ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΓΝΩΡΙΣΗ ΑΚΟΥΣΤΙΚΗΣ ΣΚΗΝΗΣ ΣΕ ΠΕΡΙΒΑΛΛΟΝΤΑ ΠΟΛΗΣ

Διπλωματική Εργασία

Ευθύμης-Πανορμίτης Κλάδης

Επιβλέπων: Γεράσιμος Ποταμιάνος

Βόλος 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΓΝΩΡΙΣΗ ΑΚΟΥΣΤΙΚΗΣ ΣΚΗΝΗΣ ΣΕ ΠΕΡΙΒΑΛΛΟΝΤΑ ΠΟΛΗΣ

Διπλωματική Εργασία

Ευθύμης-Πανορμίτης Κλάδης

Επιβλέπων: Γεράσιμος Ποταμιάνος

Βόλος 2021

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

ACOUSTIC SCENE RECOGNITION IN URBAN ENVIRONMENTS

Diploma Thesis

Efthimis-Panormitis Kladis

Supervisor: Potamianos Gerasimos

Volos 2021

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων **Γεράσιμος Ποταμιάνος**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Μπέλλας Νικόλαος**
Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Τσαλαπάτα Χαρίκλεια**
Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ημερομηνία έγκρισης: 8-7-2021

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Γεράσιμο Ποταμιάνο για τις υποδείξεις και την καθοδήγηση που μου προσέφερε, με αποτέλεσμα να ολοκληρώσω επιτυχώς την διπλωματική μου εργασία, αλλά και να εμπλουτίσω το γνωστικό μου υπόβαθρο.

Επίσης, ένα μεγάλο ευχαριστώ σε όλους τους καθηγητές του τμήματος, οι οποίοι είχαν τεράστια συνεισφορά όλα αυτά τα χρόνια των σπουδών μου στην άνθηση του πνευματικού υποβάθρου και στην διαμόρφωση του τρόπου σκέψης μου ως μηχανικού.

Τέλος, είμαι κάτι παραπάνω από ευγνώμων στην οικογενειά μου για την κάθε είδους υποστήριξη που μου προσέφερε όλα αυτά τα χρόνια, καθώς και στα φιλικά μου πρόσωπα για την οποιαδήποτε βοήθεια ή συμβουλή με οδήγησε να εξελιχθώ ώς Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο/Η Δηλών/ούσα

Ευθύμης-Πανορμίτης Κλάδης

8-7-2021

Περίληψη

Τα προβλήματα ταξινόμησης είναι μερικά από τα πιο συνηθισμένα στον τομέα της μηχανικής μάθησης. Η αναγνώριση ακουστικής σκηνής ή η ταξινόμηση ήχου, όπως αναφέρεται από πολλούς, καθιστά ένα αναδυόμενο θέμα τα τελευταία χρόνια. Ένα από τα πιο σημαντικά ζητήματα ταξινόμησης ήχου είναι η αναγνώριση σκηνών και πηγών ήχου από περιβάλλοντα όπως το αστικό, το οποίο είναι και το αντικείμενο αυτής της διπλωματικής εργασίας. Τα πιο επιτυχημένα μοντέλα ταξινόμησης ήχου αποτελούνται συνήθως από μία ή περισσότερες τεχνικές εξαγωγής χαρακτηριστικών (feature extraction) και μια ή περισσότερες αρχιτεκτονικές νευρωνικών δικτύων. Πιο συγκεκριμένα, σε αυτή τη διπλωματική, στόχος μας είναι η αναγνώριση ακουστικών σκηνών σε περιβάλλοντα πόλης, το οποίο σημαίνει ότι πρέπει να κατηγοριοποιήσουμε κάθε εγγραφή του συνόλου δοκιμής (test set) σε μια από τις προκαθορισμένες τάξεις ακουστικών σκηνών. Για τον σκοπό αυτό χρησιμοποιούμε το σύνολο δεδομένων ανάπτυξης TAU urban acoustic scenes 2019 που παρέχεται στον ιστότοπο της κοινότητας DCASE 2019 [1]. Το σύνολο αυτό περιλαμβάνει ένα σέτ δεδομένων εκπαίδευσης (training set), ένα σέτ δεδομένων επικύρωσης (validation set) και ένα σέτ με δεδομένα δοκιμών. Το γενικό σύνολο των δεδομένων περιλαμβάνει 10 ακουστικές σκηνές που έχουν καταγραφεί σε 12 διαφορετικές Ευρωπαϊκές πόλεις. Αρχικά για την αντιμετώπιση του προβλήματος εξετάσαμε διάφορα χαρακτηριστικά ήχου, όπως τους MFCC (Mel Frequency Cepstral Coefficients), τον μηδενικό ρυθμό διέλευσης (zero crossing rate), την ρίζα της μέσης τετραγωνικής ενέργειας (root mean square energy) καθώς και ποικίλα μοντέλα ταξινόμησης, με τα πιο αποδοτικά να καθίστανται αυτά που είναι βασισμένα στα βαθιά νευρωνικά δίκτυα. Για αυτόν τον λόγο, αναπτύχθηκε μια πολυεπίπεδη αρχιτεκτονική πυκνού (dense) νευρωνικού δικτύου χρησιμοποιώντας το Keras API της βιβλιοθήκης TensorFlow. Τα αρχικά μας πειράματα απέδωσαν 70% ακρίβεια ταξινόμησης, η οποία βελτιώθηκε περαιτέρω σε 90 % χρησιμοποιώντας ένα μοντέλο συνολικής μάθησης (ensemble learning) της ίδιας αρχιτεκτονικής που αντιμετώπισε το πρόβλημα της ανισορροπίας τών τάξεων ανάμεσα στο σύνολο

εκπαίδευσης και στο σύνολο δοκιμής.

Abstract

Classification problems are some of the most common ones in the field of machine learning. Acoustic scene recognition or sound classification, as referred by many, has been an emerging topic in the last few years. One of the most important sound classification tasks is recognizing scenes and sound sources from environments such as an urban one, which is also the subject of this thesis. The most successful sound classification models usually consist of one or more feature extraction techniques and a deep neural network architecture. In this thesis, our task is to recognize acoustic scenes of urban environments which means we have to classify a test recording into one of the predefined acoustic scene classes. For that purpose we use the TAU urban acoustic scenes 2019 development dataset which is provided on the DCASE community 2019 challenge site [1]. The dataset contains a training, a validation, and a test set involving 10 acoustic scenes recorded in 12 different European cities. To address the problem, we have considered various audio features from the audio samples such as MFCCs (Mel Frequency Cepstral Coefficients), zero crossing rate, root mean square energy and others, as well as various classification models, with deep-learning based ones being the best choice for a classification system. Thus, a multi-layer dense neural network architecture was developed using the Keras API from the TensorFlow library. Our initial experiments yielded a 70% classification accuracy, which was further improved to 90% by employing an ensemble learning model with the same architecture, which addressed the problem of class inbalance between the data training and test sets.

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Αναγνώριση ακουστικής σκηνής

Σαν ανθρώπινα όντα ακούμε καθημερινά ποικίλους ήχους και είμαστε σε θέση να τους ξεχωρίζουμε εύκολα ως μέρος της ζωής μας. Ωστόσο, για έναν υπολογιστή δεν είναι τόσο απλό όσο φαίνεται. Το πεδίο της αναγνώρισης ακουστικών σκηνών πραγματεύεται με την διάκριση της τοποθεσίας μιας σκηνής με βάση τους ήχους του περιβάλλοντός της. Οι ήχοι μεταφέρουν ποικίλες πληροφορίες για το περιβάλλον μας και για τα γεγονότα που λαμβάνουν χώρα σε αυτό. Εξάγοντας αυτές τις πληροφορίες έχουμε τη δυνατότητα να αναγνωρίζουμε σκηνές (αεροδρόμιο, έναν πολυσύχναστο δρόμο, κτλ.) και πηγές ήχου (το γάβγισμα ενός σκύλου, ένα διερχόμενο αυτοκίνητο, κτλ.). Η ανάπτυξη μεθόδων επεξεργασίας σήματος για την εξαγωγή αυτών των πληροφοριών έχει πλήθος δυνατοτήτων σε εφαρμογές, όπως η δημιουργία συσκευών που αναγνωρίζουν δραστηριότητες στα περιβάλλοντά τους (κινητά τηλέφωνα, αυτοκίνητα, ρομπότ κτλ.) ή η δημιουργία ευφυών συστημάτων παρακολούθησης. Ωστόσο, υπάρχει ακόμη πολλή δουλειά που πρέπει να γίνει για να είναι εφικτή η αναγνώριση ακουστικών σκηνών για ηχητικά τοπία, όπου παρουσιάζονται πολλαπλοί ήχοι ταυτόχρονα.

Τα πιο επιτυχημένα μοντέλα που ταξινομούν ένα ηχητικό συμβάν στην κατάλληλη κλάση είναι αυτά της βαθιάς μάθησης. Η ταξινόμηση ήχου από ένα νευρωνικό δίκτυο μπορεί να γίνει είτε χρησιμοποιώντας ένα πλαίσιο δεδομένων των χαρακτηριστικών που εξάγονται από τα ηχητικά δείγματα ή χρησιμοποιώντας εικόνες φασματογραμμάτων που απεικονίζουν τις αλλαγές στη συχνότητα (Hz) και την ένταση (db) με την πάροδο του χρόνου. Για την ταξινόμηση εικόνων τα πιο διαδεδομένα και αποτελεσματικά μοντέλα είναι τα συνελικτικά νευρωνικά δίκτυα (CNNs). Εάν το σύνολο των ήχων ενός περιβάλλοντος θεωρείται αρκετά

ξεχωριστό, τότε οι διαφορές που παρουσιάζονται σε ένα φασματόγραμμα είναι αρκετές για να πραγματοποιηθεί μια ακριβής ταξινόμηση από ένα CNN. Παρά το ενδιαφέρον που παρουσιάζουν τα συνελικτικά νευρωνικά δίκτυα, πέραν μιας μικρής εισαγωγής στο κεφάλαιο 2, δεν θα περιγραφούν σε βάθος σε αυτή τη διπλωματική.

1.2 Η συνεισφορά της διπλωματικής εργασίας

Σκοπός της παρούσας διπλωματικής εργασίας είναι να δημιουργήσει ένα σύστημα που πραγματοποιεί ταξινόμηση ακουστικών σκηνών σε αστικά περιβάλλοντα, με όσο το δυνατόν μεγαλύτερη ακρίβεια. Ο στόχος κάθε συστήματος αναγνώρισης ακουστικών σκηνών είναι να προβλέπει την κλάση κάθε αρχείου ήχου που ανήκει στο σύνολο δοκιμών. Το σύνολο δεδομένων μας είναι το TAU urban acoustic scenes 2019 της κοινότητας DCASE, το οποίο περιέχει ήχους αστικών ηχοτοπίων και αποτελείται από ένα σύνολο εκπαίδευσης με 9185 δείγματα ήχου, ένα σύνολο δοκιμών 4185 δειγμάτων ήχου και ένα σύνολο επικύρωσης που είναι το ίδιο με το σύνολο δοκιμών για τους σκοπούς της ταξινόμησης. Ο αριθμός των προκαθορισμένων κλάσεων είναι 10 και περιέχονται ηχητικά αρχεία από 12 μεγάλες ευρωπαϊκες πόλεις. Περισσότερες λεπτομέρειες σχετικά με το σύνολο δεδομένων θα δοθούν στο κεφάλαιο 5.

Αρχικά, ξεκινήσαμε μελετώντας τα πιο κατάλληλα χαρακτηριστικά για εξαγωγή από τα ηχητικά μας δείγματα. Επιλέχθηκαν ποικίλα χαρακτηριστικά όπως ο μηδενικός ρυθμός διέλευσης, η ρίζα της μέσης τετραγωνικής ενέργειας, οι μετρήσεις φασματογράφων κλίμακας Mel (Mel scaled spectrograms), το φασματικό κέντρο (spectral centroid) και άλλα φασματικά χαρακτηριστικά, αλλά το πιο σημαντικό από όλα αυτά αποδείχθηκε οτι είναι οι MFCCs. Όλα τα χαρακτηριστικά εξήχθησαν μετά από ανάλυση σύντομου χρόνου (short time analysis) στο ηχητικό σήμα, χρησιμοποιώντας την συνάρτηση παραθύρωσης Hann με μέγεθος παραθύρου 2048 σημείων FFT (93 ms περίπου), συχνότητα δειγματοληψίας 22.05 kHz και μετατόπιση παραθύρου 23 ms (512 δείγματα FFT). Για κάθε ηχητικό αρχείο υπολογίζουμε τον μέσο όρο (mean) και την τυπική απόκλιση (standard deviation) των χαρακτηριστικών που εξαγάγουμε. Είναι επίσης σημαντικό να αναφερθεί ότι όλα τα χαρακτηριστικά εξήχθησαν χρησιμοποιώντας το Librosa, μια open source βιβλιοθήκη της Python κατάλληλη για την ανάλυση ηχητικών σημάτων. Στη συνέχεια, με σκοπό να μειώσουμε τις διαστάσεις των χαρακτηριστικών εξαγωγής που θα δωθούν σαν είσοδος στο νευρωνικό δίκτυο, χρησιμοποιήσαμε την μέθοδο της ανάλυσης γραμμικής διάκρισης (Linear Discriminant Analysis). Έτσι,

τα ομαλοποιημένα χαρακτηριστικά εξαγωγής συμβάλουν στην μείωση της τυπικής απόκλισης λάθους του νευρωνικού δικτύου με αποτέλεσμα η σωστή πρόβλεψη να επιτευχθεί πιο γρήγορα.

Έπειτα από πολλά πειράματα με διαφορετικά μοντέλα ταξινόμησης (SVM, K-πλησιέστεροι γείτονες, κλπ.) και αρχιτεκτονικές βαθιάς μάθησης (CNNs, RNNs, ResNets), αυτό που λειτούργησε για εμάς ήταν ένα νευρωνικό μοντέλο δικτύου με πλήρως συνδεδεμένα (πυκνά) στρώματα. Αυτό το μοντέλο αποτελείται από 1 στρώμα εισόδου (input layer), 1 στρώμα εξόδου (output layer), 3 κρυμμένα (hidden layers) πυκνά συνδεδεμένα (dense) στρώματα, καθώς και απο στρώματα εγκατάλειψης (dropout layers) και κανονικοποίησης παρτίδας (batch normalization) για να αποφευχθεί το φαινόμενο της υπερ-προσαρμογής ή αλλίως overfitting όπως είναι γνωστό στον ευρύτερο χώρο της επιστήμης των νευρωνικών δικτύων. Το πρόβλημα της υπερ-προσαρμογής ήταν μεγάλο στην περίπτωσή μας και αυτό οφείλεται κυρίως στις μεγάλες ανισσοροπίες μεταξύ του συνόλου εκπαίδευσης και του συνόλου δοκιμών. Για το σκοπό αυτό, χρησιμοποιήσαμε επίσης την τεχνική πρόωρης διακοπής της εκπαίδευσης (early stopping). Τα στρώματα εγκατάλειψης και η πρόωρη διακοπή εμπόδισαν το μοντέλο να ταιριάξει υπερβολικά καλά στο σύνολο εκπαίδευσης, κάτι που θα οδηγούσε σε μή επιθυμητή απόδοση κατά τη δοκιμή του συστήματος. Τα αρχικά μας πειράματα απέδωσαν με ακρίβεια 70% στο σύνολο δοκιμών. Αν και τα αποτελέσματα αυτού του συστήματος σημείωσαν βελτίωση κατά 7,5% σε σύγκριση με το σύστημα βάσης του DCASE,τα περιθώρια βελτιστοποίησης ήταν σημαντικά. Ένα μοντέλο συνολικής μάθησης (ensemble learning) ήταν η λύση στο πρόβλημά μας. Διαχωρίσαμε το σύνολο δεδομένων μας σε πολλαπλά υποσύνολα και δημιουργήσαμε ένα βασικό μοντέλο για καθένα από αυτά. Το βασικό μας μοντέλο εκπαιδεύτηκε και δοκιμάστηκε σε υποσύνολα που περιείχαν δύο κλάσεις το καθένα, κάτι το οποίο οδήγησε στην δημιουργία 5 βασικών μοντέλων. Τα μοντέλα εκτελέστηκαν παράλληλα και ανεξάρτητα το ένα από το άλλο. Συνοψίζοντας, οι τελικές μας προβλέψεις προσδιορίστηκαν συνδυάζοντας τις προβλέψεις από τα 5 βασικά μοντέλα και αυτό είχε ως αποτέλεσμα 90% ακρίβεια ταξινόμησης στο αρχικό μας σύνολο δοκιμών.

1.3 Σχετικές εργασίες

Πολλές προσεγγίσεις έχουν γίνει από τους συμμετέχοντες του διαγωνισμού της κοινότητας DCASE 2019 οι οποίες διαφέρουν σε ορισμένα σημεία με το σύστημά μας, αλλά

είναι παρόμοιες σε κάποια άλλα. Για παράδειγμα, στο [12] χρησιμοποιήθηκε ένα μοντέλο διασταυρούμενης επικύρωσης (cross-validation) παρόμοιο με το δικό μας ensemble learning μοντέλο. Το μοντέλο εκπαιδεύεται σε διαφορετικούς φακέλους δεδομένων και οι προβλέψεις των μοντέλων που προκύπτουν υπολογίζονται κατά μέσο όρο. Επίσης, σε αυτή την προσέγγιση, όπως και σε πολλές άλλες, χρησιμοποιήθηκε μια σύντομη χρονική ανάλυση παραθύρου μεγέθους 2048 FFT με μια συχνότητα δειγματοληψίας 22.05 kHz. Τα χαρακτηριστικά εισόδου κανονικοποιούνται χρησιμοποιώντας τον μέσο όρο και την τυπική απόκλιση ακριβώς όπως στην περίπτωσή μας. Η αρχιτεκτονική που χρησιμοποιείται, είναι ένα CNN με ικανότητα ανίχνευσης συχνότητας που λαμβάνει ως είσοδο φασματογράμματα δυο καναλιών. Το σύστημα ανίχνευσης συχνότητας προκύπτει από φίλτρα τα οποία είναι εξειδικευμένα στην αναγνώριση ορισμένων συχνοτήτων με την βοήθεια ενός καναλιού το οποίο περιλαμβάνει πλήθος φασματικών πληροφοριών. Η προσπάθεια αυτού του συμμετέχοντα απέδωσε ακρίβεια ταξινόμησης 85%.

Μια ακόμα ενδιαφέρουσα προσέγγιση ήταν η [13] στην οποία εφαρμόστηκε ένα ειδικό σχήμα αύξησης δεδομένων. Για να αντιμετωπιστούν οι ανισορροπίες μεταξύ του σετ εκπαίδευσης και του σετ δοκιμών που οφείλονται στις δοκιμές σε πόλεις στις οποίες το μοντέλο δεν έχει εκπαιδευτεί, ο συμμετέχων χρησιμοποίησε μια γεννήτρια δημιουργίας ψεύτικων δειγμάτων, τα οποία πρόσθεσε στη βάση δεδομένων με σκοπό την αύξηση της δειγματικής ποικιλίας. Η γεννήτρια δημιούργησε επίσης ψεύτικους ακουστικούς χάρτες χαρακτηριστικών με ετικέτες σκηνών παρόμοιες με αυτές των πραγματικών. Ένα σύστημα διάκρισης προστέθηκε ακόμα στο σχήμα με σκοπό να ξεχωρίζει τα πραγματικά από τα ψεύτικα χαρακτηριστικά εξαγωγής καθώς και τις ετικέτες σκηνής (κλάσεις). Το σύστημα αυτό αποτελείται από δύο ταξινομητές, ένα 2D πλήρως συνελικτικό νευρωνικό δίκτυο και ένα βαθύ CNN, το οποίο μαζί με μια στρατηγική διασταυρούμενης επικύρωσης οδήγησε σε ακρίβεια 85%.

Τεχνικές διασταυρούμενης επικύρωσης παρόμοιες με αυτή που θα παρουσιάσουμε σε αυτήν την διατριβή, χρησιμοποιήθηκαν από την πλειοψηφία των διαγωνιζομένων. Ένα σύστημα που ξεχώρισε, όχι τόσο με βάση την ακρίβεια των αποτελεσμάτων του αλλά για τους καινοτόμους μηχανισμούς που προτείνει ήταν το [14]. Η φιλοσοφία αυτής της προσέγγισης αποκαλείται μάθηση πολλαπλών περιπτώσεων (Multiple Instance Learning). Οι συμμετέχοντες διάσπασαν κάθε ηχητικό αρχείο των 10 δευτερολέπτων σε υποτμήματα του 1 δευτερολέπτου, δημιουργώντας μια ”τσάντα” με περιπτώσεις οπως χαρακτηριστικά την αποκαλούν. Ένα φασματόγραμμα Mel εξήχθη από κάθε περίπτωση μετά την εκτέλεση ενός STFT χρησι-

μοποιώντας Hann παράθυρο μήκους 40 ms με 50% επικάλυψη. Επίσης, χρησιμοποιήθηκαν τρία διαφορετικά μοντέλα CNN βασισμένα στην φιλοσοφία MIL, ένα παρόμοιο με το βασικό, μια αρχιτεκτονική με περιφραγμένα στρώματα και ένα μοντέλο με διασταλμένα. Οι πίνακες πιθανοτήτων που προέκυψαν από αυτά τα τρία μοντέλα συνδυάστηκαν και τροφοδοτήθηκαν σε έναν MLP ταξινομητή ο οποίος πραγματοποίησε τις τελικές προβλέψεις. Η αρχιτεκτονική αυτή οδήγησε σε μια ακρίβεια ταξινόμησης 72.3%.

Μια άλλη αρχιτεκτονική πολλαπλών μοντέλων παρουσιάζεται και στο [15], όπου οι επιμέρους προβλέψεις των 4 βασικών μοντέλων συνδυάστηκαν για να επιτευχθεί τελική πρόβλεψη. Οι αρχιτεκτονικές που χρησιμοποιήθηκαν σε αυτή την προσέγγιση ονομαστικά ήταν οι VGG12, ResNet50, AcINet και AcISincNet και είναι αρχιτεκτονικές στρωμάτων συνέλιξης. Για λόγους επίτευξης όσο το δυνατόν ακριβέστερων αποτελεσμάτων, αυτά τα τέσσερα μοντέλα προ-εκπαιδεύτηκαν με τα δεδομένα Audioset της Google, τα οποία περιέχουν πολυάριθμα ηχητικά κλίπ των 10 δευτερολέπτων, τα οποία καλύπτουν ένα ευρύ φάσμα ήχων όπως μουσικά όργανα, ανθρώπινους, ζωικούς και περιβαλλοντικούς ήχους. Είναι επίσης σημαντικό να αναφερθεί ότι το ηχητικό σήμα είχε δειγματοληφθεί στην συχνότητα των 16 kHz. Κάθε μοντέλο εκπαιδεύτηκε ξεχωριστά χρησιμοποιώντας μια στρατηγική διασταυρούμενης επικύρωσης των 5 φακέλων και η βαθμολογία εξόδου ήταν ο μέσος όρος της βαθμολογίας των 5 αυτών μοντέλων βάσης. Έπειτα, οι βαθμολογίες εξόδου των 4 αρχικών μοντέλων συγχωνεύθηκαν και αποδείχθηκαν συμπληρωματικές η μια της άλλης, κάτι που οφείλεται στην ποικιλομορφία των χαρακτηριστικών του ”front-end” συστήματος. Συμπερασματικά, ο μέσος όρος των αποτελεσμάτων του μοντέλου φάνηκε να βελτιώνει την ακρίβεια ταξινόμησης στο 83% σε σύγκριση με την ατομική απόδοση κάθε μοντέλου μοντέλου ξεχωριστά, η οποία έφτασε το 77.9% .

Μια ενδιαφέρουσα προσέγγιση που επικεντρώνεται στα προεπεξεργασμένα χαρακτηριστικά περιγράφεται στο [16]. Σε αυτή την περίπτωση, μόνο το σήμα εισόδου στερεοφωνικού μικροφώνου επεξεργάστηκε από τα ”front-end” συστήματα της αρχιτεκτονικής. Πιο συγκεκριμένα, το πρώτο χαρακτηριστικό που εξήχθη είναι το λογαριθμικό Mel φασματόγραμμα από μονοφωνικά, αρμονικά και κρουστικά σήματα. Στη συνέχεια, ο διαγωνιζόμενος προχώρησε στην εξαγωγή φασματογραμμάτων χαμηλής συχνότητας καθώς και στο λογαριθμικό Mel φασματλογραμμα ενός απομακρυσμένου σήματος. Τελευταίο αλλά εξίσου σημαντικό χαρακτηριστικό εξαγωγής αποτελεί το σήμα εισόδου στερεοφωνικού μικροφώνου το οποίο επεξεργάστηκε έπειτα από γειτονικά φίλτρα καθώς και από τον μετασχηματισμό φάσης δια-

σταυρούμενης συσχέτισης (GCC-PHAT) [17]. Η ”back-end” αρχιτεκτονική αυτού του συστήματος αποτελείται από 8 υπο-μοντέλα των οποιών οι κατά μέσο όρο προβλέψεις οδηγούν σε ένα τελικό μοντέλο συνόλου. Οι προτεινόμενες αρχιτεκτονικές ανήκουν στην ευρύτερη κατηορία των CNN και είναι: ένα συμβατικό CNN, ένα υποφασματικό νευρωνικό δίκτυο και ένα ResNet. Το τελικό σύστημα συνόλου απέδωσε ακρίβεια ταξινόμησης 80.4%

Το σύστημα που παρουσιάζεται στο [18] αποτέλεσε πηγή έμπνευσης για πλήθος πειραμάτων στην διατριβή μας. Για την εξαγωγή χαρακτηριστικών, εφαρμόστηκε μετασχηματισμός Fourier σύντομου χρονικού διαστήματος (STFT) σε συνδυασμό με Mel φίλτρα. Ο αλγόριθμος αρμονικού και κρουστικού διαχωρισμού πηγής (HPSS) [19] χρησιμοποίηθηκε ως ένας επιπρόσθετος εξαγωγέας χαρακτηριστικών. Η αρχιτεκτονική περιλαμβάνει ένα κύριο CNN και ένα υποστηρικτικό CNN εκπαιδευμένο σε ομαδοποιημένες τάξεις το οποίο διακρίνει μεταξύ των πιο δύσκολων προς ταξινόμηση σκηνών. Με αυτό τον τρόπο, δημιουργήθηκαν δύο συμπλέγματα, ένα που εντάχθηκε στις κλάσεις του πεζόδρομου και της δημόσιας πλατείας και ένα που περιλαμβάνει τις κλάσεις των μετρό και τραμ. Έτσι, το κύριο CNN προέβλεπε μεταξύ των μη ομαδοποιημένων τάξεων και το υποστηρικτικό CNN πραγματοποιούσε προβλέψεις για τις κλάσεις που συχνά ταξινομούνται εσφαλμένα. Η τελική πρόβλεψη επιτεύχθηκε χρησιμοποιώντας τους αλγορίθμους soft voting και Random Forest. Ο αλγόριθμος soft voting οδήγησε σε ακρίβεια 81.6% στο σύνολο επικύρωσης, ενώ ο αλγόριθμος Random Forest είχε αποτέλεσμα 80.6 % στο leaderboard σετ που δόθηκε στους συμμετέχοντες του διαγωνισμού.

Η τελευταία προσέγγιση του DCASE 2019 challenge που θα αναφέρουμε είναι η [20]. Σε αυτή την προσπάθεια, προτάθηκε ένα διαφορετικό χαρακτηριστικό εξαγωγής που είναι γνωστό ως συντελεστής φιλτραρίσματος μονής συχνότητας (SFFCC). Τα αρχικά πειράματα αφορούσαν ένα μοντέλο που συνδύαζε τα SFFCCs και ένα DNN μοντέλο ως ”front-end” και ”back-end” συστήματα αντίστοιχα, οδηγώντας σε ακρίβεια ταξινόμησης 65.3%. Στη συνέχεια, επιχειρήθηκε μια παρόμοια αρχιτεκτονική ”back-end” συστήματος, αλλά αυτή τη φορά τα εξαγόμενα χαρακτηριστικά ήταν οι λογαριθμικές Mel ενέργειες, οι οποίες οδήγησαν σε ακρίβεια 66.8%. Το τελικό πείραμα περιελάμβανε την συννένωση των δύο αυτών συστημάτων και ήταν αυτό που απέδωσε την καλύτερη ακρίβεια ταξινόμησης με ποσοστό επιτυχούς πρόβλεψης 70.4%, το οποίο αποτελεί βελτίωση κατά 7.9% σε σύγκριση με το σύστημα βάσης του διαγωνισμού.

1.4 Δομή διπλωματικής εργασίας

Σε αυτή την ενότητα θα περιγράψουμε την οργάνωση της διπλωματικής μας εργασίας. Εκτός από την εισαγωγή, η δομή της διπλωματικής μας είναι:

1. Το κεφάλαιο 2 δίνει μια μικρή εισαγωγή στα νευρωνικά δίκτυα, αναλύει μερικές από τις πιο διαδεδομένες συναρτήσεις ενεργοποίησης και περιγράφει τους τρεις πιο συνηθισμένους τύπους νευρωνικών δικτύων που χρησιμοποιούνται για την αναγνώριση ακουστικών σκηνών.

2. Το κεφάλαιο 3 αναλύει τη διαδικασία εξαγωγής χαρακτηριστικών που εφαρμόσαμε.

3. Το κεφάλαιο 4 εμπεριέχει όλες τις αρχιτεκτονικές του συστήματος ταξινόμησης που δημιουργήσαμε.

4. Το κεφάλαιο 5 ξεκινάει με μια ανάλυση της βάσης δεδομένων και στην συνέχεια αναλύει και συγκρίνει τα αποτελέσματα των διαφόρων αρχιτεκτονικών και αλγορίθμων που δοκιμάσαμε. Κλείνοντας, καταλήγει στα τελικά αποτελέσματα της εργασίας μας.

5. Το κεφάλαιο 6 περιλαμβάνει μία σύνοψη της διπλωματικής εργασίας και συζητά ορισμένες ιδέες για μελλοντική δουλειά.

Κεφάλαιο 2

Τα νευρωνικά δίκτυα στην αναγνώριση ακουστικών σκηνών

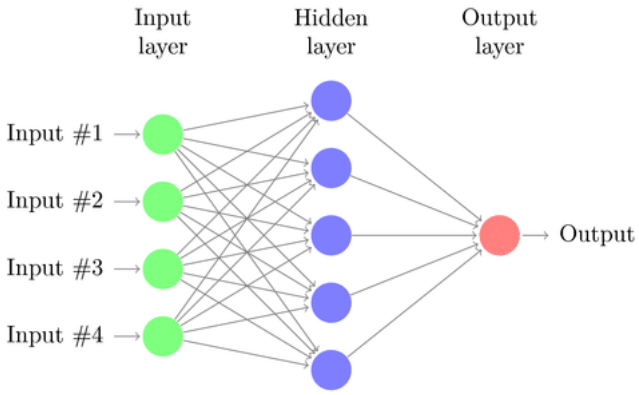
2.1 Εισαγωγή στα νευρωνικά δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks) αποτελούν τη βάση της βαθιάς μάθησης, η οποία καθιστά ένα πεδίο της μηχανικής μάθησης. Τα νευρωνικά δίκτυα είναι υπολογιστικά συστήματα που βασίζονται στα βιολογικά νευρωνικά δίκτυα που συμβάλλουν στον σχηματισμό του ανθρώπινου και ζωικού εγκεφάλου. Έχουν την ικανότητα να προσαρμόζονται, να μαθαίνουν, να συσσωρεύουν και να οργανώνουν τεράστιες ποσότητες δεδομένων. Αυτό επιτυγχάνεται με τη λήψη δεδομένων στην είσοδό τους, την εκπαίδευση τους για την αναγνώριση των μοτίβων και στη συνέχεια με την πρόβλεψη των εξόδων για ένα νέο σύνολο παρόμοιων δεδομένων. Ένα καλό νευρωνικό δίκτυο είναι αυτό που επιτυγχάνει υψηλή ακρίβεια πρόβλεψης για δεδομένα που είναι παρόμοια με τα δεδομένα εκπαίδευσης, αλλά όχι ακριβώς τα ίδια, ενώ ένα δίκτυο που αποδίδει ικανοποιητικά μόνο για τα δεδομένα που εκπαιδεύεται μπορεί να χαρακτηριστεί ως μη λειτουργικό. Τα νευρωνικά δίκτυα είναι πολύ διαδεδομένα στον τομέα της μηχανικής μάθησης και χρησιμοποιούνται σε διάφορα ζητήματα, συμπεριλαμβανομένων των ηχητικών ταξινομήσεων.

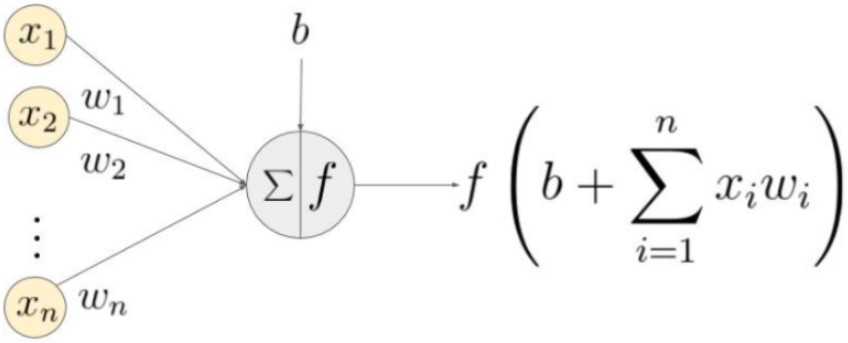
Ένα ANN βασίζεται σε ένα άθροισμα συνδεδεμένων κόμβων που ονομάζονται νευρώνες και μοντελοποιούν αυτούς ενός βιολογικού εγκεφάλου. Οι συνδέσεις μεταξύ των νευρώνων τους επιτρέπουν να επικοινωνούν με σήματα. Κάθε νευρώνας μπορεί να επεξεργαστεί και να μεταδώσει ένα σήμα σε άλλους νευρώνες που συνδέονται σε αυτόν. Το σήμα είναι ένας πραγματικός αριθμός και αντιπροσωπεύει κάποιες πληροφορίες σχετικά με τα δεδομένα που

λαμβάνει το δίκτυο ως είσοδο. Όπως μπορούμε να δούμε στο σχήμα 2.1, μια αρχιτεκτονική νευρωνικού δικτύου αποτελείται από τρία ή περισσότερα στρώματα νευρώνων. Το σήμα ταξιδεύει από το στρώμα εισόδου (πρώτο στρώμα) στο στρώμα εξόδου (τελευταίο στρώμα) αφού πιθανώς διασχίσει μερικά ενδιάμεσα στρώματα, τα οποία ονομάζονται κρυμμένα στρώματα και εκτελούν τους περισσότερους υπολογισμούς.

Αξίζει να αναλυθεί ο τρόπος διάδοσης των σημάτων μέσω ενός νευρωνικού δικτύου. Οι νευρώνες ενός στρώματος συνδέονται με νευρώνες του επόμενου στρώματος μέσω καναλιών και κάθε κανάλι έχει μια αριθμητική τιμή, η οποία είναι γνωστή ως βάρος. Τα δεδομένα που τροφοδοτούνται ως εισροές στο πρώτο στρώμα, πολλαπλασιάζονται με τα αντίστοιχα βάρη και το άθροισμά τους αποστέλλεται ως είσοδος στους συνδεδεμένους νευρώνες του κρυμμένου στρώματος. Κάθε ένας από αυτούς τους νευρώνες συνδέεται με μια αριθμητική τιμή που ονομάζεται πόλωση (bias), η οποία προστίθεται στο άθροισμα της εισόδου. Αυτή η τιμή μεταβιβάζεται στη συνέχεια σε μια συνάρτηση κατωφλίου (threshold), γνωστή ως συνάρτηση ενεργοποίησης (activation function). Οι υπολογισμοί που μόλις εξηγήσαμε φαίνονται στο σχήμα 2.2. Το αποτέλεσμα της συνάρτησης ενεργοποίησης καθορίζει εάν ο συγκεκριμένος νευρώνας θα ενεργοποιηθεί ή όχι, κάτι που σημαίνει ότι θα είναι σε θέση να μεταδώσει το σήμα στους συνδεδεμένους νευρώνες του επόμενου στρώματος. Τέλος, όταν φτάσουμε στο στρώμα εξόδου, ο νευρώνας με την υψηλότερη τιμή καθορίζει την έξοδο. Αυτές οι τιμές αντιπροσωπεύουν μια πιθανότητα και για αυτόν τον λόγο η μεγαλύτερη από αυτές δηλώνει την έξοδο που προβλέπεται από το νευρωνικό δίκτυο. Η διαδικασία που περιγράψαμε ονομάζεται διάδοση προς τα εμπρός (forward propagation). Κατά τη διάρκεια της διαδικασίας εκπαίδευσης, μαζί με την είσοδο, το δίκτυο τροφοδοτείται επίσης και με την έξοδο των δεδομένων. Η προβλεπόμενη έξοδος στη συνέχεια συγκρίνεται με την πραγματική έξοδο για να συνειδητοποιήσουμε το μέγεθος του σφάλματος στην πρόβλεψη, το οποίο ουσιαστικά δίνει την πληροφορία για το πόσο λανθασμένη είναι η πρόβλεψη του συστήματός μας. Αυτές οι πληροφορίες μεταφέρονται στη συνέχεια προς τα πίσω μέσω του δικτύου και με βάση αυτό ρυθμίζονται τα βάρη. Αυτή η διαδικασία είναι γνωστή ως οπισθοδιάδοση (backpropagation). Ο κύκλος της διάδοσης προς τα εμπρός και πίσω πραγματοποιείται με πολλαπλές εισόδους έως ότου το δίκτυό μας να είναι έτοιμο να προβλέψει σωστά την πλειοψηφία των εξόδων ή τουλάχιστον ένα ικανοποιητικό ποσοστό αυτών. Στη συνέχεια, είμαστε σε θέση να υποστηρίξουμε ότι το δίκτυό μας είναι καλά εκπαιδευμένο και έτοιμο να κάνει προβλέψεις σε ένα σύνολο δεδομένων παρόμοιο με αυτό των δεδομένων εκπαίδευσης, το οποίο αποτελεί



Σχήμα 2.1: Οπτικοποίηση ενός νευρωνικού δικτύου. (Εικόνα από [2]).



Σχήμα 2.2: Υπολογισμοί συνάρτησης ενεργοποίησης. (Εικόνα από [3]).

το σύνολο δοκιμών του συστήματός μας.

2.2 Συναρτήσεις ενεργοποίησης (activation functions)

Όπως αναφέρθηκε προηγουμένως, οι συναρτήσεις ενεργοποίησης είναι συναρτήσεις κατωφλίου που λαμβάνουν μια τιμή ως είσοδο και υπολογίζουν ένα αποτέλεσμα, το οποίο θεωρείται ως η έξοδος του νευρώνα που καθορίζει εάν ο νευρώνας πρέπει να μεταφέρει δεδομένα στους συνδεδεμένους νευρώνες του επόμενου στρώματος. Το κύριο όφελος της έννοιας της συνάρτησης ενεργοποίησης είναι ότι παρέχει μη γραμμικότητα στην έξοδο του νευρώνα [21]. Αυτό μας επιτρέπει να δουλεύουμε με τεράστια και πολύπλοκα δεδομένα καθώς και να μαθαίνουμε πιο περίπλοκες πληροφορίες από τα δεδομένα μας. Δίχως τις λειτουργίες ενεργοποίησης, όλες οι εξόδοι θα ήταν γραμμικές κάτι που θα καθιστούσε τις δυνατότητες του δικτύου μας πολύ περιορισμένες. Δεδομένου ότι οι συναρτήσεις ενεργοποίησης είναι τόσο σημαντικές για ένα σύστημα νευρωνικού δικτύου, αξίζει να αναλύσουμε μερικούς από τους πιο συνηθισμένους τύπους που αποτέλεσαν επίσης μέρος των πειραμάτων μας [4].

1. **SIGMOID**

   Όταν χρησιμοποιείται η συνάρτηση Sigmoid, η είσοδος x μεταβιβάζεται στην εξίσωση 2.1 για να παράγει μια έξοδο μεταξύ 0 και 1.

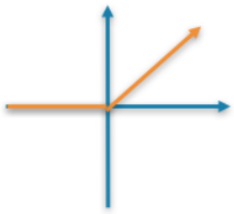
$$sigmoid(x) = \frac{1}{1 + \exp(-x)} \tag{2.1}$$

   Όταν μια πραγματικά μεγάλη είσοδος μεταβιβάζεται στη συνάρτηση, ο παρονομαστής ισούται με 1 επειδή το $exp(-x)$ αποτελεί έναν τόσο μικρό αριθμό που θεωρείται μηδενικός. Αυτό σημαίνει πώς για πραγματικά μεγάλες εισόδους, η έξοδος της συνάρτησης Sigmoid ισούται με την μονάδα. Αντιθέτως, για πολύ μικρές αρνητικές εισόδους, ο παρονομαστής είναι ένας πολύ μεγάλος αριθμός και δεδομένου ότι ο αριθμητής είναι 1, το αποτέλεσμα είναι ένας αριθμός που σε μια αριθμομηχανή αναπαριστάται ώς 0. Έτσι, η έξοδος για πραγματικά μικρές αρνητικές εισόδους είναι 0. Παρ'όλα αυτά, η συνάρτηση Sigmoid έχει πολύ χαμηλή χρήση και αυτό οφείλεται στο ότι δεν είναι κεντραρισμένη στο μηδέν, γεγονός που καθιστά εξαιρετικά δύσκολη τη βελτιστοποίηση του νευρωνικού δικτύου. Επίσης, δεν επιλύει το πρόβλημα της εξαφάνισης της κλίσης (vanishing gradient). Στο πρόβλημα αυτό, οι κλίσεις συρρικνώνονται επειδή το δίκτυο κατά τη διάρκεια της διαδικασίας της οπισθοδιάδοσης δίνει τιμές τόσο μικρές ώστε τα βάρη μετά βίας ενημερώνονται στα αρχικά στρώματα. Στην Sigmoid, η μέγιστη κλίση είναι 0,25 που έχει ως αποτέλεσμα τα βάρη να συρρικνώνονται καθώς πολλαπλασιάζονται με τόσο μικρές τιμές. Αυτό έχει ως αποτέλεσμα το δίκτυο να μην εκπαιδεύεται περαιτέρω έτσι ώστε να κάνει ακριβείς προβλέψεις.

2. **SOFTMAX**

   Η συνάρτηση ενεργοποίησης Softmax χρησιμοποιείται ιδιαίτερα σε ζητήματα που περιλαμβάνουν πολλαπλές τάξεις, όπως στην περίπτωσή μας. Η Softmax ομαλοποιεί τις εξόδους για κάθε τάξη μεταξύ των τιμών 0 και 1 και διαιρεί με το άθροισμα τους, έτσι ώστε ως αποτέλεσμα να δίνεται η πιθανότητα που έχει μια είσοδος να ανήκει σε μια συγκεκριμένη κλάση. Αυτός είναι ο λόγος για τον οποίο χρησιμοποιείται πιο συχνά στο στρώμα εξόδου ενός νευρωνικού δικτύου, προκειμένου να ταξινομηθούν οι είσοδοι στίς διάφορες κατηγορίες.

$$softmax(x) = \frac{\exp(x)}{\sum_{k=1}^{K} \exp(x)} \tag{2.2}$$



Σχήμα 2.3: ReLU. (Εικόνα από [4]).

3. **TANH / HYPERBOLIC TANGENT**

   Η TanH είναι επίσης μια μη γραμμική συνάρτηση που μπορεί να συλλάβει πιο σύνθετα πρότυπα δεδομένων και κυμαίνεται από -1 έως 1, όπως μπορούμε να επαληθεύσουμε παρατηρώντας την εξίσωση 2.3. Επιπρόσθετα, είναι κεντραρισμένη στο μηδεν, κάτι που επιτρέπει στο δίκτυο να μοντελοποιεί εισόδους έντονα αρνητικές, ουδέτερες και έντονα θετικές. Ωστόσο, αυτή η λειτουργία ενεργοποίησης μπορεί επίσης να υποφέρει από το πρόβλημα εξαφάνισης της κλίσης που εξηγήσαμε προηγουμένως.

$$tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{2.3}$$

4. **ReLU (RECTIFIED LINEAR UNIT)**

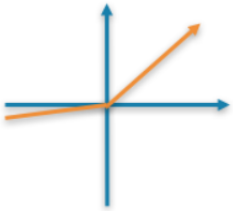
   Η συνάρτηση ενεργοποίησης ReLU χρησιμοποιείται ιδιαίτερα συχνά επειδή επιτρέπει στο νευρωνικό δίκτυο να συγκλίνει γρήγορα και όπως μπορούμε να δούμε από το σχήμα 2.3 κυμαίνεται από το μηδέν έως το άπειρο παίρνοντας πολύ μεγάλες τιμές. Παρ'όλα αυτά, όταν οι τιμές εισόδου είναι αρνητικές ή κοντά στο μηδέν, η κλίση της συνάρτησης γίνεται μηδενική, κάτι που εμποδίζει το δίκτυο να επεξεργαστεί και να μάθει περαιτέρω πληροφορίες απο τα δεδομένα. Αυτό το πρόβλημα είναι γνωστό ως "dying ReLU".

5. **LEAKY ReLU**

   Η Leaky ReLU είναι μια παραλλαγή της ReLU, η οποία έχει μια μικρή θετική κλίση στην αρνητική περιοχή, προκειμένου να αποφευχθεί το "dying Relu" πρόβλημα και να ενεργοποιηθεί η διαδικασία της οπισθοδιάδοσης του δικτύου, ακόμη και για αρνητικές τιμές. Το μειονέκτημα της Leaky ReLU είναι ότι δεν παρέχει συνεπείς προβλέψεις για αρνητικές τιμές.

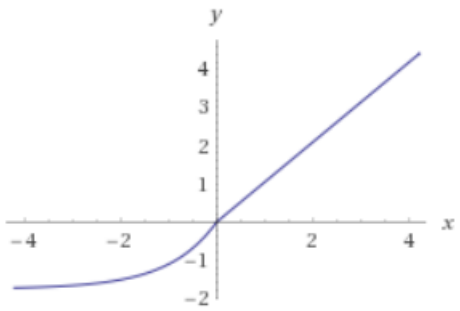
6. **SELU (SCALED EXPONENTIAL LINEAR UNITS)**

   Η τελευταία συνάρτηση ενεργοποίησης που θα αναλύσουμε είναι και αυτή που τελικά



Σχήμα 2.4: Leaky ReLU. (Εικόνα από [4]).

χρησιμοποιήσαμε στο νευρωνικό μας δίκτυο και ονομάζεται SELU. Η SELU εισάγει την ιδέα της κανονικοποίησης (normalization) για να αντιμετωπίσει τα προβλήματα που αναφέραμε προηγουμένως. Υπάρχουν τρεις τύποι κανονικοποίησης στα νευρωνικά δίκτυα: η κανονικοποίηση εισόδου, η κανονικοποίηση παρτίδας και η εσωτερική κανονικοποίηση. Στην εσωτερική κανονικοποίηση η SELU ξεχωρίζει από τις άλλες συναρτήσεις ενεργοποίησης. Η βασική έννοια της εσωτερικής κανονικοποίησης είναι πως κάθε στρώμα διατηρεί τον μέσο όρο και τη διακύμανση (variance) των τιμών από τα προηγούμενα στρώματα. Για να μετατοπιστεί ο μέσος όρος, μια συνάρτηση ενεργοποίησης πρέπει να έχει τόσο αρνητικές όσο και θετικές τιμές ως εξόδους, μια απαίτηση που ικανοποιεί η SELU όπως μπορούμε να δούμε στο σχήμα 2.5. Ο καλύτερος τρόπος για να ρυθμιστεί η διακύμανση, είναι με τη χρήση της κλίσης. Παρατηρώντας την εξίσωση 2.4 είναι λογικό να πούμε ότι η SELU μοιάζει με την ReLU για τιμές μεγαλύτερες από το μηδέν. Ωστόσο, υπάρχει μια επιπλέον παράμετρος λ που είναι ο λόγος για τον ”κλιμακωτό (scaled)” όρο στον ορισμό της συνάρτησης. Όταν το λ είναι μεγαλύτερο από 1, η κλίση είναι επίσης μεγαλύτερη από 1, οπότε η συνάρτηση ενεργοποίησης μπορεί να αυξήσει τη διακύμανση. Από την άλλη πλευρά, όταν το λ είναι κοντά στο μηδέν, η κλίση είναι κοντά στο μηδέν, κάτι που μειώνει τη διακύμανση των τιμών. Αυτό είναι κάτι που θα προκαλούσε ένα πρόβλημα εξαφάνισης της κλίσης σε ορισμένες άλλες συναρτήσεις ενεργοποίησης, αλλά είναι ζωτικής σημασίας για την εσωτερική κανονικοποίηση. Συνοψίζοντας, η SELU επιλύει τα δύο κύρια προβλήματα των συναρτήσεων ενεργοποίησης, χρησιμοποιώντας εσωτερική κανονικοποίηση και επίσης με βάση τα πειράματα μας εκπαιδεύεται ταχύτερα και πιο ορθά από άλλες συναρτήσεις ενεργοποίησης [5], ακόμη και στην περίπτωση της κανονικοποίησης παρτίδας. Αυτός μάλιστα είναι ο κύριος λόγος για τον οποίο προτιμήθηκε στο ζήτημα της αναγνώρισης ακουστικών σκηνών που αντιμετωπίζουμε σε αυτήν την διπλωματική.



Σχήμα 2.5: Γραφική παράσταση της SELU. (Εικόνα από [5]).

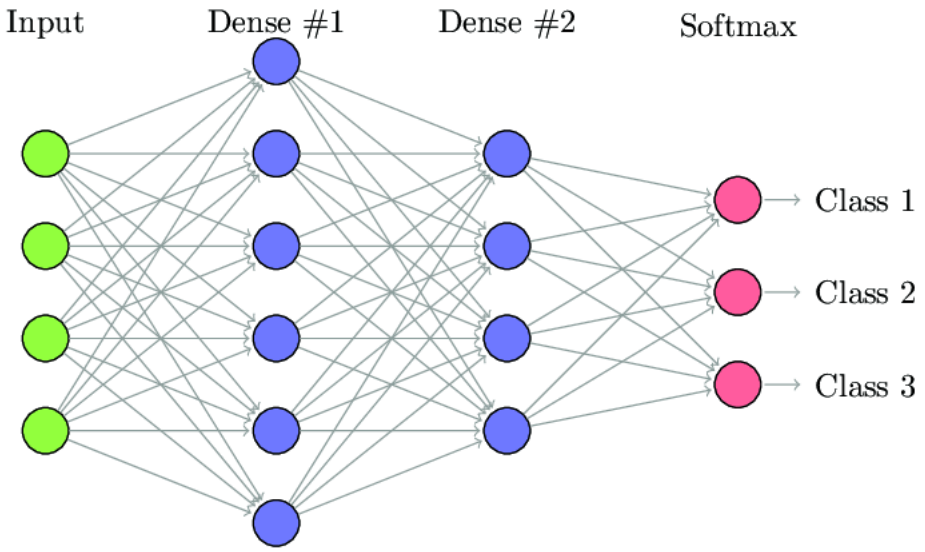
$$f(x) = \lambda \begin{cases} \alpha(\exp(x) - 1) & x < 0 \\ x & 0 \leq x \end{cases} \tag{2.4}$$

2.3 Τα νευρωνικά δίκτυα στην ταξινόμηση ήχου

Πριν προχωρήσουμε στο σύστημα αναγνώρισης αστικών ηχητικών σκηνών, είναι σημαντικό να αναλύσουμε τους κύριους τύπους νευρωνικών δικτύων που χρησιμοποιούνται στη μηχανική μάθηση για προβλήματα ταξινόμησης ήχου και αποτελούν επίσης μέρος της έρευνας μας. Με βάση τη μελέτη μας, όταν πρόκειται για επιλογή αρχιτεκτονικής ανώτατου επιπέδου, οι πιο συνηθισμένες είναι: πυκνά ή πλήρως συνδεδεμένα νευρωνικά δίκτυα (Dense ή fully connected), υπολειμματικά νευρωνικά δίκτυα (ResNets) και συνελικτικά νευρωνικά δίκτυα (CNNs). Ας δούμε τους ορισμούς αυτών των τύπων νευρωνικών δικτύων και γιατί χρησιμοποιούνται τόσο ευρέως στην ταξινόμηση του ήχου.

2.3.1 Πυκνά νευρωνικά δίκτυα (Dense neural networks)

Ένα πυκνό νευρωνικό δίκτυο, όπως υποδηλώνει το όνομά του, αποτελείται από πυκνά ή αλλιώς πλήρως συνδεδεμένα στρώματα [22]. Τα πυκνά ή πλήρως συνδεδεμένα στρώματα είναι γραμμικές λειτουργίες, όπου κάθε νευρώνας λαμβάνει ως είσοδο τις εξόδους όλων των νευρώνων από το προηγούμενο στρώμα. Με απλούστερους όρους, όλοι οι νευρώνες σε ένα στρώμα συνδέονται με όλους από το επόμενο στρώμα. Το κύριο πλεονέκτημα ενός δικτύου με πυκνά στρώματα και ο λόγος για τον οποίο αυτή η αρχιτεκτονική είναι τόσο διαδεδομένη, είναι ότι εκπαιδεύεται και μαθαίνει από όλους τους συνδυασμούς χαρακτηριστικών στα προ-

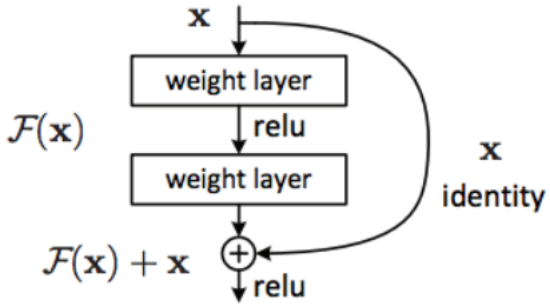


Σχήμα 2.6: Παράδειγμα ενός πυκνού νευρωνικού δικτύου. (Εικόνα από [6]).

ηγούμενα στρώματα, κάτι που σημαίνει πως είναι αδύνατο να χαθεί οποιαδήποτε σημαντική πληροφορία. Επίσης, τα πλήρως συνδεδεμένα δίκτυα είναι ”δομικά αγνωστικιστικά” δηλαδή ο τύπος εισόδου τους δεν καθορίζεται, σε αντίθεση με τα συνελικτικά δίκτυα όπως θα εξηγήσουμε αργότερα. Επίσης, είναι σημαντικό να αναφέρουμε ότι ένα πλεονέκτημα αυτής της δομής είναι το χαμηλό υπολογιστικό κόστος και ο μικρός χρόνος εκπαίδευσης, κάτι που ήταν πραγματικά βοηθητικό για τα πειράματά μας και ένας από τους κύριους λόγους επιλογής ενός δικτύου με πυκνά στρώματα.

2.3.2 Υπολλειματικά νευρωνικά δίκτυα (Residual neural networks)

Τα βαθύτερα νευρωνικά δίκτυα συνήθως αποδίδουν καλύτερα από τις ρηχότερες εκδόσεις του εαυτού τους καθώς αποτελούνται από περισσότερα στρώματα. Ωστόσο, ένα βαθύ νευρωνικό δίκτυο είναι πιο δύσκολο να βελτιστοποιηθεί και απαιτεί περισσότερο υπολογιστικό κόστος και χρόνο, κάτι που καθιστά τη διαδικασία εκπαίδευσης δυσχερέστερη. Αυτό συμβαίνει διότι, είναι πιο δύσκολο για τον αλγόριθμο βελτιστοποίησης να βρει τις σωστές παραμέτρους. Επίσης, από τα πειράματά μας παρατηρήσαμε ότι καθώς προσθέτουμε στρώματα σε ένα δίκτυο, τα αποτελέσματα βελτιώνονται μέχρι κάποιο σημείο όπου η ακρίβεια αρχίζει να μειώνεται. Τα υπολειμματικά δίκτυα είναι δομές που επιλύουν αυτό το ζήτημα προσθέτοντας συνδέσεις παράλειψης (skip connections). Οι συνδέσεις παράλειψης προσθέτουν αντιστοιχίσεις ταυτότητας (identity mappings) από ένα σημείο του δικτύου σε ένα προς



Σχήμα 2.7: Παράδειγμα ενός υπολειμματικού μπλόκ. (Εικόνα από [7]).

τα εμπρός σημείο και αφήνουν το δίκτυο να μάθει τις επιπλέον πληροφορίες που παραλείψαμε. Ένα μπλοκ με συνδέσεις παράλειψης όπως αυτό στο σχήμα 2.7 ονομάζεται υπολειμματικό μπλοκ. Ένα υπολειμματικό δίκτυο δεν είναι παρά μια ακολουθία τέτοιων υπολειμματικών μπλοκ. Επίσης, η δομή των υπολειμματικών δικτύων είναι αρκετά παρόμοια με τη δομή του ανθρώπινου εγκεφάλου, αφού για παράδειγμα οι νευρώνες φλοιώδους στιβάδας VI λαμβάνουν είσοδο από το στρώμα I παρακάμπτοντας όλα τα ενδιάμεσα στρώματα [7].

2.3.3 Συνελικτικά νευρωνικά δίκτυα (CNNs)

Τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks) είναι μια κατηγορία βαθέων νευρωνικών δικτύων που χρησιμοποιούνται ευρέως στην ανάλυση εικόνας [23]. Ένα CNN έχει την εξειδίκευση να ανιχνεύει μοτίβα και να βγάζει συμπεράσματα από αυτά, γι'αυτό χρησιμοποιείται συχνά στην ανάλυση εικόνων. Το στοιχείο που διαφοροποιεί ένα CNN από ένα πολυστρωματικό δίκτυο perceptron είναι τα κρυμμένα στρώματα που ονομάζονται συνελικτικά στρώματα. Ένα CNN μπορεί επίσης να έχει και μη συνελικτικά στρώματα, αλλά η βάση ενός συνελικτικού δικτύου είναι συνήθως αυτά. Ακριβώς όπως κάθε άλλο στρώμα, ένα συνελικτικό στρώμα λαμβάνει μια είσοδο την οποία τροποποιεί και στη συνέχεια την εξάγει στο επόμενο στρώμα. Η διαφορά σε αυτήν την περίπτωση είναι ότι αυτός ο μετασχηματισμός είναι μια λειτουργία συνέλιξης και κάθε ένα από αυτά τα στρώματα έχει φίλτρα που ανιχνεύουν τα μοτίβα που αναφέραμε προηγουμένως. Λέγοντας μοτίβα, εννοούμε τις διαφορετικές άκρες, σχήματα, αντικείμενα που φαίνεται να εμφανίζονται σε μια εικόνα. Τα φίλτρα είναι μικροί πίνακες των οποίων οι τιμές είναι πραγματικοί αριθμοί που περιέχουν τα βάρη του στρώματος και το bias. Για παράδειγμα, ας υποθέσουμε ότι θέλουμε ένα στρώμα να περιέχει ένα φίλτρο με διαστάσεις 4x4, το φίλτρο θα ολισθήσει πάνω από κάθε σύνολο 4x4 εικονοστοιχείων της εικόνας εισόδου. Αυτή η ολίσθηση είναι γνωστή ως συνέ-

λιξη. Καθώς εισχωρούμε βαθύτερα στο δίκτυο, τα φίλτρα γίνονται πιο εξελιγμένα, το οποίο σημαίνει ότι τα πρώτα στρώματα είναι σε θέση να ανιχνεύσουν μόνο άκρες και σχήματα, ενώ σε μεταγενέστερα στρώματα μπορεί να είναι σε θέση να ανιχνεύσουν συγκεκριμένα αντικείμενα όπως ένα αυτοκίνητο ή ένα δέντρο. Εκτός από τα συνελικτικά στρώματα ένα CNN μπορεί να έχει και άλλα στρώματα. Αυτά τα στρώματα μπορούν να είναι στρώματα συγκέντρωσης (pooling) ή πλήρως συνδεδεμένα (dense) που εμφανίζονται σε ένα πυκνό νευρωνικό δίκτυο όπως αναφέραμε προηγουμένως. Τα στρώματα συγκέντρωσης είναι χρήσιμα επειδή μειώνουν τις διαστάσεις των δεδομένων συνδυάζοντας τις εξόδους κάθε νευρώνα ενός στρώματος σε έναν μόνο νευρώνα του επόμενου στρώματος. Υπάρχουν δύο τύποι συγκέντρωσης: η μέγιστη συγκέντρωση (max pooling) και η μέση συγκέντρωση (average pooling). Η μέγιστη συγκέντρωση χρησιμοποιεί τη μέγιστη τιμή κάθε ομάδας νευρώνων του προηγούμενου στρώματος ενώ η μέση συγκέντρωση χρησιμοποιεί τη μέση τιμή. Τα συνελικτικά νευρωνικά δίκτυα εισάγουν επίσης τον όρο του υποδεκτικού πεδίου (receptive field). Το υποδεκτικό πεδίο είναι μια συγκεκριμένη περιοχή σε κάθε στρώμα οπού η έξοδος λαμβάνεται αποκλειστικά ως είσοδος των επόμενων στρωμάτων.

Λαμβάνοντας υπόψη όλα αυτά, μπορεί να μην φαίνεται πρακτικό να χρησιμοποιήσουμε ένα συνελικτικό νευρωνικό δίκτυο για την πραγματοποίηση της ηχητικής ταξινόμησης, αφού είναι χρήσιμο μόνο για εικόνες. Παρ'όλα αυτά, κάθε ήχος μπορεί να αναπαρασταθεί με εικόνες φασματογραμμάτων που απεικονίζουν τις αλλαγές στην συχνότητα σε Hz ή τις αλλαγές στην ένταση σε dB. Εάν οι ομαδοποιημένοι ήχοι διαφέρουν αρκετά μεταξύ τους, ένα συνελικτικό δίκτυο θα πρέπει να είναι σε θέση να τους ξεχωρίσει με βάση τις διαφορές στα φασματογράμματα. Ωστόσο, ένα CNN πιθανότατα δεν θα είναι σε θέση να ταξινομήσει τις λέξεις σε μια πρόταση, αλλά θα είναι ιδιαιτέρως κατάλληλο να διακρίνει το γάβγισμα ενός σκύλου από μια συνομιλία ή στην περίπτωσή μας τους ήχους ενός αεροδρομίου από τους ήχους ενός πεζοδρόμου.

Κεφάλαιο 3

Εξαγωγή χαρακτηριστικών

Η εξαγωγή χαρακτηριστικών είναι μια διαδικασία η οποία προσδιορίζει και χρησιμοποιεί τα πιο σημαντικά χαρακτηριστικά από μεγάλα σύνολα δεδομένων [24] για να τα περιγράψει με όσο το δυνατόν μεγαλύτερη ακρίβεια. Στην περίπτωσή μας, τα χαρακτηριστικά εξαγωγής των ηχητικών αρχείων θα πρέπει να είναι σε θέση να περιγράψουν επαρκώς κάθε ακουστική σκηνή και να εντοπίσουν τις διαφορές μεταξύ αυτών, έτσι ώστε το νευρωνικό μας δίκτυο να λάβει τις κατάλληλες εισόδους που θα το οδηγήσουν σε σωστές προβλέψεις. Για τους σκοπούς εξαγωγής χαρακτηριστικών χρησιμοποιήσαμε την Librosa [25], μια open source βιβλιοθήκη της Python εξειδικευμένη για ανάλυση ήχου. Για την σωστή εξαγωγή των χαρακτηριστικών, αρχικά προηγήθηκε η ανάλυση του σήματος κάθε αρχείου ήχου σε παράθυρα (frames). Η Librosa περιέχει συναρτήσεις που όχι μόνο εξάγουν τα χαρακτηριστικά της προτίμησής μας από τα αρχεία ήχου, αλλά επίσης είναι σε θέση να εκτελέσουν την ανάλυση του σήματος σε παράθυρα με βάση τις παραμέτρους που επιλέγουμε. Το σήμα κάθε αρχείου ήχου δειγματολήφθηκε χρησιμοποιώντας την αντίστοιχη συνάρτηση της Librosa, σε συχνότητα 22.05 KHz. Στη συνέχεια, εξάγαμε τα χαρακτηριστικά μας ανά παράθυρο χρησιμοποιώντας μέγεθος παραθύρου περίπου 93 ms (2048 FFT) και μέγεθος αναπήδησης 23 ms (512 FFT). Το μέγεθος αναπήδησης ή αλλιώς το βήμα παραθύρου είναι ο αριθμός των δειγμάτων μεταξύ δύο διαδοχικών παραθύρων και πρέπει πάντα να είναι μικρότερος από το μέγεθος του παραθύρου για να εξασφαλιστεί η επικάλυψη.

Η συνάρτηση παραθύρωσης που χρησιμοποιήθηκε για την ανάλυσή είναι η συνάρτηση Hann [26] και δίνεται από την ακόλουθη σχέση με το L να είναι το μήκος του παραθύρου:

$$w_0(x) \triangleq \left\{ \begin{array}{cc} \frac{1}{2}\left(1+\cos\left(\frac{2\pi x}{L}\right)\right) = \cos^2\left(\frac{\pi x}{L}\right), & |x| \leq L/2 \\ 0, & |x| > L/2 \end{array} \right\}$$

Σε αυτό το σημείο της διπλωματικής θα αναλύσουμε λεπτομερώς τα χαρακτηριστικά τα οποία αποφασίσαμε ότι είναι τα πιο κατάλληλα για εξαγωγή με σκοπό την επιτυχή αναγνώριση ακουστικών σκηνών. Συγκεκριμένα, χρησιμοποιήσαμε:

- **Ρίζα μέσης τετραγωνικής ενέργειας**

  Ένα από τα πιο κοινά χαρακτηριστικά των ηχητικών σημάτων για εξαγωγή είναι η ρίζα της μέσης τετραγωνικής ενέργειας (Root Mean Square Energy) του σήματος. Η RMSE είναι βασικά η τετραγωνική ρίζα του μέσου τετραγωνικού πλάτους ενέργειας ενός παραθύρου [27] και υπολογίζεται από τη σχέση 3.1. Η ενέργεια είναι μια μονάδα μέτρησης της έντασης του σήματος, η οποία μπορεί να διαφέρει από το ένα ηχητικό τοπίο στο άλλο και αυτό είναι κάτι πραγματικά χρήσιμο σε ένα πρόβλημα ταξινόμησης ακουστικών σκηνών.

$$RMSE(x) \quad = \quad \sqrt{\frac{1}{N}\sum_{n}|x(n)|^2} \tag{3.1}$$

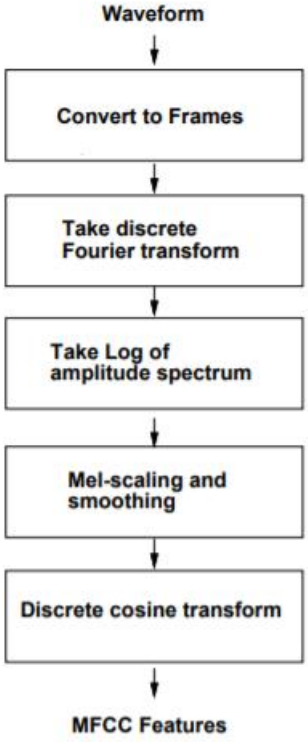
- **Συντελεστές συχνότητας Mel (Mel Frequency Cepstral Coefficients).**

  Οι MFCCs είναι μακράν τα πιο κυρίαρχα χαρακτηριστικά που χρησιμοποιούνται σε θέματα που αφορούν ηχητικά σήματα όπως η αναγνώριση ομιλητή, η ανίχνευση ομιλίας ή η αναγνώριση σκηνών όπως στην εργασία μας. Οι MFCCs είναι φασματικά βασισμένα χαρακτηριστικά και είναι ιδιαιτέρως διαδεδομένα λόγω της ικανότητάς τους να αντιπροσωπεύουν ένα ευρύ φάσμα ήχου σε μια τέτοια στερεή μορφή. Παρόλο που εξάγαμε αυτά τα χαρακτηριστικά όπως και όλα τα άλλα μέσω μιας συνάρτησης Librosa, είναι σημαντικό να αναλύσουμε όλα τα βήματα που απαιτούνται για τον υπολογισμό των συντελεστών συχνότητας Mel όπως παρουσιάζονται συνοπτικά στο σχήμα 3.1. Αρχικά, πρέπει να εκτελέσουμε μια ανάλυση σύντομου χρόνου. Στη συνέχεια, πρέπει να υπολογίσουμε το φάσμα ισχύος κάθε πλαισίου χρησιμοποιώντας ένα περιοδόγραμμα που εμπνέεται από ένα συγκεκριμένο όργανο στο ανθρώπινο αυτί που ονομάζεται κοχλίας [27]. Ο κοχλίας δονείται σε πολλαπλά σημεία με βάση τη συχνότητα των ήχων που λαμβάνει ως είσοδο. Για να προσομοιώσουμε κάτι τέτοιο και να κατασκευάσουμε το σωστό περιοδόγραμμα ξεκινάμε υπολογίζοντας τον διακριτό μετασχη-

ματισμό Fourier κάθε πλαισίου όπως φαίνεται στην εξίσωση 3.2 με το $s_i(n)$ να είναι το παραθυρωμένο σήμα του πλαισίου i, N ο αριθμός των δειγμάτων σε ένα παράθυρο Hann, h(n) το παράθυρο Hann. Στη συνέχεια, για να υπολογίσουμε το περιοδόγραμμα φάσματος ισχύος εφαρμόζουμε τον τύπο 3.3. Στη συνέχεια, παίρνουμε τον λογάριθμο του φάσματος ισχύος αφού αναλογιζόμαστε την ένταση ως λογαριθμική τιμή. Για τον καλύτερο διαχωρισμό των στενά διαχωρισμένων συχνοτήτων περνάμε το σήμα από την τράπεζα φίλτρων Mel (Mel filterbank). Το Mel filterbank είναι ένα σύνολο 26 τριγωνικών φίλτρων τα οποία αποτελούνται κυρίως από μηδενικά αλλά σχηματίζουν ένα μη μηδενικό τρίγωνο σε μια συγκεκριμένη περιοχή συχνοτήτων. Εφαρμόζουμε αυτά τα φίλτρα στον λογάριθμο του φάσματος ισχύος που υπολογίσαμε στο προηγούμενο βήμα σχηματίζοντας 26 λογαριθμικές ενέργειες Mel. Για το τελικό βήμα, εφαρμόζουμε ένα διακριτό μετασχηματισμό συνημίτονου (Discrete Cosine Transform) [28] στις προκύπτουσες ενέργειες με στόχο την αναίρεση οποιωνδήποτε συσχετισμών μεταξύ τους. Το αποτέλεσμα είναι 26 συντελεστές που αποτελούν τους MFCCs, από τους οποίους μόνο ένα μέρος είναι χρήσιμο. Συνήθως σε θέματα αναγνώρισης ομιλίας, οι MFCCs χρησιμοποιούνται για την ταξινόμηση των φωνημάτων και χρειάζονται μόνο οι πρώτοι 12-13 συντελεστές. Ωστόσο, στην αναγνώριση ακουστικών σκηνών και μετά από πολλαπλά πειράματα, καταλήξαμε στο συμπέρασμα ότι οι πρώτοι 20 MFCCs έχουν μεγάλη σημασία για την επίτευξη καλύτερων αποτελεσμάτων.

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)\exp\frac{-j2kn\pi}{N} \tag{3.2}$$

$$P_i(k) = \frac{1}{N}|S_i(k)|^2 \tag{3.3}$$



Σχήμα 3.1: Βήματα για την δημιουργία των MFCCs. (Εικόνα από [8]).

- **Μηδενικός ρυθμός διέλευσης (Zero Crossing Rate)**

  Ο μηδενικός ρυθμός διέλευσης (ZCR) είναι ένα από τα απλούστερα και ευκολότερα χαρακτηριστικά ώς πρός τον υπολογισμό του σε ένα ηχητικό σήμα. Ο ρυθμός αυτός ενός παραθυρωμένου σήματος είναι ο αριθμός των φορών που το σήμα αλλάζει από θετικές τιμές σε αρνητικές και το αντίθετο, διαιρούμενο με το μήκος του παραθύρου [29]. Με άλλα λόγια, ο ρυθμός με τον οποίο το πλάτος του σήματος περνάει από την μηδενική τιμή. Ο μηδενικός ρυθμός διέλευσης είναι συνήθως μια μέτρηση που σχετίζεται με το πόσο θορυβώδες μπορεί να είναι ένα σήμα. Τα θορυβώδη σήματα τείνουν να επιδεικνύουν υψηλότερες τιμές ενώ τα σήματα που περιλαμβάνουν φωνήεντα για παράδειγμα παρουσιάζουν χαμηλότερες τιμές. Η ευκολία στον υπολογισμό του σε συνδυασμό με τις πληροφορίες θορύβου ενός σήματος καθιστούν το ZCR ένα χαρακτηριστικό που δεν μπορεί να λείψει από οποιαδήποτε εργασία ταξινόμησης ήχου.

- **Τέμπο**

  Το τέμπο είναι βασικά ο ρυθμός ενός ηχητικού σήματος και μετράται σε παλμούς ανά λεπτό (BPM). Είναι ένας όρος που σχετίζεται με τη μουσική κυρίως και συνήθως χρησιμοποιείται σε ζητήματα ταξινόμησης μουσικών ειδών. Παρ'όλα αυτά, η καταγραφή

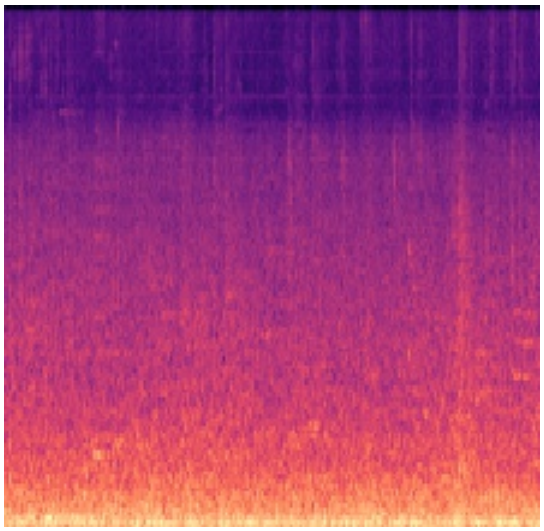
του ρυθμού των ηχητικών σημάτων από διαφορετικές ακουστικές σκηνές είναι χρήσιμη για την αναγνώριση μοτίβων που συμβάλλουν στην διάκριση ορισμένων ηχητικών τοπίων από κάποια άλλα [27].

- **Φασματόγραμμα κλίμακας Mel**

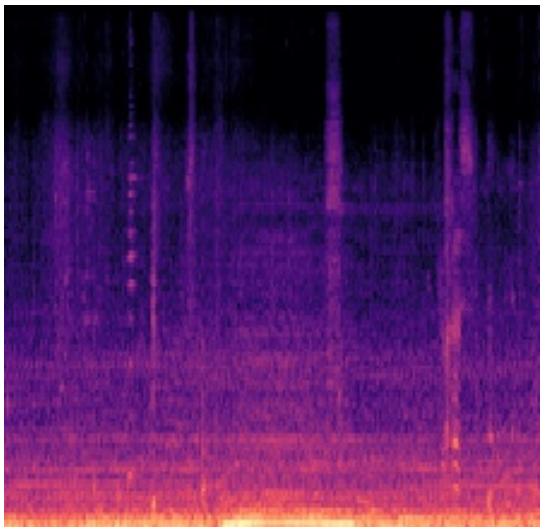
  Κάθε σήμα αποτελείται από διάφορα ηχητικά κύματα συχνότητας που περιλαμβάνουν πολύτιμες πληροφορίες για οποιοδήποτε ηχητικό περιβάλλον. Αυτές οι συχνότητες μπορούν να απεικονιστούν χρησιμοποιώντας ένα φασματόγραμμα. Για να δημιουργηθεί ένα φασματόγραμμα, πρώτα πρέπει να έχει προηγηθεί η μετατροπή του σήματος από το πεδίου του χρόνου στο πεδίο της συχνότητας χρησιμοποιώντας τον γρήγορο μετασχηματισμό Fourier (Fast Fourier Transform), έναν αλγόριθμο που υπολογίζει τον μετασχηματισμό Fourier ενός σήματος. Δεδομένου ότι τα ηχητικά σήματα είναι μη περιοδικά σήματα, πρέπει επίσης να εκτελέσουμε μια ανάλυση σύντομου χρόνου με σκοπό την αναπαράσταση των διαφόρων συχνοτήτων κατά τη διάρκεια των χρονικών παραθύρων. Ένα φασματόγραμμα κλίμακας Mel είναι ένα φασματόγραμμα με συχνότητες που μετατρέπονται στην κλίμακα Mel. Αυτή η κλίμακα επινοήθηκε διότι οι άνθρωποι είναι ικανοί να ανιχνεύουν διαφορές στον ήχο πιο καλά στις χαμηλότερες συχνότητες από ότι στις υψηλότερες. Για παράδειγμα, είναι εύκολο για κάποιον να διακρίνει την διαφορά μεταξύ δύο συχνοτήτων 300 και 800 Hz, αλλά πολύ δύσκολο να διακρίνει μια συχνότητα 15000 Hz από μια συχνότητα 15500 Hz, παρόλο που τα δύο ζεύγη έχουν την ίδια διαφορά. Συμπερασματικά, ένα φασματόγραμμα Mel περιέχει πληροφορίες ιδιαίτερα χρήσιμες για τη διάκριση μεταξύ δύο διαφορετικών ακουστικών σκηνών, όπως μπορούμε να παρατηρήσουμε στα σχήματα 3.2 και 3.3. Παρομοίως με όλα τα άλλα εξαγόμενα χαρακτηριστικά, η διαδικασία για τη δημιουργία ενός φασματογράμματος Mel όπως περιγράφηκε προηγουμένως μπορεί να επιτευχθεί με την βοήθεια της βιβλιοθήκης Librosa.

- **Φασματικό κεντροειδές**

  Το φασματικό κεντροειδές είναι μια μονάδα μέτρησης της θέσης και του σχήματος του φάσματος. Υποδεικνύει το κέντρο βαρύτητας του φάσματος [29] και συνήθως οι υψηλές τιμές του φασματικού κεντροειδούς αντιστοιχούν σε οξείς ήχους. Με τον όρο αυτόν εννοούμε ήχους που μπορεί να είναι ενοχλητικοί για το ανθρώπινο αυτί, όπως για παράδειγμα ένα αεροπλάνο που απογειώνεται. Το φασματικό κεντροειδές υπολογίζεται από τον τύπο 3.4 με X_i(k), k=0,1,...N-1 να είναι οι διακριτοί συντελεστές



Σχήμα 3.2: Φασματόγραμμα Mel από ακουστική σκηνή αεροδρομίου στο Ελσίνκι



Σχήμα 3.3: Φασματόγραμμα Mel από ακουστική σκηνή λεωφορείου στην Βαρκελώνη

μετασχηματισμού Fourier του σήματος x_i(n), n=0,1...N-1 του i παραθύρου [30].

$$C_i = \frac{\sum_{k=0}^{N-1} k|X_i(k)|}{\sum_{k=0}^{N-1} |X_i(k)|} \tag{3.4}$$

- **Φασματικό εύρος ζώνης**

  Το φασματικό εύρος ζώνης κάθε τμήματος ήχου εξάγεται μέσω της Librosa σε κάθε παράθυρο του σήματος και καθορίζει το εύρος μεταξύ των υψηλότερων και χαμηλότερων τιμών συχνότητας του σήματος [31]. Η μονάδα μέτρησης του φασματικού εύρους ζώνης είναι τα Hertz.

- **Χαρακτηριστική εξασθένηση φάσματος (spectrum roll-off)**

  Σύμφωνα με το [30], μόνο ένα ποσοστό κοντά στο 80-90% της κατανομής μεγέθους φάσματος συγκεντρώνεται στο φάσμα. Η συχνότητα κάτω από την οποία συγκεντρώνεται αυτό το ποσοστό στο φάσμα είναι γνωστή ως χαρακτηριστική εξασθένηση φάσματος και δίνει πολλές πληροφορίες σχετικά με το ηχητικό σήμα.

- **Φασματική αντίθεση**

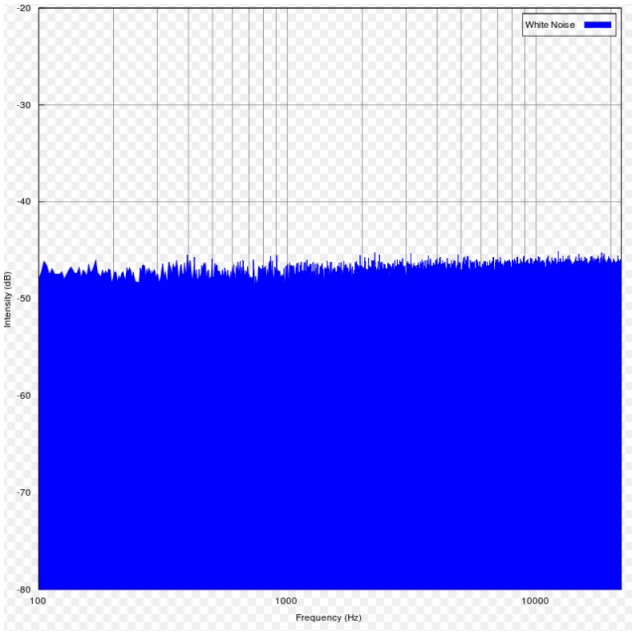
  Ένα άλλο χρήσιμο χαρακτηριστικό του τομέα συχνοτήτων που περιέχει πολλές πληροφορίες είναι η φασματική αντίθεση [32]. Η φασματική αντίθεση είναι βασικά η διαφορά στα dB μεταξύ κορυφών και κοιλάδων στο φάσμα. Οι κορυφές συνήθως αντιπροσωπεύουν αρμονικούς ήχους του περιβάλλοντος, όπως για παράδειγμα το κελαήδισμα των πουλιών, ενώ οι φασματικές κοιλάδες αντιστοιχούν σε θορυβώδη μέρη του σήματος [33]. Έτσι, η εξαγωγή της φασματικής αντίθεσης μας βοηθά στη διάκριση του θορύβου από τους αρμονικούς ήχους του φυσικού περιβάλλοντος.

- **Φασματική επιπεδότητα (Spectral flatness)**

  Η φασματική επιπεδότητα είναι επίσης ένα χαρακτηριστικό που βοηθά στη διάκριση των συνηθισμένων ήχων από τον θόρυβο. Ονομάζεται επίσης συντελεστής τόνου [9] και περιγράφει τον αριθμό των υφιστάμενων κορυφών σε ένα φάσμα ισχύος σε αντίθεση με το επίπεδο φάσμα ενός λευκού θορύβου. Ο λευκός θόρυβος όπως φαίνεται στο σχήμα 3.4 έχει την ίδια ποσότητα κατονομής ισχύος σε όλες τις φασματικές ζώνες και οι φασματικές τιμές επιπεδότητας πλησιάζουν τη μονάδα. Αντιθέτως, μια φασματική επιπεδότητα κοντά στο μηδέν αντιστοιχεί σε ημιτονοειδή σήματα που συνήθως αντιπροσωπεύουν αρμονικούς ήχους. Η φασματική επιπεδότητα υπολογίζεται από τη διαίρεση του γεωμετρικού μέσου όρου του φάσματος ισχύος με τον αριθμητικό μέσο όρο του φάσματος.

- **Τονικά κεντροειδή χαρακτηριστικά**

  Με αφορμή μια πρόταση που συναντήσαμε στο [34] επιλέξαμε να εξάγουμε τα τονικά κεντροειδή χαρακτηριστικά των ηχητικών σημάτων και να τα προσθέσουμε στο πλαίσιο δεδομένων με στόχο την επίτευξη της μέγιστης δυνατής ακρίβειας ταξινόμησης. Τα χαρακτηριστικά αυτά σχετίζονται κυρίως με μουσικά σήματα και η εξαγωγή τους από την Librosa επιτεύχθηκε χρησιμοποιώντας την αντίστοιχη συνάρτηση. Η συνάρτηση αυτή χρησιμοποιεί έναν αλγόριθμο που περιγράφεται στο [35] και ανιχνεύει τις



Σχήμα 3.4: Φάσμα λευκού θορύβου. (Εικόνα από [9]).

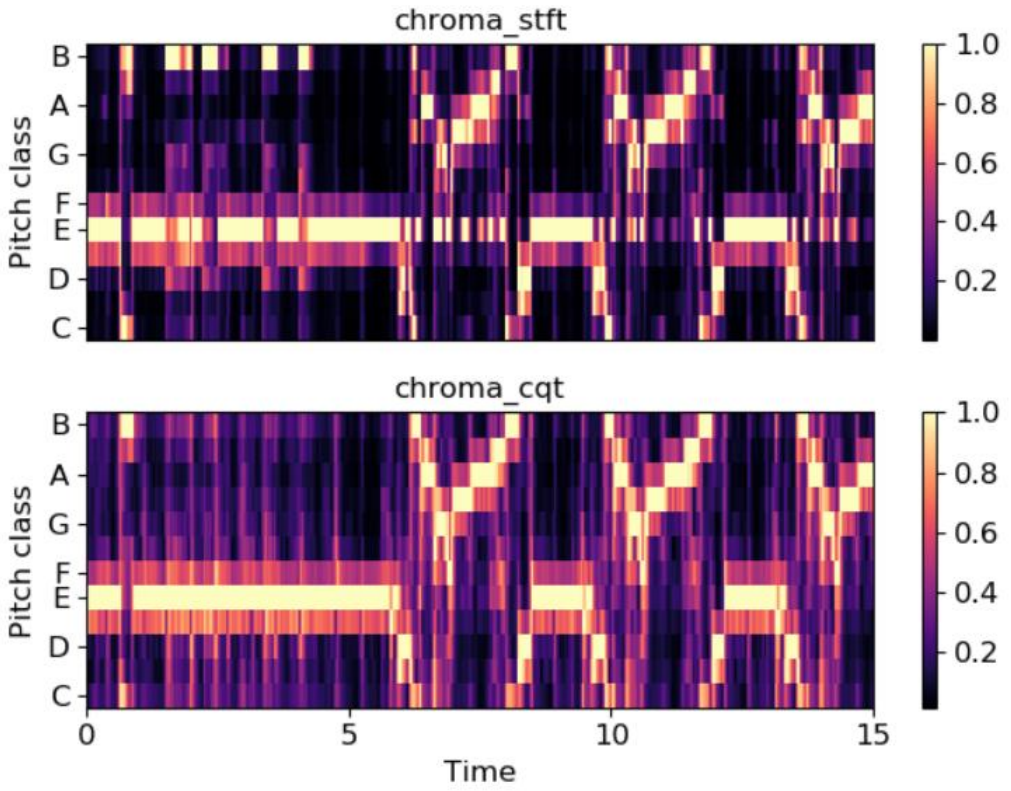
αρμονικές αλλαγές σε ένα ηχητικό σήμα. Η αρμονία ενός ηχητικού σήματος ορίζεται επίσης ως συγχρονισμένες δομές βήματος (synchronous pitch structures) [36].

- **Χρωματόγραμμα (chromagram)**

  Το χρωματόγραμμα αφορά τις 12 διαφορετικές κλάσεις περιοδικής συχνότητας (pitch) στις οποίες μπορεί να κατηγοριοποιηθεί ένα ηχητικό σήμα (C, C#, D, D#, E, F, F#, G, G#, A, A#, B) [37]. Το χαρακτηριστικό αυτό, παρέχει έναν τρόπο αναγνώρισης των περιοδικών συχνοτήτων που διαφέρουν κατά μια οκτάβα, αφού η ανθρώπινη ακοή δεν διαθέτει αυτή την ικανότητα. Με τον υπολογισμό του μετασχηματισμού Fourier σύντομου χρονικού διαστήματος του χρωματογράμματος (STFT) και την προσθήκη του ως χαρακτηριστικό εξαγωγής, ενισχύουμε την ευρωστία του συστήματός μας σε παραλλαγές στην χροιά, η οποία μπορεί να δείξει στοιχεία της ”συμπεριφοράς” της κάθε ακουστικής σκηνής.

- **Χρωματόγραμμα σταθεράς Q (Constant Q chromagram)**

  Ένα χρωματόγραμμα που μετασχηματίζεται στο πεδίο των συχνοτήτων χρησιμοποιώντας την μετατροπή σταθεράς Q (Constant Q Transform) αντί του STFT ονομάζεται χρωματόγραμμα σταθεράς Q. Ο μετασχηματισμός Q [38] παρομοίως με την κλίμακα Mel χρησιμοποιεί έναν άξονα λογαριθμικών τιμών, παρέχοντας μια καλύτερη προσομοίωση του ανθρώπινου ακουστικού συστήματος όπου σε χαμηλές συχνότητες η φασματική ανάλυση είναι πιο ακριβής.



Σχήμα 3.5: Σύγκριση STFT και CQT χρωματογραμμάτων του ίδιου ηχητικού αρχείου. (Εικόνα από [10]).

- **Κανονικοποιημένες στατιστικές χρωματικής ενέργειας (CENS)**

  Οι CENS είναι ορισμένα ιδιαιτέρως κοινά χαρακτηριστικά που χρησιμοποιούνται κυρίως σε ζητήματα αντιστοίχισης ήχου. Η βασική ιδέα είναι ότι η καταγραφή των στατιστικών της κατανομής ενέργειας σε μεγάλα χρονικά παράθυρα, εξομαλύνει τυχόν αποκλίσεις στο ρυθμό ή στην άρθρωση του ήχου [39].

Όπως αναφέρθηκε προηγουμένως, όλα αυτά τα χαρακτηριστικά που μόλις αναλύσαμε εξήχθησαν ανά παράθυρο με μέγεθος παραθύρου περίπου 93 ms και βήμα αναπήδησης στα 23 ms. Πιο συγκεκριμένα κάθε αρχείο ήχου (14400 συνολικά) των 10 δευτερολέπτων χωρίζεται σε 431 παράθυρα και για κάθε ένα από αυτά τα παράθυρα έχουμε τιμές 14 διαφορετικών εξαγόμενων χαρακτηριστικών. Φυσικά αυτό οδηγεί σε ένα τεράστιο και πολύπλοκο πλαίσιο δεδομένων που απαιτεί μεγάλο χρόνο εκπαίδευσης και έχει επίσης μεγάλο υπολογιστικό κόστος. Έτσι, για να μειώσουμε τον χρόνο εκπαίδευσης και την πολυπλοκότητα των υπολογισμών, αποσπάσαμε την μέση τιμή και την τυπική απόκλιση των παραθύρων της κάθε ηχητικής εγγραφής.

3.1 Μείωση διαστάσεων

Είναι επίσης σημαντικό να αναφέρουμε ότι προτού τροφοδοτήσουμε τα επεξεργασμένα δεδομένα στο back-end σύστημα του νευρωνικού δικτύου, υπάρχει ανάγκη να προσθέσουμε έναν αλγόριθμο μείωσης διαστάσεων, προκειμένου να μειώσουμε την πολυπλοκότητα των εισόδων, διατηρώντας παράλληλα όσο το δυνατόν περισσότερες πληροφορίες. Οι δύο πιο συνηθισμένες τεχνικές μείωσης διαστάσεων και αυτές που προσπαθήσαμε να εφαρμόσουμε στα δεδομένα εισόδου μας είναι η ανάλυση κύριων συνιστωσών (Principal Component Analysis) και η ανάλυση γραμμικών διακρίσεων (Linear Discriminant Analysis). Στην PCA, πραγματοποιείται γραμμική χαρτογράφηση των δεδομένων σε χώρο χαμηλότερων διαστάσεων με αποτέλεσμα τη μεγιστοποίηση της διακύμανσης των δεδομένων [40]. Η LDA από την άλλη πλευρά, προσπαθεί να μοντελοποιήσει τις διαφορές μεταξύ των τάξεων βρίσκοντας έναν γραμμικό συνδυασμό χαρακτηριστικών που περιγράφουν μια κλάση και την διαχωρίζουν από τις άλλες [41]. Τώρα, όσον αφορά το σύστημα ταξινόμησής μας, η PCA δεν φάνηκε να βελτιώνει την ακρίβεια ταξινόμησης που επιτεύχθηκε, σε αντίθεση με την LDA η οποία φάνηκε να ταιριάζει εξαιρετικά με το σύνολο των δεδομένων μας βελτιώνοντας την ακρίβεια του συστήματός μας κατά 5%.

Κεφάλαιο 4

Αρχιτεκτονικές συστήματος

4.1 Αρχική αρχιτεκτονική

Σε αυτό το κεφάλαιο θα πραγματοποιηθεί λεπτομερής ανάλυση όλων των στοιχείων της αρχιτεκτονικής νευρωνικού δικτύου που κατασκευάστηκε με σκοπό την αναγνώριση ακουστικών σκηνών. Ένα διαδοχικό (sequential), πλήρως συνδεδεμένο (fully connected), πυκνό (dense) μοντέλο νευρωνικού δικτύου δημιουργήθηκε χρησιμοποιώντας το Keras, ένα API της Python με εξειδίκευση στην βαθιά μάθηση. Ο όρος πλήρως συνδεδεμένο είναι γνωστός στον αναγνώστη από το κεφάλαιο 2 και σημαίνει ότι κάθε νευρώνας σε ένα στρώμα συνδέεται με κάθε νευρώνα στο επόμενο στρώμα. Το ”διαδοχικό” μέρος σημαίνει ότι το δίκτυό μας αποτελεί μια στοίβα στρώσεων όπου κάθε στρώμα έχει ακριβώς έναν τανυστή εισόδου και έναν τανυστή εξόδου. Ένας τανυστής (tensor) είναι μια δομή δεδομένων που χρησιμοποιείται στη μηχανική μάθηση και μπορεί να θεωρηθεί ως ένας πολυδιάστατος πίνακας και χρησιμοποιείται από το δίκτυο για εκπαιδευτικούς και λειτουργικούς σκοπούς. Το διαδοχικό μοντέλο είναι κατάλληλο για προβλήματα ταξινόμησης που εμπεριέχουν πολλαπλές κλάσεις. Κάθε στρώμα πρέπει να επικοινωνεί με τα άλλα και όπου η ύπαρξη μιας μη γραμμικής τοπολογίας είναι αναγκαία, επιλογές διακλάδωσης (branch) δίνονται στο δίκτυο [42].

Το μοντέλο που δημιουργήσαμε για την ταξινόμησή μας αποτελείται από 5 πυκνά στρώματα συμπεριλαμβανομένων αυτών της εισόδου και της εξόδου. Έπειτα από πολυάριθμα πειράματα, καταλήξαμε στο συμπέρασμα ότι το στρώμα εισόδου πρέπει να αποτελείται από 512 νευρώνες και ότι η συνάρτηση ενεργοποίησης που μας βοηθά να επιτύχουμε τη μέγιστη δυνατή ακρίβεια ταξινόμησης είναι SELU, η οποία περιγράφηκε εκτενώς στο κεφάλαιο 2. Η SELU χρησιμοποιήθηκε επίσης ως συνάρτηση ενεργοποίησης των κρυφών στρωμά-

των, αλλά ο αριθμός των νευρώνων διακυμάνθηκε από το ένα στρώμα στο άλλο έτσι ώστε να επιτευχθεί η μέγιστη ακρίβεια ταξινόμησης. Το πρώτο κρυφό στρώμα αποτελείται από 256 νευρώνες ενώ τα υπόλοιπα δύο περιλαμβάνουν από 128 νευρώνες το καθένα. Τέλος, το στρώμα εξόδου που είναι επίσης πυκνό, θα πρέπει πάντα να έχει τον ίδιο αριθμό νευρώνων με τις κλάσεις στις οποίες πρέπει να ταξινομήσουμε τα δεδομένα, δηλαδή 10 στην περίπτωσή μας. Η συνάρτηση ενεργοποίησης που επιλέχθηκε για το στρώμα εξόδου, με την οποία ο αναγνώστης είναι επίσης εξοικειωμένος από το κεφάλαιο 2, είναι η Softmax. Οι δύο πιο κοινές συναρτήσεις που χρησιμοποιούνται στα στρώματα εξόδου είναι η Sigmoid και η Softmax, οι οποίες δίνουν ως έξοδο έναν πραγματικό αριθμό μεταξύ 0 και 1 που αντιπροσωπεύει την πιθανότητα ενός αρχείου εισόδου να ανήκει σε μια συγκεκριμένη κλάση. Η συνάρτηση ενεργοποίησης Softmax απέδωσε ιδιαιτέρως ικανοποιητικά αποτελέσματα συγκριτικά με την Sigmoid και αυτός είναι ο λόγος για τον οποίο προτιμήθηκε. Επιπρόσθετα, στα κρυμμένα στρώματα ο αρχικοποιητής βάρους LecunNormal [43] προστέθηκε στο σύστημα, καθώς αποδείχθηκε ότι βελτιώνει την απόδοση του συστήματος όταν συνδυάζεται με την συνάρτηση ενεργοποίησης SELU.

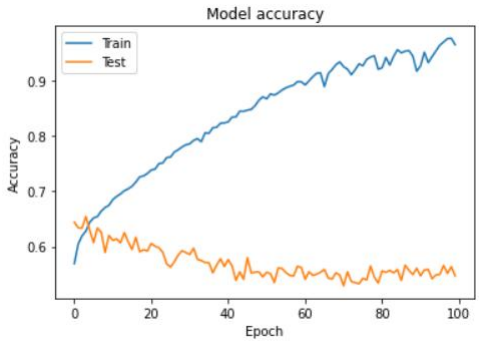
Κάθε νευρωνικό δίκτυο βαθιάς μάθησης εκπαιδεύεται χρησιμοποιώντας έναν στοχαστικό αλγόριθμο καθόδου κλίσης (stochastic gradient descent). Αυτός ο αλγόριθμος υποδεικνύει ότι το σφάλμα πρόβλεψης πρέπει να εκτιμάται σε κάθε κατάσταση του μοντέλου. Για το λόγο αυτό, το σύστημα χρειάζεται μια συνάρτηση σφάλματος ή αλλιώς γνωστή ως **loss function** δηλαδή συνάρτηση απώλειας σε ελεύθερη μετάφραση, για να βοηθήσει τα βάρη να ενημερωθούν σωστά, προκειμένου να μειωθούν οι απώλειες στις προσεχείς αξιολογήσεις. Η συνάρτηση απώλειας που ταιριάζει με το σύστημά μας είναι αυτή της αραιής κατηγορικής διασταυρούμενης εντροπίας (Sparse Categorical Cross-Entropy). Οι συναρτήσεις απώλειας διασταυρούμενης εντροπίας χρησιμοποιούνται σε προβλήματα ταξινόμησης πολλαπλών κλάσεων [44]. Σε αυτές τις συναρτήσεις, η προβλεπόμενη πιθανότητα κλάσης συγκρίνεται με την πραγματική κλάση και μια τιμή ποινής υπολογίζεται με βάση το πόσο απέχει η πρόβλεψή μας από το να είναι ορθή [45]. Η απώλεια υπολογίζεται από τον τύπο 4.1 με το t_i να αντιστοιχεί στην πραγματική ετικέτα της κλασης και p_i να είναι πιθανότητα που υπολογίζει η Softmax για την κλάση i. Οι τιμές απώλειας κοντά στο 0 δηλώνουν πολύ μικρές διαφορές μεταξύ των προβλεπόμενων και των σωστών ετικετών, ενώ τιμές κοντά στο 1 δείχνουν να υπάρχει σημαντική απόκλιση της προβλεπόμενης ετικέτας σε σχέση με την πραγματική.

$$L = -\sum_{i=1}^{10} t_i \log p_i \tag{4.1}$$

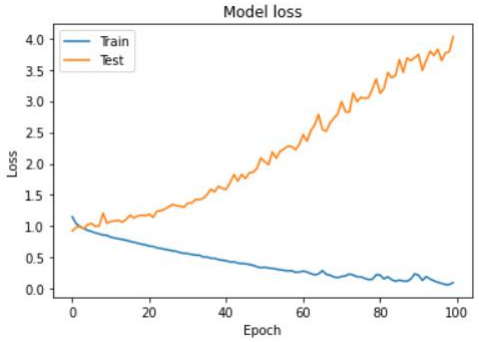
Όπως ήδη αναφέρθηκε παραπάνω, κάθε νευρωνικό δίκτυο χρησιμοποιεί έναν στοχαστικό αλγόριθμο καθόδου κλίσης για την ενημέρωση των βαρών του δικτύου με βάση τα δεδομένα εκπαίδευσης, κατά τη διάρκεια της διαδικασίας της οπισθοδιάδοσης. Ο αλγόριθμος που εφαρμόσαμε στο δίκτυο ονομάζεται **Adam** και είναι μια βελτιωμένη έκδοση του στοχαστικού αλγορίθμου καθόδου κλίσης (SGD) [46]. Ο Adam αποδεικνύεται κατάλληλος για προβλήματα με μεγάλα δεδομένα και ”θορυβώδεις” κλίσεις, συνδυάζοντας τα πλεονεκτήματα δύο άλλων αλγορίθμων βελτιστοποίησης, του αλγορίθμου προσαρμοστικής κλίσης (Adagrad) και της μέσης τετραγωνικής διάδοσης ρίζας (RMSProp) [47] [48]. Όπως θα δούμε αργότερα σε αυτή τη διπλωματική, τα αποτελέσματα του δικτύου μας ήταν συντριπτικά υπέρ του Adam σε σύγκριση με τους υπόλοιπους αλγόριθμους βελτιστοποίησης.

4.1.1 Πρόβλημα overfitting

Η αρχιτεκτονική του συστήματός μας μέχρι στιγμής αποτελείται από το σύστημα front-end που περιγράφεται στο κεφάλαιο 3 και το μοντέλο ταξινόμησης πυκνού νευρωνικού δικτύου όπως περιγράφεται στις παραπάνω παραγράφους. Τα αποτελέσματα της αξιολόγησης αυτού του συστήματος φαίνονται στο σχήμα 4.1, όπου μπορούμε να παρατηρήσουμε την ακρίβεια των συνόλων εκπαίδευσης και δοκιμής του συστήματος σε 100 εποχές (epochs) και στο σχήμα 4.2 που απεικονίζει την γραφική παράσταση των απωλειών των δύο συνόλων δεδομένων στον ίδιο αριθμό εποχών. Όσον αφορά την πρώτη γραφική παράσταση, παρατηρούμε ότι καθώς η ακρίβεια του συνόλου εκπαίδευσης αυξάνεται κατά τη διάρκεια των εποχών, η ακρίβεια ταξινόμησης του συνόλου δοκιμών μειώνεται δραματικά. Αντίθετα, όταν οι απώλειες του σετ εκπαίδευσης μειώνονται, οι απώλειες του σετ δοκιμών αυξάνονται, πράγμα που σημαίνει ότι το σφάλμα πρόβλεψης γίνεται όλο και μεγαλύτερο με κάθε εποχή. Αυτό το πρόβλημα είναι κοινό στον τομέα της μηχανικής μάθησης και είναι γνωστό ως **overfitting**. Το φαινόμενο αυτό παρουσιάζεται όταν ένα σύστημα νευρωνικού δικτύου ”ταιριάζει” (fits) υπερβολικά καλά στο σετ εκπαίδευσης. Αυτό έχει ως αποτέλεσμα την εκμάθηση κάθε λεπτομέρειας και θορύβου με τρόπο που επηρεάζει αρνητικά την απόδοση του συστήματος σε οποιαδήποτε δεδομένα διαφορετικά από αυτά στα οποία εκπαιδεύεται [49].

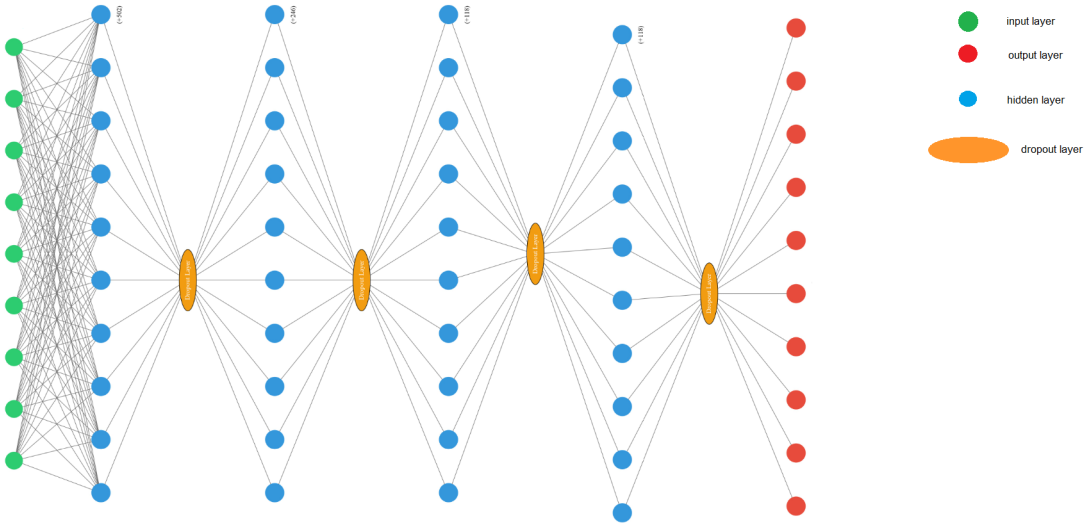


Σχήμα 4.1: Ακρίβεια ταξινόμησης σε 100 εποχές



Σχήμα 4.2: Απώλειες ταξινόμησης σε 100 εποχες

Δεδομένου ότι το μοντέλο μας δεν έχει ικανότητες προσαρμογής σε νέα δεδομένα, δημιουργήθηκε μεγάλη ανάγκη να απαλλαγούμε από το φαινόμενο του overfitting. Ένας υπολογιστικά φθηνός και αποτελεσματικός τρόπος αντιμετώπισης του προβλήματος αυτού είναι η προσθήκη **dropout** στρωμάτων ή αλλιώς στρωμάτων εγκατάλειψης σε ελληνική ορολογία. Τα στρώματα αυτά αφαιρούν προσωρινά τους νευρώνες από το δίκτυο κατά τη διάρκεια της εκπαιδευτικής διαδικασίας δίνοντας μηδενικές τιμές στις μονάδες εισόδου. Αυτό έχει ως αποτέλεσμα να καταστεί η διαδικασία της εκπαίδευσης θορυβώδης, αναγκάζοντας ορισμένους νευρώνες να αναλάβουν πιθανώς μεγαλύτερη ευθύνη για τις εισροές, ενώ κάποιοι άλλοι παραμένουν ανενεργοί [50]. Βασικά, ένα στρώμα dropout προσομοιώνει μια αραιή ενεργοποίηση των εισόδων που λαμβάνει από ένα άλλο στρώμα, το οποίο έχει ως αποτέλεσμα το δίκτυο να μαθαίνει μια αραιή αναπαράσταση των δεδομένων εισόδων. Επίσης, με την προσθήκη αυτών των στρωμάτων μειώνεται η χωρητικότητα του δικτύου κατά τη διάρκεια της εκπαίδευσης, ενώ τα βάρη γίνονται μεγαλύτερα από το συνηθισμένο. Η συχνότητα με την οποία οι μονάδες εισόδου είναι ρυθμισμένες στο 0 καθορίζεται από μια υπερπαράμετρο P που κυμαίνεται μεταξύ 0 και 1 και αντιπροσωπεύει το ρυθμό εγκατάλειψης (dropout rate). Στην αρχιτεκτονική μας, προσθέσαμε dropout στρώματα έπειτα από κάθε κρυφό στρώμα.



Σχήμα 4.3: Οπτικοποίηση της αρχιτεκτονικής του νευρωνικού δικτύου

Στο πρώτο στρώμα εγκατάλειψης που είναι το πλησιέστερο στο στρώμα εισόδου, η υπερπαράμετρος P συνήθως λαμβάνει τις μεγαλύτερες τιμές της για να ομαλοποιήσει τις αρχικές εισόδους, οπότε της δώσαμε την τιμη 0.8. Στα άλλα 3 στρώματα dropout δόθηκε ρυθμός εγκατάλειψης της τάξης του 0.5. Στο σχήμα 4.3 βλέπουμε μια απεικόνιση του νευρωνικού μας δικτύου μετά την εισαγωγή του dropout. Μπορούμε εύκολα να παρατηρήσουμε την επίδραση των στρωμάτων εγκατάλειψης συγκρίνοντας τις μονάδες εισόδου που λαμβάνονται από το πρώτο και το δεύτερο κρυφό στρώμα αντίστοιχα. Επίσης, είναι φανερό ότι οι διαστάσεις του στρώματος εισόδου μειώνονται λόγω του LDA αλγορίθμου που εφαρμόσαμε στα εξαγόμενα χαρακτηριστικά μας, τα οποία σε συνδυασμό με τα στρώματα εγκατάλειψης καθιστούν το δίκτυό μας λιγότερο περίπλοκο, πιο υπολογιστικά αποδοτικό και πιο ακριβές στις προβλέψεις του.

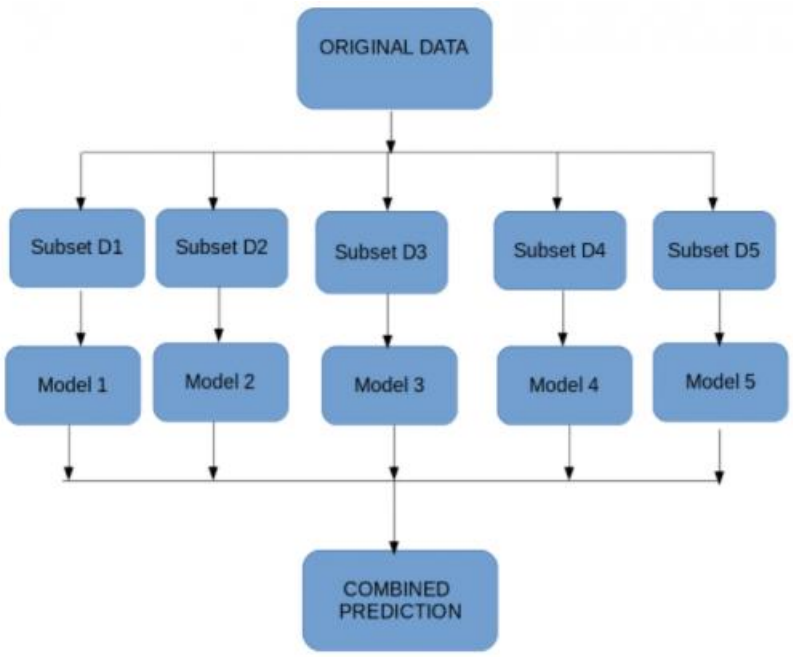
Επιπλέον, στην προσπάθειά μας να αντιμετωπίσουμε το πρόβλημα του overfitting και να επιτύχουμε τη μέγιστη δυνατή ακρίβεια ταξινόμησης, προσθέσαμε στο σύστημά μας μερικά επιπλέον στοιχεία. Πρώτον, επαναπροσδιορίσαμε το μοντέλο εφαρμόζοντας **batch normalization** ή κανονικοποίηση παρτίδας. Χρησιμοποιώντας κανονικοποίηση παρτίδας κατά τη διάρκεια της εκπαιδευτικής διαδικασίας, η κατανομή των εισόδων κάθε στρώματος διαφοροποιείται καθώς αλλάζουν οι παράμετροι του προηγούμενου στρώματος, καθιστώντας τις κλίσεις πιο ακριβείς στην πρόβλεψη [51]. Επίσης, χρησιμοποιήσαμε μια συνάρτηση της Python που ονομάζεται ”ModelCheckpoint”, η οποία μετά από κάθε εποχή, φορτώνει τα βάρη του μοντέλου με τη μέγιστη ακρίβεια στο σύνολο επικύρωσης μέχρι εκείνη τη στιγμή. Τελευταίο αλλά εξίσου σημαντικό εφαρμόσαμε στο σύστημά μας την λειτουργία του **early stopping**. Το

early stopping ουσιαστικά αποτελεί μια πρόωρη διακοπή της εκπαίδευσης του μοντέλου έτσι ώστε να αποφευχθεί το overfitting. Θέσαμε ένα όριο υπομονής 30 εποχών από την στιγμή που το μοντέλο παρουσιάσει τις ελάχιστες απώλειες στο σύνολο επικύρωσης, κάτι που σημαίνει ότι σε αυτό το σημείο το μοντέλο σταματάει να βελτιώνει την απόδοσή του και αρχίζει να ”ταιριάζει” υπερβολικά στα δεδομένα εκπαίδευσης. Οι ελάχιστες απώλειες επικύρωσης ενός μοντέλου, τις περισσότερες φορές ισοδυναμούν με μέγιστη ακρίβεια επικύρωσης.

4.2 Συνολική μάθηση (ensemble learning)

Τα αποτελέσματα της παρούσας αρχιτεκτονικής, τα οποία θα αναλυθούν εκτενώς στο επόμενο κεφάλαιο, ήταν ικανοποιητικά αλλά υπήρξε αρκετό περιθώριο για βελτίωση. Για αυτό τον λόγο, υιοθετήσαμε στην αρχιτεκτονική μας την φιλοσοφία της συνολικής μάθησης η οποία είναι γνωστή στον χώρο της μηχανικής μάθησης ως **ensemble learning** [11]. Η φιλοσοφία αυτής της προσέγγισης ουσιαστικά υποστηρίζει τον συνδυασμό των αποφάσεων πολλαπλών μοντέλων με σκόπο την βελτίωση της συνολικής απόδοσης του συστήματος. Οι τεχνικές συνολικής μάθησης είναι αρκετές και η πιο συνηθισμένη από αυτές είναι ο αλγόριθμος max voting όπου οι προβλέψεις της πλειοψηφίας των μοντέλων θεωρούνται ως η τελική πρόβλεψη του συστήματος. Γνωστή επίσης μέθοδος ensemble learning είναι το averaging το οποίο ουσιαστικά είναι μια επέκταση του max voting. Σε αυτήν την περίπτωση υπολογίζεται ο μέσος όρος απο τις προβλέψεις των επιμέρους μοντέλων με σκόπο την επίτευξη της τελικής πρόβλεψης.

Ωστόσο στο συστημά μας χρησιμοποίησαμε μια άλλη τεχνική συνολική μάθησης, η οποία είναι γνωστή ως ”bagging”. Σύμφωνα με αυτήν την τεχνική, τα αποτελέσματα πολλαπλών μοντέλων συνδυάζονται για να βγεί το τελικό αποτέλεσμα. Παρ’όλ’αυτά, για να έχει νόημα αυτό, τα επιμέρους μοντέλα θα πρέπει να διενεργήσουν σε διαφορετικά τμήματα του συνόλου των δεδομένων. Έτσι χωρίσαμε το αρχικό σύνολο δεδομένων σε 5 μικρότερα υποσύνολα ίδιου μεγέθους, τα οποία περιλαμβάνουν δεδομένα από δυο κλάσεις το καθένα (1837 αρχεία το σύνολο εκπαίδευσης και 837 το σύνολο δοκιμής). Ο διαχωρισμός έγινε έτσι ώστε οι κλάσεις που προβλέπονται εσφαλμένα πιο συχνά να ομαδοποιηθούν με κάποια κλάση που σημειώνει υψηλά ποσοστά επιτυχούς πρόβλεψης, με σκοπό τον διαμοιρασμό των σωστών προβλέψεων. Πιο συγκεκριμένα η κατανομή των δεδομένων παρουσιάζεται στον πίνακα 4.1. Έπειτα απο τον σχηματισμό των υποσυνόλων, ένα μοντέλο βάσης ίδιας αρχιτεκτονικής με



Σχήμα 4.4: Οπτική αναπαράσταση του αλγορίθμου συνολικής μάθησης. (Εικόνα από [11]).

το αρχικό μοντέλο δημιουργήθηκε για κάθε ένα από τα υποσύνολα. Τα 5 πλέον μοντέλα βάσης λειτούργησαν παράλληλα και ανεξάρτητα το ένα με το άλλο, δηλαδή εκπαιδεύτηκαν και πραγματοποίησαν προβλέψεις το καθένα ανάλογα με το υποσύνολό του. Η τελική πρόβλεψη του συστηματός μας προήλθε από τον συνδυασμό των προβλέψεων των μοντέλων βάσης. Για την καλύτερη κατανόηση του ensemble learning αλγορίθμου που εντάξαμε στο σύστημά μας στο σχήμα 4.4 παρουσιάζεται μια οπτική αναπαράστασή του.

Πίνακας 4.1: Κατανομή δεδομένων συνολικής μάθησης

Υποσύνολο	Κλάσεις
1	Αεροδρόμιο-Λεωφορείο
2	Μετρό-Σταθμός μετρό
3	Πάρκο-Δημόσια πλατεία
4	Εμπορικό κέντρο-Πεζόδρομος
5	Δρόμος οδικής κυκλοφορίας-Τράμ

Κεφάλαιο 5

Περιγραφή βάσης δεδομένων και αποτελέσματα

5.1 Περιγραφή βάσης δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήσαμε στα πλαίσια αυτής της διατριβής είναι το TAU Urban Acoustic Scenes 2019 [1]. Τα δεδομένα καταγραφής συλλέχθηκαν από το Πανεπιστήμιο τεχνολογίας του Τάμπερε και ανήκουν σε 10 διαφορετικές ακουστικές σκηνές που θα αποτελέσουν τις 10 προκαθορισμένες τάξεις. Αυτές οι ακουστικές σκηνές είναι: αεροδρόμιο, εμπορικό κέντρο αγορών, σταθμός μετρό, πεζόδρομος, δημόσια πλατεία, δρόμος με μέσο επίπεδο κυκλοφορίας, τραμ, λεωφορείο, μετρό, πάρκο. Τα δεδομένα καταγράφηκαν σε 12 μεγάλες ευρωπαϊκές πόλεις οι οποίες είναι: Άμστερνταμ, Βαρκελώνη, Ελσίνκι, Λισαβόνα, Λονδίνο, Λυών, Μαδρίτη, Μιλάνο, Πράγα, Παρίσι, Στοκχόλμη, Βιέννη. Για κάθε ακουστική σκηνή, ο ήχος καταγράφηκε σε διαφορετικές τοποθεσίες, πολλαπλά πάρκα, πολλαπλά αεροδρόμια, πολλαπλούς σταθμούς μετρό κλπ. Για κάθε τοποθεσία υπάρχουν 5-6 λεπτά από ήχητικά αρχεία, τα οποία χωρίστηκαν σε τμήματα των 10 δευτερολέπτων και παρέχονται σε μεμονωμένα αρχεία. Οι συσκευές που χρησιμοποιήθηκαν για τους σκοπούς της εγγραφής ήταν ένα μικρόφωνο δυο διάυλων και μια συσκευή εγγραφής ήχου σε συχνότητα δειγματοληψίας 48kHz και ανάλυση 24 bit. Οποιαδήποτε ηχητικά τμήματα που περιείχαν κοντινές συνομιλίες εξαλείφθηκαν, ενώ ορισμένες παρεμβολές από κινητά τηλέφωνα παρέμειναν στην εγγραφή καθώς θεωρήθηκαν μέρος των ηχοτοπίων.

Ο συνολικός αριθμός των ηχητικών αρχείων που περιέχονται στο σύνολο δεδομένων είναι 14400 και πιο συγκεκριμένα 1440 από κάθε σκηνή/τάξη, τα οποία έχουν καταγραφεί

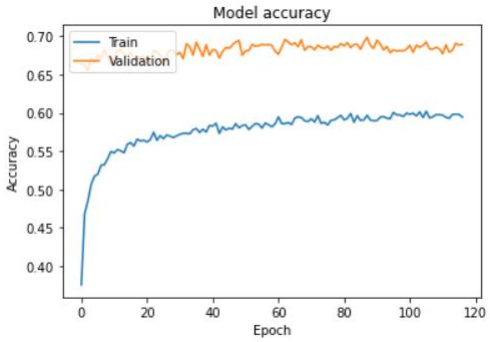
σε 514 διαφορετικές τοποθεσίες. Από αυτά τα τμήματα, 9185 προορίζονται για την εκπαίδευση του δικτύου, 4185 για την δοκιμή ενώ υπάρχουν και 1030 αχρησιμοποίητα τμήματα. Ο διαχωρισμός έχει γίνει με τέτοιο τρόπο ώστε τα τμήματα που καταγράφονται στην ίδια τοποθεσία να ανήκουν στο ίδιο υποσύνολο, κάτι που δημιουργεί ανισορροπία μεταξύ του συνόλου εκπαίδευσης και του συνόλου δοκιμών. Το εκπαιδευτικό σύνολο του συστήματός μας προέρχεται από τοποθεσίες σε 9 διαφορετικές πόλεις, ενώ το σετ που θα δοκιμαστεί το σύστημά μας περιλαμβάνει εγγραφές και από τις 10 πόλεις. Τα δεδομένα από το Μιλάνο χρησιμοποιούνται μόνο για σκοπούς δοκιμής, κάτι το οποίο έχει ως αποτέλεσμα ένα μέρος τους να μένει αχρησιμοποίητο δεδομένου ότι υπάρχει μια ισορροπημένη ποσότητα υλικού και από τις δέκα πόλεις. Πιο λεπτομερείς πληροφορίες σχετικά με τον διαχωρισμό των δεδομένων παρέχονται στον πίνακα 5.1.

Πίνακας 5.1: Στατιστικά στοιχεία δεδομένων

Ακουστική σκηνή	Εγγραφές εκπαίδευσης	Τοποθεσίες εγγραφών εκπαίδευσης	Εγγραφές δοκιμής	Τοποθεσίες εγγραφών δοκιμής	Αχρησιμοποίητες εγγραφές	Τοποθεσίες αχρησιμοποίητων εγγραφών
Αεροδρόμιο	911	25	421	12	108	3
Λεωφορείο	928	46	415	20	97	5
Μετρό	902	41	433	20	105	6
Σταθμός μετρό	897	37	435	17	108	3
Πάρκο	946	27	386	11	108	3
Δημόσια πλατεία	945	28	387	12	108	3
Εμπορικό κέντρο	896	24	441	10	103	2
Πεζόδρομος	924	29	429	14	87	3
Δρόμος, οδικής κυκλοφορίας	942	27	402	12	96	4
Τράμ	894	41	436	21	110	8
Σύνολο	9185	325	4185	149	1030	40

5.2 Αρχικά αποτελέσματα

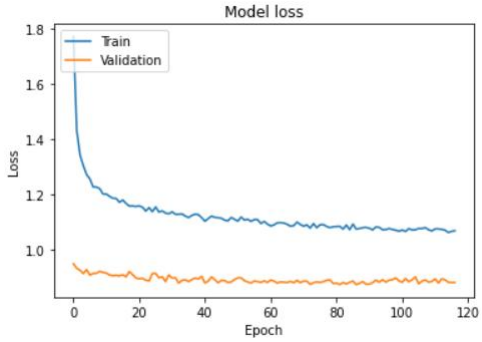
Σε αυτήν την ενότητα θα αναλύσουμε τα αποτελέσματα της αρχικής αρχιτεκτονικής του συστήματός μας, δηλαδή πρίν την ένταξη του ensemble learning αλγορίθμου. Όπως ανα-



Σχήμα 5.1: Γράφημα ακρίβειας μοντέλου

φέρθηκε στο κεφάλαιο 4, η αρχιτεκτονική μας περιλαμβάνει μηχανισμό πρόωρης διακοπής, ο οποίος σταματάει την εκπαίδευση αφού σημειωθούν 30 εποχές από την παρουσιάση των ελάχιστων απωλειών του μοντέλου στο σύνολο επικύρωσης. Στην περίπτωσή μας, οι ελάχιστες απώλειες στο σύνολο επικύρωσης είχαν τιμή 0.8728 και αυτό επιτεύχθηκε στον αριθμό εποχής 87 με ακρίβεια επικύρωσης 69.84%. Έτσι, λαμβάνοντας υπόψη την πρόωρη διακοπή, η εκπαιδευτική διαδικασία έληξε στην εποχή με αριθμό 117. Στο σχήμα 5.1 παρατηρούμε τα γραφήματα της ακρίβειας του μοντέλου στα σύνολα εκπαίδευσης και επικύρωσης ενώ το σχήμα 5.2 αντιπροσωπεύει τις απώλειες του μοντέλου σε αυτά τα σύνολα στις ίδιες εποχές. Όσον αφορά τα διαγράμματα ακρίβειας, παρατηρούμε ότι η απόδοση στο σύνολο επικύρωσης κυμαίνεται σταθερά μεταξύ 65% και 70% από τις πρώτες εποχές και αυτό οφείλεται κυρίως στην κανονικοποίηση παρτίδας. Όσον αφορά το γράφημα των απωλειών, παρατηρούμε ότι ενώ οι απώλειες στο σετ εκπαίδευσης μειώνονται εκθετικά φτάνοντας σε μια ελάχιστη τιμή που είναι πάνω από τη μονάδα, οι απώλειες στο σετ επικύρωσης παραμένουν σταθερά μεταξύ των τιμών 0.8 και 0.95 καθώς σημειώνουν μια μικρή μείωση όσο προχωρούν οι εποχές. Οι μεγάλες απώλειες στο σύνολο εκπαίδευσης οφείλονται στα dropout στρώματα, τα οποία είναι επίσης υπεύθυνα για τη βελτίωση της ακρίβειας επικύρωσης και τη μείωση των απωλειών. Σε γενικές γραμμές, αν συγκρίνουμε τα σχήματα 5.1 και 5.2 με τα 4.1 και 4.2, όπου το μοντέλο δεν περιλάμβανε dropout, batch normalization και την πρόωρη διακοπή της εκπαίδευσης, μπορούμε με ασφάλεια να φτάσουμε στο συμπέρασμα ότι η υπερβολικά καλή εκμάθηση των δεδομένων εισόδου στο σύνολο εκπαίδευσης ουσιαστικά έβλαπτε την απόδοση στο σύνολο επικύρωσης και κατ'επέκταση στο σύνολο δοκιμής.

Δεδομένου ότι η εκπαίδευση του νευρωνικού μας δικτύου ολοκληρώθηκε και το μοντέλο μας επικυρώθηκε, ήρθε η ώρα να ελέγξουμε την απόδοση του συστήματος μας στο σετ δοκιμών. Για τον σκοπό αυτόν, χρησιμοποιήσαμε συστήματα μέτρησης από τη βιβλιο-



Σχήμα 5.2: Γράφημα απωλειών μοντέλου

θήκη sklearn της Python. Αρχικά, χρησιμοποιήθηκε η αναφορά ταξινόμησης (classification report) [52] η οποία παρουσιάζεται στο σχήμα 5.3 και περιλαμβάνει τις κύριες μετρήσεις ταξινόμησης που είναι η ακρίβεια, η ανάκληση, η βαθμολογία f1 και η μια ακόμη στήλη που υποδεικνύει τον αριθμό των ηχητικών εγγραφών που ανήκουν σε κάθε κλάση. Αυτές οι μετρήσεις υπολογίζονται χρησιμοποιώντας αληθινά αρνητικά (true negatives) και ψευδώς θετικά (false positives). Η ακρίβεια για κάθε κατηγορία ορίζεται ως ο λόγος των πραγματικών θετικών στο σύνολο των πραγματικών θετικών και ψευδών θετικών ή με άλλα λόγια σε ποιο ποσοστό οι προβλέψεις μας ήταν σωστές. Το μοντέλο μας σημείωσε ακρίβεια 71% στο σύνολο δοκιμών. Η ανάκληση από την άλλη πλευρά, εκφράζει την αναλογία των αληθινών θετικών στο άθροισμα των αληθινών θετικών και των ψευδών αρνητικών και το σύστημά μας απέδωσε με 70%. Τελευταίο αλλά εξίσου σημαντικό, το μοντέλο μας έφτασε επίσης σε 70% ακρίβεια στο F1-score που είναι η βασική μετρική όταν πρόκειται για τη σύγκριση της ακρίβειας των μοντέλων ταξινόμησης. Έτσι, από την αναφορά ταξινόμησης παρατηρούμε ότι οι σκηνές του σταθμού του μετρό και της δημόσιας πλατείας είναι αυτές που συχνά ταξινομούνται εσφαλμένα ενώ οι κλάσεις των λεωφορείων, πάρκων, δρόμων οδικής κυκλοφορίες προβλέπονται σωστά σε ποσοστό άνω του 80%. Οι προβλέψεις για τις κλάσεις του αεροδρομίου, του μετρό, του εμπορικού κέντρου, του πεζόδρομου και του τραμ είναι πιο κοντά στη μέση ακρίβεια του μοντέλου μας, η οποία είναι περίπου 70% με βάση την αναφορά.

Για να επιβεβαιώσουμε τα αποτελέσματα της αναφοράς ταξινόμησης, δοκιμάσαμε ακόμα μια μετρική της βιβλιοθήκης sklearn η οποία είναι γνωστή ως confusion matrix [53]. Ο confusion matrix του συστήματος ταξινόμησης μας παρουσιάζεται στο σχήμα 5.4. Οι σειρές του πίνακα αυτού αντιπροσωπεύουν την ετικέτα που προέβλεψε το σύστημά μας ενώ οι στήλες αντιπροσωπεύουν την πραγματική ετικέτα της κλάσης. Για παράδειγμα, η πρόβλεψη που έκανε το δίκτυό μας σχετικά με την κλάση του πάρκου ήταν 88.6% σωστή. Ο confusion ma-

Classification Report

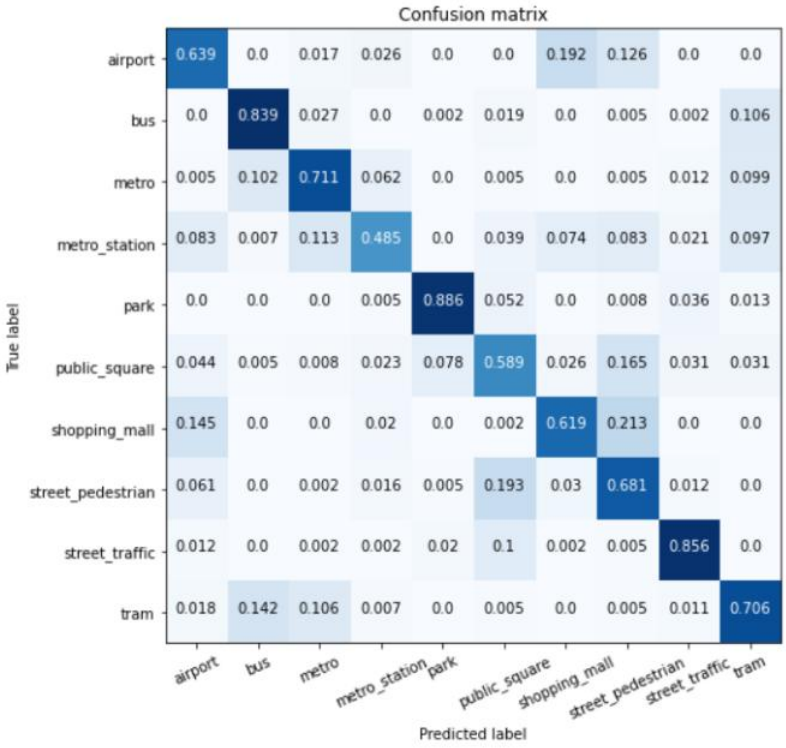
	precision	recall	f1-score	support
airport	0.63	0.64	0.63	421
bus	0.76	0.84	0.80	415
metro	0.72	0.71	0.72	433
metro_station	0.75	0.49	0.59	435
park	0.89	0.89	0.89	386
public_square	0.57	0.59	0.58	387
shopping_mall	0.67	0.62	0.64	441
street_pedestrian	0.53	0.68	0.60	429
street_traffic	0.87	0.86	0.86	402
tram	0.68	0.71	0.69	436
accuracy			0.70	4185
macro avg	0.71	0.70	0.70	4185
weighted avg	0.71	0.70	0.70	4185

Σχήμα 5.3: Αναφορά ταξινόμησης

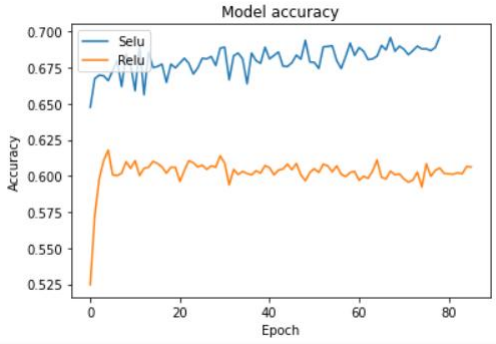
trix έρχεται να επιβεβαιώσει τα αποτελέσματα της αναφοράς ταξινόμησης, καθώς οι κλάσεις που προβλέπονται περισσότερο ή λιγότερο επιτυχώς, παραμένουν οι ίδιες.

5.2.1 Διάφορα πειράματα

Έως ότου το νευρωνικό μας δίκτυο να φτάσει τη μέγιστη δυνατή ακρίβεια ταξινόμησης, πειραματιστήκαμε με διαφορετικές συναρτήσεις ενεργοποίησης και αλγορίθμους βελτιστοποίησης. Αρχικά, δοκιμάσαμε διαφορετικούς συνδυασμούς συναρτήσεων ενεργοποίησης για την είσοδο και τα κρυμμένα στρώματα με τις ReLU και SELU να είναι μακράν οι πιο αποτελεσματικές. Στη συνέχεια, αποφασίσαμε να συγκρίνουμε αυτές τις δύο συναρτήσεις για να αποφασίσουμε ποια θα μας οδηγήσει στα καλύτερα αποτελέσματα. Στο σχήμα 5.5 παρατηρούμε την απόδοση στα δεδομένα συνόλου επικύρωσης των δύο συναρτήσεων που χρησιμοποιούνται στα κρυμμένα στρώματα και στο στρώμα εισόδου επίσης. Είναι αξιοσημείωτο ότι η SELU αποδίδει καλύτερα, παρόλο που ενεργοποιεί την πρόωρη διακοπή εκπαίδευσης σε πιο πρόσφατη εποχή. Για το στρώμα εξόδου, μόνο δύο συναρτήσεις ενεργοποίησης θα μπορούσαν να είναι κατάλληλες και δεν είναι άλλες από τις Sigmoid και Softmax. Το διάγραμμα 5.6 δείχνει ότι η Sigmoid δεν είναι σε θέση να ανταγωνιστεί την Softmax όταν πρόκειται για την τελική πρόβλεψη του δικτύου. Όσον αφορά τους αλγορίθμους βελτιστοποίησης, αναφέραμε και νωρίτερα ότι ο Adam είναι ο πιο κατάλληλος αλγόριθμος όταν πρόκειται για μεγάλα δεδομένα και θορυβώδεις κλίσεις καθώς συνδυάζει τα οφέλη του Adagrad και του RMSProp. Έτσι, στο διάγραμμα 5.7 παρουσιάζουμε μια σύγκριση του Adam με αυτούς τους δύο αλγορίθμους όπως και επίσης με τον κλασικό αλγόριθμο SGD αλλά και τον Adadelta.



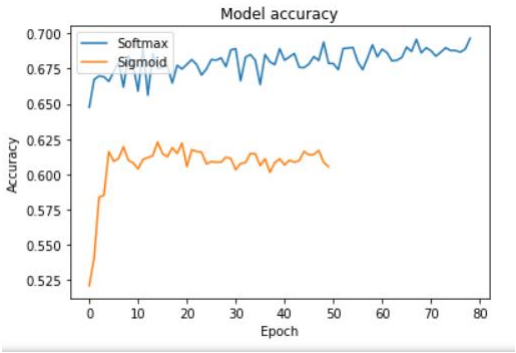
Σχήμα 5.4: Confusion matrix



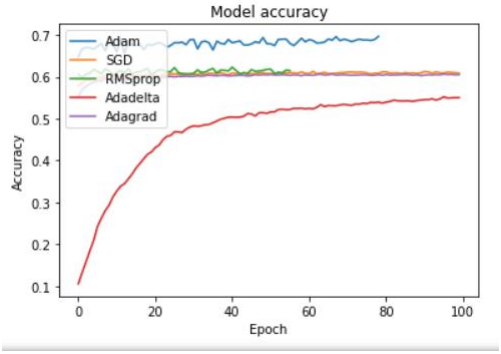
Σχήμα 5.5: Σύγκριση ReLU και SELU

Παρατηρούμε ότι, εκτός από τον Adadelta που έχει εκθετική βελτίωση στην απόδοσή του αλλά φτάνει μόνο λίγο πάνω από το 50%, οι άλλοι 3 αλγόριθμοι έχουν περίπου την ίδια συμπεριφορά που οδηγεί σε ακρίβεια ταξινόμηση της τάξης του 60%. Έτσι, ο Adam ως αλγόριθμος βελτιστοποίησης, η SELU ως συνάρτηση ενεργοποίησης της εισόδου και των κρυφών στρωμάτων και η Softmax να ενεργοποιεί το στρώμα εξόδου πραγματοποίησαν τον τέλειο συνδυασμό που οδήγησε το σύστημά μας στα αποτελέσματα που περιγράφονται στην παραπάνω ενότητα.

Παρόλο που τα αποτελέσματα του νευρωνικού μας δικτύου σημείωσσαν περίπου 10% βελτίωση σε σύγκριση με το σύστημα βασικής γραμμής του DCASE, σκηνές όπως ο σταθμός



Σχήμα 5.6: Σύγκριση Sigmoid και Softmax

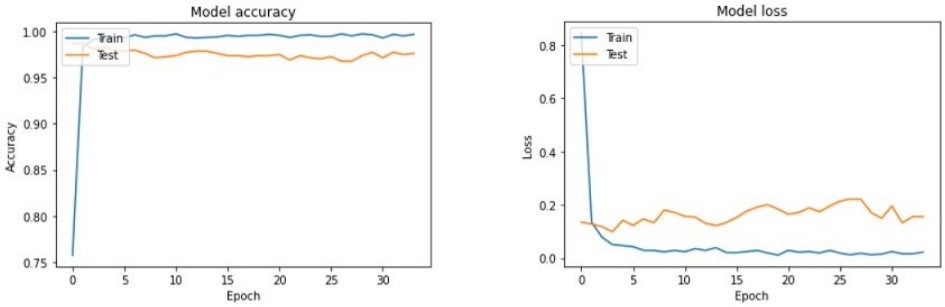


Σχήμα 5.7: Σύγκριση αλγορίθμων βελτιστοποίησης

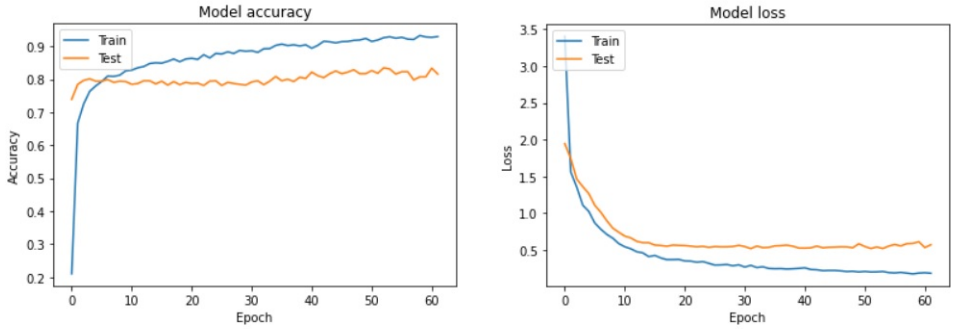
του μετρό συχνά ταξινομήθηκαν εσφαλμένα με ακρίβεια πρόβλεψης κάτω από 50%. Έτσι, υπήρχαν πολλά περιθώρια για περαιτέρω βελτίωση.

5.3 Αποτελέσματα μοντέλου συνολικής μάθησης

Με σκοπό την βελτίωση των παραπάνω αποτελεσμάτων, δημουργήσαμε ένα μοντέλο συνολικής μάθησης όπως ακριβώς περιγράφεται στο κεφάλαιο 4. Τα μοντέλα βάσης λειτούργησαν ανεξάρτητα το ένα από το άλλο και το καθένα παρουσιάζει τα δικά του αποτελέσματα στο αντίστοιχο υποσύνολο. Το πρώτο μοντέλο όπως και τα υπόλοιπα, εκπαιδεύτηκε συνολικά σε 1837 ηχητικά αρχεία και δοκιμάστηκε σε 837. Τα δεδομένα που χρησιμοποιήθηκαν απο το μοντέλο αυτό ανήκουν στις κλάσεις του αεροδρομίου και του λεωφορείου, όπως φαίνεται και στον πίνακα 4.1. Στο σχήμα 5.8 παρουσιάζονται τα γραφήματα ακρίβειας και απωλειών του μοντέλου. Παρατηρούμε ότι, ο μηχανισμός πρόωρης διακοπής της εκπαίδευσης ενεργοποιήθηκε στην εποχή 34, κάτι το οποίο σημαίνει πως οι ελάχιστες απώλειες του μοντέλου παρουσιάστηκαν στην εποχή 4 και συγκεκριμένα είχαν την τιμή 0.091. Η ακρίβεια των προβλέψεων του μοντέλου είναι της τάξης του 98%, ποσοστό ιδιαιτέρως υψηλό για τα δεδομένα

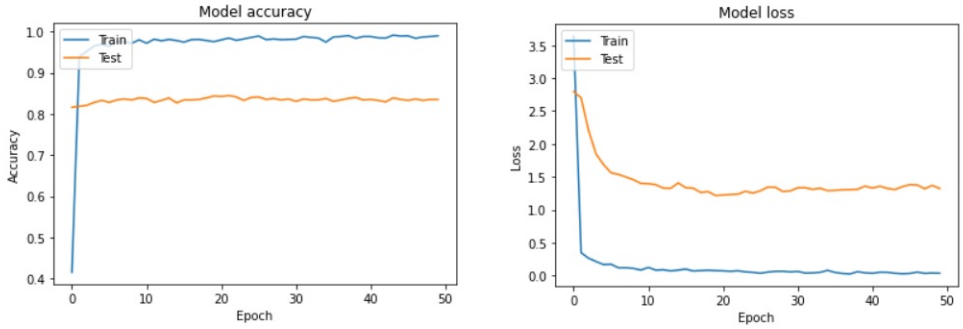


Σχήμα 5.8: Γραφήματα ακρίβειας ταξινόμησης και απωλειών στα σύνολα εκπαίδευσης και δοκιμής - υπομοντέλο 1

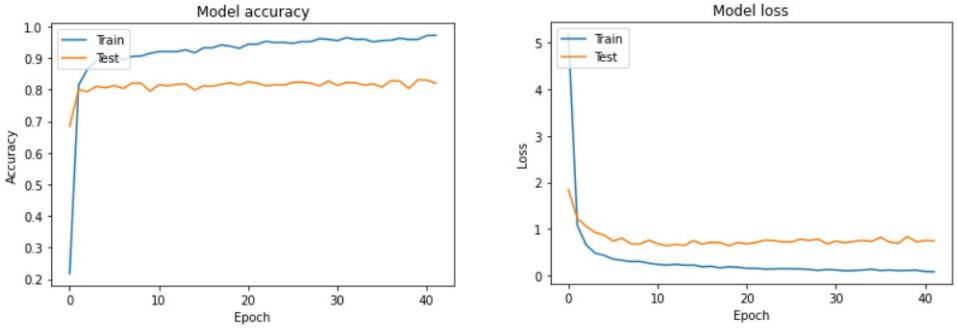


Σχήμα 5.9: Γραφήματα ακρίβειας ταξινόμησης και απωλειών στα σύνολα εκπαίδευσης και δοκιμής - υπομοντέλο 2

μιας ταξινόμησης ήχου. Το δεύτερο υπομοντέλο διενέργησε σε δεδομένα που ανήκουν στις κλάσεις του μετρό και του σταθμού του μετρό, με την δεύτερη να αποτελεί κλάση με ποσοστό επιτυχημένης πρόβλεψης 49% με την προηγούμενη αρχιτεκτονική. Η τεχνική ”bagging” της συνολικής μάθησης φάνηκε να λειτούργησε επιτυχημένα σε αυτήν την περίπτωση, αφού το υπομοντέλο προέβλεψε με επιτυχία περίπου 83% την έξοδο των αρχείων του υποσυνόλου δοκιμής, όπως παρατηρούμε στο γράφημα ακρίβειας του σχήματος 5.9. Τα υπομοντέλα 3 και 4, τα οποία εκπαιδεύτηκαν και δοκιμάστηκαν στις κλασεις πάρκο-δημόσια πλατεία και εμπορικό κέντρο-πεζόδρομος αντίστοιχα παρουσίασαν ανάλογη συμπεριφορά με την ακρίβεια ταξινόμησης στο σύνολο δοκιμής να φτάνει περίπου στο 83%, όπως παρατηρούμε και στα διαγράμματα 5.10 και 5.11. Η μόνη διαφορά που αξίζει να αναφερθεί είναι ότι το 3ο υπομοντέλο είναι το μοναδικό που παρουσιάζει απώλειες που ξεπερνούν την μονάδα. Τέ-



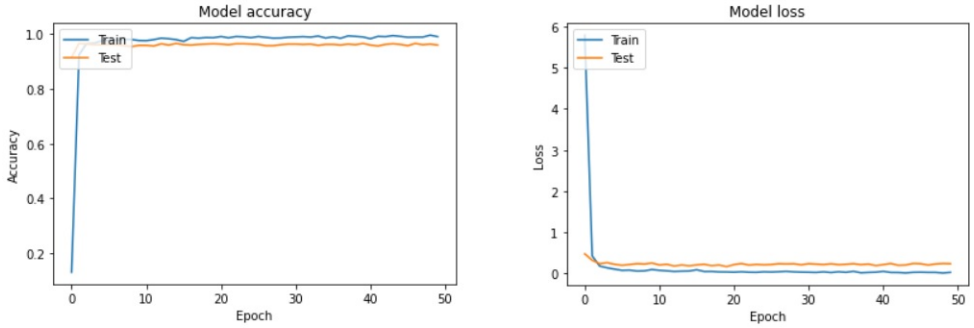
Σχήμα 5.10: Γραφήματα ακρίβειας ταξινόμησης και απωλειών στα σύνολα εκπαίδευσης και δοκιμής - υπομοντέλο 3



Σχήμα 5.11: Γραφήματα ακρίβειας ταξινόμησης και απωλειών στα σύνολα εκπαίδευσης και δοκιμής - υπομοντέλο 4

λος, το 5ο και τελευταίο μοντέλο βάσης το οποίο λειτουργεί με δεδομένα των κλάσεων του δρόμου οδικής κυκλοφορίας και του τραμ σημειώνει εξαιρετικά υψηλά ποσοστά ακρίβειας της τάξεως του 96% (σχήμα 5.12).

Η τελική μας πρόβλεψη προέρχεται απο την συγχώνευση των μοντέλων βάσης. Το τελικό μας σύστημα καλείται να πραγματοποιήσει τις προβλέψεις του στα δεδομένα του συνόλου δοκιμής, το οποίο περιέχει αρχεία που ανήκουν και στις 10 κλάσεις. Στο 5.13 απεικονίζεται αναλυτικά η αναφορά ταξινόμησης των ακουστικών σκηνών, με όλες τις μετρικές να παρουσιάζουν 89% επιτυχία ταξινόμησης στο σύνολο των δεδομένων δοκιμής. Αν παρατηρήσουμε την σύγκριση των confusion matrices στο σχήμα 5.14, με τον αριστερά πίνακα να αντιπροσωπεύει την ταξινόμηση που πραγματοποιήθηκε από το μοντέλο συνολικής μάθησης και δεξιά να παρουσιάζεται ο πίνακας του αρχικού μας μοντέλου, μπορούμε να διακρίνουμε την μεγάλη βελτίωση στο ποσοστό επιτυχημένης πρόβλεψης σε όλες τις ακουστικές σκηνές. Οι



Σχήμα 5.12: Γραφήματα ακρίβειας ταξινόμησης και απωλειών στα σύνολα εκπαίδευσης και δοκιμής - υπομοντέλο 5

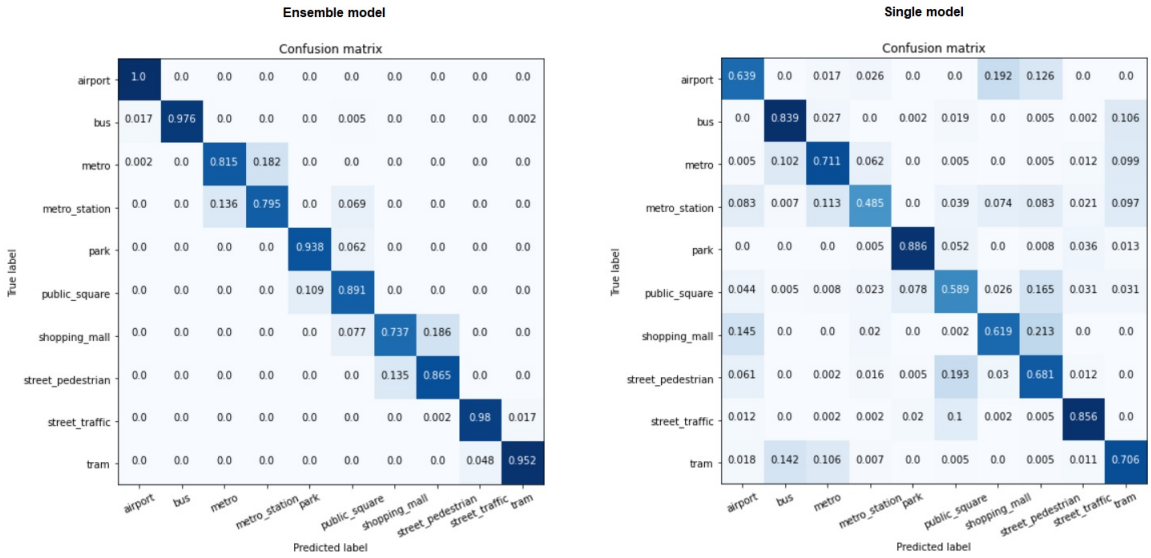
Classification Report

	precision	recall	f1-score	support
airport	0.98	1.00	0.99	421
bus	1.00	0.98	0.99	415
metro	0.86	0.82	0.84	433
metro_station	0.81	0.80	0.80	435
park	0.90	0.94	0.92	386
public_square	0.79	0.89	0.84	387
shopping_mall	0.85	0.74	0.79	441
street_pedestrian	0.82	0.86	0.84	429
street_traffic	0.95	0.98	0.96	402
tram	0.98	0.95	0.97	436
accuracy			0.89	4185
macro avg	0.89	0.89	0.89	4185
weighted avg	0.89	0.89	0.89	4185

Σχήμα 5.13: Αναφορά ταξινόμησης μοντέλου συνολικής μάθησης

κλάσεις που ξεχωρίζουν για την βελτίωση του ποσοστού επιτυχημένης πρόβλεψης είναι:

1. Το αεροδρόμιο, με βελτίωση 36.1%.
2. Ο σταθμός μετρό, μια κλάση που είχε ποσοστό επιτυχούς πρόβλεψης κάτω απο 50% με το νέο μοντέλο σημείωσε 79.5%.
3. Η δημόσια πλατεία που παρουσίασε βελτίωση της τάξης του 30%.
4. Το τράμ με ποσοστό επιτυχούς πρόβλεψης 95.2% από 70.6%.



Σχήμα 5.14: Σύγκριση των confusion matrices του μοντέλου συνολικής μάθησης και της αρχικής αρχιτεκτονικής

Κεφάλαιο 6

Σύνοψη και μελλοντική δουλειά

Στα πλαίσια του πλήθους των πειραμάτων της μελέτης μας καταλήξαμε ότι η εξαγωγή ποικίλων ηχητικών χαρακτηριστικών σε συνδυασμό με ένα πυκνό νευρωνικό δίκτυο 5 στρωμάτων ήταν η ιδανική αρχιτεκτονική για την ταξινόμηση των ακουστικών σκηνών σε περιβάλλοντα πόλης στα πλαίσια της διατριβής μας. Τα αρχικά πειράματα, οδήγησαν στο συμπέρασμα ότι οι ανισορροπίες μεταξύ των συνόλων εκπαίδευσης και δοκιμής αποτελούν ανασταλτικό παράγοντα στην επίτευξη υψηλών ποσοστών επιτυχίας στην ταξινόμηση των ακουστικών σκηνών από το δίκτυό μας. Πιο συγκεκριμένα, δημιουργήθηκε το πρόβλημα του overfitting οπού το δίκτυο μας εκπαιδεύτηκε υπερβολικά καλά στα δεδομένα εκπαίδευσης με αποτέλεσμα να έχει χαμηλή απόδοση στα δεδομένα του συνόλου δοκιμής, τα οποία διέφεραν σε ορισμένα σημεία όπως για παράδειγμα στις τοποθεσίες εγγραφής των ηχητικών αρχείων. Για την αντιμετώπιση του προβλήματος αυτού προσθέσαμε στο σύστημα μας διάφορους μηχανισμούς, με τους πιο σημαντικούς να είναι τα dropout στρώματα, η κανονικοποίηση παρτίδας και η πρόωρη διακοπή της εκπαιδευτικής διαδικασίας. Έπειτα, δοκιμάσαμε το σύστημα μας στο αντίστοιχο σύνολο όπου απέδωσε με 70% ακρίβεια ταξινόμησης στην πλειοψηφία των μετρικών, ποσοστό που είχε αρκετά περιθώρια βελτίωσης. Η βελτίωση αυτή επιτεύχθηκε με την δημιουργία ενός μοντέλου συνολικής μάθησης (ensemble learning), το οποίο αποτελούνταν από 5 μοντέλα βάσης ίδιας αρχιτεκτονικής με αυτήν που περιγράψαμε. Τα μοντέλα αυτά εκπαιδεύτηκαν και πραγματοποίησαν τις προβλέψεις τους σε υποσύνολα του συνόλου δεδομένων, με την τελική πρόβλεψη να προέρχεται από τον συνδυασμό των προβλέψεων των μοντέλων βάσης. Το τελικό σύστημα απέδωσε με 89% στο σύνολο δοκιμής, ποσοστό ιδιαιτέρως ικανοποιητικό για την ταξινόμηση των ακουστικών σκηνών.

Σε μελλοντικά πειράματα αποσκοπούμε στην επίτευξη υψηλών ποσοστών ακρίβειας τα-

ξινόμησης ακουστικών σκηνών με την χρήση συνελικτικών νευρωνικών δικτύων. Αυτό μπορεί να επιτευχθεί με την εξαγωγή φασματογραμμάτων, τα οποία απεικονίζουν το εύρος των συχνοτήτων του κάθε ηχητικού αρχείου και η ταξινόμηση θα πραγματοποιηθεί με βάση αυτά. Μετά την απαραίτητη εξοικείωση με ζητήματα ταξινόμησης ήχου, υπάρχουν πλάνα δημιουργίας διαδικτυακής εφαρμογής αναγνώρισης ήχου, η οποία διαμορφώνει αυτόματα το ηχητικό προφίλ της ”smart” συσκευής του χρήστη (smartphones, tablets κτλ.) ανάλογα με το περιβάλλον στο οποίο βρίσκεται. Για παράδειγμα, σε έναν πολυσύχναστο δρόμο το ηχητικό προφίλ της συσκευής του χρήστη πρέπει να βρίσκεται στην επιλογή ”δυνατό”, έτσι ώστε να γίνονται αντιληπτές ειδοποιήσεις όπως κλήσεις ή μηνύματα. Αντιθέτως, σε ένα ήσυχο περιβάλλον η αυτόματη επιλογή του ανάλογου προφίλ συμβάλει στην καταστολή των ενοχλητικών θορύβων των ειδοποιήσεων πρός όφελος του ίδιου του χρήστη και του περίγυρού του αλλά και ικανοποιεί σκοπούς εξοικονόμησης μπαταρίας στην συσκευή.

Βιβλιογραφία

[1] Acoustic scene classification - DCASE. http://dcase.community/challenge2019/task-acoustic-scene-classification.

[2] Neural network models. https://otexts.com/fpp2/nnetar.html#.

[3] Topic dl01: Activation functions and its types in artifical neural network. https://abhigoku10.medium.com/activation-functions-and-its-types-in-artifical-neural-network-14511f3080a8.

[4] Stacey Ronaghan. Deep learning: Overview of neurons and activation functions | by stacey ronaghan | medium. https://srnghn.medium.com/deep-learning-overview-of-neurons-and-activation-functions-1d98286cf1e4.

[5] Timo Bohm. A first introduction to SELUs and why you should start using them as your activation functions | by timo böhm | towards data science. https://towardsdatascience.com/gentle-introduction-to-selus-b19943068cd9.

[6] Temporal convolutional neural network for the classification of satellite image time series - scientific figure on researchgate. https://www.researchgate.net/figure/Example-of-fully-connected-neural-network_fig2_331525817.

[7] Dorian Lazar. Building a resnet in Keras. https://towardsdatascience.com/building-a-resnet-in-keras-e8f1322a49ba, 3 2020.

[8] MFCC Based Speech Retrieval. https://1library.net/document/ozlo1n2z-mfcc-based-speech-retrieval.html.

[9] Spectral flatness - Wikipedia. https://en.wikipedia.org/wiki/Spectral_flatness#cite_note-3.

[10] https://librosa.org/doc/main/generated/librosa.feature.chroma_cqt.html.

[11] Aishwarya Singh. A comprehensive guide to Ensemble Learning (with Python codes). https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/, June 2018.

[12] Koutini Hamid Khaled, Eghbal-zadeh, and Widmer Gerhard. CP-JKU submissions to DCASE'19: Acoustic scene classification and audio tagging with receptive-field-regularized CNNs. http://dcase.community/documents/challenge2019/technical_reports/DCASE2019_Eghbal-zadeh_99.pdf, 2019.

[13] Chen Hangting, Liu Zuozhen, Liu Zongming, Zhang Pengyuan, and Yan Yonghong. Integrating the data augmentation scheme with various classifiers for acoustic scene modeling. http://dcase.community/documents/challenge2019/technical_reports/DCASE2019_Zhang_34.pdf, 2019.

[14] Valentin Bilot and Quang Khanh Ngoc Duong. Acoustic scene classification with multiple instance learning and fusion. Technical report, DCASE2019 Challenge, June 2019.

[15] Jonathan Huang, Paulo Lopez Meyer, Hong Lu, Hector Cordourier Maruri, and Juan Del Hoyo. Acoustic scene classification using deep learning-based ensemble averaging. Technical report, DCASE2019 Challenge, June 2019.

[16] Seo Hyeji and Park Jihwan. Acoustic scene classification using various pre-processed features and convolutional neural networks. Technical report, DCASE2019 Challenge, June 2019.

[17] R. Lee, M. Kang, B. Kim, K. Park, S. Q. Lee, and H. Park. Sound source localization based on gcc-phat with diffuseness mask in noisy and reverberant environments. *IEEE Access*, 8:7373–7382, 2020.

[18] Marcin Plata. Deep neural networks with supported clusters preclassification procedure for acoustic scene recognition. Technical report, DCASE2019 Challenge, June 2019.

[19] K. Drossos, P. Magron, S. I. Mimilakis, and T. Virtanen. Harmonic-percussive source separation with deep neural networks and phase recovery. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 421–425, 2018.

[20] Chandrasekhar Paseddula and Suryakanth V.Gangashetty. DCASE 2019 task 1a: Acoustic scene classification by sffcc and DNN. Technical report, DCASE2019 Challenge, June 2019.

[21] Joshua Payne. Activation functions in artificial neural networks | by joshua payne | the startup | medium. `https://medium.com/swlh/activation-functions-in-artificial-neural-networks-8aa6a5ddf832`, 1 2020.

[22] Mohammed Rampurawala. Classification with tensorflow and dense neural networks | by mohammed rampurawala | heartbeat. `https://heartbeat.fritz.ai/classification-with-tensorflow-and-dense-neural-networks-8299327a818a`, 2 2019.

[23] S. Albawi, T. A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, 2017.

[24] Oluwatobi Ayodeji Akanbi, Iraj Sadegh Amiri, and Elahe Fazeldehkordi. *Feature Extraction*. Elsevier, 1 2015.

[25] Librosa — Librosa 0.8.0 documentation. `https://librosa.org/doc/latest/index.html`.

[26] Hann function - Wikipedia. `https://en.wikipedia.org/wiki/Hann_function`.

[27] Maël Fabien. Sound feature extraction. `https://maelfabien.github.io/machinelearning/Speech9/#`, 12 2019.

[28] Discrete cosine transform - Wikipedia. `https://en.wikipedia.org/wiki/Discrete_cosine_transform`.

[29] Theodoros Giannakopoulos and Aggelos Pikrakis. Chapter 4 - audio features. In Theodoros Giannakopoulos and Aggelos Pikrakis, editors, *Introduction to Audio Analysis*, pages 59–103. Academic Press, Oxford, 2014.

[30] Nilesh Kulkarni and Vinayak Bairagi. *Role of Different Features in Diagnosis of Alzheimer Disease*. Elsevier, 1 2018.

[31] Bandwidth (signal processing) - Wikipedia. https://en.wikipedia.org/wiki/Bandwidth_(signal_processing).

[32] Jun Yang, Fa Long Luo, and Arye Nehorai. Spectral contrast enhancement: Algorithms and comparisons. *Speech Communication*, 39:33–46, 1 2003.

[33] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature 1. Technical report, 2002.

[34] Saeed Aaqib. Urban sound classification, part 1 - aaqib saeed. http://aqibsaeed.github.io/2016-09-03-urban-sound-classification-part-1/.

[35] Christopher Harte, Mark Sandler, and Martin Gasser. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, AMCMM '06, page 21–26, New York, NY, USA, 2006. Association for Computing Machinery.

[36] Juan Pablo Bello. Chroma and tonality el9173 selected topics in signal processing: Audio content analysis nyu poly. Technical report, 2018.

[37] Chroma feature - Wikipedia. https://en.wikipedia.org/wiki/Chroma_feature.

[38] Constant-Q transform - Wikipedia. https://en.wikipedia.org/wiki/Constant-Q_transform.

[39] Chroma. https://musicinformationretrieval.com/chroma.html.

[40] Dimensionality reduction - Wikipedia. https://en.wikipedia.org/wiki/Dimensionality_reduction#Linear_discriminant_analysis_(LDA).

[41] Linear discriminant analysis - Wikipedia. https://en.wikipedia.org/wiki/Linear_discriminant_analysis.

[42] The sequential model. https://keras.io/guides/sequential_model/.

[43] tf.keras.initializers.Lecunnormal | tensorflow core v2.4.1. https://www.tensorflow.org/api_docs/python/tf/keras/initializers/LecunNormal.

[44] How to choose loss functions when training deep learning neural networks. https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/.

[45] Cross-entropy loss function. A loss function used in most… | by kiprono elijah koech | towards data science. https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e.

[46] Stochastic Gradient Descent — clearly explained !! | by aishwarya v srinivasan | towards data science. https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31.

[47] Gentle introduction to the Adam optimization algorithm for deep learning. https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/.

[48] Adam — latest trends in deep learning optimization. | by vitaly bushaev | towards data science. https://towardsdatascience.com/adam-latest-trends-in-deep-learning-optimization-6be9a291375c.

[49] Overfitting and underfitting with machine learning algorithms. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/.

[50] A gentle introduction to dropout for regularizing deep neural networks. https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/.

[51] A gentle introduction to batch normalization for deep neural networks. https://machinelearningmastery.com/batch-normalization-for-training-of-deep-neural-networks/.

[52] Understanding the classification report through sklearn – muthukrishnan. https://muthu.co/understanding-the-classification-report-in-sklearn/.

[53] Understanding confusion matrix | by sarang narkhede | towards data science. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62.