# Αναγνώριση Είδους Μουσικής με Συνελικτικά Νευρωνικά Δίκτυα

Χαρίκλεια Δ. Ράπτη, Μελίνα Γ. Αποστολίδου

***Abstract*— Σε αυτή την εργασία παρουσιάζεται μία προσέγγιση για την ταξινόμηση του είδους μουσικής με τη χρήση συνελικτικών νευρωνικών δικτύων. Έγινε αξιολόγηση διαφορετικών αρχιτεκτωνικών χρησιμοποιώντας το σύνολο δεδομένων GTZAN και τα φασματογραφήματα του κάθε τραγουδιού ως είσοδο στο συνελικτικό νευρωνικό δίκτυο. Τέλος, παρουσιάζονται τα καλύτερα αποτελέσματα και η αρχιτεκτονική μέσω της οποίας επιτεύχθηκαν αυτά. Μία μελλοντική εργασία θα μπορούσε να περιλαμβάνει την οπτικοποίηση κάθε στρώματος του νευρωνικού δικτύου καθώς και πειραματισμό με περισσότερες αρχιτεκτονικές με σκοπό την βελτίωση της ακρίβειας.**

***Keywords:*** Ταξινόμηση είδους μουδικής, συνελικτικά νευρωνικά δίκτυα, Φασματογράφημα , CNN, Deep Learning.

## I. Εισαγωγη

Τα τελευταία χρόνια με τη ραγδαία ανάπτυξη της ψηφιακής τεχνολογίας ένας μεγάλος όγκος μουσικής είναι όλο και περισσότερο διαθέσιμος. Συνεπώς, η δόμηση και διαχείρηση της μουσικής αποτελεί πλεόν ένα θεμελιώδες πρόβλημα. Ένας τρόπος για να δομηθεί ο όλο και αυξανόμενος όγκος μουσικής είναι η ταξινόμηση του είδους μουσικής. Επομένως, για την δόμηση και οργάνωση μεγάλων αρχείων μουσικής απαιτείται μία αποτελεσματική και ακριβής ταξινόμηση του μουσικού είδους.

Τα συνελικτικά νευρωνικά δίκτυα είναι γνωστό ότι έχουν εξαιρετικά αποτελέσματα στον τομέα της «όρασης» του υπολογιστή. Αυτά τα νευρωνικά δίκτυα αποτελούνται από ένα συνελικτικό στρώμα που ακολουθείται από ένα συγκεντρωτικό στρώμα (pooling). Τα δίκτυα αυτά μαθαίνουν να αναγνωρίζουν τα διαφορετικά χαρακτηριστικά της εισόδου και καθώς στοιβάζονται το ένα μετά το άλλο μαθαίνουν να αναγνωρίζουν πιο περίπλοκα χαρακτηριστικά. Με την πάροδο των τελευταίων χρόνων έχουν εισαχθεί ορισμένες βελτιστοποιήσεις όπως το Dropout ώστε να αποφευχθεί η υπερβολική προσαρμογή (overfitting).

Τα περισσότερα προβλήματα ταξινόμησης είδους μουσικής αποτελούνται κυρίως από δύο τμήματα. Το ένα είναι η προεπεξεργασία ακατέργαστων δεδομένων ήχου και το δεύτερο είναι ο σχεδιασμός του μοντέλου ταξινόμησης. Ως ένα βασικό τμήμα του συστήματος τα προεπεξεργασμένα δεδομένα είναι το κλειδί για την τελική ακρίβεια ταξινόμησης. Γενικά, υπάρχουν τρεις βασικοί τρόποι προεπεξεργασίας ακατέργαστου ήχου. Ο πρώτος είναι η εξαγωγή των ακουστικών χαρακτηριστικών, ο δεύτερος είναι ο μετασχηματισμός σε φασματογραφήματα και ο τρίτος η χρήση του ακατέργαστου ήχου. Πριν την άνθηση του Deep Learning ένας συνήθης τρόπος ήταν η εξαγωγή συγκεκριμένων ακουστικών χαρακτηριστικών και η συγκέντρωσή τους ως είσοδος χρησιμοποιώντας διάφορους αλγορίθμους μηχανικής μάθησης. Ωστόσο αυτή η μέθοδος απαιτεί σημαντική μηχανική προσπάθεια και επαγγελματικές γνώσεις. Με την ταχεία ανάπτυξη του Deep Learning το συνελικτικό νευρωνικό δίκτυο (CNN) έχει λάβει μεγάλη επιτυχία και αναγνώριση και έχει δοκιμαστεί στο πεδίο της αναγνώρισης του είδους μουσικής. Από την άλλη μεριά ο μουσικός ήχος με ετικέτα είναι πραγματικά ανεπαρκής σε αυτόν τον τομέα λόγω του υψηλού κόστους επαγγελματικής σήμανσης ειδικών.

Στη συγκεκριμένη εργασία χρησιμοποιήθηκαν κυρίως συνελικτικά στρώματα (convolutional layers) και average pooling layers. Χρησιμοποιήθηκαν ιδέες νευρωνικών δικτύων για την όραση του υπολογιστή με εισόδο φασματογραφήματα αντί για εικόνες. Για την εργασία αυτή χρησιμοποιήθηκε ένα δημοφιλές σύνολο δεδομένων. Το σύνολο δεδομένων GTZAN περιέχει τραγούδια από δέκα διαφορετικά μουσικά είδη. Αυτό το σύνολο δεδομένων είναι μικρό αλλά έχει χρησιμοποιηθεί σε αρκετές εργασίες για αναγνώριση είδους μουσικής οπότε και επιλέχθηκε και για την παρούσα εργασία.

.

## II. Συνολο Δεδομενων και Χαρακτηριστικα

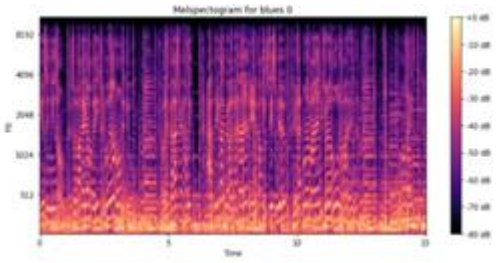
Δεδομένου ότι η μουσική προστατεύεται από πνευματικά δικαιώματα, η εύρεση ενός καλού συνόλου δεδομένων είναι αρκετά περίπλοκη. Τα τρία πιο δημοφιλή σύνολα δεδομένων για την ταξινόμηση είδους μουσικής είναι το GTZAN, το MagnaTagATune και το Million Song Dataset.

Στην παρούσα εργασία χρησιμοποιήθηκε το σύνολο δεδομένων GTZAN, το οποίο συλλέχθηκε από τους Tzanetakis και Cook. Αποτελείται από 1000 ηχητικά κομμάτια 30

δευτερολέπτων και περιέχει 10 είδη μουσικής (Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, Rock) με 100 κομματια ανά είδος. Όλα τα κομμάτια είναι 22,050Hz, Mono 16-bit αρχεία ήχου σε .au μορφή. Διαπιστώθηκε ότι το μικρό μέγεθος του συνόλου δεδομένων καθιστά δύσκολη τη σύγκλιση όταν χρησιμοποιούνται βαθύτερα μοντέλα. Για αυτό το σύνολο δεδομένων δημιουργήσαμε τα φασματογραφήματα για όλα τα τραγούδια και έπειτα σειριοποιήσαμε όλα τα φασματογραφήματα ως ένα numpy array. Ακολούθως, αυτά τα δεδομένα φορτώθηκαν στη μνήμη για εκπαίδευση. Λόγω του μικρού μεγέθους του συνόλου δεδομένων ήταν απλό να φορτωθούν τα δεδομένα σε numpy arrays στη μνήμη και να τροφοδοτηθεί το μοντέλο Tensorflow Keras.

Αυτή η μορφή Tensorflow αναπαριστά μία ακολουθία δυαδικών συμβολοσειρών. Η μορφή είναι χρήσιμη για ροή μεγάλου όγκου δεδομένων διαδοχικά. Ως εκ τούτου κάθε τραγούδι είχε το φασματογράφημα του και την ετικέτα του. Αρχικά, όλα τα τραγούδια προεπεξεργάζονται ως φασματογραφήματα. Για τον υπολογισμό των φασματογραφημάτων έγινε εκτεταμένη χρήση της βιβλιοθήκης librosa για επεξεργασία ήχου.

.



Εικόνα 1. Φασματογράφημα τραγουδιού της κλάσης blues

## III. Σχετικες εργασιες

Η βαθιά μάθηση (Deep Learning) και ειδικά τα συνελικτικά νευρωνικά δίκτυα (CNNs) εφαρμόστηκαν πρόσφατα με επιτυχία στην «όραση» του υπολογιστή. Έχει υπάρξει αρκετό ενδιαφέρον για τη διερεύνηση της εκμάθησης χαρακτηριστικών χρησιμοποιώντας βαθιά νευρωνικά δίκτυα στο πρόβλημα της ταξινόμησης του είδους μουσικής. Υπάρχουν αρκετές εργασίες για τη χρήση νευρωνικών δικτύων (όχι μόνο συνελικτικά αλλά και απλά πλήρως συνδεδεμένα και επαναλαμβανόμενα δίκτυα) για την ταξινόμηση του μουσικού είδους μεταξύ άλλων προσεγγίσεων χωρίς Deep Learning. Υπάρχουν διαφορετικές μέθοδοι που χρησιμοποιούν κυρίως πλήρως συνελικτικά νευρωνικά δίκτυα καθώς και επαναλαμβανόμενα νευρωνικά δίκτυα. Το μεγαλύτερο μέρος των εργασιών που χρησιμοποιούν το σύνολο δεδομένων GTZAN χρησιμοποιεί φασματογραφήματα για την προεπεξεργασία των τραγουδιών. Αυτό δεν επικεντρώνεται στην αναγνώριση του είδους, αλλά στην προσομοίωση τραγουδιών για συστάσεις μουσικής. Ωστόσο, ήταν σημαντικό να αναφερθεί για τη χρήση συνελικτικών νευρωνικών δικτύων με ενεργοποίηση ReLU σε κλιπ τραγουδιών που έχουν υποστεί προεπεξεργασία ως φασματογραφήματα. Στην συγκεκριμένη εργασία, για την προεπεξεργασία των τραγουδιών προτείνεται επίσης η χρήση φασματογραφημάτων.

Υπάρχουν άλλες παρόμοιες προσεγγίσεις που χρησιμοποιούν πλήρως συνελικτικά νευρωνικά δίκτυα για την επίλυση αυτού του προβλήματος. Αυτές οι αρχιτεκτονικές από ένα συνελικτικό στρώμα ακολουθούμενο από ένα max pooling στρώμα n φορές και τέλος ένα πλήρως συνδεδεμένο στρώμα. Όλες αυτές οι εργασίες έχουν μικρές διαφορές ως προς τον αριθμό των επιπέδων, τις υπερπαραμέτρους κλπ. Η βασική ιδέα των εργασιών αυτών είναι η χρήση πλήρως συνδεδεμένων νευρωνικών δικτύων.

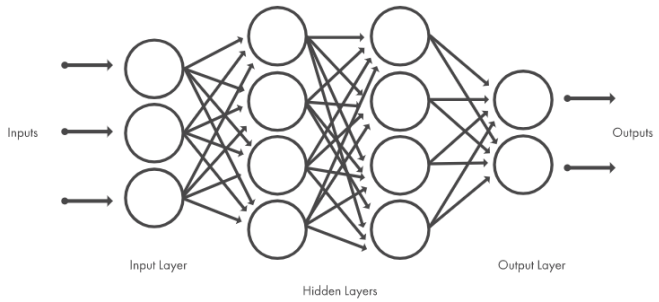
Ένα συνελικτικό deep belief δίκτυο (CDBN) προτάθηκε για τη βελτίωση της ταξινόμησης του είδους μουσικής με τη χρήση φασματογραφημάτων ήχου και MFCC (Mel-frequency cepstral coefficients) χαρακτηριστικά, από τους H. Lee, P. Pham, Y. Largman, και A. Y. Ng. Τα χαρακτηριστικά που έχουν εξαχθεί από ηχητικά δεδομένα χωρίς ετικέτα φαίνονται να έχουν πολύ καλή απόδοση σε αρκετά προβλήματα ταξινόμησης μουσικής. Οι T. L. Li, A. B. Chan, and A. Chun χρησιμοποίησαν CNN για να εξάγουν χαρακτηριστικά μουσικών μοτίβων ηχητικών κλιπ. Η εργασία τους απέδειξε ότι τα συνελικτικά νευρωνικά δίκτυα (CNN) είχαν την δυνατότητα να εξάγουν σημαντικά χαρακτηριστικά από ένα εύρος μουσικών προτύπων με ελάχιστες προαπαιτούμενες γνώσεις. Ωστόσο, τα πειραματικά τους αποτελέσματα έδειξαν ότι τα προτεινόμενα μοντέλα δεν γενικεύτηκαν πολύ καλά για μη εκπαιδευμένα δεδομένα. Οι P. Zhang, X. Zheng, W. Zhang, S. Li, S. Qian, W. He, S. Zhang, and Z. Wang ασχολήθηκαν με CNN με k-max pooling επίπεδα για σημασιολογική μοντελοποίηση της μουσικής. Η προτεινόμενη μέθοδος θα μπορούσε να παράγει περισσότερο ισχυρές μουσικές αναπαραστάσεις προσθέτοντας περισσότερα επίπεδα. Οι C. Zhang, G. Evangelopoulos, S. Voinea, L. Rosasco, and T. Poggio δημιούργησαν μία ιεραρχική αρχιτεκτονική για την εξαγωγή αμετάβλητων και διακριτών ηχητικών αναπαραστάσεων. Οι S. Dieleman and B. Schrauwen διερεύνησαν την απόδοση μοντέλων που εξήγαγαν χαρακτηριστικά από ακατέργαστα ηχητικά σήματα χρησιμοποιώντας CNN. Διαπίστωσαν ότι τα δίκτυα μπορούσαν να ανακαλύπτουν αυτόματα τις αποσυνθέσεις της συχνότητας. Ωστόσο, η μέθοδος αυτή δεν ξεπέρασε την απόδοση προσεγγίσεων που έκαναν χρήση φασματογραφημάτων.

## IV. Επεξεργασια δεδομενων

Το φασματογράφημα αναπαριστά τις ακολουθίες φασμάτων που διαφοροποιούνται κατά μήκος του άξονα του χρόνου. Η μετατροπή των τραγουδιών σε φασματογραφήματα αποτελεί το κλειδί για την επιτυχή εφαρμογή του συνελικτικού νευρωνικού δικτύου (CNN) στην ταξινόμηση των ειδών μουσικής. Με αυτόν τον τρόπο η επισήμανση ετικετών σε ηχητικά κομμάτια μουσικής αναδιαμορφώνεται ως εργασία ταξινόμησης εικόνας. Επίσης, το μέγεθος του φασματογραφήματος είναι μία υπερπαράμετρος. Τα φασματογραφήματα που εισάγονται στο CNN έχουν διάσταση 150×150.

## V. Μεθοδολογια

Όπως και άλλα νευρωνικά δίκτυα, ένα συνελικτικό νευρωνικό δίκτυο (CNN) αποτελείται από ένα επίπεδο εισόδου, ένα επίπεδο εξόδου και πολλά κρυφά επίπεδα ενδιάμεσα.




Εικόνα 2. Επίπεδα Συνελικτικού Νευρωνικού Δικτύου

Αυτά τα επίπεδα εκτελούν λειτουργίες που αλλάζουν τα δεδομένα με σκοπό την εκμάθηση συγκεκριμένων χαρακτηριστικών των δεδομένων. Τα επίπεδα Συνελικτικών Νευρωνικών Δικτύων από τα οποία αποτελείται και το νευρωνικό δίκτυο της συγκεκριμένης εργασίας παρουσιάζονται παρακάτω:

- **Convolution:** Αυτό το επίπεδο περνάει την εικόνα μέσα από ένα σύνολο συνελικτικών φίλτρων, καθένα από τα οποία ενεργοποιεί ορισμένα χαρακτηριστικά από τις εικόνες. Είναι το πρώτο επίπεδο που εξάγει χαρακτηριστικά από μία εικόνα εισόδου. Το επίπεδο αυτό διατηρεί τη σχέση μεταξύ των pixels με την εκμάθηση των χαρακτηριστικών της εικόνας χρησιμοποιώντας μικρά τετράγωνα των δεδομένων εισόδου. Είναι μία μαθηματική λειτουργία που δέχεται δύο εισόδους, έναν πινακα εικόνας και ένα φίλτρο ή kernel.

- **Pooling:** Αυτό το επίπεδο απλοποιεί την έξοδο εκτελώντας μη γραμμική δειγματοληψία, μειώνοντας τον αριθμό των παραμέτρων που πρέπει να μάθει το δίκτυο, όταν η εικόνα είναι πολύ μεγάλη. Υπάρχουν διαφορετικοί τύποι αυτού του επιπέδου. Συγκεκριμένα, το επίπεδο max pooling παίρνει το μεγαλύτερο στοιχείο από τον διορθωμένο χάρτη χαρακτηριστικών, ενώ το επίπεδο average pooling παίρνει το μέσο στοιχείο.

- **Dense:** Το επίπεδο Dense είναι ένα πλήρως συνδεδεμένο επίπεδο, που σημαίνει ότι όλοι οι νευρώνες σε ένα επίπεδο είναι συνδεδεμένοι με αυτούς στο επόμενο επίπεδο. Ένα πυκνά συδεδεμένο επίπεδο παρέχει χαρακτηριστικά εκμάθησης από όλους τους συνδυασμούς των χαρακτηριστικών του προηγούμενου επιπέδου.

- **Rectified Linear Unit (ReLU):** Το επίπεδο αυτό επιτρέπει την ταχύτερη και αποτελεσματικότερη εκπαίδευση, χαρτογραφώντας τις αρνητικές τιμές στο 0 και διατηρώντας τις θετικές τιμές. Αυτό μερικές φορές αναφέρεται ως ενεργοποίηση, επειδή μόνο τα ενεργοποιημένα χαρακτηριστικά μεταφέρονται στο επόμενο επίπεδο. Συγκεκρικένα, είναι μία μη γραμμική λειτουργία και η έξοδος είναι $f(x)=\max(0,x)$.

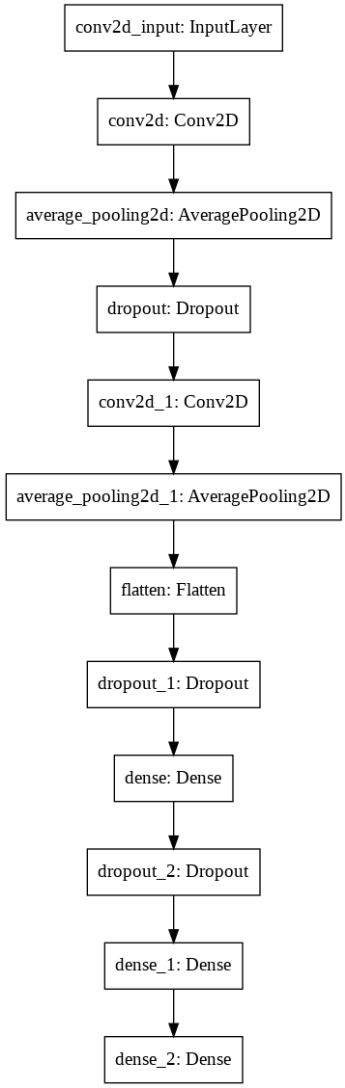
- **Dropout:** Το Dropout είναι μία μέθοδος κανονικοποίησης που προσεγγίζει την εκπαίδευση μεγάλου αριθμού νευρωνικών δικτύων με διαφορετικές αρχιτεκτονικές παράλληλα. Κατά τη διάρκεια της εκπαίδευσης κάποιος αριθμός εξόδων επιπέδου αγνοείται τυχαία ή απορρίπτεται. Αυτό έχει σαν αποτέλεσμα το επίπεδο να αντιμετωπίζεται σαν ένα επίπεδο με διαφορετικό αριθμό κόμβων και συνεκτικότητα με το προηγούμενο επίπεδο. Πιο συγκεκριμένα, η απομάκρυνση μίας μονάδας (κόμβων) σημαίνει η προσωρινή της αφαίρεση από το δίκτυο, μαζί με όλες τις εισερχόμενες και εξερχόμενες συνδέσεις της.

- **Flatten:** Μεταξύ του συνελικτικού επιπέδου και του πλήρως συνδεδεμένου επιπέδου υπάρχει ένα επίπεδο Flatten. Το επίπεδο αυτό μετατρέπει έναν δισδιάστατο πίνακα χαρακτηριστικών σε ένα διάνυσμα το οποίο μπορεί να τροφοδοτηθεί σε έναν πλήρως συνδεδεμένο ταξινομητή νευρωνικών δικτύων.

Για το σύνολο δεδομένων GTZAN χρησιμοποιήθηκε η παρακάτω αρχιτεκτονική. Αρχικά, το συνελικτικό νευρωνικό δίκτυο αποτελείται από δύο μπλοκ συνελεύσεων. Κάθε συνελικτικό μπλοκ περιέχει ένα επίπεδο Conv2D ακολουθούμενο από ένα average pool επίπεδο. Το πρώτο συνελικτικό μπλοκ έχει 32 φίλτρα όπως επίσης και το δεύτερο. Τα δύο συνελικτικά φίλτρα είναι 3×3. Τα δύο average pool επίπεδα έχουν μέγεθος (2,2). Επιπλέον, το πρώτο συνελικτικό μπλοκ περιέχει ένα επίπεδο Dropout με πιθανότητα 20%.

Μετά τα δύο συνελικτικά μπλοκ υπάρχει ένα flatten επίπεδο ακολουθούμενο από ένα πλήρως συνδεδεμένο επίπεδο με 64 νευρώνες με ενεργοποίηση relu και ένα επίπεδο Dropout με πιθανότητα 40%. Έπειτα, υπάρχει ένα ακόμα πλήρως συνδεδεμένο επίπεδο με 32 νευρώνες με ενεργοποίηση relu και ένα επίπεδο Dropout με πιθανότητα 30%. Το συνελικτικό νευρωνικό δίκτυο (CNN) εξάγει πιθανότητες κλάσης με βάση 10 κλάσεις.

Στο παρακάτω σχήμα παρουσιάζεται η αρχιτεκτονική του Συνελικτικού Νευρωνικού Δικτύου που δημιουργήθηκε. Συγκεκριμένα, παρουσιάζονται τα επίπεδα του νευρωνικού δικτύου με τη σειρά που είναι συνδεδεμένα.




Εικόνα 3. Μοντέλο για το σύνολο δεδομένων GTZAN

Κατά τη διάρκεια εκπόνησης αυτής της εργασίας δοκιμάστηκαν αρκετές προσεγγίσεις χρησιμοποιώντας παραλλαγές των υπερπαραμέτρων, περισσότερα επίπεδα, batch normalization, dropout κλπ. Αλλά με αυτές τις παραλλαγές η ακρίβεια του μοντέλου δεν κατάφερε να ξεπεράσει το 47%.

## VI. Αποτελεσματα

Το σύνολο δεδομένων GTZAN χωρίζεται σε 20% για επικύρωση (validation) και το υπόλοιπο 80% για εκπαίδευση (train). Η εκπαίδευση γίνεται χρησιμοποιώντας τον adam optimizer (βελτιστοποιητής). Ο βελτιστοποιητής adam είναι μία προσαρμοστική μέθοδος learning rate, που σημαίνει ότι υπολογίζει τα μεμονωμένα ποσοστά μάθησης για διαφορετικές παραμέτρους. Το όνομά του προέρχεται από την προσαρμοστική εκτιμηση της στιγμής (adaptive moment estimation) και ο λόγος είναι επειδή χρησιμοποιεί εκτιμήσεις της πρώτης και της δεύτερης στιγμής της κλίσης για να προσαρμόσει τον ρυθμό εκμάθησης (learning rate) για κάθε βάρος του νευρωνικού δικτύου.

Μετά από αυτήν την εκπαίδευση το μοντέλο έφτασε σε ακρίβεια 55% στο σύνολο εκπαίδευσης και 43.5% στο σύνολο επικύρωσης. Στα παρακάτω σχήματα φαίνεται η ακρίβεια και η απώλεια αυτού του μοντέλου.


Εικόνα 4. Ακρίβεια και απώλεια των συνόλων εκπαίδευσης και επικύρωσης

Στο πρώτο σχήμα φαίνεται η ακρίβεια του συνόλου εκπαίδευσης (μπλε) και η ακρίβεια του συνόλου επικύρωσης (πορτοκαλί). Η διαφορά μεταξύ τους είναι 0.115 όχι ιδιαίτερα μεγάλη ώστε να θεωρηθεί ότι το μοντέλο έχει υπερπροσαρμοστεί στα δεδομένα (overfiiting).

Παρακάτω παρουσιάζεται μία σύγκριση μεταξύ του συγκεκριμένου μοντέλου και ενός από τα μοντέλα που δοκιμάστηκαν. Το μοντέλο αυτό είναι ίδιο με το τελικο μοντέλο με την αλλαγή ότι τα average pool επίπεδα έχουν αντικατασταθεί με max pool.

| | loss | accuracy | val_loss | val_accuracy |
|---|---|---|---|---|
| average pool | 1.2223 | 0.5500 | 1.6307 | 0.4350 |
| max pool | 1.4306 | 0.4613 | 1.6130 | 0.4500 |

Στα παρακάτω σχήματα φαίνεται η ακρίβεια και η απώλεια του μοντέλου με τα max pool στρώματα.


Εικόνα 5. Ακρίβεια και απώλεια των συνόλων εκπαίδευσης και επικύρωσης του μοντέλου με τα max pool επίπεδα.

Από τα παραπάνω σχήματα φαίνεται ότι η ακρίβεια του συνόλου εκπαίδευσης δεν έχει σχεδόν καθόλου απόλκιση από την ακρίβεια του συνόλου επικύρωσης το οποίο σημαίνει ότι δεν έχει γίνει υπερπροσαρμογή στα δεδομένα (overfitting). Το ίδιο ισχύει και για το διάγραμμα του σφάλματος όπου η απόκλιση του σφάλματος του συνόλου εκπαίδευσης από το σφάλμα του συνόλου επικύρωσης είναι ελάχιστη. Παρ' όλα αυτά δεν επιλέχτηκε το συγκεκριμένο μοντέλο καθώς η ακρίβεια του μοντέλου, η οποία ήταν 46.13% ήταν αρκέτα χαμηλή.

## VII. Συμπερασματα και μελλοντικες εργασιες

Στην εργασία αυτή διερευνήθηκε η αποτελεσματικότητα του μοντέλου με τη χρήση συνελικτικού νευρωνικού δικτύου (CNN) για ταξινόμηση του μουσικού είδους. Τα πειραματικά αποτελέσματα που προέκυψαν δείχνουν ότι οι δύο ακόλουθοι τρόποι είναι αποτελεσματικοί για τη βελτίωση της ταξινόμησης του μουσικού είδους με CNN. Ο πρώτος τρόπος είναι συνδυάζοντας το max pooling και το average pooling ώστε να παρέχονται περισσότερες στατιστικές πληροφορίες σε νευρωνικά δίκτυα ανώτερου επιπέδου. Ο δεύτερος τρόπος είναι χρησιμοποιώντας συνδέσεις συντόμευσης (shortcut connections) για να παραλειφθεί ένα ή περισσότερα επίπεδα. Η μέθοδος αυτή είναι εμπνευσμένη από την υπολειμματική μάθηση (residual learning).

Μελλοντικά, μπορεί να γίνει προσπάθεια να συνδυαστούν νέες μέθοδοι όπως multi-scale convolution και pooling με υπολειμματική μάθηση (π.χ. inception resnet). Δεδομένου ότι το κάθε τραγούδι έχει μετατραπεί σε φασματογράφημα σε μία μελλοντική εργασία θα μπορούσαν να εξαχθούν τα χαρακτηριστικά από τα ακατέργαστα ηχητικά κλιπ κατευθείαν.

## VIII. Βιβλιογραφια

[1] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in Advances in neural information processing systems, 2009, pp. 1096–1104.

[2] T. L. Li, A. B. Chan, and A. Chun, "Automatic musical pattern feature extraction using convolutional neural network," in Proc. Int. Conf. Data Mining and Applications, 2010.

[3] P. Zhang, X. Zheng, W. Zhang, S. Li, S. Qian, W. He, S. Zhang, and Z. Wang, "A deep neural network for modeling music," in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, 2015, pp. 379–386.

[4] C. Zhang, G. Evangelopoulos, S. Voinea, L. Rosasco, and T. Poggio, "A deep representation for invariance and music classification," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 6984–6988.

[5] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 6964– 6968.

[6] Yandre M G Costa, Luiz S Oliveira, Alessandro L Koericb, and Fabien Gouyon, "Music genre recognition using spectrograms," IEEE International Conference on Systems, Signals and Image Processing(IWSSIP), pp. 1– 4, 2011.

[7] B. Zhen, X. Wu, Z. Liu, and H. Chi, "On the importance of components of the mfcc in speech and speaker recognition." in INTERSPEECH. ISCA, 2000, pp. 487–490.