύ των επιπέδων παράγονται λιγότερα περιττά χαρακτηριστικά κάνοντας χρήση λιγότερων παραμέτρων (Huang et al., 2017).

2.2.11 Δίκτυα Λεωφόρων (Highway Networks)

Το βάθος των νευρωνικών δικτύων όπως έχει ήδη αναφερθει αποτελεί βασικό συστατικό για την επιτυχία τους. Ωστόσο, η εκπαίδευση τους καθίσταται δυσκολότερη με την αύξηση του βάθους. Μια καινοτόμα αρχιτεκτονική που έχει σχεδιαστεί για να διευκολύνει την εκπαίδευση των πολύ βαθιών δικτύων αποτελούν τα δίκτυα Λεωφόρων (Highway Networks). Αναφέρονται έτσι καθώς επιτρέπουν την απρόσκοπτη ροή πληροφορίας στο δίκτυο. Η αρχιτεκτονική τους χρησιμοποιεί πύλες (gates) που μαθαίνουν να ρυθμίζουν τη ροή της πληροφορίας στο δίκτυο. Η αρχιτεκτονική τους είναι εμπνευσμένη από τα δίκτυα Νευρώνων Μακράς-Βραχείας Μνήμης (Long Short-Term Memory Units ή LSTMs). Τα Highway Δίκτυα μπορούν να εκπαιδευτούν με εκατοντάδες επίπεδα απευθείας μέσω του αλγόριθμου στοχαστικής μείωσης κλίσης (Stochastic Gradient Decent) και με μεγάλη ποικιλία ως προς τις επιλογές των συναρτήσεων ενεργοποίησης. Έστω ένα νευρωνικό δίκτυο το οποίο αποτελείται από L επίπεδα όπου το επίπεδο $l \in \{1,2, \dots, L\}$ εφαρμόζει στα δεδομένα εισόδου του x_l τον μη-γραμμικό μετασχηματισμό H_{W_l} όπου ως W_l

συμβολίζονται τα βάρη του και παράγει ως έξοδο του το y_l. Συνεπώς x_1 είναι η είσοδος στο δίκτυο και y_L είναι η έξοδος του δικτύου. Για λόγους απλότητας η έξοδος κάθε επιπέδου είναι:

$$y = H(x, W_H) \tag{2.14}$$

Στην περίπτωση των Highway δικτύων σε κάθε επίπεδο προστίθενται δύο ακόμα μη γραμμικοί μετασχηματισμοί $T(x, W_T)$ και $C(x, W_C)$ έτσι ώστε:

$$y = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C) \tag{2.15}$$

Ο μετασχηματισμός T αναφέρεται ως πύλη Μετασχηματισμού (Transform gate) ενώ ο C ως πύλη Μεταφοράς (Carry gate) καθώς εκφράζουν πόση πληροφορία παράγεται από τον μετασχηματισμό της εισόδου και πόση μεταφέρεται χωρίς καμία τροποποίηση, αντίστοιχα. Για απλότητα και μείωση του πλήθους των παραμέτρων προτείνεται η παραλλαγή $C = 1 - T$ για την πύλη Μεταφοράς. Συνεπώς ένα επίπεδο Highway ορίζεται ως:

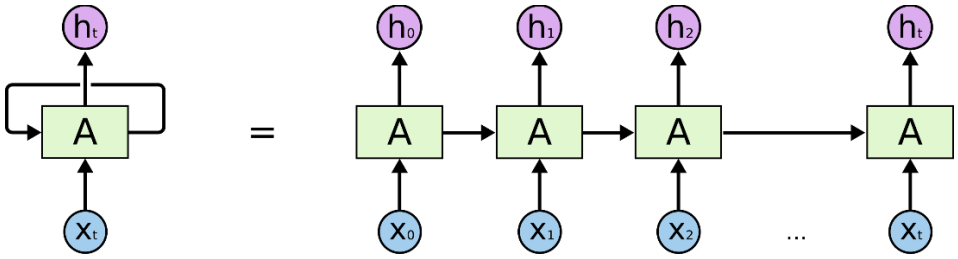
$$y = H(x, W_H) \cdot T(x, W_T) + x \cdot \big(1 - T(x, W_T)\big) \tag{2.16}$$

Έτσι στην ακραία περίπτωση που $T(x, W_T) = 0$ τότε η έξοδος είναι ίδια με την είσοδο, ενώ αν $T(x, W_T) = 1$ τότε η έξοδος είναι ο μετασχηματισμός $H(x, W_H)$.
Σημειώνεται επίσης ότι η διάσταση των x, y , $H(x, W_H)$ και $T(x, W_T)$ πρέπει να είναι ίδια. Σε περιπτώσεις όπου είναι επιθυμητό να αλλάξει το μέγεθος της αναπαράστασης, μπορεί κανείς να αντικαταστήσει το x με το $\hat{x}$ που λαμβάνεται κατάλληλα είτε με υπο-δειγματοληψία ή Padding. Μια άλλη εναλλακτική λύση είναι να χρησιμοποιηθεί ένα πλήρως συνδεδεμένο επίπεδο (χωρίς highway επίπεδο) για να τροποποιηθεί η διάσταση και στη συνέχεια να ξανά χρησιμοποιηθούν highway επίπεδα. Τα δίκτυα αυτά μπορούν να χρησιμοποιηθούν συμπληρωματικά με πλήρως συνδεδεμένα επίπεδα και με δομή των $H(x, W_H)$ και $T(x, W_T)$ επίσης ως πλήρως συνδεδεμένα δίκτυα.
Τα συνελικτικά Highway επίπεδα κατασκευάζονται με όμοιο τρόπο με τα πλήρως συνδεδεμένα επίπεδα. Οι μετασχηματισμοί $H(x, W_H)$ και $T(x, W_T)$ σε αυτή την περίπτωση αποτελούν συνελικτικά επίπεδα όπως ορίστηκαν σε προηγούμενη υποενότητα (Srivastava et al., 2015).

2.3 Ανατροφοδοτούμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks)

2.3.1 Ορισμός

Τα ανατροφοδοτούμενα νευρωνικά δίκτυα είναι μια ειδική κατηγορία τεχνητών νευρωνικών δικτύων. Τα δίκτυα αυτά διαφέρουν από τα απλά προωθητικά δίκτυα, υπό την έννοια ότι προωθούν την έξοδό τους ξανά προς στην είσοδό τους. Επίσης έχουν δυναμική αρχιτεκτονική, το οποίο τους επιτρέπει να ξεπεράσουν ορισμένους από τους περιορισμούς άλλων δικτύων, όπως Συνελικτικά ή τα Πλήρως Συνδεδεμένα Δίκτυα, όπως ότι μπορούν να επεξεργάζονται δεδομένα μεταβλητού μήκους και ότι διαθέτουν "μνήμη", που τους επιτρέπει να ανακαλύπτουν εξαρτήσεις μεταξύ των δεδομένων εισόσου. Τα δίκτυα αυτά επεξεργάζονται αποτελεσματικά κάθε είδους ακολουθιακά δεδομένα και παραδείγματα αποτελούν η φωνή, η γραφή, η οπτική πληροφορία που προκύπτει από μία κίνηση.



Σχήμα 2.12: 'Ξεδιπλωμένο' Ανατροφοδοτούμενο Νευρωνικό Δίκτυο [5].

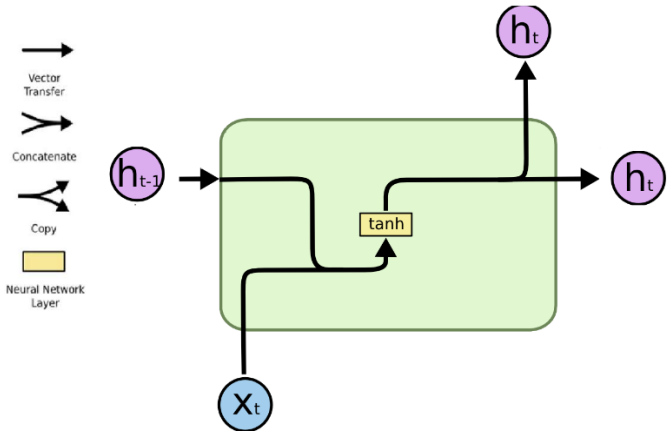
Ένα αναδρομικό δίκτυο μετασχηματίζει κάθε νέα είσοδο με τρόπο που εξαρτάται τόσο από την ίδια την είσοδο όσο και από τις προηγούμενες εισόδους που έχει δεχτεί. Συγκεκριμένα ένα ανατροφοδοτούμενο νευρωνικό δίκτυο f ενός επιπέδου που δέχεται τη χρονική στιγμή t την είσοδο x_t παράγει την έξοδο h_t ως εξής:

$$h_t = f(x_t, h_{t-1}) = f\big(x_t, f(x_{t-1}, h_{t-2})\big) = \cdots = f\Big(x_t, f\big(x_{t-1}, \dots f(x_1, h_0)\big)\Big) \qquad (2.17)$$

Επίσης η γενική μορφή ενός τέτοιου επιπέδου είναι:

$$h_t = f(x_t, h_{t-1}) = f_{W,U}(x_t, h_{t-1}) = f(W \cdot x_t + U \cdot h_{t-1} + b) \qquad (2.18)$$

όπου $f(\cdot)$ μία μη γραμμική συνάρτηση ενεργοποίησης, W ο πίνακας παραμέτρων που επιδρούν πάνω στην είσοδο x_t, U ο πίνακας παραμέτρων που επιδρούν πάνω στην έξοδο h_t του επιπέδου την προηγούμενη χρονική στιγμή και b ένα διάνυσμα πόλωσης. Σημειώνεται πως η έξοδος h_t συχνά αναφέρεται και ως κατάσταση του δικτύου.
Ένα παράδειγμα ένα τέτοιου επιπέδου εμφανίζεται στο επόμενο Σχήμα 2.13 όπου η συνάρτηση ενεργοποίησης f είναι η υπερβολική εφαπτομένη $\tanh(\cdot)$ (Elman 1990).



Σχήμα 2.13: Δομικό στοιχείο ενός απλού Ανατροφοδοτούμενου Νευρωνικού Δικτύου ενός επιπέδου [5] (Μετά από επεξεργασία).

2.3.2 RNN Διπλής Κατεύθυνσης

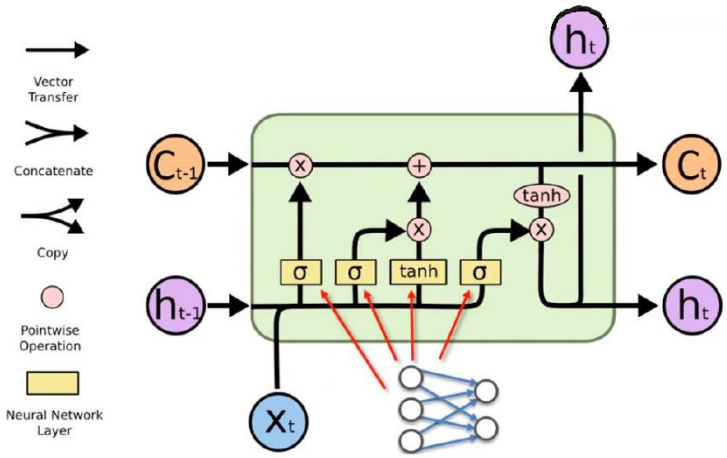
Ένα διπλής κατεύθυνσης RNN (bidirectional RNN) αποτελείται από τον συνδυασμό δύο διαφορετικών RNN, όπου το κάθε ένα επεξεργάζεται την ακολουθία με διαφορετική φορά. Το κίνητρο αυτής της τεχνικής, είναι η δημιουργία μιας καλύτερης αναπαράστασης της εξόδου h_t. Για τον υπολογισμό της αναπαράστασης αυτής χρησιμοποιείται η αναπαράσταση του απλού ανατροφοδοτούμενου δικτύου $\overrightarrow{h_t}$ αλλά και ενός αντίστροφου ανατροφοδοτούμενου δικτύου όπου τα δεδομένα εισόδου τροφοδοτούνται με την αντίθετη φορά ως προς το χρόνο t και συμβολίζεται $\overleftarrow{h_t}$. Έτσι σε κάθε χρονική στιγμή η έξοδος του δικτύου είναι $h_t = (\overrightarrow{h_t}, \overleftarrow{h_t})$ (Schuster et al., 1997).

2.3.3 Οπισθοδιάδοση στο Χρόνο (Backpropagation Through Time)

Για την εκπαίδευση των ανατροφοδοτούμενων νευρωνικών δικτύων χρησιμοποιούνται τεχνικές παρεμφερείς με αυτές που περιγράφηκαν στην περίπτωση των απλών προωθητικών νευρωνικών δικτύων. Η εκπαίδευση των ανατροφοδοτούμενων νευρωνικών δίκτυων βασίζεται σε μία επέκταση της οπισθοδιάδοσης που ονομάζεται οπισθοδιάδοση στον χρόνο (backpropagation through time, or BPTT) . Ένα βασικό βήμα για την κατανόηση του αλγορίθμου αυτού είναι το ‘ξεδίπλωμα’ του δικτύου ως προς τον άξονα του χρόνου όπως στο Σχήμα 2.12 και στη σχέση 2.17. Στην περίπτωση αυτήν, ο χρόνος εκφράζεται σαν μία διατεταγμένη σειρά υπολογισμών, συνδέοντας το ένα χρονικό βήμα με το επόμενο. Γίνεται αντιληπτό ότι όπως και τα προωθητικά δίκτυα έτσι και τα ανατροφοδοτούμενα αποτελούνται από εμφωλευμένες συναρτήσεις. Συνεπώς η προσθήκη του στοιχείου του χρόνου, απλώς επεκτείνει τη σειρά των υπολογισμών των παραγώγων της συνάρτησης κόστους ως προς τα βάρη του δικτύου μέσω του κανόνα της αλυσίδας (Mozer, 1995).

2.3.4 Δίκτυα Νευρώνων Μακράς-Βραχείας Μνήμης (Long Short-Term Memory Units ή LSTM)

Τα RNN όπως παρουσιάστηκαν προσπαθούν να ανακαλύψουν συσχετίσεις μεταξύ της εξόδου του με παρατηρήσεις αρκετών βημάτων πίσω στο χρόνο. Συνεπώς ασχέτως του πλήθους των επιπέδων ενός RNN, λόγω της αναδρομικής τους λειτουργίας αποτελεί μία βαθιά αρχιτεκτονική δικτύων. Η τελευταία παρατήρηση γίνεται πιο κατανοητή με το 'ξεδίπλωμα' του δικτύου ως προς τον χρόνο. Για το λόγο αυτό, όπως και τα απλά βαθιά προωθητικά δίκτυα πάσχουν από πρόβλημα των εξαφανιζόμενων/ανατινασσόμενων κλίσεων έτσι και τα RNN αντιμετωπίζουν το ίδιο πρόβλημα κατά την εκπαίδευση τους. Συνεπώς τα RNN εμφανίζουν προβλήματα στο να "μάθουν" μακροσκελείς αλληλουχίες παρατηρήσεων.
Στα μέσα του 1990, μια παραλλαγή των ανατροφοδοτούμενων νευρωνικών δικτύων έκαναν την εμφάνισή τους, με το όνομα Δίκτυα Νευρώνων Μακράς-Βραχειας Μνήμης (Long Short-Term Memory Units ή LSTMs). Τα δίκτυα αυτά αποτέλεσαν μία λύση στο πρόβλημα των Εξαφανιζόμενων/Ανατινασσόμενων Κλίσεων. Τα LSTMs βοηθάνε στη διατήρηση του σφάλματος το οποίο μπορεί να οπισθοδιαδοθεί μέσω του χρόνου και των επιπέδων. Με το να διατηρούμε ένα πιο σταθερό σφάλμα, δίνουμε την δυνατότητα στα επανατροφοδοτούμενα δίκτυα να μαθαίνουν για πολλά βήματα χρόνου.



Σχήμα 2.14: Δομικό στοιχείο ενός LSTM ενός επιπέδου [5] (Μετά από επεξεργασία).

Η βασική δομή μιας μονάδας LSTM βασίζεται στο κύτταρο μνήμης (memory cell) c_t το οποίο είναι υπεύθυνο για την διατήρηση της μνήμης της ακολουθίας δεδομένων. Σε κάθε βήμα το LSTM δέχεται ως είσοδο μία παρατήρηση (διάνυσμα) x_t , ενημερώνει το κύτταρο μνήμης c_t και παράγει ως έξοδο την κατάσταση ή κρυφή κατάσταση (hidden state) h_t. Η διαδικασία ενημέρωσης χρησιμοποιεί έναν μηχανισμό που βασίζεται σε πύλες (gates). Μία πύλη λήθης f_t (forget gate) που ελέγχει πόση πληροφορία από το παρελθόν c_{t-1} διατηρείται. Μία πύλη εισόδου i_t (input gate) που ελέγχει το πόση πληροφορία από την είσοδο x_t θα θα ενημερώσει την κατάσταση c_t ώστε να διατηρηθεί για τα επόμενα βήματα. Μία πύλη εξόδου o_t (output gate)

η οποία ελέγχει την ποσότητα πληροφορίας που της που τροφοδοτείται από την μνήμη στην έξοδο ως κρυφή κατάσταση h_t. Συγκεκριμένα η διαδικασία που ακολουθείται είναι η εξής :

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \quad (2.19)$$
$$f_t = \sigma\big(W_{xf}x_t + W_{hf}h_{t-1} + b_f\big), \quad (2.20)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \quad (2.21)$$
$$c_t = f_t\, c_{t-1} + i_t\, tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (2.22)$$
$$h_t = o_t\, \tanh(c_t), \quad (2.23)$$

όπου i, f, o, c είναι οι πύλες εισόδου, λήθης, εξόδου και μνήμης αντίστοιχα. Οι πίνακες των βαρών W και τα διανύσματα των τιμών κατωφλίων b είναι παράμετροι που προσαρμόζονται κατά την εκπαίδευση του δικτύου (Hochreiter et al., 1997).

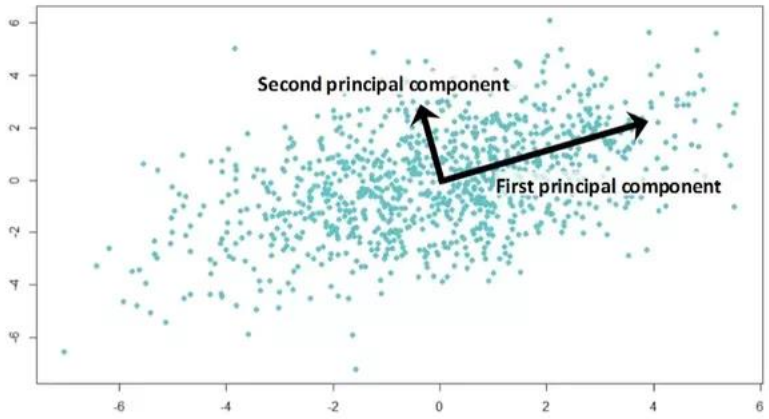
2.4 Τεχνικές Μείωσης Διάστασης

Η υψηλή διάσταση των δεδομένων περιορίζει σημαντικά τον τρόπο με τον οποίο τα δεδομένα μπορούν να επεξεργαστούν και να αναλυθούν. Στην εργασία αυτή παρουσιάστηκαν τέτοια προβλήματα όπου δεδομένα διαστάσεων μερικών χιλιάδων χρειάστηκαν να αναλυθούν. Σε αυτό το πρόβλημα δίνουν λύσεις οι τεχνικές μείωσης διάστασης. Πρόκειται για μεθοδολογίες που προβάλουν ένα σύνολο από διανύσματα υψηλής διάστασης σε ένα χώρο χαμηλότερης διάστασης με σκοπό να διατηρεί όσο περισσότερο γίνεται η δομή των δεδομένων (Roweis et al., 2000).

2.4.1 Ανάλυση Κύριων Συνιστωσών (PCA)

Σε ένα σύνολο *n* παρατηρήσεων με πλήθος μεταβλητών *p* απαιτούνται όλες οι *p* μεταβλητές για την ερμηνεία της συνολική μεταβλητότητα του συνόλου δεδομένων, συχνά όμως, η περισσότερη από αυτή τη μεταβλητότητα μπορεί να ερμηνευτεί από ένα μικρό αριθμό k κύριων συνιστωσών. Αν πράγματι συμβεί αυτό, τότε, υπάρχει (σχεδόν) τόση πληροφορία στις *k* συνιστώσες, όση υπάρχει στις *p* αρχικές μεταβλητές. Οι *k* κύριες συνιστώσες μπορούν τότε να αντικαταστήσουν τις αρχικές *p* μεταβλητές, και το αρχικό σύνολο δεδομένων που αποτελείται από *n* παρατηρήσεων των *p* μεταβλητών, μειώνεται σε ένα σύνολο δεδομένων που αποτελείται από *n* παρατηρήσεων των *k* μεταβλητών.
Η μέθοδος PCA (Ανάλυση Κύριων Συνιστωσών), αποτελεί μία γραμμική μέθοδο μείωσης της διάστασης των δεδομένων διατηρώντας το μεγαλύτερο ποσοστό της συνολικής μεταβλητότητας και συνίσταται από τον επαναπροσδιορισμό των συντεταγμένων του συνόλου δεδομένων σε ένα άλλο σύστημα συντεταγμένων μικρότερης διάστασης.
Αλγεβρικά, οι κύριες συνιστώσες είναι γραμμικοί συνδυασμοί των *p* τυχαίων μεταβλητών $X_1, X_2, \ldots, X_p$. Γεωμετρικά, αυτοί οι γραμμικοί συνδυασμοί παριστάνουν την επιλογή του νέου συστήματος συντεταγμένων που λαμβάνεται με την περιστροφή του αρχικού συστήματος με $X_1, X_2, \ldots, X_p$ ως άξονες συντεταγμένων. Οι νέοι άξονες παριστάνουν τις διευθύνσεις με την

μεγαλύτερη μεταβλητότητα και παρέχουν μια απλούστερη περιγραφή της δομής της συνδιασποράς (Hotelling, 1933).



Σχήμα 2.15: PCA σε δεδομένα δύο διαστάσεων [6].

Οι κύριες συνιστώσες εξαρτώνται μόνο από τον πίνακα συνδιασπορών $\boldsymbol{\Sigma}$ των $X_1, X_2, \ldots, X_p$. Έστω το τυχαίο διάνυσμα $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]$ που έχει πίνακα συνδιασπορών $\boldsymbol{\Sigma}$ με ιδιοτιμές $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.
Θεωρούμε τους γραμμικούς συνδυασμούς:

$$Y_i = l_i \cdot \boldsymbol{X} = l_{1i} \cdot X_1 + l_{2i} \cdot X_2 + \cdots + l_{pi} \cdot X_p \quad (2.24)$$

$$Var(Y_i) = l_i \Sigma l_i \quad (2.25)$$

$$Cov(Y_i, Y_k) = l_i \Sigma l_k, \text{ όπου } i, k = 1,2, \ldots, p \quad (2.26)$$

Οι κύριες συνιστώσες είναι αυτοί οι ασυσχέτιστοι γραμμικοί συνδυασμοί Y_i των οποίων οι διασπορές $Var(Y_i)$ είναι οι μεγαλύτερες δυνατές. Συνεπώς ορίζεται ως:

- Πρώτη κύρια συνιστώσα αποτελεί τον γραμμικό συνδυασμό $l_1 \cdot \boldsymbol{X}$ που μεγιστοποιεί την $Var(l_1 \cdot \boldsymbol{X})$ υπό την προϋπόθεση ότι $\|l_1\|$= 1.
- Η i κύρια συνιστώσα αποτελεί τον γραμμικός συνδυασμό $l_i \cdot \boldsymbol{X}$ που μεγιστοποιεί την $Var(l_i \cdot \boldsymbol{X})$ υπό την προϋπόθεση ότι $\|l_i\|$= 1 και $Cov(l_i \cdot \boldsymbol{X}, l_k \cdot \boldsymbol{X}) = 0$ για κάθε $k < i$

2.4.2 Στοχαστική Εμφύτευση Γειτόνων (t-distributed Stochastic Neighbor Embedding ή t-SNE)

Σε αντίθεση με το PCA η τεχνική t-SNE αποτελεί μία μη γραμμική πιθανοτική τεχνική μείωσης της διάστασης που αφορά την εμφύτευση δεδομένων από έναν χώρο πολλών διαστάσεων σε ένα χώρο δύο ή τριών διαστάσεων. Η συγκεκριμένη τεχνική έχει σχεδιαστεί με γνώμονα την μετέπειτα οπτικοποίηση των δεδομένων σε μια προσπάθεια να γίνει κατανοητή η δομή των

δεδομένων στον χώρο των πολλών διαστάσεων. Ουσιαστικά όμοια δεδομένα στον αρχικό χώρο μοντελοποιούνται με μεγάλη πιθανότητα σε γειτονικά σημεία στον νέο χώρο μικρότερης διάστασης και ανόμοια τοποθετούνται σε μεγάλες αποστάσεις μεταξύ τους.
Ο αλγόριθμος t-SNE επικεντρώνεται στην τοπική δομή των δεδομένων. Αυτή η ικανότητα ομαδοποίησης δειγμάτων με βάση την τοπική δομή μπορεί να είναι ωφέλιμη για την οπτική απεικόνιση ενός συνόλου δεδομένων που περιλαμβάνει πολλές πολλαπλότητες ταυτόχρονα (Maaten, 2008).
Ο αλγόριθμος περιλαμβάνει δύο κύρια στάδια. Αρχικά κατασκευάζει μία κατανομή πιθανότητας μέσω ενός πίνακα ομοιότητας (Similarity Matrix) NxN για τα N δεδομένα x_i του αρχικού χώρου μεγάλης διάστασης. Ο πίνακας ομοιότητας κατασκευάζεται από τα ζεύγη δεδομένων κατά τέτοιο τρόπο ώστε όμοια δεδομένα να έχουν μεγάλη πιθανότητα να επιλεγούν μαζί, ενώ αντίθετα τα ανόμοια δεδομένα έχουν εξαιρετικά μικρή πιθανότητα. Ο πίνακας ομοιότητας περιέχει τις πιθανότητες:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{2.27}$$

όπου

$$p_{j|i} = \frac{\exp(-\left||x_i - x_j|\right|_2^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||_2^2 / 2\sigma_i^2)} \tag{2.28}$$

Επίσης για τον υπολογισμό των τιμών ομοιότητας μεταξύ των δεδομένων είναι αναγκαία η εκτίμηση των διασπορών σ_i των Γκαουσιανών κατανομών γύρω από τα x_i για κάθε $i = 1,2, \dots, N$. Καθώς η πυκνότητα των δεδομένων από περιοχή σε περιοχή μπορεί θεωρητικά να μεταβάλλεται είναι αναγκαία η επιλογή μικρών τιμών σ_i για περιοχές με μεγάλη πυκνότητα και μεγάλων τιμών σ_i για περιοχές όπου τα δεδομένα βρίσκονται σε μεγάλες αποστάσεις μεταξύ τους. Αντίστοιχα η εντροπία της Γκαουσιανής κατανομής γύρω από το x_i αυξάνεται καθώς αυξάνεται το σ_i. Έτσι για τον υπολογισμό των σ_i ο αλγόριθμος εκτελεί μια αναζήτηση των σ_i έτσι ώστε η ποσότητα της *Σύγχυσης (*P*erplexity)* της Γκαουσιανής κατανομής γύρω από το x_i να είναι όση ορίζεται από τον χρήστη (Παράμετρος του αλγορίθμου). Ορίζεται:

$$Perplexity(p) = 2^{H(p)}, \tag{2.29}$$

όπο*υ* $H(p)$ η συνολική εντροπία της κατανομής πάνω στα δεδομένα. Σημειώνεται ότι η ποσότητα P*erplexity* ουσιαστικά λειτουργεί ως ένα μέτρο του αριθμού των γειτόνων και επιλέγεται από τον χρήστη.
Στο δεύτερο στάδιο ο αλγόριθμος ορίζει μία όμοια κατανομή πιθανοτήτων μέσω ενός πίνακα ομοιότητας πάνω σε σημεία y_i που περιέχονται σε ένα χώρο μικρότερης διάστασης (Συνήθως 2 ή 3). Ο πίνακας αυτός περιέχει τις πιθανότητες:

$$q_{ij} = \frac{f(|y_i - y_j|)}{\sum_{k \neq i} f(|y_i - y_k|)} \text{ με } f(z) = \frac{1}{1+z^2} \tag{2.30}$$

Τονίζεται ότι ενώ τα δεδομένα του αρχικού πίνακα ομοιότητας (p_{ij}) είναι πλέον σταθερά, τα σημεία του πίνακα ομοιότητας (q_{ij}) εξαρτώνται από τα σημεία y_i. Συνεπώς τελικός στόχος του αλγορίθμου είναι οι επιλογή εκείνων των y_i έτσι ώστε οι δύο πίνακες ομοιότητας (κατανομές πιθανοτήτων) να είναι όσο το δυνατό πιο 'κοντά' μεταξύ τους.
Συνεπώς για το σκοπό αυτό ο αλγόριθμος τελικά ελαχιστοποιεί την μετρική *Kullback–Leibler divergence* μεταξύ των δύο αυτών κατανομών ως προς την τοποθεσία των σημείων y_i. Η ελαχιστοποίηση γίνεται μέσω του αλγορίθμου Μείωσης Κλίσης (Gradient Decent) και η παράγωγος τη KL μετρικής ως προς τα y_i υπολογίζεται αναλυτικά ως:

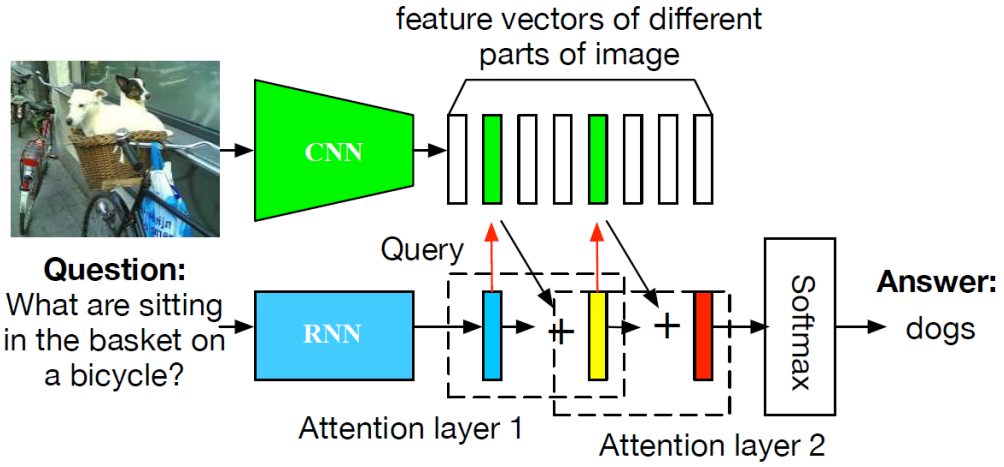
$$\frac{\partial KL(P||Q)}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \left||y_i - y_j\right|\right|^2)^{-1} \tag{2.31}$$

3. Υλοποίηση και Σχεδιασμός Μοντέλου

Στο κεφάλαιο αυτό αναλύονται τα μοντέλα που χρησιμοποιήθηκαν για το πρόβλημα απάντησης σε ερώτηση που αφορά οπτικό περιεχόμενο (VQA). Σε πρώτη φάση παρουσιάζεται μία επισκόπηση της αρχιτεκτονικής που χρησιμοποιήθηκε. Έπειτα παρουσιάζονται τα μοντέλα που σχετίζονται με την αναπαράσταση-περιγραφή της εικόνας. Στη συνέχεια παρουσιάζονται τα μοντέλα που αφορούν την αναπαράσταση της ερώτησης. Τέλος παρουσιάζεται το μοντέλο που χρησιμοποιήθηκε για την εκτίμηση της σωστής απάντησης κάνοντας χρήση των αναπαραστάσεων της εικόνας και της ερώτησης.

3.1 Επισκόπηση Αρχιτεκτονικής του Συστήματος

Το συνολικό σύστημα που χρησιμοποιήθηκε βασίζεται σε επιμέρους μοντέλα που αποτελούνται από νευρωνικά δίκτυα. Το σύστημα δέχεται σαν είσοδο μία εικόνα και μία ερώτηση σε μορφή κειμένου και έχει σαν έξοδο μία απάντηση. Συνολικά το σύστημα περιλαμβάνει τέσσερα επιμέρους μοντέλα. Τα μοντέλα αυτά αποτελούνται από το μοντέλο αναπαράστασης της εικόνας, το μοντέλο αναπαράστασης της ερώτησης, το μοντέλο που με βάση την ερώτηση και την εικόνα παράγει νέα χαρακτηριστικά και το μοντέλο ταξινόμησης που εκτιμάει την σωστή απάντηση.
Το επιμέρους μοντέλο που χρησιμοποιήθηκε για την εικόνα έχει ως στόχο να εξάγει χαρακτηριστικά από διαφορετικές περιοχές της εικόνας σε μορφή διανυσμάτων και αποτελείται από ένα βαθύ συνελικτικό νευρωνικό δίκτυο (CNN). Για την εξαγωγή των χαρακτηριστικών της εικόνας δοκιμάστηκαν δύο διαφορετικά συνελικτικά δίκτυα τα οποία θα αναλυθούν στη συνέχεια. Όσον αφορά την αναπαράσταση της ερώτησης σε μορφή διανύσματος έγινε πειραματισμός με δύο διαφορετικά ακολουθιακά δίκτυα (RNN). Έπειτα με βάση την αναπαράσταση της ερώτησης εφαρμόζεται ένας μηχανισμός εστίασης πάνω σε χαρακτηριστικά της εικόνας με σκοπό να εντοπιστούν οι περιοχές εκείνες που σχετίζονται με την ερώτηση ώστε να εξαχθούν νέα χαρακτηριστικά με τα οποία το μοντέλο ταξινόμησης θα εκτιμήσει την σωστή απάντηση (Σχήμα 3.1).



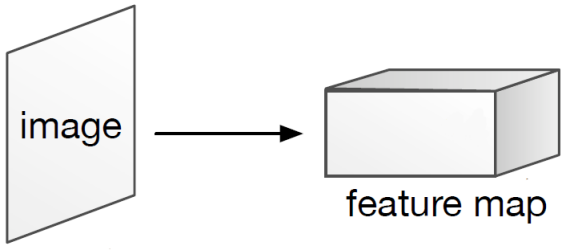
Σχήμα 3.1: Δίκτυο Πολλαπλών Εστιάσεων για VQA (Yang et al., 2016).

Το μοντέλο που εστιάζει σε περιοχές της εικόνας που σχετίζονται με την ερώτηση ονομάζεται Δίκτυο Πολλαπλών Εστιάσεων (Stacked Attention Network) (Yang et al., 2016). Το δίκτυο αυτό αποτελείται από ένα ή περισσότερα επίπεδα εστίασης. Θεωρητικά όσο περισσότερα είναι τα επίπεδα τόσο περισσότερη εστίαση επιτυγχάνεται σε περιοχές της εικόνας που σχετίζονται με την ερώτηση. Συγκεκριμένα το δίκτυο αυτό χρησιμοποιεί το διάνυσμα χαρακτηριστικών της ερώτησης στο πρώτο επίπεδο για να επιλέξει τις περιοχές που σχετίζονται με την ερώτηση. Έπειτα συνδυάζει τα διανύσματα χαρακτηριστικών αυτών των περιοχών και εκείνα της ερώτησης ώστε να δημιουργηθεί ένα νέο βελτιωμένο διάνυσμα χαρακτηριστικών το οποίο χρησιμοποιείται σαν διάνυσμα ερώτησης στο επόμενο επίπεδο εστίασης, έτσι ώστε να εστιαστούν ακόμα περισσότερο οι περιοχές που σχετίζονται με την ερώτηση. Τέλος, συνδυάζονται τα χαρακτηριστικά των περιοχών που έχουν εντοπιστεί από το τελευταίο επίπεδο εστίασης και το τελευταίο δυάνυσμα ερώτησης ώστε να γίνει εκτίμηση της απάντησης από το μοντέλο ταξινόμησης.

Σχήμα 3.2: Παράδειγμα από τις περιοχές εστίασης δύο Επιπέδων Εστίασης (Yang et al., 2016).

3.2 Αναπαράσταση Εικόνας

Τα μοντέλα που δοκιμάστηκαν για την αναπαράσταση της εικόνας εισόδου βασίζονται σε Συνελικτικά Νευρωνικά Δίκτυα (CNN). Συγκεκριμένα τα δύο διαφορετικά μοντέλα είναι ένα VGGNet-19 και ένα DenseNet-161, τα οποία χρησιμοποιήθηκαν ώστε να εξάγουν για την εικόνα εισόδου I μία απεικόνιση χαρακτηριστικών f_I.



Σχήμα 3.3: Μοντέλο αναπαράστασης εικόνας (Yang et al., 2016 Μετά από επεξεργασία).

$$f_I = CNN(I) \qquad (3.1)$$

3.2.1 Προσαρμογή του VGGNet/Densenet στο συνολικό σύστημα

Παρόλο που είναι συνήθης πρακτική σε προβλήματα μηχανικής μάθησης που σχετίζονται με την αναπαράσταση μιας εικόνας η εξαγωγή χαρακτηριστικών πάνω στο σύνολο της, στη συγκεκριμένη προσέγγιση υπολογίζονται χαρακτηριστικά πάνω σε επιμέρους περιοχές της εικόνας. Συγκεκριμένα και για τα δύο διαφορετικά μοντέλα VGGNet-19 και DenseNet-161 τα χαρακτηριστικά f_I που επιλέχθηκαν είναι εκείνα των τελευταίων συγκεντρωτικών επιπέδων (pooling layers), όπως φαίνονται στις εικόνες 2.7 και 2.11, και όχι τα χαρακτηριστικά που παράγονται από το αμέσως επόμενο πλήρως συνδεδεμένο επίπεδο (Fully Connected Layer) όπως συνηθίζεται. Ο λόγος είναι ότι τα χαρακτηριστικά από τα συγκεντρωτικά επίπεδα διατηρούν την χωρική πληροφορία έναντι εκείνων των πλήρως διασυνδεδεμένων επιπέδων τα οποία αποτελούν χαρακτηριστικά στο σύνολο της εικόνας. Συνεπώς η εικόνα χωρίζεται σε ίσα κομμάτια και πάνω σε κάθε ένα από αυτά εξάγονται χαρακτηριστικά σε μορφή διανυσμάτων.

Συγκεκριμένα τα βήματα που ακολουθήθηκαν ήταν τα εξής:

- Αρχικά γίνεται τροποποίηση της κλίμακας των εικόνων ως προ το ύψος και πλάτος (rescaling) με δύο διαφορετικούς τρόπους:
    a) 224x224
    b) 448x448

- Έπειτα επειδή τα προ-εκπαιδευμένα δίκτυα VGGNet και DenseNet δέχονται σαν είσοδο μια έγχρωμη *RGB (Red-Green-Blue)* εικόνα και αναμένουν συγκεκριμένες κατανομές των RGB τιμών έγινε η εξής τροποποίηση:
    a) Για το VGGNet-19 οι τιμές των εικόνων εισόδου πρέπει να είναι γύρω από το 0. Συνεπώς αφαιρέθηκαν οι μέσοι όροι των τιμών όλων των εικονοστοιχείων (pixels) όλων των εικόνων για κάθε χρωματικό κανάλι, δηλαδή σε κάθε εικόνα αφαιρέθηκαν οι τιμές [103.939, 116.779, 123.68] (Οι αρχικές RBG εικόνες πρέπει να παίρνουν τιμές στο σύνολο $[0, 255]^3$).
    b) Για το DenseNet-161 οι τιμές των εικόνων εισόδου κανονικοποιούνται με μέσο όρο [0.485, 0.456, 0.406] και τυπική απόκλιση [0.229, 0.224, 0.225] (Οι αρχικές RBG εικόνες πρέπει να παίρνουν τιμές στο σύνολο $[0, 1]^3$).
- Στην συνέχεια γίνεται εξαγωγή των χαρακτηριστικών από το τελευταίο συγκεντρωτικό επίπεδο και ανάλογα με το μοντέλο και τις διαστάσεις των εικόνων εισόδου:
    a) Για VGGNet-19 με 224x224 εικόνα οι διαστάσεις των χαρακτηριστικών είναι 512x7x7
    b) Για VGGNet-19 με 448x448 εικόνα οι διαστάσεις των χαρακτηριστικών είναι 512x14x14
    c) Για DenseNet-161 με 224x224 εικόνα οι διαστάσεις των χαρακτηριστικών είναι 2208x7x7
    d) Για DenseNet-161 με 448x448 εικόνα οι διαστάσεις των χαρακτηριστικών είναι 2208x14x14

  Ουσιαστικά οι διαστάσεις 7x7 (όμοια οι 14x14) αποτελούν τον αριθμό των περιοχών στην εικόνα και οι διαστάσεις 512 για το VGGNet (όμοια 2208 για το DenseNet) αποτελούν τα χαρακτηριστικά της κάθε μίας περιοχής. Παρατηρείται επίσης ότι το κάθε διάνυσμα χαρακτηριστικών για όλες τις περιπτώσεις αντιστοιχεί σε μια περιοχή 32x32 εικονοστοιχείων της εικόνας εισόδου. Ορίζεται λοιπόν με f_i το διάνυσμα χαρακτηριστικών της κάθε περιοχής εικόνας όπου $i \in [0,48]$ στην περίπτωση όπου η είσοδος είναι διάστασης 224x224 και $i \in [0,195]$ στην περίπτωση όπου η είσοδος είναι διάστασης 448x448.
- Τέλος για λόγους μοντελοποίησης γίνεται χρήση ενός απλού επιπέδου Perceptron ώστε κάθε διάνυσμα χαρακτηριστικών f_i να μετασχηματιστεί σε ένα νέο διάνυσμα ίσης διάστασης με τα διανύσματα χαρακτηριστικών τα οποία παράγονται από το μοντέλο αναπαράστασης της ερώτησης (περιγράφονται στην επόμενη ενότητα):

$$u_I = tahn(W_I f_I + b_I), \qquad (3.2)$$

όπου u_I είναι ο ένας πίνακας και κάθε $i - οστή$ κολόνα u_i αποτελεί την τελική αναπαράσταση την περιοχής i (Yang et al., 2016).

3.3 Αναπαράσταση Λέξεων

Τα LSTM δίκτυα και γενικότερα τα βαθιά Αναδρομικά Νευρωνικά Δίκτυα (RNN) έχουν εμφανίσει εξαιρετικά αποτελέσματα στο να αναπαριστούν το σημασιολογικό νόημα ενός κειμένου. Για το λόγο αυτό για την αναπαράσταση της ερώτησης εισόδου δοκιμάστηκαν δύο διαφορετικά μοντέλα τα οποία βασίζονται σε LSTM δίκτυα. Το πρώτο μοντέλο βασίζεται σε εξαγωγή χαρακτηριστικών από *Σακούλες Λέξεων (Bag-of-Words)* (Harris, 1954) του συνόλου των λέξεων των ερωτήσεων. Το δεύτερο μοντέλο κάνει χρήση του μοντέλου *ELMo* όπως έχει προταθεί από τους Peters et al. (2018) το οποίο βασίζεται σε διπλής κατευθύνσεως LSTM (bi-LSTM). Στην επόμενη ενότητα παρουσιάζεται η προεπεξεργασία που εφαρμόστηκε στο σύνολο των λέξεων των ερωτήσεων και των απαντήσεων, καθώς και στη συνέχεια αναλύονται οι δύο διαφορετικές τεχνικές που χρησιμοποιήθηκαν για την αναπαράσταση της ερώτησης.

3.3.1 Προεπεξεργασία του Συνόλου Ερωτήσεων και Απαντήσεων

Το σύνολο δεδομένων εκπαίδευσης αποτελείται από εικόνες και από τις αντίστοιχες ερωτήσεις και απαντήσεις τους στην αγγλική γλώσσα. Λόγω του μεγάλου πλήθους ερωτήσεων και απαντήσεων αλλά και λόγω του ότι οι 1000 απαντήσεις με τη μεγαλύτερη συχνότητα στο σύνολο εκπαίδευσης εμφανίζονται σε ποσοστό 87,47% των ερωτήσεων (τα οποία θα αναλυθούν περαιτέρω στο επόμενο κεφάλαιο) επιλέχθηκαν μόνο οι ερωτήσεις που αντιστοιχούν σε αυτές τις 1000 απαντήσεις για την εκπαίδευση των μοντέλων. Οι λέξεις οι οποίες χρησιμοποιήθηκαν προκειμένου να μπορεί επιλέξει το τελικό μοντέλο την σωστή απάντηση για το ζεύγος εικόνας-ερώτησης προέρχονται λοιπόν από το σύνολο των 1000 αυτών λέξεων, το οποίο λειτουργεί σαν ***λεξικό (vocabulary)*** για τις πιθανές απαντήσεις. Συνολικά, το λεξικό απαντήσεων αποτελείται από *1000 λέξεις*, ταξινομημένες σε φθίνουσα σειρά, με βάση τη συχνότητα εμφάνισής τους στο σύνολο δεδομένων εκπαίδευσης. Για λόγους μοντελοποίησης και για τη χρήση των λέξεων στις διαδικασίες της εκπαίδευσης του μοντέλου, οι απαντήσεις αναπαρίστανται με ακέραιους αριθμούς. Οι ακέραιοι αριθμοί αυτοί επιλέχθηκαν να είναι η σειρά της απάντησης ως προς τον αριθμό εμφάνισης της. Η μορφή του λεξικού των απαντήσεων είναι εκείνη που παρουσιάζεται στον επόμενο πίνακα 3.1.

Αρ. Σειράς	Αρ. Εμφάνισεων	Απάντηση
1	84978	yes
2	82516	no
3	12540	1
4	12215	2
5	8916	white
6	6536	3
...	...	...
999	27	roman numerals
1000	27	peace

Πίνακας 3.1: Λεξικό απαντήσεων το οποίο βασίζεται στις 1000 πιο συχνές απαντήσεις.

Έτσι για παράδειγμα, η απάντηση ‘yes’ απεικονίζεται στον αριθμό 1 και η απάντηση ‘2’ στον αριθμό 4.

3.3.2 Εμφύτευση Λέξεων Ερωτήσεων σε Διανύσματα (Word Embedding Vectors)

Στην ενότητα αυτή παρουσιάζεται το πρώτο μοντέλο που δοκιμάστηκε για την αναπαράσταση των ερωτήσεων καθώς και η αντίστοιχη επεξεργασία που χρειάστηκε να επιτευχθεί στο σύνολο των ερωτήσεων.
Η διαδικασία για την κατασκευή του λεξικού ερωτήσεων το οποίο βασίζεται στο σύνολο των ερωτήσεων είναι όμοια με εκείνη που περιγράφηκε στην προηγούμενη ενότητα για τη δημιουργία του λεξικού απαντήσεων. Το *λεξικό ερωτήσεων που* κατασκευάστηκε αποτελείται από όλες τις διαφορετικές λέξεις που εμφανίζονται στις ερωτήσεις οι οποίες αντιστοιχούν στις επιλεγμένες 1000 πιο συχνές απαντήσεις του συνόλου δεδομένων εκπαίδευσης, χωρίς κάποιον περιορισμό ως προς την συχνότητα εμφάνισής τους. Αναφέρεται ότι το ανώτατο όριο στο μήκος της κάθε ερώτησης θεωρείται το 26, δηλαδή για τις ερωτήσεις που αποτελούνται από περισσότερες από 26 λέξεις επιλέγονται μόνο οι 26 πρώτες λέξεις της. Η επιλογή έγινε με βάση κάποια στατιστικά στοιχεία που παρουσιάζονται σε επόμενη ενότητα σχετικά με το σύνολο των δεδομένων. Για την διευκόλυνση δημιουργίας του λεξικού, οι ερωτήσεις μετασχηματίστηκαν σε λίστες από λέξεις, χωρισμένες με κόμμα μεταξύ τους με βάση τους εξής χαρακτήρες - . " ' , : ? ! $ # @ ~ ( ) * & \ ^ % ; [ ] / + < > \n, όπως φαίνεται στο παράδειγμα που ακολουθεί:

'How many slices of pizza are there?' → ['How', 'many', 'slices', 'of', 'pizza', 'are', 'there', '?']

Συνολικά, το λεξικό ερωτήσεων αποτελείται από *12915* λέξεις. Όμοια με το λεξικό απαντήσεων οι λέξεις των ερωτήσεων αναπαρίστανται με ακέραιους αριθμούς για την χρήση τους στις διαδικασίες εκπαίδευσης του μοντέλου. Έτσι δημιουργήθηκε μία απεικόνιση μεταξύ των λέξεων του λεξικού και των ακεραίων αριθμών από το 0 έως το 12914.
Στη συνέχεια η αναπαράσταση αυτή μετασχηματίστηκε σε *one-hot* διανύσματα τα οποία έχουν αρκετά μεγάλες διαστάσεις, των οποίων ο αριθμός είναι ίσος με το μέγεθος του λεξικού. Ένα παράδειγμα παρουσιάζεται στον επόμενο πίνακα 3.2.

Αναπαράσταση με Ακέραιους		Αναπαράσταση με One-hot διανύσματα															
7		0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0
4		0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0
11		0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
11		0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
0		1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
12914	→	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1
22		0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0
14		0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
17		0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
11		0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
3		0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0

Πίνακας 3.2: Παράδειγμα μετασχηματισμού σε one-hot διανύσματα.

Στην φάση αυτή γίνεται εμφανές ότι η κάθε λέξη του λεξικού ερωτήσεων απεικονίζεται σε έναν άξονα στον 12915-διάστασης χώρο το οποίο παρουσιάζει τα εξής προβλήματα:

α) δεν λαμβάνεται υπόψιν η σημασιολογική ομοιότητα μεταξύ των λέξεων (καθώς όλες απέχουν μεταξύ τους ίσες αποστάσεις)
β) η κάθε λέξη απεικονίζεται σε ένα διάνυσμα εξαιρετικά μεγάλης διάστασης το οποίο δημιουργεί προβλήματα στους υπολογισμούς και στην μνήμη που απαιτούν οι αλγόριθμοι.
γ) δεν λαμβάνεται υπόψιν η σειρά εμφάνισής τους σε μία ερώτηση

Για την επίλυση των προβλημάτων α) και β) γίνεται χρήση των *διανυσμάτων εμφύτευσης σταθερού μήκους (word embedding vectors)* και για το γ) το αναδρομικό δίκτυο LSTM που παρουσιάζεται σε επόμενη ενότητα. *Τα διανύσματα εμφύτευσ*ης είναι μια μέθοδος κατασκευής, χαμηλών διαστάσεων, διανυσμάτων αναπαράστασης λέξεων, τα οποία διατηρούν μια σημασιολογική ομοιότητα περιεχομένου μεταξύ των λέξεων. Η βασική ιδέα των διανυσμάτων αυτών είναι να μετατρέπουν τις λέξεις σε διανύσματα σταθερού μήκους, τα οποία περιέχουν πραγματικούς αριθμούς και όχι απλά 0 και 1. Έτσι δημιουργούν αναπαραστάσεις των λέξεων, των οποίων το μέγεθος είναι ανεξάρτητο από το μέγεθος του λεξικού και συνεπώς μπορούν να έχουν αρκετά μικρότερο αριθμό διαστάσεων. Ένα υποθετικό παράδειγμα στον χώρο των 3ών διαστάσεων με μόλις τρεις λέξεις είναι το ακόλουθο. Κάθε μία από τις λέξεις “dog”, “cat” και “pizza” απεικονίζεται στις εξής τρεις διαστάσεις/διανύσματα (one-hot διανύσματα):

- v("dog") = (1, 0, 0)
- v("cat") = (0, 1, 0)
- v("pizza") = (0, 0, 1)

Παρατηρείται λοιπόν ότι δεν μπορεί να αποτυπωθεί το γεγονός ότι οι λέξεις “dog” και “cat” είναι όμοιες. Δηλαδή η συγκεκριμένη αναπαράσταση θεωρεί εξίσου όμοιες μεταξύ τους τις τρεις αυτές λέξεις.
Αντίθετα με τα *διανύσματα εμφύτευσ*ης τα διανύσματα αναπαράστασης των λέξεων “dog”, “cat” και “pizza” μπορούν για παράδειγμα να είναι:

- v("dog") = (0.8, 0.2)
- v("cat") = (0.7, 0.1)
- v("pizza") = (0.1, 0.7)

Όπου ουσιαστικά η πρώτη διάσταση συμπυκνώνει την έννοια του ζώου ενώ η δεύτερη την έννοια του φαγητού.

Η αναπαράσταση του λεξικού των ερωτήσεων λοιπόν σε *διανύσματα εμφύτευσ*ης γίνεται μέσω μιας απεικόνισης που ονομάζεται *πίνακας εμφύτευσης* (embedding matrix). Ο πίνακας αυτός αποτελείται από βάρη τα όποια προσαρμόζονται κατά την διαδικασία εκπαίδευσης του μοντέλου μέσω του αλγορίθμου οπισθοδιάδοσης του λάθους.
Δοθέντος λοιπόν μίας ερώτησης $\mathrm{q} = [q_1, \dots, q_T]$ όπου $q_t \in \{0,1\}^{12915}$ είναι ένα one-hot διάνυσμα αναπαράστασης μίας λέξης στην θέση t και $T \leq 26$, η εμφύτευση της λέξης q_t επιτυγχάνεται μέσω της απεικόνισης W_e:

$$x_t = W_e q_t, t \in \{1, 2, \dots, T\} \qquad (3.3)$$

όπου $W_e \in \mathbb{R}^{N_e} \times \mathbb{R}^{12915}$ και N_e είναι το μέγεθος των διανυσμάτων εμφύτευσης το όποιο είναι παραμετροποιήσιμο. Συνεπώς οι 12915 διαστάσεις των αρχικών λέξεων απεικονίζονται σε $N_e \ll 12915$. Να σημειώσουμε ότι για λόγους βελτιστοποίησης στην πραγματικότητα η βιβλιοθήκη Pytorch που έχει επιλεχθεί για την υλοποίηση των μοντέλων παρέχει συγκεκριμένες δομές που ονομάζονται Επίπεδα Εμφύτευσης (Embedding Layers) και ουσιαστικά για καλύτερη διαχείριση της μνήμης δεν απαιτεί τα προς εμφύτευση διανύσματα εισόδου q_t να είναι σε μορφή one-hot αλλά σαν ακέραιοι αριθμοί και λειτουργεί σαν Πίνακας Αναζήτησης (Look-up Table). Παρόλα αυτά για λόγους μοντελοποίησης και φορμαλισμού κρίνεται απαραίτητη η προηγούμενη διατύπωση με one-hot διανύσματα (Mikolov et al., 2013) (Yang et al., 2016).
Η αναπαράσταση των λέξεων μιας ερώτησης μέσω των εμφυτευμένων διανυσμάτων x_t δίνει στο μοντέλο την ικανότητα να αναπαραστήσει τη σημασιολογική ομοιότητα μεταξύ των λέξεων μέσω ενός διανύσματος μικρότερης διάστασης. Δοθέντος λοιπόν των διανυσμάτων x_t στόχος πλέον είναι να υπάρξει μια συνολική αναπαράσταση της ερώτησης σε μορφή διανύσματος. Για το σκοπό αυτό γίνεται χρήση ενός δικτύου LSTM το οποίο αποτελεί σημαντικό μέρος του συνολικού συστήματος και παρουσιάζεται σε επόμενη ενότητα, αφού πρώτα παρουσιαστεί ένας εναλλακτικός τρόπος για την εμφύτευση των διανυσμάτων λέξεων.

3.3.3 Εμφύτευση Λέξεων Ερωτήσεων σε Διανύσματα με Χρήση Γλωσσικού Μοντελου (Word Embedding Vectors) – Μοντέλο ELMo

Η εμφύτευση των λέξεων ερωτήσεων σε διανύσματα όπως παρουσιάστηκε στην προηγούμενη ενότητα, πάσχει από κάποια προβλήματα και δέχεται κάποιες απλουστεύσεις για την μοντελοποίηση των λέξεων. Η κάθε λέξη απεικονίζεται σε ένα μοναδικό διάνυσμα ανεξάρτητα από τη θέση που παρουσιάζεται και από τον τρόπο που χρησιμοποιείται καθώς οι εμφυτεύσεις βασίζονται σε έναν Πίνακα Αναζήτησης (Look-up table). Υπάρχουν όμως φορές που μία λέξη ή φράση έχει πολλαπλές και συσχετιζόμενες σημασίες. Για παράδειγμα η λέξη **γράμμα** μπορεί να σημαίνει:

1. στοιχείο του αλφαβήτου: «Το πρώτο **γράμμα** στο αλφάβητο της Ελληνικής γλώσσας είναι το άλφα (Α)»
2. επιστολή : «Έστειλα ένα **γράμμα** στον δήμαρχο»
3. νομικός όρος: «Το **γράμμα** του νόμου»

Το φαινόμενο αυτό ονομάζεται πολυσημία. Ένα άλλο παράδειγμα στην Αγγλική γλώσσα είναι αν υπάρχει η λέξη hot" πριν από τη λέξη "dog". Στην περίπτωση αυτή η λέξη "dog" εμφανίζεται πιο όμοια με την λέξη "pizza" παρά με την λέξη "cat".
Με βάση τα παραπάνω η σημασία μιας λέξης εξαρτάται από τις προηγούμενες αλλά και τις επόμενες λέξεις εντός μια πρότασης. Γίνεται κατανοητό ότι η ανθρώπινη γλώσσα είναι αρκετά δυναμική και δεν μπορεί να μοντελοποιηθεί μόνο από Πίνακες Αναζήτησης (Look-up tables). Είναι επιθυμητό λοιπόν ένα μοντέλο που θα λαμβάνει υπόψιν ολόκληρη την πρόταση και έπειτα

θα παράγει αναπαραστάσεις των λέξεων που την απαρτίζουν, κάτι το οποίο θα μπορούσε να μοντελοποιεί καλύτερα τα φαινόμενα αυτά.

Μια τέτοια προσπάθεια αποτελεί και το μοντέλο ELMo (Embeddings from Language Models) (Peters et al., 2018). Το ELMo αποτελεί ένα δυναμικό Γλωσσικό Μοντέλο με την έννοια ότι οι εμφυτεύσεις για την ίδια λέξη είναι διαφορετικές ανάλογα με το συνολικό περιεχόμενο της πρότασης στην οποία βρίσκονται. Ένα Γλωσσικό Μοντέλο είναι ένα στατιστικό μοντέλο το οποίο υπολογίζει κατανομές πιθανοτήτων σε ακολουθίες λέξεων. Έτσι δοθέντος κάποιες προτάσεις ή ακολουθίες λέξεων υπολογίζει τις πιθανότητες εμφάνισης επόμενων λέξεων. Το ELMo κάνει χρήση βαθιών αναδρομικών νευρωνικών δικτύων RNN και συγκεκριμένα διπλής κατεύθυνσης LSTM (biLSTM) για την εκπαίδευση του επίσης διπλής κατεύθυνσης Γλωσσικού Μοντέλου (biLM).

Σχήμα 3.4: Ο Elmo (χαρακτήρας του Muppet Show).

Γενικότερα για τα Γλωσσικά μοντέλα διπλής κατεύθυνσης, δοθείσας μιας ακολουθίας από N λέξεις $(t_1, t_2, \dots, t_N)$ ένα Εμπρόσθιο Γλωσσικό Μοντέλο υπολογίζει την πιθανότητα αυτής της ακολουθίας λέξεων μοντελοποιώντας την πιθανότητα εμφάνισης της t_k λέξης δοθεισών των προηγούμενων $(t_1, t_2, \dots, t_{k-1})$ λέξεων:

$$P(t_1, t_2, \dots, t_N) = \prod_{k=1}^{N} P(t_k | t_1, t_2, \dots, t_{k-1}) \qquad (3.4)$$

Αντίστοιχα ένα Οπίσθιο Γλωσσικό Μοντέλο υπολογίζει την πιθανότητα της ίδιας ακολουθίας μοντελοποιώντας την πιθανότητα εμφάνισης της t_k δοθεισών των επόμενων $(t_{k+1}, t_{k+2}, \dots, t_N)$ λέξεων:

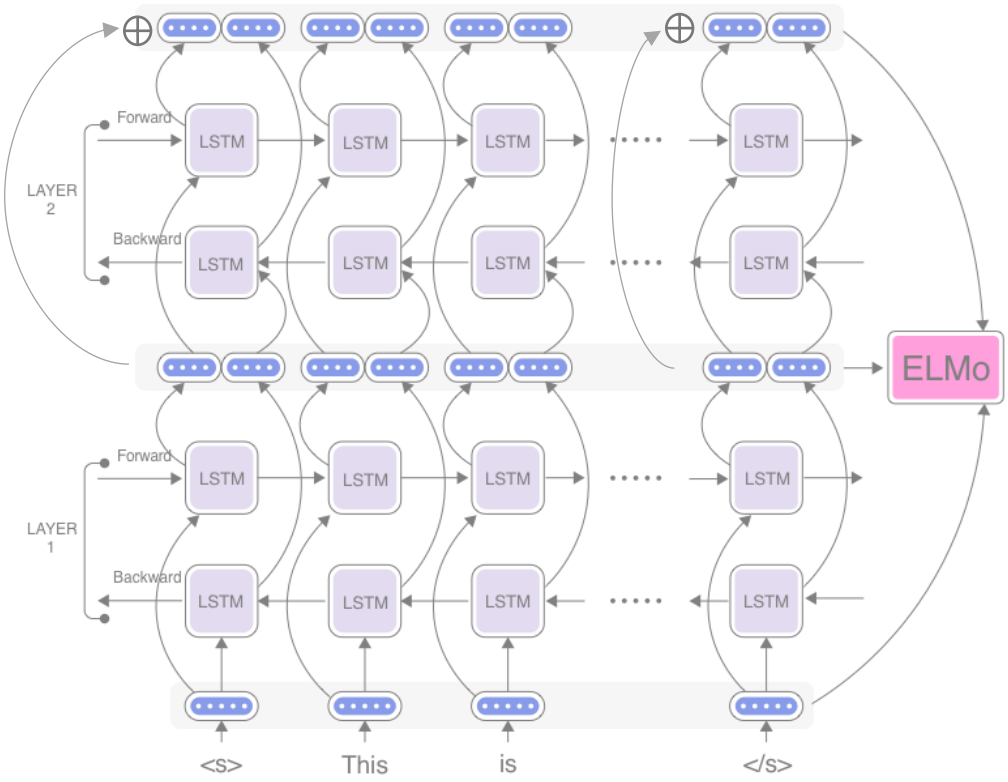
$$P(t_1, t_2, \dots, t_N) = \prod_{k=1}^{N} P(t_k | t_{k+1}, t_{k+2}, \dots, t_N) \qquad (3.5)$$

Το ELMo κατά την εκπαίδευση του συνδυάζει και τα δύο αυτά Γλωσσικά Μοντέλα με στόχο να μεγιστοποιήσει την ποσότητα:

$$\sum_{k=1}^{N} \begin{pmatrix} logP(t_k | t_1, t_2, \dots, t_{k-1}; \Theta_x, \vec{\Theta}_{LSTM}, \Theta_s) + \\ logP(t_k | t_{k+1}, t_{k+2}, \dots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \end{pmatrix} \qquad (3.6)$$

Όπου Θ_x είναι η αναπαράσταση των λέξεων πριν δοθούν στο LSTM μοντέλο, $\overleftarrow{\Theta}_{LSTM}$ και $\vec{\Theta}_{LSTM}$ είναι οι παράμετροι των LSTM μοντέλων οι οποίοι είναι διαφορετικοί ως προς κάθε κατεύθυνση και Θ_s είναι το Softmax Επίπεδο που χρησιμοποιείται για την εκτίμηση των πιθανοτήτων. Οι παράμετροι αυτοί και η αρχιτεκτονική του ELMo αναλύονται στην συνέχεια (Peters et al., 2018).

Όπως αναφέρθηκε το ELMo χρησιμοποιεί ένα Γλωσσικό Μοντέλο το οποίο ονομάζεται Διπλής Κατεύθυνσης Γλωσσικό Μοντέλο (biLM) το οποίο συνδυάζει τα Εμπρόσθιο και Οπίσθιο Γλωσσικά Μοντέλα που παρουσιάστηκαν. Στο επόμενο Σχήμα 3.5 παρουσιάζεται η γενικότερη αρχιτεκτονική του μοντέλου ELMo.



Σχήμα 3.5: Το ELMo εξάγει χαρακτηριστικά τα οποία αποτελούνται από τις εσωτερικές αναπαραστάσεις ενός πολύεπίπεδου biLM [7].

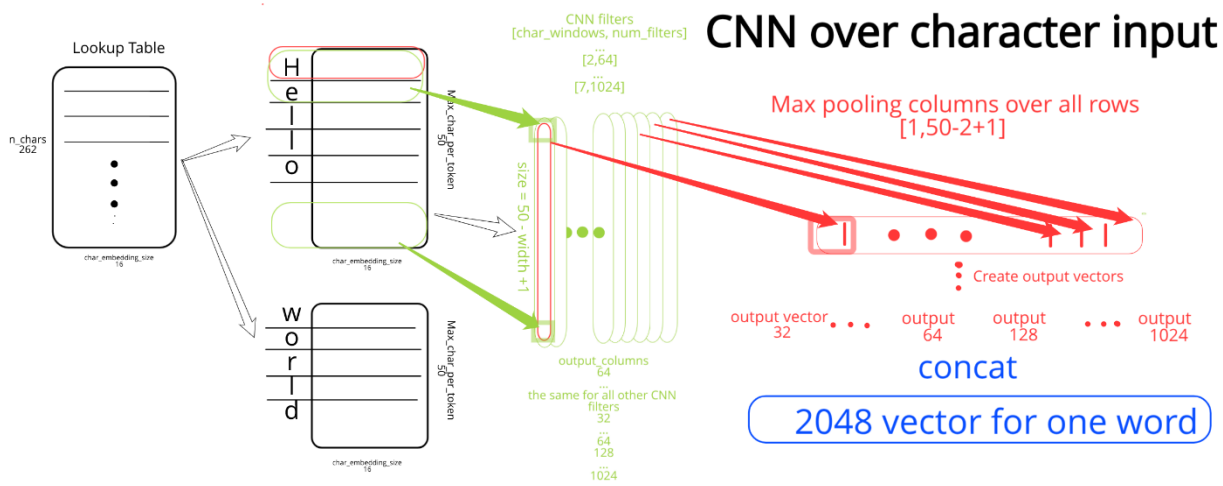
Το biLM μοντέλο που χρησιμοποιείται από το ELMo αποτελείται ουσιαστικά από ένα μοντέλο biLSTM δύο επιπέδων. Παρόλα αυτά υπάρχει μία σημαντική διαφορά μεταξύ του μοντέλου biLSTM και του biLM που χρησιμοποιείται από το ELMo. Το biLM του ELMo ουσιαστικά χρησιμοποιεί δύο διαφορετικά LSTM Γλωσσικά Μοντέλα, ένα που σαρώνει την πρόταση από την αρχή προς το τέλος και ένα αντίστροφα και έπειτα απλά συνενώνει σειριακά (concatenation) τις εξόδους του κάθε επιπέδου. Από τη άλλη ένα biLSTM κάνει κάτι παραπάνω από μια απλή συνένωση δύο διαφορετικών LSTM. Σε ένα biLSTM οι έξοδοι του κάθε επιπέδου συνενώνονται πριν δοθούν ως είσοδοι στο επόμενο επίπεδο, ενώ σε ένα biLM, οι έξοδοι του κάθε επιπέδου συνενώνονται από δύο ανεξαρτήτως εκπαιδευμένα δίκτυα (LM - Language Model).
Ένα τέτοιο δίκτυο δέχεται σαν είσοδο μία πρόταση λέξη-λέξη, έτσι η εσωτερική του κατάσταση (Hidden state) ενημερώνεται, η οποία αποτελεί το “περιεχόμενο” της πρότασης που έχει προσπελαστεί μέχρι στιγμής. Συνεπώς η εσωτερική κατάσταση κατά το προς τα εμπρός πέρασμα σε μία λέξη αντανακλά την ίδια την λέξη και το οτιδήποτε πριν από αυτή, ενώ κατά το προς τα πίσω πέρασμα αντανακλά την ίδια την λέξη και οτιδήποτε μετά από αυτή. Οι εσωτερικές καταστάσεις (Hidden states) και για τα δύο αυτά περάσματα για κάθε επίπεδο ξεχωριστά έπειτα συνενώνονται και παράγουν μια ενδιάμεση αναπαράσταση της κάθε λέξης. Σημειώνεται ότι η κάθε αναπαράσταση συμπυκνώνει την έννοια της λέξης, αλλά ταυτόχρονα και το πως

χρησιμοποιείται με βάση την υπόλοιπη πρόταση. Το ELMo χρησιμοποιεί δύο επίπεδα, έτσι τα διανύσματα αναπαραστάσεων που παράγονται από το 1ο επίπεδο δίδονται σαν είσοδο στο 2ο. Κατά τη διαδικασία αυτή παράγονται πιο αφηρημένες αναπαραστάσεις των λέξεων, π.χ από αναπαράσταση φράσεων στο 1ο και θεματικής ή συναισθήματος στο 2ο. Επίσης το biLM περιέχει παραλείπουσες συνδέσεις από το 1ο στο 2ο επίπεδο όπως περιγράφηκαν σε προηγούμενο κεφάλαιο. Ακόμα το κάθε επίπεδο περιέχει 4096 LSTM μονάδες και απεικονίζει την έξοδο του σε 512 χαρακτηριστικά (Hagiwara 2019) (Mihail 2018).

Πέρα από τη γενικότερη αρχιτεκτονική του ELMo που βασίζεται σε LSTM, το μοντέλο χρησιμοποιεί επίσης ένα Συνελικτικό Νευρωνικό Δίκτυο για την μοντελοποίηση των χαρακτήρων των λέξεων και τον υπολογισμό των εμφυτεύσεων των λέξεων που δίδονται ως είσοδοι στο 1ο επίπεδο του biLM, όπως έχει προταθεί αρχικά από τους (Kim, 2014), (Kim et al., 2015) και έπειτα από τους (Wieting et al., 2016) (CHARAGRAM) και (Bojanowski et al., 2017) (fastText) (Peters et al., 2018).

Η είσοδος υπολογίζεται μόνο στους χαρακτήρες και σε συνδυασμούς τους και δεν βασίζεται σε Πίνακες Αναζήτησης (Look-up tables) ολόκληρων των λέξεων. Η συγκεκριμένη πρακτική μπορεί να εντοπίσει την εσωτερική δομή της λέξης και συνεπώς να δημιουργήσει μια αναπαράσταση όπου για παράδειγμα οι λέξεις "dog" και "doggy" συσχετίζονται. Επίσης επειδή δεν βασίζεται σε Σακούλες Λέξεων (Bag of Words) μπορεί να χρησιμοποιηθεί ακόμα και για άγνωστες λέξεις που δεν έχουν συμπεριληφθεί κατά τη διαδικασία της εκπαίδευσης.

Η διαδικασία που ακολουθεί το μοντέλο χαρακτήρων εμφανίζεται στο επόμενο Σχήμα 3.6.



Σχήμα 3.6: Το Μοντέλο Εμφύτευσης Χαρακτήρων που χρησιμοποιεί το ELMo [8].

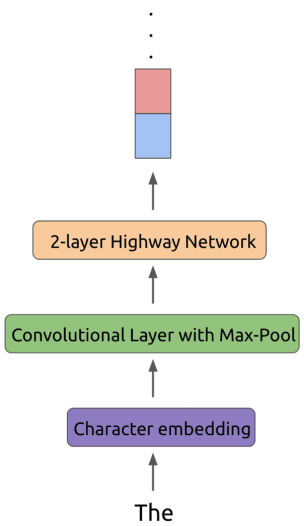
Συγκεκριμένα το μοντέλο για τους χαρακτήρες είναι κατασκευασμένο πάνω σε ένα Λεξικό χαρακτήρων – Σακούλα Χαρακτήρων μεγέθους 262 το οποίο αποτελείται από τα εξής:

- 0-255 για τους χαρακτήρες σε byte της κωδικοποίησης utf-8 που εμφανίζονται στο διπλανό Σχήμα 3.7 και εμπεριέχουν το Αγγλικό αλφάβητο
- 256-262 σε ειδικούς χαρακτήρες σχετικά με την αρχή και τέλος της κάθε πρότασης, για την αρχή και το τέλος της κάθε λέξης και για γέμισμα λέξεων (padding).

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F	
0-	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI	0-
1-	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US	1-
2-		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/	2-
3-	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	3-
4-	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	4-
5-	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_	5-
6-	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	6-
7-	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL	7-
8-	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3	8-
9-	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC	9-
A-		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	¯	A-
B-	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿	B-
C-	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	C-
D-	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß	D-
E-	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	E-
F-	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ	F-
	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F	

Σχήμα 3.7: Οι πρώτοι 256 UTF-8 χαρακτήρες [9].

Έπειτα κάθε γράμμα εμφυτεύεται σε ένα διάνυσμα διάστασης 16 όπως όμοια είχε περιγραφεί στην σχέση (3). Στη συνέχεια για κάθε λέξη σχηματίζεται ένας πίνακας διάστασης 16x50 όπου 50 είναι το μέγιστο μήκος χαρακτήρων (για κάθε λέξη μικρότερου μήκους γίνεται γέμισμα με βάση την εμφύτευση του ειδικού χαρακτήρα). Στον σχηματισμένο αυτό πίνακα κάθε λέξης εφαρμόζονται τα εξής συνελικτικά φίλτρα [1, 32], [2, 32], [3, 64], [4, 128], [5, 256], [6, 512] και [7, 1024] όπου η πρώτη διάσταση αναφέρεται στο μέγεθος του φίλτρου δηλαδή ανά πόσα γράμματα (n-gram) και η δεύτερη για το πλήθος των διαφορετικών φίλτρων που παράγονται. Συνολικά λοιπόν παράγονται 2048 φίλτρα και αφού εφαρμοστεί ένα Συγκεντρωτικό Επίπεδο Μεγίστου (Max Pooling) στην έξοδο του κάθε φίλτρου προκύπτει η τελική αναπαράσταση για κάθε λέξη. Πριν από την είσοδο του κάθε εμφυτευμένου διανύσματος στο biLM χρησιμοποιούνται δύο επίπεδα από Δίκτυα Λεωφόρων (Highway Networks) ώστε να επιτευχθεί ομαλότερη ροή και επιλογή της πληροφορίας και έπειτα ένας απλός γραμμικός μετασχηματισμός ώστε οι 2048 διαστάσεις να απεικονιστούν σε 512.

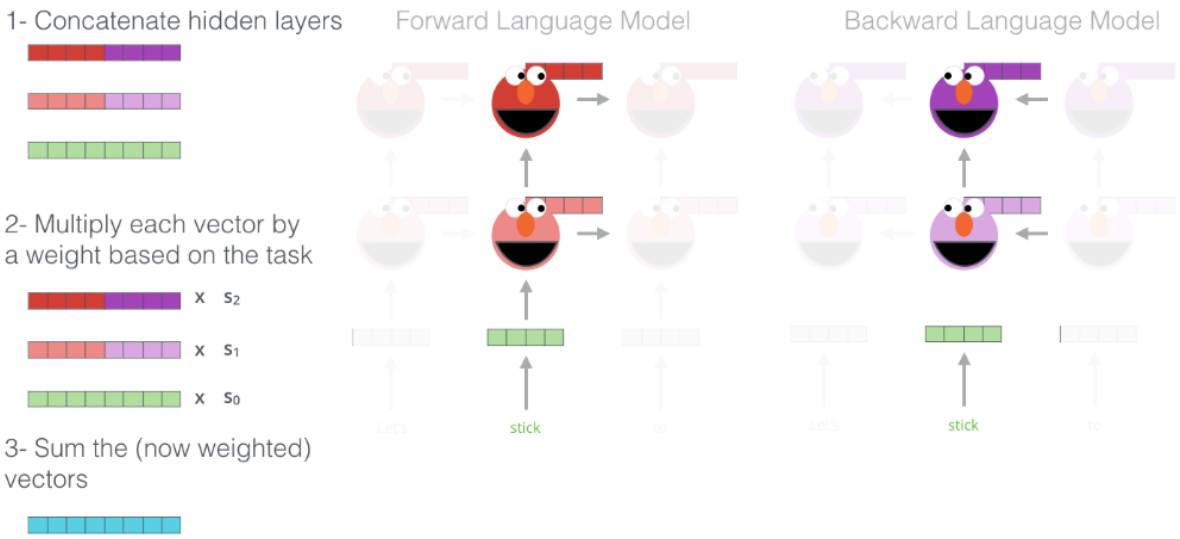


Σχήμα 3.8: Οι μετασχηματισμοί που εφαρμόζονται σε κάθε λέξη πριν τροφοδοτηθούν στο LSTM [10].

Καταλήγοντας, η τελική αναπαράσταση της κάθε λέξης που παράγεται από το ELMo είναι ο σταθμισμένος συνδυασμός των διαφορετικών 2+1 αναπαραστάσεων, δηλαδή των εξόδων από τα δύο επίπεδα του biLM και της εισόδου του που παράγεται από το Συνελικτικό Δίκτυο χαρακτήρων. Επίσης τα βάρη της στάθμισης προσαρμόζονται κατά την εκπαίδευση του μοντέλου της συγκεκριμένης εργασίας και δεν είναι μέρος του προ-εκπαιδευμένου μοντέλου ELMo. Συγκεκριμένα η τελική αναπαράσταση της k λέξης εντός μιας πρότασης λαμβάνεται ως εξής:

$$ELMo_k^{task} = \gamma^{task} \cdot (s_0^{task} \cdot x_k + s_1^{task} \cdot h_{1,k} + s_2^{task} \cdot h_{2,k}) \qquad (3.7)$$

Όπου x_k είναι η αναπαράσταση από το Συνελικτικό Δίκτυο που βασίζεται στους χαρακτήρες της, $h_{1,k}$ και $h_{2,k}$ οι αναπαραστάσεις που προέρχεται από το 1ο και 2ο επίπεδο του biLM αντίστοιχα και οι παράμετροι $s_0^{task}, s_1^{task}, s_2^{task}$, γ^{task} που προσαρμόζονται κατά την εκπαίδευση του δικτύου για την συγκεκριμένη εργασία. Ένα παράδειγμα για την λέξη “stick” εμφανίζεται στο επόμενο Σχήμα 3.9.



Σχήμα 3.9: Εμφύτευση της λέξης ‘stick’ μέσο του ELMo [11].

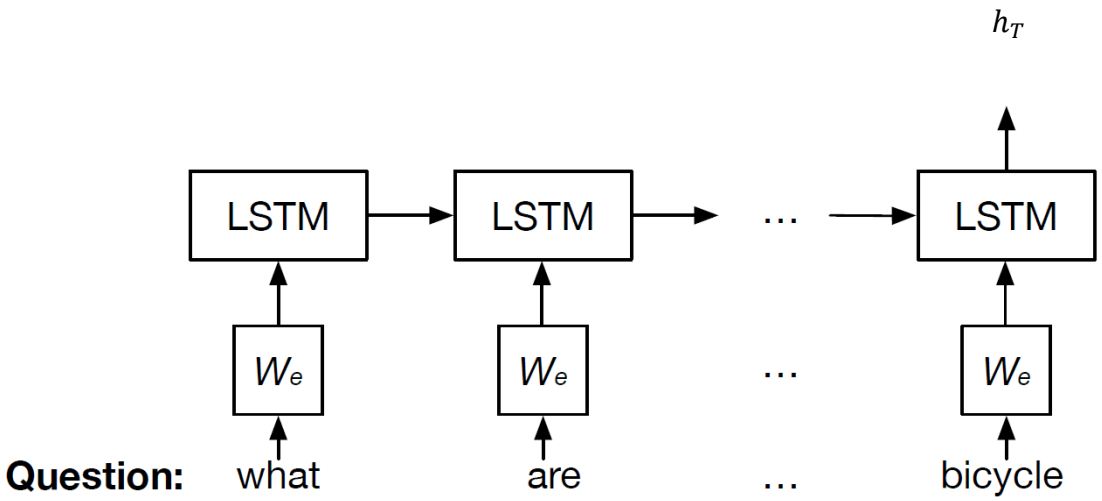
Αναφέρεται επίσης ότι το δίκτυο αποτελείται συνολικά από 93.6 εκατομμύρια παραμέτρους οι οποίες έχουν προσαρμοστεί πάνω σε 5.5 δισεκατομμύρια λέξεις και προέρχονται 1.9 δισεκατομμύρια από την Wikipedia και 3.6 δισεκατομμύρια από δεδομένα ροών ειδήσεων στην αγγλική που προέρχονται από το WMT (Workshop on Statistical Machine Translation) 2008-2012.

3.3.4 Εμφύτευση Ερωτήσεων Μέσω Μοντέλου LSTM

Η αναπαράσταση των λέξεων μιας ερώτησης μέσω των απλών εμφυτευμένων διανυσμάτων x_t ή των πιο σύνθετων διανυσμάτων $ELMo_k$ δίνει στο μοντέλο την ικανότητα να αναπαραστήσει τη σημασία των λέξεων μέσω ενός διανύσματος. Δοθέντος λοιπόν των διανυσμάτων x_t (θα χρησιμοποιείται αυτός ο συμβολισμός για τα διανύσματα $ELMo_k$) στόχος πλέον είναι να υπάρξει μια συνολική αναπαράσταση της ερώτησης σε μορφή διανύσματος. Για το σκοπό αυτό γίνεται χρήση ενός δικτύου LSTM. Η επιλογή του δικτύου LSTM, έγινε λόγω της ικανότητας των δικτύων αυτών να αξιοποιούν την σειρά εμφάνισης των λέξεων σε μια ερώτηση καθώς επίσης έχουν εμφανίσει εξαιρετικά αποτελέσματα στο πεδίο της Επεξεργασίας Φυσικής Γλώσσας (NLP) όταν υπάρχει σχετικά μεγάλος όγκος δεδομένων. Συνεπώς για κάθε λέξη εντός μιας ερώτησης (χρονικό βήμα t) το εμφυτευμένο διάνυσμα της ερώτησης δίδεται ως είσοδο στο LSTM:

$$h_t = LSTM(x_t), t \in \{1, 2, \dots, T\} \qquad (3.8)$$

Ως τελική αναπαράσταση u_Q της ερώτησης θεωρείται η έξοδος του LSTM (κρυφή κατάσταση – hidden state) μετά από το πέρασμα της τελευταίας λέξης της ερώτησης, δηλαδή $u_Q = h_T$ (Yang et al., 2016).



Σχήμα 3.10: Μοντέλο Αναπαράστασης της ερώτησης που βασίζεται σε LSTM (Yang et al., 2016).

Ένα παράδειγμα για την ερώτηση “what are sitting in the basket on a bicycle” παρουσιάζεται στο Σχήμα 3.10.

3.4 Δίκτυο Πολλαπλών Εστιάσεων (Stacked Attention Network)

Έχοντας την αναπαράσταση της εικόνας u_I και την αναπαράσταση της ερώτησης u_Q , το Δίκτυο Πολλαπλών Εστιάσεων (SAN) κάνει εκτίμηση της σωστής απάντησης μέσω πολλαπλών βημάτων. Τα βήματα αυτά προσπαθούν να προσεγγίσουν την ανθρώπινη συλλογιστική πορεία που θα ακολουθούταν για την απάντηση μιας ερώτησης που βασίζεται σε οπτικό περιεχόμενο. Σε πολλές περιπτώσεις η απάντηση μπορεί να σχετίζεται μόνο με μία μικρή περιοχή της εικόνας. Για παράδειγμα στο Σχήμα 3.1 παρόλο που εμφανίζονται πολλαπλά αντικείμενα (Ποδήλατα, καλάθι, παράθυρο, δρόμος, σκυλιά) μόνο η λέξη σκυλιά σχετίζεται με την απάντηση στην ερώτηση. Συνεπώς κάνοντας χρήση μόνο μιας αναπαράστασης για το σύνολο της εικόνας (global feature vector) για την εκτίμηση της σωστής απάντησης θα οδηγούσε σε μη βέλτιστα αποτελέσματα λόγω του θορύβου που θα δημιουργούσαν οι περιοχές της εικόνας που δεν σχετίζονται με την απάντηση. Αντίθετα η σταδιακή εστίαση μέσω κατάλληλων επιπέδων εστίασης (Attention Layers) είναι ικανή να φιλτράρει τον θόρυβο αυτό και να εντοπίσει περιοχές που σχετίζονται με την απάντηση.

Δοθείσης λοιπόν της αναπαράστασης της εικόνας μέσω ενός πίνακα χαρακτηριστικών $u_I \in \mathbb{R}^{d\times m}$ και της αναπαράστασης της ερώτησης $u_Q \in \mathbb{R}^d$, όπου d είναι η διάσταση των αναπαραστάσεων και m οι διαφορετικές περιοχές της εικόνας, οι αναπαραστάσεις αυτές δίδονται ως είσοδοι σε ένα Πλήρως Συνδεδεμένο Επίπεδο και έπειτα σε μία Softmax συνάρτηση ώστε να παραχθεί η κατανομή της εστίασης πάνω στις περιοχές της εικόνας:

$$h_A = \tanh(W_{I,A}u_I \oplus (W_{Q,A}u_Q + b_A)), \quad (3.9)$$
$$p_I = softmax(W_P h_A + b_P), \quad (3.10)$$

Όπου $W_{I,A}, W_{Q,A} \in \mathbb{R}^{k\times d}$, $b_A \in \mathbb{R}^k$, $b_P \in \mathbb{R}^m$ και $W_P \in \mathbb{R}^{1\times k}$ αποτελούν τα βάρη του δικτύου που προσαρμόζονται κατά την εκπαίδευση και συνεπώς το $p_I = (p_1, p_2, \dots, p_m) \in \mathbb{R}^m$ αποτελεί τις πιθανότητες εστίασης σε κάθε μία από τις m περιοχές της εικόνας. Σημειώνεται ότι με $\oplus$ συμβολίζεται η πρόσθεση μεταξύ ενός διανύσματος και ενός πίνακα όπου ουσιαστικά η πρόσθεσή τους εκτελείται προσθέτοντας σε κάθε κολώνα του πίνακα το διάνυσμα (Broadcasting).

Στη συνέχεια με βάση την κατανομή πιθανοτήτων εστίασης υπολογίζεται το σταθμισμένο άθροισμα των διανυσμάτων της εικόνας $\tilde{u}_I$ από την κάθε περιοχή u_i όπως εμφανίζεται στην εξίσωση (7). Έπειτα συνδυάζεται το διάνυσμα $\tilde{u}_I$ με εκείνο της ερώτησης u_Q, με σκοπό τη δημιουργία ενός νέου βελτιωμένου διανύσματος u που περιέχει πληροφορία και από την ερώτηση αλλά και από χαρακτηριστικά περιοχών της εικόνας που σχετίζονται με την ερώτηση, όπως εμφανίζεται στην εξίσωση (8):

$$\tilde{u}_I = \sum_i p_i u_i, \quad (3.11)$$
$$u = \tilde{u}_I + u_Q. \quad (3.12)$$

Συγκριτικά με μοντέλα που απλά συνδυάζουν τα διανύσματα ερωτήσεων με διάνυσμα χαρακτηριστικών από το σύνολο της εικόνας, το μοντέλο εστίασης δημιουργεί μία αναπαράσταση που εμπεριέχει πιο ακριβή πληροφορία καθώς μεγαλύτερα βάρη πιθανοτήτων

αντιστοιχούν σε περιοχές της εικόνας που σχετίζονται με την ερώτηση. Ωστόσο για αρκετά πολύπλοκες ερωτήσεις, ένα μόνο Επίπεδο Εστίασης όπως περιγράφηκε προηγουμένως δεν είναι επαρκές για τον εντοπισμό των περιοχών που σχετίζονται με την σωστή απάντηση. Για παράδειγμα η ερώτηση 'what are sitting in the basket on a bicycle' όπως παρουσιάστηκε στο Σχήμα 3.1 αναφέρεται σε μια πολύ λεπτομερή συσχέτισή των πολλαπλών αντικειμένων που βρίσκονται στην εικόνα. Για το λόγο αυτό απαιτείται μία πολλαπλή διαδικασία εστίασης η οποία σε κάθε στάδιο θα εξάγει μια πιο λεπτομερή πληροφορία εστίασης. Ουσιαστικά σε κάθε στάδιο χρησιμοποιείται η αναπαράσταση u ως ένα βελτιωμένο διάνυσμα για την αναπαράσταση της ερώτησης. Συγκεκριμένα, για το k-οστό Επίπεδο Εστίασης υπολογίζονται τα εξής:

$$h_A^k = \tanh(W_{I,A}^k u_I \oplus (W_{Q,A}^k u^{k-1} + b_A^k)), \quad (3.13)$$

$$p_I^k = softmax\left(W_P^k h_A^k + b_P^k\right). \quad (3.14)$$

Όπου $u^0 = u_Q$. Έπειτα όπως περιγράφηκε και προηγουμένως το σταθμισμένο διάνυσμα χαρακτηριστικών της εικόνας με βάση τις υπολογισμένες πιθανότητες εστίασης προστίθεται στο αποτέλεσμα του προηγούμενου επιπέδου εστίασης, ώστε να δημιουργηθεί το διάνυσμα χαρακτηριστικών u^k του Επιπέδου Εστίασης k:

$$\tilde{u}_I^k = \sum_i p_i^k u_i, \quad (3.15)$$

$$u^k = \tilde{u}_I^k + u^{k-1}. \quad (3.16)$$

Η διαδικασία αυτή επαναλαμβάνεται K φορές και τέλος γίνεται χρήση του u^K ώστε να γίνει εκτίμηση της σωστής απάντησης μέσω ενός Πλήρως Συνδεδεμένου Επιπέδου με συνάρτηση ενεργοποίησης την Softmax:

$$p_{ans} = softmax(W_u u^K + b_u). \quad (3.17)$$

Στο Σχήμα 3.1 εμφανίζεται η διαδικασία που ακολουθείται. Στο πρώτο επίπεδο το μοντέλο εντοπίζει τις περιοχές που σχετίζονται με τις λέξεις basket, bicycle και sitting in, ενώ στο δεύτερο επίπεδο το μοντέλο εστιάζει με μεγαλύτερη λεπτομέρεια σε περιοχές που σχετίζονται με την απάντηση dogs.

Σημειώνεται ότι το $p_{ans} \in [0,1]^{1000}$ αποτελεί την κατανομή των πιθανοτήτων για κάθε δυνατή απάντηση που βρίσκεται στο Λεξικό Απαντήσεων όπως έχει περιγραφεί σε προηγούμενη ενότητα. Συνεπώς η τελική απάντηση είναι εκείνη που αντιστοιχεί στην διάσταση (από τις 1000) που εμφανίζει την μεγαλύτερη πιθανότητα (Yang et al., 2016).

3.5 Προγραμματιστικής Πλατφόρμες & Εργαλεία

Η υλοποίηση της παρούσας εργασίας βασίστηκε στην γλώσσα προγραμματισμού Python και τα δύο υπολογιστικά προγραμματιστικά περιβάλλοντα που χρησιμοποιήθηκαν είναι τα PyCharm και Jupyter Notebook. Το πρώτο χρησιμοποιήθηκε για την κεντρική εκπαίδευση των μοντέλων και το δεύτερο για την αρχική δημιουργία των μοντέλων καθώς και για την οπτικοποίηση των αποτελεσμάτων. Η βασική πλατφόρμα που χρησιμοποιήθηκε για την υλοποίηση των μοντέλων είναι η PyTorch (Paszke et al. 2017). Η πλατφλορμα *PyTorch*, δημιουργήθηκε από ερευνητές και μηχανικούς της *Facebook*, για τις ανάγκες έρευνας πάνω σε συστήματα μηχανικής και βαθιάς μάθησης. Η ευέλικτη αρχιτεκτονική της δίνει το δικαίωμα να παραταχθούν υπολογισμοί σε περισσότερες από μια CPUs ή GPUs με ένα μόνο API. Σημειώνεται ότι το προ-εκπαιδευμένο μοντέλο DenseNet-161 που χρησιμοποιήθηκε είναι εκείνο που παρέχεται μέσω του Pytorch και ότι το προ-εκπαιδευμένο μοντέλο ELMo υλοποιημένο στην PyTorch παρέχεται από τους ίδιους τους δημιουργούς του μέσω της ανοικτού κώδικα βιβλιοθήκης AllenNLP (Gardner et al., 2017). Παράλληλα με την PyTorch αξιοποιήθηκε και η πλατφόρμα Caffe για την χρήση των προ-εκπαιδευμένων μοντέλων VGGNet-19 όπως παρέχονται από το Visual Geometry Group του Πανεπιστημίου της Οξφόρδης. Η Caffe αποτελεί μια πλατφόρμα για συστήματα μηχανικής και βαθιάς μάθησης και έχει δημιουργηθεί από την ομάδα Berkeley AI Research. Η επεξεργασία των δεδομένων, η υλοποίηση των μοντέλων και κυρίως η εκπαίδευσή τους αποτελούν εργασίες που απαιτούν πολλούς υπολογιστικούς πόρους. Για το λόγο αυτό χρησιμοποιήθηκαν δύο κάρτες γραφικών (GPU) Nvidia GeForce GTX 1080 με 8GB μνήμη οι οποίες ανήκουν στο εργαστήριο ISLAB - Εργαστήριο Ευφυών Συστημάτων, σε συνδυασμό με το λογισμικό 'CUDA' της Nvidia το οποίο χρησιμοποιεί η PyTorch και η Caffe (Jia et al., 2014). Το 'NVIDIA CUDA Toolkit' παρέχει ένα περιβάλλον για τη δημιουργία υψηλής απόδοσης GPU εφαρμογών. Οι βιβλιοθήκες CUDA επιτρέπουν την υπολογιστική επιτάχυνση σε πολλαπλούς τομείς, όπως η επεξεργασία εικόνων και βίντεο και η βαθιά μηχανική μάθησης (deep learning).

Μαζί με τις βασικές βιβλιοθήκες PyTorch και Caffe, χρησιμοποιήθηκαν οι ακόλουθες βιβλιοθήκες:

- H5py (Collette 2013) : API για δεδομένα σε μορφή HDF5 που χρησιμοποιήθηκαν για την διαχείριση και αποθήκευση δεδομένων
- Numpy (Oliphant 2006) : υπολογιστικές δυνατότητες και δομές
- Matplotlib (Hunter 2007) : δημιουργία γραφικών παραστάσεων
- PIL (Alex Clark and Contributors) / OpenCV (Bradski 2000) : επεξεργασία και διαχείριση εικόνων
- Sklearn (Pedregosa et al., 2011) : μετρικές απόδοσης και εργαλεία μηχανικής μάθησης
- NLTK (Loper 2002) : επεξεργασία και διαχείριση των ερωτήσεων και απαντήσεων

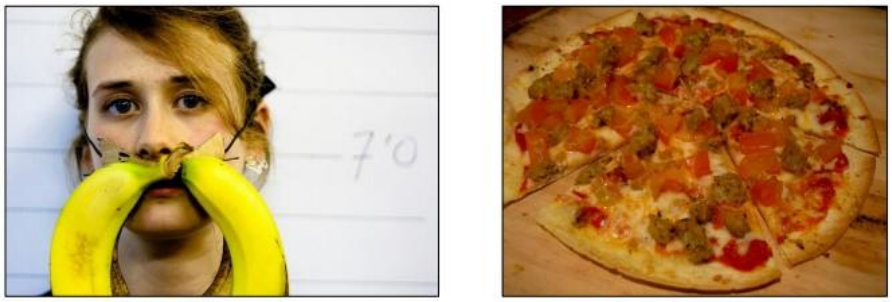
4. Εκπαίδευση Μοντέλων και Αποτελέσματα

4.1 Σύνολο Δεδομένων

Στο κεφάλαιο αυτό αναλύονται τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων που δοκιμάστηκαν, καθώς και για την αξιολόγηση της απόδοσής τους. Επίσης παρουσιάζονται στατιστικά στοιχεία σχετικά με το σύνολο των δεδομένων εκπαίδευσης που οδήγησαν στην επιλογή των βασικών παραμέτρων των μοντέλων.

4.1.1 VQA V.2

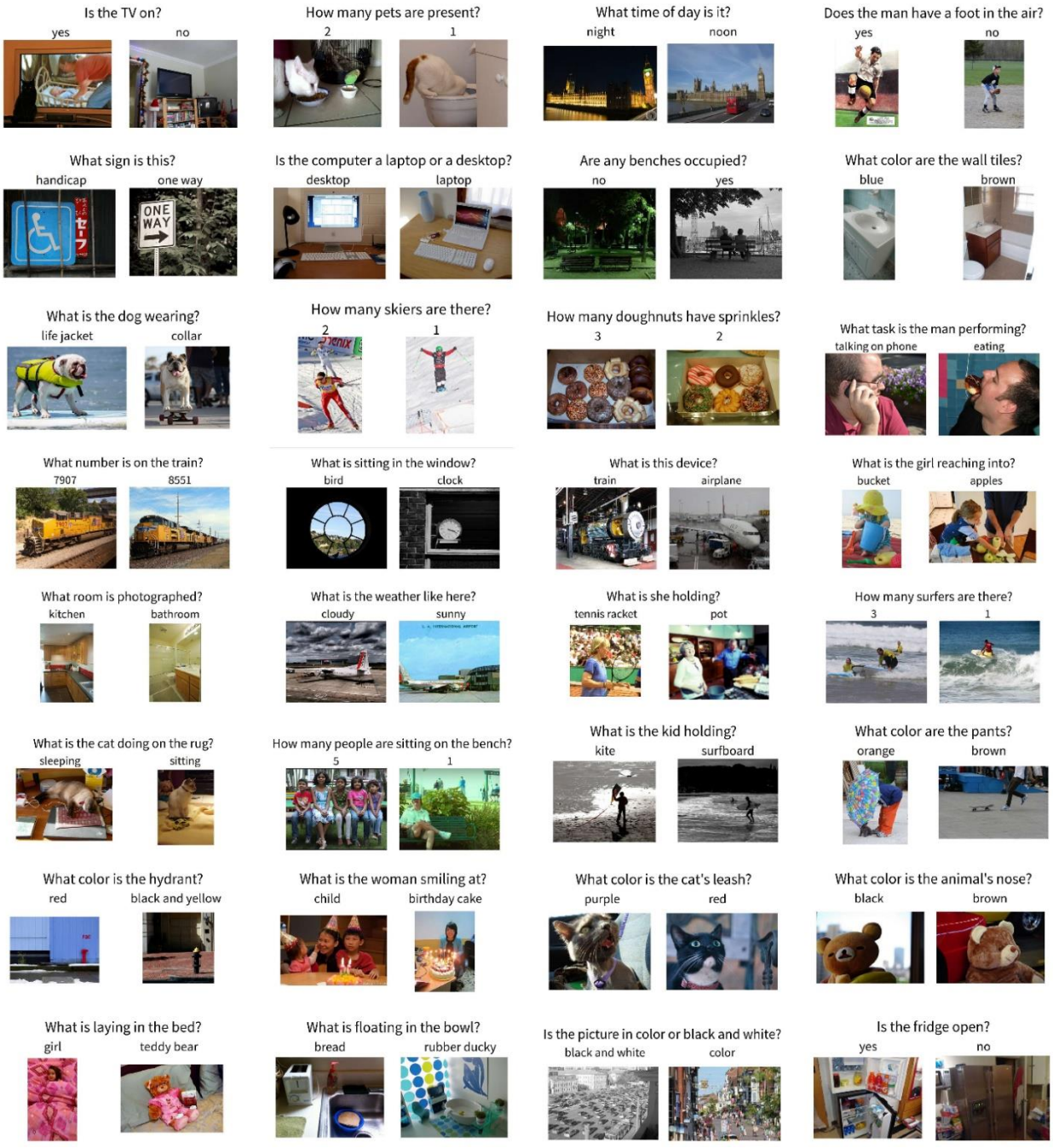
Το σύνολο των δεδομένων που χρησιμοποιήθηκε ονομάζεται VQA v2. Το VQA v2 αποτελεί μία διευρυμένη έκδοση του VQA v1. Το VQA v1 σύνολο δεδομένων αποτελείται από εικόνες, ερωτήσεις και απαντήσεις πάνω σε αυτές τις εικόνες. Για κάθε εικόνα ζητήθηκε από τουλάχιστον τρεις διαφορετικούς χρήστες να θέσουν μία ερώτηση για τη συγκεκριμένη εικόνα. Μάλιστα κάθε χρήστης γνώριζε τις προηγούμενες ερωτήσεις ώστε να υπάρξει ποικιλία στις ερωτήσεις. Έπειτα για κάθε μία από τις ερωτήσεις αυτές, ζητήθηκε από 10 διαφορετικούς χρήστες κάθε φορά να την απαντήσουν. Στο επόμενο Σχήμα 4.1 εμφανίζονται δύο παραδείγματα από το VQA v1 σύνολο δεδομένων (Antol et al., 2015).



Σχήμα 4.1: Παράδειγμα από το σύνολο δεδομένων VQA (Antol et al., 2015).

Το VQA v1 είναι ένα σύνολο δεδομένων το οποίο συνεισέφερε στην πρόοδο του συγκεκριμένου πεδίου. Παρόλα αυτά αποδείχθηκε ότι το VQA v1 ωθεί τα μοντέλα να είναι μεροληπτικά ως προς το γλωσσικό περιεχόμενο και ουσιαστικά να μην χρησιμοποιούν το οπτικό περιεχόμενο. Ουσιαστικά αυτό συμβαίνει διότι για αρκετές παρόμοιες ερωτήσεις υπάρχουν ίδιες απαντήσεις και συνεπώς τα μοντέλα δεν αξιοποιούν την οπτική πληροφορία. Για παράδειγμα, σε όλες τις ερωτήσεις σχετικά με το αν υπάρχει ένα ρολόι σε μία εικόνα, οι απαντήσεις των χρηστών ήταν πάντα ‘ναι’. Συνεπώς όλα τα μοντέλα μάθαιναν για τη συγκεκριμένη ερώτηση να απαντούν πάντα ‘ναι’. Η διεύρυνση λοιπόν του VQA v2 έγινε ως προς την κατεύθυνση ότι για κάθε εικόνα

I να υπάρχει μία παρόμοια με αυτήν εικόνα I’, έτσι ώστε η απάντηση A στην ερώτηση Q που διατυπώθηκε για την αρχική εικόνα I να είναι διαφορετική από την ίδια ερώτηση Q για την παρόμοια εικόνα I’ (Goyal et al., 2017).



Σχήμα 4.2: Παράδειγμα από το σύνολο δεδομένων VQA v2. Για κάθε ερώτηση υπάρχουν δύο όμοιες εικόνες με δύο διαφορετικές απαντήσεις (Goyal et al., 2017).

Το VQA v2 σύνολο δεδομένων αποτελείται από 204,721 εικόνες, 1,105,904 ερωτήσεις και 11,059,040 απαντήσεις. Το σύνολο των εικόνων βασίζεται στο MS COCO 2015 σύνολο δεδομένων. Λόγω της επέκτασής του από το VQA v1, περιέχει τουλάχιστον τρεις ερωτήσεις για κάθε εικόνα (5,4 κατά μέσο όρο). Το σύστημα το οποίο χρησιμοποιήθηκε για τη συλλογή των ερωτήσεων και απαντήσεων είναι το Amazon Mechanical Turk (AMT). Στη συνέχεια παρουσιάζονται κάποια δείγματα από το σύνολο δεδομένων VQA v2. Για κάθε ερώτηση υπάρχουν τουλάχιστον δύο παρόμοιες εικόνες με διαφορετικές απαντήσεις.

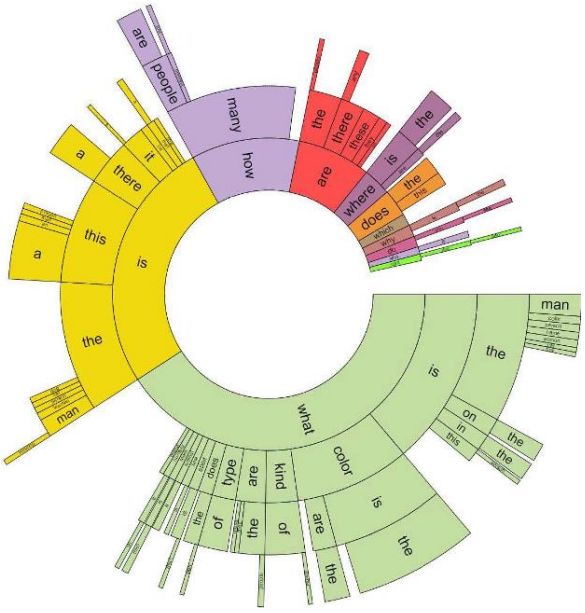
Το σύνολο δεδομένων περιλαμβάνει 3 υποσύνολα:

- Σύνολο εκπαίδευσης
  82,783 εικόνες
  443,757 ερωτήσεις
  4,437,570 απαντήσεις
- Σύνολο αξιολόγησης
  40,504 εικόνες
  214,354 ερωτήσεις
  2,143,540 απαντήσεις
- Σύνολο ελέγχου
  81,434 εικόνες
  447,793 ερωτήσεις
  Δεν παρέχονται οι απαντήσεις

Λόγω του μεγάλου και ποικίλου αριθμού παραδειγμάτων εκπαίδευσης το συγκεκριμένο σύνολο δεδομένων χρησιμοποιείται ευρέως για την αξιολόγηση μοντέλων ερωτήσεων-απαντήσεων πάνω σε οπτικό περιεχόμενο.

4.1.2 Διερευνητική Ανάλυση Δεδομένων

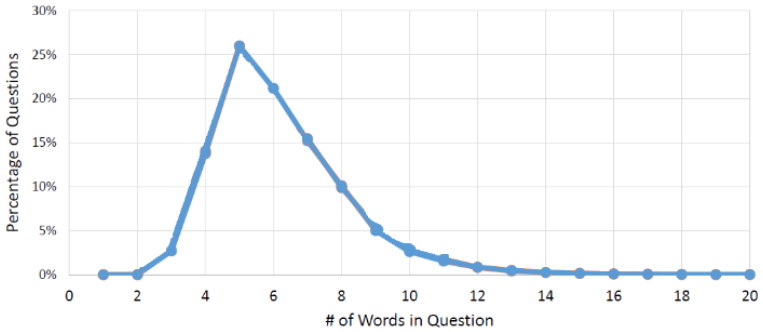
Στη συνέχεια παρουσιάζονται κάποια στατιστικά στοιχεία σχετικά με το σύνολο δεδομένων VQA. Όσον αφορά τις ερωτήσεις, στο επόμενο Σχήμα 4.3 εμφανίζεται η κατανομή τους ως προς τις πρώτες τέσσερις λέξεις που περιέχουν.



Σχήμα 4.3: Η κατανομή των ερωτήσεων ως προς τις πρώτες τέσσερις λέξεις τους για ένα τυχαίο δείγμα 60000 ερωτήσεων (Antol et al., 2015).

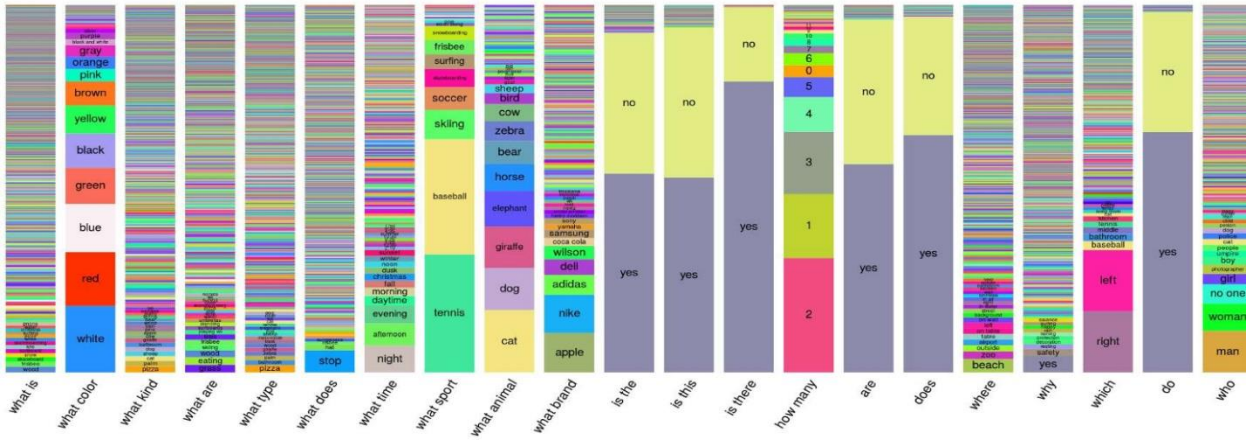
Η κατανομή είναι υπολογισμένη πάνω σε ένα τυχαίο δείγμα 60,000 ερωτήσεων. Παρατηρείται πολύ μεγάλη ποικιλία ερωτήσεων καθώς και ότι το μεγαλύτερο ποσοστό των ερωτήσεων αρχίζει με τις λέξεις ‘what..’, ‘is..’, ‘are..’ και ‘how..’. Επίσης παρατηρείται ότι η ερώτηση ‘what is..’ έχει αρκετά μεγάλο ποσοστό εμφάνισης κάτι που είναι αρκετά ενδιαφέρον καθώς αποτελεί μία ερώτηση που η σωστή απάντηση μπορεί να προέρχεται από ένα τεράστιο πλήθος πιθανών απαντήσεων.

Όσον αφορά την κατανομή του μήκους των ερωτήσεων παρατηρείται στο επόμενο Σχήμα 4.4 ότι το μεγαλύτερο ποσοστό έχει μήκος μέχρι 10 λέξεις. Επίσης αναφέρεται ότι στο σύνολο εκπαίδευσης το μεγαλύτερο μήκος ερώτησης είναι 23 ενώ στο σύνολο ελέγχου είναι 25. Για το λόγο αυτό επιλέχθηκε το μήκος 26 για τον μέγιστο αριθμό που μπορούν να υποστηρίξουν οι δομές των μοντέλων της συγκεκριμένης εργασίας.



Σχήμα 4.4: Ποσοστό των ερωτήσεων ανά πλήθος λέξεων (Antol et al., 2015).

Στο επόμενο Σχήμα 4.5 εμφανίζεται η κατανομή των απαντήσεων για κάθε κατηγορία ερώτησης. Όπως αναφέρουν και οι δημιουργοί του VQA v2 το συγκεκριμένο σύνολο δεδομένων παρότι δεν είναι τέλεια ισορροπημένο είναι πολύ περισσότερο σε σύγκριση με εκείνο του VQA v1.



Σχήμα 4.5: Κατανομή των απαντήσεων για κάθε κατηγορία ερώτησης από ένα τυχαίο δείγμα 60000 ερωτήσεων του συνόλου VQA v.2 (Goyal et al., 2017).

Όσον αφορά την κατανομή των απαντήσεων, το 98% των απαντήσεων δεν υπερβαίνει τις τρεις λέξεις και συγκεκριμένα η κατανομή των απαντήσεων που περιέχουν μία, δύο ή τρεις λέξεις είναι 89,32%, 6,91% και 2,74% αντίστοιχα. Ένα άλλο σημαντικό χαρακτηριστικό είναι ότι οι 1000 απαντήσεις με τη μεγαλύτερη συχνότητα στο σύνολο εκπαίδευσης εμφανίζονται σε ποσοστό 87,47% των ερωτήσεων. Το τελευταίο ποσοστό δηλώνει ότι η προσπάθεια για σωστή εκτίμηση αυτών των απαντήσεων αποτελεί στην ουσία το βασικό στόχο για αυτό το σύνολο δεδομένων. Για το λόγο αυτό έγινε επιλογή ώστε το Λεξικό των απαντήσεων να έχει μέγεθος 1000, όπως παρουσιάστηκε σε προηγούμενο κεφάλαιο. Αξίζει επίσης να αναφερθεί ότι σε ποσοστό 38,37% των ερωτήσεων οι απαντήσεις είναι “yes”/”no” και σε ποσοστό 12,31% κάποιος αριθμός. Η τελευταία παρατήρηση δηλώνει ότι ένα πολύ μεγάλο ποσοστό των απαντήσεων επικεντρώνεται σε δύο μόνο κατηγορίες απαντήσεων (Goyal et al., 2017).

4.1.3 Μετρική Τελικής Αξιολόγησης

Οι δημιουργοί του VQA συνόλου δεδομένων διεξάγουν κάθε χρόνο έναν διαγωνισμό πάνω στο συγκεκριμένο σύνολο δεδομένων. Αυτή τη στιγμή διεξάγεται ο 4ος κατά σειρά διαγωνισμός με όνομα VQA Challenge 2019. Οι διαγωνιζόμενοι αξιολογούνται ως προς την απόδοση των μοντέλων τους πάνω στο Σύνολο ελέγχου. Η αξιολόγηση τους γίνεται με την μετρική:

$$Acc(\text{απάντηση}) = \min\{\frac{\#\ \text{άνθρωποι που απάντησαν 'απάντηση'}}{3}, 1\} \qquad (4.1)$$

Η μετρική αξιολόγησης όπως είναι προφανές παίρνει τιμές από 0 έως 1 και συγκεκριμένα τις τιμές [0, 0.33, 0.66, 1]. Ουσιαστικά λοιπόν, τέλεια ακρίβεια 1 σε μία απάντηση δίδεται όταν τουλάχιστον τρεις από τους ερωτηθέντες έδωσαν την ίδια απάντηση. Η μετρική αυτή αντανακλά πολύ καλύτερα την έννοια της ανθρώπινης 'επικρατούσας άποψης' σε σχέση με την συνήθη μετρική της ακρίβειας. Για τον λόγο αυτό αλλά και για να είναι εφικτή η σύγκριση των αποτελεσμάτων με άλλες προσπάθειες που χρησιμοποιήθηκε το VQA v2 σύνολο δεδομένων, η αξιολόγηση των αποτελεσμάτων έγινε με την προαναφερθείσα μετρική (Antol et al., 2015).

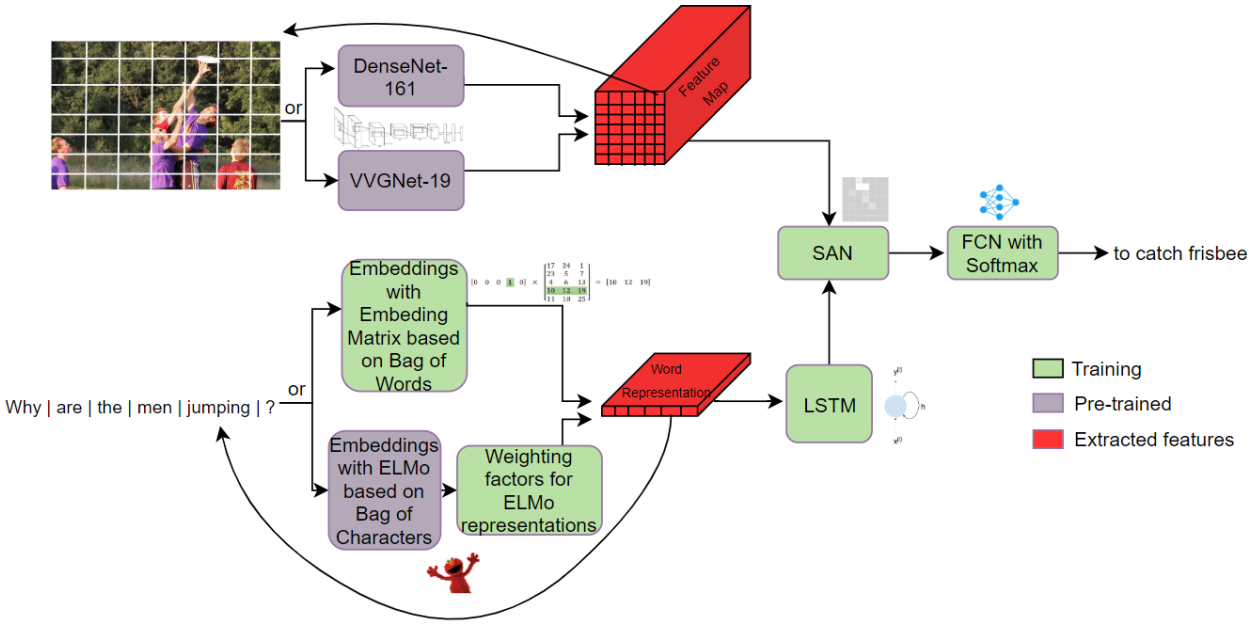
4.2 Εκπαίδευση Μοντέλων

Στην ενότητα αυτή παρουσιάζεται αρχικά μια επισκόπηση των μοντέλων που εκπαιδεύτηκαν και γίνεται διάκριση των προ-εκπαιδευμένων τμημάτων τους και εκείνων που τα βάρη τους προσαρμόστηκαν κατά την διαδικασία εκπαίδευσης στην παρούσα εργασία. Έπειτα παρουσιάζονται οι υποψήφιες υπέρ-παράμετροι των διαφορετικών μοντέλων. Στη συνέχεια αναλύεται η προς ελαχιστοποίηση συνάρτηση κόστους που επιλέχθηκε για την εύρεση των βέλτιστων υπέρ-παραμέτρων καθώς και τεχνικές που χρησιμοποιήθηκαν. Τέλος, παρουσιάζεται η βέλτιστη επιλογή των υπέρ-παραμέτρων του κάθε μοντέλου που δοκιμάστηκε.

4.2.1 Εκπαιδευόμενες Μεταβλητές

Το συνολικό σύστημα όπως παρουσιάστηκε και στο Σχήμα 3.1 αποτελείται από επιμέρους μοντέλα που σχετίζονται με την αναπαράσταση της εικόνας, της ερώτησης καθώς και τον συνδυασμό αυτών των δύο για την τελική εκτίμηση της απάντησης. Στο Σχήμα 4.6 εμφανίζεται η γενική αρχιτεκτονική του συστήματος που υλοποιήθηκε με τις εναλλακτικές επιλογές που δοκιμάστηκαν. Συγκεκριμένα για την αναπαράσταση της εικόνας οι δύο διαφορετικές προσεγγίσεις είναι τα VVGNet-19 και DenseNet-161 και για την αναπαράσταση της ερώτησης η μία επιλογή είναι η αναπαράσταση των λέξεων της ερώτησης μέσω Λεξικού Λέξεων και απεικόνιση τους μέσω Πίνακα Εμφύτευσης, ενώ η δεύτερη είναι η χρήση του μοντέλου ELMo το οποίο βασίζεται σε Λεξικό Γραμμάτων. Κάποια από αυτά τα μοντέλα παρέχονται από τις ομάδες που τα πρότειναν προ-εκπαιδευμένα. Επίσης η προσαρμογή των συγκεκριμένων μοντέλων έχει γίνει σε εκατομμύρια δεδομένα και παρουσιάζουν εξαιρετικές επιδόσεις. Συγκεκριμένα τα μοντέλα VVGNet-19, DenseNet-161 και ELMo παρέχονται με τα προ-εκπαιδευμένα βάρη τους από το Visual Geometry Group του πανεπιστημίου της Οξφόρδης, το Pytorch Hub (Αποθετήριο μοντέλων) και το AllenNLP αντίστοιχα. Τα τρία αυτά μοντέλα έχουν υιοθετηθεί σε ένα πολύ μεγάλο εύρος από διαφορετικές εργασίες που σχετίζονται με την αναπαράσταση εικόνας ή κειμένου και εμφανίζουν εξαιρετικά αποτελέσματα. Λόγω λοιπόν της εξαιρετικής τους γενίκευσης αποφασίστηκε να χρησιμοποιηθούν και στην παρούσα εργασία και μάλιστα με τα προ-εκπαιδευμένα βάρη τους.
Στο επόμενο Σχήμα 4.6 εμφανίζονται με γκρι χρώμα τα μοντέλα εκείνα τα οποία αρχικοποιήθηκαν με τα ήδη εκπαιδευμένα βάρη τους και τα οποία δεν προσαρμόζονται κατά την εκπαίδευση του συνολικού συστήματος. Αντίθετα με πράσινο χρώμα εμφανίζονται τα μέρη του συστήματος στα οποία τα βάρη τους προσαρμόζονται κατά την διαδικασία εκπαίδευσης.

Επίσης με κόκκινο χρώμα εμφανίζονται τα χαρακτηριστικά που έχουν εξαχθεί από την εικόνα και την ερώτηση μέσω των μοντέλων αναπαράστασης της εικόνας και της ερώτησης αντίστοιχα. Ακόμα με βέλη προς τα πίσω αποτυπώνεται ένα παράδειγμα που συνδέει ένα διάνυσμα χαρακτηριστικών εικόνας με την αντίστοιχη περιοχή στην αρχική εικόνα και ένα διάνυσμα χαρακτηριστικών λέξης με την αντίστοιχη λέξη της ερώτησης.



Σχήμα 4.6: Γενική αρχιτεκτονική του συστήματος.

Αναλυτικά, τα βάρη που δεν προσαρμόζονται κατά την εκπαίδευση του συνολικού συστήματος είναι:

- Του δικτύου VVGNet-19 όπου αποτελείται από:
  a) Δύο 3x3 Συνελικτικά Επίπεδα που παράγουν 64 χάρτες χαρακτηριστικών
  b) Δύο 3x3 Συνελικτικά Επίπεδα που παράγουν 128 χάρτες χαρακτηριστικών
  c) Τέσσερα 3x3 Συνελικτικά Επίπεδα που παράγουν 256 χάρτες χαρακτηριστικών
  d) Οκτώ 3x3 Συνελικτικά Επίπεδα που παράγουν 512 χάρτες χαρακτηριστικών
  e) Δύο Πλήρως Συνδεδεμένα Επίπεδα μεγέθους 4096
  f) Ένα Πλήρως Συνδεδεμένο Επίπεδο μεγέθους 1000
- Του δικτύου DenseNet-161 όπου αποτελείται από:
  a) Ογδόντα ένα 1x1 Συνελικτικά Επίπεδα που παράγουν 48 χάρτες χαρακτηριστικών
  b) Εβδομήντα οκτώ 3x3 Συνελικτικά Επίπεδα που παράγουν 48 χάρτες χαρακτηριστικών
  c) Ένα 7x7 Συνελικτικό Επίπεδό που παράγει 48 χάρτες χαρακτηριστικών
  d) Ένα Πλήρως Συνδεδεμένο Επίπεδο μεγέθους 1000
- Του δικτύου ELMo όπου αποτελείται από:

a) Ένα 1x16 Συνελικτικό Επίπεδο που παράγει 32 χαρακτηριστικά
b) Ένα 2x16 Συνελικτικά Επίπεδα που παράγουν 32 χαρακτηριστικά
c) Ένα 3x16 Συνελικτικά Επίπεδα που παράγουν 64 χαρακτηριστικά
d) Ένα 4x16 Συνελικτικά Επίπεδα που παράγουν 128 χαρακτηριστικά
e) Ένα 5x16 Συνελικτικά Επίπεδα που παράγουν 256 χαρακτηριστικά
f) Ένα 6x16 Συνελικτικά Επίπεδα που παράγουν 512 χαρακτηριστικά
g) Ένα 7x16 Συνελικτικά Επίπεδα που παράγουν 1024 χαρακτηριστικά
h) Δύο Επίπεδα Λεωφόρων (Highway Layers)
i) Τρία Πλήρως Συνδεδεμένα Επίπεδα μεγέθους 512
j) Ένα διπλής κατεύθυνσης LSTM με δύο επίπεδα που το κάθε ένα αποτελείται από 4096 LSTM μονάδες

Αναλυτικά τα βάρη που προσαρμόζονται κατά την εκπαίδευση του συνολικού συστήματος είναι:

- Τα βάρη του Πίνακα εμφύτευσης (Στην περίπτωση που δεν γίνεται χρήση του ELMo)
- Οι συντελεστές βαρύτητας των αναπαραστάσεων που παράγει το ELMo
- Τα βάρη του LSTM δικτύου που χρησιμοποιήθηκε για την αναπαράσταση της ερώτησης
- Τα βάρη από τα Πολλαπλά Επίπεδα Εστίασης
- Ένα Πλήρως Συνδεδεμένο Επίπεδο μεγέθους 1000

Σημειώνεται ότι το ακριβές πλήθος και μέγεθος των επιπέδων που προσαρμόζονται κατά την εκπαίδευση, συμπεριλαμβάνεται στις υπερπαραμέτρους του συστήματος και συνεπώς δεν είναι εξαρχής καθορισμένα.

4.2.2 Συνάρτηση Κόστους και Τεχνικές Γενίκευσης

Η προσαρμογή των βαρών κατά την εκπαίδευση γίνεται με στόχο την αύξηση της απόδοσης του συνολικού συστήματος. Η αύξηση της απόδοσης αυτής μετασχηματίζεται σε πρόβλημα βελτιστοποίησης μέσω της ελαχιστοποίησης μια συνάρτησης κόστους. Η συνάρτηση κόστους που έχει επιλεχθεί είναι η διασχιζόμενη εντροπία (cross-entropy loss) όπως έχει παρουσιαστεί και στο Κεφάλαιο 2.
Ο αλγόριθμος με βάση τον οποίο γίνεται η διόρθωση των βαρών που ελαχιστοποιούν την συνάρτηση κόστους συμπεριλαμβάνεται επίσης στις υπερπαραμέτρους που δοκιμάστηκαν.
Μία ακόμα τεχνική που χρησιμοποιήθηκε κατά την διάρκεια της εκπαίδευσης για την επίτευξη καλύτερης γενίκευσης και αποφυγής της υπερπροσαρμογής του μοντέλου είναι ο Περιορισμός Ενεργοποίησης (dropout) όπως επίσης έχει παρουσιαστεί και στο Κεφάλαιο 2. Η συγκεκριμένη τεχνική εφαρμόστηκε στις εισόδους της σχέσης (5) του Επιπέδου Εστίασης και συγκεκριμένα στις αναπαραστάσεις της εικόνας u_I και της ερώτησης u_Q. Επίσης χρησιμοποιείται στην είσοδο της σχέσης (6), δηλαδή στις τιμές του χάρτη χαρακτηριστικών h_A. Η πιθανότητα να χρησιμοποιηθεί η συγκεκριμένη τεχνική σε μια τιμή αποτελεί υπερπαράμετρο του συστήματος και παρουσιάζεται στη συνέχεια.

4.2.3 Υπερπαράμετροι Συστήματος

Η επιλογή των υπερπαραμέτρων του συνολικού συστήματος όπως για παράδειγμα το πλήθος και μέγεθος των διαφόρων επιπέδων και η επιλογή του αλγορίθμου βελτιστοποίησης παίζουν σημαντικό ρόλο στην τελική του απόδοση. Για το λόγο αυτό προκειμένου το σύστημα να παράγει όσο το δυνατόν καλύτερα αποτελέσματα, θα πρέπει να αναζητηθούν οι βέλτιστες υπερπαράμετροί του. Η αναζήτηση των βέλτιστων παραμέτρων για τα διάφορα μοντέλα έγινε μέσω δοκιμών πάνω σε ένα σύνολο πιθανών επιλογών. Οι υπερπαράμετροι που δοκιμάστηκαν χωρίζονται σε δύο κατηγορίες, εκείνες που σχετίζονται α) με την αρχιτεκτονική και το μέγεθος του συστήματος και με εκείνες που σχετίζονται β) με την διαδικασία εκπαίδευσης και βελτιστοποίησης.
Οι υπερπαράμετροι που σχετίζονται με την πρώτη κατηγορία αφορούν τα ακόλουθα μέρη του συστήματος:

- Εικόνες εισόδου ανάλυσης 224x224 ή 448x448
- Το μοντέλο για την αναπαράσταση της εικόνας: VGGNet-19 το οποίο συνεπάγεται 512 χάρτες χαρακτηριστικών ή DenseNet-161 με 2208 χάρτες χαρακτηριστικών για την κάθε περιοχή της εικόνας.
- Το μοντέλο για την αναπαράσταση των λέξεων της ερώτησης, δηλαδή αν θα γίνει χρήση του Πίνακα Εμφύτευσης σε Λεξικό Λέξεων ή του μοντέλου ELMo. Και στις δύο περιπτώσεις τα παραγόμενα χαρακτηριστικά είναι 1024.
- Η πιθανότητα ένα από τα χαρακτηριστικά της εικόνας να είναι μηδέν (Dropout) στο εύρος [0.2, 0.5]
- Το μέγεθος του LSTM που χρησιμοποιείται για την αναπαράσταση της ερώτησης. Αντίστοιχα και το μέγεθος του Πλήρως Συνδεδεμένου Επιπέδου το οποίο χρησιμοποιείται ώστε κάθε διάνυσμα χαρακτηριστικών f_i της αναπαράστασης της εικόνας να μετασχηματιστεί σε ένα νέο διάνυσμα ίσης διάστασης με τα διανύσματα χαρακτηριστικών της ερώτησης. Έγιναν δοκιμές με 512, 1024 και 2048 πλήθος μονάδων LSTM.
- Το πλήθος των LSTM επιπέδων για την αναπαράσταση της ερώτησης. Από 1 έως 2.
- Η πιθανότητα ένα από τα χαρακτηριστικά της αναπαράστασης της ερώτησης να είναι μηδέν (Dropout) στο εύρος [0.2, 0.5]
- Από 1 έως 3 Επίπεδα Πολλαπλών Εστιάσεων.
- Το μέγεθος k του κάθε Επιπέδου Πολλαπλών Εστιάσεων (Stacked Attention Networks) όπως αυτό παρουσιάστηκε σε προηγούμενη ενότητα. Ουσιαστικά αποτελεί το μέγεθος των Πλήρως Συνδεδεμένων επιπέδων (Fully connected layers) που υπάρχουν εσωτερικά των Επιπέδων Πολλαπλών Εστιάσεων.
- Η πιθανότητα ένα από τα χαρακτηριστικά που παράγονται εσωτερικά ενός Επιπέδου Εστίασης να είναι μηδέν (Dropout) στο εύρος [0.2, 0.5] (πριν το Πλήρως Συνδεδεμένο Επίπεδο με συνάρτηση ενεργοποίησης την Softmax).

Οι υπερπαράμετροι που σχετίζονται με τη δεύτερη κατηγορία αφορούν τις εξής ρυθμίσεις:

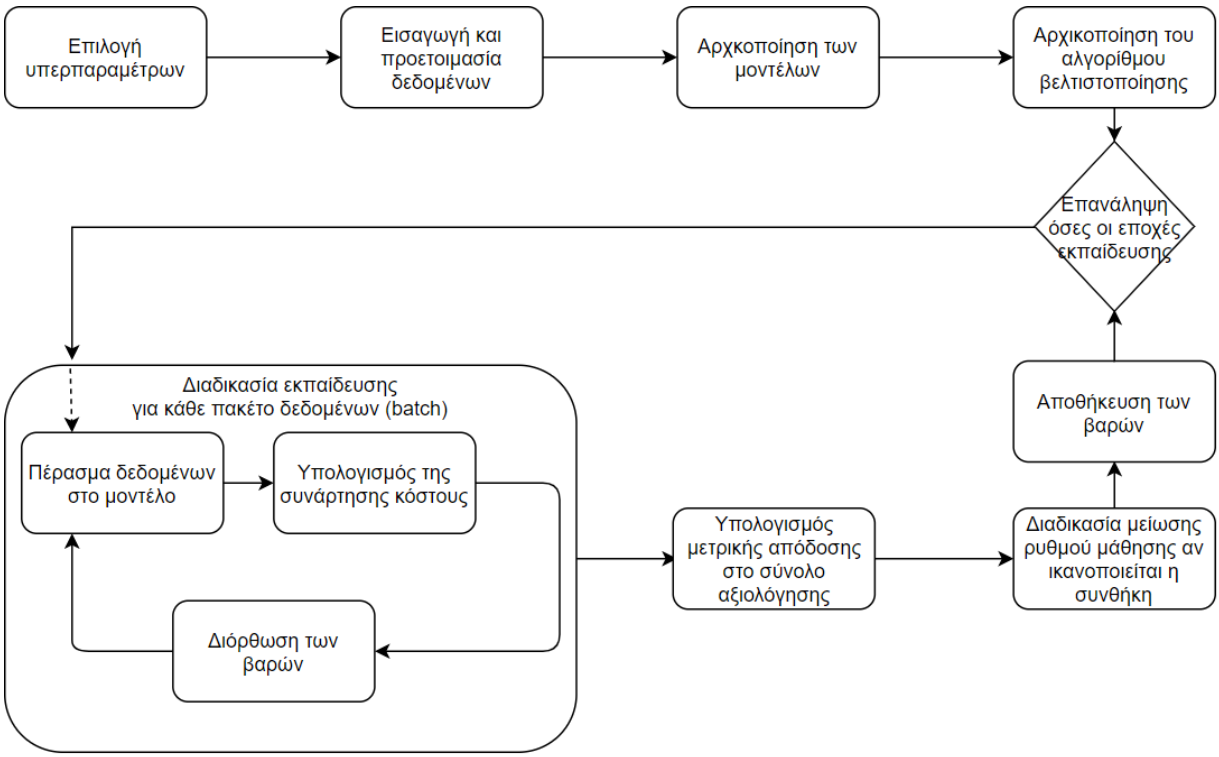
- Ο αριθμός των δεδομένων τα οποία θα επεξεργάζονται ταυτόχρονα σε κάθε βήμα (batch size). Δοκιμάστηκαν τιμές στο εύρος [32, 320]. Τιμές μεγαλύτερες αυτών δημιουργούν προβλήματα στη διαδικασία εκπαίδευσης καθώς απαιτούν εξαιρετικά υψηλούς πόρους μνήμης.
- Μέχρι 50 εποχές (epochs) εκπαίδευσης του συστήματος.
- Ο αλγόριθμος βελτιστοποίησης. Έγινε πειραματισμός μεταξύ των SGD, RMSprop και Adam.
- Ο ρυθμός μάθησης (Learning rate) που χρησιμοποιούν οι αλγόριθμοι. Έγιναν δοκιμές στο εύρος [1e-7, 2e-4]
- Ο όρος ορμής (Momentum) για τους αλγορίθμους SGD και RMSprop. Έγιναν δοκιμές στο εύρος [0.85, 0.95]

4.2.4 Διαδικασία Μείωσης του Ρυθμού Μάθησης

Είναι συνηθισμένη πρακτική να μειώνεται ο ρυθμός μάθησης κατά την εκπαίδευση του μοντέλου σε περίπτωση που δε βελτιώνεται κάποια μετρική απόδοσης. Για το λόγο αυτό στην παρούσα εργασία έγιναν δοκιμές και με αυτή την προσέγγιση πέραν των σταθερών ρυθμών μάθησης για την εκπαίδευση των μοντέλων. Συγκεκριμένα, αν μετά από 2 εποχές δεν είχε βελτιωθεί ο μέσος όρος στην ακρίβεια με βάση τη σχέση 4.1 για το σύνολο ελέγχου τότε μειωνόταν ο ρυθμός μάθησης πολλαπλασιαζόμενος με το παράγοντα 0.5. Σημειώνεται ότι στις δοκιμές που χρησιμοποιήθηκε η διαδικασία αυτή, η αρχική τιμή του ρυθμού μάθησης ήταν 0.0002.

4.2.5 Αλγόριθμος Εκπαίδευσης

Η διαδικασία εκπαίδευσης ήταν η ίδια για όλα τα μοντέλα. Σε πρώτη φάση γίνεται η επιλογή των υπερπαραμέτρων και των μοντέλων που θα δοκιμαστούν. Έπειτα αρχικοποιούνται τα επιλεγμένα μοντέλα και ο αλγόριθμος βελτιστοποίησης. Στη συνέχεια πραγματοποιείται η εκπαίδευση των μοντέλων για το σύνολο των εποχών εκπαίδευσης. Σε κάθε εποχή τα δεδομένα που επεξεργάζονται ταυτόχρονα (batch) περνάνε στα μοντέλα τα οποία με την σειρά τους παράγουν μία εκτίμηση της σωστής απάντησης. Έτσι υπολογίζεται η συνάρτηση κόστους και επιτελείται η διόρθωση των βαρών. Μετά το τέλος κάθε εποχής το συνολικό σύστημα αξιολογείται στο σύνολο αξιολόγησης με βάση την μετρική ακρίβεια της σχέσης 4.1. Στην περίπτωση που η απόδοση φθίνει για δύο συνεχόμενες εποχές ενεργοποιείται η διαδικασία μείωσης του ρυθμού μάθησης. Τέλος αποθηκεύονται τα προσαρμοσμένα βάρη. Στο διάγραμμα που ακολουθεί περιγράφεται η γενική πορεία εκπαίδευσης των δικτύων.



Σχήμα 4.7: Διαδικασία εκπαίδευσης του μοντέλου.

4.3 Αποτελέσματα – Μετρήσεις

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα και οι μετρήσεις της απόδοσης του συστήματος κατά τη διαδικασία εύρεσης των βέλτιστων υπερπαραμέτρων για τα μοντέλα που δοκιμάστηκαν. Αρχικά παρουσιάζονται τα αποτελέσματα πάνω στο σύνολο αξιολόγησης κατά τη διάρκεια εκπαίδευσης. Έπειτα, για την επιλογή των βέλτιστων παραραμέτρων παρουσιάζονται τα αποτελέσματα πάνω στο σύνολο ελέγχου που αξιολογήθηκαν από τον εξυπηρετητή (server) αξιολόγησης που παρέχεται από τους δημιουργούς του συνόλου δεδομένων. Για τα βέλτιστα αποτελέσματα παρουσιάζονται επίσης συγκριτικά αποτελέσματα σε σχέση με άλλες αντίστοιχες αρχιτεκτονικές που έχουν εκπαιδευτεί στο ίδιο σύνολο δεδομένων. Οι συνδυασμοί των μοντέλων που αναλύονται είναι:

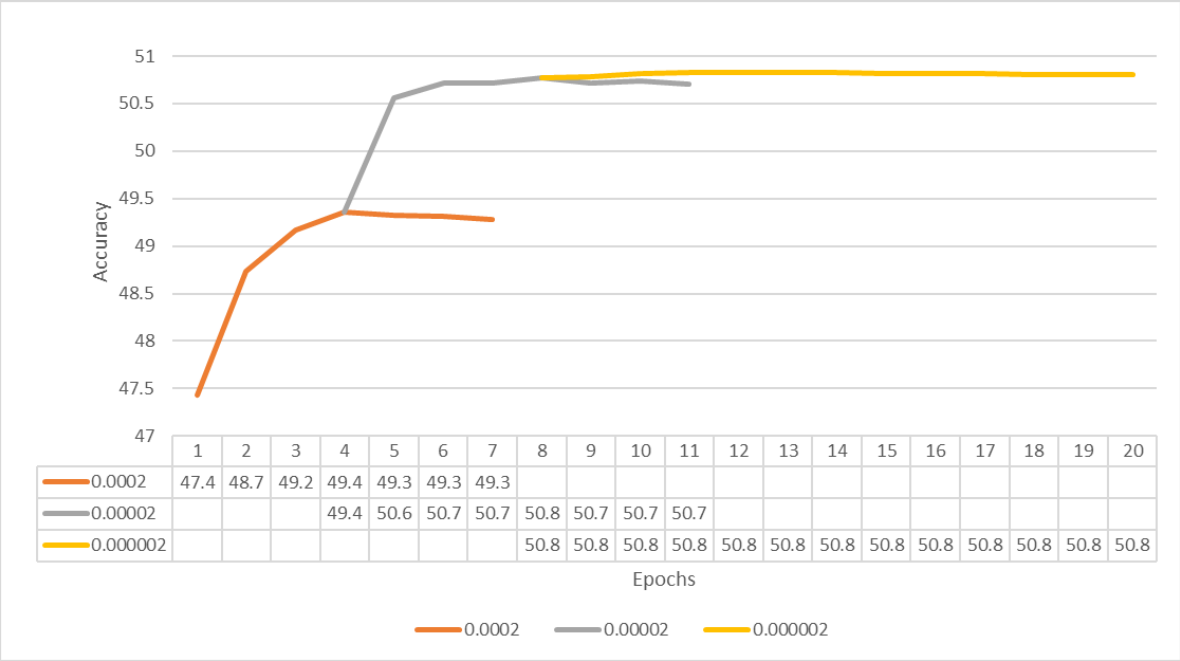
	Αναπαράσταση Εικόνας	Αναπαράσταση Ερώτησης
1	VGGNet-19 (224x224)	Πίνακας εμφύτευσης πάνω σε Λεξικό Λέξεων
2	VGGNet-19 (448x448)	Πίνακας εμφύτευσης πάνω σε Λεξικό Λέξεων
3	DenseNet-161 (224x224)	Πίνακας εμφύτευσης πάνω σε Λεξικό Λέξεων
4	DenseNet-161 (448x448)	Πίνακας εμφύτευσης πάνω σε Λεξικό Λέξεων
5	DenseNet-161 (448x448)	ELMo

Πίνακας 4.1: Συνδυασμοί Μοντέλων που δοκιμάστηκαν.

4.3.1 Αποτελέσματα με VGGNET (224X224) και Πίνακα Εμφύτευσης

Ρυθμός Μάθησης

Ανεξαρτήτως της επιλογής μοντέλων και υπερπαραμέτρων παρατηρήθηκε ότι η επιλογή του ρυθμού μάθησης παίζει πολύ σημαντικό ρόλο. Συγκεκριμένα σε όλα τα πειράματα παρατηρήθηκε ότι η μείωση του ρυθμού μάθησης κατά τη διάρκεια εκπαίδευσης των μοντέλων ήταν αναγκαία για την καλύτερη προσαρμογή τους. Ένα τέτοιο παράδειγμα πάνω στο σύνολο αξιολόγησης εμφανίζεται και στο επόμενο Σχήμα 4.8 για την περίπτωση των μοντέλων VGGNet-19 με είσοδο εικόνας σε ανάλυση 224X224 και Πίνακα Εμφύτευσης για την αναπαράσταση της ερώτησης. Στην περίπτωση αυτή πραγματοποιήθηκαν τρεις διαδοχικές εκπαιδεύσεις του συστήματός με σταθερές όλες τις υπερπαραμέτρους, όπου καθώς συνέκλινε η απόδοσή του συστήματος, ο ρυθμός μάθησης μειωνόταν στο 1/10 της τιμής του. Για τη συγκεκριμένη εκπαίδευση του μοντέλου έγινε χρήση του αλγορίθμου Adam και ο αρχικός ρυθμός μάθησης που εμφανίζεται με πορτοκαλί χρώμα είναι 0.0002. Κατά την τροποποίηση του ρυθμού μάθησης οι τιμές των βαρών ήταν εκείνες όπου επιτυγχάνονταν η προηγούμενη βέλτιστη απόδοση.

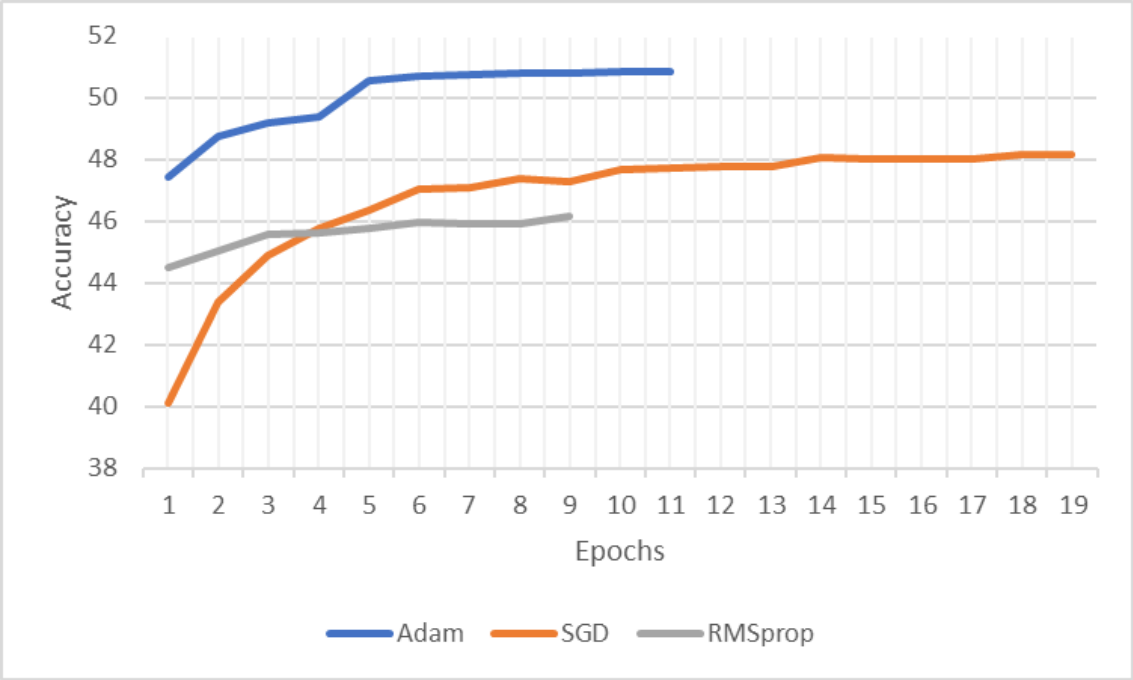


Σχήμα 4.8: Απόδοση στο σύνολο αξιολόγησης ανά εποχές εκπαίδευσης και για διαφορετικό ρυθμό μάθησης.

Λόγω του φαινομένου αυτού όλες οι διαδικασίες εκπαίδευσης που παρουσιάζονται κάνουν χρήση της διαδικασίας μείωσης του ρυθμού μάθησης που παρουσιάστηκε στην προηγούμενη υποενότητα.

Αλγόριθμος Βελτιστοποίησης

Μία ακόμα υπερπαράμετρος που η επιλογή της παίζει σημαντικό ρόλο είναι ο αλγόριθμος βελτιστοποίησης που εμφανίστηκε να μην αλληλεπιδρά με τις υπόλοιπες. Συγκεκριμένα σε όλα τα πειράματα παρατηρήθηκε ότι ο αλγόριθμος Adam εμφανίζει τη βέλτιστη και πιο γρήγορη σύγκλιση σε σχέση με τους υπόλοιπους. Παρατηρήθηκε επίσης ότι ο SGD παρουσιάζει αρκετά ομαλή και σταθερή συμπεριφορά κατά τη σύγκλιση, όμως απαιτεί μεγάλο αριθμό εποχών, χωρίς ταυτόχρονα να βελτιώνει τα αποτελέσματα σε σχέση με τον Adam. Ακόμα ο αλγόριθμος RMSprop δεν εμφάνισε ικανοποιητικά αποτελέσματα ούτε ως προς την τελική απόδοση αλλά ούτε ως προς την ταχύτητα σύγκλισης. Στο Σχήμα 4.9 εμφανίζεται η απόδοση στο σύνολο αξιολόγησης κατά τη διάρκεια εκπαίδευσης για τους διάφορους αλγορίθμους βελτιστοποίησης. Σημειώνεται ότι για κάθε αλγόριθμο βελτιστοποίησης έχουν επιλεχθεί οι βέλτιστες υπερπαράμετροι.



Σχήμα 4.9: Απόδοση στο σύνολο αξιολόγησης ανά εποχές εκπαίδευσης και για διαφορετικό αλγόριθμο βελτιστοποίησης.

Γίνεται σαφές ότι η βέλτιστη επιλογή είναι η εκπαίδευση μέσω του Adam. Όμοια συμπεριφορά παρατηρείται και για τα υπόλοιπα μοντέλα.

Βέλτιστες Υπερπαράμετροι

Για την εύρεση των βέλτιστων υπερπαραμέτρων πραγματοποιήθηκαν αρκετά πειράματα ώστε να εξεταστούν οι πιθανές αλληλεπιδράσεις μεταξύ των παραμέτρων. Η τελική επιλογή των βασικών υπερπαραμέτρων που παρουσιάστηκαν σε προηγούμενη ενότητα εμφανίζεται στον επόμενο πίνακα.

Υπερπαράμετρος	Τιμή
LSTM επίπεδα (Ερώτηση)	1
Hidden Size (Εσωτερική Αναπαράσταση Εικόνας και Ερώτησης)	1024
Επίπεδα Εστίασης	2
Μέγεθος Επιπέδου Εστίασης(Εσωτερική Αναπαράσταση του επιπέδου εστίασης)	512
Dropout (Εικόνα)	0.3
Dropout (Ερώτηση)	0.3
Dropout (Επίπεδο Εστίασης)	0.5
Μέγεθος πακέτων δεδομένων(Batch Size)	64
Αλγόριθμος βελτιστοποίησης	Adam
Αρχικός ρυθμός μάθησης	0.0002
Ορμή	0.9
Εποχές	11

Πίνακας 4.2: Βέλτιστες υπερπαράμετροι για την περίπτωση VGGNET (224X224) και Πίνακα Εμφύτευσης.

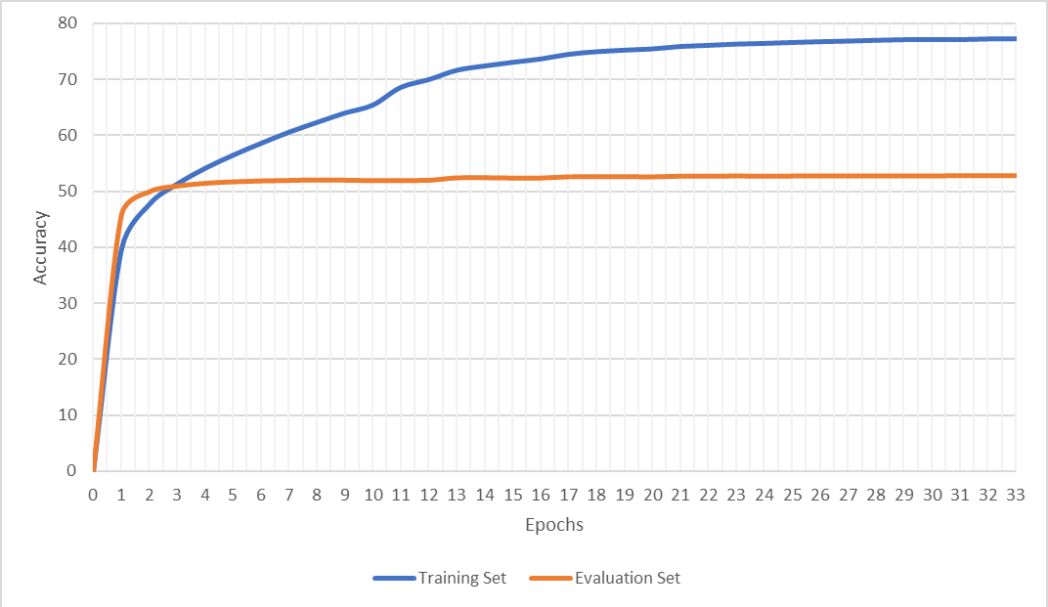
Για τη συγκεκριμένη επιλογή υπερπαραμέτρων ο βέλτιστος αριθμός εποχών ήταν 11. Μετά τις 11 εποχές το σύστημα άρχισε να παρουσιάζει το πρόβλημα της υπερπροσαρμογής. Στον επόμενο πίνακα εμφανίζονται τα αποτελέσματα της ακρίβειας του μοντέλου όπως υπολογίστηκαν από τα διαθέσιμα δεδομένα του συνόλου αξιολόγησης και από το ‘κρυφό’ σύνολο ελέγχου μέσω του εξυπηρετητή (server) που παραχωρούν οι δημιουργοί των δεδομένων για την τελική αξιολόγηση. Τονίζεται ότι με βάση τις βέλτιστες υπερπαραμέτρους το μοντέλο εκπαιδεύτηκε από την αρχή στο σύνολο των δεδομένων εκπαίδευσης και αξιολόγησης πριν περάσουν στην τελική αξιολόγηση από τον επίσημο server. Η μετρική που υπολογίζονται τα αποτελέσματα είναι εκείνη της σχέσης 4.1.

Κατηγορία Απαντήσεων	Σύνολο Αξιολόγησης	Σύνολο Ελέγχου
Ναι/Όχι	70.02	67.83
Αριθμός	33.2	32.56
Άλλο	40.89	40.39
Συνολικά	50.83	**50.91**

Πίνακας 4.3: Απόδοση μοντέλου VGGNET (224X224) και Πίνακα Εμφύτευσης στα σύνολα αξιολόγησης και ελέγχου.

4.3.2 Αποτελέσματα με VGGNET (448X448) και Πίνακα Εμφύτευσης

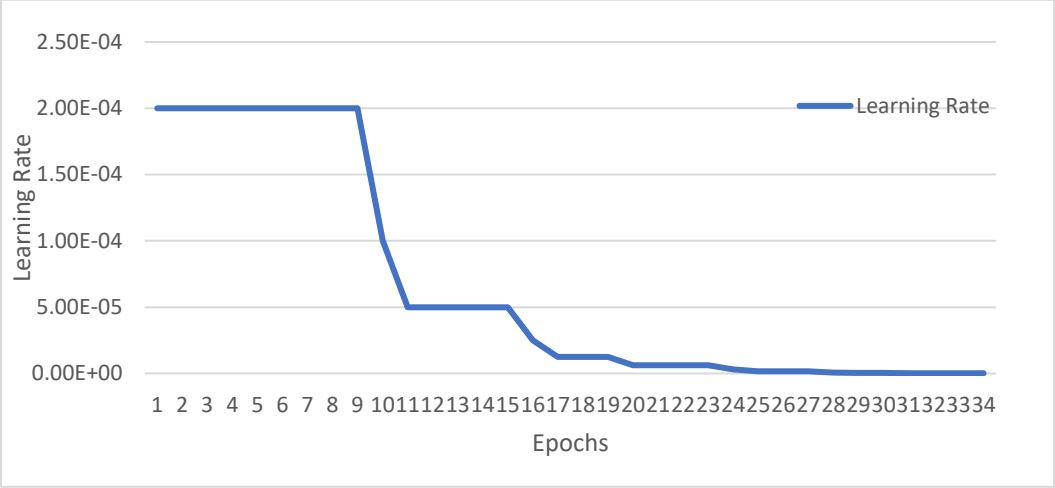
Όπως είναι αναμενόμενο, το σύστημα που χρησιμοποιεί τις εικόνες σε μεγαλύτερη χωρική ανάλυση 442x442 εμφανίζει βελτιωμένα αποτελέσματα σε σχέση με το προηγούμενο που χρησιμοποιεί 224x224. Παρόλα αυτά, η έκδοση του δικτύου με τη μεγαλύτερη ανάλυση εμφάνισε πιο έντονα το πρόβλημα της υπερπροσαρμογής. Ο λόγος είναι ότι στην περίπτωση αυτή το δίκτυο είναι πιο μεγάλο με αρκετές περισσότερες εκπαιδευόμενες παραμέτρους. Από την μία το δίκτυο είναι σε θέση να αναπαραστήσει αντικείμενα τα οποία απαιτούν μεγαλύτερη λεπτομέρεια όμως από την άλλη είναι πιο επιρρεπές στον θόρυβο που περιέχεται στο σύνολο δεδομένων. Επίσης μετά από αναζήτηση των βέλτιστων παραμέτρων υπήρξε ανάγκη για μεγαλύτερο μέγεθος στα επίπεδα πολλαπλών εστιάσεων από 512 (για 224x224) σε 2048 (για 448x448). Λόγω των παραπάνω η βέλτιστη επιλογή του αριθμού των δεδομένων που επεξεργάζονται (batch size) ταυτόχρονα αυξήθηκε από 64 σε 300. Το τελευταίο είναι αναμενόμενο καθώς με μεγαλύτερο batch size τα μοντέλα τείνουν να επιτυγχάνουν πιο αργή σύγκλιση αλλά είναι λιγότερο επιρρεπή σε τοπικά μέγιστα. Στο επόμενο Σχήμα 4.10 εμφανίζεται η επίδοση του συστήματος με τις βέλτιστες παραμέτρους κατά τη διάρκεια εκπαίδευσης. Παρατηρείται ότι υπάρχει αύξηση της απόδοσης και για τα δύο σύνολα δεδομένων έως και την εποχή 33. Επίσης ο ρυθμός αύξησης στην περίπτωση του συνόλου ελέγχου εμφανίζεται να φθίνει δραματικά μετά την 4^{η} εποχή κάτι το οποίο δεν συμβαίνει στο σύνολο εκπαίδευσης.



Σχήμα 4.10: Απόδοση στο σύνολο αξιολόγησης και εκπαίδευσης ανά εποχές εκπαίδευσης για το VGGNET (448X448) και Πίνακα Εμφύτευσης.

Από το γράφημα γίνεται αντιληπτό ότι τα δεδομένα που δίδονται για εκπαίδευση δεν είναι αντιπροσωπευτικά του συνολικού προβλήματος και δεν παρέχουν επαρκή πληροφορία που να σχετίζεται με το σύνολο αξιολόγησης. Ο λόγος είναι ότι όχι μόνο δεν παρατηρείται μείωση της

απόδοσης στο σύνολο αξιολόγησης, αλλά αντίθετα μια σταδιακή μικρή αύξηση. Η αιτία που παρατηρείται αυτή η συμπεριφορά αφορά την επιλογή που έχει γίνει στο σύνολο εκπαίδευσης να χρησιμοποιούνται μόνο οι ερωτήσεις που αντιστοιχούν στις 1000 πιο συχνές απαντήσεις. Υπενθυμίζεται ότι αντιστοιχούν στο 87,47% των ερωτήσεων του συνόλου εκπαίδευσης και μόνο αυτό το υποσύνολο χρησιμοποιείται για εκπαίδευση. Ουσιαστικά ένα υποσύνολο του συνόλου αξιολόγησης περιέχει παραδείγματα για τα οποία δεν υπάρχουν αντίστοιχα στο σύνολο εκπαίδευσης. Συνεπώς είναι αναμενόμενο να εμφανίζεται πιο αυξημένη η απόδοση στο σύνολο εκπαίδευσης και επομένως δεν χαρακτηρίζεται ως υπερπροσαρμογή σε αυτά.
Στη συνέχεια στο Σχήμα 4.11 απεικονίζεται η μείωση του ρυθμού μάθησης κατά τη διάρκεια εκπαίδευσης.



Σχήμα 4.11: Μείωση του ρυθμού μάθησης κατά τη διάρκεια εκπαίδευσης για το VGGNET (448X448) και Πίνακα Εμφύτευσης.

Στον επόμενο πίνακα εμφανίζονται οι βέλτιστες υπερπαράμετροι και στη συνέχεια τα αποτελέσματα από τα δεδομένα του συνόλου αξιολόγησης και τα αποτελέσματα από τα δεδομένα του συνόλου ελέγχου μέσω του εξυπηρετητή (server) για την τελική αξιολόγηση.

Υπερπαράμετρος	LSTM επίπεδα (Ερώτηση)	Hidden Size (Εσωτερική Αναπαράσταση Εικόνας και Ερώτησης)	Επίπεδα Εστίασης	Μέγεθος Επιπέδου Εστίασης(Εσωτερική Αναπαράσταση του επιπέδου εστίασης)	Dropout (Εικόνα)	Dropout (Ερώτηση)	Dropout (Επίπεδο Εστίασης)	Μέγεθος πακέτων δεδομένων (Batch Size)	Αλγόριθμος βελτιστοποίησης	Αρχικός ρυθμός μάθησης	Ορμή	Εποχές
Τιμή	1	1024	2	2048	0.3	0.3	0.5	300	Adam	0.0002	0.9	33

Πίνακας 4.4: Βέλτιστες υπερπαράμετροι για την περίπτωση VGGNET (448X448) και Πίνακα Εμφύτευσης.

Κατηγορία Απαντήσεων	Σύνολο Αξιολόγησης	Σύνολο Ελέγχου
Ναι/Όχι	71.38	70.33
Αριθμός	33.39	32.82
Άλλο	43.92	44.39
Συνολικά	52.85	**53.88**

Πίνακας 4.5: Απόδοση μοντέλου VGGNET (448X448) και Πίνακα Εμφύτευσης στα σύνολα αξιολόγησης και ελέγχου.

Με βάση τα αποτελέσματα από τον επίσημο server η συνολική απόδοση του μοντέλου αυξήθηκε από 50.91 (224x224) σε 53.88 (448x448) και κρίνεται ικανοποιητική. Υπενθυμίζεται ότι κατά την τελική αξιολόγηση στον επίσημο server τα μοντέλα έχουν προσαρμοστεί εκ' νέου στο σύνολο των διαθέσιμων δεδομένων (εκπαίδευσης και αξιολόγησης).

4.3.3 Αποτελέσματα με DenseNet (224X224) και Πίνακα Εμφύτευσης

Στη συνέχεια παρουσιάζονται τα αποτελέσματα των βέλτιστων παραμέτρων όπου αντί για το μοντέλο VGGNet-19 χρησιμοποιείται το DenseNet-161 με ανάλυση 224x224 για την εικόνα εισόδου.

Υπερπαράμετρος	LSTM επίπεδα (Ερώτηση)	Hidden Size (Εσωτερική Αναπαράσταση Εικόνας και Ερώτησης)	Επίπεδα Εστίασης	Μέγεθος Επιπέδου Εστίασης(Εσωτερική Αναπαράσταση του επιπέδου εστίασης)	Dropout (Εικόνα)	Dropout (Ερώτηση)	Dropout (Επίπεδο Εστίασης)	Μέγεθος πακέτων δεδομένων (Batch Size)	Αλγόριθμος βελτιστοποίησης	Αρχικός ρυθμός μάθησης	Ορμή	Εποχές
Τιμή	1	1024	2	512	0.3	0.3	0.5	64	Adam	0.0002	0.9	22

Πίνακας 4.6: Βέλτιστες υπερπαράμετροι για την περίπτωση DenseNet (224X224) και Πίνακα Εμφύτευσης.

Επίσης τα αποτελέσματα στο σύνολο αξιολόγησης και στο τελικό σύνολο ελέγχου είναι:

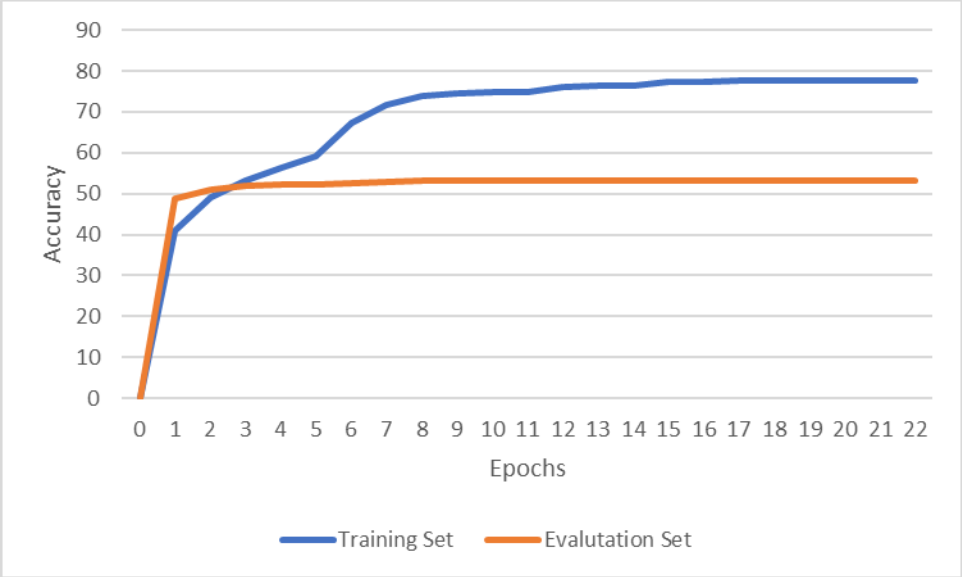
Κατηγορία Απαντήσεων	Σύνολο Αξιολόγησης	Σύνολο Ελέγχου
Ναι/Όχι	69.5	66.31
Αριθμός	32.68	31.49
Άλλο	42.15	42.14
Συνολικά	51.19	**51**

Πίνακας 4.7: Απόδοση μοντέλου DenseNet (224X224) και Πίνακα Εμφύτευσης στα σύνολα αξιολόγησης και ελέγχου.

Υπενθυμίζεται ότι το DenseNet-161 παράγει πολύ περισσότερα χαρακτηριστικά (2208) σε σχέση με το VGGNet-19 (512) και για το λόγο αυτό η αύξηση της συνολικής απόδοσης από 50.91 σε 51.00 δεν κρίνεται ικανοποιητική.

4.3.4 Αποτελέσματα με DenseNet (448X448) και Πίνακα Εμφύτευσης

Στην έκδοση αυτή που γίνεται χρήση του DenseNet με 448x448 ανάλυση εικόνας, τα αποτελέσματα εμφανίζονται αρκετά πιο βελτιωμένα και σε σχέση με εκείνα των 224x224 αλλά και με τα αντίστοιχα 448x448 με το VGGNet μοντέλο. Η πορεία που ακολούθησε η επίδοση του μοντέλου κατά τη διάρκεια της εκπαίδευσης στα σύνολα εκπαίδευσης και αξιολόγησης εμφανίζεται στο επόμενο Σχήμα 4.12. Όμοια με πριν, η επίδοση και στα δύο σύνολα αυξάνεται κατά τη διάρκεια της εκπαίδευσης. Και σε αυτή την περίπτωση παρατηρείται η αυξημένη απόδοση στο σύνολο εκπαίδευσης σε σχέση με εκείνης του συνόλου αξιολόγησης, που αιτιολογείται με την επιλογή να χρησιμοποιηθεί το 87,47% των ερωτήσεων εκπαίδευσης.



Σχήμα 4.12: Απόδοση στο σύνολο αξιολόγησης και εκπαίδευσης ανά εποχές εκπαίδευσης για το DenseNet (448X448) και Πίνακα Εμφύτευσης.

Η επιλογή των βέλτιστων υπερπαραμέτρων η οποία ταυτίζεται με τη περίπτωση του VGGNet-19 είναι:

Υπερπαράμετρος	LSTM επίπεδα (Ερώτηση)	Hidden Size (Εσωτερική Αναπαράσταση Εικόνας και Ερώτησης)	Επίπεδα Εστίασης	Μέγεθος Επιπέδου Εστίασης(Εσωτερική Αναπαράσταση του επιπέδου εστίασης)	Dropout (Εικόνα)	Dropout (Ερώτηση)	Dropout (Επίπεδο Εστίασης)	Μέγεθος πακέτων δεδομένων (Batch Size)	Αλγόριθμος βελτιστοποίησης	Αρχικός ρυθμός μάθησης	Ορμή	Εποχές
Τιμή	1	1024	2	2048	0.3	0.3	0.5	300	Adam	0.0002	0.9	22

Πίνακας 4.8: Βέλτιστες υπερπαράμετροι για την περίπτωση DenseNet (448X448) και Πίνακα Εμφύτευσης.

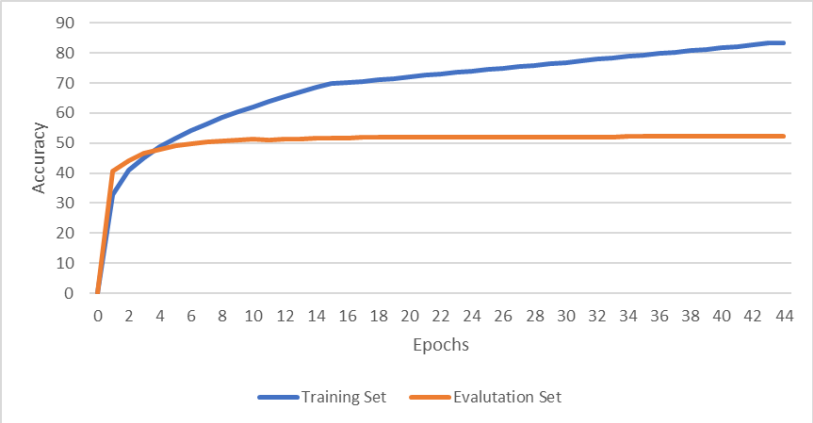
Στον επόμενο πίνακα εμφανίζονται τα τελικά αποτελέσματα για το μοντέλο DenseNet-161 με 448x448 εικόνα εισόδου. Παρατηρείται αύξηση 4.23 μονάδων σε σχέση με το σύστημα που κάνει χρήση του DenseNet με 224x224 ανάλυση εικόνας εισόδου και η αύξηση αυτή κρίνεται αρκετά ικανοποιητική.

Κατηγορία Απαντήσεων	Σύνολο Αξιολόγησης	Σύνολο Ελέγχου
Ναι/Όχι	70.63	71.31
Αριθμός	33.82	33.83
Άλλο	45.15	46.13
Συνολικά	53.24	**55.23**

Πίνακας 4.9: Απόδοση μοντέλου DenseNet (448X448) και Πίνακα Εμφύτευσης στα σύνολα αξιολόγησης και ελέγχου.

4.3.5 Αποτελέσματα με DenseNet (448X448) και ELMo

Ο τελευταίος συνδυασμός μοντέλων που δοκιμάστηκαν είναι του DenseNet με ανάλυση εικόνας εισόδου 448x448 και του ELMo αντί για Πίνακα εμφύτευσης. Στην συγκεκριμένη περίπτωση παρατηρήθηκε ελαφρώς ομαλότερη σύγκλιση αλλά χρειάστηκαν περισσότερες εποχές εκπαίδευσης (Σχήμα 13). Επίσης τα αποτελέσματα στο σύνολο αξιολόγησης δεν εμφανίζουν αύξηση αλλά αντιθέτως μια μικρή μείωση.



Σχήμα 4.13: Απόδοση στο σύνολο αξιολόγησης και εκπαίδευσης ανά εποχές εκπαίδευσης για το DenseNet (448X448) και ELMo.

Στη συνέχεια εμφανίζεται η βέλτιστη επιλογή υπερπαραμέτρων η όποια συμπίπτει με την αντίστοιχη επιλογή όπου αντί για το ELMo χρησιμοποιήθηκε ο Πίνακας Εμφύτευσης. Η μόνη διαφορά είναι στις εποχές εκπαίδευσης όπου αντί για 22 χρειάστηκαν 44 εποχές.

Υπερπαράμετρος	LSTM επίπεδα (Ερώτηση)	Hidden Size (Εσωτερική Αναπαράσταση Εικόνας και Ερώτησης)	Επίπεδα Εστίασης	Μέγεθος Επιπέδου Εστίασης(Εσωτερική Αναπαράσταση του επιπέδου εστίασης)	Dropout (Εικόνα)	Dropout (Ερώτηση)	Dropout (Επίπεδο Εστίασης)	Μέγεθος πακέτων δεδομένων (Batch Size)	Αλγόριθμος βελτιστοποίησης	Αρχικός ρυθμός μάθησης	Ορμή	Εποχές
Τιμή	1	1024	2	2048	0.3	0.3	0.5	300	Adam	0.0002	0.9	44

Πίνακας 4.10: Βέλτιστες υπερπαράμετροι για την περίπτωση DenseNet (448X448) και ELMo.

Στον επόμενο πίνακα εμφανίζονται τα αποτελέσματα για τις βέλτιστες παραμέτρους στο σύνολο αξιολόγησης και τελικού ελέγχου.

Κατηγορία Απαντήσεων	Σύνολο Αξιολόγησης	Σύνολο Ελέγχου
Ναι/Όχι	70.38	69.79
Αριθμός	33.3	32.46
Άλλο	43.47	44.96
Συνολικά	52.25	**53.89**

Πίνακας 4.11: Απόδοση μοντέλου DenseNet (448X448) και ELMo στα σύνολα αξιολόγησης και ελέγχου.

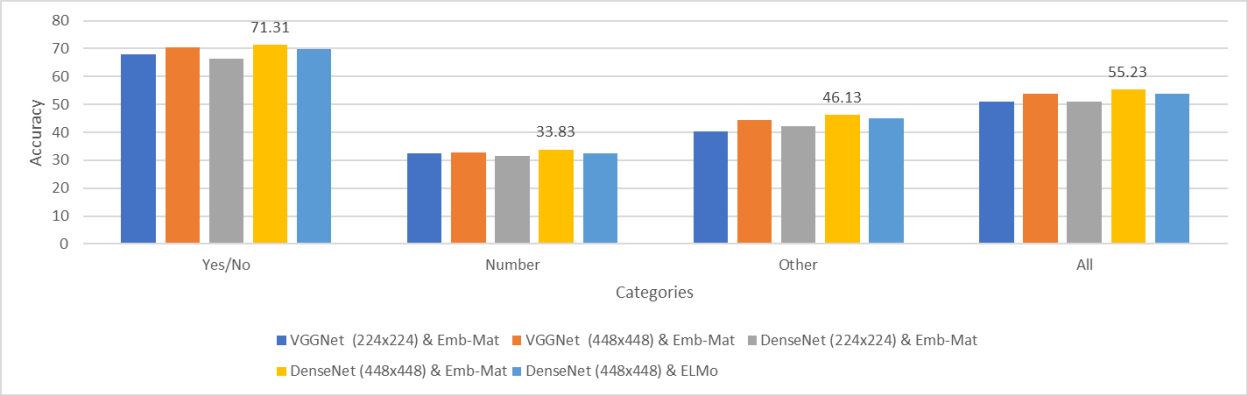
Τα αποτελέσματα δεν είναι ιδιαίτερα ικανοποιητικά καθώς με τη χρήση ενός αρκετά πιο σύνθετου μοντέλου αναπαράστασης των λέξεων επιτυγχάνεται μια ελαφρώς μειωμένη απόδοση κατά 1.34.

4.3.6 Συγκριτικά Αποτελέσματα και Συνδυασμός Μοντέλων

Με βάση τα προηγούμενα αποτελέσματα ο βέλτιστος συνδυασμός μοντέλων αναπαράστασης της εικόνας και της ερώτησης είναι μέσω του DenseNet-161 (448x448) και του Πίνακα Εμφύτευσης. Τα συνολικά αποτελέσματα πάνω στο τελικό σύνολο ελέγχου καθώς και η οπτική τους αναπαράσταση για κάθε κατηγορία απάντησης εμφανίζονται στον επόμενο πίνακα 4.12 και Σχήμα 4.14.

	VGGNet (224x224) & Emb-Mat	VGGNet (448x448) & Emb-Mat	DenseNet (224x224) & Emb-Mat	DenseNet (448x448) & Emb-Mat	DenseNet (448x448) & ELMo
Yes/No	67.83	70.33	66.31	71.31	69.79
Number	32.56	32.82	31.49	33.83	32.46
Other	40.39	44.39	42.14	46.13	44.96
All	**50.91**	**53.88**	**51**	**55.23**	**53.89**

Πίνακας 4.12: Συγκεντρωτικά αποτελέσματα στο σύνολο ελέγχου για κάθε κατηγορία απάντησης.



Σχήμα 4.14: Συγκεντρωτικά αποτελέσματα στο σύνολο ελέγχου για κάθε κατηγορία απάντησης.

Από τα αποτελέσματα προκύπτει ότι το σύστημα με DenseNet (448x448) και Πίνακα Εμφύτευσης παρουσιάζει καλύτερα αποτελέσματα σε σχέση με τα υπόλοιπα σε όλες τις επιμέρους κατηγορίες απαντήσεων.

Για την αξιοποίηση του συνόλου των 5 διαφορετικών συστημάτων, αποφασίστηκε να δοκιμαστεί η απόδοση τους σε ένα σύστημα που συνδυάζει τις αποφάσεις τους. Έτσι έγινε πειραματισμός με δύο εκδοχές. Η πρώτη αφορά ένα σύστημα που επιστρέφει ως την τελική απάντηση εκείνη που επικρατεί μεταξύ των διαφορετικών απαντήσεων που επιστρέφουν τα 5 επιμέρους συστήματα (Majority Ensemble Models). Ουσιαστικά τα 5 διαφορετικά συστήματα ψηφίζουν για την τελική απάντηση. Η δεύτερη αφορά ένα σύστημα όπου η τελική απάντηση προκύπτει πάλι μέσω 'ψηφοφορίας' όμως στην περίπτωση αυτή υπάρχει και στάθμιση (weights) στην ψήφο του κάθε επιμέρους συστήματος. Η στάθμιση αυτή προκύπτει από την συνολική απόδοση του κάθε συστήματος πάνω στο σύνολο αξιολόγησης (Weighted Ensemble Models). Έτσι για παράδειγμα η εκτίμηση που κάνει το DenseNet (448x448) με Πίνακα εμφύτευσης έχει μεγαλύτερη συνεισφορά στην τελική απάντηση από εκείνη που κάνει το VGGNet (224x224) με Πίνακα εμφύτευσης.

Η συνολική αύξηση της απόδοσης για το σύστημα χωρίς στάθμιση σε σχέση με το βέλτιστο μεμονωμένο υποσύστημα είναι 1.21% ενώ για το σύστημα με στάθμιση είναι 1.40%.

Τέλος παρουσιάζονται τα αποτελέσματα των δύο συστημάτων που κάνουν χρήση των 5 επιμέρους συστημάτων συγκριτικά με το βέλτιστο μεμονωμένο σύστημα. Επίσης εμφανίζονται συγκριτικά αποτελέσματα σε σχέση με μοντέλα που προτείνουν οι δημιουργοί των δεδομένων και αναφέρονται ως βασικό σημείο αναφοράς (Baseline) και σύγκρισης σε σχέση με άλλα. Τα αποτελέσματα αυτά αφορούν το τελικό σύνολο ελέγχου μέσω του επίσημου server. Επίσης αναφέρεται ότι η εκπαίδευση τους έχει γίνει στα σύνολα εκπαίδευσης και αξιολόγησης μαζί.

Approach	Yes/No	Number	Other	All
VQAteam 'Yes' Model	61.26	0	0	25.98
VQAteam Question Only	66.79	31.75	27.64	44.34
VQAteam Deeper LSTM Q + I	72.99	35.52	41.91	54.08
DenseNet (448x448) & LSTM	71.31	33.83	46.13	55.23
Majority Ensemble Models	71.13	35.18	48.51	56.44
Weighted Ensemble Models	71.14	35.5	48.84	**56.63**

Πίνακας 4.13: Αποτελέσματα των Ensemble μοντέλων στο σύνολο ελέγχου για κάθε κατηγορία

Το πρώτο μοντέλο που αναφέρεται ως VQAteam ‘Yes’ Model ουσιαστικά παράγει σταθερά την απάντηση ‘Yes’ ανεξαρτήτως της εισόδου του. Όπως είναι αναμενόμενο έχει πολύ χαμηλή συνολική απόδοση.
Στη συνέχεια το μοντέλο που αναγράφεται ως VQAteam Question Only δέχεται σαν είσοδο μόνο την ερώτηση και όχι την εικόνα. Συγκεκριμένα η αναπαράσταση της ερώτησης γίνεται μέσω Πίνακα Εμφύτευσης για τις λέξεις και μέσω LSTM για το σύνολο της ερώτησης όπως και στην παρούσα εργασία. Η τελική του έξοδος προκύπτει μέσω ενός Πλήρως Συνδεδεμένου Επιπέδου με συνάρτηση ενεργοποίησης την Softmax. Η συνολική του απόδοση είναι 44.34% και αξίζει να τονιστεί ότι ουσιαστικά αποτελεί έναν δείκτη για την δυνατότητα να απαντηθούν οι ερωτήσεις του συνόλου ελέγχου χωρίς καμία οπτική πληροφορία.
Ως VQAteam Deeper LSTM Q+I αναφέρεται το μοντέλο που αποτελεί το ουσιαστικό σημείο αναφοράς του διαγωνισμού και μαζί με την ερώτηση κάνει χρήση και της εικόνας εισόδου. Συγκεκριμένα είναι ισοδύναμο με το μοντέλο που χρησιμοποιεί VGGNet (448x448) και Πίνακα Εμφύτευσης στην παρούσα εργασία με τη διαφορά ότι για το συνδυασμό των αναπαραστάσεων της εικόνας και της ερώτησης δεν κάνει χρήση των Επιπέδων Εστίασης αλλά απλώς οι δύο αναπαραστάσεις πολλαπλασιάζονται κατά σημείο. Μία ακόμα διαφορά είναι ότι οι αναπαραστάσεις της εικόνας κανονικοποιούνται με βάση την L2 νόρμα.
Το τελικό σύστημα που χρησιμοποιεί τα 5 επιμέρους συστήματα με στάθμιση στις εκτιμήσεις του καθενός μεμονωμένου επιτυγχάνει το 56.63 που είναι το βέλτιστο αποτέλεσμα της παρούσας εργασίας.

4.4 Αξιολόγηση – Διαγνωστικός Έλεγχος

Στην ενότητα αυτή γίνεται αξιολόγηση του βέλτιστου μεμονωμένου μοντέλου (DenseNet161 (448x448) & Πίνακα Εμφύτευσης) μέσω συγκεκριμένων παραδειγμάτων. Επίσης παρουσιάζονται αποτελέσματα που σχετίζονται με ενδιάμεσες αναπαραστάσεις στο εσωτερικό του συστήματος. Συγκεκριμένα παρουσιάζονται διαγράμματα σχετικά με την ικανότητα του μοντέλου να αναπαραστήσει μια ερώτηση και την ικανότητα να εστιάσει σε περιοχές που σχετίζονται με την ερώτηση. Τέλος γίνεται ανάλυση των σφαλμάτων που παρουσιάζει το σύστημα και γίνεται προσπάθεια για την κατηγοριοποίηση τους.

4.4.1 Έλεγχος Λειτουργίας

Στο σημείο αυτό γίνεται ο έλεγχος λειτουργίας μέσα από κάποια παραδείγματα που παρουσιάζονται παρακάτω. Στα παραδείγματα αυτά υπάρχουν δεδομένα από το Σύνολο Ελέγχου αλλά και γενικότερα δεδομένα εκτός του συνόλου VQA v2. Το σύνολο των ερωτήσεων που θέτονται προς το σύστημα διαμορφώθηκαν στην παρούσα εργασία και δεν αποτελούν μέρος του VQA v2 συνόλου δεδομένων. Στη συνέχεια εμφανίζονται οι εικόνες και οι αντίστοιχες ερωτήσεις που τέθηκαν. Με πράσινο χρώμα παρουσιάζεται η απάντηση που ήταν σωστή και με κόκκινο η απάντηση που ήταν λάθος.
Στην πρώτη εικόνα η οποία εμπεριέχεται στο Σύνολο Ελέγχου, εμφανίζεται ένας άντρας να παίζει μπέιζμπολ. Οι ερωτήσεις που τέθηκαν προς το σύστημα εμφανίζονται στην επόμενη εικόνα.

Q: What sport is this? A: Baseball

Q: What color is his shoes? A: Black

Q: What color is the ground? A: Brown

Q: Does he wear glasses? A: No

Q: Where is the ball? A: In the air

Q: What color is his shirt? A: Gray

Q: How many people are there? A: 3

Σχήμα 4.15: Άντρας παίζει μπέιζμπολ.

Παρατηρείται ότι το σύστημα αποδίδει ικανοποιητικά καλά καθώς αναγνωρίζει το άθλημα και αρκετά στοιχεία της εικόνας. Από την άλλη, αδυνατεί να εκτιμήσει σωστά το χρώμα της μπλούζας το οποίο για έναν ανθρώπινο παρατηρητή είναι το λευκό αλλά αποδίδεται ως γκρι. Ακόμα δεν εκτιμάει σωστά το πλήθος των ανθρώπων που απεικονίζονται.

Ένα άλλο παράδειγμα είναι το επόμενο, όπου στην πολύ γενική ερώτηση ‘τι είναι αυτό’ η απάντηση δίδεται σωστά και είναι η ‘τραπεζαρία’. Μάλιστα η δεύτερη πιο πιθανή απάντηση που δόθηκε είναι η λέξη ‘λίμνη’ και η τρίτη στη σειρά η λέξη ‘νερό’. Ακόμα αναγνωρίζει την ύπαρξη πιάτων αλλά αδυνατεί να εκτιμήσει σωστά το πλήθος τους. Επίσης στην ερώτηση αν υπάρχει φαγητό στο τραπέζι, η απάντηση είναι λανθασμένα ‘ναι’.

Q: What is this? A: Dining room

Q: Are there any dishes? A: Yes

Q: How many glasses are there? A: 6

Q: Is there any food at the table? A: Yes

Σχήμα 4.16: Τραπεζαρία χωρίς φαγητό, δίπλα στη θάλασσα

Στη συνέχεια παρατηρούμε μία εικόνα από μια λεωφόρο. Το σύστημα εμφανίζει και σε αυτή την περίπτωση αρκετά ικανοποιητικά αποτελέσματα. Αξίζει να αναφερθεί ότι στην συγκεκριμένη εικόνα είχε μεγάλη σημασία ο τρόπος του τίθονταν οι ερωτήσεις. Για παράδειγμα στην περίπτωση που αντί της ερώτησης ‘τι χρώμα είναι το αμάξι στην μέση?’ τίθονταν η ερώτηση ‘τι χρώμα είναι το αμάξι στα αριστερά’ το σύστημα αδυνατούσε να απαντήσει σωστά. Επίσης αν η ερώτηση σχετικά με της καιρικές συνθήκες ήταν διατυπωμένη διαφορετικά θα εμφανιζόταν η σωστή απάντηση ‘αίθριος’ (sunny).

Q: What is the color of the car in the middle?
A: White

Q: What is the color of the right car? A: Red

Q: What is this in front of the car? A: Car

Q: What time is it? A: Evening

Q: What are the weather conditions? A: Cloudy

Q: How many cars are there? A: 3

Σχήμα 4.17: Λεωφόρος

Στην επόμενη εικόνα εμφανίζεται ένας αστερίας. Είναι αναμενόμενο ότι η απάντηση ‘starfish’ δεν εμπεριέχεται στις επιλεγμένες 1000 απαντήσεις. Παρόλα αυτά το σύστημα στην

προσπάθεια να εκτιμήσει σωστά την απάντηση επιστρέφει την απάντηση ‘star’. Η απάντηση αυτή δεν μπορεί να θεωρηθεί ως λανθασμένη στην συγκεκριμένη γενική ερώτηση καθώς αποτελεί το σχήμα του αντικειμένου που αναπαρίσταται. Στις επόμενες δύο ερωτήσεις το σύστημα δίνει τη σωστή απάντηση και στην τελευταία ερώτηση που δεν έχει καμία σχέση με την εικόνα το σύστημα δίνει λάθος απάντηση. Παρόλα αυτά στην τελευταία αυτή ερώτηση η τρίτη πιο πιθανή απάντηση είναι ‘nowhere’ , δηλαδή πουθενά, το οποίο αποτελεί τη σωστή απάντηση.

Q: What is this? A: Star

Q: What is under the starfish? A: Rocks

Q: Is there any hand? A: Yes

Q: Where is the monkey? A: On ground

Σχήμα 4.18: Αστερίας πάνω σε πέτρα.

Στην επόμενη εικόνα τίθονται αρχικά ερωτήσεις απλής λογικής και στη συνέχεια ερωτήσεις που απαιτούν να ακολουθηθεί μια πιο σύνθετη συλλογιστική διαδικασία ώστε να απαντηθούν.

Q: What is the animal on the left? A: Dog

Q: What color is the ground? A: Gray

Q: What type of ball is this? A: Soccer

Q: Where is the wood? A: Floor

Q: What is sitting inside the plate? A: Cats

Q: How many animals are there? A: 2

Q: What is in front of the dog? A: People

Q: What does the cat do? A: Jump

Σχήμα 4.19: Ένας σκύλος, μία γάτα μέσα σε ένα μπολ και μία μπάλα ποδοσφαίρου.

Στην προηγούμενη εικόνα το σύστημα εμφάνισε ιδιαίτερα ικανοποιητική απόδοση, καθώς μπόρεσε να ανταποκριθεί σωστά σε ένα μεγάλο εύρος ερωτήσεων. Επίσης στην ερώτηση ‘τι βρίσκεται μέσα στο πιάτο’ η όποια σε καμία περίπτωση δεν μπορεί να απαντηθεί σωστά χωρίς να εξαχθεί σωστά η οπτική πληροφορία, το σύστημα απάντηση σωστά ‘γάτα’.
Τέλος εμφανίζεται μία εικόνα που απεικονίζει ένα γαϊδούρι. Στις 1000 απαντήσεις που έχει την δυνατότητα να απαντήσει δεν εμπεριέχεται αυτή η επιλογή, παρόλα αυτά το σύστημα είναι σε θέση να απαντήσει το άλογο που είναι συγγενικό ζώο και για αυτό το λόγο η απάντηση γίνεται δεκτή.

Q: What animal is this? A: Horse

Q: What color is his t-shirt? A: Green

Q: What is on the ground? A: Dirt

Q: What color is his pants? A: Blue

Q: What does he hold? A: Dirt

Q: Does he wear glasses? A: No

Σχήμα 4.20: Ένας άντρας και ένας γάιδαρος.

4.4.2 Αναπαράσταση Ερωτήσεων σε Μικρότερη Διάσταση

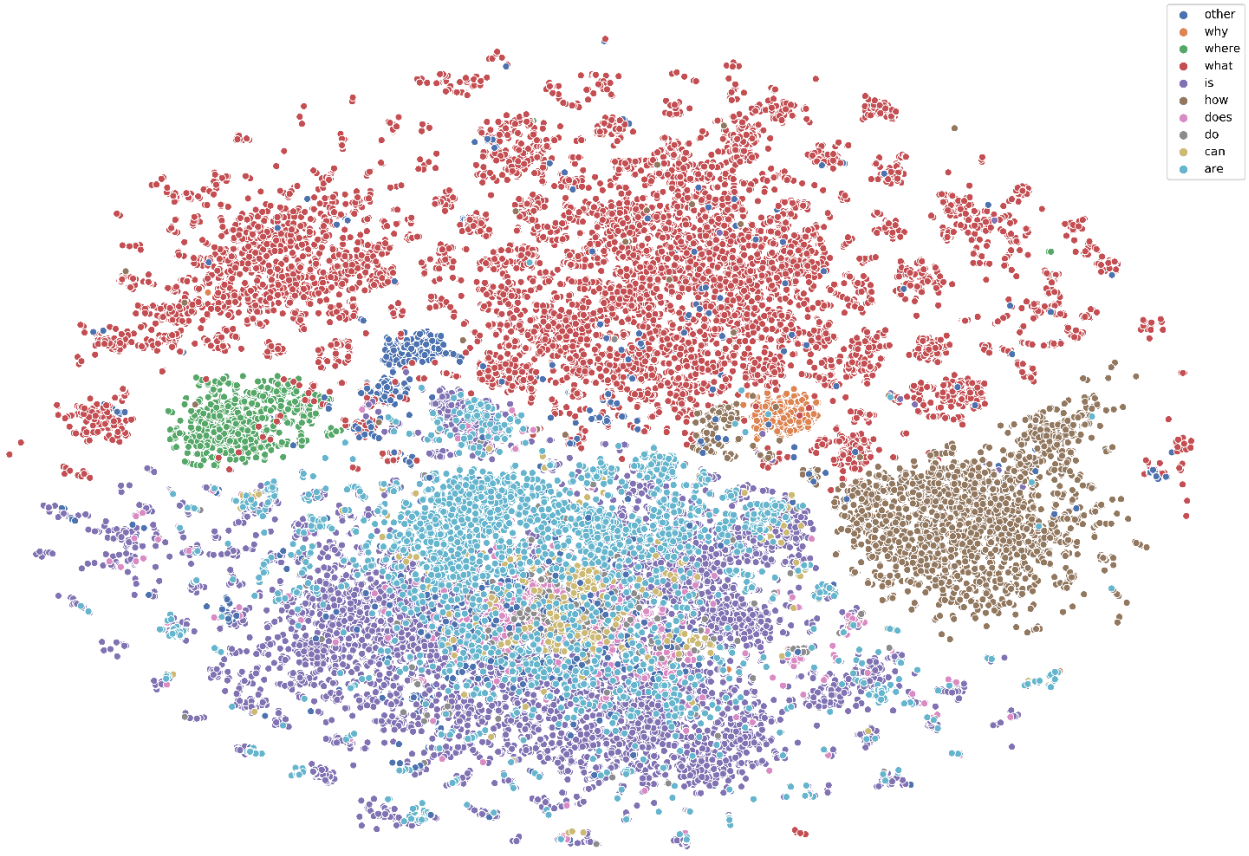
Για την καλύτερη κατανόηση της λειτουργίας του συστήματος έγινε προσπάθεια για οπτικοποίηση των αναπαραστάσεων των διαφορετικών κατηγοριών ερωτήσεων. Υπενθυμίζεται ότι μετά την εμφύτευση των λέξεων των ερωτήσεων σε διανύσματα σταθερού μεγέθους, το σύστημα χρησιμοποιεί ένα LSTM για την εξαγωγή της αναπαράστασης ολόκληρης της ερώτησης σε ένα διάνυσμα διάστασης 1024. Στόχος λοιπόν είναι η οπτικοποίηση της πληροφορίας που εμπεριέχεται στις 1024 διαστάσεις για όλες τις ερωτήσεις του τελικού συνόλου ελέγχου. Για το λόγο αυτό χρησιμοποιήθηκαν κάποιες τεχνικές που ως στόχο έχουν την προβολή και απεικόνιση των διανυσμάτων αυτών σε ένα χώρο μικρότερης διάστασης. Οι τεχνικές αυτές πέρα από τη μείωση της διάστασης προσπαθούν να δημιουργήσουν μια πιστή αναπαράσταση του αρχικού χώρου και συνεπώς να διατηρούν την ομοιότητα και τις αποστάσεις των διανυσμάτων στο νέο

αυτό χώρο μικρότερης διάστασης. Συγκεκριμένα από τις 447,793 ερωτήσεις του συνόλου ελέγχου επιλέχθηκε ένα τυχαίο δείγμα μεγέθους 60000, και για αυτό το δείγμα όπου οι ερωτήσεις αναπαρίστανται σε 1024 διαστάσεις μετασχηματίζονται σε 2 διαστάσεις ώστε να οπτικοποιηθούν.

Σε πρώτη φάση εφαρμόστηκε η τεχνική της Σταδιακής Ανάλυσης σε Κύριες Συνιστώσες (Incremental Principal Components Analysis -IPCA) ώστε ο πίνακας 60000x1024 (Ερωτήσεις x Διαστάσεις) να μετασχηματιστεί σε 60000x50, δηλαδή να γίνει μείωση από τις 1024 στις 50 διαστάσεις. Ο αλγόριθμος IPCA είναι μια εναλλακτική του αλγορίθμου PCA και μπορεί να χρησιμοποιηθεί στην περίπτωση που το σύνολο δεδομένων δεν μπορεί να χωρέσει ολόκληρο στην μνήμη. Ουσιαστικά η προσπέλαση των δεδομένων γίνεται σε πακέτα (batches) και έτσι μπορεί να επιτευχθεί καλύτερη διαχείριση της μνήμης που απαιτείται (Ross et al., 2008).

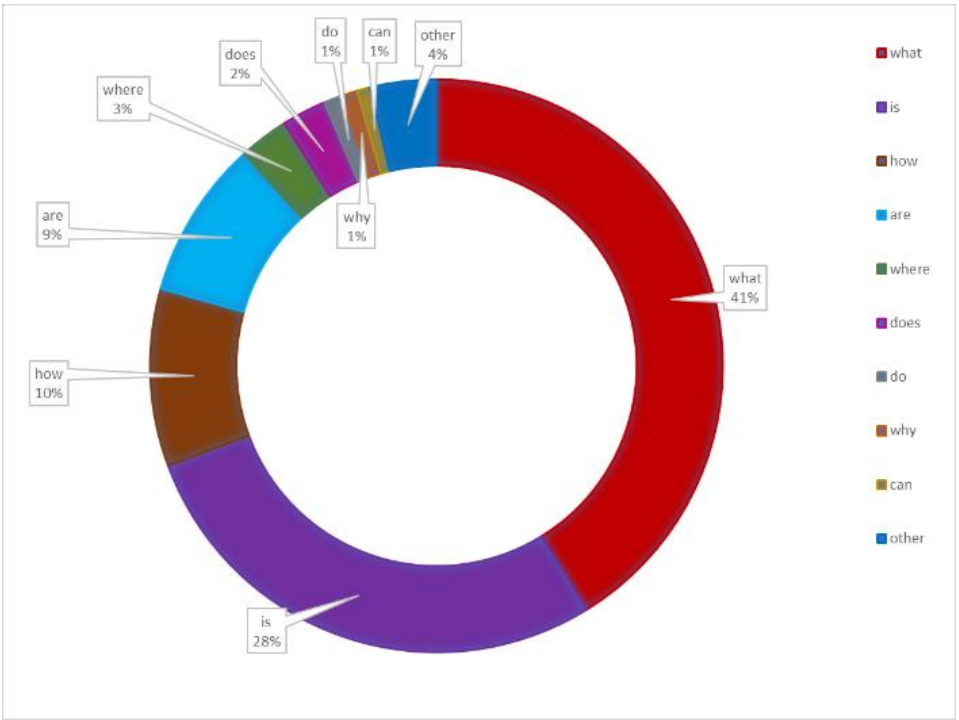
Σε δεύτερη φάση εφαρμόστηκε ο αλγόριθμος t-SNE (t-distributed stochastic neighbor embedding) ώστε να επιτευχθεί περαιτέρω μείωση από τις 50 διαστάσεις σε 2 με σκοπό την τελική οπτικοποίηση των αναπαραστάσεων. Σημειώνεται πως για μεγάλο αριθμό διαστάσεων προτείνεται να χρησιμοποιείται πριν από τον αλγόριθμο t-SNE κάποιος άλλος αλγόριθμος μείωσης διάστασης όπως ο PCA καθώς έτσι μειώνεται ο πιθανός θόρυβος και επιταχύνονται οι υπολογισμοί των αποστάσεων μεταξύ των σημείων.

Στη συνέχεια παρουσιάζεται ο τελικός μετασχηματισμός των αναπαραστάσεων των 60000 ερωτήσεων στις 2 διαστάσεις (Σχήμα 4.21).



Σχήμα 4.21: Δείγμα 60000 αναπαραστάσεων ερωτήσεων οι οποίες έχουν εμφυτευθεί σε δύο διαστάσεις μέσω του αλγορίθμου t-SNE. Η κάθε ερώτηση είναι χρωματισμένη ανάλογα με την πρώτη λέξη της.

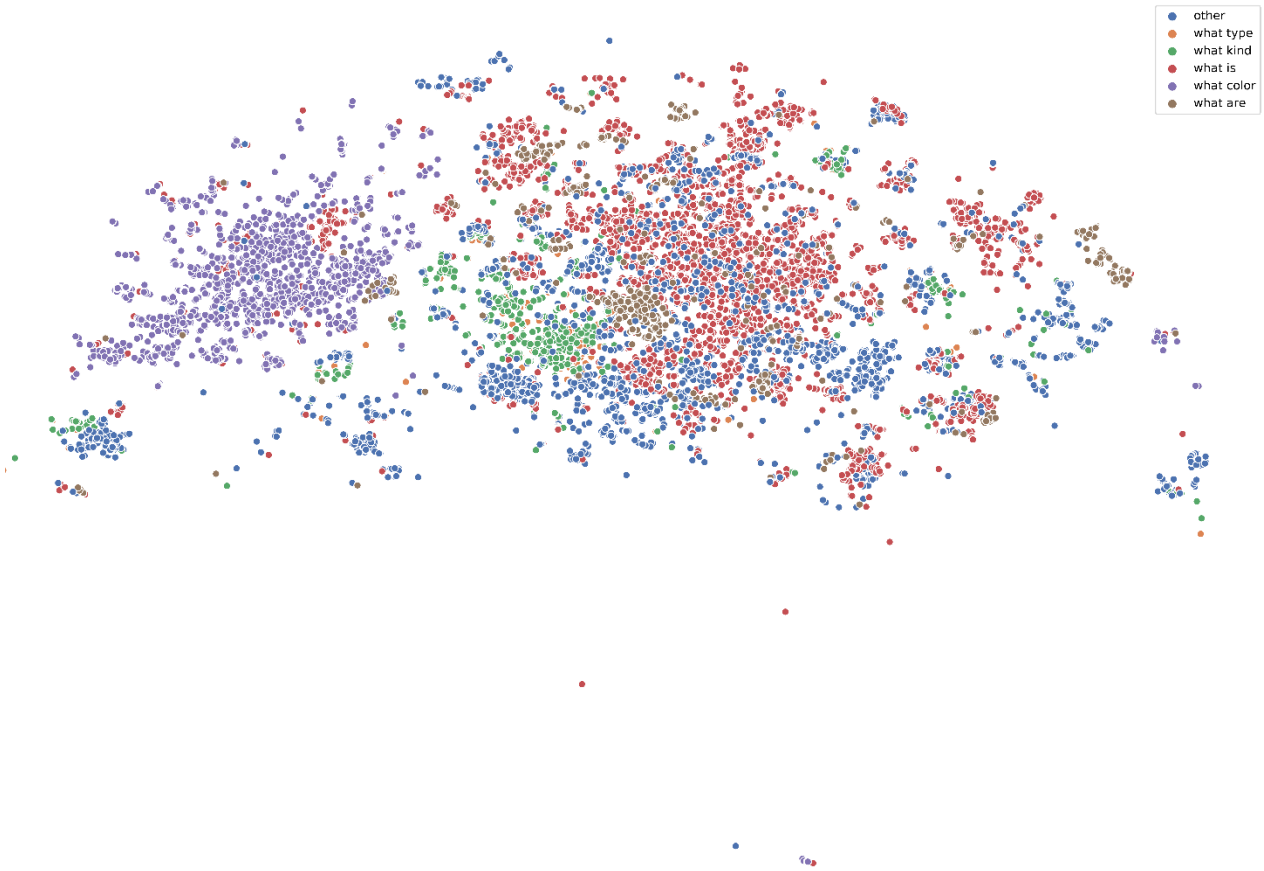
Κάθε σημείο δηλώνει την αναπαράσταση μιας ερώτησης και είναι χρωματισμένο με βάση την πρώτη λέξη της ερώτησης. Οι κατηγορίες των ερωτήσεων ως προς την πρώτη λέξη έχουν επιλεχθεί με βάση τη συχνότητα εμφάνιση τους. Συνεπώς έχουν επιλεχθεί οι 9 πιο συχνές ερωτήσεις με βάση την πρώτη λέξη και οι υπόλοιπες έχουν χαρακτηριστεί ως ‘άλλες’ (other).
Στη συνέχεια παρουσιάζονται κάποια στατιστικά σχετικά με τη συχνότητα εμφάνισης τους.



Σχήμα 4.22: Η κατανομή του προηγούμενου δείγματος 60000 ερωτήσεων.

Η οπτικοποίηση των αναπαραστάσεων αναδεικνύει ότι η αναπαράσταση της κάθε ερώτησης κυριαρχείται σε μεγάλο βαθμό από την πρώτη λέξη, γεγονός που είναι πολύ σηματικό. Παρατηρείται ισχυρή δομή στα δεδομένα και αρκετά ομοιόμορφες περιοχές. Συγκεκριμένα παρατηρούνται δύο εξαιρετικά καλοσχηματισμένες ομαδοποιήσεις των ερωτήσεων που αρχίζουν από τις λέξεις ‘how’, ‘where’ και ‘why’ καθώς επίσης και αρκετές ομάδες που σχετίζονται με την λέξη ‘what’. Κάτι ακόμα που εμφανίζει ενδιαφέρον είναι το γεγονός ότι οι ομάδες ‘are’, ‘is’, ‘can’, ‘do’ και ‘does’ εμφανίζουν κάποιες ελαφρώς ασαφείς υπο-ομάδες οι οποίες όμως συγκροτούν μια μεγαλύτερη ισχυρή ομάδα. Οι ερωτήσεις αυτές έχουν ένα κοινό χαρακτηριστικό σε σχέση με τις υπόλοιπες, ότι οι απαντήσεις που κατά πάσα πιθανότητα αναμένονται είναι ή ‘yes’ ή ‘no’. Συνεπώς η μεγάλη αυτή ομάδα συμβαδίζει και με την ανθρώπινη διαίσθηση της συγκεκριμένης κατηγορίας ερώτησης.
Ένα επιπλέον συμπέρασμα που προκύπτει είναι ότι το σύστημα δίνει μεγάλη έμφαση στην πρώτη λέξη της ερώτησης. Η παρατήρηση αυτή σε κάθε περίπτωση δεν είναι αντίθετη με τη γενικότερη

ανθρώπινη λογική που απαιτείται για να δοθεί μια απάντηση σε μια ερώτηση. Είναι εμφανές ότι η πρώτη λέξη καθορίζει σε πολύ μεγάλο βαθμό την τελική επιλογή της απάντησης.
Κάτι που επίσης έχει ενδιαφέρον είναι η διερεύνηση της επιρροής των υπόλοιπων λέξεων στη σειρά καθώς και μια περαιτέρω διερεύνηση στις διάφορες υπο-ομάδες της κατηγορίας 'what' που αποτελεί και την κατηγορία με την μεγαλύτερη συχνότητα εμφάνισης. Για το λόγο αυτό παρουσιάζεται το επόμενο γράφημα που απεικονίζει τις ομαδοποιήσεις με βάση τη δεύτερη λέξη μόνο για τις ερωτήσεις που αρχίζουν με την λέξη 'what'. Επίσης εμφανίζονται και το πλήθος ερωτήσεων ανά κατηγορία.



Σχήμα 4.23: Αναπαραστάσεις των ερωτήσεων 'what' οι οποίες έχουν εμφυτευθεί σε δύο διαστάσεις μέσω του αλγορίθμου t-SNE. Η κάθε ερώτηση είναι χρωματισμένη ανάλογα με την δεύτερη λέξη της.

Κατηγορίες Ερωτήσεων	Πλήθος
what is	9481
other	6460
what color	4725
what kind	1683
what are	1440
what type	996

Πίνακας 4.14: Κατανομή των ερωτήσεων τύπου 'what'

Παρατηρείται λοιπόν ότι και η επιλογή της δεύτερης λέξης επιδρά εξίσου σημαντικά στη διαμόρφωση της αναπαράστασης. Συγκεκριμένα σχηματίζεται μια αρκετά συμπαγής ομάδα των ερωτήσεων που ξεκινούν με τις λέξεις ‘what color’ καθώς επίσης και μια περιοχή που κυριαρχείται από τις ομάδες ‘what kind’ και ‘what type’ (επικαλύπτεται από την ομάδα ‘what kind’ σε μεγάλο βαθμό). Μια σημαντική επίσης παρατήρηση είναι το γεγονός ότι οι ερωτήσεις που ξεκινούν με ‘what’ και σαν δεύτερη λέξη έχουν κάποια διαφορετική λέξη από τις παραπάνω κατηγορίες (κατηγορία other) και εκείνες που ξεκινούν με τις λέξεις ‘what is’ εμφανίζουν τη μεγαλύτερη διασπορά. Το τελευταίο είναι επίσης λογικό καθώς οι απαντήσεις στις ερωτήσεις αυτές έχουν ένα τεράστιο εύρος. Συνεπώς γίνεται εμφανές ότι έχουν σημασία και οι επόμενες λέξεις, καθώς στην περίπτωση των ερωτήσεων τύπου ‘what is’ η αναπαράσταση εξαρτάται από το σύνολο των λέξεων της ερώτησης.

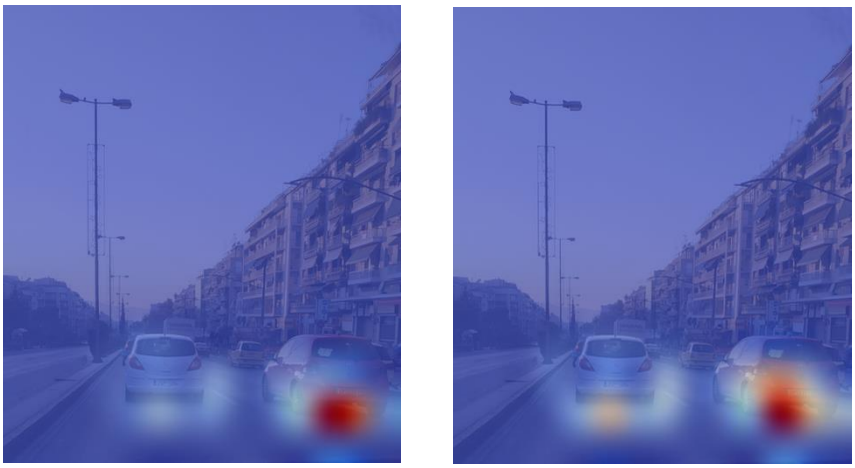
4.4.3 Ανάλυση Επιπέδων Εστίασης

Ένα από τα πιο βασικά μέρη της αρχιτεκτονικής του συστήματος αποτελούν τα επίπεδα εστίασης. Για την καλύτερη κατανόηση και ερμηνεία των αποτελεσμάτων είναι αναγκαία η ανάλυση της ικανότητας εστίασης του συστήματος στις περιοχές που σχετίζονται με την ερώτηση. Στην συνέχεια παρουσιάζονται αποτελέσματα από περιοχές που εστίασε το σύστημα στην προσπάθεια να εντοπίσει τις συσχετιζόμενες με την ερώτηση περιοχές. Οι εικόνες και οι ερωτήσεις που χρησιμοποιούνται αποτελούν μέρος εκείνων που τέθηκαν κατά τον έλεγχο λειτουργίας σε προηγούμενη υποενότητα. Για κάθε εικόνα και ερώτηση εμφανίζονται με κόκκινο χρώμα οι περιοχές που εμφάνισαν αυξημένη πιθανότητα εστίασης και με μπλε οι περιοχές με μικρή πιθανότητα. Εμφανίζονται αποτελέσματα και για τα δύο επίπεδα εστίασης.
Στην πρώτη φάση εμφανίζονται οι ερωτήσεις που τέθηκαν για την εικόνα που απεικονίζεται μια λεωφόρος. Παρατηρείται ότι το σύστημα αντιλαμβάνεται πολύ καλά τους τοπικούς προσδιορισμούς που αναφέρονται εντός της ερώτησης και εντοπίζει ικανοποιητικά τα διάφορα αντικείμενα.

Q: What is the color of the car in the middle?

A: White

Σχήμα 4.24: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση ‘Τι χρώμα είναι το αυτοκίνητο στην μέση;’



Q: What is the color of the right car?

A: Red

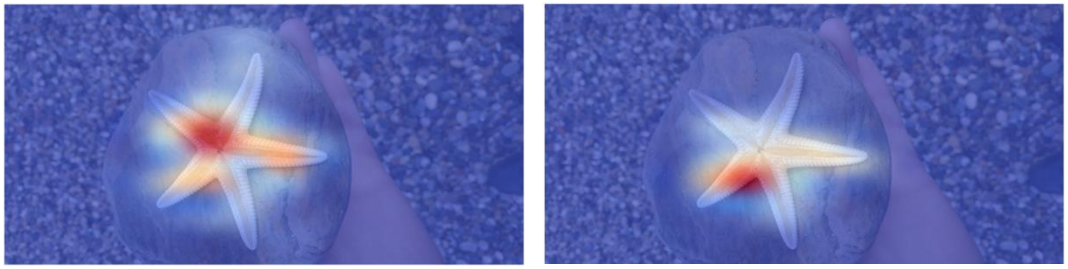
Σχήμα 4.25: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση ‘Τι χρώμα είναι το αυτοκίνητο στα δεξιά;’

Q: What is this in front of the car?

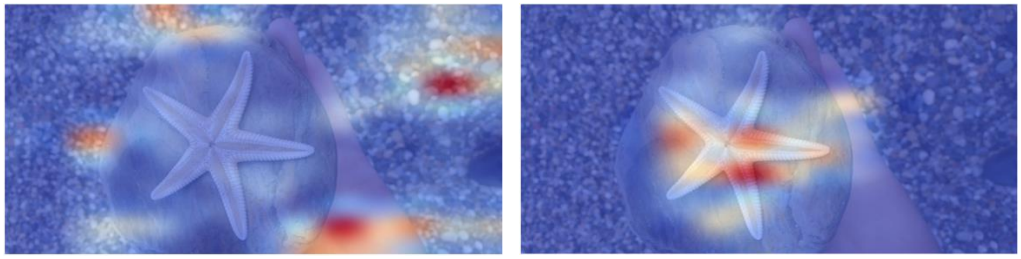
A: Car

Σχήμα 4.26: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση ‘Τι βρίσκεται μπροστά από το αυτοκίνητο;’

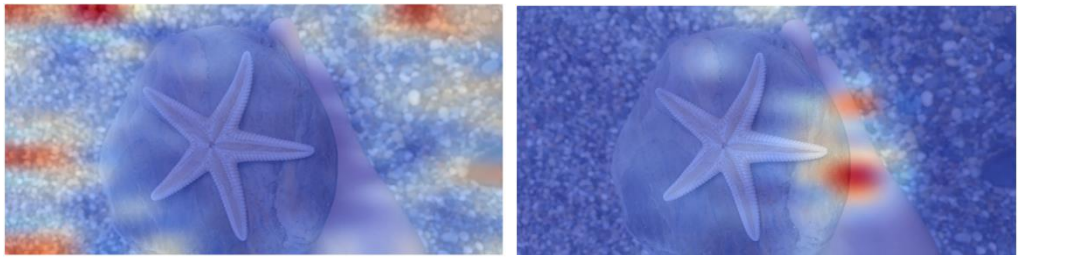
Εξίσου καλή συμπεριφορά παρατηρείται και στις επόμενες εικόνες που απεικονίζεται ο αστερίας.



Σχήμα 4.27: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση 'Τι είναι αυτό;'



Σχήμα 4.28: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση 'Τι βρίσκεται κάτω από τον αστερία;'



Σχήμα 4.29: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση 'Υπάρχει κάποιο χέρι;'

Ενδιαφέρον παρουσιάζει η περίπτωση του αστερία στην ερώτηση που τίθεται για το αν υπάρχει κάποιο χέρι στην εικόνα. Στο πρώτο επίπεδο εστίασης το σύστημα ουσιαστικά κάνει μια συνολική αναζήτηση στο σύνολο της εικόνας και στο επόμενο επίπεδο εστιάζει περαιτέρω στο χέρι και τελικά απαντάει σωστά.
Στη συνέχεια εμφανίζεται η εικόνα με το γαϊδούρι και παρατηρείται επίσης καλή λειτουργία. Με βάση τα αποτελέσματα της εστίασης γίνεται κατανοητό ότι το σύστημα είναι σε θέση να εντοπίζει τα διαφορετικά αντικείμενα που τίθενται στην ερώτηση.

Q: What animal is this? A: Horse

Σχήμα 4.30: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση ‘Τι ζώο είναι αυτό;’

Q: What color is his t-shirt? A: Green

Σχήμα 4.31: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση ‘Τι χρώμα είναι η μπλούζα του;’

Q: What color is his pants? A: Blue

Σχήμα 4.32: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση ‘Τι χρώμα είναι το παντελόνι του;’

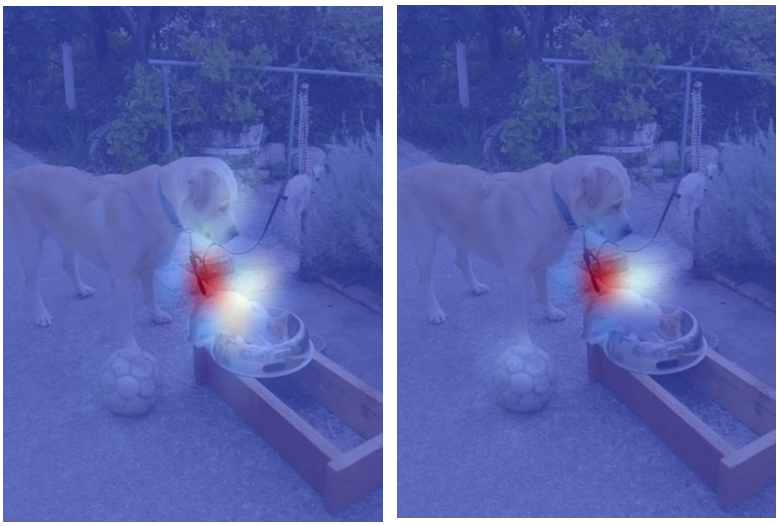
Τελευταία εικόνα είναι εκείνη που απεικονίζει έναν σκύλο, μία μπάλα και μία γάτα η οποία βρίσκεται εντός ενός μπολ. Παρατηρείται ότι στην ερώτηση για το ποιο ζώο απεικονίζεται στο αριστερό μέρος της εικόνας το σύστημα στο πρώτο επίπεδο εντοπίζει αποτελεσματικά τον σκύλο. Στο δεύτερο επίπεδο φαίνεται να μην επιτυγχάνεται περαιτέρω εστίαση στον σκύλο, παρόλα αυτά το σύστημα είναι σε θέση να δώσει σωστή απάντηση.
Στην εικόνα που σχετίζεται με το πάτωμα παρατηρείται εξαιρετικός εντοπισμός της συνολικής επιφάνειας του πατώματος στο δεύτερο επίπεδο εστίασης.
Τέλος στην ερώτηση τι βρίσκεται μέσα στο μπολ το σύστημα είναι σε θέση να εντοπίσει την γάτα εντός του μπολ από το πρώτο επίπεδο εστίασης.

Q: What is the animal on the left? A: Dog

Σχήμα 4.33: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση ‘Τι ζώο είναι αυτό στα αριστερά;’

Q: What color is the ground? A: Gray

Σχήμα 4.34: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση ‘Τι χρώμα είναι το πάτωμα;’



Q: What is sitting inside the plate? A: Cats

Σχήμα 4.35: Οπτικοποίηση των Επιπέδων Εστίασης για την ερώτηση ‘Τι βρίσκεται μέσα στο πιάτο;’

4.4.4 Ανάλυση Σφαλμάτων

Για την περαιτέρω κατανόηση της λειτουργίας του τελικού συστήματος έγινε προσπάθεια ανάλυσης και κατηγοριοποίησης των σφαλμάτων που παρουσιάζονται. Για το λόγο αυτό αξιοποιήθηκε το βέλτιστο μεμονωμένο σύστημα που αποτελείται από το DenseNet-161 (448x448) & Πίνακα Εμφύτευση και έχει εκπαιδευτεί μόνο στα δεδομένα του συνόλου εκπαίδευσης. Έτσι επιλέχθηκε ένα τυχαίο δείγμα μεγέθους 100 από το σύνολο αξιολόγησης, για το οποίο διατίθενται και οι αντίστοιχες απαντήσεις ώστε να αναλυθούν οι περιπτώσεις που εμφανίζονται σφάλματα. Το σύστημα απάντησε με απόλυτη ακρίβεια τις 48/100 ερωτήσεις. Παραδείγματα των ερωτήσεων που δεν μπόρεσαν να απαντηθούν με απόλυτη ακρίβεια εμφανίζονται στη συνέχεια. Επίσης παρατηρήθηκαν τρεις διαφορετικές κατηγορίες οι οποίες θα μπορούσαν να κατηγοριοποιήσουν τα σφάλματα αυτά.

<u>Λανθασμένη περιοχή εστίασης</u>

Στην πρώτη κατηγορία ταξινομήθηκαν τα σφάλματα στα οποία το σύστημα δεν μπόρεσε να εστιάσει στην σωστή περιοχή. Με βάση την παραπάνω δειγματοληψία, 20/52 σφάλματα ανήκουν σε αυτή την κατηγορία. Ένα παράδειγμα εμφανίζεται στη συνέχεια στο οποίο το σύστημα θα έπρεπε να εστιάσει στην χαίτη του αλόγου ενώ αντίθετα εστιάζει στο τρίχωμα. Στην προκειμένη περίπτωση γίνεται κατανοητό ότι η λανθασμένη απάντηση ‘καφέ’ οφείλεται λόγω της λανθασμένης εστίασης.

Σχήμα 4.36: Οπτικοποίηση του μέσου όρου των Επιπέδων Εστίασης για την ερώτηση ‘Τι χρώμα είναι η χαίτη του αλόγου;’

<u>Σωστή περιοχή εστίασης</u>

Η δεύτερη κατηγορία σφαλμάτων αποτελείται από εκείνα στα οποία το σύστημα εστίασε στις περιοχές που σχετίζονται με τις ερωτήσεις αλλά παρόλα αυτά έδωσε λανθασμένη απάντηση. Από το δείγμα που εξετάστηκε το μεγαλύτερο μέρος των σφαλμάτων 25/52 προήλθε από αυτή

την κατηγορία. Ένα παράδειγμα είναι το επόμενο όπου το σύστημα εστιάζει αποτελεσματικά στον πυροσβεστικό κρουνό αλλά αδυνατεί να κωδικοποιήσει σωστά το χρώμα.

Q: What color is the fire hydrant?

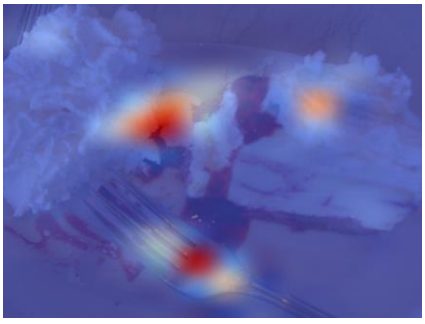
Prediction: gray

Ground Truth: silver and red

Σχήμα 4.37: Οπτικοποίηση του μέσου όρου των Επιπέδων Εστίασης για την ερώτηση 'Τι χρώμα είναι ο πυροσβεστικός κρουνός;'

<u>Παρόμοια απάντηση με βάση την ανθρώπινη αντίληψη (Όμοια πρόβλεψη)</u>

Η τρίτη κατηγορία αφορά απαντήσεις που είναι σχεδόν ίδιες με την σωστή απάντηση. Λόγω του ότι η αξιολόγηση γίνεται με βάση την απόλυτη ταύτιση των λέξεων που έδωσαν οι χρήστες με εκείνη του συστήματος, υπάρχει η πιθανότητα η εκτίμηση της απάντησης να εμφανίζεται ως λανθασμένη. Το πλήθος τέτοιων απαντήσεων ήταν 7/52. Δύο τέτοια παραδείγματα είναι τα ακόλουθα.



Q: What is on the plate?

Prediction: cake

Ground Truth: cheesecake

Σχήμα 4.38: Οπτικοποίηση του μέσου όρου των Επιπέδων Εστίασης για την ερώτηση 'Τι βρίσκεται στο πιάτο;'

Q: What is parked next t o the bike?

Prediction: bike

Ground Truth: motorcycle

Σχήμα 4.39: Οπτικοποίηση του μέσου όρου των Επιπέδων Εστίασης για την ερώτηση ‘Τι είναι παρκαρισμένο δίπλα στο ποδήλατο/μηχανή;’

5. Επίλογος

5.1 Σύνοψη και συμπεράσματα

Στην παρούσα εργασία υλοποιήθηκε ένα σύστημα ερωτοαπαντήσεων βάσει οπτικού περιεχομένου με τη χρήση τεχνικών βαθιάς μηχανικής μάθησης. Έγιναν πειραματισμοί με τέσσερις διαφορετικές προσεγγίσεις για την εξαγωγή των χαρακτηριστικών της εικόνας και δύο για την αναπαράσταση των ερωτήσεων σε ένα διάνυσμα σταθερού μεγέθους. Σε κάθε περίπτωση ο συνδυασμός των αναπαραστάσεων έγινε μέσω των Δικτύων Πολλαπλών Εστιάσεων (Stacked Attention Networks) (Yang et al., 2016).
Παρατηρήθηκε ότι η ανάλυση της εικόνας εισόδου 224x224 ή 448x448 παίζει σημαντικό ρόλο στο αποτέλεσμα, και συγκεκριμένα με την αύξηση της ανάλυσης εμφανίζεται και αύξηση της απόδοσης. Επίσης παρατηρήθηκε ότι το υπερσύγχρονο δίκτυο DenseNet-161 (Huang et al., 2017) εμφανίζει καλύτερα αποτελέσματα από ότι το VGGNet-19 (Simonyan et al., 2015). Παρόλα αυτά η αύξηση δεν κρίνεται εξαιρετικά σημαντική αναλογικά με τη διαφορά στο βάθος των δύο δικτύων.
Ακόμα σημειώνεται ότι η χρήση του μοντέλου ELMo (Peters et al., 2018) δεν αύξησε την απόδοση του μοντέλου σε σχέση με την πιο απλή προσέγγιση του Πίνακα Εμφύτευσης για την αναπαράσταση των λέξεων. Παρατηρήθηκε ότι η αναπαράσταση των ερωτήσεων που προέρχονται από εμφυτεύσεις των λέξεων μέσω ενός Πίνακα Εμφύτευσης εμφανίζει τα καλύτερα αποτελέσματα. Οι αναπαραστάσεις αυτές αξιολογήθηκαν περαιτέρω με τεχνικές μείωσης της διάστασης όπου εντοπίστηκαν αρκετά συμπαγείς ομάδες για κάθε κατηγορία ερωτήσεων στον πολυδιάστατο χώρο της αναπαράστασης των ερωτήσεων. Μάλιστα οι αποστάσεις μεταξύ των ομάδων εμφανίστηκαν ανάλογες με τη σημασιολογική ομοιότητα τους όπως γίνεται αντιληπτή με βάση την ανθρώπινη αντίληψη.
Για όλα τα μοντέλα που δοκιμάστηκαν έγινε αξιολόγηση της τελικής τους απόδοσης στο ‘κρυφό’ σύνολο ελέγχου που παρέχεται από τους δημιουργούς του συνόλου δεδομένων VQA v2 (Goyal et al., 2017). Το μοντέλο που εμφάνισε τα βέλτιστα αποτελέσματα ήταν εκείνο που κάνει χρήση του DenseNet-161 με ανάλυση εικόνας εισόδου 448x448 για την εξαγωγή των χαρακτηριστικών της εικόνας και χρησιμοποιεί έναν Πίνακα Εμφύτευσης για την αναπαράσταση των λέξεων των ερωτήσεων οι οποίες έπειτα τροφοδοτούνται σε ένα LSTM. Για τη συγχώνευση των αναπαραστάσεων παρατηρήθηκε ότι με δύο Επίπεδα Εστίασης επιτυγχάνεται το βέλτιστο αποτέλεσμα. Το βέλτιστο μοντέλο εμφανίστηκε επίσης να εντοπίζει αρκετά αποτελεσματικά τις περιοχές της εικόνας που σχετίζονται με την ερώτηση. Ο εντοπισμός αυτός έγινε αντιληπτός με βάση τις πιθανότητες της κατανομής που παράγουν τα επίπεδα εστίασης.
Πέραν της αξιολόγησης των μεμονωμένων αυτών μοντέλων, πραγματοποιήθηκε και αξιολόγηση συστημάτων που συνδυάζουν τις αποφάσεις των επιμέρους συστημάτων που δοκιμάστηκαν. Η βέλτιστη προσέγγιση ήταν αυτή όπου η τελική απάντηση προκύπτει μέσω ‘ψηφοφορίας’ με συντελεστές στάθμισης (weights) στην ψήφο του κάθε επιμέρους συστήματος (Weighted Ensemble Models). Η στάθμιση αυτή προκύπτει από τη συνολική απόδοση του κάθε συστήματος πάνω στο σύνολο αξιολόγησης (Validation Set). Το σύστημα αυτό αποτελεί το σύστημα με την βέλτιστη απόδοση στην παρούσα εργασία του οποίου η απόδοση του είναι 71.14% σε

απαντήσεις ‘ΝΑΙ’/’ΟΧΙ’ , 35,5% σε απαντήσεις που απαιτείται κάποιος αριθμός, 48,84% σε οποιαδήποτε άλλη κατηγορία και 56.63% συνολικά για όλες μαζί της ερωτήσεις του συνόλου ελέγχου.

5.2 Μελλοντικές επεκτάσεις

Επέκταση του συνόλου δεδομένων

Στην παρούσα εργασία έγινε χρήση του συνόλου δεδομένων VQA v2 το οποίο αποτελεί ένα από τα πιο δημοφιλή σύνολα δεδομένων για το πρόβλημα της αυτόματης απάντησης σε ερωτήσεις φυσικής γλώσσας που αναφέρεται στο περιεχόμενο μιας εικόνας (Visual Question Answering). Κατά τη διαδικασία των αποτελεσμάτων παρατηρήθηκε ότι η αύξηση του συνόλου των απαντήσεων θα μπορούσε να οδηγήσει σε ακόμα καλύτερα αποτελέσματα. Συνεπώς κρίνεται αναγκαία μια πιθανή επέκταση του συνόλου των δεδομένων για την εκπαίδευση των συστημάτων.
Αρκετά σύνολα δεδομένων πέραν αυτού έχουν αναπτυχθεί την τελευταία περίοδο με ίσως το πιο σημαντικό το Visual Genome (Krishna et al., 2017). Το σύνολο αυτό δεδομένων αποτελεί αυτή τη στιγμή το μεγαλύτερο σύνολο δεδομένων για το πρόβλημα VQA. Επίσης σε σχέση με το VQA v2 όπου το 38% των απαντήσεων είναι τύπου ‘ΝΑΙ’/’ΟΧΙ’, στο Visual Genome για την καταγραφή μεγαλύτερης ποικιλίας και πολυπλοκότητας ερωτήσεων οι δημιουργοί του υπέδειξαν στους χρήστες που επισήμαιναν τις εικόνες να μην χρησιμοποιούν καθόλου τέτοιου τύπου ερωτήσεις και απαντήσεις. Ακόμα το 89% των απαντήσεων στο VQA αποτελείται από μία λέξη ενώ στο Visual Genome το αντίστοιχο ποσοστό είναι μόλις 57%. Επίσης οι 1000 ερωτήσεις με τη μεγαλύτερη συχνότητα καλύπτουν το 87% στο VQA έναντι του 65% στο Visual Genome. Το Visual Genome συνεπώς εμφανίζεται πιο πλήρες, παρόλα αυτά δεν αξιοποιήθηκε στην παρούσα εργασία καθώς οι περισσότερες μέθοδοι αυτή τη στιγμή αξιολογούνται και συγκρίνονται πάνω στο VQA v.2 σύνολο δεδομένων το οποίο συνεισέφερε σημαντικά στην ανάπτυξη του συγκεκριμένου επιστημονικού πεδίου.
Συνεπώς μια μελλοντική επέκταση θα αφορούσε την επέκταση του συνόλου δεδομένων αξιοποιώντας ταυτόχρονα περισσότερα από ένα σύνολα δεδομένων όπως τα DAQUAR (Malinowski et al., 2014), COCO-QA (Ren et al., 2015), The VQA Dataset (Goyal et al., 2017), FM-IQA (Gao et al., 2015), Visual7W (Zhu et al., 2016), VisDial (Das et al., 2017) και Visual Genome (Krishna et al., 2017). Επίσης για περαιτέρω αύξηση των συνόλων θα μπορούσαν να αξιοποιηθούν και τεχνικές αύξησης των δεδομένων (Data Augmentation) όπως αντικατοπτρισμός των εικόνων (Image Mirroring) για το διπλασιασμό του συνόλου των εικόνων.

Ενίσχυση των χαρακτηριστικών της εικόνας

Στην παρούσα εργασία η εκτίμηση των περιοχών που σχετίζονται με την ερώτηση επιτυγχάνεται μέσω των Πολλαπλών Επιπέδων Εστίασης (Stacked Attention Network). Οι περιοχές αυτές ορίζονται πάνω σε ένα ομοιόμορφο πλέγμα που χωρίζει την εικόνα. Ουσιαστικά η εικόνα

χωρίζεται είτε σε 7x7 είτε σε 14x14 περιοχές και για κάθε περιοχή υπολογίζεται μια πιθανότητα της συσχέτισης της με την ερώτηση. Ένα πρόβλημα που παρατηρείται είναι ότι τα αντικείμενα σε μια εικόνα μπορεί να βρίσκονται σε τελείως διαφορετική κλίμακα μεταξύ τους, λόγω του βάθους που μπορεί να έχει μια εικόνα. Στις περιπτώσεις αυτές μια τέτοια προσέγγιση είναι πιθανό να μην μπορεί να εξάγει αποτελεσματικά τα χαρακτηριστικά όλων των αντικειμένων. Μία λύση για τη βελτίωση των φαινομένων αυτών μπορεί να είναι η αύξηση της ανάλυσης της εικόνας ώστε να παραχθούν περιοχές, όπως για παράδειγμα 28x28. Παρόλα αυτά ίσως μια τέτοια επέκταση να προσθέτει αρκετό θόρυβο στα δεδομένα και να μην οδηγεί σε βελτίωση των αποτελεσμάτων. Μια άλλη πιο διαισθητικά ορθή λύση που δείχνει να κερδίζει έδαφος είναι εκείνη όπου οι περιοχές δεν δημιουργούνται πάνω σε ένα πλέγμα αλλά εντοπίζονται αρχικά μέσω ενός μοντέλου Εντοπισμού Αντικειμένων (Object Detection) στην εικόνα. Ένα παράδειγμα είναι αυτό των (Peters et al., 2018) στο οποίο κάνουν χρήση ενός Faster R-CCN (Ren et al., 2015) δικτύου για τον εντοπισμό των αντικειμένων εντός της εικόνας. Συνεπώς μια πιθανή μελλοντική προσέγγιση είναι ο εντοπισμός αυτών των περιοχών, η εξαγωγή των χαρακτηριστικών τους και έπειτα η τροφοδότηση τους σε ένα Επίπεδο Εστίασης.

Ενίσχυση της αναπαράστασης της ερώτησης

Στην παρούσα εργασία πέραν της αναπαράστασης των λέξεων μέσω ενός Πίνακα Εμφύτευσης, έγινε πειραματισμός και με το μοντέλο ELMo (Peters et al., 2018). Παρατηρήθηκε ότι η αναπαράσταση των λέξεων μέσω του ELMo δεν εμφάνισε καλύτερα αποτελέσματα. Παρόλα αυτά όπως προτείνουν οι (Peters et al., 2018) (Reimers et al., 2019) ένα διαφορετικό σχήμα συνδυασμού των αναπαραστάσεων είναι πολύ πιθανό να οδηγήσει σε καλύτερα αποτελέσματα. Για παράδειγμα προτείνεται αντί να γίνει στάθμιση των τριών αναπαραστάσεων του ELMo να γίνει στάθμιση μόνο των δύο πρώτων ή να εφαρμοστεί σειριακή επέκταση (concatenation) των αναπαραστάσεων. Υπενθυμίζεται ότι οι συντελεστές στάθμισης των αναπαραστάσεων προσαρμόζονται κατά την εκπαίδευση του μοντέλου. Μια άλλη εναλλακτική που προτείνεται για τη χρήση του ELMo είναι ο συνδυασμός των αναπαραστάσεων του ELMo με εκείνες άλλων μεθόδων όπως του Πίνακα Εμφύτευσης. Η τελευταία μέθοδος θεωρείται η πιο υποσχόμενη για τη βελτίωση των αποτελεσμάτων.
Μία ακόμα επέκταση που θα μπορούσε να ωφελήσει την απόδοση του μοντέλου είναι η χρήση Επιπέδων Εστίασης και στις αναπαραστάσεις των λέξεων ή των φράσεων εντός των ερωτήσεων. Όπως προτείνεται από τους (Lu et al., 2016), δοθείσας της εικόνας, γίνεται εστίαση στις λέξεις που σχετίζονται περισσότερο με αυτή. Η τεχνική αυτή ονομάζεται Co-Attention.

Βελτίωση στο συνδυασμό των δύο αναπαραστάσεων (Εικόνας/Ερώτησης)

Εναλλακτικές προσεγγίσεις στον τρόπο που συνδυάζονται οι αναπαραστάσεις της εικόνας και της ερώτησης είναι πιθανό να οδηγήσουν σε αύξηση της απόδοσης του συστήματος. Η απλή κατά σημείο πρόσθεση των χαρακτηριστικών της εικόνας και της ερώτησης που εφαρμόζεται στην παρούσα εργασία ίσως είναι περιοριστική για τη βέλτιστη συγχώνευση των δύο αναπαραστάσεων. Είναι πιθανό να υπάρχουν αλληλεπιδράσεις μεταξύ των χαρακτηριστικών οι οποίες δεν αποτυπώνονται ικανοποιητικά με μία τέτοια μέθοδο. Τεχνικές που εστιάζουν την

προσοχή τους στον καλύτερο συνδυασμό ονομάζονται Πολυτροπική Συγχώνευση Χαρακτηριστικών (Multi-Modal Feature Fusion) και αρκετές επικεντρώνονται σε Διγραμμικές Απεικονίσεις. Ένα παράδειγμα μιας τέτοιας μεθόδου αποτελεί ο υπολογισμός του Εξωτερικού Γινομένου των δύο διανυσμάτων χαρακτηριστικών και ο οποίος μπορεί να ανιχνεύσει όλες τις αλληλεπιδράσεις $2^{ης}$ τάξης μεταξύ δύο διανυσμάτων χαρακτηριστικών. Με τον όρο $2^{ης}$ τάξης αναφέρεται ότι αν υπάρχουν αλληλεπιδράσεις μεταξύ τριών και περισσότερων χαρακτηριστικών αυτές δεν αποτυπώνονται άμεσα με το εξωτερικό γινόμενο. Μια τέτοια τεχνική δεν μπορεί να εφαρμοστεί άμεσα σε κάθε περίπτωση καθώς απαιτεί εξαιρετικά υψηλούς πόρους σε μνήμη (Fukui et al., 2016). Άλλη τεχνική μπορεί να είναι ο μετασχηματισμός ξεχωριστά για την κάθε αναπαράσταση μέσω πινάκων που προσαρμόζονται κατά την εκπαίδευση, έπειτα πολλαπλασιάζονται κατά σημείο και στη συνέχεια γίνεται εφαρμογή κάποιου Συγκεντρωτικού Επιπέδου Αθροίσματος (Yu et al., 2017).
Σε κάθε περίπτωση ο περαιτέρω πειραματισμός για το συνδυασμό των αναπαραστάσεων είναι αναγκαίος και πιθανότατα να βελτιώνει αισθητά τα αποτελέσματα.

Αύξηση του αριθμού των επιμέρους μοντέλων (Ensembling)

Στα περισσότερα προβλήματα μηχανικής μάθησης έχει παρατηρηθεί ότι σε περιπτώσεις που οι αποφάσεις λαμβάνονται σε συνδυασμό από αρκετά μεμονωμένα μοντέλα (Ensemble Models) τα αποτελέσματα είναι αρκετά καλύτερα. Ένα παράδειγμα σε ένα πρόβλημα Ταξινόμησης μπορεί να είναι ότι το τελικό αποτέλεσμα είναι εκείνο που εμφανίζεται πιο συχνά μεταξύ πολλών επιμέρους μοντέλων.
Στο VQA πρόβλημα έχει παρατηρηθεί ότι η αύξηση των μοντέλων που αποφασίζουν βελτιώνει εξαιρετικά πολύ τα αποτελέσματα. Οι (Teney et al., 2018) για ένα μοντέλο που εκπαιδεύτηκε στο ίδιο σύνολο δεδομένων με αυτό της παρούσας εργασίας, παρατήρησαν ότι με την αύξηση του πλήθους των μοντέλων αυξάνεται δραματικά η τελική απόδοση. Μάλιστα η απόδοση συνέκλινε για τον αριθμό των 30 μεμονωμένων δικτύων. Συνεπώς μια μελλοντική επέκταση αφορά την εκπαίδευση ακόμα περισσότερων μοντέλων πέραν των 5 που χρησιμοποιούνται για την παρούσα εργασία.

6. Βιβλιογραφία

Albelwi S. and Mahmood A., 2017. “A framework for designing the architectures of deep convolutional neural networks,” Entropy, vol. 19, pp. 1–20.

Anderson P., He X., Buehler C., Teney D., Johnson M., Gould S., Zhang L., 2018. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv: 1707.07998.

Antol S., Agrawal A., Lu J., Mitchell M., Batra D., Zitnick C.L., Parikh D., 2015. Vqa: Visual question answering. arXiv preprint arXiv:1505.00468.

Bengio Y., 2009. "Learning Deep Architectures for AI" (PDF). Foundations and Trends in Machine Learning.

Bradski G. 2000. The OpenCV Library. Dr Dobb’s Journal of Software Tools.

Chen X. and Zitnick C.L., 2015. Mind’s Eye: A Recurrent Visual Representation for Image Caption Generation. In CVPR.

Collette A., 2013. Python and HDF5.

Das A., Kottur S., Gupta K., Singh A., Yadav D., Moura J., Parikh D., Batra D., 2017. Visual Dialog, CVPR.

Donahue J., Hendricks L.A., Guadarrama S., Rohrbach M., Venugopalan S., Saenko K., Darrell T., 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In CVPR.

Elman J., 1990. “Finding Structure in Time”. en. In: Cognitive Science 14.2. 00000, pp. 179–211. issn: 03640213. doi: 10.1016/0364-0213(90)90002-E.

Alpaydin Et., 2010. Introduction to Machine Learning. MIT Press. p. 9. ISBN 978-0-262-01243-0.

Fang H., Gupta S., Iandola F.N., Srivastava R., Deng L., Doll´ar P., Gao J., He X., Mitchell M., Platt J.C., Zitnick C., Zweig G., 2015. From Captions to Visual Concepts and Back. In CVPR.

Fukui A., Park D.H., Yang D., Rohrbach A., Darrell T., Rohrbach M., 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In EMNLP.

Huang G., Liu Z., Van Der Maaten L., Weinberger K., 2017. "Densely Connected Convolutional Networks." In CVPR, vol. 1, no. 2, p. 3.

Gao H., Mao J., Zhou J., Huang Z., Wang L., Xu W., 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. arXiv preprint arXiv:1505.05612.

Gardner M., Grus J., Neumann M., Tafjord O., Dasigi P., Liu N., Peters M., Schmitz M., Zettlemoyer L., 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform.

Geman S., Bienenstock E., DoursatR., 1992. "Neural networks and the bias/variance dilemma" (PDF). Neural Computation.

Goodfellow I., Bengio Y., Courville A., 2016. Deep Learning. MIT Press.

Goyal Y., Khot T., Summers-Stay D., Batra D., Parikh D., 2017 Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In CVPR.

Hagiwara M. "Real-World Natural Language Processing". To be published by Manning Publications in 2019.

Haykin S., 1999. Neural Networks: A Comprehensive Foundation, Prentice Hall, ISBN 0-13-273350-1.

Haykin S., 2009. Neural Networks and Learning Machines, 3rd Edition.

He K., Zhang X., Ren S., Sun J., 2016. Deep Residual Learning for Image Recognition, CVPR.

Hochreiter., 1991. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut f. Informatik, Technische Univ. Munich.

Hochreiter S., Schmidhuber J., 1997. Long Short-Term Memory, Neural Computation 9(8):1735{1780.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417–441, and 498–520.

Hunter J.D., 2007. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95.

Kiros R., SalakhutdinovR., Zemel R.S., 2015. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. TACL.

Krishna R., Zhu Y., Groth O., Johnson J., Hata K., Kravitz J., Chen S., Kalantidis Y., Li L., Shamma D.A. et al., 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations," International Journal of Computer Vision, vol. 123, no. 1, pp. 32-73.

Krizhevsky A., Sutskever I., Hinton G., 2012. ImageNet Classification with Deep Convolutional Neural Networks, NIPS.

LeCun Y., Bengio Y., Hinton G., 2015. Deep learning. Nature, 521(7553), pp.436-444.

LeCun Y., Bottou L., Bengio Y., Haffner P., 1998. Gradient-based learning applied to document recognition, Proc. IEEE 86(11): 2278–2324.

Loper E. and Bird S., 2002. NLTK: The Natural Language Toolkit, Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.

Lu J., Yang J., Batra D., Parikh D., 2016. Hierarchical question-image co-attention for visual question answering. arXiv preprint arXiv:1606.00061.

Malinowski M., Rohrbach M., Fritz M., 2015. Ask your neurons: A neural-based approach to answering questions about images. arXiv preprint arXiv:1505.01121.

Malinowski M. and Fritz M., 2014. A multi-world approach to question answering about real world scenes based on uncertain input," in Advances in Neural Information Processing Systems (NIPS).

Mao J., Xu W., Yang Y., Wang J., Yuille A.L., 2014.Explain Images with Multimodal Recurrent Neural Networks. CoRR, abs/1410.1090.

Mikolov T., Sutskever I., Chen K., Corrado G., Jeffrey D., 2013. "Distributed Representations of Words and Phrases and their Compositionality".

Mohri M., Rostamizadeh A., Talwalkar A., 2012. Foundations of Machine Learning, The MIT Press.

Mozer, M. C., 1995. "A Focused Backpropagation Algorithm for Temporal Pattern Recognition".

In Chauvin, Y.; Rumelhart, D. (eds.). Backpropagation: Theory, architectures, and applications. ResearchGate. Hillsdale, NJ: Lawrence Erlbaum Associates. pp. 137–169. Retrieved 2017-08-21.
Oliphant T.E, 2006. A guide to NumPy, USA: Trelgol Publishing.

Paszke A., Gross S., Chintala S., Chanan G., Yang E., DeVito Z., Lin Z., Desmaison A., Antiga L., Lerer A., 2017. Automatic differentiation in PyTorch.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., 2011. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830.

Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L., 2018. Deep contextualized word representations, Proc. of NAACL.

Reimers N. and Gurevych I., 2019. Alternative weighting schemes for elmo embeddings.

Ren M., Kiros R., and Zemel R., 2015. Exploring models and data for image question answering. arXiv preprint arXiv:1505.02074.

Ren S., He K., Girshick R., Sun J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS.

Robbins H. and Monro S., 1951. A Stochastic Approximation Method. The Annals of Mathematical Statistics, 22(3):400–407.

Ross D., Lim J., Lin R., Yang M., 2008. Incremental Learning for Robust Visual Tracking, International Journal of Computer Vision, Volume 77, Issue 1-3, pp. 125-141.

Roweis, S. T.; Saul, L. K., 2000. "Nonlinear Dimensionality Reduction by Locally Linear Embedding". Science. 290 (5500): 2323–2326.

Russell S. and Norvig P., 1995. Artificial Intelligence: A Modern Approach. Upper Saddle River. New Jersey.

Schmidhuber J., 2015. Deep learning in neural networks: An overview. Neural Networks, 61: 85–117.

Schuster M., and Paliwal K., 1997. "Bidirectional recurrent neural networks." Signal Processing, IEEE Transactions on 45.11: 2673-2681.2. Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan

Simon P., 2013. Too Big to Ignore: The Business Case for Big Data. Wiley, p. 89. ISBN 978-1-118-63817-0.

Simonyan K. and Zisserman A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR.

Srivastava R.K., Greff K., Schmidhuber J., 2015. “Highway Networks.” CoRR abs/1505.00387 (2015): n. pag.

Szegedy C. et al., 2015. "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 1-9.

Teney D., Anderson P., He X., Van den Hengel A., 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In CVPR.

Van der Maaten L.J.P., Hinton G.E., 2008. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9:2579-2605.

Vinyals O., Toshev A., Bengio S., Erhan D., 2014. Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555.

Xu K., Ba J., Kiros R., Courville A., Salakhutdinov R., Zemel R., Bengio Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044.

Yang Z., He X., Gao J., Deng L., Smola A., 2016. Stacked Attention Networks for Image Question Answering. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Yangqing J., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Guadarrama S., Darrell, T., 2014. Caffe: Convolutional Architecture for Fast Feature Embedding.

Yu Z., Yu J., Fan J., Tao D., 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In ICCV.

Zeiler M.D., Fergus R., 2014. Visualizing and Understanding Convolutional Networks. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol 8689. Springer, Cham

Harris. Z., 1954. "Distributional Structure". Word. 10 (2/3): 146–62.

Zhu Y., Groth O., Bernstein M., Fei-Fei L., 2016. Visual7w: Grounded question answering in images, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Πηγές Σχημάτων (Διαδίκτυο)

[1]: https://optima-systems.co.uk/neural-networks-apl/

[2]: https://medium.com/@krishnakalyan3/introduction-to-exponential-linear-unit-d3e2904b366c

[3]: http://cs231n.github.io/convolutional-networks/

[4]: https://medium.com/coinmonks/paper-review-of-vggnet-1st-runner-up-of-ilsvlc-2014-image-classification-d02355543a11

[5]: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[6]: http://evelinag.com/blog

[7]: http://www.realworldnlpbook.com/blog/improving-sentiment-analyzer-using-elmo.html

[8]: https://petrlorenc.github.io/ELMO/

[9]: http://mrinitialman.com/Library/HTML/Chapters/Appendices/Appendices-Characters.html

[10]: https://www.mihaileric.com/posts/deep-contextualized-word-representations-elmo/

[11]: http://jalammar.github.io/illustrated-bert/