Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Αναγνώριση συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής βασισμένη σε αρχιτεκτονικές βαθιάς μηχανικής μάθησης

Διπλωματικη Εργασια

του

ΚΩΝΣΤΑΝΤΙΝΟΥ ΠΥΡΟΒΟΛΑΚΗ

Επιβλέπων: Γεώργιος Στάμου
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Παρασκευή Τζούβελη
Ε.ΔΙ.Π.

Εργαστηριο Συστηματων Τεχνητης Νοημοσυνης και Μαθησης

Αθήνα, Μάρτιος 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Αναγνώριση συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής βασισμένη σε αρχιτεκτονικές βαθιάς μηχανικής μάθησης

Διπλωματικη Εργασια

του

ΚΩΝΣΤΑΝΤΙΝΟΥ ΠΥΡΟΒΟΛΑΚΗ

Επιβλέπων: Γεώργιος Στάμου
Αναπληρωτής Καθηγητής Ε.Μ.Π.
Συνεπιβλέπων: Παρασκευή Τζούβελη
Ε.ΔΙ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 3η Μαρτίου 2020.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γεώργιος Στάμου	Στέφανος Κόλλιας	Ανδρέας Σταφυλοπάτης
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2020

(Υπογραφή)

.....

ΠΥΡΟΒΟΛΑΚΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

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

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Περίληψη

Το θέμα της παρούσας διπλωματικής εργασίας είναι η αναγνώριση συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής βασισμένη σε αρχιτεκτονικές βαθιάς μηχανικής μάθησης (Deep Learning). Το θέμα της αυτόματης αναγνώρισης συναισθήματος από την μουσική αποτελεί ένα ενεργό πρόβλημα στον τομέα της MIR (Music Information Retrieval) και συνδέεται με πολλές ερευνητικές μελέτες τα τελευταία χρόνια.

Η ραγδαία εξάπλωση της πληροφορίας, που προσφέρει η σύγχρονη εποχή, επηρεάζει τους περισσότερους τομείς της κοινωνικής και οικονομικής δραστηριότητας. Είναι επόμενο πως αυτή η ευκολία στη σύνθεση και τη διάδοση της πληροφορίας δεν θα άφηνε ανεπηρέαστη τη μουσική και τη μουσική βιομηχανία. Η ψηφιακή εποχή έχει αλλάξει τον τρόπο με τον οποίο η μουσική παράγεται και διαδίδεται, έτσι δημιουργούνται νέες ανάγκες για την αυτόματη και αποτελεσματική διαχείριση μεγάλου όγκου μουσικών κομματιών, όπως απαιτεί και η αναγνώριση του συναισθήματος.

Προκειμένου να προσεγγίσουμε το πρόβλημα της αναγνώρισησ συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής, αρχικά θα πραγματοποιήσουμε ξεχωριστές προσπάθειες στην ανάλυση των στίχων και στην ανάλυση ηχητικού σήματος. Στη συνέχεια, θα εφαρμοστεί μια πολυκαναλική ανάλυση που θα αξιοποιεί τόσο τους στίχους όσο και το ηχητικό σήμα. Τα διαθέσιμα δεδομένα για την εκπαίδευση και την επαλήθευση των συστημάτων προέρχονται από ένα σύνολο 2.000 τραγουδιών ταξινομημένα σε τέσσερις συναισθηματικές κλάσεις {happy, angry, sad, relaxed}.

Για την αποτελεσματική αξιοποίηση των στίχων και του ηχητικού σήματος, στα πλαίσια του προβλήματος, θα εφαρμοστούν μέθοδοι Επεξεργασίας Φυσικής Γλώσσας και Ψηφιακής Επεξεργασίας Σήματος. Τελικά, θα προταθούν και θα αναπτυχθούν ξεχωριστά συστήματα ανάλυσης στίχων και ηχητικού σήματος από τα οποία θα επιλεχθούν τα αποτελεσματικότερα για την κατασκευή ενός ενιαίου πολυκαναλικού συστήματος συναισθηματικής ταξινόμησης.

Λέξεις Κλειδιά

Αναγνώριση Συναισθήματος, Επεξεργασία Φυσικής Γλώσσας, Ψηφιακή Επεξεργασία Σήματος, Αντίληψη Μουσικής Έκφρασης, Μηχανική Μάθηση, Βαθιά Μηχανική Μάθηση, Μεταφορική Μάθηση.

Abstract

Subject of this diploma thesis is Music Mood Classification utilizing Deep Learning techniques. The subject of automatic emotion recognition from music tracks has been an active problem in the field of MIR (Music Information Retrieval) and it is associated with a lot of research studies in the recent years.

The rapid spread of information, that modern times offer, affects in most aspects of the social and economic activity. It is expected that the convinience in creating and spreading information would not leave music and music industry untouched. Digital age has changed the way music have been produced and distibuted, this creates new needs for autonomous and effective management of music tracks in large volumes.

In order to approach the task of Music Mood Classification, we will make separate approaches for lyric analysis and audio signal analysis. Later on, a multimodal analysis will be applied based on the audio signal and the lyrics of a track. The dataset consist of 2.000 tracks classified in four mood classes {happy, angry, sad, relaxed}.

For the effective use of audio signal and lyrics, in the context of the task, methods of Natural Language Processing and Digital Signal Processing will be used. Finally, different models will be proposed and developed for lyric analysis and audio analysis, from this models we will chose the most effective to construct a multimodal model for mood classification.

Keywords

Mood Classification, Natural Language Processing, Digital Signal Processing, Machine Learning, Deep Learning, Transfer Learning.

Ευχαριστίες

Ευχαριστώ τον καθηγητή Γεώργιο Στάμου και τα μέλη του εργαστηρίου Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης για την ευκαιρία που μου δόθηκε να εργαστώ στο συγκεκριμένο θέμα της διπλωματικής εργασίας μου.

Ευχαριστώ ιδιαίτερα την κυρία Δρ. Παρασκευή Τζούβελη, που μου προσέφερε την κάθε δυνατή βοήθεια και σωστή καθοδήγηση κατά την εκπόνηση της διπλωματικής μου εργασίας όπως επίσης και το μάθημα Νευρωνικών Δικτύων που αποτέλεσε βασικό γνωσιακό υπόβαθρο για την εκτέλεση της παρούσας εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τους καθηγητές Στέφανο Κόλλια και Ανδρέα-Γεώργιο Σταφυλοπάτη που συμμετέχουν στην τριμελή επιτροπή.

Τέλος, θέλω να ευχαριστήσω πολύ τους φίλους μου για την υποστήριξη που μου προσέφεραν καθ΄ όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας και την οικογένειά μου που είναι πάντα δίπλα μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Οι όροι μουσική και συναίσθημα είναι δύο έννοιες άρρηκτα συνδεδεμένες από την πρώτη κιόλας στιγμή που ο άνθρωπος εφηύρε τη μουσική. Η μουσική είναι παρούσα σε κάθε πτυχή της ζωής, ενισχύοντας τις ανθρώπινες δραστηριότητες είτε αφορούν τη διασκέδαση, τη ξεκούραση ή ακόμα και τη δουλεία. Επομένως, παρουσιάζει ιδιαίτερο επιστημονικό ενδιαφέρον η μελέτη του τρόπου που η μουσική προκαλεί συναισθηματική διέγερση στον ακροατή. Κατά βάση, ένα μουσικό κομμάτι θα προκαλέσει τα ίδια συναισθήματα σε ένα σύνολο ακροατών, αν και πολλοί είναι παράγοντες που ευθύνονται για τον τρόπο που θα γίνει αντιληπτό ένα μουσικό κομμάτι και διαφέρουν από ακροατή σε ακροατή.

Η σύνδεση της μουσικής με το συναίσθημα μπορεί να έχει πρακτική εφαρμογή σε έναν μεγάλο αριθμό πεδίων. Ένα παράδειγμα θα μπορούσε να είναι μια επιχείρηση που έχει άμεση επαφή με τον καταναλωτή, όπως ένα εστιατόριο ή ένα πολυκατάστημα, και μέσα από τη μουσική μπορεί να δημιουργήσει μια ταυτότητα ή ακόμα και να κατευθύνει τις καταναλωτικές αποφάσεις. Ακόμα, η αναγνώριση συναισθήματος από μουσική είναι ένα αντικείμενο που απασχολεί σε μεγάλο βαθμό τη μουσική βιομηχανία, καθώς μπορεί να βοηθήσει σε αρκετά προβλήματα όπως είναι οι μουσικές προτάσεις ή η κατηγοριοποίηση κομματιών και η δημιουργία μουσικών λιστών.

1.1 Αντικείμενο της διπλωματικής εργασίας

Η άντληση πληροφορίας από μουσική (Music Information Retrieval - MIR) είναι ένα πεδίο μελέτης με συνεχή ανάπτυξη τα τελευταία χρόνια, οδηγούμενο από την ανάγκη αυτόματης επεξεργασίας μεγάλου όγκου μουσικών τραγουδιών. Πολλές είναι οι επιστήμες και τα επιστημονικά πεδία από τα όποια η MIR αντλεί γνώση, όπως η ψυχολογία, η μουσικολογία, η επεξεργασία σήματος, η μηχανική μάθηση και πολλά άλλα. Μερικά αντικείμενα μελέτης στην MIR είναι η αναγνώριση του είδους μουσικής στο οποία ανήκει ένα μουσικό κομμάτι, ο διαχωρισμός φωνής και ενόργανης μουσικής, η παρακολούθηση ρυθμού και τονικού ύψους, η αναγνώριση μουσικών οργάνων αλλά και η αναγνώριση συναισθήματος ενός μουσικού κομματιού.

Ουσιαστικά, η αναγνώριση του συναισθήματος που προέρχεται από την μουσική αναφέρεται

στην αυτόματη ανίχνευση του συναισθήματος που προκαλείται από την ακρόαση ενός μουσικού κομματιού.

Στο πλαίσιο της παρούσας διπλωματικής θα προσεγγίσουμε το πρόβλημα της αναγνώρισης του μουσικού συναισθήματος εφαρμόζοντας μια πολυκαναλική ανάλυση βασισμένη στους στίχους και στο μουσικό σήμα. Για την αντιμετώπιση του προβλήματος θα εφαρμόσουμε τεχνικές Βαθιάς Μηχανικής Μάθησης και θα συγκρίνουμε τα αποτελέσματα που προκύπτουν από τις τρεις προσεγγίσεις, της ανάλυσης στίχων, της ανάλυσης ηχητικού σήματος και του συνδυασμού τους. Τα διαθέσιμα δεδομένα, για την ανάπτυξη και την επαλήθευση των συστημάτων, προέρχονται από το MoodyLyrics Dataset [18] και αποτελούνται από ένα σύνολο 2.000 τραγουδιών μαζί με τις αντίστοιχες ετικέτες (labels) τους, που προέρχονται από ένα σύνολο τεσσάρων βασικών συναισθημάτων {happy, angry, sad, relaxed}.

Αφού γίνει η παραδοχή ότι κάθε τραγούδι θα ανήκει σε μια από τις τέσσερις συναισθηματικές κλάσεις {happy, angry, sad, relaxed}, όπως προκύπτουν από το συναισθηματικό μοντέλο του Russel [57], το πρόβλημα που θα προσπαθήσουμε να επιλύσουμε μετατρέπεται σε ένα πρόβλημα ταξινόμησης.

Στην προσπάθεια της ανάπτυξης ενός συστήματος Βαθιάς Μηχανικής Μάθησης που θα αντιμετωπίζει το πρόβλημα της αναγνώρισης συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής κρίθηκε απαραίτητη η μελέτη και η εφαρμογή μεθόδων Επεξεργασίας Φυσικής Γλώσσας και Ψηφιακής Επεξεργασίας Σήματος. Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP) φάνηκε ιδιαίτερα χρήσιμη στο πρόβλημα της ανάλυσης των στίχων, ενώ η Ψηφιακή Επεξεργασίας Σήματος (Digital Signal Processing - DSP) εφαρμόστηκε στην ανάλυση του ηχητικού σήματος.

1.2 Παρεμφερείς εργασίες

Η αυτόματη αναγνώριση μουσικού συναισθήματος αποτελεί ένα ενεργό πεδίο έρευνας στην MIR τα τελευταία τριάντα χρόνια. Μια πρώτη προσπάθεια ξεκινά από το 1990 όπου ο Gordon Bruner [7] ερεύνησε τον τρόπο που επιδρά το μουσικό συναίσθημα στον τομέα του μάρκετινγκ. Αργότερα, οι Menno van Zaanen και Pieter Kanters στην εργασία τους [68] αναπτύσσουν ένα ένα μοντέλο Μηχανικής Μάθησης για την συναισθηματική ταξινόμηση μουσικών κομματιών, αξιοποιώντας τους στίχους και την μέθοδο ενσωμάτωσης TF-IDF. Προσπάθειες για την συναισθηματική ταξινόμηση βασισμένη στα χαρακτηριστικά ηχητικού σήματος θα γίνουν από τους Γιώργο Τζανετάκη [64] και Geoffroy Peeters [52]. Οι προηγούμενοι, στην προσπάθεια τους να αντιμετωπίσουν τα προβλήματα ταξινόμησης του MIREX (Music Information Retrieval Evaluation eXchange), θα χρησιμοποιήσουν κλασσικά χαρακτηριστικά ήχου, όπως τα Mel Frequency Cepstral Coefficients (MFCCs), για να τα τροφοδοτήσουν αργότερα σε Support Vector Machines (SVM).

Το 2016, πάλι στα πλαίσια του MIREX, οι Thomas Lidy και Alexander Schindler [33] θα προχωρήσουν στην εκπαίδευση Συνελικτικών Νευρωνικών Δικτύων (CNNs), για την συναισθηματική ταξινόμηση, αξιοποιώντας χαρακτηριστικά ήχου όπως τις συχνότητες Mel. Αρκετές είναι και οι προσπάθειες αντιμετώπισης του προβλήματος ταξινόμησης με πολυκαναλικές

προσεγγίσεις. Μερικά αξιοσημείωτα παραδείγματα, ανάλυσης στίχων και ηχητικού σήματος, είναι οι εργασίες των Xiao Hu [27] και Rémi Delbouys [14] οι οποίοι καταλήγουν στο συμπέρασμα πως οι πολυκαναλικές προσεγγίσεις ξεπερνούν τις επιδόσεις των απλών μοντέλων.

1.3 Οργάνωση του εγγράφου

Στην παρούσα διπλωματική εργασία αντιμετωπίζουμε την αναγνώριση συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής ως ένα πρόβλημα που μπορεί να επιλυθεί αποτελεσματικά αξιοποιώντας γλωσσικά χαρακτηριστικά και χαρακτηριστικά ήχου. Αρχικά, στο κεφάλαιο 2 θα αναπτυχθούν έννοιες και ορισμοί της Μηχανικής Μάθησης ως θεωρητικό υπόβαθρο πάνω στο οποίο θα βασιστεί η δόμηση των συστημάτων Βαθιάς Μηχανικής Μάθησης της εργασίας. Στο κεφάλαιο 3 θα γίνει μια περιγραφή της δημιουργίας του συνόλου των δεδομένων εισόδου μέσα από την επεξεργασία κειμένου και ηχητικού σήματος, που προηγήθηκαν.

Αργότερα, στο κεφάλαιο 4 θα αναλυθούν οι προτεινόμενες αρχιτεκτονικές και οι προτεινόμενες ενσωματώσεις δεδομένων που συνθέτουν τα συστήματα συναισθηματικής κατηγοριοποίησης που υλοποιήσαμε. Ενώ, στο κεφάλαιο 5 θα εξηγηθεί λεπτομερώς η πειραματική διαδικασία που εφαρμόστηκε για την εκπαίδευση και την καταγραφή της επίδοσης των προτεινόμενων συστημάτων. Τέλος, στο κεφάλαιο 6 καταλήγουμε στα βασικά συμπεράσματα της διπλωματικής εργασίας και δίνουμε προτάσεις που θα μπορούσαν να εφαρμοστούν στο μέλλον για την βελτίωση της επίδοσης των συστημάτων μας αλλά και για την χρήση τους σε άλλες εφαρμογές.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

2.1 Τεχνητή Νοημοσύνη

Η τεχνητή νοημοσύνη (Artificial Intelligence-AI) είναι ένας τομέας της πληροφορικής που ασχολείται με την σχεδίαση και την ανάπτυξη συστημάτων που μιμούνται την ανθρώπινη συμπεριφορά. Όπως υποδηλώνει και το όνομα της, τα συστήματα τεχνητής νοημοσύνης προσπαθούν να μοντελοποιήσουν και να προσομοιώσουν νοήμονες συμπεριφορές που χαρακτηρίζουν τον άνθρωπο. Επίσης, η ανάπτυξη τέτοιων συστημάτων προϋποθέτει την μελέτη και άλλων επιστημών, όπως για παράδειγμα της ψυχολογίας και της ιατρικής, συνθέτοντας έτσι ένα σημείο τομής μεταξύ πληροφορικής και των επιστημών αυτών. Η τεχνητή νοημοσύνη επεκτείνεται σε πολλές κατηγορίες όπως η επεξεργασία φυσικής γλώσσας (NTLK), η ρομποτική (Robotics), η όραση υπολογιστών (Computer Vision), η μηχανική μάθηση (Machne Learning) κ.α.

2.2 Μηχανική Μάθηση

Η μηχανική μάθηση (Machine Learning-ML) αποτελεί υποκατηγορία της τεχνητής νοημοσύνης με σκοπό την ανάπτυξη συστημάτων που μπορούν να μαθαίνουν. Το όνομα μηχανική μάθηση αποδόθηκε από τον Arthut Samuel το 1959 [59], πολύ αργότερα το 1997 ο Tom M. Mitchell θα δώσει έναν πιο επίσημο και συχνά αναφερόμενο ορισμό της μηχανικής μάθησης. Σύμφωνα με τον προηγούμενο, ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από μια εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης Π, αν η επίδοσή του στις εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E [44, p. 2]. Ουσιαστικά, οι αλγόριθμοι μηχανικής μάθησης χτίζουν ένα μαθηματικό μοντέλο με την χρήση κάποιων δεδομένων, γνωστά και ως δεδομένα εκπαίδευσης (training data), ώστε να μπορέσουν να κάνουν προβλέψεις και να πάρουν αποφάσεις χωρίς να έχουν προγραμματιστεί ρητά για το συγκεκριμένο έργο. Οι αλγόριθμοι μηχανικής μάθησης ταξινομούνται σε τρεις κύριες κατηγορίες, βάσει του τρόπου εκμάθησης, την επιβλεπόμενη μάθηση (Supervised learning), τη μη επιβλεπόμενη μάθηση (Unsupervised learning) και την ενισχυμένη μάθηση (Reinforcement learning).

2.2.1 Επιβλεπόμενη Μάθηση

Στην επιβλεπόμενη μάθηση, ή αλλιώς επιτηρούμενη μάθηση, το σύνολο των δεδομένων (dataset) που χρησιμοποιεί ο εκάστοτε αλγόριθμος αποτελείται από ζευγάρια εισόδου (input) και της επιθυμητής εξόδου (output). Συχνά τα δεδομένα εξόδου αναφέρονται και ως ετικέτες (labels). Είναι σύνηθες, στη διαδικασία της επιβλεπόμενης μάθησης το σύνολο των δεδομένων να χωρίζεται σε δύο υποσύνολα, τα δεδομένα εκπαίδευσης (traing data) και τα δεδομένα δοκιμής (test data). Με τα δεδομένα εκπαίδευσης επιτυγχάνεται η εκπαίδευση του αλγορίθμου ενώ με τα δεδομένα δοκιμής αξιολογείται η επιτυχία του. Σκοπός ενός αλγορίθμου επιβλεπόμενης μάθησης είναι η προσέγγιση μιας συνάρτησης, με τη βοήθεια της εκπαίδευσης, που θα αντιστοιχεί μία ετικέτα σε κάθε δεδομένο εισόδου. Οι ετικέτες των δεδομένων εισόδου είναι αυτές που θα καθοδηγήσουν τον αλγόριθμο να γενικεύσει την αντιστοίχηση εισόδου-ετικέτας, να μάθει να αποδίδει δηλαδή τις σωστές ετικέτες σε οποιαδήποτε είσοδο. Οι ετικέτες μπορούν να πάρουν είτε ποιοτικές ή ποσοτικές τιμές. Για τον λόγο αυτό, τα προβλήματα που επιλύει η επιβλεπόμενη μάθηση χωρίζονται σε δύο κατηγόριες, τα προβλήματα κατηγοριοποίησης (Classification) και τα προβλήματα παλινδρόμησης (Rergression).

- Κατηγοριοποίηση (Classification) Στα προβλήματα κατηγοριοποίησης, οι τιμές των ετικετών ανήκουν σε ένα περιορισμένο σύνολο, οι ετικέτες δηλαδή εκφράζουν ένα ποιοτικό χαρακτηριστικό. Ο αλγόριθμος έχει στόχο να ταξινομήσει τα δεδομένα της εισόδου στην κατηγορία που ανήκουν. Για παράδειγμα, προβλήματα ταξινόμησης μπορεί να είναι η απόφαση αν σε μια εικόνα υπάρχει μια γάτα ή ο διαχωρισμός τραγουδιών σε μουσικά είδη.

- Παλινδρόμηση (Regression) Στα προβλήματα παλινδρόμησης, οι τιμές των ετικετών παίρνουν συνεχείς τιμές, οι ετικέτες δηλαδή εκφράζουν ένα ποσοτικό χαρακτηριστικό. Ο αλγόριθμος σε αυτή την περίπτωση έχει ως στόχο να προβλέψει μια τιμή για τα δεδομένα της εισόδου. Προβλήματα παλινδρόμησης θα μπορούσαν να είναι η πρόβλεψη της τιμής μιας μετοχής ή η πρόβλεψη της θερμοκρασίας της ακόλουθης μέρας.

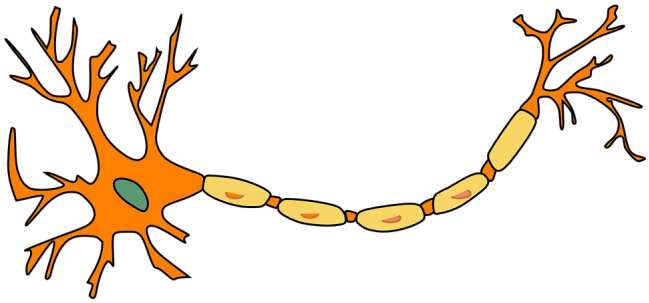
2.2.2 Μη Επιβλεπόμενη Μάθηση

Στη μη επιβλεπόμενη μάθηση, το σύνολο των δεδομένων δεν είναι επισημασμένα, δεν έχουν δηλαδή ετικέτες, σε αντίθεση με την εποβλεπόμενη μάθηση. Σκοπός της μη επιβλεπόμενης μάθησης είναι να αναγνωρίσει κοινά χαρακτηριστικά των δεδομένων εισόδου και να τα ομαδοποιήσει σε κατηγορίες. Αυτή η τεχνική ομαδοποίησης ονομάζεται συσταδοποίηση (Clustering) καθώς επικεντρώνεται στην ταξινόμηση των δεδομένων σε συστάδες. Η ανάλυση της συσταδοποίησησς τοποθετεί στην ίδια ομάδα παρατηρήσεις με κοινά χαρακτηριστικά, σύμφωνα με κάποια προκαθορισμένα κριτήρια, ενώ παρατηρήσεις με ανόμοια χαρακτηριστικά τις απομακρύνει. Ανάλογα με τη φύση του προβλήματος, ο αριθμός των συστάδων, που προσπαθούν οι αλγόριθμοι μη επιβλεπόμενης μάθησης να ταξινομήσουν τα δεδομένα, δεν είναι πάντα γνωστός εκ των προτέρων. Έστω ότι έχουμε εικόνες από ζώα και θέλουμε να τα κατηγοριοποιήσουμε σε ομάδες ώστε κάθε ομάδα να περιέχει εικόνες ίδιων ζώων. Είναι φανερό ότι

ο αριθμός των ομάδων σε αυτό το πρόβλημα είναι αρχικά άγνωστος. Αντίθετα, στο πρόβλημα διαχωρισμού αλληλογραφίας σε κακόβουλη ή όχι ο αριθμός των συστάδων είναι γνωστός εκ των προτέρων.

2.2.3 Ενισχυτική Μάθηση

Στην ενισχυτική μάθηση τα δεδομένα δεν είναι επισημασμένα, όπως στην προηγούμενη κατηγορία, όμως η διαδικασία εκμάθησης διαφέρει αρκετά από τις προηγούμενες κατηγορίες. Η βασικής τεχνική που χρησιμοποιεί ένας αλγόριθμος ενισχυτικής μάθησης είναι η ανταμοιβή και η τιμωρία. Πιο συγκεκριμένα, ένας πράκτορας (agent) ανταμοίβεται ή τιμωρείται για τις κινήσεις (actions) που επιλέγει να κάνει σε ένα περιβάλλον (environment) προσπαθώντας να μεγιστοποιήσει μια αθροιστική μεταβλητή (reward). Το αντικείμενο στο οποίο επικεντρώνεται η ενισχυτική μάθηση είναι η ισορροπία ανάμεσα στην απόκτηση νέας γνώσης μέσα από την εξερεύνηση του περιβάλλοντος και την εκμετάλευση της ήδη υπάρχουσας γνώσης. Λόγω τις γενικότητας τους, οι αλγόριθμοι ενισχυτικής μάθησης χρησιμοποιούνται και σε πολλούς άλλους τομείς, πέρα από την μηχανική μάθηση, όπως η θεωρία ελέγχου, η θεωρία παιγνίων, οι γενετικοί αλγόριθμοι κ.α.[47]. Ακόμα, η χρήση των αλγόριθμων αυτών είναι συχνή σε προβλήματα που απαιτούν μακροπρόθεσμη έναντι βραχυπρόθεσμης ανταμοιβής όπως για παράδειγμα τα αυτοοδηγούμενα οχήματα και τα προγράμματα που έχουν αναπτυχθεί για να αντιμετωπίζουν τον άνθρωπο σε παιχνίδια (Σκάκι , Ντάμα , Go κ.α.).



Πηγή: en.wikipedia.org/wiki/Neuron
Σχήμα 2.1: Η δομή ενός βιολογικού νευρώνα

2.2.4 Μεταφορική Μάθηση

Η μεταφορική μάθηση (Transfer Learning - TL) είναι μια τεχνική μηχανικής μάθησης. Η τεχνική αυτή επικεντρώνεται στην αποθήκευση της γνώσης που αποκτήθηκε, επιλύοντας ένα πρόβλημα και στη συνέχεια εφαρμόζοντας τη γνώση αυτή για την επίλυση ενός διαφορετικού αλλά σχετικού προβλήματος. Το 1993, ο Lorien Pratt παρουσίασε στη διατριβή του πάνω στη μεταφορά στη μηχανική μάθηση τον αλγόριθμο DBT [53]. Ουσιαστικά, η μεταφορική μάθηση είναι μια διαδικασία βελτιστοποίησης, μια παράκαμψη εξοικονόμησης χρόνου ή επίτευξης καλύτερης απόδοσης. Ωστόσο, στη γενική περίπτωση δεν είναι από πριν προφανές ότι η μεταφορική μάθηση θα είναι ωφέλιμη στον τομέα εργασίας αλλά μετά την ανάπτυξη και την αξιολόγηση του μοντέλου. Για το λόγο αυτό η χρήση των μηχανισμών μεταφορικής μάθησης

δεν πρέπει να γίνεται αυθαίρετα αλλά μόνο μετά από προσεκτική μελέτη. Μερικά πεδία εφαρμογής είναι η κατηγοριοποίησης κειμένων [16], η αναγνώριση ψηφίων [37] και η αναγνώριση κακόβουλης αλληλογραφίας [5].

2.3 Τεχνητά Νευρωνικά Δίκτυα

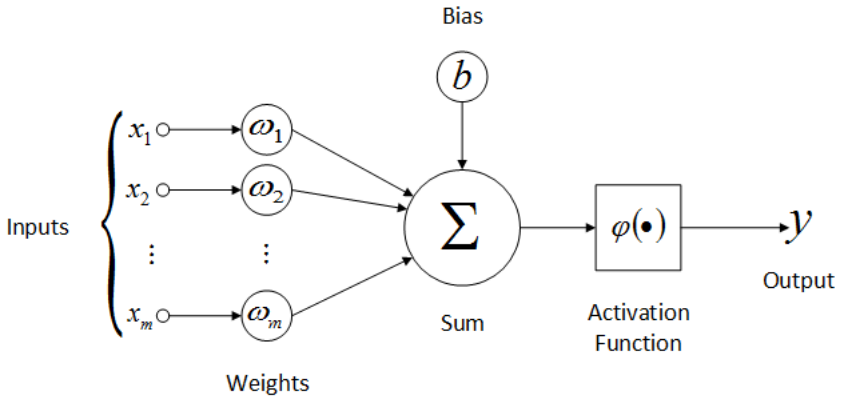
Τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks-ANN) είναι υπολογιστικά συστήματα εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα του εγκεφάλου. Τα βιολογικά νευρωνικά δίκτυα συνθέτονται από συνάψεις μονάδων (νευρώνων) με σκοπό να εκτελέσουν συγκεκριμένες λειτουργίες όταν ενεργοποιηθούν. Οι συνάψεις των νευρώνων σε ένα βιολογικό δίκτυο είναι συνδυασμός χημικών και ηλεκτρικών λειτουργιών που μαθαίνουν να εκτελούν διεργασίες χωρίς να έχουν προγραμματιστεί ρητά για το σκοπό αυτό.Τα τεχνητά νευρωνικά δίκτυα, επομένως, προσπαθούν να συνθέσουν ένα απλοποιημένο μοντέλο των βιολογικών και όχι να τα αντιγράψουν.

Ιστορικά, ο πρώτος τεχνητός νευρώνας δημιουργείται το 1943[39] όπου οι Warren McCulloch και Walter Pitts παρουσίασαν ένα μοντέλο νευρώνα με δυαδική συνάρτηση ενεργοποίησης, ικανό να αναπαριστά boolean συναρτήσεις. Η απλότητα όμως του μοντέλου των McCulloch και Pitts δημιουργούσε ορισμένα προβλήματα καθώς επέτρεπε μόνο δυαδικές εισόδους, χρησιμοποιούσε μόνο τη βηματική σαν συνάρτηση ενεργοποίησης, δεν ενσωμάτωνε βάρη στην είσοδο και δεν μπορούσε να περιγράψει μη γραμμικές συναρτήσεις.

Μερικά από τα προβλήματα αυτά μπόρεσαν να αντιμετωπιστούν το 1958 όπου ο Frank Rosenblatt παρουσίασε μια πιο ολοκληρωμένη μοντελοποίηση του τεχνητού νευρώνα, το perceptron (αντίληπτρο) [56]. Τα επόμενα χρόνια ήταν γεμάτα ανακαλύψεις, αξιοσημείωτη είναι ωστόσο η δημοσίευση των Minsky και Papert το 1969 [43], όπου αποδεικνύουν ότι οι νευρώνες perceptrons μπορούν να εκπαιδευτούν για την επίλυση γραμμικών διαχωριζόμενων προβλημάτων και μόνο.

Ένα από τα πιο χαρακτηριστικά παραδείγματα μη γραμμικών διαχωριζόμενων προβλημάτων είναι το αποκλειστικό OR (XOR)[1]. Προκειμένου να επιλυθεί επιτυχώς το πρόβλημα XOR πρέπει η έξοδος του δικτύου να είναι αληθής αν και μόνο αν μία είσοδος του δικτύου είναι αληθής. Οι Minsky και Papert γνώριζαν ότι αυτό μπορούσε να αντιμετωπιστεί μόνο με τη χρήση πολυεπίπεδων δικτύων, η μέχρι τότε γνώση όμως δεν επέτρεπε την εκπαίδευση ενός τέτοιου δικτύου. Λύση στο πρόβλημα αυτό βρέθηκε να δώσει ο αλγόριθμος backpropagation του Paul Jhon Werbos το 1975.

[1] https://en.wikipedia.org/wiki/XOR_gate



Πηγή: researchgate.net/figure/Single-Perceptron-Haykin-1999_fig1_294583744
Σχήμα 2.2: Η δομή ενός νευρώνα Perceptron

2.3.1 Ο αλγόριθμος Perceptron

Ο νευρώνας perceptron, ή αλγόριθμος perceptron, αποτελεί το βασικό δομικό στοιχείο των σύγχρονων τεχνητών νευρωνικών δικτύων. Στην ουσία, ο αλγόριθμος perceptron είναι ένας δυαδικός ταξινομητής, δηλαδή είναι μια συνάρτηση που μπορεί να αποφανθεί αν η είσοδος ανήκει σε μία κλάση ή όχι. Στο σχήμα 2.2 παρουσιάζεται γραφικά η εσωτερική δομή του νευρώνα, στη συνέχεια περιγράφεται αναλυτικότερα ο αλγόριθμος για ένα perceptron ενός επιπέδου.

Όπως αναφέρθηκε ο αλγόριθμος perceptron αποτελεί επί της ουσίας μία συνάρτηση που αντιστοιχεί ένα διάνυσμα εισόδου $x = [x_1, x_2, x_3, ..., x_n] \in \mathbb{R}^n$ σε μία τιμή εξόδου $\varphi(x)$. Η τιμή της εξόδου υπολογίζεται από τη σχέση:

$$y = \varphi(w \cdot x + b) \tag{2.1}$$

όπου το $w \cdot x = \sum_{i=1}^{n} x_i w_i$ είναι το εσωτερικό γινόμενο του διανύσματος εισόδου x με το διάνυσμα βαρών $w = [w_1, w_2, ..., w_n] \in \mathbb{R}^n$ και b είναι η πόλωση του νευρώνα. Η τιμή της εξόδου y εξαρτάται από τη συνάρτηση ενεργοποίησης φ, συνήθως παίρνει τιμές στο $\mathbb{R}$, ή στο $[0, 1]$ ή ακόμα και στο $\{0, 1\}$.

Η φύση του αλγόριθμου, ως γραμμικός ταξινομητής, του επιτρέπει να ταξινομήσει σωστά όλα τα διανύσματα εισόδου σε δύο κλάσεις, μόνο εάν αυτά είναι γραμμικά διαχωρίσιμα. Σε αντίθετη περίπτωση, αν τα διανύσματα εισόδου δεν είναι γραμμικά διαχωρίσιμα όπως στο πρόβλημα XOR που αναφέρθηκε στην προηγούμενη ενότητα, χρειάζονται περισσότερα επίπεδα νευρώνων για να μπορέσουν να ταξινομηθούν σε δύο κλάσεις. Τα πολυεπίπεδα νευρωνικά δίκτυα θα περιγραφούν με λεπτομέρεια στη συνέχεια (2.3.3).

2.3.2 Θεώρημα Καθολικής Προσέγγισης

Η σημαντικότητα του θεωρήματος καθολικής προσέγγισης (Universal Approximation Theorem) συνδέεται άμεσα με την ισχύ των νευρωνικών δικτύων, συγκεκριμένα των εμπρόσθια τροφοδοτούμενων δικτύων (feedforward networks). Ένα εμπρόσθια τροφοδοτούμενο δίκτυο

είναι ένα νευρωνικό δίκτυο στο οποίο οι συνδέσεις μεταξύ των νευρώνων δεν σχηματίζουν κύκλο. Το θεώρημα καθολικής προσέγγισης αυτό που αποδεικνύει είναι ότι ένα εμπρόσθια τροφοδοτούμενο δίκτυο με μόνο ένα κρυφό επίπεδο, πέρα από τα επίπεδα της εισόδου και της εξόδου, και πεπερασμένο αριθμό νευρώνων μπορεί να προσεγγίσει οποιαδήποτε συνεχή συνάρτηση ορισμένη σε ένα κλειστό σύνολο των πραγματικών αριθμών, με οσοδήποτε μικρό σφάλμα.

Μία πιο επίσημη έκφραση του θεωρήματος καθολικής προσέγγισης διατυπωμένη με μαθηματικούς όρους είναι η παρακάτω. Έστω μια συνάρτηση $\varphi : \mathbb{R} \to \mathbb{R}$ η οποία είναι μη σταθερή, συνεχής και φραγμένη, γνωστή ως συνάρτηση ενεργοποίησης. Συμβολίζουμε με I_m τον m-διάστατο υπερκύβο στο χώρο $[0,1]^m$ και με $C(I_m)$ την κλάση όλων των πραγματικών συνεχών συναρτήσεων στο I_m. Τότε, δεδομένου οποιοδήποτε αριθμού $\varepsilon > 0$ και οποιαδήποτε συνάρτηση $f \in I_m$, υπάρχει ένας ακέραιος N, σταθερές r_i, $b_i \in \mathbb{R}$ και διανύσματα $w_i \in \mathbb{R}^m$ για $i = 1, 2, ..., N$ τέτοια ώστε να οριστεί η συνάρτηση $F(x)$:

$$F(x) = \sum_{i=1}^{N} r_i \varphi({w_i}^T x + b_i) \tag{2.2}$$

ως μια προσέγγιση της συνάρτησης f, τέτοια ώστε να ισχύει:

$$|F(x) - f(x)| < \varepsilon \tag{2.3}$$

για όλα τα $x \in I_m$.

2.3.3 Πολυεπίπεδα Νευρωνικά Δίκτυα

Στις προηγούμενες υποενότητες έγινε αναφορά στα πολυεπίπεδα νευρωνικά δίκτυα (Multilayer Perceptrons-MLP) και στη σημαντικότητα της χρήσης τους. Όπως ειπώθηκε, ένας νευρώνας Perceptron μπορεί να προσεγγίσει μόνο γραμμικές συναρτήσεις, μη γραμμικά προβλήματα απαιτούν την χρήση μη γραμμικών συναρτήσεων, οι οποίες μπορούν να προσεγγιστούν από πολλαπλά επίπεδα νευρώνων Perceptron. Ωστόσο πολλές φορές, ο όρος πολυεπίπεδα νευρωνικά δίκτυα χρησιμοποιείται γενικότερα για να περιγράψει όλα τα εμπρόσθια τροφοδοτούμενα νευρωνικά δίκτυα και όχι μόνο δίκτυα που αποτελούνται από νευρώνες Perceptrons. Τα πολυεπίπεδα νευρωνικά δίκτυα συνθέτονται από τουλάχιστον τρία επίπεδα νευρώνων όπου το πρώτο επίπεδο ονομάζεται επίπεδο εισόδου και περιγράφει την είσοδο του δικτύου, το τελευταίο ονομάζεται επίπεδο εξόδου και περιγραφεί την έξοδο του δικτύου ενώ όλα τα ενδιάμεσα ονομάζονται ¨κρυφά' επίπεδα (hidden layers). Επί της ουσίας, κάθε επίπεδο έχει ως είσοδο την έξοδο του προηγούμενου και τροφοδοτεί με την εξοδό του το επόμενο. Συγκεκριμένα, στις οριακές περιπτώσεις, είσοδο στο επίπεδο εισόδου αποτελούν τα δεδομένα εισόδου ενώ η έξοδος στο επίπεδο εξόδου συνθέτει την έξοδο του δικτύου.

Στην περίπτωση που τα ενδιάμεσα επίπεδα είναι παραπάνω από ένα τότε αναφερόμαστε στην κατηγορία των βαθιά νευρωνικών δικτύων (Deep Neural Networks-DNN). Τα βαθιά νευρωνικά δίκτυα χρησιμοποιούνται κατά κόρον σε εφαρμογές βαθιάς μηχανικής μάθησης (Deep Learning), όπου η αρχιτεκτονική τους βοηθά στην εξαγωγή αφηρημένων χαρακτηριστικών. Για παράδειγμα, σε ένα νευρωνικό δίκτυο με τρία κρυφά επίπεδα, όπου ως είσοδο θα είχε

εικόνες προσώπων και θα προσπαθούσε να τα αναγνωρίσει, η διαδικασία εξαγωγής χαρακτηριστικών θα γινόταν διαδοχικά. Δηλαδή, το πρώτο κρυφό επίπεδο θα μπορούσε να αναγνωρίζει απλά χαρακτηριστικά όπως γραμμές και απλές καμπύλες, το δεύτερο κρυφό επίπεδο θα μπορούσε να αναγνωρίζει πιο σύνθετα χαρακτηριστικά του προσώπου όπως τα μάτια ή η μύτη και τέλος το τρίτο κρυφό επίπεδο να αναγνωρίζει ολόκληρη την σύνθεση ενός προσώπου.

Τα πολυεπίπεδα νευρωνικά δίκτυα εκπαιδεύονται, δηλαδή ανανεώνουν τα βάρη τους, κάνοντας χρήση μια τεχνική επιβλεπόμενης μάθησης που ονομάζεται οπισθοδιάδοση (Backpropagation). Στην περίπτωση ενός προβλήματος ταξινόμησης της εισόδου σε τρεις κατηγορίες, το δίκτυο θα ανανέωνε τα βάρη του με τέτοιο τρόπο ώστε το επίπεδο εξόδου να αποδώσει τις κατάλληλες τιμές και η είσοδος να ταξινομηθεί στη σωστή κατηγορία. Οι τιμές που θα δώσει η έξοδος αλλά και κάθε επίπεδο του δικτύου εξαρτάται από την συνάρτηση ενεργοποίησης που χρησιμοποιεί κάθε επίπεδο. Στο παράδειγμα ταξινόμησης της εισόδου σε τρεις κατηγορίες το επίπεδο εξόδου θα αποτελούνταν από τρεις νευρώνες, όπου κάθε ένας θα αντιπροσώπευε μία κατηγορία, και η συνάρτηση ενεργοποίησης των νευρώνων 2.16 θα ήταν υπεύθυνη για την ταξινόμηση της εισόδου στις τρεις κατηγορίες.

2.3.4 Αλγόριθμοι Εκπαίδευσης και Βασικές Συναρτήσεις

2.3.4.1 Συνάρτηση Κόστους

Η συνάρτηση κόστους (Cost function) είναι μία μετρική της επίδοσης των συστημάτων μηχανικής μάθησης. Η επίδοση στα συστήματα μηχανικής μάθησης συνήθως μετράται ως τη διαφορά που έχουν οι προβλεπόμενες τιμές του συστήματος σε σχέση με τις αναμενόμενες τιμές. Σκοπός λοιπόν της συνάρτησης κόστους είναι να υπολογίσει το σφάλμα μεταξύ προβλεπόμενων και πραγματικών τιμών. Οι τιμές που αποκτά το σφάλμα συνηθίζεται να είναι πραγματικοί αριθμοί. Ο σκοπός της συνάρτησης κόστους είναι τις περισσότερες φορές να ελαχιστοποιήσει την τιμή της, άλλα ενδέχεται να πρέπει και να την μεγιστοποιήσει, ανάλογα με τη φύση του προβλήματος. Για παράδειγμα, αν η συνάρτηση κόστους υπολογίζει μία ανταμοιβή για το σύστημα τότε επιθυμητό θα ήταν να μεγιστοποιήσει την τιμή της. Ενώ σε αντίθετη περίπτωση, αν η συνάρτηση κόστους υπολογίζει ένα κόστος του συστήματος, όπως δηλώνει και το όνομά της, τότε η τιμή της θα πρέπει να ελαχιστοποιηθεί.

Ένα παράδειγμα συνάρτησης που προσπαθεί να ελαχιστοποιήσει την τιμή της είναι η συνάρτηση μέσου τετραγωνικού σφάλματος που περιγράφεται με την σχέση 2.4. Στη συνάρτηση μέσου τετραγωνικού σφάλματος (Mean Squarred Error) μετράται το τετραγωνικό σφάλμα μεταξύ της προβλεπόμενης ($f(x_i|\vartheta)$) και της αναμενόμενης (y_i) τιμής της εξόδου για N δεδομένα εισόδου (x_i) και παραμέτρους δικτύου (ϑ).

$$MSE(\vartheta) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i|\vartheta) - y_i)^2 \tag{2.4}$$

Πρέπει να γίνει κατανοητό ότι η επιτυχία ενός νευρωνικού δικτύου συνδέεται με την ικανότητα του να γενικεύσει. Δηλαδή, για ένα σύνολο δεδομένων εκπαίδευσης το δίκτυο θα πρέπει να είναι σε θέση να προβλέψει με επιτυχία και για το σύνολο των δεδομένων δοκιμής.

Η τιμή της συνάρτησης κόστους μειώνεται όσο το δίκτυο εκπαιδεύεται με τα δεδομένα εκπαίδευσης. Η εκτεταμένη εκπαίδευση του δικτύου συχνά μπορεί να οδηγήσει και στο πρόβλημα της υπερπροσαρμογής (overfitting), αυτό ουσιαστικά σημαίνει ότι το δίκτυο μαθαίνει να προβλέπει πολύ καλά για τα ΄γνωστά΄ δεδομένα εκπαίδευσης ενώ αδυναμεί στις προβλέψεις των ΄άγνωστων΄ δεδομένων δοκιμής. Σκοπός του δικτύου,όπως αναφέρθηκε, είναι να μπορέσει να γενικεύσει να μπορεί δηλαδή να προβλέψει με μεγάλη ακρίβεια για άγνωστα, στο δίκτυο, δεδομένα και όχι να προσεγγίσει την συνάρτηση που διέπει τα δεδομένα εκπαίδευσης.

2.3.4.2 Αλγόριθμος Backpropagation

Ο αλγόριθμος Backpropagation είναι ένα από τα σημαντικότερα δομικά στοιχεία ενός νευρωνικού δικτύου. Ο όρος backpropagation και η γενική χρήση του στα νευρωνικά δίκτυα αρχικά εδραιώθηκε το 1986 [13] από τους Rumelhart, Hinton και Williams, ενώ η ιδέα για την διαδικασία που περιγράφει ο αλγόριθμος υπήρχε ήδη από τη δεκαετία του 60. Ο αλγόριθμος ουσιαστικά χρησιμοποιείται για να εκπαιδεύσει ένα δίκτυο κάνοντας χρήση του κανόνα της αλυσίδας. Μια απλή περιγραφή είναι ότι μετά από κάθε εμπρόσθιο πέρασμα (forward pass) στο δίκτυο ο αλγόριθμος πραγματοποιεί ένα πέρασμα με αντίθετη φορά (backward pass) υπολογίζοντας τις μερικές παραγώγους των παραμέτρων του δικτύου, δηλαδή τα βάρη και τις πολώσεις. Παρακάτω γίνεται μια μαθηματική περιγραφή της διαδικασίας αυτής.

Βασικός στόχος του αλγορίθμου είναι να υπολογίσει τη μερική παράγωγο μιας συνάρτησης κόστους, έστω J αναλυτική αναφορά γίνεται στην υποενότητα 2.3.4.1, ως προς τις παραμέτρους του δικτύου. Έστω ένας κόμβος με βάρος $w^l_{j,k}$ και πόλωση b^l_j, όπου τα j,k δηλώνουν την θέση του στοιχείου στο αντίστοιχο διάνυσμα ενώ το l τον αριθμό του επιπέδου στο οποίο αναφερόμαστε. Η μερική παράγωγος του βάρους $w^l_{j,k}$ υπολογίζεται με τον κανόνα της αλυσίδας ως:

$$\frac{\partial J}{\partial w^l_{j,k}} = \frac{\partial J}{\partial z^l_j}\frac{\partial z^l_j}{\partial w^l_{j,k}} \tag{2.5}$$

Εξ ορισμού, έστω m ο αριθμός των κόμβων στο επίπεδο $l-1$:

$$z^l_j = \sum_{k=1}^{m} w^l_{j,k} a^{l-1}_k + b^l_j \tag{2.6}$$

λόγω παραγώγισης προκύπτει:

$$\frac{\partial J}{\partial w^l_{j,k}} = a^{l-1}_k \tag{2.7}$$

Άρα η τελική τιμή της μερικής παραγώγου είναι:

$$\frac{\partial J}{\partial w^l_{j,k}} = \frac{\partial J}{z^l_j} a^{l-1}_k \tag{2.8}$$

Η ίδια διαδικασία ακολουθείται και για τον υπολογισμό της μερικής παραγώγου της πόλωσης b^l_j:

$$\frac{\partial J}{\partial b^l_j} = \frac{\partial J}{\partial z^l_j}\frac{\partial z^l_j}{\partial b^l_j} \tag{2.9}$$

Λόγω της εξίσωσης 2.6 προκύπτει:

$$\frac{\partial J}{\partial b_j^l} = 1 \tag{2.10}$$

και τελικά σε αυτή την περίπτωση η τιμή της μερικής παραγώγου είναι:

$$\frac{\partial J}{\partial b_j^l} = \frac{\partial J}{z_j^l} \tag{2.11}$$

Στη συνέχεια, οι τιμές των μερικών παραγώγων που υπολογίστηκαν από τον αλγόριθμο backpropagation θα χρησιμοποιηθούν από τον αλγόριθμο κατάβασης κλήσης για την ανανέωση των παραμέτρων του δικτύου. Η διαδικασία αυτή αναλύεται στην επόμενη ενότητα 2.3.4.3.

2.3.4.3 Αλγόριθμος Κατάβασης Κλίσης

Ο αλγόριθμος κατάβασης κλίσης (Gradient Descent Algorithm) είναι ένας αλγόριθμος βελτιστοποίησης και στον τομέα της μηχανικής μάθησης η χρήση του είναι πολύ συχνή. Πιο συγκεκριμένα, ο αλγόριθμος προσπαθεί να ελαχιστοποιήσει μια συνάρτηση κόστους, έστω $J(\vartheta)$ όπου ϑ είναι το διάνυσμα που αντιπροσωπεύει τις παραμέτρους του δικτύου, εν προκειμένω τα βάρη και τις πολώσεις. Ο αλγόριθμός κατάβασης κλίσης είναι επαναληπτικός και σε κάθε επανάληψη αφαιρεί μια μικρή ποσότητα από τις παραμέτρους του συστήματος. Την ποσότητα αυτή συνθέτει το γινόμενο της μερικής παραγώγου της συνάρτησης κόστους ως προς την παράμετρο που ανανεώνεται, όπως υπολογίστηκε από τον αλγόριθμο Backpropagation, επί την παράμετρο r που ονομάζεται ρυθμός μάθησης (learning rate) και καθορίζει το μέγεθος του βήματος εκμάθησης που θα κάνει κάθε φόρα ο αλγόριθμος. Με την παρακάτω σχέση 2.12 περιγράφεται πως γίνεται η ανανέωση της παραμέτρου θ_j τη χρονική στιγμή $t+1$

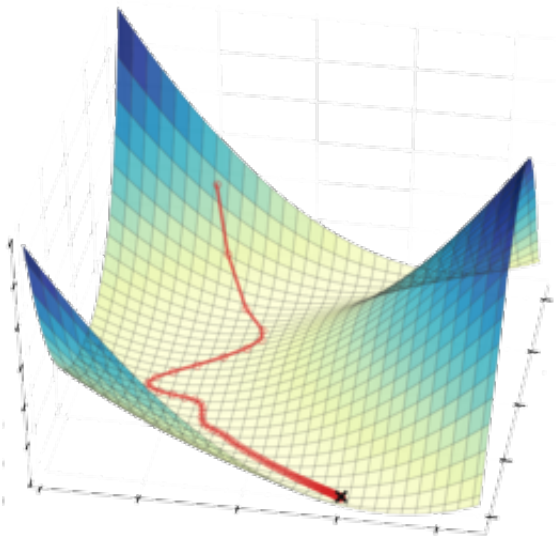
$$\vartheta_{j,t+1} = \vartheta_{j,t} - r\frac{\partial}{\partial \vartheta_{j,t}} J(\vartheta) \tag{2.12}$$

Στην παρακάτω εικόνα 2.3 γίνεται παρουσίαση ενός παραδείγματος της διαδρομής που ακολουθεί ο αλγόριθμος κατάβασης κλίσης προσεγγίζοντας την ελάχιστη τιμή της συνάρτησης κόστους.

Πολλές φορές ο αλγόριθμος κατάβασης κλίσης συναντάται και σε διάφορες εκδοχές οι οποίες, σε αντίθεση με την γενική περίπτωση (Batch Gradient Descent - BGD), χρησιμοποιούν ένα μέρος των δεδομένων εισόδου κάθε φορά και όχι το σύνολό τους. Συγκεκριμένα, αυτές οι εκδοχές του αλγόριθμου, χωρίζουν τα δεδομένα εισόδου σε τεμάχια (batches), όπου κάθε τέτοιο τεμάχιο χρησιμοποιείται με την σειρά για την ανανέωση των παραμέτρων του δικτύου.

Ο Batch Gradient Descent θα υπολογίσει την παράγωγο σε όλα τα δεδομένα και θα πραγματοποιήσει μια ανανέωση. Για το λόγο αυτό μπορεί να είναι πολύ αργός και για μεγάλα δεδομένα εισόδου που δεν χωρούν στη μνήμη ο έλεγχος να προκύψει πολύ δύσκολος.

Μια εκδοχή του αλγόριθμου κατάβασης κλήσης είναι ο Stohastic Gradient Descent (SGD), ο οποίος σε αντίθεση με τον προηγούμενο πραγματοποιεί ανανέωση παραμέτρων για κάθε δεδομένο εκπαίδευσης. Λόγω των συχνών ανανεώσεων, η ενημέρωση των παραμέτρων παρουσιάζει υψηλή διακύμανση και με την σειρά της η συνάρτηση κόστους λαμβάνει κυμαινόμενες τιμές.



Πηγή: datasciencecentral.com/profiles/blogs/alternatives-to-the-gradient-descent-algorithm

Σχήμα 2.3: Γραφική απεικόνιση του αλγόριθμου κατάβασης κλίσης

Αυτό είναι εξαιρετικά χρήσιμο, επειδή επιτρέπει στην συνάρτηση κόστους να εντοπίσει νέα τοπικά ελάχιστα. Όμως, οι συνεχείς διακυμάνσεις μπορούν να προκαλέσουν την συνεχή προσπέραση του ολικού ελαχίστου, έτσι τελικά η σύγκλιση του αλγορίθμου να προκύψει αρκετά αργή.

Μια άλλα εκδοχή του αλγορίθμου που αντιμετωπίζει το πρόβλημα αυτό είναι ο Mini Batch Gradient Descent (MBGD), που χρησιμοποιεί τα προτερήματα των δύο προηγούμενων μεθόδων. Σύμφωνα με αυτή την τεχνική τα δεδομένα εκπαίδευσης χωρίζονται σε μικρά τεμάχια που τροφοδοτούνται στο δίκτυο με τη σειρά και η ανανέωση των παραμέτρων γίνεται για κάθε τεμάχιο. Αυτή η τεχνική προκαλεί μικρότερες διακυμάνσεις στην ανανέωση των παραμέτρων με αποτέλεσμα η σύγκλιση του αλγόριθμου να είναι αρκετά πιο γρήγορη και αποτελεσματική.

2.3.4.4 Αλγόριθμοι Βελτιστοποίησης

Αντικείμενο της βελτιστοποίησης (Optimization) στα συστήματα μηχανικής μάθησης είναι η ελαχιστοποίηση, πιο σπάνια η μεγιστοποίηση, μιας συνάρτησης στόχου. Παραδείγματος χάρη ο αλγόριθμος κατάβασης κλίσης που παρουσιάστηκε στην προηγούμενη υποενότητα 2.3.4.3 είναι και αυτός ένας αλγόριθμος βελτιστοποίησης που ελαχιστοποιεί μια συνάρτηση κόστους. Για την αποδοτική και αποτελεσματική εκπαίδευση του συστήματος ο σωστός υπολογισμός και η ανανέωση των εσωτερικών παραμέτρων παίζει πολύ σημαντικό ρόλο. Για τον λόγο αυτό στρατηγικές και αλγόριθμοι βελτιστοποίησης εφαρμόζονται για τον υπολογισμό των παραμέτρων του συστήματος που θα επηρεάσουν την διαδικασία εκπαίδευσης και την έξοδό του.

Οι αλγόριθμοι βελτιστοποίησης χωρίζονται σε δύο κατηγορίες, με βάση την τάξη της παραγώγου που χρησιμοποιούν. Οι αλγόριθμοι βελτιστοποίησης *Πρώτης Τάξης (First Order*

Optimization Algorithms) υπολογίζουν και χρησιμοποιούν την πρώτη παράγωγο της συνάρτησης στόχου με σκοπό την βελτιστοποίηση της. Η παράγωγος πρώτης τάξης μιας συνάρτησης σε ένα σημείο ουσιαστικά υπολογίζει την τιμή της κλίσης της στο σημείο αυτό, δηλαδή αν η συνάρτηση έχει την τάση να αυξήσει ή να μειώσει την τιμή της στο επόμενο σημείο. Ένας αλγόριθμός βελτιστοποίησης πρώτης τάξης είναι για παράδειγμα ο αλγόριθμος κατάβασης κλίσης.

Η δεύτερη κατηγορία αλγόριθμων βελτιστοποίησης είναι οι *Αλγόριθμοι Βελτιστοποίησης Δεύτερης Τάξης (Second Order Optimization Algorithms)*. Όπως είναι αναμενόμενο, οι αλγόριθμοι αυτοί υπολογίζουν και χρησιμοποιούν τις παραγώγους δεύτερης τάξης της συνάρτησης στόχου. Η δεύτερη παράγωγος μιας συνάρτησης σε ένα σημείο εκφράζει την κλίση της πρώτης παραγώγου στο σημείο αυτό, δηλαδή την κυρτότητα της συνάρτησης. Όμως ο υπολογισμός της δεύτερης παραγώγου είναι υπολογιστικά πολύ πιο δαπανηρός, έτσι οι αλγόριθμοι αυτοί δεν χρησιμοποιούνται συχνά στην πράξη. Το πλεονέκτημα των αλγόριθμων δεύτερης τάξης είναι ότι είναι αποτελεσματικότεροι καθώς δεν αγνοούν την καμπυλότητα της επιφάνειας.

Στη συνέχεια, με σκοπό την αντιμετώπιση των έντονων ταλαντώσεων που εμφανίζει ο αλγόριθμος κατάβασης κλίσης και κατά συνέπεια την δυσκολία σύγκλισης, παρουσιάζεται μια τεχνική βελτιστοποίησης που ονομάζεται *Ορμή (Momentum)* . Η τεχνική αυτή εισάγει έναν όρο γ στην 2.12 με αποτέλεσμα να επιταχύνει την διαδικασία σύγκλισης οδηγώντας τις παραμέτρους προς την σχετική κατεύθυνση και εξομαλύνοντας τις ταλαντώσεις στις άσχετες κατευθύνσεις. Όταν χρησιμοποιείται και η ορμή για την ανανέωση παραμέτρων ϑ η σχέση που τις ανανεώνει δίνεται από τις εξισώσεις:

$$V(t) = \gamma V(t-1) + r\nabla J(\vartheta) \tag{2.13}$$

$$\vartheta = \vartheta - V(t) \tag{2.14}$$

, μία τυπική τιμή για την ορμή είναι $\gamma = 0.9$.

Τρεις από τους πιο συνηθισμένους αλγόριθμους βελτιστοποίησης που χρησιμοποιούν την τεχνική της ορμής είναι οι παρακάτω:

- *Adagrad*

  Ο Adagrad [36] επιτρέπει στον ρυθμό μάθησης r να προσαρμόζεται σε κάθε παράμετρο βασιζόμενος στις παρελθοντικές κλίσεις (gradients) του. Έτσι για παραμέτρους που ανανεώνουν τις τιμές τους συχνά, ο αλγόριθμος κάνει μικρά βήματα στην ενημέρωση αυτών των παραμέτρων. Ενώ σε αντίθετη περίπτωση κάνει μεγάλα βήματα για παραμέτρους που οι τιμές τους ανανεώνονται λιγότερο συχνά. Ένα πλεονέκτημα του αλγόριθμου είναι ότι δεν απαιτείται να γίνει χειροκίνητη ρύθμιση του ρυθμού εκμάθησης, ενώ μειονέκτημα είναι ότι ο ρυθμός εκμάθησης παίρνει συνεχώς μικρότερες τιμές περιορίζοντας σημαντικά τη διαδικασία εκμάθησης.

- *AdaDelta*

Ο αλγόριθμος AdaDelta [69] είναι μία βελτίωση του Adagrad που αντιμετωπίζει το πρόβλημα του φθίνοντος ρυθμού εκμάθησης. Χρησιμοποιεί μόνο ένα τμήμα των προηγούμενων τιμών της κλίσης της συνάρτησης για την ανανέωση των παραμέτρων σε αντίθεση με τον Adagrad που χρειαζόταν όλες τις προηγούμενες τιμές της κλίσης. Ένα ακόμη προτέρημα του αλγόριθμου είναι ότι δεν χρειάζεται προκαθορισμένη τιμή για το ρυθμό εκμάθησης.

- *Adam*

  Ο αλγόριθμός Adam (Adaptive Moment Estimation) [30] είναι και αυτός μία μέθοδος προσαρμοστική ως προς το ρυθμό εκμάθησης. Ο αλγόριθμός χρησιμοποιεί τις τιμές των ροπών πρώτης (μέσος όρος) και δεύτερης (διακύμανσης) τάξης για την ανανέωση των παραμέτρων. Σε πολλά σύγχρονα συστήματα μηχανικής μάθησης γίνεται χρήση του αλγόριθμου Adam λόγω της ικανότητας του να συγκλίνει πολύ γρήγορα και να αντιμετωπίζει τα προβλήματα που εμφανίζουν οι προηγούμενες τεχνικές.

2.3.4.5 Συνάρτηση Ενεργοποίησης

Η συνάρτηση ενεργοποίησης (Activation function) είναι η συνάρτηση που σε ένα τεχνητό νευρωνικό δίκτυο θα εφαρμοστεί στην έξοδο κάθε κόμβου και η τιμή της θα τροφοδοτηθεί στο επόμενο επίπεδο του δικτύου. Η χρήση της συνάρτησης ενεργοποίησης στα νευρωνικά δίκτυα είναι απαραίτητη καθώς εισάγει μη γραμμικότητα στο σύστημα αλλά και κανονικοποιεί την έξοδο των νευρώνων σε ένα κλειστό σύνολο.

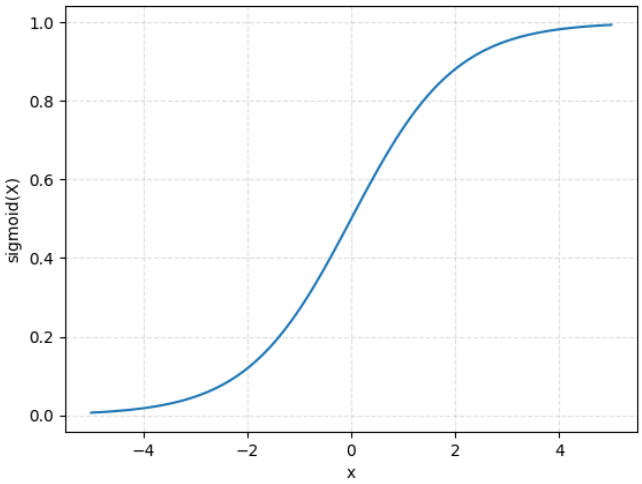
Για να γίνει πιο κατανοητή η σημασία της κανονικοποίησης της εξόδου, ας υποθέσουμε ότι έχουμε ένα νευρωνικό δίκτυο και εξετάζουμε έναν κόμβο του δικτύου με είσοδο το διάνυσμα x, βάρη w και πόλωση b. Η τιμή της εξόδου y, όπως φαίνεται στη σχέση 2.15, μπορεί να παίρνει τιμές στο $(-\infty, +\infty)$. Με την εφαρμογή της συνάρτησης ενεργοποίησης 2.16, έστω f καταφέρνουμε να αντιστοιχίσουμε την τιμή της εξόδου του νευρώνα z στο επιθυμητό σύνολο, συνήθως στο $[0, 1]$. Αυτή η διαδικασία είναι απαραίτητη ώστε ο νευρώνας να μπορέσει να διαχωρίσει τις περιπτώσεις που θα πρέπει να μεταδώσει την τιμή στα επόμενα επίπεδα ή να την αποκόψει. Τιμές κοντά στο ένα έχουν αποτέλεσμα την πυροδότηση (ενεργοποίηση) του νευρώνα ενώ τιμές κοντά στο μηδέν την αποκοπή του.

$$y = w \cdot x + b, y \in (-\infty, +\infty) \tag{2.15}$$

$$z = f(y) = f(w \cdot x + b), z \in [0, 1] \tag{2.16}$$

Μια άλλη σκοπιμότητα της συνάρτησης ενεργοποίησης είναι η εισαγωγή μη γραμμικότητας στο σύστημα, όπως αναφέρθηκε. Η μη γραμμικότητα σε ένα σύστημα μηχανικής μάθησης είναι αναγκαία ώστε να μπορέσει να προσεγγίσει μη γραμμικές συναρτήσεις. Η φύση πολλών συναρτήσεων ενεργοποίησης ως μη γραμμικές εξυπηρετούν τη σκοπιμότητα αυτή. Στη συνέχεια γίνεται μια σύντομη παρουσίαση των συνηθέστερων συναρτήσεων ενεργοποίησης:

- *Σιγμοειδής Συνάρτηση*

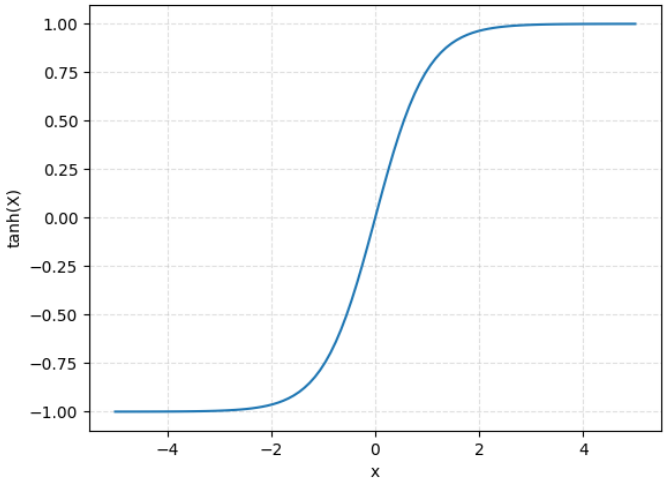


Σχήμα 2.4: Σιγμοειδής συνάρτηση

Η γραφική αναπαράσταση της σιγμοειδούς συνάρτησης Sigmoid function φαίνεται στο σχήμα 2.8, γίνεται εύκολα κατανοητή αλλά στην πράξη δεν χρησιμοποιείται συχνά. Η σιγμοειδής έχει πεδίο τιμών το $(0, 1)$ αλλά αυτό δημιουργεί πρόβλημα γιατί δεν είναι κεντραρισμένο στο μηδέν. Δύο ακόμα προβλήματα που εμφανίζει είναι η εξαφάνιση της κλίσης (Vanishing gradient problem) όπως επίσης και η αργή σύγκλιση που την χαρακτηρίζει στο στάδιο της εκπαίδευσης.

$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.17}$$

- *Υπερβολική εφαπτομένη*

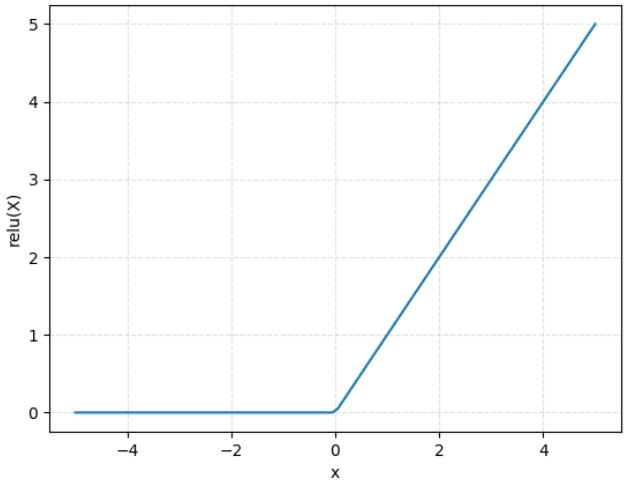


Σχήμα 2.5: Υπερβολική εφαπτομένη

Η υπερβολική εφαπτομένη (Hyperbolic tangent) 2.5 μοιάζει αρκετά με την σιγμοειδή συνάρτηση. Υπερτερεί σε σχέση με την σιγμοειδή καθώς το πεδίο τιμών της παίρνει τιμές στο $(-1, 1)$ και είναι κεντραρισμένο στο μηδέν. Ένα ακόμα προτέρημα είναι η μεγαλύτερη κλίση της που βοήθα στην ταχύτερη σύγκλιση. Όπως και η σιγμοειδής δεν καταφέρνει να αντιμετωπίσει το πρόβλημα της εξαφάνισης κλίσης.

$$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1 = 2\sigma(2x) - 1 \tag{2.18}$$

- *Rectified Linear Unit (ReLU)*

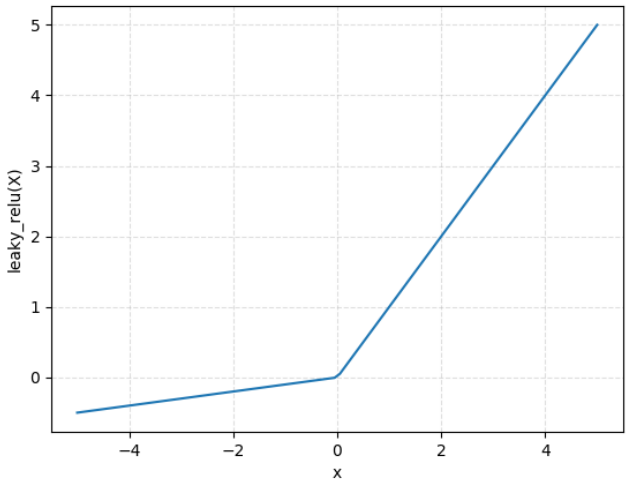


Σχήμα 2.6: Συνάρτηση ReLU

Η συνάρτηση ReLU παρουσιάζει αρκετά προτερήματα και η χρήση της στα σύγχρονα συστήματα μηχανικής μάθησης είναι συχνή. Αντιμετωπίζει με επιτυχία το πρόβλημα της εξαφάνισης κλίσης και η σύγκλισή της είναι ταχύτερη από τις δύο προηγούμενες μεθόδους. Η μαθηματική εξίσωση της ReLU 2.19 είναι πολύ απλή και αποδοτική, ο μόνος περιορισμός που εμφανίζει είναι η χρήση της μόνο στα ενδιάμεσα επίπεδα ενός νευρωνικού δικτύου.

$$f(x) = max(0, x) = \begin{cases} x, & x > 0 \\ 0, & otherwise \end{cases} \tag{2.19}$$

- *Leaky Rectified Linear Unit (Leaky ReLU)*

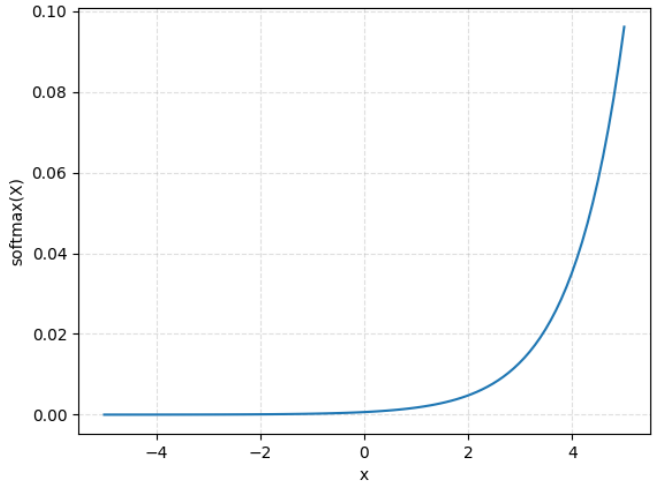


Σχήμα 2.7: Συνάρτηση Leaky ReLU

Η συνάρτηση Leaky ReLU είναι μια βελτιωμένη εκδοχή της ReLU που καταφέρνει να δώσει λύση στο πρόβλημα των Νεκρών Νευρώνων (Dead Neurons) που μπορεί να προκαλέσει η ReLU. Σύμφωνα με το πρόβλημα αυτό η ReLU μπορεί κατά την διάρκεια εκπαίδευσης να οδηγήσει στην ανανέωση κάποιου βάρους και να προκαλέσει απενεργοποίηση της εξόδου για όλα τα δεδομένα εισόδου. Όπως φαίνεται στο σχήμα 2.7, σε αντίθεση με την ReLU, διατηρεί μια πολύ μικρή κλίση για της αρνητικές τιμές αποφεύγοντας έτσι το πρόβλημα των νεκρών νευρώνων.

$$f(x) = \begin{cases} x, & x > 0 \\ \alpha x, & x \leq 0 \end{cases} \tag{2.20}$$

- *Συνάρτηση Softmax*



Σχήμα 2.8: Συνάρτηση Softmax

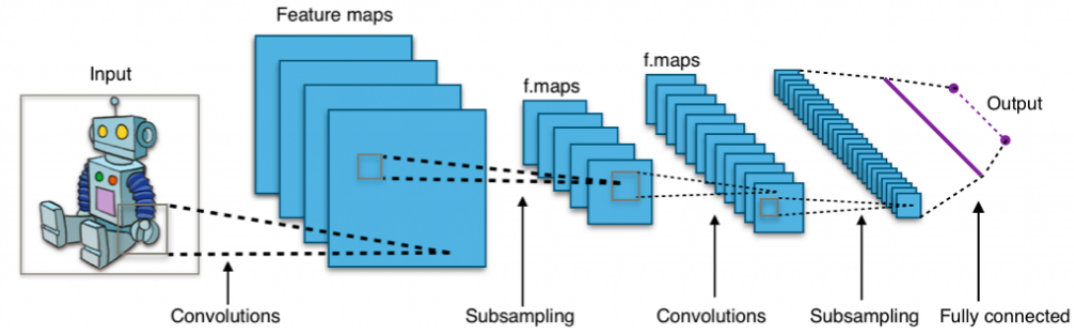
Η συνάρτηση softmax ξεχωρίζει από τις προηγούμενες κατηγορίες συναρτήσεων καθώς εκφράζει πιθανότητες στην έξοδο της. Χρησιμοποιείται συχνά για την κατηγοριοποίηση των δεδομένων εισόδου σε κλάσεις, για αυτό τον λόγο συναντάται συνήθως ως συνάρτηση ενεργοποίησης του τελευταίου επιπέδου σε ένα νευρωνικό δίκτυο. Πιο συγκεκριμένα, η softmax κανονικοποιεί τις εξόδους για κάθε κλάση και στη συνέχεια διαιρεί με το άθροισμα όλων το εξόδων, έτσι εκφράζει την πιθανότητα κάθε είσοδος να ανήκει στην αντίστοιχη κλάση. Η σχέση 2.21 δίνει την δυνατότητα οι έξοδοι να εκφράζονται ως πιθανότητες.

$$\sum_{k=1}^{K} y_k = \sum_{k=1}^{K} Softmax(z)_k = 1 \tag{2.21}$$

2.4 Συνελικτικά Νευρωνικά Δίκτυα (CNN)

Τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNNs) αποτελούν μία κλάση των τεχνητών νευρωνικών δικτύων. Η χρήση τους είναι πολύ συχνή στην ανάπτυξη συστημάτων βαθιάς μηχανικής μάθησης και πιο συγκεκριμένα στην ανάπτυξη εφαρμογών που απαιτούν την ανάλυση εικόνας και βίντεο. Το όνομά ΄συνελικτικά νευρωνικά δίκτυα΄ προέρχεται από την μαθηματική πράξη της συνέλιξης. Η συνέλιξη είναι μια ειδική γραμμική πράξη και τα συνελικτικά νευρωνικά δίκτυα την χρησιμοποιούν ως μια γενική μέθοδο πολλαπλασιασμού πινάκων.

Ιστορικά, το 1968 η εργασία των D. Hubel και T. Wiesel [12] για τον διαχωρισμό των οπτικών κυττάρων του εγκεφάλου σε απλά και σύνθετα άνοιξε το δρόμο για την δημιουργία των συνελικτικών νευρωνικών δικτύων. Αργότερα, το 1980 ο Kunihiko Fukushima εμπνευσμένος από τους προηγούμενους θα είναι ο πρώτος που θα παρουσιάσει το τεχνητό νευρωνικό



Πηγή: unite.ai/what-are-convolutional-neural-networks/
Σχήμα 2.9: Συνελικτικό Νευρωνικό Δίκτυο

δίκτυο neocognitron [20] που θα περιλαμβάνει συνελικτικά και υποδειγματοληπτικά επίπεδα. Για να γίνουν δημοφιλή τα CNNs στα συστήματα βαθιάς μηχανικής μάθησης χρειάστηκαν να περάσουν αρκετά χρόνια. Το 2012 ο Alex Krizhevsky με την αρχιτεκτονική που παρουσίασε ως AlexNet [2] κατάφερε να κερδίσει τον διαγωνισμό αναγνώρισης εικόνων ImageNet. Συστήματα που βασίζονται στην αρχιτεκτονική AlexNet αναπτύσσονται μέχρι και σήμερα.

2.4.1 Τρόπος λειτουργίας

Τα συνελικτικά νευρωνικά δίκτυα, όπως προαναφέρθηκε, έχουν σχεδιαστεί για την ανάλυση εικόνων. Η αρχιτεκτονική τους είναι ανάλογη με τα συνδετικά πρότυπα ενός ανθρώπινου εγκέφαλου και η οργάνωση τους είναι εμπνευσμένη από τον οπτικό φλοιό. Σε αντιστοίχιση με τον ανθρώπινο εγκέφαλο τα CNNs λαμβάνοντας μια εικόνα καταφέρνουν να ξεχωρίσουν και να αναγνωρίσουν τα σημεία ενδιαφέροντος. Τα σημεία αυτά σε μία εικόνα είναι δισδιάστατα σχήματα και τα CNNs καταφέρνουν να τα αναγνωρίσουν ακόμη και όταν οι παραμορφώσεις της εικόνας είναι υψηλές. Για να το καταφέρουν αυτό τα συνελικτικά νευρωνικά δίκτυα εκπαιδεύονται με έναν επιβλεπόμενο τρόπο ακολουθώντας συγκεκριμένα βήματα εκπαίδευσης. Τα βήματα που ακολουθούν είναι:

- *Εξαγωγή Χαρακτηριστικών (Feature Extraction)*

  Σαν πρώτο βήμα κάθε νευρώνας εξάγει τοπικά χαρακτηριστικά αξιοποιώντας τις εισόδους που παίρνει από το προηγούμενο επίπεδο. Για κάθε χαρακτηριστικό που εξάγεται διατηρείται η πληροφορία για τη σχετική θέση του και μειώνεται η σημαντικότητα της, έτσι ώστε να αποκτούν προτεραιότητα τα χαρακτηριστικά που δεν έχουν αναγνωριστεί ακόμα.

- *Αντιστοίχιση Χαρακτηριστικών (Feature Mapping)*

  Κάθε χαρακτηριστικό που υπολογίστηκε από την προηγούμενη διαδικασία αποθηκεύεται στη συνέχεια σε έναν χάρτη χαρακτηριστικών. Πολλοί τέτοιοι χάρτες εξάγονται από κάθε συνελικτικό επίπεδο σε ένα CNN. Κάθε χάρτης προέρχεται από την συνέλιξη ενός φίλτρου και του διανύσματος που δέχεται ως είσοδο το επίπεδο. Η διαδικασία αυτή, του φιλτραρίσματος, συνήθως μειώνει τις ελεύθερες παραμέτρους και ακολουθείται από μια

διαδικασία ενεργοποίησης. Ως συνάρτηση ενεργοποίησης στο βήμα αυτό χρησιμοποιείται συνήθως η ReLU.

- *Υποδειγματοληψία (Downsampling)*

  Μετά από κάθε συνελικτικό επίπεδο υπάρχει και ένα συγκεντρωτικό επίπεδο. Σκοπός του επιπέδου αυτού είναι να μειώσει τις διαστάσεις των χαρτών χαρακτηριστικών ώστε ελαττωθεί η υπολογιστική ισχύς που απαιτείται για την επεξεργασία των δεδομένων.

- *Αντιστοίχιση Προβλέψεων (Prediction Mapping)*

  Στο τέλος ενός συνελικτικού νευρωνικού δικτύου βρίσκεται μια σειρά από πλήρως συνδεδεμένα επίπεδα (Fully Connected Layers). Τα πλήρως συνδεδεμένα επίπεδα ακολουθούν τα επίπεδα που περιγράφηκαν προηγουμένως με σκοπό τη μετατροπή της πληροφορίας που ρέει στο δίκτυο στην επιθυμητή έξοδο και την εξαγωγή προβλέψεων.

Ο τρόπος λειτουργίας των συνελικτικών νευρωνικών δικτύων ακολουθεί τα πρότυπα των πολυεπίπεδων νευρωνικών δικτύων. Ένα προωθητικό πέρασμα (forward pass), που τροφοδοτεί τα δεδομένα εισόδου στο δίκτυο, ακολουθείται από ένα οπισθοδρομικό (backward pass) πέρασμα και η διαδικασία αυτή επαναλαμβάνεται αρκετές φορές. Στο προωθητικό πέρασμα η ροή της πληροφορίας, από την είσοδο στην έξοδο, περνάει από τα διαδοχικά επίπεδα επεξεργασίας. Κάθε επίπεδο δέχεται ως είσοδο την έξοδο του προηγούμενου επιπέδου, τη μετασχηματίζει και την προωθεί στο επόμενο επίπεδο. Όταν τελικά γίνει μια πρόβλεψη για την έξοδο του δικτύου, η τιμή αυτή θα διαδοθεί στο δίκτυο με το οπισθοδρομικό πέρασμα και με σκοπό την ανανέωση των παραμέτρων του δικτύου για την επίτευξη καλύτερης πρόβλεψης.

2.4.2 Επίπεδα επεξεργασίας

Από αρχιτεκτονικής σκοπιάς τα συνελικτικά νευρωνικά δίκτυα δομούνται από ένα σύνολο σειριακών επιπέδων επεξεργασίας. Ένα σύνολο από ομαδοποιημένους νευρώνες συνθέτει κάθε επίπεδο και καθορίζει την λειτουργία του. Στη συνέχεια παρουσιάζεται ο τρόπος κατασκευής και η λειτουργία των διάφορων επιπέδων επεξεργασίας σε ένα συνελικτικό νευρωνικό δίκτυο.

2.4.2.1 Επίπεδο Εισόδου

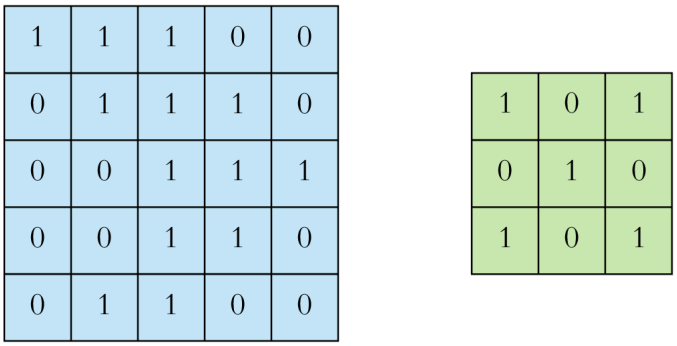
Το επίπεδο εισόδου (Input layer) είναι το πρώτο επίπεδο επεξεργασίας σε ένα τεχνητό νευρωνικό δίκτυο υπεύθυνο για την τροφοδότηση των δεδομένων εισόδου στο δίκτυο. Οι διαστάσεις των δεδομένων εισόδου καθορίζουν και τις διαστάσεις του επιπέδου εισόδου. Για παράδειγμα, σε ένα συνελικτικό νευρωνικό δίκτυο όπου τα δεδομένα εισόδου είναι ψηφιακές εικόνες με αναπαράσταση RGB διαστάσεων 1024x768 pixels, το επίπεδο εισόδου θα έχει μήκος 1024 νευρώνες, πλάτος 768 και ύψος 3, όσα είναι δηλαδή τα κανάλια σε μια RGB εικόνα (Red, Green, Blue).

2.4.2.2 Συνελικτικό Επίπεδο

Το συνελικτικό επίπεδο (Convolutional layer) είναι το κύριο δομικό στοιχείο ενός συνελικτικού νευρωνικού δικτύου αλλά και το πιο σύνθετο. Δέχεται σαν είσοδο την έξοδο του προηγούμενου επιπέδου, την μετασχηματίζει, εξάγει χαρακτηριστικά και δημιουργεί πίνακες (χάρτες) χαρακτηριστικών. Για να γίνει πιο εύκολα κατανοητός ο τρόπος λειτουργίας του επιπέδου αυτού στη συνέχεια γίνεται παρουσίαση των διαδικασιών και των εννοιών που το συνθέτουν:

- *Συνέλιξη (Convolution)*

  Η συνέλιξη είναι μια μαθηματική πράξη που συνδέει δύο διαφορετικά σύνολα πληροφορίας. Στην περίπτωση των συνελικτικών επιπέδων σε ένα CNN, η πράξη της συνέλιξης γίνεται μεταξύ των δεδομένων εισόδου και ενός φίλτρου (filter) ή αλλιώς πυρήνα (kernel). Με απλά λόγια, η πράξη της συνέλιξης είναι η ολίσθηση ενός παραθύρου, εν προκειμένω του πυρήνα, πάνω στα δεδομένα εισόδου. Το παράθυρο αυτό ολισθαίνοντας θα περάσει πάνω από όλα τα στοιχεία της εισόδου. Σε κάθε βήμα της ολίσθησης πραγματοποιείται ένα-προς-ένα πολλαπλασιασμός πινάκων μεταξύ του πυρήνα και των στοιχείων που καλύπτει το παράθυρο στον πίνακα των δεδομένων εισόδου. Στο σχήμα 2.10 δίνεται ένα παράδειγμα ενός πίνακα εισόδου και ενός πίνακα πυρήνα.

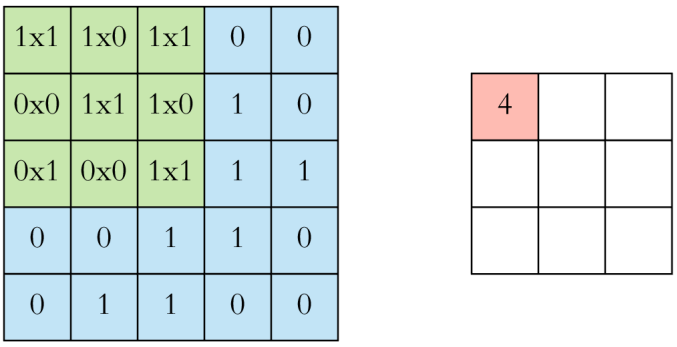


Πηγή: towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Σχήμα 2.10: Παράδειγμα εισόδου και πυρήνα

- *Πίνακες Χαρακτηριστικών (Feature Maps)*

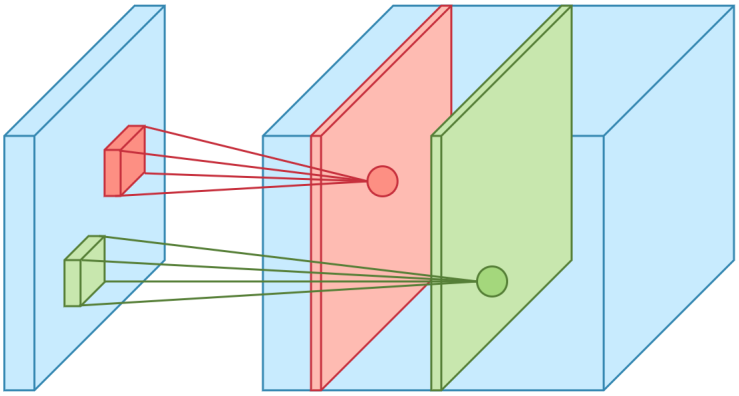
  Το αποτέλεσμα του πολλαπλασιασμού των πινάκων εισόδου και πυρήνα αποθηκεύεται σε έναν πίνακα χαρακτηριστικών, όπως φαίνεται στο σχήμα 2.11.



Πηγή: towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Σχήμα 2.11: Κατασκευή πίνακα χαρακτηριστικών

Ο αριθμός και οι διαστάσεις των φίλτρων ορίζεται στις παραμέτρους του επιπέδου. Για την ολίσθηση κάθε φίλτρου δημιουργείται ένας ξεχωριστός δισδιάστατος πίνακας χαρακτηριστικών. Κάθε τέτοιος πίνακας περιγράφει ξεχωριστά χαρακτηριστικά για την είσοδο. Αυτοί οι πίνακες στοιβάζονται (βλέπε σχήμα 2.12), έτσι στην έξοδο του επιπέδου θα σχηματιστεί ένας τρισδιάστατος πίνακας με ύψος όσος είναι ο αριθμός των πυρήνων που έχει καθοριστεί.



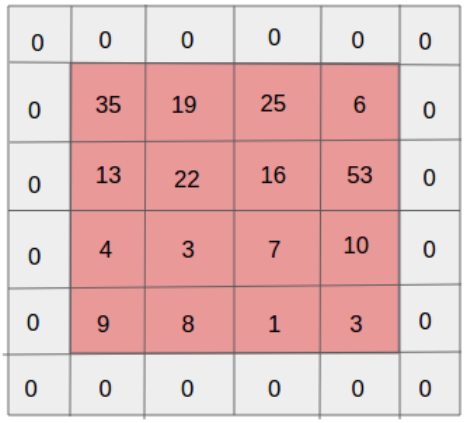
Πηγή: towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Σχήμα 2.12: Στοιβαγμένοι πίνακες χαρακτηριστικών

- *Υπερπαράμετροι (Hyperparameters)*

  Οι υπερπαράμετροι του επιπέδου συνέλιξης θα καθορίσουν το μέγεθος των πινάκων χαρακτηριστικών και την φύση της πληροφορίας που περιέχουν. Οι τέσσερις βασικές υπερπαράμετροι που πρέπει να καθοριστούν είναι το μέγεθος του πυρήνα (kernel size), ο αριθμός των πυρήνων (kernel count), το βήμα ολίσθησης (stride) και η επέκταση του περιθωρίου (padding). Όπως αναφέρθηκε ο αριθμός των πυρήνων θα καθορίσει

την τρίτη διάσταση στη έξοδο του συνελικτικού επιπέδου, όπως επίσης και το πλήθος των διαφορετικών χαρακτηριστικών. Το μέγεθος του πυρήνα καθορίζει τον όγκο της πληροφορίας που αναπαριστά κάθε στοιχείο στον πίνακα εξόδου. Τέλος, το βήμα ολίσθησης καθορίζει το μήκος και το πλάτος της εξόδου, αν το βήμα είναι πάνω από 1 αυτές οι διαστάσεις είναι μικρότερες από ότι είναι στην είσοδο του επιπέδου. Για να αποφύγουμε την μείωση των διαστάσεων μπορούμε να χρησιμοποιήσουμε την τελευταία υπερπαράμετρο για να επεκτείνουμε το περιθώριο της εισόδου όπως στο σχήμα 2.13.



Πηγή: towardsdatascience.com/introduction-to-convolutional-neural-networks-cnn-with-tensorflow-57e2f4837e18

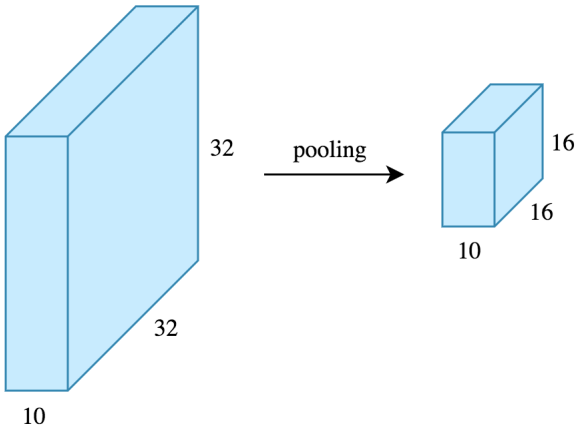
Σχήμα 2.13: Padding στα δεδομένα εισόδου

- *Μη-γραμμικότητα (Non-linearity)*

  Η εισαγωγή της μη-γραμμικότητας είναι αναγκαία ώστε ένα τεχνητό νευρωνικό δίκτυο να είναι ισχυρό. Σε ένα CNN η έννοια της μη-γραμμικότητας εφαρμόζεται στο συνελικτικό επίπεδο. Συγκεκριμένα, το αποτέλεσμα της συνέλιξης περνάει πρώτα από μία μη-γραμμική συνάρτηση ενεργοποίησης, πριν αποθηκευτεί. Στην πράξη, η συνάρτηση ενεργοποίησης που χρησιμοποιείται είναι η ReLU. Τελικά, αυτό που καταφέρνει η ReLU, σύμφωνα με τη σχέση 2.19, είναι να μηδενίσει όλες τις αρνητικές τιμές σε έναν πίνακα χαρακτηριστικών.

2.4.2.3 Συγκεντρωτικό Επίπεδο

Ένα συγκεντρωτικό επίπεδο (Pooling layer) συνηθίζεται να ακολουθεί κάθε συνελικτικό επίπεδο με σκοπό την μείωση των διαστάσεων του χώρου που αναπαριστά τα δεδομένα. Ουσιαστικά, όπως φαίνεται και από το σχήμα 2.14, το συγκεντρωτικό επίπεδο δειγματοληπτεί τους πίνακες χαρακτηριστικών, μειώνοντας το μήκος και το πλάτος τους, ενώ το ύψος τους παραμένει ανέπαφο. Αυτό, επιτρέπει να ελαττωθεί ο αριθμός των παραμέτρων του συστήματος, το οποίο επιταχύνει την διαδικασία εκπαίδευσης και αντιμετωπίζει το πρόβλημα της υπερπροσαρμογής.



Πηγή: towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Σχήμα 2.14: Εφαρμογή υποδειγματοληψίας

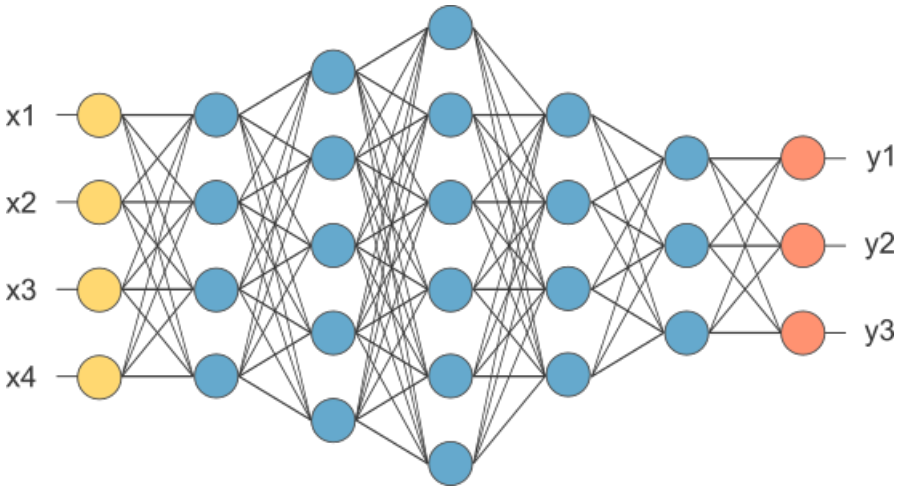
Ο τρόπος που γίνεται η υποδειγματοληψία στα επίπεδα αυτά θυμίζει πολύ την διαδικασία της συνέλιξης. Συγκεκριμένα, πάλι ένα παράθυρο ολισθαίνει πάνω από όλα τα δεδομένα εισόδου και υπολογίζει σε κάθε βήμα μία τιμή για τα στοιχεία που καλύπτει. Το μέγεθος του παραθύρου και το βήμα ολίσθησης καθορίζονται στις υπερπαραμέτρους του συγκεντρωτικού επιπέδου. Η τιμή που θα υπολογιστεί σε κάθε βήμα ολίσθησης καθορίζεται από την μέθοδο pooling που θα εφαρμοστεί. Μερικές από τις πιο συνήθεις μεθόδους είναι το max pooling που επιλέγει την μέγιστη τιμή στην περιοχή του παραθύρου, το average pooling που υπολογίζει την μέση τιμή των στοιχείων της περιοχής και το sum pooling που υπολογίζει το άθροισμα των στοιχείων της περιοχής.

2.4.2.4 Πλήρως Συνδεδεμένο Επίπεδο

Τα πλήρως συνδεδεμένα επίπεδα (Fully connected layers) εμφανίζονται σε ομάδες. Σε ένα CNN μια ομάδα από πληρώς συνδεδεμένα επίπεδα συναντάται στο τέλος του δικτύου μετά από μια αλληλουχία συνελικτικών και συγκεντρωτικών επιπέδων. Σε ένα πλήρως συνδεδεμένο επίπεδο κάθε νευρώνας του επιπέδου συνδέεται με όλους του νευρώνες του προηγούμενου επιπέδου, όπως στο σχήμα 2.15. Ένα τέτοιο επίπεδο αναμένει να πάρει ως είσοδο ένα διάνυσμα δεδομένων μίας διάστασης, τα συνελικτικά και συγκεντρωτικά επίπεδα όμως δίνουν διανύσματα τριών διαστάσεων. Για τον λόγο αυτό σε ένα CNN πριν από το πλήρως συνδεδεμένο επίπεδο εφαρμόζεται ένας μετασχηματισμός των δεδομένων που ονομάζεται ισοπέδωση (flattening) και μετατρέπει τα τρισδιάστατα διανύσματα σε μονοδιάστατα, χωρίς να χάνεται καμία πληροφορία.

Σκοπός μιας ομάδας πλήρως συνδεδεμένων επιπέδων είναι να αξιοποιήσουν τα χαρακτηριστικά που έχουν προέλθει από τα συνελικτικά και συγκεντρωτικά επίπεδα, ώστε να παράξουν την επιθυμητή έξοδο. Για παράδειγμα, σε ένα πρόβλημα ταξινόμησης εικόνων τα πλήρως συνδεδεμένα επίπεδα του δικτύου αναλαμβάνουν να επιλέξουν σε ποια κλάση ανήκει κάθε εικόνα της εισόδου. Τέλος, η χρησιμότητα των επιπέδων αυτών δεν περιορίζεται μόνο στην ταξινόμηση της εισόδου αλλά και στην ικανότητα εκμάθησης μη γραμμικών συνδυασμών των

χαρακτηριστικών της εισόδου.



Πηγή: quora.com/What-is-a-fully-connected-layer

Σχήμα 2.15: Πλήρως συνδεδεμένα επίπεδα

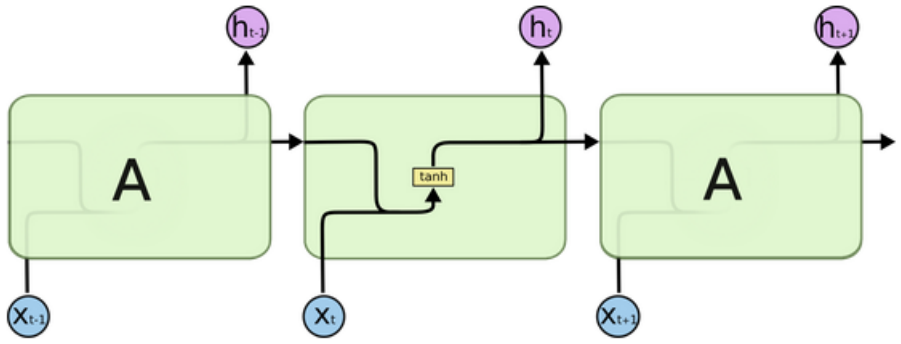
2.5 Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNN)

Τα επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Newral Networks) αποτελούν και αυτά μία κλάση των τεχνητών νευρωνικών δικτύων. Η ανάγκη για τη δημιουργία των (RNNs) προέκυψε από την αδυναμία των απλών νευρωνικών δικτύων να αξιοποιήσουν σωστά ακολουθιακά δεδομένα. Για παράδειγμα δεδομένα όπως τα καρέ σε ένα βίντεο ή οι λέξεις σε ένα κείμενο που έχουν ακολουθιακή φύση. Ο άνθρωπος για να επεξεργαστεί τέτοιου είδους δεδομένα χρησιμοποιεί την μνήμη του ώστε να διατηρήσει την χρονική πληροφορία που τα χαρακτηρίζει. Τα RNNs για να μπορέσουν να αξιοποιήσουν την ακολουθιακή πληροφορία των δεδομένων μιμούνται τον άνθρωπο και εισάγουν την έννοια την μνήμης.

Τα βασικά feed forward νευρωνικά δίκτυα έχουν και αυτά μνήμη, υπό την έννοια ότι θυμούνται όσα έχουν μάθει κατά τη διαδικασία εκπαίδευσης. Τα RNNs σε επέκταση αυτών, καταφέρνουν να θυμούνται εισόδους που χρησιμοποιήθηκαν ήδη για την παραγωγή εξόδων προηγουμένως. Έτσι, κάθε έξοδος του δικτύου δεν εξαρτάται μόνο από δεδομένα της εισόδου εκείνη την χρονική στιγμή και τους παραμέτρους του δικτύου αλλά και από ένα διάνυσμα 'κρυφής' κατάστασης (hidden state) που αναπαριστά όλες τις παρελθοντικές εισόδους στο δίκτυο. Επομένως, ένα RNN για δύο περιπτώσεις με την ίδια είσοδο μπορεί να αποδώσει διαφορετικές εξόδους όταν τα προηγούμενα δεδομένα εισόδου είναι διαφορετικά σε κάθε περίπτωση. Η πρώτη παρουσίαση των επαναλαμβανόμενα νευρωνικών δικτύων και της έννοιας της εσωτερικής κατάστασης έγινε το 1982 από τον J. Hopfield με τα Hopfield Networks [26].

2.5.1 Τρόπος λειτουργίας

Ο υπολογισμός της εξόδου σε ένα επαναλαμβανόμενα νευρωνικό δίκτυο δεν εξαρτάται μόνο από το διάνυσμα εισόδου αλλά και από το διάνυσμα εσωτερικής κατάστασης (hidden



Πηγή: http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Σχήμα 2.16: Δομή νευρώνα RNN

state), όπως αναφέρθηκε. Στην εικόνα 2.16 ο υπολογισμός της επόμενης κρυφής κατάστασης εξαρτάται από την είσοδο και την προηγούμενη κρυφή κατάσταση. Με τις σχέσεις 2.22 και 2.23 φαίνεται αναλυτικά ο τρόπος υπολογισμού της κρυφής κατάστασης h_t και της εξόδου y_t για είσοδο x_t. Στη βιβλιογραφία, οι διαστάσεις του διανύσματος h αναφέρονται ως hidden size και ουσιαστικά εκφράζει το μέγεθος της εσωτερικής μνήμη του κυττάρου (νευρώνα). Οι μεταβλητές W_h, U_h, W_y, b_h, b_y είναι διανύσματα παραμέτρων και οι διαστάσεις τους εξαρτώνται από το input size και το hidden size. Τέλος, τα σ_h, σ_y είναι συναρτήσεις ενεργοποίησης, στο παράδειγμα της εικόνας $\sigma_h = tanh$.

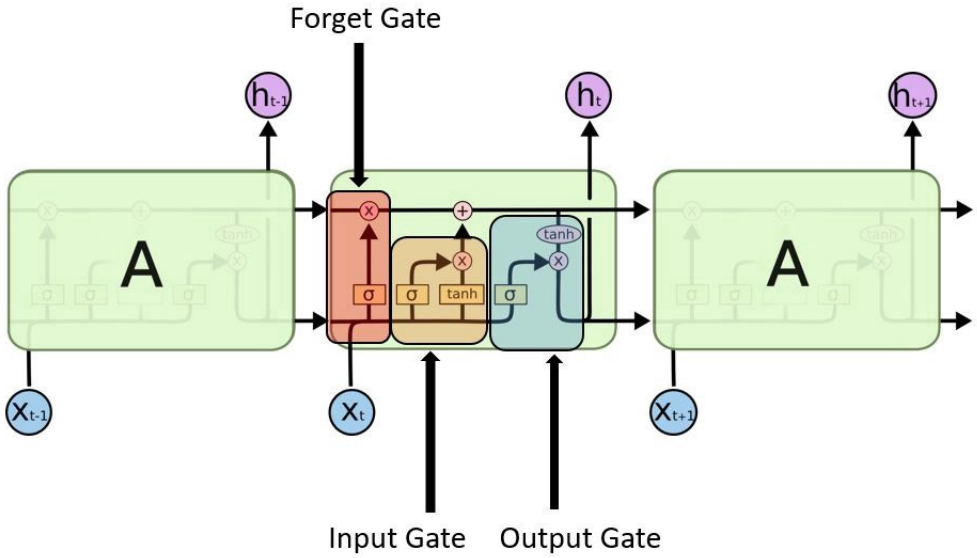
$$h_t = \sigma_h(W_h x_t + U_h h_{t-1} + b_h) \tag{2.22}$$

$$y_t = \sigma_y(W_y h_t + b_y) \tag{2.23}$$

2.5.2 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (LSTM)

Τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory) είναι μια τροποποιημένη εκδοχή των απλών RNNs, με την ικανότητα να ανακαλούν πιο εύκολα παρελθοντικά δεδομένα. Τα LSTMs μπορούν να διατηρούν πληροφορία για μεγάλα χρονικά διαστήματα αντιμετωπίζοντας το πρόβλημα της μακροπρόθεσμης εξάρτησης. Επιπλέον, τα LSTM αντιμετωπίζουν με αποτελεσματικότητα το πρόβλημα της εξαφανιζόμενης κλίσης που συναντάται στα απλά RNNs. Η αρχιτεκτονική τους προτάθηκε το 1997 από του Hochreiter και Schmidhuber [25]. Η βασική διαφορά που παρουσιάζουν τα δίκτυα LSTM είναι ότι εκτός από την κρυφή κατάσταση (hidden state) διαθέτουν και την κατάσταση κυττάρου (cell state).

Μια ακόμα ειδοποιός διαφορά είναι ότι πέρα από τις πύλες εισόδου και εξόδου, τα κύτταρα LSTM διαθέτουν και μια επιπλέον πύλη που ονομάζεται πύλη άγνοιας (forget gate). Στην εικόνα 2.17 δίνεται η εσωτερική δομή ενός κυττάρου (νευρώνα) σε ένα δίκτυο LSTM. Παρακάτω δίνονται οι μαθηματικές σχέσεις που περιγράφουν τις εσωτερικές διεργασίες στο κύτταρο:



Πηγή: http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Σχήμα 2.17: Δομή κυττάρου LSTM

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{2.24}$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{2.25}$$
$$i'_t = \sigma_{i'}(W_{i'} x_t + U_{i'} h_{t-1} + b_{i'}) \tag{2.26}$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{2.27}$$
$$c_t = f_t \circ c_{t-1} + i_t \circ i'_t \tag{2.28}$$
$$h_t = o_t \circ \sigma_h(c_t) \tag{2.29}$$

Συγκεκριμένα, x_t είναι το διάνυσμα εισόδου, h_t και c_t είναι τα διανύσματα κρυφής κατάστασης και κατάστασης κυττάρου αντίστοιχα. Ακόμα, i_t, f_t και o_t είναι τα διανύσματα ενεργοποίησης της πύλης εισόδου, της πύλης άγνοιας και της πύλης εξόδου. Ως i'_t συμβολίζεται ένα βοηθητικό διάνυσμα ενεργοποίησης της πύλης εισόδου, με μοναδική διαφορά στην συνάρτηση ενεργοποίησης. Οι παράμετροι του δικτύου συμβολίζονται ως W, U, b και οι συναρτήσεις ενεργοποίησης με σ. Τέλος, ο τελεστής $\circ$ εκφράζει τον πολλαπλασιασμό πινάκων ένα-προς-ένα.

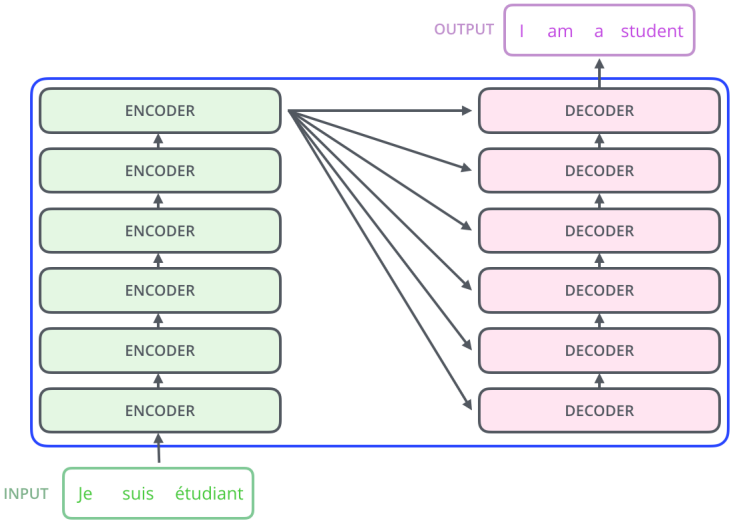
2.6 Transformers

Οι Tranformers [66] είναι μια αρχιτεκτονική τεχνητών νευρωνικών δικτύων που αποκτά μεγάλη δημοτικότητα τα τελευταία χρόνια. Κορυφαίες εταιρείες στον τομέα της τεχνητής

νοημοσύνης όπως η Google και η OpenAI χρησιμοποιούν Transformers για την ανάπτυξη συστημάτων επεξεργασίας φυσικής γλώσσας. Οι Transformers αναπτύχθηκαν για την αντιμετώπιση ζητημάτων που σχετίζονται με το πρόβλημα της μεταγωγής ακολουθίας (sequence transduction). Δηλαδή, οι διαδικασίες που παράγουν μια ακολουθία εξόδου από μια ακολουθία εισόδου, όπως για παράδειγμα η αναγνώριση φωνής και η μετάφραση κειμένου.

2.6.1 Τρόπος λειτουργίας

Η λειτουργία των Transformers βασίζεται σε αρχιτεκτονικές κωδικοποιητή - αποκωδικοποιητή (encoder - decoder) και σε μηχανισμούς προσοχής (Attention mechanisms). Συγκεκριμένα, ο κωδικοποιητής χρησιμοποιώντας μηχανισμούς προσοχής παράγει μια κρυφή αναπαράσταση της εισόδου και στη συνέχεια ο αποκωδικοποιητής λαμβάνει την κρυφή αναπαράσταση για να παράξει μια ακολουθία εξόδου, χρησιμοποιώντας και αυτός μηχανισμούς προσοχής. Το μοντέλο σε κάθε βήμα χρησιμοποιεί τις τιμές που έχουν υπολογιστεί στα προηγούμενα βήματα για να παράξει την νέα τιμή. Η εσωτερική οργάνωση ενός Transformer φαίνεται στο σχήμα 2.18. Στη συνέχεια γίνεται αναλυτικότερη περιγραφή της εσωτερικής δομής των Tranformers και των μηχανισμών που χρησιμοποιεί:



Πηγή: towardsdatascience.com/transformers-141e32e69591

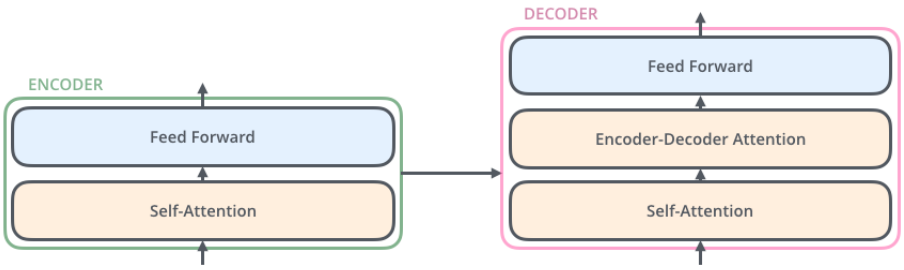
Σχήμα 2.18: Εσωτερική δομή Transformer

- *Στάδια κωδικοποίησης - αποκωδικοποίησης*

  Κωδικοποίηση (Encoding): Το στάδιο κωδικοποίησης συντίθεται από έξι πανομοιότυπους κωδικοποιητές. Όλοι οι κωδικοποιητές έχουν κοινή αρχιτεκτονική και αποτελούνται από δύο επίπεδα. Το πρώτο επίπεδο είναι ένας μηχανισμός προσοχής self-attention, ενώ το δεύτερο επίπεδο είναι ένα πλήρως συνδεδεμένο feed-forward δίκτυο.

  Αποκωδικοποίηση (Decoding): Το στάδιο της αποκωδικοποίησης μοιάζει αρκετά με το στάδιο της κωδικοποίησης και συντίθεται και αυτό από έξι πανομοιότυπους αποκωδικο-

ποιητές. Η εσωτερική δομή των αποκωδικοποιητών επεκτείνει την δομή των κωδικοποιητών με ένα επιπλέον επίπεδο που παρεμβάλλεται μεταξύ του μηχανισμού προσοχής self-attention και του feed-forward δικτύου. Το τρίτο επίπεδο είναι πάλι ένας μηχανισμός προσοχής που παρακολουθεί την έξοδο του σταδίου κωδικοκοποίησης. Όπως φαίνεται στο σχήμα 2.18, κάθε κωδικοποιητής όπως και κάθε αποκωδικοποιητής συνδέεται με τον προηγούμενό του και ακόμα η έξοδος του τελευταίου κωδικοποιητή μεταδίδεται και στους έξι κωδικοποιητές.



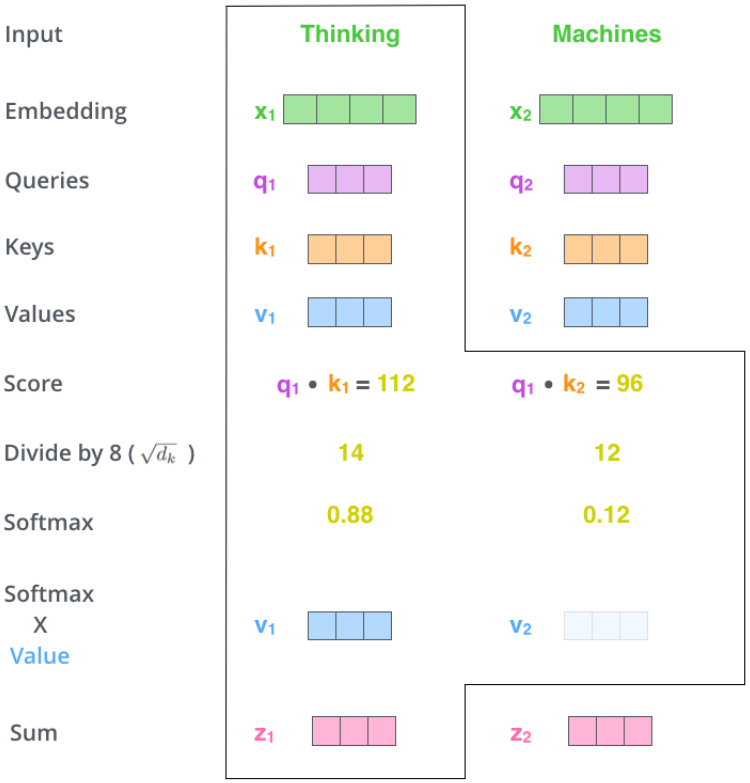
Πηγή: towardsdatascience.com/transformers-141e32e69591
Σχήμα 2.19: Εσωτερική δομή Encoder - Decoder

- *Μηχανισμός προσοχής*

  Ο μηχανισμός προσοχής (Attention mechanism) [35] είναι μια τεχνική που εφαρμόζεται στα τεχνητά νευρωνικά δίκτυα και ουσιαστικά υλοποιεί αυτό που δηλώνει το όνομα του, εστιάζει την προσοχή του σε ένα τμήμα του συνόλου που δέχεται σαν είσοδο κάθε φορά. Συγκεκριμένα, αυτό που κάνει ο μηχανισμός προσοχής είναι να εφαρμόζει μια συνάρτηση πάνω σε τρία διανύσματα q (query), k (key), v (value) για την παραγωγή του διανύσματος εξόδου. Τα τρία αυτά διανύσματα προέρχονται από τον πολλαπλασιασμό του διανύσματος που αναπαριστά κάθε λέξη στην είσοδο με έναν από τους τρεις πίνακες W^Q, W^K, W^V αντίστοιχα.

  Συγκεκριμένα, οι Transformers χρησιμοποιούν τον μηχανισμό προσοχής που ονομάζεται self-attention. Σε πρώτη φάση αυτό που κάνει ο μηχανισμός αυτός είναι να υπολογίσει τα διανύσματα q, k, v για κάθε λέξη της εισόδου. Στη συνέχεια, για κάθε λέξη το διάνυσμα q πολλαπλασιάζεται με κάθε διάνυσμα k όλων των λέξεων της εισόδου, κάθε τέτοιος πολλαπλασιασμός υπολογίζει μια τιμή s (score) που στη συνέχεια θα διαιρεθεί με την ρίζα του μήκους του διανύσματος λέξης και στην τιμή που θα προκύψει εφαρμόζεται η συνάρτηση softmax. Τέλος, κάθε διάνυσμα v πολλαπλασιάζεται με την έξοδο της softmax και τα αποτελέσματα αθροίζονται στο διάνυσμα z. Για καλύτερη διαίσθηση η παραπάνω διαδικασία φαίνεται στο σχήμα 2.20. Περιληπτικά, αυτό που κάνει ο μηχανισμός self-attention είναι να υπολογίζει, για μια ακολουθία λέξεων της εισόδου, την ΄σχετικότητα΄ μεταξύ των λέξεων, δηλαδή από νοηματικής σκοπιάς πόσο ΄κοντά΄ ή ΄μακριά΄ είναι οι λέξεις ανά δύο.

  Στην πράξη οι Transformers κάνουν παράλληλη χρήση πολλών μηχανισμών προσοχής



Πηγή: towardsdatascience.com/transformers-141e32e69591
Σχήμα 2.20: Παραγωγή διανυσμάτων εξόδου με τον μηχανισμό self-attention

self-attention με μια τεχνική που ονομάζεται multihead attention. Αυτή η τεχνική επιτρέπει στους Transformers να εστιάζουν την προσοχή τους σε πολλές λέξεις για κάθε είσοδο και όχι μόνο σε μια, κάτι που εμπλουτίζει σημαντικά την πληροφορία που έχουμε για κάθε ακολουθία εισόδου.

- *Κωδικοποίηση θέσης*

  Ένα ακόμα σημαντικό βήμα για το μοντέλο των Transformers είναι η εισαγωγή πληροφορίας για την θέση της κάθε λέξης στην ακολουθία της εισόδου. Για την επίτευξη αυτού του σκοπού εισάγεται στα διανύσματα των λέξεων μια επιπλέον πληροφορία σχετική με τη θέση κάθε λέξης που το μοντέλο μαθαίνει να αναγνωρίζει. Η διαίσθηση είναι ότι ένα διάνυσμα που εκφράζει θέση προστίθεται σε κάθε λέξη στο πρώτο επίπεδο κωδικοκοποίησης. Η κωδικοποίηση θέσης εισάγει σημαντική πληροφορία στο μοντέλο, που στη συνέχεια θα χρησιμοποιήσει για να αποδώσει σημασία στις λέξεις.

2.6.2 Bidirectional Encoder Representations from Trasformers (BERT)

Το BERT είναι ένα μοντέλο αναπαράστασης γλώσσας με εντυπωσιακή ακρίβεια σε πολλές εφαρμογές στο τομέα της Επεξεργασίας Φυσικής Γλώσσας, όπως το Question Answering, το Natural Language Inference και πολλά άλλα. Παρουσιάστηκε πρόσφατα από τους ερευνητές

της Google AI[2] στη δημοσίευσή τους με όνομα *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* [15] προκαλώντας αναστάτωση στην παγκόσμια κοινότητα της Μηχανικής Μάθησης. Σύμφωνα με την Google το 15% των αναζητήσεων που γίνονται καθημερινά στη μηχανή αναζήτησης δεν έχουν πραγματοποιηθεί ξανά. Επομένως, το πρόβλημα που εγείρεται δεν αφορά τις ερωτήσεις που μπορεί να κάνουν οι χρήστες αλλά με τους πόσους διαφορετικούς τρόπους μπορεί να γίνει μία αναζήτηση. Για την αντιμετώπιση αυτού η Google απομακρύνεται από την χρήση λέξεων-κλειδιά για την κατανόηση των αναζητήσεων και τη χρήση πιο σύνθετων μοντέλων όπως το BERT.

Το σημείο κλειδί στην τεχνική καινοτομία του BERT είναι η ικανότητα του να εφαρμόσει την αμφίδρομη εκπαίδευση των Transformers και έναν δημοφιλή μηχανισμό προσοχής στη μοντελοποίηση της φυσικής γλώσσας. Αυτό έρχεται σε αντίθεση με προηγούμενες προσπάθειες που εξέταζαν μία ακολουθία κειμένου από τα αριστερά προς τα δεξιά (και αντιστρόφως). Τα αποτελέσματα της δημοσίευσης [15] έδειξαν ότι ένα μοντέλο φυσικής γλώσσας που είναι αμφίδρομα εκπαιδευμένο μπορεί να αποκτήσει καλύτερη κατανόηση για το περιεχόμενο ενός κειμένου και την ροή του λόγου σε σύγκριση με ένα μοντέλο μονής κατεύθυνσης. Η αμφίδρομη εκπαίδευσή, που αδυνατούσαν τα προηγούμενα μοντέλα να εφαρμόσουν, επιτυγχάνεται στο BERT με την τεχνική που ονομάζεται Masked LM.

Ο τρόπος λειτουργίας του BERT στηρίζεται, όπως αναφέρθηκε, στη χρήση Transformers και σε ένα μηχανισμό προσοχής που αναγνωρίζει από τα συμφραζόμενα τις σχέσεις μεταξύ των λέξεων σε ένα κείμενο. Όπως περιγράφηκε στην προηγούμενη υποενότητα, η απλή μορφή των Transformers εσωτερικά δομείται από δύο στάδια επεξεργασίας, το στάδιο κωδικοποίησης που χρησιμεύει για την ανάγνωση της εισόδου και το στάδιο της αποκωδικοποίησης υπεύθυνο για την παραγωγή προβλέψεων. Από τη στιγμή που το BERT σαν σκοπό έχει την παραγωγή ενός μοντέλου φυσικής γλώσσας το στάδιο της αποκωδικοποίησης δεν είναι χρήσιμο. Ο κωδικοποιητής διαβάζει μια ολόκληρη ακολουθία λέξεων με τη μια και για το λόγο αυτό το μοντέλο χαρακτηρίζεται ως αμφίδρομο. Αυτό το χαρακτηριστικό επιτρέπει στο μοντέλο να χρησιμοποιεί τις περιβάλλουσες λέξεις μιας λέξης-στόχου για να μάθει το εννοιολογικό περιεχόμενο της .

Η εκπαίδευση (προεκπαίδευση) του BERT στην κατανόηση της φυσικής γλώσσας στηρίζεται σε δύο στρατηγικές εκπαίδευσης, τη στρατηγική Masked LM και τη στρατηγική Next Sentece Prediction. Στη συνέχεια, η γνώση που έχει δημιουργηθεί από αυτές τις στρατηγικές μπορεί να χρησιμοποιηθεί για την προσαρμογή του μοντέλου σε μια συγκεκριμένη εφαρμογή μέσα από μια διαδικασία επιμέρους εκπαίδευσης που ονομάζεται fine-tuning.

- Masked LM (MLM)

  Πριν την τροφοδότηση των ακολουθιών λέξεων στο BERT, 15% των λέξεων που συνθέτουν μια ακολουθία αντικαθίστανται από μία μάσκα [MASK] αποκρύπτοντας έτσι την αρχική τους τιμή. Στη συνέχεια το μοντέλο προσπαθεί να προβλέψει την αρχική τιμή των λέξεων με μάσκα βασιζόμενο στο περιεχόμενο των υπόλοιπων λέξεων στην ακολουθία που δεν έχουν μάσκα. Από τεχνικής πλευράς, η πρόβλεψη των όρων [MASK] προ-

[2] https://ai.google/

ϋποθέτει τη προσθήκη ενός επιπέδου ταξινόμησης (classification layer) στην έξοδο του κωδικοποιητή. Ακολούθως, η έξοδος του επιπέδου ταξινόμησης μεταχηματίζεται στις διαστάσεις του λεξιλογίου αφού πολλαπλασιαστεί με τον πίνακα embedding και τελικά εφαρμόζεται η συνάρτηση softmax στο διάνυσμα που έχει παραχθεί για τον υπολογισμό της πιθανότητας κάθε λέξη να αντιπροσωπεύει την αρχική τιμή του όρου [MASK].

- Next Sentence Prediction (NSP)

  Η επόμενη στρατηγική για την εκπαίδευση του μοντέλου είναι η πρόβλεψη επόμενης ακολουθίας. Ουσιαστικά, το μοντέλο τροφοδοτείται με ζευγάρια ακολουθιών και προσπαθεί να προβλέψει αν η δεύτερη ακολουθία βρίσκεται αμέσως μετά την πρώτη στο αρχικό κείμενο. Κατά την διαδικασία εκπαίδευσης το 50% των ζευγαριών αποτελείται από διαδοχικές ακολουθίες στο κείμενο, ενώ το υπόλοιπο 50% αποτελείται από ζευγάρια τυχαία επιλεγμένων ακολουθιών. Και σε αυτή τη στρατηγική ένα επίπεδο ταξινόμησης προστίθεται στην έξοδο του κωδικοποιητή, προσπαθώντας να ταξινομήσει τις εξόδους στις κλάσεις IsNext, NotNext εφαρμόζοντας πάλι τη συνάρτηση softmax.

- Fine-tuning

  Τέλος, μόλις ολοκληρωθεί η διαδικασία της προεκπαίδευσης (pretraining) μπορεί να προστεθεί στο τέλος του μοντέλου ένας αποκωδικοποιητής ή ένας ταξινομητής (ρηχό πλήρως συνδεδεμένο δίκτυο) για την προσαρμογή του σε κάποια συγκεκριμένη εφαρμογή. Κατά την διαδικασία του λεπτού χειρισμού (fine-tuning) μόνο τα επιπρόσθετα επίπεδα ανανεώνουν τις παραμέτρους τους και ίσως μερικά από τα τελευταία επίπεδα του δικτύου, ενώ οι περισσότερες από τις παραμέτρους του συστήματος δεν επηρεάζονται.

Τα δεδομένα εκπαίδευσης που συνοδεύουν κάθε εφαρμογή θα χρησιμοποιηθούν κατά την διαδικασία fine-tuning για την εκπαίδευση των τελευταίων επιπρόσθετων επιπέδων. Το BERT παρέχει προεκπαιδευμένα μοντέλα έτοιμα για χρήση, με μοναδική απαίτηση το fine-tuning. Στην παρούσα διπλωματική θα χρησιμοποιηθεί το BERT base uncased προεκπαιδευμένο σε ολόκληρη την αγγλική Wikipedia και σε 40.000 αγγλόφωνα βιβλία.

2.7 Μετρικές Αξιολόγησης

Η προετοιμασία των δεδομένων και εκπαίδευση είναι αντικείμενα υψίστης σημασίας για την ανάπτυξη ενός μοντέλου μηχανικής μάθησης, εξίσου σημαντική είναι όμως και η παρακολούθηση της επίδοσης του μοντέλου. Δηλαδή, πόσο καλά μπορεί να γενικεύει το μοντέλο για άγνωστα δεδομένα. Οι μετρικές αξιολόγησης εξυπηρετούν αυτόν το σκοπό, συγκεκριμένα μετρούν κάποιο μέγεθος του εκπαιδευμένου μοντέλου ως προς κάποιο χαρακτηριστικό. Χωρίς την χρήση μετρικών αξιολόγησης η βελτίωση της προβλεπτικής ικανότητας του μοντέλου ή η σύγκρισή του με άλλα μοντέλα δεν θα ήταν εφικτή. Ακόμα, αξίζει να αναφερθεί ότι η χρήση των μετρικών αξιολόγησης δεν είναι καθολική αλλά η φύση του κάθε προβλήματος ενθαρρύνει την χρήση διαφορετικών μετρικών αξιολόγησης. Στη συνέχεια παρουσιάζονται μερικές από

τις σημαντικότερες μετρικές αξιολόγησης. Πρώτα, όμως, γίνεται αναφορά στις κλάσεις που μπορούν να ανήκουν οι προβλέψεις του συστήματος:

- *True Positive (TP)*:

  Το σύνολο τις εξόδου για το οποίο η πρόβλεψη είναι σωστή και η προβλεπόμενη κλάση είναι θετική.

- *True Negative (TN)*:

  Το σύνολο τις εξόδου για το οποίο η πρόβλεψη είναι σωστή και η προβλεπόμενη κλάση είναι αρνητική.

- *False Positive (FP)*:

  Το σύνολο τις εξόδου για το οποίο η πρόβλεψη είναι λανθασμένη και η προβλεπόμενη κλάση είναι θετική.

- *False Negative (FN)*:

  Το σύνολο τις εξόδου για το οποίο η πρόβλεψη είναι λανθασμένη και η προβλεπόμενη κλάση είναι αρνητική.

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

Σχήμα 2.21: Κλάσεις προβλέψεων

1. *Accuracy*

   Η μετρική αξιολόγησης Accuracy ή ακρίβεια όπως φαίνεται και από τη σχέση 2.30, περιγράφει τον λόγο των σωστά ταξινομημένων δειγμάτων προς το σύνολο όλων των δειγμάτων.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.30}$$

   Η μετρική Accuracy είναι η πιο σημαντική μετρική και δίνει μια άμεση και απλή αξιολόγηση του μοντέλου. Η χρήση της συνίσταται για δεδομένα που είναι καλά ισορροπημένα.

2. *Precision*

Η μετρική αξιολόγησης Precision (βλέπε σχέση 2.31) που εκφράζει τον λόγο των σωστά ταξινομημένων θετικών προβλέψεων προς το σύνολο των προβλέψεων που έχουν ταξινομηθεί ως θετικές.

$$Precision = \frac{TP}{TP + FP} \tag{2.31}$$

Η μετρική Precision χρησιμοποιείται σε προβλήματα που η εγκυρότητα της πρόβλεψης είναι μεγάλης σημασίας.

3. *Recall*

   Η μετρική αξιολόγησης Recall (βλέπε σχέση 2.32) που εκφράζει το λόγο των σωστά ταξινομημένων θετικών προβλέψεων προς το σύνολο των θετικών προβλέψεων.

$$Recall = \frac{TP}{TP + FN} \tag{2.32}$$

   Η μετρική Recall χρησιμοποιείται σε προβλήματα που έχουν ως σκοπό την μεγιστοποίηση των θετικών προβλέψεων.

4. *F1 Score*

   Η μετρική αξιολόγησης F1 Score (βλέπε σχέση 2.33) που εκφράζει τον αρμονικό μέσο όρο των μετρικών Precision και Recall.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.33}$$

   Η χρήση της μετρικής F1 Score γίνεται όταν το πρόβλημα απαιτεί καλό Precision και Recall.

5. *Crossentropy*

   Η μετρική Crossentropy ή Log Loss λαμβάνει υπόψιν την αβεβαιότητα της πρόβλεψης βασιζόμενη στο πόσο διαφέρει από την πραγματική τιμή. Χρησιμοποιείται σε προβλήματα δυαδικής ταξινόμησης και υπολογίζεται από τον τύπο 2.34.

$$Crossentropy = -(y \log(p) + (1 - y) \log(1 - p)) \tag{2.34}$$

   Όπου p είναι η πιθανότητα η πρόβλεψη να είναι 1 και y είναι η πρόβλεψη του μοντέλου. Η μετρική αυτή χρησιμοποιείται όταν η έξοδος του μοντέλου είναι πιθανοτικές προβλέψεις.

6. *Categorical Crossentropy*

   Αυτή η μετρική αξιολόγησης είναι ίδια με την μετρική Crossentropy η μόνη διαφορά είναι ότι χρησιμοποιείται σε προβλήματα που οι κλάσεις ταξινόμησης είναι παραπάνω από δύο.

$$CategoricalCrossentropy = \frac{-1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} * \log(p_{ij}) \tag{2.35}$$

Όπου το y_{ij} είναι 1 αν το δείγμα i ανήκει στην κλάση j, αλλιώς είναι 0 και p_{ij} είναι η πιθανότητα το μοντέλο να προβλέψει ότι το δείγμα i ανήκει στην κλάση j.

2.8 Μουσική και Συναίσθημα

Η ικανότητα αντίληψης συναισθήματος από τη μουσική λέγεται ότι αναπτύσσεται από τα πρώτα, κιόλας, χρόνια της ζωής του ανθρώπου και εξελίσσεται με την πάροδο του χρόνου. Πέρα από την ηλικία πολύ σημαντικό ρόλο στην αντίληψη συναισθήματος από μουσική έχουν και οι πολιτισμικές επιρροές [63], αν και στις περισσότερες περιπτώσεις η ακρόαση και η πρόκληση συναισθημάτων έχουν παγκόσμια ερμηνεία. Εκτός από τα χαρακτηριστικά του ακροατή, τα συναισθήματα που σχετίζονται με τη μουσική εξαρτώνται και από άλλους παράγοντες. Πολλές μελέτες ερευνούν τη φύση των συναισθημάτων που δημιουργούνται με την ακρόαση μιας μουσικής σύνθεσης αλλά και με τα χαρακτηριστικά της μουσικής που επηρεάζουν το συναίσθημα.

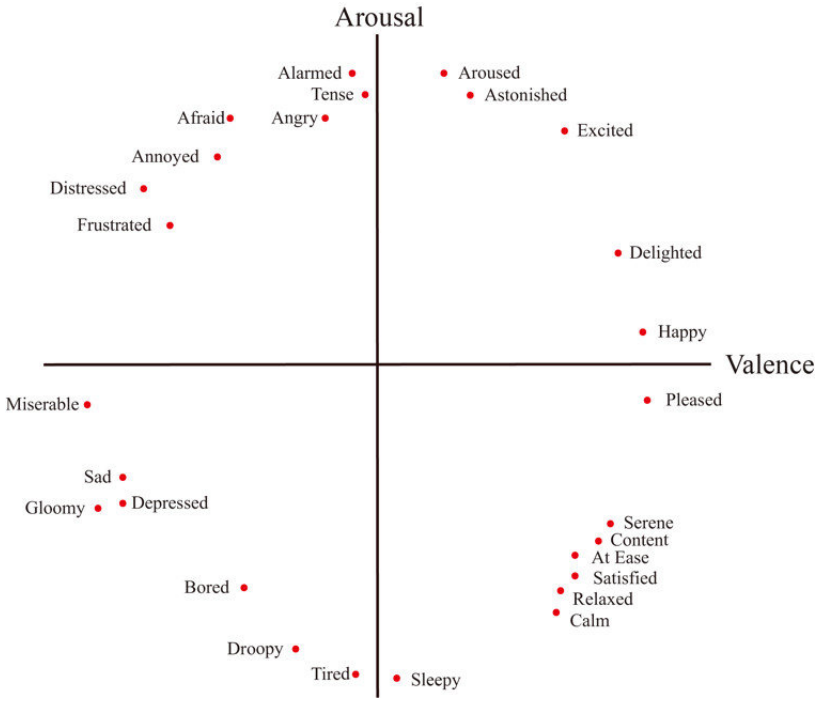
Τα δομικά χαρακτηριστικά της μουσικής που συνεισφέρουν στην αντίληψη της μουσικής έκφρασης χωρίζονται σε δύο κατηγορίες τα τμηματικά χαρακτηριστικά και τα υπερκαλυπτικά χαρακτηριστικά. Τα τμηματικά χαρακτηριστικά είναι οι ακουστικές δομές που συνθέτουν την μουσική, όπως η διάρκεια, το εύρος και το τονικό ύψος. Από την άλλη, τα υπερκαλυπτικά χαρακτηριστικά αφορούν τα δομικά στοιχεία ενός μουσικού κομματιού όπως ο ρυθμός, το τέμπο και η μελωδία. Ορισμένα από τα δομικά χαρακτηριστικά της μουσικής συνδέονται άμεσα με την έκφραση συγκεκριμένων συναισθημάτων στον ακροατή [21]. Παρακάτω παρουσιάζεται ο πίνακας συσχέτισης των μουσικών χαρακτηριστικών με τα αντίστοιχα συναισθήματα:

Δομικά Χαρακτηριστικά	**Ορισμός**	**Σχετικά Συναισθήματα**
Τέμπο	Η ταχύτητα ή ο ρυθμός ενός μουσικού κομματιού.	Γρήγορο τέμπο: ευτυχία, ενθουσιασμός, θυμός. Αργό τέμπο: λύπη, ηρεμία.
Κλίμακα	Ο τύπος της μουσικής κλίμακας.	Κλίμακα ματζόρε: ευτυχία, χαρά. Κλίμακα μινόρε: λύπη.
Ακουστικότητα	Η φυσική δύναμη και το εύρος του ήχου.	Ένταση, θυμός.
Μελωδία	Η διαδοχή μουσικών φθόγγων και ήχων, που ενοποιημένοι συνθέτουν ένα ηχητικό αποτέλεσμα.	Συμπληρωματικές αρμονίες: ευτυχία, χαλάρωση, ηρεμία. Μη-συμπληρωματικές αρμονίες: ενθουσιασμός, θυμός, δυσαρέσκεια.
Ρυθμός	Το επαναλαμβανόμενο μοτίβο στο τέμπο ενός μουσικού κομματιού.	Απαλός ρυθμός: ευτυχία, γαλήνη. Άτακτος ρυθμός: ανησυχία. Μεταβαλλόμενος ρυθμός: χαρά.

Πίνακας 2.1: Συσχέτιση μουσικών χαρακτηριστικών με συναισθήματα

Ένα ακόμη αντικείμενο που χρειάζεται μελέτη είναι η κατηγοριοποίηση των συναισθημάτων, υπό την έννοια της διάκρισης και του διαχωρισμού των συναισθημάτων. Η προσέγγιση στην κατηγοριοποίηση των συναισθημάτων μελετάται από δύο οπτικές. Η πρώτη οπτική αντιμετωπίζει τα συναισθήματα ως διακριτές μονάδες, ενώ η δεύτερη τα κατηγοριοποιεί σε ομάδες σε μια διαστασιακή βάση. Τα διαστασιακά μοντέλα προσπαθούν να ερμηνεύσουν τα ανθρώπινα συναισθήματα από το διάνυσμα της θέσης που τα αντιπροσωπεύει σε έναν χώρο δύο ή τριών διαστάσεων. Ακόμα προτείνουν ότι ένα διασυνδεδεμένο νευρικό σύστημα είναι υπεύθυνο για την απόδοση των συναισθηματικών καταστάσεων του ανθρώπου. Αρκετά διαστατικά μοντέλα έχουν αναπτυχθεί στον τομέα της συναισθηματικής ανάλυσης αλλά λίγα από αυτά διατηρούν την κυριαρχία τους και είναι αποδεκτά μέχρι και σήμερα.

Ένα ισχυρό διαστατικό μοντέλο, που θα χρησιμοποιηθεί στο πλαίσιο της διπλωματικής, είναι το Circumplex, ανεπτυγμένο το 1980 από τον James Russel [57]. Το μοντέλο αυτό υποστηρίζει ότι τα συναισθήματα είναι κατανεμημένα σε έναν δισδιάστατο κυκλικό χώρο, άξονες του οποίου συνθέτουν το σθένος (valence) και η διέγερση (arousal). Ο κάθετος άξονας αντιπροσωπεύει την διέγερση, ο οριζόντιος το σθένος, ενώ το κέντρο του κύκλου χαρακτηρίζει συναισθήματα ουδέτερου σθένους και μέτριας διέγερσης. Στο μοντέλο Circumplex τα συναισθήματα αναπαριστώνται ως ένα ζεύγος θετικών ή αρνητικών τιμών σθένους και διέγερσης. Ο Russel υποστηρίζει ότι οποιοδήποτε συναίσθημα πρωτόγνωρο ή όχι μπορεί να αναπαρασταθεί στον δισδιάστατο χώρο ως ένα ζεύγος τιμών.



Πηγή: mdpi.com/2079-9292/8/2/164/htm

Σχήμα 2.22: Μοντέλο Circumplex

Κεφάλαιο 3

Επεξεργασία Δεδομένων

Τα συστήματα μηχανικής μάθησης βασίζονται σε υπολογιστικές πράξεις μεταξύ αριθμών, οπότε όσο περίπλοκο να είναι ένα τέτοιο σύστημα, εσωτερικά εκτελούνται πράξεις πρόσθεσης και πολλαπλασιασμού. Πολλές φορές, τα δεδομένα που θα χρησιμοποιηθούν για την ανάπτυξη ενός συστήματος μηχανικής μάθησης είναι από την φύση τους αριθμητικά και είναι εύκολο να χρησιμοποιηθούν, όπως για παράδειγμα οι τιμές των μετοχών στο χρηματιστήριο ή οι τιμές καρδιολογικών εξετάσεων ενός συνόλου ανθρώπων. Σε αυτή την περίπτωση τα δεδομένα δεν χρειάζονται ιδιαίτερη επεξεργασία για την τροφοδότησή τους στο σύστημα. Άλλες φορές όμως, υπάρχουν περιπτώσεις που τα δεδομένα δεν μπορούν να χρησιμοποιηθούν άμεσα στη μορφή που συναντώνται στον πραγματικό κόσμο, παραδείγματος χάρη το σύνολο των λέξεων ενός κειμένου ή οι κυματομορφές ήχου μουσικών τραγουδιών. Η αναπαράσταση αυτού του είδους των δεδομένων στον υπολογιστή μπορεί πάλι να γίνεται με 0 και 1, άλλα η πληροφορία που περιέχουν δεν μπορεί να γίνει άμεσα αντιληπτή από τον υπολογιστή. Σε τέτοιου είδους δεδομένα η επεξεργασία για την μετατροπή τους και την ενσωμάτωση στα συστήματα μηχανικής μάθησης είναι απαραίτητη για την επίτευξη των επιθυμητών αποτελεσμάτων.

3.1 Επεξεργασία Κειμένου

Στα μοντέλα που επεξεργάζονται κείμενα οι λέξεις που συνθέτουν τα κείμενα ή οι χαρακτήρες που συνθέτουν τις λέξεις δεν παρουσιάζουν κάποια άμεση αριθμητική αντιστοιχία. Στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP) η πιο διαδεδομένη τεχνική στην αναπαράσταση του λεξιλογίου ενός κειμένου είναι η ενσωμάτωση λέξεων (Word Embeddings). Σύμφωνα με αυτή την τεχνική, η αναπαράσταση κάθε λέξης στο λεξιλόγιο ενός κειμένου, γίνεται με ένα διάνυσμα. Για τον υπολογισμό των διανυσμάτων αυτών υπάρχουν διάφορες μεθοδολογίες, στις επόμενες υποενότητες παρουσιάζονται οι πιο γνωστές.

3.1.1 Bag of Words

Η μέθοδος Bag of Words (BoW) [70] είναι η πιο απλή τεχνική μετατροπής κειμένου σε διάνυσμα. Αυτό που υλοποιεί ουσιαστικά η μέθοδος αυτή είναι ότι κάθε κείμενο αναπαριστάται

από το σύνολο (σακίδιο) των λέξεων του. Συγκεκριμένα, δοσμένου ενός λεξιλογίου $V = [v_1, v_2, ..., v_v]$ με πλήθος λέξεων $|V| = v$ και ενός κειμένου C με πλήθος λέξεων $|C| = c \leq v$, ο αλγόριθμος BoW κατασκευάζει ένα διάνυσμα $w = [w_1, w_2, ..., w_v]$ μήκους v, όπου κάθε στοιχείο w_i του διανύσματος εκφράζει πόσες φορές βρίσκεται στο κείμενο C η λέξη v_i. Έστω για παράδειγμα το λεξιλόγιο $V = [today, yesterday, my, dog, cat, a]$ και κείμενο το C = 'Yesterday my dog met a dog', τότε το παραγόμενο διάνυσμα θα είναι το $w = [0, 1, 1, 1, 2, 0, 1]$.

Για πολλά χρόνια ο αλγόριθμος Bag of Words χρησιμοποιόταν με μεγάλη επιτυχία σε πολλά προβλήματα που απασχολούσαν τον τομέα της επεξεργασίας της φυσικής γλώσσας. Η απλότητα της μεθόδου αυτής επιτρέπει την χρήση της μέχρι και σήμερα με μικρές παραλλαγές. Από την άλλη η μέθοδος Bow έχει και αρκετά μειονεκτήματα καθώς η μόνη πληροφορία που διατηρεί είναι το πλήθος εμφάνισης των λέξεων. Το πιο σημαντικό μειονέκτημα είναι ότι όλη η πληροφορία για την συντακτική δομή του κειμένου χάνεται. Ένα παράδειγμα αυτού του προβλήματος είναι για της προτάσεις 'Make love not war' και 'Make war not love' ο αλγόριθμος θα δώσει ακριβώς το ίδιο διάνυσμα για την αναπαράσταση τους, παρόλο που η σημασία τους είναι εντελώς διαφορετική.

3.1.2 Bag of n-grams

Η μέθοδος Bag of n-grams [32] αποτελεί μια επέκταση της μεθόδου BoW η οποία αντιμετωπίζει εν μέρη το πρόβλημα της συντακτικής δομής. Ένα n-gram είναι ουσιαστικά μια ακολουθία από n λέξεις που βρίσκονται σε σειρά. Έτσι, για ένα κείμενο ο αλγόριθμος θα εξάγει, πέρα από τις μεμονωμένες λέξεις, όλες τις ακολουθίες που αποτελούνται από n λέξεις και θα τις αναζητήσει ως ξεχωριστά αντικείμενα στο λεξιλόγιο. Στη βιβλιογραφία τα n-grams για $n = 1$ αναφέρονται ως unigrams (1-grams), για $n = 2$ ως bigrams (2-grams) και για $n = 3$ ως trigrams (3-grams). Έστω για παράδειγμα η πρόταση 'Yesterday my dog met a dog' τότε ως bigrams ο αλγόριθμος θα εξάγει τις ακολουθίες 'Yesterday my', 'my dog', 'dog met', 'met a', 'a dog', ενώ αντίστοιχα για trigrams θα εξάγει τις ακολουθίες 'Yesterday my dog', 'my dog met', 'dog met a', 'met a dog'. Αυτή η μέθοδος δίνει παραπάνω χρήσιμη πληροφορία από την μέθοδο BoW, επειδή για κάθε λέξη αξιοποιεί και συμφραζόμενα της. Το πρόβλημα με αυτή τη διαδικασία είναι ότι το λεξιλόγιο αυξάνεται εκθετικά και κατά συνέπεια τα διανύσματα αναπαράστασης θα είναι εκθετικά μεγαλύτερα και αραιά.

3.1.3 TF-IDF

Η μεθοδολογία TF-IDF (Term Frequency - Inverse Document Frequency) [54] είναι μια στατιστική διαδικασία που υπολογίζει την σχετικότητα μιας λέξης σε ένα κείμενο από ένα σύνολο κειμένων. Η τιμή που θα υπολογιστεί για μία λέξη σε ένα έγγραφο εξαρτάται από δύο μεγέθη, την συχνότητα του όρου (Term Frequency) και την αντίστροφη συχνότητα των εγγράφων. Το μέγεθος TF υπολογίζεται με αρκετούς τρόπου με τον πιο απλό να είναι η απαρίθμηση των φορών που η λέξη εμφανίζεται στο έγγραφο. Το μέγεθος IDF εκφράζει την συχνότητα της λέξης στο σύνολο των εγγράφων και υπολογίζεται ως τον λογάριθμο του λόγου του συνολικού αριθμού των εγγράφων προς τον αριθμό των εγγράφων στα οποία εμφανίζεται

η λέξη.

Η τιμή TF-IDF για μία λέξη w σε ένα κείμενο d από ένα σύνολο από κείμενα D υπολογίζεται από τον τύπο 3.1, τιμές κοντά στο 0 σημαίνουν ότι η σημαντικότητα της λέξης στο κείμενο είναι μικρή, ενώ τιμές κοντά στο 1 δηλώνουν μεγάλη σημαντικότητα.

$$TF - IDF(w, d, D) = TF(w, d) \cdot IDF(w, d, D) \tag{3.1}$$

Όπου,

$$TF(w, d) = freq(w, d) \tag{3.2}$$

$$IDF(w, d, D) = \log(\frac{N}{count(d \in D : w \in d)}) \tag{3.3}$$

Κύριο προτέρημα που παρουσιάζει η μέθοδος TD-IDF είναι ότι εξετάζει τη σημαντικότητα των λέξεων απέναντι σε ένα μεγάλο σύνολο από έγγραφα. Για παράδειγμα λέξεις που εμφανίζονται συχνά σε ένα κείμενο όπως οι *'is', 'what', 'the'* είναι μικρής σημασία για την σημασιολογία ενός εγγράφου και ο αλγόριθμος καταφέρνει να το αναγνωρίζει. Αντίθετα, η εμφάνιση της λέξης *'bug'* θα είναι συχνή σε ένα έγγραφο που αναλύει έντομα ή ασχολείται με την αποσφαλμάτωση κώδικα και ο αλγόριθμος είναι σε θέση να της αποδώσει την απαραίτητη σημασία. Ωστόσο, ένα πρόβλημα με την μεθοδολογία αυτή είναι ότι αγνοεί τη συντακτική δομή ενός κειμένου, όπως και η μεθοδολογία BoW. Ένα ακόμα πρόβλημα που παρουσιάζει είναι ότι για να είναι αποδοτική αυτή η μεθοδολογία χρειάζεται μεγάλο αριθμό εγγράφων.

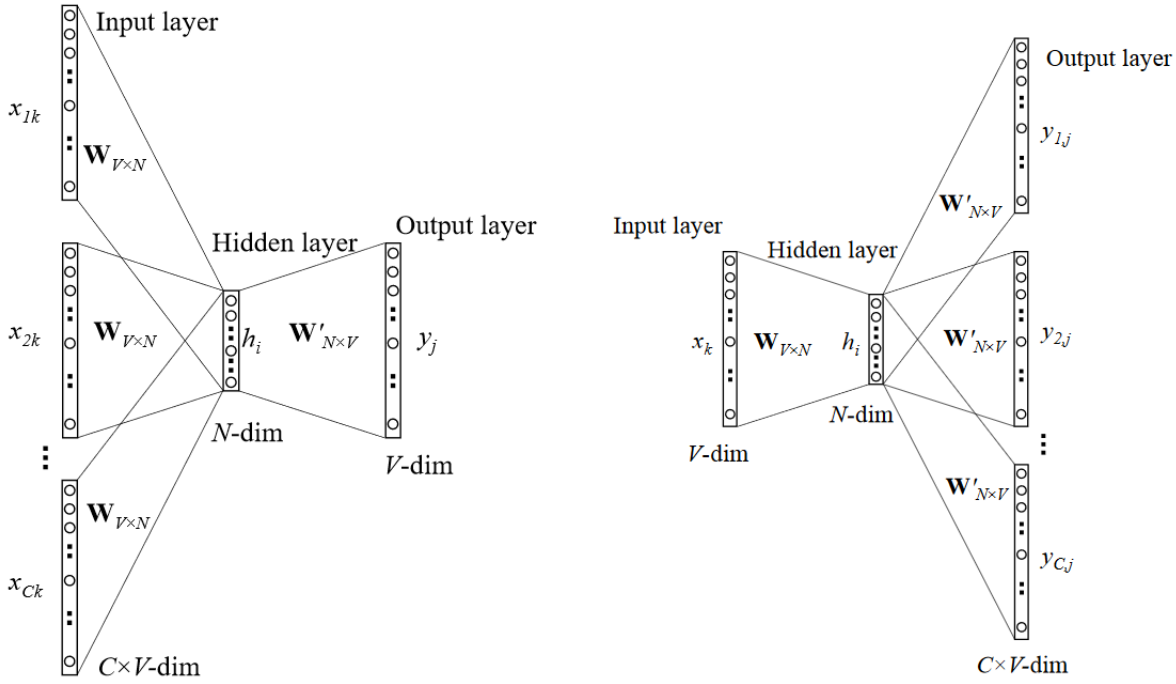
3.1.4 Word2Vec

Ο όρος Word2Vec [42] [22] αναφέρεται σε μια ομάδα μοντέλων που χρησιμοποιούνται για τη διανυσματική αναπαράσταση λέξεων που συνθέτουν ένα κείμενο. Ένα μοντέλο Word2Vec δέχεται σαν είσοδο έναν μεγάλο όγκο εγγράφων (corpus) και παράγει έναν διανυσματικό χώρο (vector space) $\mathbb{R}^N$, τυπικά το N ανήκει στην τάξη των εκατοντάδων. Κάθε μοναδική λέξη που ανήκει στο corpus αναπαριστάται στο διανυσματικό χώρο από ένα διάνυσμα, η διανυσματική αναπαράσταση γίνεται με τέτοιο τρόπο ώστε λέξεις με παρόμοια συμφραζόμενα να βρίσκονται κοντά. Με αυτόν τον τρόπο επιτυγχάνεται η συσχέτιση των λέξεων που έχουν παρόμοια εννοιολογική σημασία.

Τα διανύσματα των λέξεων στην τεχνική Word2Vec είναι κωδικοποιημένα στη μορφή One-hot, που σημαίνει ότι κάθε διάνυσμα έχει μήκος V, όσο και το μέγεθος του λεξιλογίου, αποτελείται από μηδενικά σε όλα τα στοιχεία του εκτός από το στοιχείο που αντιπροσωπεύει την συγκεκριμένη λέξη στο λεξιλόγια, στη θέση του οποίου έχει 1. Στο κρυφό επίπεδο, αθροίζονται τα γινόμενα των εισόδων με τους πίνακες παραμέτρων και στο επίπεδο της εξόδου, εφαρμόζεται η συνάρτηση softmax που προσπαθεί να προβλέψει τις σωστές θέσεις των άσσων, δηλαδή τις σωστές λέξεις, στα διανύσματα One-hot της εξόδου.

Τα μοντέλα Word2Vec είναι στην ουσία τεχνητά νευρωνικά δίκτυα που αποτελούνται από δύο επίπεδα νευρώνων. Για την διανυσματική αναπαράσταση των λέξεων τα μοντέλα αυτά καλούνται να διαλέξουν ανάμεσα σε δύο αρχιτεκτονικές μοντελοποιήσεις, τα CBoW και τα Skip-gram. Το μοντέλο CBoW (Continuou Bag of Words) δέχεται σαν είσοδο τις γειτονικές

λέξεις μιας λέξης-στόχου και προσπαθεί να προβλέψει στην έξοδο την λέξη-στόχο. Σε αντιδιαστολή, το μοντέλο Skip-gram δέχεται σαν είσοδο μια λέξη και προβλέπτει στην έξοδο τις γειτονικές της λέξεις. Όπως φαίνεται από το σχήμα 3.1 η δομή του δικτύου Skip-gram είναι στην στην ουσία η δομή του δικτύου CBoW ανεστραμμένη.



Πηγή: devopedia.org/word2vec

Σχήμα 3.1: Δομή δικτύων CBoW (αριστερά) και Skip-gram (δεξιά)

Και τα δύο μοντέλα έχουν προτερήματα και μειονεκτήματα. Το μοντέλο Skip-gram δουλεύει καλύτερα για μικρό όγκο δεδομένων και καταφέρνει να αναπαραστήσει σπάνιες λέξεις καλύτερα. Από την άλλη το μοντέλο CBoW αναπαριστά καλύτερα πιο συχνές λέξεις και είναι πιο γρήγορο.

3.1.5 GloVe

Η μέθοδος GloVe (Global Vectors) [51], [38] είναι και αυτή, όπως η Word2Vec (βλέπε ενότητα 3.1.4), μια διαδικασία που ακολουθείται για την διανυσματική αναπαράσταση λέξεων. Το πλεονέκτημα της μεθόδου αυτής είναι ότι δεν χρησιμοποιεί μόνο τοπικά (local) χαρακτηριστικά, όπως τα συμφραζόμενα των λέξεων, αλλά ενσωματώνει και κάποια καθολικά (global) στατιστικά για να εξάγει τα διανύσματα. Το μοντέλο GloVe, εξαιτίας της καθολικότητας, έχει την δύναμη να αναγνωρίσει λέξεις που αν και εμφανίζονται στα συμφραζόμενα μιας άλλης δεν προσδίδουν κανένα εννοιολογικό περιεχόμενο. Για παράδειγμα, στην πρόταση ΄The dog chased the cat΄ η λέξη 'the' προηγείται των λέξεων 'dog' και 'cat' αλλά δεν τους προσδίδει καμία πληροφορία κάτι που είναι απαραίτητο να αναγνωριστεί και από ένα σύστημα NLP.

Όπως και τα μοντέλα Word2Vec, έτσι και ένα μοντέλο GloVe δημιουργεί έναν διανυσματικό χώρο αξιοποιώντας ένα σύνολο από έγγραφα. Το πετυχαίνει αυτό υπολογίζοντας και αξιοποιώντας την συνεμφάνιση (co-occerrance) λέξεων σε έναν όγκο εγγράφων (corpus). Το βασικό δομικό στοιχείο της μεθόδου είναι ο πίνακας συνεμφάνισης C διαστάσεων VxV για λεξιλόγιο μεγέθους V, όπου κάθε στοιχείο C_{ij} του πίνακα εκφράζει τις πιθανότητες συνεμφάνισης της λέξης w_j στα συμφραζόμενα της λέξης w_i.

Στη συνέχεια αξιοποιεί τον πίνακα συνεμφάνισης και υπολογίζει λόγους από τις τιμές συνεμφάνισης δύο λέξεων για να ανακαλύψει την εννοιολογική τους σχέση. Έστω ότι το $P(k|w)$ εκφράζει τη συνεμφάνιση, δηλαδή την πιθανότητα η λέξη k να ανήκει στα συμφραζόμενα της λέξης w. Για παράδειγμα αν έχουμε δύο λέξεις $k_1 = ice$ και $k_2 = steam$, τότε ο λόγος $P(solid|ice)/P(solid|steam)$ θα ήταν μεγάλος γιατί ο όρος $P(solid|ice)$ θα είχε τιμή κοντά στο 1 ενώ ο όρος $P(solid|steam)$ τιμή κοντά στο 0. Αντίστοιχα ο λόγος $P(gas|ice)/P(gas|steam)$ θα είχε πολύ μικρή τιμή, ενώ ο λόγος $P(water|ice)/P(water|steam)$ θα είχε τιμή κοντά στο 1.

Η διαδικασία που η μέθοδος GloVe ακολουθεί για να προβλέψει συμφραζόμενες λέξεις είναι με την χρήση των λόγων πιθανοτικής συνεμφάνισης με την εκτέλεση λογιστικής παλινδρόμησης. Εν κατακλείδι, το GloVe δημιουργεί μια συνάρτηση f που δέχεται ως όρισμα μια λέξη και παράγει ένα διάνυσμα αυτής της λέξης. Το χαρακτηριστικό αυτών των διανυσμάτων είναι ότι η πρόσθεση και η αφαίρεση τους δεν έχει αντιστοιχία μόνο στον διανυσματικό χώρο αλλά και στον εννοιολογικό, ένα τέτοιο παράδειγμα είναι η πρόταση: $f('queen') = f('king') - f('man') + f('woman')$.

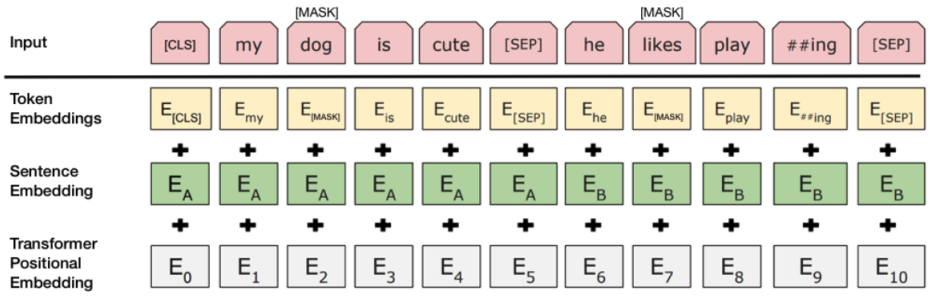
3.1.6 BERT word embeddings

Το μοντέλο BERT (Bidirectional Encoder Representations from Transformers) [15] θα χρησιμοποιηθεί για την εφαρμογή της μεταφορικής μάθησης (TL) στην παρούσα διπλωματική. Με σκοπό την προσέγγιση ενός προβλήματος NLP, ένα μοντέλο BERT απαιτεί συγκεκριμένη διανυσματική αναπαράσταση των λέξεων και των προτάσεων που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου, η αναπαράσταση αυτή ονομάζεται BERT word embeddings.

Τα διανύσματα που παράγονται με την μεθοδολογία BERT υπερτερούν έναντι των προηγούμενων μεθόδων επειδή η αναπαράσταση κάθε λέξης δεν γίνεται στατικά, όπως για παράδειγμα με τη μέθοδο Word2Vec, όπου η αναπαράσταση είναι ανεξάρτητη από τα συμφραζόμενα, αλλά κάθε αναπαράσταση ενημερώνεται δυναμικά από τις περιβάλλουσες λέξεις. Για παράδειγμα, στις δυο προτάσεις 'The man was accused of robbing a bank.' και 'The man went fishing by the river bank.' η μέθοδος Word2Vec θα δώσει το ίδιο διάνυσμα για την αναπαράσταση της λέξης 'bank' και στις δυο προτάσεις. Αντίθετα, το BERT θα συνυπολογίσει και τα συμφραζόμενα της λέξης 'bank' και θα δώσει διαφορετικές αναπαραστάσεις για την κάθε περίπτωση.

Εκτός από τις προφανείς διαφορές που αναγνωρίζει η μεθοδολογία αυτή, όπως η πολυσημία, αναγνωρίζει και άλλες μορφές πληροφορίας που θα οδηγήσουν σε πιο ακριβείς αναπαραστάσεις χαρακτηριστικών και κατά συνέπεια σε καλύτερη επίδοση συστήματος. Η ικανότητα του BERT βασίζεται στην αρχιτεκτονική του, για να μπορέσει όμως να είναι αποδοτική η

εκπαίδευση, το μοντέλο αναμένει μια συγκεκριμένη αναπαράσταση της εισόδου, όπως προαναφέρθηκε. Η μορφή της εισόδου που αναμένει το μοντέλο φαίνεται στο σχήμα 3.2, ουσιαστικά η αναπαράσταση μίας πρότασης γίνεται από τρία διανύσματα. Πριν γίνει η ανάλυση των διανυσμάτων, πρέπει να πρώτα να γίνει κατανοητή η διαδικασία που μια πρόταση μετατρέπεται σε μια ακολουθία από μονάδες (tokenization) στο επίπεδο εισόδου. Το BERT διαθέτει δικό του tokenizer το οποίο, δεδομένης μια πρότασης στην είσοδο του, θα κατασκευάσει στη έξοδο μια ακολουθία από τις λέξεις (tokens) που θα βρει στο λεξιλόγιο. Το BERT χρησιμοποιεί την τεχνική WordPiece [67] για tokenization, σύμφωνα με αυτή η λέξη 'playing' θα χωριστεί στα tokens 'play' και '##ing'. Αυτός ο διαχωρισμός σκοπεύει στην ανάλυση μιας λέξης στα συνθετικά της και γίνεται για την κάλυψη του φάσματος λέξεων εκτός του λεξιλογίου (Out-Of-Vocabulary OVV). Στη συνέχεια θα εισαχθούν στην ακολουθία και τα ειδικά tokens, το token [CLS] θα τοποθετηθεί στην αρχή κάθε ακολουθίας και το token [SEP] θα χρησιμοποιηθεί στην αρχή κάθε πρότασης, στην περίπτωση που μια ακολουθία εισόδου αποτελείται πάνω από μια πρόταση. Τώρα που η είσοδος έχει μετασχηματιστεί στη κατάλληλη μορφή, κάθε token στην ακολουθία της εισόδου θα αναλυθεί σε τρία τμήματα, όπως φαίνεται στην εικόνα 3.2. Το πρώτο τμήμα, τα *token embeddings*, είναι τα IDs του κάθε token στο λεξιλόγιο. Τα *sentence embeddings* είναι μια δυαδική τιμή για των διαχωρισμό των προτάσεων. Τέλος, τα *tranformer positional mebeddings* υποδεικνύουν τη θέση κάθε λέξης στην ακολουθία.

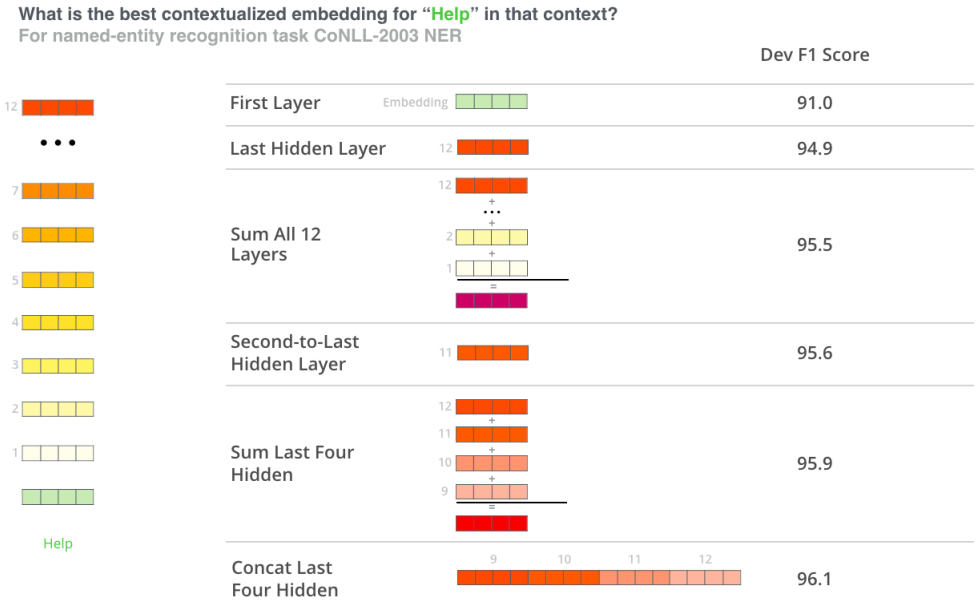


Πηγή: towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b

Σχήμα 3.2: Δομή της εισόδου στο μοντέλο BERT

Η διαδικασία που έχει περιγραφεί μέχρι τώρα δεν έχει ως αποτέλεσμα τη διανυσματική αναπαράσταση των λέξεων αλλά τη μετατροπή της εισόδου σε μορφή που μπορεί να γίνει κατανοητή από το σύστημα. Για να γίνει η διανυσματική αναπαράσταση θα βοηθήσει η αρχιτεκτονική του μοντέλου. Από αρχιτεκτονικής σκοπιάς, ένα μοντέλο BERT αποτελείται από δώδεκα στοιβαγμένα επίπεδα κωδικοποιήτων, όπως αναφέρθηκαν στην ενότητα 2.6. Κάθε επίπεδο για κάθε λέξη παράγει ένα διάνυσμα (hidden state), αυτά τα διανύσματα και ο συνδυασμός τους μπορούν να χρησιμοποιηθούν ως τα διανύσματα που θα αναπαραστήσουν κάθε λέξη. Κάθε επίπεδο του BERT κωδικοποιεί διαφορετικά είδη πληροφορίας. Στην εικόνα 3.3 φαίνονται διάφοροι συνδυασμοί αυτών των διανυσμάτων και τα f1 scores τους. Η επιλογή του συνδυασμού εξαρτάται κάθε φορά από το πρόβλημα, εμπειρικά οι εμπνευστές του BERT προτείνουν ότι το άθροισμα των διανυσμάτων από τα τέσσερα τελευταία επίπεδα είναι μια καλή

επιλογή για υψηλή απόδοση.

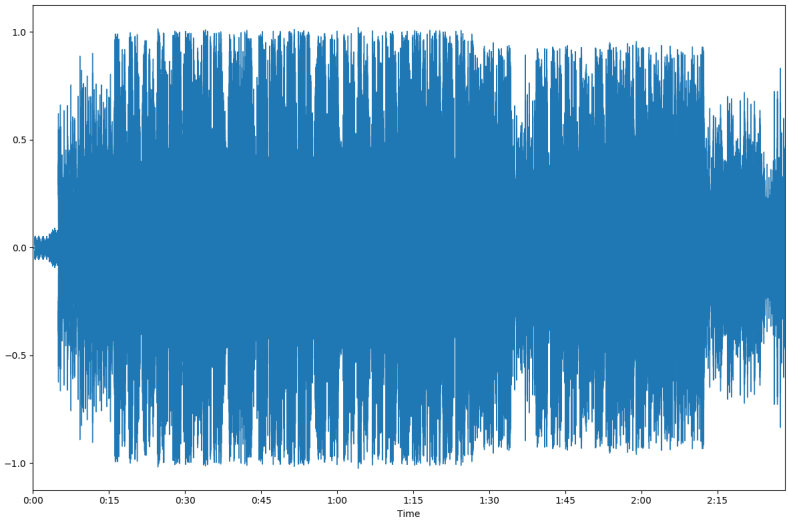


Πηγή: towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b

Σχήμα 3.3: Συνδυασμός διανυσμάτων εξόδου από τα επίπεδα του BERT

3.2 Επεξεργασία Σήματος Ήχου

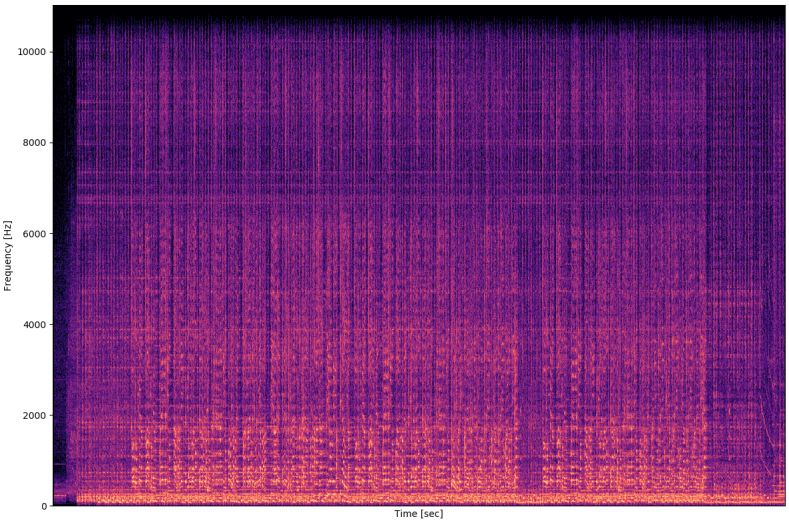
Η επεξεργασία σήματος ήχου είναι ένα υποπεδίο της επεξεργασία σήματος που ασχολείται με την ηλεκτρονική διαχείριση ηχητικών σημάτων. Τα ηχητικά σήματα είναι η ηλεκτρονική αναπαράσταση των μεταβολών πίεσης που διαδίδονται στον αέρα, δηλαδή των ηχητικών κυμάτων. Η κυματομορφή ενός ηχητικού σήματος όμως, δεν περιέχει πληροφορία που να μπορεί να αξιοποιηθεί άμεσα από ένα σύστημα μηχανικής μάθησης, όπως στην προκείμενη διπλωματική. Για να μπορέσει ένα τέτοιο σύστημα να αξιοποιήσει ηχητικά σήματα είναι απαραίτητο να προηγηθεί ανάλυση των σημάτων. Η ανάλυση ηχητικών σημάτων αναφέρεται στην εξαγωγή γνώσης που σχετίζεται με το περιεχόμενο και τη φύση των σημάτων αυτών. Για την ανάλυση ενός σήματος είναι αναγκαία η χρήση κάποιων μεθόδων από το πεδίο της ψηφιακής επεξεργασίας σήματος που ως στόχο έχουν την εξαγωγή εκείνων των χαρακτηριστικών (features) που θα εμπεριέχουν την επιθυμητή πληροφορία. Στη συνέχεια παρουσιάζονται ορισμένες από τις μεθόδους που φάνηκαν χρήσιμες για την εξαγωγή χαρακτηριστικών από ένα μουσικό κομμάτι σχετικές με το συναίσθημα που αυτό δίνει.



Σχήμα 3.4: Ηχητικό σήμα μουσικού κομματιού

3.2.1 Short-Time Fourier Transform

Ο μετασχηματισμός Short-Time Fourier (STFT) χρησιμοποιείται για να εκφράσει την ημιτονοειδή συχνότητα και το φασικό περιεχόμενο σε ένα τμήμα του σήματος καθώς αυτό μεταβάλλεται στο χρόνο. Στην πράξη, για να υπολογιστεί ο STFT το σήμα διαιρείται σε τμήματα ίσου μήκους στα οποία εφαρμόζεται ο μετασχηματισμός Fourier, αναδεικνύοντας έτσι το Fourier φάσμα. Η διαδικασία απεικόνισης των μεταβλητών φασμάτων σαν συνάρτηση του χρόνου είναι γνωστή ως φασματογράφημα (spectogram).



Σχήμα 3.5: φασματογράφημα ηχητικού σήματος

Στην περίπτωση του συνεχούς χρόνου, η συνάρτηση που θέλουμε να μετασχηματίσουμε, έστω $x(t)$, πολλαπλασιάζεται με μια συνάρτηση παράθυρο (window function) η οποία είναι μη μηδενική για ένα μικρό χρονικό διάστημα. Για την συνάρτηση παραθύρου χρησιμοποιείται

συνήθως το παράθυρο τύπου Hann ή Γκαουσιανό παράθυρο και συμβολίζεται με $w(t)$. Ο μετασχηματισμός Fourier, για τον STFT, υπολογίζεται καθώς το παράθυρο ολισθαίνει πάνω στο σήμα και μαθηματικά υπολογίζεται από τη σχέση 3.4.

$$STFT\{x(t)\} \equiv X(\tau, w) = \int_{-\infty}^{+\infty} x(t)w(t-\tau)e^{-iwt}dt \tag{3.4}$$

3.2.2 Χαρακτηριστικά βασισμένα σε Mel φίλτρα

Τα φίλτρα τύπου Mel μιμούνται την απόκριση του ανθρώπινου ακουστικού συστήματος με καλύτερη απόδοση από της γραμμικές μπάντες συχνοτήτων. Ουσιαστικά, ο όρος Mel αναφέρεται σε μια κλίμακα τονικού ύψους (pitch) δημιουργημένη από πειράματα σε ακροατές με σκοπό την αναγνώριση τονικών μεταβολών από το ανθρώπινο αυτί. Η απόδοση της ονομασίας οφείλεται στους S. Stevens, J. Volkmann, E. Newman [58] και προέρχεται από τη λέξη melody (μελωδία). Για την εξαγωγή χαρακτηριστικών στην παρούσα διπλωματική χρησιμοποιήθηκαν τρεις παραλλαγές μετασχηματισμών Mel:

- *Mel Spectogram*

  Για τον υπολογισμό του Mel φασματογραφήματος πρώτα υπολογίζεται το απλό φασματογράφημα από το ηχητικό σήμα με τον STFT μετασχηματισμό και στη συνέχεια γίνεται αντιστοίχιση των συχνοτήτων στην κλίμακα Mel σύμφωνα με τον τύπο 3.5:

$$mel(f) = \frac{1000}{\log 2} log(1 + \frac{f}{1000}) \tag{3.5}$$

- *Log-Mel Spectogram*

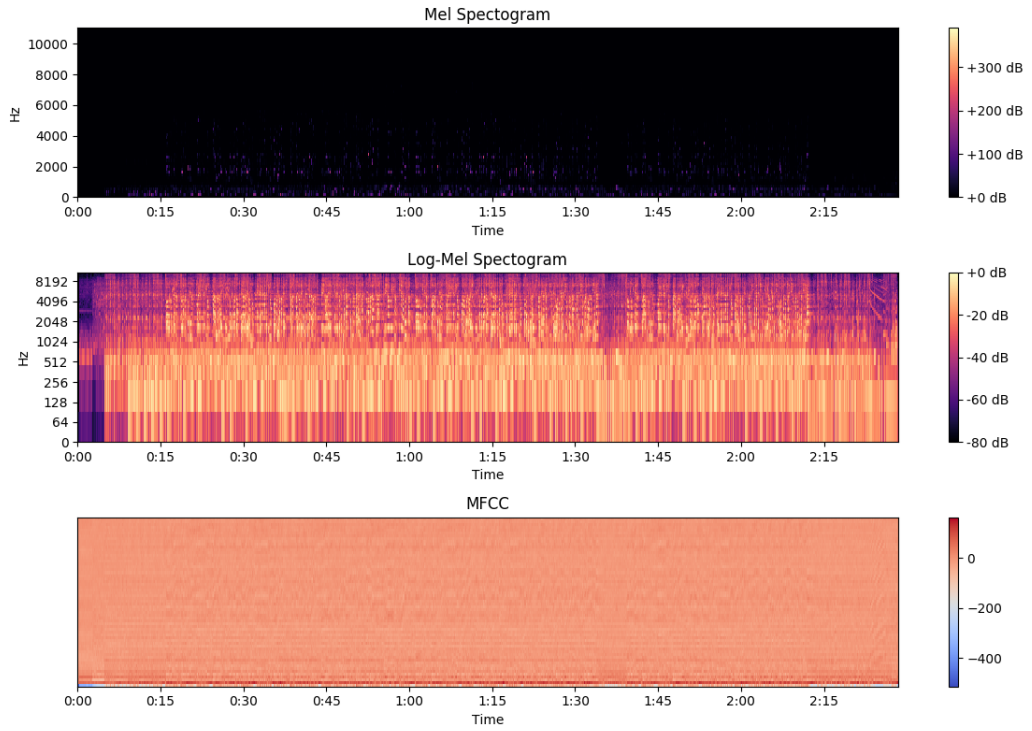
  Ο υπολογισμός του Log-Mel σπεκτρογράμματος επεκτείνει την προηγούμενη διαδικασία εφαρμόζοντας έναν λογαριθμικό μετασχηματισμό στο Mel φασματογράφημα.

- *MFCC*

  Ακολούθως, ο υπολογισμός του MFCC (Mel Frequeny Cepstral Coefficients) επεκτείνει τον προηγούμενο υπολογισμό εφαρμόζοντας μετασχηματισμό DCT στο Log-Mel φασματογράφημα για την εξαγωγή σαφματικών (cepstral) χαρακτηριστικών.

$$MFCC = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} X_m(i) \cos\left(\frac{c\pi(m-\frac{1}{2})}{M_m}\right) \tag{3.6}$$

  όπου, Q_m είναι η λογαριθμική ενέργεια του m-οστού Log-Mel φασματογραφήματος, c είναι ο δείκτης του σαφματικού συντελεστή.



Σχήμα 3.6: Mel Spectgram, Log-Mel Spectogram, MFCC ηχητικού σήματος

3.2.3 Γενικά χαρακτηριστικά συχνότητας

- *Chroma*

  Τα χαρακτηριστικά chroma ή αλλιώς αντιπροσωπευτικές κλάσεις τονικού ύψους (pitch class profiles) χρησιμοποιούνται ευρέως σε εργασίες που ασχολούνται με την μουσική ανάλυση και αναγνώριση μουσικής. Τα χαρακτηριστικά chroma δεν επηρεάζονται από μεταβολές στη τονική ποιότητα (timbre) και σχετίζονται άμεσα με την μουσική αρμονία. Ο Müller [45] αναφέρει τα chromas ως ισχυρά χαρακτηριστικά μεσαίου επιπέδου ικανά να ανακτήσουν σημαντική πληροφορία για θέματα ηχητικής ανάκτησης (audio retrieval).

  Αν υποθέσουμε την ισοβαθμισμένη τονική κλίμακα που χρησιμοποιείται στην δυτική μουσική 3.7, τότε είναι εύκολο να περιγραφεί η αντιστοίχιση ενός ηχητικού σήματος σε χαρακτηριστικά chroma. Στην πράξη, πρώτα εφαρμόζεται ένας STFT μετασχηματισμός στο σήμα και στη συνέχεια, για κάθε παράθυρο, υπολογίζεται ένα διάνυσμα $x = [x_1, x_2, ..., x_{12}]$ όπου κάθε στοιχείο x_i αντιπροσωπεύει το αντίστοιχο στοιχείο της κλίμακας 3.7.

$$\{C, C^{\#}, D, D^{\#}, E, F, F^{\#}, G, G^{\#}, A, A^{\#}, B\} \tag{3.7}$$

- *Tonnetz*

  Τα κεντροειδή τονικά χαρακτηριστικά (tonnetz) [11] είναι μια αναπαράσταση του τονικού ύψους ενός ηχητικού σήματος. Το διάνυσμα tonnetz t_n, στο χρονικό παράθυρο n είναι

αποτέλεσμα πολλαπλασιασμού του διανύσματος chroma c_n και ενός πίνακα μετασχηματισμού T. Στη συνέχεια, το διάνυσμα t_n διαιρείται με την L_1 νόρμα του διανύσματος c_n για την αποφυγή αριθμητικής ανισορροπίας. Το διάνυσμα tonnetz δίνεται από την σχέση:

$$t_n(d) = \frac{1}{\|c_n\|_1} \sum_{i=1}^{11} T(d,l) c_n(l) \tag{3.8}$$

όπου το $d \in [0,5]$ αναφέρεται στον δείκτη του στοιχείου που υπολογίζεται και το $l \in [0,11]$ αναφέρεται στον δείκτη του διανύσματος chroma.

- *Spectral Contrast*

  Τα χαρακτηριστικά φασματικής αντίθεσης (spectral contrast) αντιπροσωπεύουν την ένταση και τις αντιθέσεις των φασματικών κορυφών και επιπέδων για ένα ηχητικό σήμα. Για την ανάκτηση αυτών των χαρακτηριστικών το σήμα τεμαχίζεται σε πλαίσια και σε κάθε πλαίσιο εφαρμόζεται μετασχηματισμός Fourier.

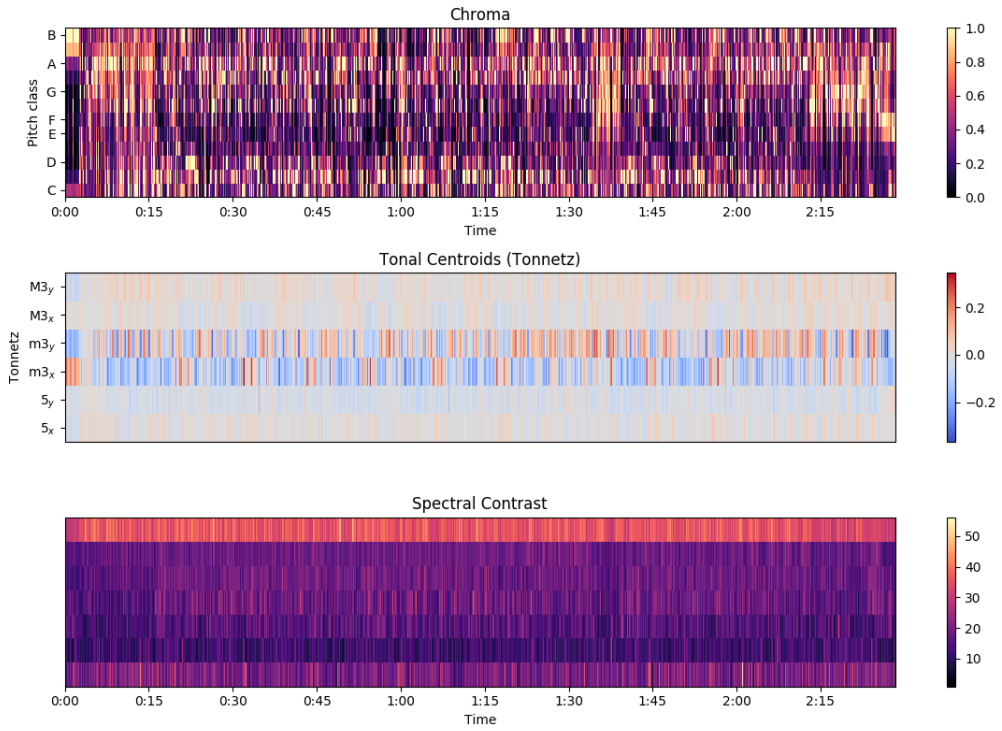
  Στη συνέχεια, το αποτέλεσμα περνά μέσα από φίλτρα οκταβικής κλίμακας, με αποτέλεσμα η συχνότητα να διαιρείται σε μπάντες συχνοτήτων και το αποτέλεσμα μεταφέρεται σε λογαριθμική κλίμακα. Τέλος, εφαρμόζεται ένας μετασχηματισμός Karhunen-loeve για την αντιστοίχιση των χαρακτηριστικών σε έναν ορθογώνιο χώρο και την εξάλειψη πληροφορίας σε ασυσχέτιστες διευθύνσεις. Το αποτέλεσμα αυτής της διαδικασίας είναι μια ισχυρή αναπαράσταση των κορυφών και των επιπέδων του φάσματος και οι αντιθέσεις μεταξύ τους μπορούν εύκολα να αναγνωριστούν.

3.3 Δημιουργία Συνόλου Δεδομένων

Στην παρούσα ενότητα, θα περιγραφεί η διαδικασία που ακολουθήθηκε για τη δημιουργία του συνόλου των δεδομένων (Dataset). Τα dataset που κατασκευάστηκαν θα χρησιμοποιηθούν για την εκπαίδευση των προτεινόμενων μοντέλων που αναπτύχθηκαν για την συναισθηματική ανάλυση μουσικών κομματιών. Η συναισθηματική ανάλυση, στα πλαίσια της διπλωματικής, προσεγγίστηκε από δύο διαφορετικές κατευθύνσεις, η πρώτη σχετίζεται με την ανάλυση των στίχων ενός μουσικού κομματιού ενώ η δεύτερη την με την ανάλυση του ηχητικού σήματος. Όπως είναι αναμενόμενο τα μοντέλα που αναπτύχθηκαν για τις δύο διαφορετικές προσεγγίσεις του θέματος απαιτούν διαφορετικού τύπου δεδομένων για την εκπαίδευση τους. Έτσι η δημιουργία του συνόλου δεδομένων βασίζεται σε έναν κοινό κορμό αλλά διασπάται στη συνέχεια σε δύο διαφορετικές διαδικασίες επεξεργασίας.

3.3.1 Κορμός συνόλου δεδομένων

Το σύνολο των δεδομένων που χρησιμοποιήθηκε ως κορμός για την παραγωγή των δεδομένων είναι το MoodyLyrics Dataset [18], το οποίο περιέχει ένα σύνολο από 2.000 τραγούδια κατηγοριοποιημένα σε τέσσερις κλάσεις: happy, sad, relaxed, angry. Η διαδικασία για την



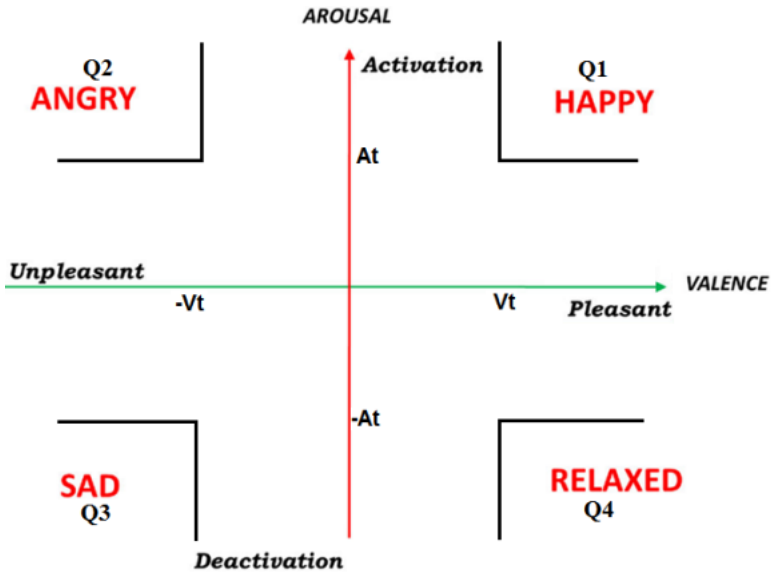
Σχήμα 3.7: Chroma, Tonnal Centroids, Spectral Contrast ηχητικού σήματος

κατηγοριοποίηση των τραγουδιών βασίστηκε στο συναισθηματικό μοντέλο του Russel (βλέπε σχ. 2.22). Αρχικά, σύμφωνα με το [18], η πρώτη διαδικασία για την κατασκευή του dataset σχετίζεται με τη συλλογή των στίχων για κάθε τραγούδι. Στη συνέχεια, σε κάθε λέξη αποδόθηκε ένας ζεύγος τιμών σθένους και διέγερσης βασισμένο στη συσχέτιση των λέξεων από τρία λεξικά που περιέχουν συναισθηματική πληροφορία, το ANEW (Affective Norms of English Words) [9], το WordNet[1] και το WordNet-Affect. Για κάθε τραγούδι υπολογίστηκε ένα γενικό ζευγάρι τιμών σθένους και διέγερσης από το σύνολο των επιμέρους τιμών των λέξεων που συνθέτουν τους στίχους. Τέλος, η κατηγοριοποίηση ολοκληρώνεται με την διακριτοποίηση των τραγουδιών στις συναισθηματικές κλάσεις σύμφωνα με τον πίνακα 3.1.

Τιμές σθένους (V) και διέγερσης (A)	**Συναίσθημα**
$A > A_t$ και $V > V_t$	Happy
$A > A_t$ και $V < -V_t$	Angry
$A < -A_t$ και $V < -V_t$	Sad
$A < -A_t$ και $V > V_t$	Relaxed

Πίνακας 3.1: Ταξινόμηση τραγουδιών

[1] https://wordnet.princeton.edu/



Σχήμα 3.8: Ταξινόμηση των τεσσάρων συναισθημάτων στο χώρο Circumplex

Πέρα από την κατηγορία της κλάσης για κάθε τραγούδι το dataset περιέχει πληροφορίες σχετικά με τον τίτλο του τραγουδιού, τον καλλιτέχνη καθώς και ένα αναγνωριστικό (ID). Οι πληροφορίες αυτές θα χρησιμοποιηθούν στη συνέχεια για την ανάκτηση των απαραίτητων δεδομένων από το διαδίκτυο.

3.3.2 Σύνολο δεδομένων κειμένου

Για την δημιουργία του συνόλου δεδομένων κειμένου χρησιμοποιήθηκαν τα ζευγάρια συμβολοσειρών (title, artist) από τις στήλες χαρακτηριστικών του MoodyLyrics. Τα ζεύγη (title, artist) χρησιμοποιήθηκαν για την αναζήτηση και την ανάκτηση των στίχων κάθε τραγουδιού από το διαδίκτυο. Στην πράξη χρησιμοποιήθηκε ένας προσαρμοσμένος Python κώδικας για την αυτόματη αναζήτηση και ανάκτηση των στίχων από την ιστοσελίδα lyrics.wikia.com[2]. Οι στίχοι που απέτυχαν να ανακτηθούν με τη διαδικασία αυτή αναζητήθηκαν χειροκίνητα στην ιστοσελίδα genius.com[3].

Για την επαύξηση του dataset και για την καλύτερη απόδοση των συστημάτων μηχανικής μάθησης που χρησιμοποιήθηκαν οι στίχοι κάθε τραγουδιού διαιρέθηκαν σε τετράστιχα. Με τη διαίρεση αυτή επιτεύχθηκε η επαύξηση του συνόλου δεδομένων κειμένου από 2.000 στοιχεία σε 18.115 στοιχεία.

Ωστόσο, στα μοντέλα Επεξεργασίας Φυσικής Γλώσσας οι λέξεις που συνθέτουν ένα κείμενο, εν προκειμένω ένα τετράστιχο, δεν παρουσιάζουν κάποια άμεση αριθμητική πληροφορία. Για να μπορέσουν τα τετράστιχα να έχουν νόημα για τα μοντέλα που αναπτύχθηκαν είναι απαραίτητη η αναπαράσταση των λέξεων από διανύσματα, γνωστά ως Word Embeddings. Ο υπολογισμός των διανυσμάτων γίνεται με διάφορες μεθοδολογίες, όπως περιγράφονται στην ενότητα 3.1, που παράγουν διανύσματα διαφορετικών διαστάσεων.

[2] https://lyrics.fandom.com/wiki/LyricWiki
[3] https://genius.com/

3.3.3 Σύνολο δεδομένων ηχητικού σήματος

Παρόμοια διαδικασία με την προηγούμενη ακολουθήθηκε για την δημιουργία του συνόλου δεδομένων ηχητικού σήματος. Τα ζεύγη συμβολοσειρών (title, artist) πάλι χρησιμοποιήθηκαν ως παράμετροι σε έναν προσαρμοσμένο Python κώδικα για την αναζήτηση και την ανάκτηση των τραγουδιών σε μορφή .mp3 από το youtube.com[4].

Επαύξηση του dataset εφαρμόστηκε και στα δεδομένα ηχητικού σήματος. Σε αυτή την περίπτωση τα αρχεία ήχου διαιρέθηκαν σε κλιπς διάρκειας 10 δευτερολέπτων, με αποτέλεσμα την επαύξηση του συνόλου δεδομένων στα 37.989 στοιχεία. Παράλληλα με την διαίρεση σε κλιπς εφαρμόστηκε και υποδειγματοληψία στα ηχητικά σήματα, με ρυθμό δειγματοληψίας από τα 44.100Hz στα 22.050Hz. Παρατηρήθηκε ότι ο υψηλός ρυθμός δειγματοληψίας δεν έπαιζε σημαντικό ρόλο στην αναγνώριση του συναισθήματος αλλά η μείωση του βοήθησε σημαντικά στην ελαχιστοποίηση των διαστάσεων εισόδου των δεδομένων.

Τα ηχητικά αρχεία, όμως, δεν βρίσκονται σε μορφή που να μπορεί να αξιοποιηθεί άμεσα για την εκπαίδευση των συστημάτων που αναπτύχθηκαν. Για την επίτευξη αυτού, θα ακολουθηθεί η διαδικασία εξαγωγής χαρακτηριστικών όπως περιγράφηκε στην ενότητα 3.2. Στην πράξη έγινε χρήση της δημοφιλούς, στην ανάπτυξη MIR (Music Information Retrieval) συστημάτων, βιβλιοθήκης librosa[5] της Python. Από την βιβλιοθήκη librosa θα χρησιμοποιηθούν συγκεκριμένα οι συναρτήσεις melspectogram(), mfcc(), chroma_stft(), tonnetz(), spectral_contrast() για την εξαγωγή των αντίστοιχων χαρακτηριστικών. Η παραπάνω διαδικασία έχει ως αποτέλεσμα την δημιουργία ενός διανύσματος εισόδου με διαστάσεις (37989, 60, 431, 6).

[4] https://www.youtube.com/
[5] https://librosa.github.io

Κεφάλαιο 4

Προτεινόμενα Συστήματα

Στο κεφάλαιο αυτό θα αναλυθεί η ανάπτυξη των προτεινόμενων συστημάτων, στο πλαίσιο προσέγγισης του προβλήματος αναγνώρισης συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής Θα γίνει μια λεπτομερής περιγραφή των αρχιτεκτονικών που δομήθηκαν αλλά και των διάφορων μεθοδολογιών ενσωμάτωσης του συνόλου δεδομένων που εφαρμόστηκαν στην πράξη. Όπως έχει ήδη αναφερθεί η προσέγγιση του συγκεκριμένου προβλήματος θα μελετηθεί από δύο οπτικές, έτσι η διαδικασία περιγραφής των προτεινόμενων συστημάτων θα γίνει ξεχωριστά για την ανάλυση κειμένου και την ανάλυση ηχητικού σήματος. Τελικός στόχος είναι ο συνδυασμός των δύο συστημάτων που παρουσιάζουν τις καλύτερες επιδόσεις σε ένα ενιαίο σύστημα.

4.1 Συστήματα Ανάλυσης Κειμένου

Η ακολουθιακή φύση των μονάδων, δηλαδή των λέξεων, που συνθέτουν ένα κείμενο στίχων και δομούν το σύνολο των δεδομένων είναι αναμενόμενο ότι πρέπει να ληφθούν υπ΄ όψιν σε ένα σύστημα Επεξεργασίας Φυσικής Γλώσσας. Με γνώμονα τη σκοπιμότητα αυτή ,τα συστήματα που θα προταθούν πρέπει να βρίσκονται σε θέση να επεξεργαστούν και να αξιοποιήσουν την ακολουθιακή φύση των λέξεων σε ένα κείμενο.

4.1.1 Προτεινόμενες αρχιτεκτονικές

- **LSTM** (T_1)

  Όπως αναφέρθηκε στην ενότητα 2.5 τα RNNs είναι επαναλαμβανόμενα (ακολουθιακά) δίκτυα των οποίων αναγνωριστικό χαρακτηριστικό είναι ότι σε κάθε βήμα (timestep) η έξοδος υπολογίζεται από την τρέχουσα είσοδο και την κρυφή κατάσταση hidden state. Το χαρακτηριστικό αυτό υποδεικνύει πως σε κάθε βήμα ο υπολογισμός της εξόδου είναι συνάρτηση όλων των εισόδων που έχουν τροφοδοτηθεί στο δίκτυο μέχρι εκείνη τη στιγμή. Αυτή η λογική αναδεικνύει την ικανότητα των RNNs να περιλαμβάνουν στην τελική τους κατάσταση όλη την πληροφορία που έχει προηγηθεί.

  Ωστόσο, τα απλά RNNs (vanilla) παρουσιάζουν αδυναμίες όταν καλούνται να επεξερ-

γαστούν ακολουθίες εισόδου μεγάλου μήκους, όπως για παράδειγμα μια μικρή παράγραφος ή ένα τετράστοιχο τραγουδιού, λόγο του προβλήματος της βραχυπρόθεσμης μνήμης (short-term memory). Λύση σε αυτό το πρόβλημα έρχονται να δώσουν τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTMs). Το χαρακτηριστικό των LSTMs που τους επιτρέπει να αντιμετωπίζουν το πρόβλημα της βραχυπρόθεσμης μνήμης είναι η εσωτερικοί μηχανισμοί που συμπεριλαμβάνουν στη δομή τους και ονομάζονται πύλες. Οι πύλες αυτές ελέγχουν την πληροφορία που ρέει μέσα από τους κόμβους του δικτύου, αναγνωρίζοντας και αξιοποιώντας τα χρήσιμα δεδομένα σε μια ακολουθία και αποκόπτοντας τα υπόλοιπα.

Λόγω των ισχυρών τους χαρακτηριστικών, τα κύτταρα LSTM επιλέχθηκαν στο πλαίσιο της διπλωματικής, ως μια πρώτη προσπάθεια για την δόμηση ενός νευρωνικού δικτύου για την ανάλυση συναισθήματος από κείμενο. Συγκεκριμένα, το δίκτυο αποτελείται από ένα επίπεδο ενσωμάτωσης των διανυσμάτων αναπαράστασης, ένα επίπεδο νευρώνων LSTM και δύο επίπεδα πλήρως συνδεδεμένων νευρώνων (βλέπε 2.4.2.4). Ωστόσο, τα διανύσματα που αποδίδουν οι μεθοδολογίες ενσωμάτωσης κειμένου Bag of Words και TF-IDF δεν παρουσιάζουν κάποια ακολουθιακή φύση, επομένως για την εκπαίδευση αυτών των δύο μοντέλων θα αφαιρεθεί το επίπεδο των νευρώνων LSTM και η τροποποιημένη αρχιτεκτονική αυτή θα αναφέρεται ως $T_1^{'}$. Από αρχιτεκτονικής πλευράς, ο συνδυασμός των τεχνικών χαρακτηριστικών του δικτύου επιλέχθηκαν έπειτα από εφαρμογή της αναζήτησης πλέγματος (grid search), μια τεχνική βελτιστοποίησης των παραμέτρων του δικτύου. Συγκεκριμένα, οι διαστάσεις του πρώτου επιπέδου ενσωμάτωσης ήταν μεταβλητό μέγεθος και επιλέχθηκαν ανάλογα με την τεχνική ενσωμάτωσης δεδομένων που εφαρμόστηκε στην κάθε περίπτωση.

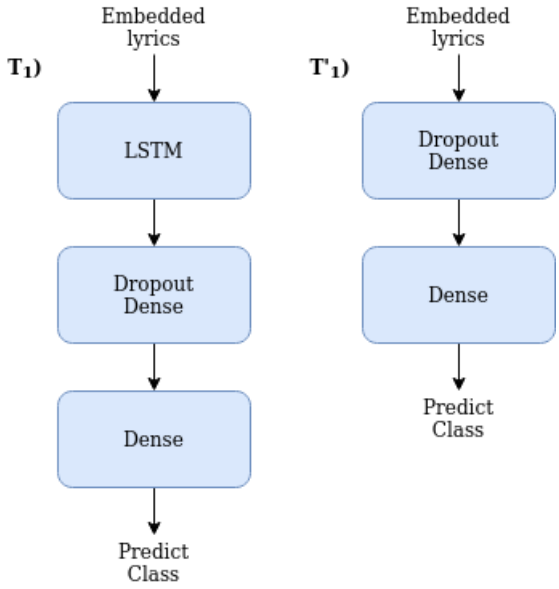
Για την κατασκευή του δεύτερου επιπέδου επιλέχθηκαν 64 νευρώνες LSTM με συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη ($tanh()$), επαναληπτική συνάρτησης ενεργοποίησης την σιγμοειδή ($sigmoid()$), βαθμό απόσυρσης (dropout) ίσο με 0.2 και επαναλαμβανόμενο βαθμό dropout ίσο με 0.2. Ο βαθμός dropout εκφράζει την πιθανότητα η έξοδος ενός νευρώνα να μηδενιστεί. Στη συνέχεια, το δεύτερο επίπεδο, των πλήρως συνδεδεμένων νευρώνων, κατασκευάστηκε από 128 νευρώνες, με μηδενικό βαθμό dropout και ως συνάρτηση ενεργοποίησης επιλέχθηκε η συνάρτηση $relu()$.

Το τρίτο και τελευταίο επίπεδο, αποτελείται από 4 πλήρως συνδεδεμένους νευρώνες με συνάρτηση ενεργοποίησης την συνάρτηση $softmax()$.

Ουσιαστικά, τα δύο τελευταία επίπεδα των πλήρως συνδεδεμένων νευρώνων συνθέτουν έναν ταξινομητή classifier, υπεύθυνο για την ταξινόμηση των δειγμάτων στις τέσσερις κλάσεις του προβλήματος αξιοποιώντας την πληροφορία που παράγει στην έξοδο του το επίπεδο των νευρώνων LSTM.

- **BERT** (T_2)

  Η αλματώδης πρόοδος που έχουν οι Transformers τα τελευταία χρόνια στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (NLP), λόγο της αμφίδρομης εκπαίδευσής τους, σε



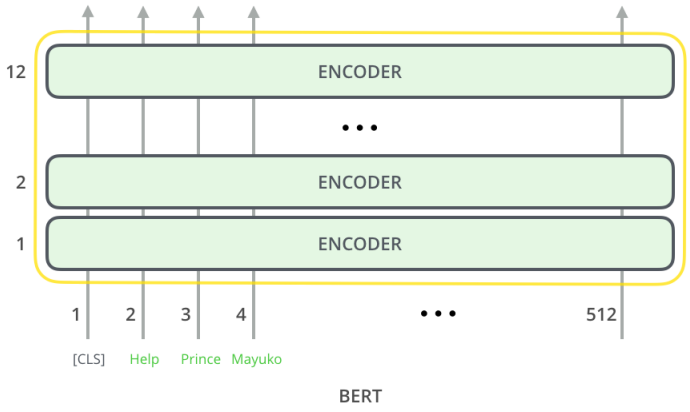
Σχήμα 4.1: Δομή των μοντέλων $T_1, T_1^{'}$

συνδυασμό με την ικανότητα που εμφανίζει η τεχνική Transfer Learning να χρησιμοποιεί την γνώση που έχει παραχθεί από την αντιμετώπιση ενός προβλήματος σε ένα άλλο, οδήγησαν στην επιλογή της επόμενης αρχιτεκτονικής. Η ικανότητα του προεκπαιδευμένου μοντέλου BERT (βλέπε 2.6.2) να συνδυάσει την αρχιτεκτονική των Tranformers με την γνώση που έχει αποκτηθεί από την εκπαίδευση του σε έναν τεράστιο όγκο αγγλικών κειμένων, το έχουν αναδείξει ως μια κορυφαία επιλογή στην ανάπτυξη ενός συστήματος NLP όπως είναι η συναισθηματική ανάλυση μουσικών κομματιών από στίχους.

Έτσι, σαν μια δεύτερη αρχιτεκτονική για την προσέγγιση του θέματος επιλέχθηκε το μοντέλο BERT.

Η Google διαθέτει μια λίστα επιλογών για το BERT, στην πράξη αυτό που επιλέχθηκε, στα πλαίσια της διπλωματικής, ήταν το BERT-base uncased. Το συγκεκριμένο μοντέλο είναι η απλή/μικρή (base) έκδοση που διατίθεται, σε αντίθεση με την δεύτερη και μεγαλύτερη (large) έκδοση του μοντέλου. Η διαφορά των δύο μοντέλων έγκειται στο μέγεθος της εσωτερικής οργάνωσης τους, το απλό μοντέλο αποτελείται από 12 επίπεδα με μέγεθος κρυφού επιπέδου (hidden layer size) ίσο με 768 και 110 εκατομμύρια συνολικές παραμέτρους, σε αντίθεση με το μεγάλο μοντέλο με 24 επίπεδα, κρυφό μέγεθος επιπέδου ίσο με 1024 και 340 εκατομμύρια συνολικές παραμέτρους. Ο όρος uncased χαρακτηρίζει την είσοδο που αναμένει το μοντέλο, οι λέξεις της εισόδου αναμένεται να είναι στα αγγλικά χωρίς κεφαλαίους χαρακτήρες. Εσωτερικά, όπως αναφέρθηκε, το μοντέλο αποτελείται από 12 πανομοιότυπα επίπεδα μεγέθους 768, κάθε επίπεδο προσομοιώνει έναν κωδικοποιητή, όπως περιγράφηκε στην ενότητα 2.6.

Συγκεκριμένα, κάθε κωδικοποιητής αποτελείται από διάφορα τμήματα επεξεργασίας των διανυσμάτων ενσωμάτωσης (embedding vectors) με πρώτο τμήμα έναν μηχανισμό προ-



Πηγή: towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b

Σχήμα 4.2: Εσωτερική δομή του μοντέλου BERT

σοχής multihead-attention με 12 σημεία εστίασης. Στη συνέχεια ακολουθεί ένα επίπεδο κανονικοποίησης (normalization layer) ακολουθούμενο από ένα εμπρόσθια τροφοδοτούμενο δίκτυο, δομημένο από δύο πλήρως συνδεδεμένα επίπεδα νευρώνων με συνάρτηση ενεργοποίησης την συνάρτηση $relu()$. Τελευταίο τμήμα του κωδικοποιητή είναι ο μηχανισμός κωδικοποίησης θέσης εισάγοντας διανύσματα ενσωμάτωσης θέσης παράλληλα με τα διανύσματα ενσωμάτωσης των λέξεων.

Στο τέλος του μοντέλου εισάγουμε και σε αυτή την περίπτωση έναν ταξινομητή δύο πλήρως συνδεδεμένων επιπέδων. Το πρώτο επίπεδο αποτελείται από 256 νευρώνες, με συνάρτηση ενεργοποίησης την σιγμοειδή $sigmoid()$ και βαθμό dropout ίσο με 0.1, ενώ το δεύτερο επίπεδο αποτελείται από 4 νευρώνες και σαν συνάρτηση ενεργοποίησης χρησιμοποιήθηκε η συνάρτηση $softmax()$.

4.1.2 Προτεινόμενες ενσωματώσεις δεδομένων

- **Bag of Words**

  Η πρώτη τεχνική ενσωμάτωσης των δεδομένων στο μοντέλο T_1 που εφαρμόστηκε στην πράξη ήταν αυτή των Bag of Words (Bow), όπως περιγράφεται στην ενότητα 3.1.2. Πριν από την ενσωμάτωση των λέξεων σε διανύσματα προηγήθηκε μια διαδικασία προεπεξεργασίας του κειμένου. Συγκεκριμένα, πρώτη εργασία της προεπεξεργασίας ήταν η αντικατάσταση όλων των κεφαλαίων χαρακτήρων του αγγλικού αλφάβητου από τους αντίστοιχους πεζούς χαρακτήρες.

  Στη συνέχεια, αφαιρέθηκαν όλα τα σημεία στίξης και οι αγγλικές stopwords, η λίστα των stopwords ανακτήθηκε από τη βιβλιοθήκη NLTK[1] της Python. Ο όρος stopwords αναφέρεται στις λέξεις που εμφανίζονται πιο συχνά στη δομή ενός κειμένου και αποδίδουν

[1] https://www.nltk.org/

ελάχιστη σημασία για την νοηματική απόδοσή του, μερικά παραδείγματα stopwords είναι τα "the", "is", "at", "which".

Τέλος, πριν την τμηματοποίηση tokenization των κειμένων εισόδου σε όρους tokens και την απόδοση αναγνωριστικών IDs στους όρους, προηγείται η διαδικασία προεπεξεργασίας κειμένου stemming. Το stemming αναφέρεται στην διαδικασία μείωσης των παράγωγων λέξεων στα βασικά τους στελέχη, για παράδειγμα το stemming θα μειώσει στην λέξη "cat" τις λέξεις "cats", "catlike", "catty".

Μετά την ενσωμάτωση των δεδομένων εισόδου (τετράστιχα) με τη μεθοδολογία Bow προκύπτει το μέγεθος του λεξιλογίου να είναι $|V| = 21.266$. Από το λεξιλόγιο επιλέγονται να χρησιμοποιηθούν για την αναπαράσταση των κειμένων οι 10.000 πιο συχνές λέξεις.

Έτσι τελικά προκύπτει κάθε τετράστιχο να αναπαριστάται από ένα διάνυσμα μήκους 10.000, όπου κάθε στοιχείο του εκφράζει τον αριθμό εμφάνισης της αντίστοιχης λέξης του λεξιλογίου στο τετράστιχο. Στο επίπεδο ενσωμάτωσης Embedding layer, του μοντέλου T_1 τα διανύσματα εισόδου συμπιέζονται σε πυκνά (dense) διανύσματα μήκους $EMBEDDING\_SIZE = 128$.

- **TF-IDF**

  Σαν δεύτερη τεχνική, λίγο πιο εξελιγμένη από την πρώτη, αναπτύχθηκε η τεχνική TF-IDF (βλέπε ενότητα 3.1.3). Η διαδικασία προεπεξεργασίας κειμένου που προηγήθηκε και σε αυτή την περίπτωσή είναι η ίδια με αυτή που προηγήθηκε και στην τεχνική BoW.

  Όπως, είναι αναμενόμενο και σε αυτή την μεθοδολογία το μέγεθος του λεξιλογίου προκύπτει $|V| = 21.266$ και πάλι διατηρούνται οι 10.000 λέξεις με την μεγαλύτερη συχνότητα. Αυτή τη φόρα όμως οι λέξεις αναπαριστώνται από τα διανύσματα που προκύπτουν σύμφωνα με τη σχέση 3.1. Πάλι στο επίπεδο ενσωμάτωσης τα διανύσματα εισόδου συμπιέζονται σε πυκνά διανύσματα μήκους $EMBEDDING\_SIZE = 128$.

- **Word2Vec**

  Μια τρίτη προσπάθεια για την ενσωμάτωση των δεδομένων, στο μοντέλο T_1, βασίστηκε στη τεχνική Word2Vec (βλέπε ενότητα 3.1.4). Πριν την εκπαίδευσή του μοντέλου ακολουθείται παρόμοια διαδικασία προεπεξεργασίας με τις προηγούμενες. Συγκεκριμένα, όλοι οι χαρακτήρες μετατρέπονται σε πεζούς, τα σημεία στίξης, οι αριθμητικοί χαρακτήρες και τα stopwords αφαιρούνται.

  Για την εκπαίδευση του μοντέλου χρησιμοποιούνται οι επεξεργασμένες ακολουθίες εισόδου και το μοντέλο παράγει διανύσματα μήκους $|s| = 408$, που αναπαριστούν τις προτάσεις κάθε ακολουθίας. Το μέγεθος $|s|$ προκύπτει ως το μέγιστο σύνολο λέξεων από τις ακολουθίες εισόδου, κάθε στοιχείο s_i αναπαριστά μια λέξη στην ακολουθία της εισόδου, ενώ τα υπόλοιπα στοιχεία αντικαθιστώνται από μηδενικά.

  Στο επίπεδο ενσωμάτωσης πάλι η διάσταση ενσωμάτωσης ορίζεται $EMBEDDING\_SIZE = 128$ ενώ αυτή τη φορά ορίζεται και αρχικοποιητής ενσωμάτωσης (embedding initializer)

ο πίνακας, διάστασης $|V|xEMBEDDING\_SIZE$, που απαρτίζεται από τις ενσωματώσεις όλων των λέξεων του λεξιλογίου και μετατρέπει κάθε λέξη της εισόδου στο αντίστοιχο διάνυσμα (embedding vector).

- **GloVe**

  Σαν τελευταία τεχνική για την ενσωμάτωση των δεδομένων εισόδου στο σύστημα T_1 θα εφαρμοστεί η μεθοδολογία GloVe, όπως περιγράφεται στην ενότητα 3.1.5. Για την προεπεξεργασία των ακολουθιών εισόδου ακολουθείται η ίδια διαδικασία με αυτή που περιγράφηκε στο προηγούμενο αντικείμενο (Word2Vec).

  Για την ενσωμάτωση των λέξεων σε αυτή την περίπτωση χρησιμοποιήθηκαν οι προεκπαιδευμένες αναπαραστάσεις που παρέχει το GloVe. Ουσιαστικά, πρόκειται για ένα λεξιλόγιο 400.000 λέξεων με προεκπαιδευμένα διανύσματα αναπαράστασης μήκους 100. Από τα έτοιμα διανύσματα κατασκευάζεται οι πίνακας συνεμφάνισης (co-occerance matrix), διάστασης $400.000x100$, που στη συνέχεια θα αντικαταστήσει τη μη-εκπαιδευόμενα βάρη του επιπέδου ενσωμάτωσης. Επειδή το μοντέλο GloVe χρησιμοποιεί πεπερασμένο λεξιλόγιο, τις λέξεις εκτός του λεξιλογίου τις αγνοεί, δηλαδή αντικαθιστά τα διανύσματα τους με μηδενικά.

- **BERT Embeddings**

  Το μοντέλο BERT (T_2) απαιτεί μια συγκεκριμένη αναπαράσταση των δεδομένων εισόδου για την ορθή λειτουργία του. Η πρώτη εργασία που εκτελέστηκε ήταν ο διαχωρισμός των προτάσεων εισόδου σε όρους, το BERT παρέχει έναν ειδικά σχεδιασμένο τμηματοποιητή (tokenizer) για αυτή τη λειτουργία.

  Συγκεκριμένα, ο tokenizer δέχεται στη είσοδο του τα κείμενα για τμηματοποίηση και δημιουργεί μια ακολουθία από όρους-λέξεις αντιστοιχίζοντας κάθε λέξη της εισόδου στον αντίστοιχο όρο που παρέχει το λεξιλόγιο του, διατηρώντας παράλληλα τη σειρά εμφάνισης των λέξεων. Για τις λέξεις της εισόδου που δεν αναγνωρίζονται στο λεξιλόγιο ο tokenizer θα προσπαθήσει να τις διασπάσει στους όρους του λεξιλογίου με μέγιστο αριθμό χαρακτήρων, στη χειρότερη περίπτωση θα διασπάσει μια λέξη στους μεμονωμένους χαρακτήρες που την απαρτίζουν. Σε περίπτωση διάσπασης μιας λέξης ο πρώτος όρος θα εμφανιστεί στην ακολουθία ως έχει, ενώ στους υπόλοιπους όρους θα προστεθεί στην αρχή τους το διπλό σύμβολο # ώστε να αναγνωρίσει το μοντέλο ότι πρόκειται για όρους στους οποίους έχει προηγηθεί διάσπαση, για παράδειγμα η λέξη "playing" θα διασπαστεί στους όρους "play" και "##ing".

  Αφού ολοκληρωθεί η διάσπαση ο tokenizer θα προσθέσει στην αρχή της ακολουθίας όρων της εξόδου τον όρο [CLS] και στο τέλος τις ακολουθίας τον όρο [SEP]. Το αποτέλεσμα του tokenizer θα χρησιμοποιηθεί στη συνέχεια για την παραγωγή των διανυσμάτων που απαιτεί το μοντέλο για την εκπαίδευση του.

  Για την ακρίβεια, το μοντέλο αναμένει τρία παράλληλα διανύσματα με σταθερό μήκος για την αναπαράσταση κάθε ακολουθίας στην είσοδο, το μήκος των διανυσμάτων επιλέχθηκε να είναι: $max\_seq\_length = 128$. Τα διανύσματα που θα κατασκευαστούν για

κάθε ακολουθία είναι τα input_ids, input_masks και segments_ids. Το διάνυσμα input_ids κατασκευάζεται από τα αναγνωριστικά IDs για κάθε όρο της ακολουθίας, όπως παρέχεται από το λεξιλόγιο του μοντέλου. Το διάνυσμα input_masks κατασκευάζεται από άσσους στις n πρώτες θέσεις, όπου n είναι το μήκος της ακολουθίας εισόδου, και από μηδενικά στις υπόλοιπες $128-n$ θέσεις. Το διάνυσμα segment_ids χρησιμοποιείται για τον διαχωρισμό των προτάσεων που συνθέτουν μια ακολουθία, για παράδειγμα αν μια ακολουθία αποτελείται από τρεις προτάσεις στο διάνυσμα segment_ids τα στοιχεία που αντιστοιχούν στους όρους τις πρώτης πρότασης θα πάρουν την τιμή 0, για την δεύτερη πρόταση τα στοιχεία θα πάρουν την τιμή 1, για την τρίτη πρόταση την τιμή 3 κ.ο.κ.. Για τις ακολουθίες εισόδου που ξεπερνούν τούς 128 όρους οι περίσσιοι όροι αποκόπτονται, ενώ οι ακολουθίες με μικρότερο μήκος συμπληρώνονται από κενούς όρους.

4.2 Συστήματα Ανάλυσης Ηχητικού Σήματος

Η φύση των δεδομένων και στην περίπτωση της ανάλυσης ηχητικού σήματος θα καθορίσει τα συστήματα που αναπτύχθηκαν. Τα ηχητικά σήματα, πριν τροφοδοτηθούν στο δίκτυο, υφίστανται συγκεκριμένες διαδικασίες επεξεργασίας με σκοπό την εξαγωγή χαρακτηριστικών. Η επεξεργασία ηχητικού σήματος στην παρούσα εργασία έχει αποτέλεσμα τη δημιουργία ορισμένων πινάκων χαρακτηριστικών, όπως περιγράφεται στην ενότητα 3.2, οι οποίοι θα χρησιμοποιηθούν ως δεδομένα εισόδου και θα τροφοδοτηθούν στο σύστημα.

4.2.1 Προτεινόμενες αρχιτεκτονικές

- **CNN** (A_1)

  Η ικανότητα των συνελικτικών νευρωνικών δικτύων να συλλάβουν και να αξιοποιήσουν τις χωροχρονικές εξαρτήσεις των δεδομένων, μέσα από την εφαρμογή κατάλληλων φίλτρων, αλλά και η παράλληλη επεξεργασία που εφαρμόζουν οδήγησε στην επιλογή τους για την ανάπτυξη της επόμενης αρχιτεκτονικής.

  Για την οργάνωση της αρχιτεκτονικής δομής του CNN που κατασκευάστηκε χρησιμοποιήθηκαν τρία ζευγάρια επιπέδων συνέλιξης, δύο διαστάσεων (Conv2D) και pooling (βλέπε ενότητα 2.4), ακολουθούμενα από έναν ταξινομητή δύο πλήρως συνδεδεμένων επιπέδων. Οι υπερπαράμετροι του δικτύου καθορίστηκαν με την βοήθεια της αναζήτησης πλέγματος (grid search), μια τεχνική που αναδεικνύει τις βέλτιστες τιμές υπερπαραμέτρων από ένα σύνολο τιμών, δοκιμάζοντας όλους τους πιθανούς συνδυασμούς αυτών και επιλέγοντάς τον συνδυασμό που πετυχαίνει την βέλτιστη επίδοση συστήματος.

  Συγκεκριμένα, για το πρώτο συνελικτικό επίπεδο επιλέχθηκε η χρήση 128 διαφορετικών φίλτρων με μέγεθος πυρήνα (kernel size) ίσο με $6x6$ και συνάρτηση ενεργοποίησης την συνάρτηση $relu()$. Για το επίπεδο pooling (συσσώρευσης) που το ακολουθεί εφαρμόστηκε η μέθοδος Max pooling με διαστάσεις παραθύρου (pool size) ίσο με $2x2$.

  Για το δεύτερο συνελικτικό επίπεδο επιλέχθηκε η χρήση 256 διαφορετικών φίλτρων με μέγεθος πυρήνα ίσο με $5x5$ και συνάρτηση ενεργοποίησης την συνάρτηση $relu()$.

Ακόμα, για στο επίπεδο εφαρμόστηκε η τεχνική συμπλήρωσης περιθωρίου $padding =$ "$same$", ώστε το μήκος της εξόδου στο επίπεδο αυτό να είναι το ίδιο με το μήκος της εισόδου. Για το δεύτερο επίπεδο pooling που το ακολουθεί εφαρμόστηκε πάλι η μέθοδος Max pooling με διαστάσεις παραθύρου ίσο με $2x2$. Για το τρίτο συνελικτικό επίπεδο επιλέχθηκε η χρήση 512 διαφορετικών φίλτρων με μέγεθος πυρήνα ίσο με $4x4$, συνάρτηση ενεργοποίησης την συνάρτηση $relu()$ και $padding =$ "$same$". Για το τρίτο επίπεδο pooling εφαρμόστηκε πάλι η μέθοδος Max pooling με διαστάσεις παραθύρου ίσο με $2x2$.

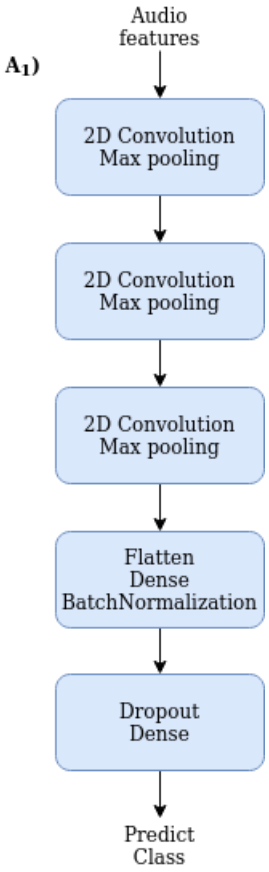
Πριν τον ταξινομητή των δύο επίπεδων παρεμβάλλεται ένα επίπεδο ισοπέδωσης (flatten) που ως αποτέλεσμα έχει την μετατροπή των τρισδιάστατων διανυσμάτων που δέχεται στην είσοδο του, σε μονοδιάστατα. Σκοπός αυτού του επιπέδου είναι η μετατροπή της πληροφορίας που ρέει στο δίκτυο σε μορφή συμβατή με τον ταξινομήτη που ακολουθεί. Για το πρώτο επίπεδο πλήρως συνδεδεμένων νευρώνων χρησιμοποιήθηκαν 16 νευρώνες με συνάρτηση ενεργοποίησης τη συνάρτηση $relu()$ και βαθμό dropout ίσο με 0.6. Για το τελευταίο επίπεδο χρησιμοποιήθηκαν τέσσερις πλήρως συνδεδεμένοι νευρώνες με συνάρτησης ενεργοποίησης τη συνάρτηση $softmax()$, ώστε να υπολογιστούν οι πιθανότητες ταξινόμησης των στοιχείων της εισόδου στις τέσσερις κλάσεις συναισθημάτων του προβλήματος.

Για την αποφυγή της υπερπροσαρμογής over-fitting των δεδομένων θα εφαρμοστούν στο δίκτυο και δύο μέθοδοι κανονικοποίησης και ομαλοποίησης. Οι κανονικοποιητές (regularizers) είναι συναρτήσεις που εφαρμόζουν ποινές στις παραμέτρους ενός επιπέδου κατά τη διάρκεια βελτιστοποίησης. Ουσιαστικά οι ποινές κανονικοποίησης εφαρμόζονται στη συνάρτηση κόστους που προσπαθεί να βελτιστοποιήσει το δίκτυο. Στην περίπτωση του δικτύου A_1 σαν συνάρτηση κανονικοποίησης θα εφαρμοστεί η συνάρτηση L2. Η συνάρτηση L2 regularization στην πράξη προσθέτει το τετράγωνο της τιμής των παραμέτρων σαν όρο ποινής στη συνάρτηση κόστους.

Για παράδειγμα, αν η συνάρτηση κόστους είναι το άθροισμα των τετραγωνικών υπολοίπων (4.1), τότε η κανονικοποίηση L2 θα προσθέσει στη συνάρτηση κόστους τον όρο που φαίνεται στην εξίσωση 4.2. Όταν το λ είναι μηδέν τότε επανερχόμαστε στην 4.1, αλλιώς όταν το λ είναι πολύ μεγάλο θα προσθέσει μεγάλη ποινή και ίσως οδηγήσει σε υποπροσαρμογή (under-fitting) των δεδομένων. Για τον λόγο αυτό η επιλογή του λ πρέπει να είναι πολύ προσεκτική, στα συνελικτικά επίπεδα του δικτύου A_1 εφαρμόζεται L2 regularization με $\lambda = 0.00001$.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \tag{4.1}$$

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4.2}$$



Σχήμα 4.3: Δομή του μοντέλου A_1

Σαν διαδικασία ομαλοποίησης των δεδομένων θα εφαρμοστεί στο πρώτο πλήρως συνδεδεμένο επίπεδο η μέθοδος ομαλοποίηση τεμαχίου Batch noramlization. Το Batch normalization [28] είναι μια τεχνική για τη βελτίωση της ταχύτητας, της επίδοσης και της ευστάθειας των τεχνητών νευρωνικών δικτύων. Για την επίτευξη των παραπάνω, στην πράξη η τεχνική αυτή ομαλοποιεί την έξοδο ενός επιπέδου ενεργοποίησης αφαιρώντας την μέση τιμή του τεμαχίου (batch) και διαιρώντας με την τυπική απόκλιση. Με αυτόν τον τρόπο διατηρείται η μέση τιμή της ενεργοποίησης διατηρείται κοντά στο 0 και η τυπική απόκλιση της ενεργοποίησης κοντά στο 1. Με αυτόν τον τρόπο ακραίες τιμές που θα αποκόπτονταν από την εκπαίδευση του δικτύου ομαλοποιούνται παίρνοντας πιο ΄φυσιολογικές΄ τιμές.

4.2.2 Προτεινόμενες ενσωματώσεις δεδομένων

Για την ενσωμάτωση των δεδομένων ηχητικού σήματος στο σύστημα A_1 εφαρμόστηκαν οι μεθοδολογίες επεξεργασίας σήματος, βασισμένες στην εργασία κατηγοριοποίησης περιβαλλοντικών ήχων [62], όπως περιγράφονται στην ενότητα 3.2. Κάθε μία από τις έξι μεθόδους παράγει έναν πίνακα χαρακτηριστικών διάστασης $60x431$ για κάθε ηχητικό κλιπ της εισόδου. Οι πίνακες αυτοί θα στοιβαχθούν και θα τροφοδοτηθούν στο δίκτυο για την εκπαίδευση του και στη συνέχεια θα μελετηθεί ο βέλτιστος συνδυασμός των πινάκων χαρακτηριστικών. Παρακάτω περιγράφεται η διαδικασία υπολογισμού του κάθε πίνακα, με τη βοήθεια των έτοιμων

συναρτήσεων της βιβλιοθήκης librosa.

- **Mel Spectogram**

  Για τον υπολογισμό του Mel Spectogram θα χρησιμοποιηθεί η συνάρτηση $melspectogram()$ που δέχεται σαν είσοδο το ηχητικό σήμα, υπολογίζει το φασματογράφημα και στη συνέχεια αντιστοιχίζει τις συχνότητες του φασματογραφήματος στην κλίμακα Mel. Για τον υπολογισμό του φασματογραφήματος εφαρμόζεται ο STFT μετασχηματισμός Fourier με παράθυρο {Ηανγινγ, μήκος παραθύρου FFT ίσο με 1024 και αριθμό βημάτων ολίσθησης ίσο με 512. Τέλος για τον μετασχηματισμό στην κλίμακα Mel επιλέγονται 60 μπάντες mel.

- **Log-Mel Spectogram**

  Για τον υπολογισμό του Log-Mel Spectogram θα χρησιμοποιηθεί η συνάρτηση $power\_to\_db()$ που δέχεται σαν είσοδο το melspectogram, όπως υπολογίστηκε προηγουμένως, και μετατρέπει την διάσταση της συχνότητας σε μονάδες ντεσιμπέλ (dB).

- **MFCC**

  Για τον υπολογισμό του MFCC θα χρησιμοποιηθεί η συνάρτηση $mfcc()$ που δέχεται σαν είσοδο το ηχητικό σήμα, υπολογίζει το Log-Mel Spectogram στο οποίο εφαρμόζει DCT μετασχηματισμό για την εξαγωγή των σαφματικών χαρακτηριστικών. Ο αριθμός των MFCC χαρακτηριστικών που θα επιστραφούν επιλέγεται ίσος με 60.

- **Chroma**

  Για τον υπολογισμό του Chroma θα χρησιμοποιηθεί η συνάρτηση $chroma\_stft()$ που δέχεται σαν είσοδο το ηχητικό σήμα, υπολογίζει το φασματογράφημα και στη συνέχεια αντιστοιχίζει τις συχνότητες του φασματογραφήματος στις κλάσεις του τονικού ύψους. Για τον υπολογισμό του φασματογραφήματος εφαρμόζεται ο STFT μετασχηματισμός Fourier με παράθυρο {Ηανγινγ, μήκος παραθύρου FFT ίσο με 1024 και αριθμό βημάτων ολίσθησης ίσο με 512. Ο αριθμός για τις τονικές κλάσεις επιλέγεται ίσος με 60.

- **Tonnetz**

  Για τον υπολογισμό του Tonnetz θα χρησιμοποιηθεί η συνάρτηση $tonnetz()$ που δέχεται σαν είσοδο το ηχητικό σήμα και παράγει τον πίνακα t_n (βλέπε σχέση 3.8). Ο πίνακας t_n έχει διαστάσεις $6x431$, με σκοπό αντιστοίχιση του με τους υπόλοιπους πίνακες χαρακτηριστικών κάθε γραμμή του πίνακα επαναλαμβάνεται 10 φορές.

- **Spectral Contrast**

  Για τον υπολογισμό του Spectral Contrast θα χρησιμοποιηθεί η συνάρτηση $spectral\_contrast()$ που δέχεται σαν είσοδο το ηχητικό σήμα, υπολογίζει το φασματογράφημα και στη συνέχεια υπολογίζει τα χαρακτηριστικά φασματικής αντίθεσης. Για τον υπολογισμό του φασματογραφήματος εφαρμόζεται ο STFT μετασχηματισμός Fourier με παράθυρο {Ηανγινγ, μήκος παραθύρου FFT ίσο με 1024 και αριθμό βημάτων ολίσθησης ίσο με

512. Στη συνέχεια το φασματογράφημα διαιρείται σε έξι μπάντες και για κάθε μπάντα υπολογίζεται η μέση τιμή της ενέργειας για τις κορυφές και τις κοιλάδες. Συγκρίνοντας αυτές τις δύο τιμές παράγεται ένας πίνακας διαστάσεων $6x431$, πάλι κάθε γραμμή του πίνακα επαναλαμβάνεται 10 φορές για την αντιστοίχιση με τους υπόλοιπους.

4.3 Συνδυασμός Συστημάτων

Στην ενότητα αυτή θα γίνει μια προσπάθεια για την ανάπτυξη ενός ενιαίου συστήματος, αξιοποίησης γλωσσικών (στίχων) και ηχητικών χαρακτηριστικών, για την αναγνώριση συναισθήματος από μουσικά κομμάτια. Όπως είναι διαπιστωμένο εμπειρικά, η συναισθηματική διέγερση που προκαλείται στον ακροατή ενός μουσικού κομματιού, οφείλεται τόσο στη σύνθεση των στίχων όσο και στη σύνθεση της μουσικής. Έτσι, στόχος της διπλωματικής αυτής ήταν εξ αρχής η αξιοποίηση και η ανάλυση των χαρακτηριστικών αυτών για την ανάπτυξη ενός συστήματος Μηχανικής Μάθησης. Για την ανάπτυξη του συστήματος αυτού θα χρησιμοποιηθούν τα συστήματα, όπως περιγράφηκαν προηγουμένως, που πετυχαίνουν τις βέλτιστες επιδόσεις τόσο στην ανάλυση κειμένου όσο και στην ανάλυση ηχητικού σήματος.

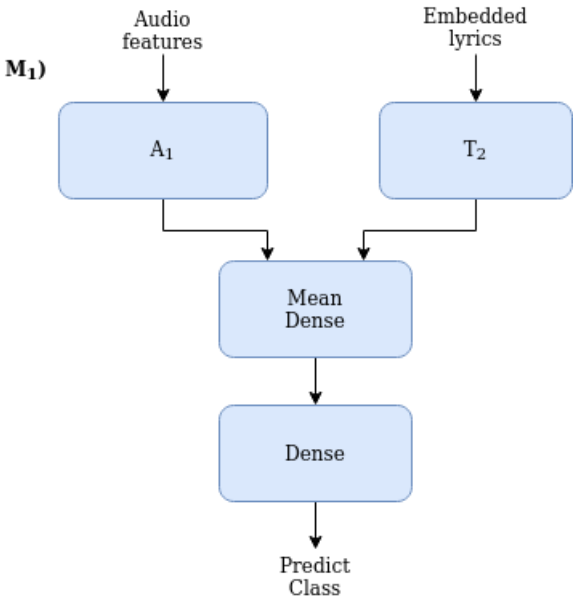
4.3.1 Προτεινόμενη αρχιτεκτονική (M_1)

Η ανάπτυξη της αρχιτεκτονικής του συστήματος θα υλοποιηθεί με τη χρήση των αρχιτεκτονικών BERT (T_2) και CNN (A_1), όπως περιγράφονται στο κεφάλαιο αυτό. Η συγχώνευση των δύο αυτών συστημάτων θα γίνει με τον συνδυασμό των εξόδων τους και την εφαρμογή της τεχνικής late fusion. Πρακτικά για να εφαρμοστεί αυτό, θα κατασκευαστεί ένας ταξινομητής που θα δέχεται τις εξόδους των δύο συστημάτων και θα πραγματοποιεί μια γενική πρόβλεψη, αξιοποιώντας την γνώση και των δύο συστημάτων. Για την κατασκευή του ενιαίου ταξινομητή, θα χρησιμοποιηθούν δύο επίπεδα πλήρως συνδεδεμένων νευρώνων, το πρώτο επίπεδο συνθέτουν 16 νευρώνες με συνάρτηση ενεργοποίησης τη συνάρτηση $relu()$. Ενώ, για το τελευταίο επίπεδο, υπεύθυνο για την ταξινόμηση των δειγμάτων στις τέσσερις συναισθηματικές κατηγορίες, χρησιμοποιήθηκαν τέσσερις πλήρως συνδεδεμένοι νευρώνες με συνάρτηση ενεργοποίησης τη συνάρτηση $softmax()$.

4.3.2 Προτεινόμενη ενσωμάτωση δεδομένων

Οι μεθοδολογίες που εφαρμόστηκαν για την ενσωμάτωση των δεδομένων στα συστήματα T_2 και A_1 είναι αυτές που περιγράφηκαν στο παρόν κεφάλαιο. Το πρόβλημα με την παραπάνω διαδικασία είναι ο συνδυασμός των ξεχωριστών δεδομένων που τροφοδοτούν τα συστήματα. Όπως έχει ήδη περιγραφεί, με σκοπό την μείωση των διαστάσεων της εισόδου αλλά και για την επαύξηση του συνόλου δεδομένων, τα δεδομένα που ανακλήθηκαν για κάθε τραγούδι διασπάστηκαν σε τμήματα. Συγκεκριμένα, οι στίχοι των τραγουδιών διασπάστηκαν σε τετράστιχα, ενώ τα ηχητικά σήματα διασπάστηκαν σε κλιπς των δέκα δευτερολέπτων.

Το ερώτημα που εγείρεται σε αυτό το σημείο είναι το πως θα καταφέρουν να συνδυαστούν τα τμήματα των δύο μεθόδων αφού δεν υπάρχει σαφής αντιστοιχία. Για την αντιμετώπιση του



Σχήμα 4.4: Δομή του μοντέλου M_1

προβλήματος αυτού αποφασίστηκε ότι τα δεδομένα πρέπει να επανέλθουν στην αρχική τους οργάνωση, για να επιτευχθεί ο συνδυασμός τους. Αυτό που εφαρμόστηκε στην πράξη ήταν πριν τον ταξινομήτη του συστήματος η συγκέντρωση των αντίστοιχων προβλέψεων από τα τετράστιχα και των αντίστοιχων προβλέψεων από τα ηχητικά κλιπς για κάθε τραγούδι.

Στη συνέχεια υπολογίστηκε μια ενιαία τιμή πρόβλεψης για τους στίχους κάθε τραγουδιού, από τον μέσο όρο των προβλέψεων για τα επιμέρους τετράστιχα, και μια ενιαία τιμή πρόβλεψης για το ηχητικό σήμα κάθε τραγουδιού, από τα επιμέρους ηχητικά κλιπς που το συνθέτουν. Τελικά, οι δύο γενικές προβλέψεις για κάθε τραγούδι συνενώνονται και τροφοδοτούνται στον ταξινομητή.

Κεφάλαιο 5

Πειραματική Διαδικασία

Στο κεφάλαιο αυτό θα αναλυθεί η διαδικασία εκπαίδευσης και επαλήθευσης των συστημάτων που περιγράφηκαν στο κεφάλαιο 4. Αφού έχουν καθοριστεί οι αρχιτεκτονικές και ο τρόπος με τον οποίο θα τροφοδοτηθούν τα δεδομένα στα μοντέλα το επόμενο βήμα που πρέπει να ολοκληρωθεί είναι η εκπαίδευση τους. Επί της ουσίας, ο όρος εκπαίδευση αναφέρεται στην επιλογή κατάλληλων τιμών για τις παραμέτρους του συστήματος (βάρη, πολώσεις), με σκοπό την παραγωγή βέλτιστων εξόδων για γνωστά, και κατ' επέκταση άγνωστα, δεδομένα εισόδου. Για την ορθή εκπαίδευση των μοντέλων στην πράξη θα εφαρμοστούν οι αλγόριθμοι Backpropagation (2.3.4.2) και Κατάβασης Κλίσης (2.3.4.3).

Τα δεδομένα σε ένα πρόβλημα επιβλεπόμενης μάθησης, όπως το πρόβλημα που προσπαθεί να αντιμετωπίσει η παρούσα διπλωματική, εμφανίζονται σε ζευγάρια (X, Υ), όπου το X αναπαριστά τα δεδομένα εισόδου ενώ το Υ την τιμή στόχο (label/target) που προσπαθεί να προβλέψει το μοντέλο. Τα εκάστοτε δεδομένα εισόδου για την σωστή εκπαίδευση και επαλήθευση των μοντέλων χωρίζονται σε τρεις κατηγορίες, τα δεδομένα εκπαίδευσης (training data),τα δεδομένα επαλήθευσης (validation data) και τα δεδομένα ελέγχου (testing data). Τα δεδομένα εκπαίδευσης αποτελούν το μεγαλύτερο μέρος των δεδομένων της εισόδου και όπως είναι προφανές από την ονομασία τους είναι τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση ενός μοντέλου. Τα δεδομένα επαλήθευσης συνήθως συνθέτουν το μικρότερο τμήμα των δεδομένων εισόδου και χρησιμοποιούνται για την επαλήθευση των αποτελεσμάτων σε κάθε στάδιο της εκπαίδευσης. Τέλος, τα δεδομένα ελέγχου χρησιμοποιούνται μετά την ολοκλήρωση της εκπαίδευσης του μοντέλου για την καταγραφή της επίδοσης του.

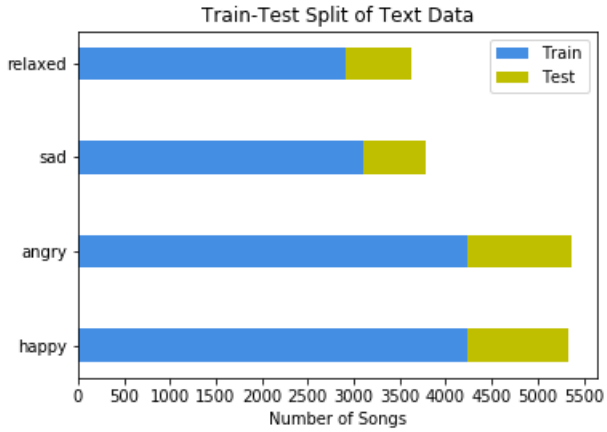
Δύο ακόμη σημαντικοί παράγοντες για την αποτελεσματική εκπαίδευση ενός μοντέλου είναι οι επιλογή των κατάλληλων τιμών για τον αριθμό των εποχών (epochs) και για το μέγεθος τεμαχίου (batch size). Ο αριθμός των εποχών αναφέρεται στον αριθμό των φορών που θα επαναληφθεί η διαδικασία εκπαίδευσης πάνω στο σύνολο των δεδομένων εκπαίδευσης, μέχρι να επιλεγούν οι τελικές τιμές των παραμέτρων του δικτύου. Το μέγεθος τεμαχίου αναφέρεται στον αριθμό των δεδομένων εισόδου που θα χρησιμοποιηθούν ως τεμάχιο για τον υπολογισμό της κλίσης και τη χρήση του για την ανανέωση των παραμέτρων του δικτύου, όπως περιγράφουν οι αλγόριθμοι Backpropagation και Κατάβασης Κλίσης.

Τελικά, μόλις ολοκληρωθεί η εκπαίδευση των μοντέλων το τελικό βήμα αφορά την αξιο-

λόγηση των εξόδων που αποδίδουν. Για την καταγραφή των αποτελεσμάτων θα χρησιμοποιηθούν ορισμένες από τις μετρικές αξιολόγησης, όπως περιγράφονται στην ενότητα 2.7, επί των δεδομένων ελέγχου. Η πειραματική διαδικασία θα αξιολογηθεί κυρίως βάση των μετρικών Accuracy, που περιγράφει των αριθμό των σωστών προβλέψεων προς το σύνολο όλων των προβλέψεων, και Loss, που περιγράφει το σφάλμα της συνάρτησης κόστους εν προκειμένω της Categorical Crossentropy. Για την καλύτερη εποπτεία της επίδοσης στη συνέχεια θα παρουσιαστούν οι τιμές των μετρικών αξιολόγησης αλλά και θα συγκριθούν οι τιμές που υπολογίζονται κατά την εκπαίδευση και την επαλήθευση.

5.1 Συστήματα Ανάλυσης Κειμένου

Στην ενότητα αυτή θα περιγραφεί η πειραματική διαδικασία εκπαίδευσης και επαλήθευσης των συστημάτων ενσωμάτωσης κειμένου με σκοπό την επιλογή εκείνου που θα πετυχαίνει την καλύτερη επίδοση στο πρόβλημα της συναισθηματικής κατηγοριοποίησης μουσικών κομματιών. Η διεξαγωγή των πειραμάτων ξεκινά από το απλό μοντέλο T_1 σε συνδυασμό με διάφορες τεχνικές ανάλυσης κειμένου και καταλήγει στο σύνθετο μοντέλο Tranfer Learning T_2. Τα διαθέσιμα δεδομένα για το συγκεκριμένο πρόβλημα αποτελούνται από 18.115 τετράστιχα εισόδου μαζί με τις αντίστοιχες ετικέτες συναισθήματος, από αυτά τα 80% θα χρησιμοποιηθεί για την εκπαίδευση των συστημάτων ενώ το υπόλοιπο 20% θα χρησιμοποιηθεί ως δεδομένα ελέγχου. Στο σχήμα 5.1 παρουσιάζεται η κατανομή των διαθέσιμων δεδομένων στις τέσσερις συναισθηματικές κλάσεις.



Σχήμα 5.1: Κατανομή διαθέσιμων δεδομένων κειμένου

5.1.1 T_1 με δημοφιλείς μεθοδολογίες ενσωμάτωσης κειμένου

Για την εκπαίδευση του μοντέλου T_1 θα εφαρμοστούν ξεχωριστά οι αναπαραστάσεις που προκύπτουν από τις τέσσερις μεθοδολογίες ενσωμάτωσης κειμένου (BoW, TFIDF, Word2Vec, GloVe), όπως περιγράφονται στην ενότητα 4.1.2. Για την εκπαίδευση του δικτύου χρησιμοποιείται ως συνάρτηση κόστους (loss function) η συνάρτηση *categorical_crossentropy*() (βλέπε

2.35). Ως βελτιστοποιητής χρησιμοποιήθηκε ο Adam με ρυθμό μάθησης (learning rate) ίσο με 0.001 και τις παραμέτρους ορμής beta_1 και beta_2 ρυθμισμένες στα 0.9 και 0.999 αντίστοιχα. Ο αριθμός των εποχών εκπαίδευσης επιλέχθηκε ίσος με 10 και το μέγεθος τεμαχίου ίσο με 128. Για την επιλογή των βέλτιστων υπερπαραμέτρων εκπαίδευσης εφαρμόστηκε η τεχνική της αναζήτησης πλέγματος (grid search).

Από τα δεδομένα εκπαίδευσης το 10% θα χρησιμοποιηθεί ως validation data για την επαλήθευση της εκπαίδευσης. Στη συνέχεια απεικονίζεται η εξέλιξη των τιμών ακρίβειας (Accuracy) και σφάλματος της συνάρτησης κόστους (Loss) κατά τη διαδοχή των εποχών εκπαίδευσης. Στο σχήμα 5.2 απεικονίζεται η σύγκριση των τιμών ακρίβειας εκπαίδευσης και ακρίβειας επαλήθευσης του μοντέλου T_1 για τις τέσσερις μεθοδολογίες ενσωμάτωσης που εφαρμόστηκαν, ενώ στο σχήμα 5.3 συγκρίνονται οι τιμές σφάλματος για τις τέσσερις μεθοδολογίες.

Τέλος στον πίνακα 5.1 παρουσιάζονται, για τις τέσσερις μεθοδολογίες ενσωμάτωσης, οι τελικές τιμές επίδοσης που πετυχαίνουν τα εκπαιδευμένα μοντέλα στα δεδομένα ελέγχου.

Μέθοδος Ενσωμάτωσης	**Loss**	**Accuracy %**
BoW	1.287	65.49
TF-IDF	1.381	67.98
Word2Vec	1.262	41.66
GloVe	1.064	53.33

Πίνακας 5.1: Τιμές επαλήθευσης μοντέλου T_1

5.1.2 T_2 με Bert Embeddings

Για την εκπαίδευση του μοντέλου T_2 θα χρησιμοποιηθούν οι αναπαραστάσεις κειμένων που προκύπτουν από την μεθοδολογία ενσωμάτωσης που απαιτεί το μοντέλο BERT, όπως περιγράφεται στην ενότητα 4.1.2. Για την εκπαίδευση του δικτύου χρησιμοποιείται, παρόμοια με προηγουμένως, ως συνάρτηση κόστους η συνάρτηση $categorical\_crossentropy()$. Ως βελτιστοποιητής χρησιμοποιήθηκε ο Adam με ρυθμό μάθησης ίσο με 0.0005 και τις παραμέτρους ορμής beta_1 και beta_2 ρυθμισμένες στα 0.9 και 0.999 αντίστοιχα. Ο αριθμός των εποχών εκπαίδευσης επιλέχθηκε ίσος με 10 και το μέγεθος τεμαχίου ίσο με 128. Η επιλογή των βέλτιστων υπερπαραμέτρων βασίστηκε πάλι στην αναζήτηση πλέγματος.

Σε αντίθεση με την προηγούμενη περίπτωση, το μοντέλο T_2 είναι ήδη προεκπταιδευμένο στη μορφή που διατίθεται. Η εκπαίδευση του συγκεκριμένου δικτύου ονομάζεται fine-tunning και επικεντρώνεται στην ανανέωση των συναπτικών βαρών και πολώσεων μόνο των τελευταίων επιπέδων, οι τιμές των παραμέτρων για τα υπόλοιπα επίπεδα παραμένουν ως έχουν. Στην περίπτωση του μοντέλου T_2 η διαδικασία fine-tunning θα επηρεάσει μονάχα το τελευταίο από τα δώδεκα επίπεδα που παρέχει το προεκπαιδευμένο μοντέλο, όπως και τα δύο πλήρως συνδεδεμένα επίπεδα του ταξινομητή που έχει προστεθεί στο τέλος του δικτύου.

Για την επαλήθευση της εκπαίδευσης θα χρησιμοποιηθεί, και σε αυτή την περίπτωση, το 10% των δεδομένων εκπαίδευσης. Στα επόμενα γραφήματα (5.4 και 5.5) απεικονίζονται οι

τιμές Accuracy και Loss για τα δεδομένα εκπαίδευσης και επαλήθευσης. Τέλος στον πίνακα 5.2 παρουσιάζονται οι τελικές τιμές επίδοσης που πετυχαίνει το εκπαιδευμένο μοντέλο T_2 σε σύγκριση με το μοντέλο T_1, στα δεδομένα ελέγχου.

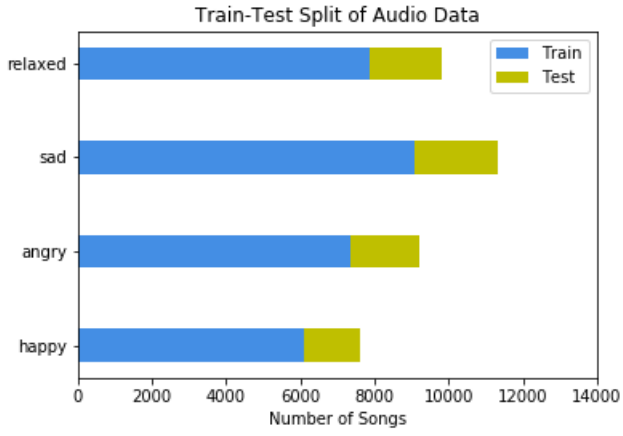
Μοντέλο	**Μέθοδος Ενσωμάτωσης**	**Loss**	**Accuracy %**
T_1'	BoW	1.287	65.49
T_1'	TF-IDF	1.381	67.98
T_1	Word2Vec	1.262	41.66
T_1	GloVe	1.064	53.33
T_2	Bert	1.353	69.11

Πίνακας 5.2: Τιμές επίδοσης μοντέλων T_1 και T_2

5.2 Συστήματα Ανάλυσης Ηχητικού Σήματος

Στην ενότητα αυτή θα περιγραφεί η πειραματική διαδικασία εκπαίδευσης και επαλήθευσης του συστήματος ανάλυσης ηχητικού σήματος. Σκοπός του πειράματος σε αυτό το στάδιο είναι η επιλογή των βέλτιστων υπερπαραμέτρων εκπαίδευσης και η ανίχνευση του συνδυασμού των πινάκων χαρακτηριστικών που πετυχαίνει την καλύτερη επίδοση στο πρόβλημα της συναισθηματικής κατηγοριοποίησης μουσικών κομματιών.

Τα διαθέσιμα δεδομένα για το συγκεκριμένο πρόβλημα αποτελούνται από 37.989 ηχητικά κλιπς των δέκα δευτερολέπτων μαζί με τις αντίστοιχες ετικέτες συναισθήματος, από αυτά τα 80% θα χρησιμοποιηθεί για την εκπαίδευση των μοντέλων ενώ το υπόλοιπο 20% θα χρησιμοποιηθεί ως δεδομένα ελέγχου. Στο σχήμα 5.6 παρουσιάζεται η κατανομή των διαθέσιμων δεδομένων στις τέσσερις συναισθηματικές κλάσεις.



Σχήμα 5.6: Κατανομή διαθέσιμων ηχητικών δεδομένων

5.2.1 A_1 και βέλτιστος συνδυασμός χαρακτηριστικών εισόδου

Για την επιλογή των βέλτιστων υπερπαραμέτρων πριν την εκπαίδευση του μοντέλου A_1 θα εφαρμοστεί και σε αυτή την περίπτωση η αναζήτηση πλέγματος με συνάρτηση κόστους την *categorical_crossentropy*(). Ο Adam επιλέγεται ως βελτιστοποιητής με τις υπερπαραμέτρους learning rate, beta_1 και beta_2 να προκύπτουν από την αναζήτηση πλέγματος ίσες με 0.001, 0.9 και 0.999 αντίστοιχα. Ο αριθμός των εποχών επανάληψης ορίζεται ίσος με 10 και το μέγεθος τεμαχίου ίσο με 16.

Σαν επόμενο βήμα, πριν την τελική εκπαίδευση και επαλήθευση του μοντέλου, πρέπει να επιλεχθεί ο βέλτιστος συνδυασμός των πινάκων χαρακτηριστικών που προκύπτουν από τις μεθοδολογίες επεξεργασίας σήματος (βλέπε ενότητα 4.2.2). Στην πράξη οι πίνακες διάστασης $60x431$ στοιβάζονται παράλληλα, σε διάφορους συνδυασμούς, και τροφοδοτούνται στην είσοδο του δικτύου.

Στον πίνακα 5.3 παρουσιάζονται στην πρώτη στήλη οι συνδυασμοί των πινάκων χαρακτηριστικών που εφαρμόστηκαν στην πράξη, καθώς και η τιμές ακρίβειας που πετυχαίνουν. Όπως φαίνεται ο συνδυασμός και των έξι χαρακτηριστικών, στην τελευταία γραμμή του πίνακα 5.3, είναι αυτός που πετυχαίνει το μεγαλύτερο ποσοστό ακρίβειας και κατά συνέπεια ο συνδυασμός των χαρακτηριστικών που θα χρησιμοποιηθεί για την τελική εκπαίδευση και επαλήθευση του δικτύου A_1.

Συνδυασμός Χαρακτηριστικών	**Accuracy %**
Mel	64.97
Mel, Log-Mel	68.38
Mel, Chroma, Tonnetz, Spectral Contrast	60.86
Log-Mel, Chroma, Tonnetz, Spectral Contrast	58.96
MFCC, Chroma, Tonnetz, Spectral Contrast	65.36
Mel, Log-Mel, MFCC, Chroma, Tonnetz	69.77
Mel, Log-Mel, MFCC, Chroma, Tonnetz, Spectral Contrast	70.34

Πίνακας 5.3: Τιμές επίδοσης πινάκων χαρακτηριστικών

Στην περίπτωση του μοντέλου A_1, πάλι, το 10% των δεδομένων εκπαίδευσης θα χρησιμοποιηθούν για την επαλήθευση της εκπαίδευσης. Στα σχήματα 5.7 και 5.8 απεικονίζονται γραφικά οι επιδόσεις του μοντέλου A_1 για τα δεδομένα εκπαίδευσης και επαλήθευσης στις μετρικές Accuracy και Loss. Στον πίνακα 5.4 παρουσιάζονται οι τελικές τιμές επίδοσης που πετυχαίνει το μοντέλο A_1, στα δεδομένα ελέγχου, σε σύγκριση με τα μοντέλα T_1 και T_2. Για το μοντέλο T_1 επιλέγεται ως μέθοδος ενσωμάτωσης εκείνη με τα μεγαλύτερα ποσοστά ακρίβειας, δηλαδή το TF-IDF.

Μοντέλο	Loss	Accuracy %
T_1'	1.381	67.98
T_2	1.353	69.11
A_1	0.743	70.51

Πίνακας 5.4: Τιμές επίδοσης μοντέλων A_1, T_1 και T_2

5.3 Συνδυασμός Συστημάτων

Στην ενότητα αυτή θα διεξαχθεί μια λίγο διαφορετική διαδικασία εκπαίδευσης από τις προηγούμενες. Για τη σύνθεση του μοντέλου F_1 θα χρησιμοποιηθούν, όπως επισημαίνεται στην ενότητα 4.3, τα συστήματα T_2 και A_1 με την προσθήκη ενός ταξινομήτη, στο τέλος του δικτύου, με σκοπό να ολοκληρωθεί ο συνδυασμός τους. Ωστόσο, η εκπαίδευση του μοντέλου F_1 περιορίζεται αποκλειστικά στην ανανέωση των συναπτικών βαρών και πολώσεων των τελευταίων επιπέδων του δικτύου που συνθέτουν τον ταξινομητή. Τα μοντέλα T_2 και A_1 θα εισαχθούν και θα χρησιμοποιηθούν με έτοιμες παραμέτρους που δεν χρειάζονται εκπαίδευση, όπως έχουν προκύψει από την διαδικασία εκπαίδευσης που περιγράφεται στις ενότητες 5.1 και 5.2.

5.3.1 M_1 και προβλέψεις μουσικών κομματιών

Η δομή του μοντέλου M_1 αποτελείται ουσιαστικά από τα μοντέλα T_2 και A_1 παράλληλα τοποθετημένα στην αρχή του, ενώ την έξοδο του μοντέλου συνθέτει ο κοινός ταξινομητής. Τα μοντέλα T_2 και A_1 τροφοδοτούνται ανεξάρτητα το ένα από το άλλο με όλα τα δεδομένα και παράγουν προβλέψεις. Οι τιμές των προβλέψεων που αποδίδουν είναι αυτές που σε συνδυασμό με τις αντίστοιχες ετικέτες θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου F_1. Οι τιμές των προβλέψεων δεν θα χρησιμοποιηθούν ως έχουν, πριν τροφοδοτηθούν στον ταξινομητή, αλλά θα υπολογιστεί ο μέσος όρος των προβλέψεων που συνθέτουν το κάθε τραγούδι, δηλαδή οι προβλέψεις των επιμέρους τετράστιχων και ηχητικών κλιπς. Επομένως, το τελικό σύστημα F_1 πραγματοποιεί προβλέψεις για κάθε τραγούδι βασιζόμενο και στους στίχους και στο ηχητικό σήμα που το συνθέτουν.

Για την εκπαίδευση του μοντέλου M_1 θα χρησιμοποιηθεί ως συνάρτηση κόστους η categorical crossentropy(). Ο Adam επιλέγεται ως βελτιστοποιητής με τις υπερπαραμέτρους learning rate, beta_1 και beta_2 να προκύπτουν από την αναζήτηση πλέγματος ίσοι με 0.001, 0.9 και 0.999 αντίστοιχα. Ο αριθμός των εποχών επανάληψης ορίζεται ίσος με 10 και το μέγεθος τεμαχίου ίσο με 16.

Οι προβλέψεις που προκύπτουν από τις εξόδους των μοντέλων T_2 και A_1 θα διασπαστούν σε δεδομένα εκπαίδευσης (80%) και δεδομένα ελέγχου (20%). Για την επαλήθευση της εκπαίδευσης θα χρησιμοποιηθεί, και σε αυτή την περίπτωση, το 10% των δεδομένων εκπαίδευσης. Στα επόμενα γραφήματα 5.9 και 5.10 απεικονίζονται οι τιμές Accuracy και Loss για τα δεδομένα εκπαίδευσης και επαλήθευσης. Στον πίνακα 5.5 παρουσιάζονται οι τελικές τιμές επίδοσης

που πετυχαίνει το εκπαιδευμένο μοντέλο M_1 σε σύγκριση με τα προηγούμενα μοντέλα.

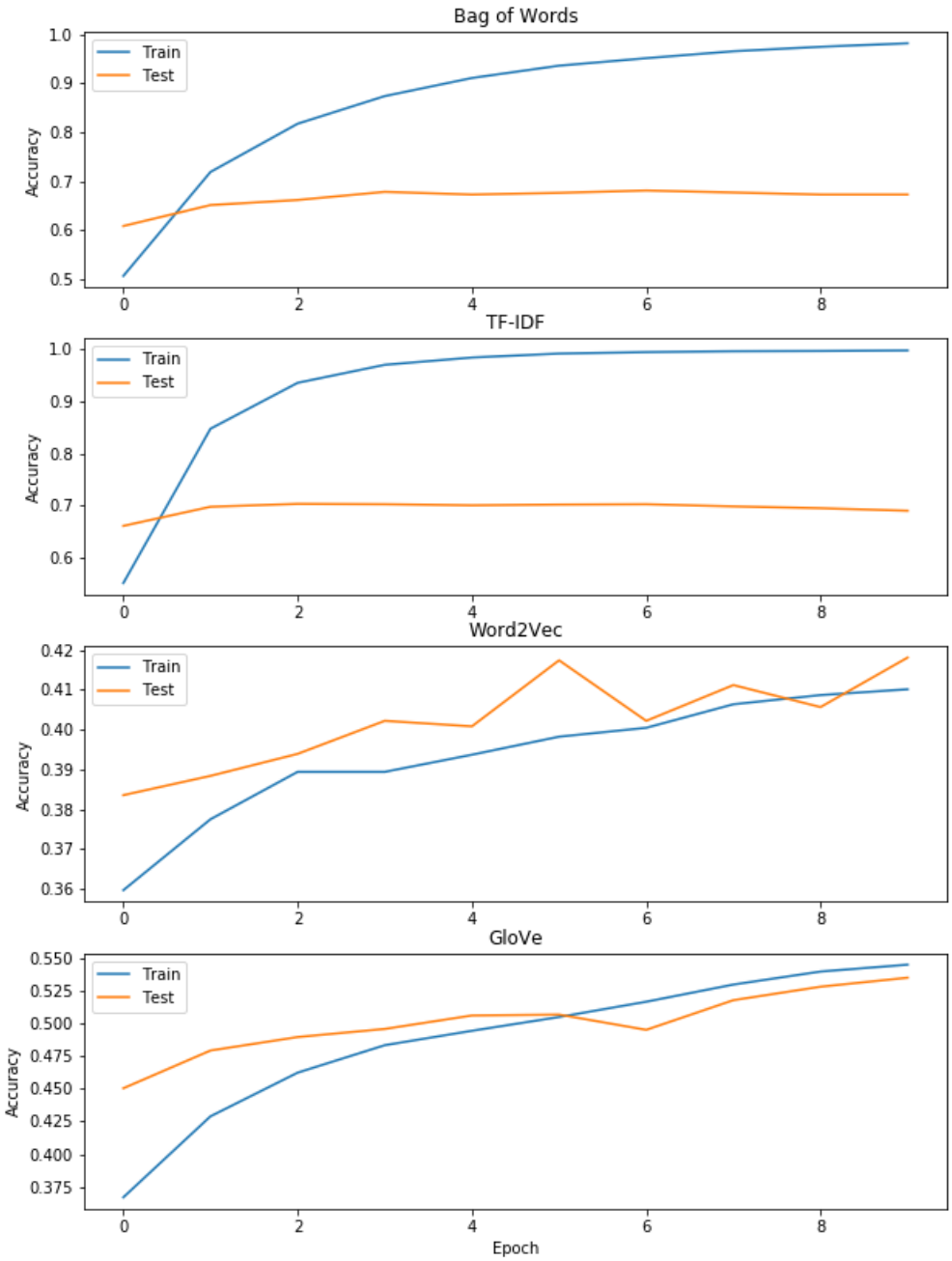
Τέλος, η πειραματική διαδικασία ολοκληρώνεται με την αξιολόγηση των μοντέλων σε ένα σύνολο 200 τραγουδιών που κατασκευάσαμε χειροκίνητα, 50 τραγούδια σε κάθε κατηγορία, από μουσικές λίστες που ανακτήσαμε από το διαδίκτυο. Στον πίνακα 5.6 εμφανίζονται οι τιμές επίδοσης Accuracy, Loss των εκπαιδευμένων μοντέλων T_2, A_1, M_1.

Μοντέλο	**Loss**	**Accuracy %**
$T_1^{'}$	1.381	67.98
T_2	1.353	69.11
A_1	0.743	70.51
M_1	0.156	94.58

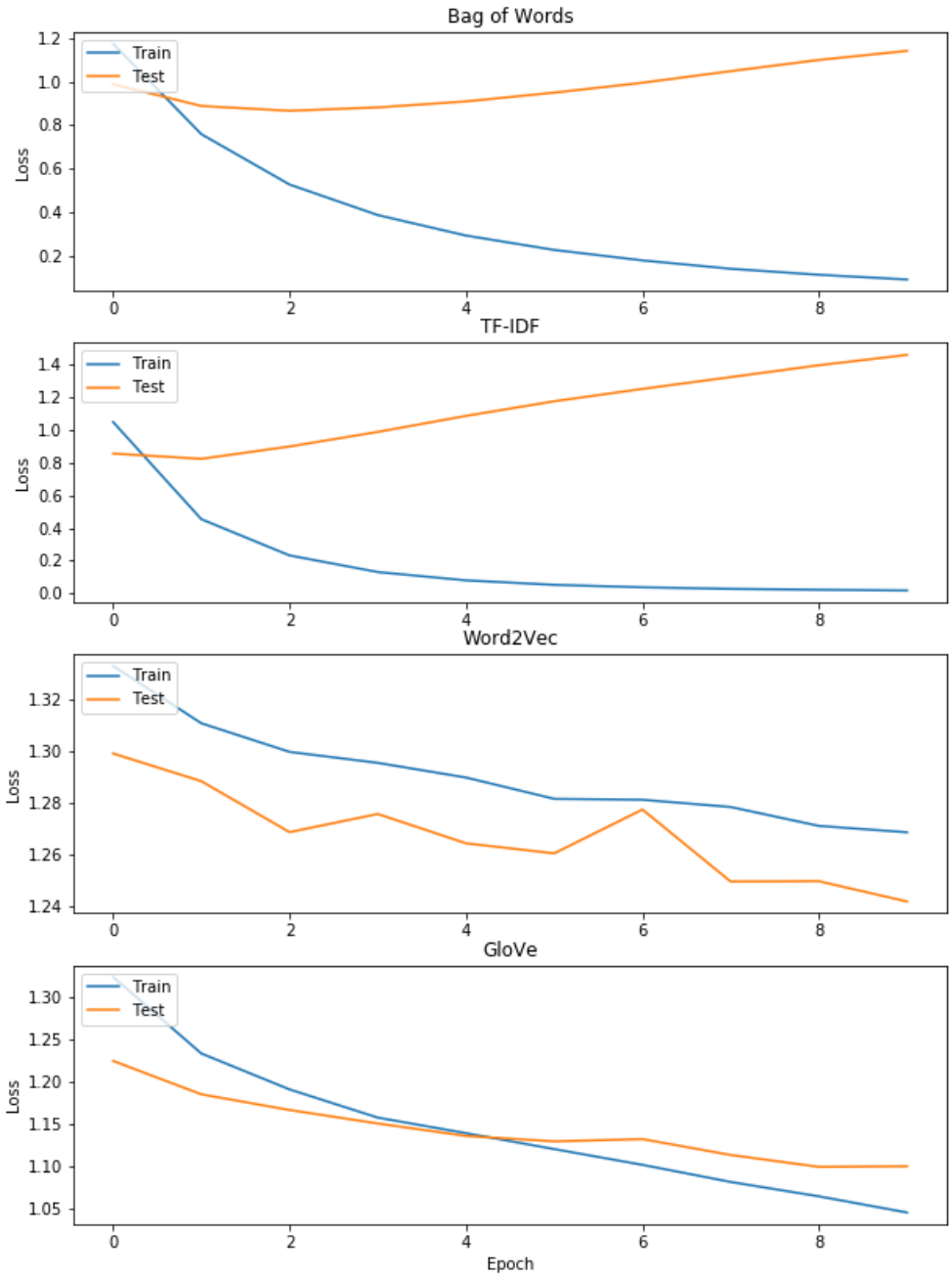
Πίνακας 5.5: Τιμές επίδοσης μοντέλων M_1, A_1, T_1 και T_2

Μοντέλο	**Loss**	**Accuracy %**
T_2	2.583	49.71
A_1	1.226	48.60
M_1	1.184	57.79

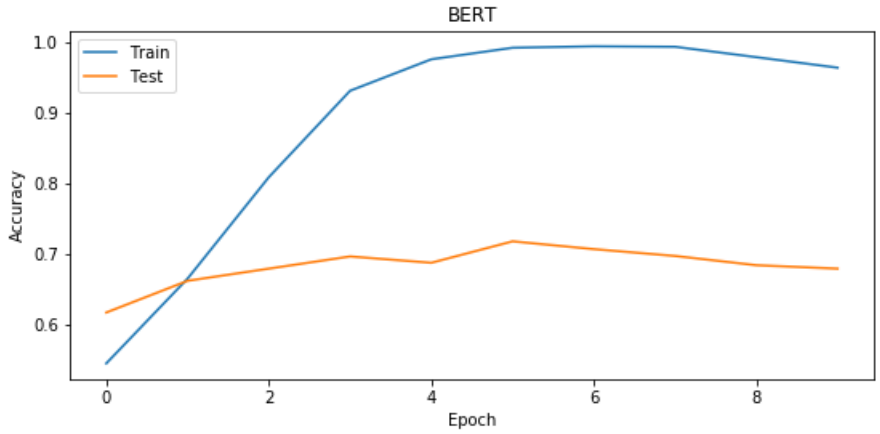
Πίνακας 5.6: Τιμές επίδοσης μοντέλων M_1, A_1, T_1 και T_2 σε ανεξάρτητο σύνολο δεδομένων



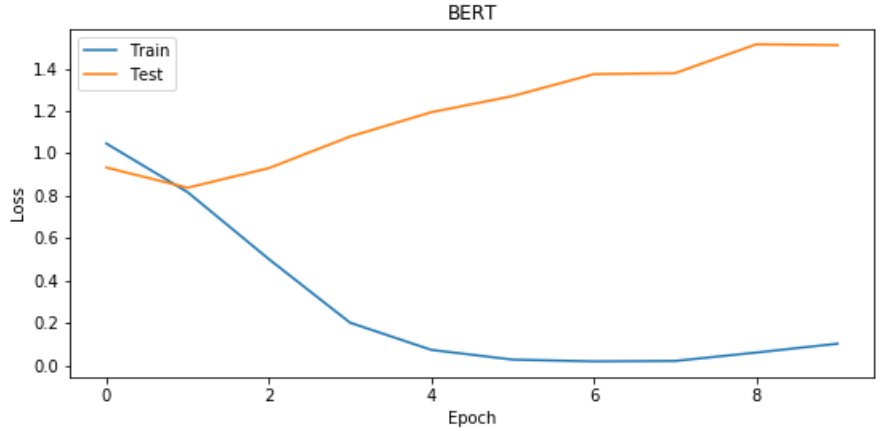
Σχήμα 5.2: Accuracy του T_1



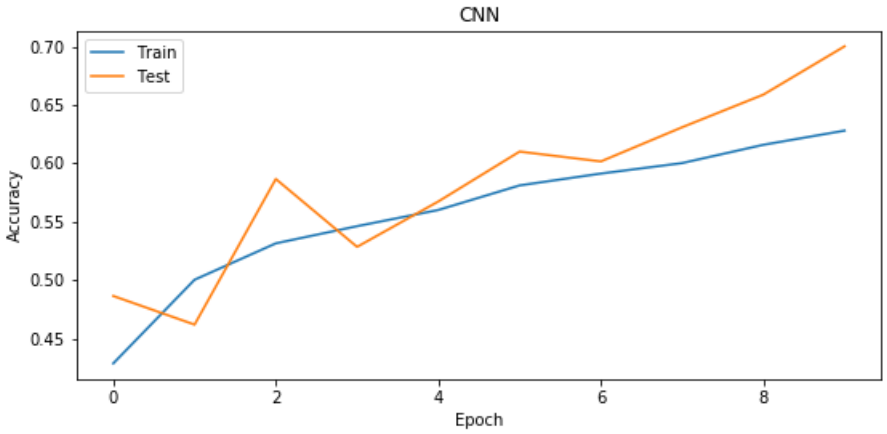
Σχήμα 5.3: Loss του T_1



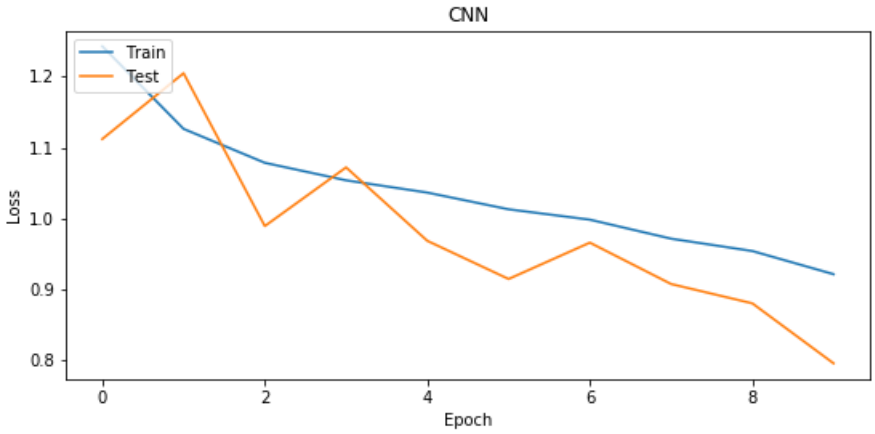
Σχήμα 5.4: Accuracy του T_2



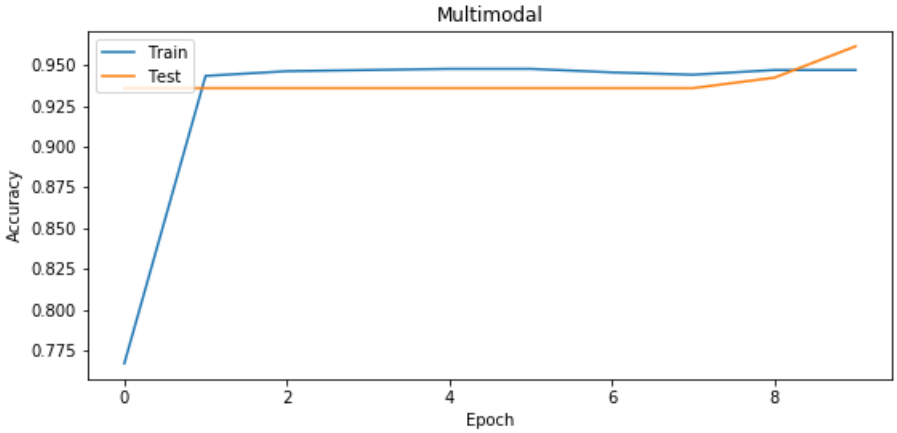
Σχήμα 5.5: Loss του T_2



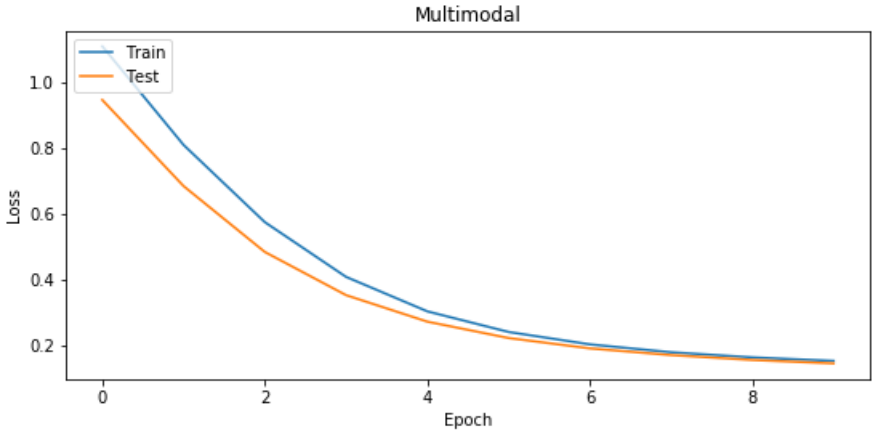
Σχήμα 5.7: Accuracy του A_1



Σχήμα 5.8: Loss του A_1



Σχήμα 5.9: Accuracy του M_1



Σχήμα 5.10: Loss του M_1

Κεφάλαιο 6

Συμπεράσματα και Προτάσεις

Στα προηγούμενα κεφάλαια έγινε μια προσπάθεια προσέγγισης του προβλήματος της αναγνώριση συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής, εφαρμόζοντας τις αρχές και τη λογική της επιβλεπόμενης μάθησης. Παρουσιάστηκαν υλοποιήσεις συστημάτων βαθιάς μηχανικής μάθησης σε συνδυασμό με κάποιες μεθοδολογίες ενσωμάτωσης δεδομένων βασισμένες σε τεχνικές επεξεργασίας φυσικής γλώσσας και ψηφιακής επεξεργασίας σήματος. Η διαδικασία ολοκληρώθηκε με την εκπαίδευση και την αξιολόγηση των συστημάτων που υλοποιήθηκαν για κάθε μια από τις διαφορετικές προσεγγίσεις. Στο παρόν κεφάλαιο συνοψίζεται το έργο που επιτελέσθηκε στη προκείμενη διπλωματική εργασία και καταγράφονται τα αποτελέσματα που εξήχθησαν, δίνοντας στη συνέχεια προτάσεις για μελλοντική έρευνα και ανάπτυξη.

6.1 Συμπεράσματα

Για την ανάπτυξη ενός συστήματος Βαθιάς Μηχανικής Μάθησης που θα πετυχαίνει μεγάλες επιδόσεις στο πρόβλημα της αναγνώρισης συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής, πραγματοποιήθηκε σε πρώτο στάδιο εκτενής μελέτη παρεμφερών εργασιών και ερευνών με σκοπό την επιλογή των κατευθύνσεων δράσης. Από τη μελέτη αυτή προέκυψε ότι για την ανάπτυξη ενός τέτοιου συστήματος η εφαρμογή ενός συνδυασμού ανάλυσης στίχων και ηχητικού σήματος είναι σε θέση να πετύχει καλύτερες επιδόσεις σε σχέση με τα συστήματα που αξιοποιούν ξεχωριστά κάθε μια από αυτές τις αναλύσεις. Με γνώμονα τα παραπάνω πραγματοποιήθηκε ξεχωριστή διαχείριση των στίχων και του ηχητικού σήματος με σκοπό την βέλτιστη επιλογή αρχιτεκτονικών και αναπαραστάσεων δεδομένων, όπως περιγράφεται στην πειραματική διαδικασία του κεφαλαίου 5.

Η πρώτη προσπάθεια μελέτης που πραγματοποιήθηκε αφορούσε την δημιουργία ενός συστήματος που θα διαχειρίζεται στίχους. Από αυτή την μελέτη προέκυψαν τρεις αρχιτεκτονικές $(T_1^{'}, T_1, T_2)$ και πέντε μεθοδολογίες αναπαράστασης κειμένου. Πιο συγκεκριμένα, για τον απλό ταξινομητή $T_1^{'}$ εφαρμόστηκαν οι μέθοδοι ενσωμάτωσης Bag of Words και TF-IDF με αρκετά ικανοποιητικά επίπεδα επίδοσης. Οι μέθοδοι ενσωμάτωσης Word2Vec και GloVe που αξιοποιούν και την σειρά εμφάνισης των λέξεων, τροφοδότησαν το επαναλαμβανόμενο δίκτυο

T_1 που τελικά δεν κατάφερε να πετύχει υψηλά επίπεδα επίδοσης, πιθανώς λόγο αυξημένης πολυπλοκότητας. Εν τέλει, όπως φαίνεται και από τον πίνακα 5.2, το μοντέλο Tranfer Learning BERT (T_2) είναι αυτό κατάφερε να επιλύσει αποτελεσματικότερα το πρόβλημα ταξινόμησης, αξιοποιώντας τους στίχους. Από τη διαδικασία που περιγράφηκε προκύπτει το συμπέρασμα πως αποτελεσματικότερη για το συγκεκριμένο πρόβλημα φάνηκε η χρήση της λογικής του Tranfer Learning και των Trasnformers, που εφαρμόζει το μοντέλο BERT, σε συνδυασμό με την μέθοδο ενσωμάτωσης που προτείνει.

Δύο βασικοί λόγοι που οδηγούν στα υψηλά επίπεδα απόδοσης και τελικά στην επιλογή αυτού του μοντέλου, είναι οι σύνθετοι μηχανισμοί προσοχής και ο μεγάλος όγκος δεδομένων που εφαρμόζει το BERT για την εκπαίδευση του. Ένα ακόμα συμπέρασμα της πειραματικής διαδικασίας είναι η αύξηση της επίδοσης και της ταχύτητας εκπαίδευσης που προκύπτουν έπειτα από την τεμαχιοποίηση των στίχων σε τετράστιχα.

Η δεύτερη διαδικασία μελέτης σχετίζεται με την αξιοποίηση του ηχητικού σήματος για την ανάπτυξη ενός συστήματος ταξινόμησης, από την οποία προέκυψε το μοντέλο CNN A_1. Αντικείμενο μελέτης σε αυτό το βήμα ήταν η εξαγωγή των κατάλληλων χαρακτηριστικών ήχου και, αφού τροφοδοτηθούν στο μοντέλο, η επιλογή του συνδυασμού αυτών που θα πετυχαίνουν την καλύτερη επίδοση.

Τα χαρακτηριστικά που αποφάσισαν να εξαχθούν, για κάθε ηχητικό κλιπ είναι τα Mel Spectogram, Log-Mel Spectogram, MFCC, Chroma, Tonnetz και Spectral Contrast. Τα χαρακτηριστικά αυτά ενσωματώθηκαν σε μορφή πινάκων και στην συνέχεια συνδυάστηκαν με διάφορους τρόπους για να τροφοδοτήσουν το μοντέλο A_1. Από τον πίνακα 5.3 προκύπτει το συμπέρασμα πως όταν γίνει χρήση όλων των πινάκων χαρακτηριστικών για την εκπαίδευση του δικτύου, τότε πετυχαίνουμε τα υψηλότερα ποσοστά επίδοσης. Ακόμα, από τον πίνακα 5.4 παρατηρείται πως το μοντέλο αξιοποίησης των ηχητικών χαρακτηριστικών πετυχαίνει μια μικρή βελτίωση σε σχέση με το μοντέλο διαχείρισης στίχων T_2.

Η φύση των CNNs επιτρέπουν την αναγνώριση περιοχών στους πίνακες χαρακτηριστικών, υπεύθυνες για τις τιμές σθένους και διέγερσης των τραγουδιών. Αξιοσημείωτη βελτίωση στην ταχύτητα και την επίδοση του μοντέλου συντέλεσε και αυτή τη φόρα η τεμαχιοποίηση των δεδομένων, δηλαδή των ηχητικών αρχείων σε κλιπς διάρκειας δέκα δευτερολέπτων.

Η τελική μελέτη της διπλωματικής εργασίας επικεντρώθηκε στον τρόπο με τον οποίο θα καταφέρουν να συνδυαστούν τα μοντέλα T_2 και A_1 ώστε να αποδώσουν ένα κοινό αποτέλεσμα. Η απόφαση που πάρθηκε για τον σκοπό αυτό ήταν τα μοντέλα να χρησιμοποιηθούν προεκπαιδευμένα, από την διαδικασία που περιγράφεται στο κεφάλαιο 5, και να εφαρμοστεί ένας ταξινομήτης ο οποίος θα δέχεται τις εξόδους τους και θα πραγματοποιεί μια κοινή πρόβλεψη. Το πρόβλημα που αντιμετωπίσαμε σε αυτή την διαδικασία ήταν ότι δεν υπήρχε καμία άμεση αντιστοιχία ανάμεσα στα τεμάχια των ηχητικών κλιπς και των στίχων. Για την επίλυση αυτού αποφασίστηκε, πριν την τροφοδότηση των προβλέψεων στον ταξινομητη, ο υπολογισμός δύο προβλέψεων για κάθε τραγούδι. Η πρώτη ως ο μέσος όρος των επιμέρους προβλεψεων από τα τετράστιχα που αντιστιχούν στο κάθε τραγούδι και ο δεύτερος ως ο μέσος όρος των ηχητικών κλιπς.

Από τον πίνακα 5.5 συμπεραίνουμε ότι ο συνδυασμός των δυο συστημάτων πετυχαίνει

τεραστία βελτίωση στην επίδοση του συστήματος M_1. Το αποτέλεσμα αυτό αφενός προκύπτει από την αξιοποίηση των στίχων και του ηχητικού σήματος, αφετέρου από τον υπολογισμό μιας συγκεντρωμένης πρόβλεψης για κάθε τραγούδι, από την αξιοποίηση όλων των επιμέρους προβλέψεων που πραγματοποιούν τα μοντέλα T_2 και A_1.

6.2 Μελλοντικές Επεκτάσεις και Προτάσεις

Ένας από τους σημαντικότερους παράγοντες στην αποτελεσματική ανάπτυξη ενός συστήματος Βαθιάς Μηχανικής Μάθησης είναι το μέγεθος του συνόλου δεδομένων που θα χρησιμοποιηθούν. Για το πρόβλημα επιβλεπόμενης μάθησης της συναισθηματικής κατηγοριοποίησης που προσπαθήσαμε να επιλύσουμε, το μεγαλύτερο dataset που μπορέσαμε να εντοπίσουμε αποτελούνταν από ένα σύνολο 2.000 τραγουδιών, αρκετά μικρό για τις απαιτήσεις ενός συστήματος βαθιάς μάθησης.

Έτσι, ένας πρώτος στόχος επέκτασης της εργασίας αυτής θα ήταν η αναζήτηση ή ακόμα και η κατασκευή ενός αρκετά μεγαλύτερου συνόλου δεδομένων κατάλληλο για την συναισθηματική κατηγοριοποίηση τραγουδιών.

Μια άλλη αντιμετώπιση του προβλήματος αυτού θα μπορούσε να είναι η εκπαίδευση ενός μοντέλου μη-επιβλεπόμενης μάθησης. Σε αυτή την περίπτωση τα δεδομένα δεν είναι απαραίτητο να συνοδεύονται από ετικέτες (labels) και μπορούν να ανακτηθούν σε μεγάλους όγκους. Επίσης όσο αφορά τα δεδομένα, ιδιαίτερο ενδιαφέρον θα εμφάνιζε η αναζήτηση και η αξιοποίηση δεδομένων που θα εμφανίζουν συγχρονισμό στίχων και ηχητικού σήματος.

Μια ακόμα κατεύθυνση για μελλοντική ανάπτυξη θα μπορούσε να αποτελέσει η καλύτερη διαχείριση του ηχητικού σήματος. Ένα παράδειγμα είναι η καλύτερη επιλογή των παραμέτρων των μετασχηματισμών που εφαρμόστηκαν, όπως το μήκος παραθύρου, το ποσοστό επικάλυψης, η επιλογή του αριθμού των Mel μπαντών κ.α.. Ένα δεύτερο παράδειγμα, αξιοποίησης, του ηχητικού σήματος θα ήταν η εξαγωγή περαιτέρω χαρακτηριστικών και μια πιο συγκεκριμένη μελέτη πάνω στον βέλτιστο συνδυασμό που θα μπορούσε να εφαρμοστεί.

Τέλος, εφικτή θα ήταν η αξιοποίηση και η εφαρμογή του συστήματος που αναπτύχθηκε στα πλαίσια αυτής της διπλωματικής εργασίας και σε άλλα προβλήματα. Παραδείγματος χάρη, η χρήση των συστημάτων και τεχνικών που εφαρμόστηκαν και σε προβλήματα συναισθηματικής ανάλυσης πέρα από την μουσική. Μια άλλη χρήση του συστήματος θα μπορούσε να είναι η εκπαίδευση και εφαρμογή του σε άλλα προβλήματα που απασχολούν το ερευνητικό πεδίο MIR όπως είναι η αναγνώριση του μουσικού είδους. Μια τελευταία αξιοποίηση του συστήματος, αν και λίγο πιο αφηρημένη, θα μπορούσε να είναι η χρήση σε εφαρμογές που απασχολούν τους τομείς Augmented Reality και Virtual Reality.

Βιβλιογραφία

[1] Mridul Agarwal and Vaneet Aggarwal. *A Reinforcement Learning Based Approach for Joint Multi-Agent Decision Making*. 2019. arXiv: 1909.02940 [cs.LG].

[2] Geoffrey E. Hinton Alex Krizhevsky Ilya Sutskever. "ImageNet classification with deep convolutional neural networks". In: *Communication of the ACM* 60 (2012), pp. 84–90. DOI: 10.1145/3065386.

[3] Mehdi Allahyari et al. *Text Summarization Techniques: A Brief Survey*. 2017. arXiv: 1707.02268 [cs.CL].

[4] Roberto Battiti. "First- and Second-Order Methods for Learning: Between Steepest Descent and Newton's Method". In: *Neural Computation* 4 (Mar. 1992). DOI: 10.1162/neco.1992.4.2.141.

[5] Steffen Bickel. *ECML-PKDD Discovery Challenge 2006 Overview*.

[6] Robin Brochier, Adrien Guille, and Julien Velcin. "Global Vectors for Node Representations". In: *The World Wide Web Conference on - WWW '19* (2019). DOI: 10.1145/3308558.3313595. URL: http://dx.doi.org/10.1145/3308558.3313595.

[7] Gordon Bruner. "Music, Mood, and Marketing". In: *Journal of Marketing* 54 (Oct. 1990), pp. 94–104. DOI: 10.2307/1251762.

[8] Christophe Cerisara, Pavel Král, and Ladislav Lenc. "On the Effects of Using word2vec Representations in Neural Networks for Dialogue Act Recognition". In: *Computer Speech Language* 47 (July 2017). DOI: 10.1016/j.csl.2017.07.009.

[9] "Characterization of the Affective Norms for English Words by discrete emotional categories". In: 2012. URL: https://doi.org/10.3758/s13428-011-0131-7.

[10] Tongfei Chen et al. *Uncertain Natural Language Inference*. 2019. arXiv: 1909.03042 [cs.CL].

[11] Richard Cohn. "Introduction to Neo-Riemannian Theory: A Survey and a Historical Perspective". In: *Journal of Music Theory* 42.2 (1998), pp. 167–180. ISSN: 00222909. URL: http://www.jstor.org/stable/843871.

[12] T.N. Wiesel D.H. Hubel. "Receptive fields and functional architecture of monkey striate cortex". In: *J Physiol* 195 (Mar. 1968), pp. 215–243. DOI: 10.1113/jphysiol.1968.sp008455.

[13] Ronald J. Williams David E. Rumelhart Geoffrey E. Hinton. "Learning representations by back-propagating errors". In: *Nature* 323 (Oct. 1986), pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0. URL: https://doi.org/10.1038/323533a0.

[14] Rémi Delbouys et al. "Music Mood Detection Based on Audio and Lyrics with Deep Neural Net". In: *ISMIR*. 2018.

[15] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv: 1810.04805 [cs.CL].

[16] Chuong B. Do and Andrew Y. Ng. "Transfer learning for text classification". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, B. Schölkopf, and J. C. Platt. MIT Press, 2006, pp. 299–306. URL: http://papers.nips.cc/paper/2843-transfer-learning-for-text-classification.pdf.

[17] Jeffrey L. Elman. "Finding Structure in Time". In: *Cognitive Science* 14.2 (1990), pp. 179–211. DOI: 10.1207/s15516709cog1402_1. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1402_1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1.

[18] Maurizio Moriso Erion Cano. "MoodyLyrics: A Sentiment Annotated Lyrics Dataset". In: 2017. DOI: 10.1145/3059336.3059340.

[19] Kawin Ethayarajh. *How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings*. 2019. arXiv: 1909.00512 [cs.CL].

[20] Dr.Kunihiko Fukushima. "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position". In: *Biological Cybermetrics* 36 (1980), pp. 193–202. DOI: doi:10.1007/BF00344251.

[21] Alf Gabrielsson and Erik Lindström. "The influence of musical structure on emotional expression". In: 2001.

[22] Yoav Goldberg and Omer Levy. *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*. 2014. arXiv: 1402.3722 [cs.CL].

[23] Cyril Goutte and Éric Gaussier. "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation". In: *ECIR*. 2005.

[24] Dani Gunawan, C Sembiring, and Mohammad Budiman. "The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents". In: *Journal of Physics: Conference Series* 978 (Mar. 2018), p. 012120. DOI: 10.1088/1742-6596/978/1/012120.

[25] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://doi.org/10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[26] J. J. Hopfield. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities". In: *Proceedings of the National Academy of Sciences of the United States of America* 79.8 (1982), pp. 2554–2558. ISSN: 00278424. URL: http://www.jstor.org/stable/12175.

[27] Xiao Hu, Kahyun Choi, and J. Downie. "A framework for evaluating multimodal music mood classification". In: *Journal of the Association for Information Science and Technology* 68 (Feb. 2016), n/a–n/a. DOI: 10.1002/asi.23649.

[28] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *CoRR* abs/1502.03167 (2015). arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

[29] Michael I. Jordan. "Chapter 25 - Serial Order: A Parallel Distributed Processing Approach". In: *Neural-Network Models of Cognition*. Ed. by John W. Donahoe and Vivian Packard Dorsel. Vol. 121. Advances in Psychology. North-Holland, 1997, pp. 471–495. DOI: https://doi.org/10.1016/S0166-4115(97)80111-2. URL: http://www.sciencedirect.com/science/article/pii/S0166411597801112.

[30] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. arXiv: 1412.6980 [cs.LG].

[31] Alexander Hanbo Li and Abhinav Sethy. *Knowledge Enhanced Attention for Robust Natural Language Inference*. 2019. arXiv: 1909.00102 [cs.CL].

[32] Bofang Li et al. "Weighted Neural Bag-of-n-grams Model: New Baselines for Text Classification". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1591–1600. URL: https://www.aclweb.org/anthology/C16-1150.

[33] Thomas Lidy and Alexander Schindler. "Parallel Convolutional Neural Networks for Music Genre and Mood Classification". In: Aug. 2016.

[34] Qun Luo, Weiran Xu, and Jun Guo. "A Study on the CBOW Model's Overfitting and Stability". In: *International Conference on Information and Knowledge Management, Proceedings* 2014 (Nov. 2014), pp. 9–12. DOI: 10.1145/2663792.2663793.

[35] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. *Effective Approaches to Attention-based Neural Machine Translation*. 2015. arXiv: 1508.04025 [cs.CL].

[36] Agnes Lydia and Sagayaraj Francis. "Adagrad - An Optimizer for Stochastic Gradient Descent". In: (May 2019).

[37] D. S. Maitra, U. Bhattacharya, and S. K. Parui. "CNN based common approach to handwritten character recognition of multiple scripts". In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. Aug. 2015, pp. 1021–1025. DOI: 10.1109/ICDAR.2015.7333916.

[38] Victor Makarenkov, Bracha Shapira, and Lior Rokach. *Language Models with Pre-Trained (GloVe) Word Embeddings*. 2016. arXiv: 1610.03759 [cs.CL].

[39] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (Dec. 1943), pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259. URL: https://doi.org/10.1007/BF02478259.

[40] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (Dec. 1943), pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259. URL: https://doi.org/10.1007/BF02478259.

[41] Tomas Mikolov et al. *Distributed Representations of Words and Phrases and their Compositionality*. 2013. arXiv: 1310.4546 [cs.CL].

[42] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL].

[43] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press, 1969.

[44] Thomas Mitchell. *Machine learning*. McGraw-Hill Pub. Co. (ISE Editions), 1997.

[45] Meinard Müller, Frank Kurth, and Michael Clausen. "Audio Matching via Chroma-Based Statistical Features." In: Jan. 2005, pp. 288–295.

[46] Feiping Nie, Hu Zhanxuan, and Xuelong Li. "An investigation for loss functions widely used in machine learning". In: *Communications in Information and Systems* 18 (Jan. 2018), pp. 37–52. DOI: 10.4310/CIS.2018.v18.n1.a2.

[47] Ann Nowe, Peter Vrancx, and Yann-Michaël De Hauwere. "Game Theory and Multi-agent Reinforcement Learning". In: Jan. 2012, p. 30. ISBN: 978-3-642-27645-3. DOI: 10.1007/978-3-642-27645-3_14.

[48] Yannis Papanikolaou, Ian Roberts, and Andrea Pierleoni. *Deep Bidirectional Transformers for Relation Extraction without Supervision*. 2019. arXiv: 1911.00313 [cs.LG].

[49] Razvan Pascanu et al. *How to Construct Deep Recurrent Neural Networks*. 2013. arXiv: 1312.6026 [cs.NE].

[50] Geoffroy Peeters. "A Generic Training and Classification System for MIREX08 Classification Tasks: Audio Music Mood, Audio Genre, Audio Artist and Audio Tag". In: 2008.

[51] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[52] Matthew E. Peters et al. *Deep contextualized word representations*. 2018. arXiv: 1802.05365 [cs.CL].

[53] L.Y. Pratt. "Discriminability-based transfer between neural networks". In: Advances in Neural Information Processing Systems (1994), pp. 204–211. URL: https://books.google.gr/books?id=6tGHlwEACAAJ.

[54] Shahzad Qaiser and Ramsha Ali. "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents". In: *International Journal of Computer Applications* 181 (July 2018). DOI: 10.5120/ijca2018917395.

[55] Zhenyue Qin and Dongwoo Kim. *Softmax Is Not an Artificial Trick: An Information-Theoretic View of Softmax in Neural Networks*. 2019. arXiv: 1910.02629 [cs.LG].

[56] F. Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain". In: *Psychological Review* (1958), pp. 65–386.

[57] J.A. Russell. "A circumplex model of affect". In: *Journal of personality and social psychology* 39.6 (1980), pp. 1161–1178. ISSN: 0022-3514.

[58] J. Volkmann S. S. Stevens and E. B. Newman. "A Scale for the Measurement of the Psychological Magnitude Pitch". In: *Journal of the Acoustical Society of America* 8 (3 1937), pp. 185–190. DOI: http://dx.doi.org/10.1121/1.1915893.

[59] Arthur L. Samuell. "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3 (1959), pp. 210–229.

[60] Alex Sherstinsky. *Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network*. 2018. arXiv: 1808.03314 [cs.LG].

[61] David Silver et al. *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*. 2017. arXiv: 1712.01815 [cs.AI].

[62] Yu Su et al. "Performance analysis of multiple aggregated acoustic features for environment sound classification". In: *Applied Acoustics* 158 (2020), p. 107050. ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2019.107050. URL: http://www.sciencedirect.com/science/article/pii/S0003682X19302701.

[63] Marco Susino and Emery Schubert. "Cross-cultural anger communication in music: Towards a stereotype theory of emotion in music". In: *Musicae Scientiae* 21.1 (2017), pp. 60–74. DOI: 10.1177/1029864916637641. eprint: https://doi.org/10.1177/1029864916637641. URL: https://doi.org/10.1177/1029864916637641.

[64] George Tzanetakis. "MARSYAS SUBMISSIONS TO MIREX 2007". In: 2007.

[65] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[66] Ashish Vaswani et al. "Attention is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964.

[67] Yonghui Wu et al. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 2016. arXiv: 1609.08144 [cs.CL].

[68] Menno Zaanen and Pieter Kanters. "Automatic Mood Classification Using TF*IDF Based on Lyrics." In: Jan. 2010, pp. 75–80.

[69] Matthew D. Zeiler. *ADADELTA: An Adaptive Learning Rate Method*. 2012. arXiv: 1212.5701 [cs.LG].

[70] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. "Understanding bag-of-words model: A statistical framework". In: *International Journal of Machine Learning and Cybernetics* 1 (Dec. 2010), pp. 43–52. DOI: 10.1007/s13042-010-0001-0.

6.2.0.0.0.1