Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Προσδιορισμός των beat μουσικών κομματιών με μηχανική μάθηση

Πετρίδης Στέφανος - Ευστράτιος

Επιβλέπον: Γεώργιος Στάμου, Αν. Καθηγητής Ε.Μ.Π

Διπλωματική Εργασία

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Αθήνα, Οκτώβριος 2019

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Προσδιορισμός των beat μουσικών κομματιών με μηχανική μάθηση

Πετρίδης Στέφανος - Ευστράτιος

Επιβλέπον: Γεώργιος Στάμου, Αν. Καθηγητής Ε.Μ.Π

Διπλωματική Εργασία

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 29η Οκτωβρίου 2019.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γεώργιος Στάμου	Ανδρέας Σταφυλοπάτης	Νικόλαος Παπασπύρου
Αναπληρωτής Καθηγητής	Καθηγητής	Καθηγητής

Αθήνα, Οκτώβριος 2019

(Υπογραφή)

. . . . . .

Στέφανος - Ευστράτιος Πετρίδης

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

Περίληψη

Το Beat Tracking είναι η ακριβής εκτίμηση της θέσης των beat μέσα σε ένα μουσικό κομμάτι. Το πρόβλημα αυτό είναι ένα από τα πολλά προβλήματα που αποσχολούν τον κλάδο της Ανάκτησης Πληροφορίας από Μουσική. Σε αντίθεση με το Tempo Estimation, που αποσκοπεί τον προσδιορισμό των διακυμάνσεων του tempo σε ένα μουσικό κομμάτι, με το Beat Tracking καλούμαστε να αναδείξουμε τις ακριβή θέση του κάθε beat.

Οι περισσότεροι άνθρωποι, ακούγοντας μουσική, μπορούν με ευκολία να προσδιορίσουν τους χτύπους της. Αρκετοί αλγόριθμοι έχουν δημιουργηθεί για την επίλυση αυτού του προβλήματος. Όμως, η μεταφορά της γνωστικής αυτής διαδικασίας σε ένα αυτοματοποιημένο σύστημα, το οποίο θα λειτουργεί με επιτυχία για πολλά και διαφορετικά είδη μουσικής δεν είναι εύκολη διαδικασία.

Στο πλαίσιο αυτής της εργασίας, προσεγγίζουμε το Beat Tracking από την σκοπιά της βαθιάς μηχανικής μάθησης, η οποία είναι αρκετά διαδεδομένη πλέον στα περισσότερα έργα ανάκτησης πληροφορίας από μουσική. Πιο συγκεκριμένα, εκπαιδεύτηκαν, 4 διαφορετικοί τύποι Αναδρομικών Νευρωνικών Δικυων Long-Short Term Memory(LSTM) καθώς και 1 αρχιτεκτονική Συνελικτικού Νευρωνικού Δικτύου(CNN).

Σαν είσοδο των νευρωνικών δικτύων, χρησιμοποιήσαμε σπεκτρογραφήματα τύπου Mel, που προέρχονται από 1124 διαφορετικά μουσικά κομμάτια.

Λέξεις Κλειδιά: Beat Tracking, Βαθιά Μηχανική Μάθηση, LSTM, CNN, Spectrogram, Mel

Abstract

Beat Tracking is the act of precisely estimating the beat positiions in a music piece. It is a problem that concerns the field of MUsic Information Retrieval(MIR). In contrast with Tempo Estimation, which aims to identify the tempo and tempo fluctuations of a music piece, with Beat Tracking we try to highligh the exact position of every individual beat.

Most humans, when they listen to music, they can easily identify its beats. Many algorithms have been developed to solve this problem. However, the replication of this congitive process to an automated system that works well for many genres of music is not a simple task.

In this thesis, we aim to solve the Beat Tracking problem using deep learning, which is very widespread in the field of Music Information Retrieval. More specificaly, we trained 4 different types of Recurrent Neural Networks using Long - Short Term Memory(LSTM) units as well as 1 type of Convolutiona Neural Network(CNN).

As input for our neural networks, we used Mel spectrograms created from 1124 different pieces of music.

Keywords: Beat Tracking, Deep Learning, LSTM, CNN, Spectrogram, Mel

Περιεχόμενα

1 Εισαγωγή **1**
1.1 Ορισμός προβλήματος . . . 3
1.2 Σκοπός . . . 3
1.3 Δομή της εργασίας . . . 3

2 Μουσική Θεωρία **5**
2.1 Ορισμός Μουσικής . . . 5
2.2 Ρυθμός . . . 5
2.3 Μουσικά Γεγονότα . . . 6

3 Στοιχεία Επεξεργασίας Σήματος **7**
3.1 Short-Time Fourier Transform . . . 7
3.2 Filter-bank . . . 8

4 Νευρωνικά Δίκτυα **10**
4.1 Τοπολογία . . . 11
4.1.1 Τεχνητός Νευρώνας . . . 11
4.1.2 Perceptron πολλών επιπέδων . . . 12
4.2 Συνελικτικά Νευρωνικά Δίκτυα . . . 13
4.2.1 Convolutional layer . . . 14
4.2.2 Pooling Layer . . . 15
4.3 Αναδρομικά Νευρωνικά Δίκτυα . . . 16
4.3.1 Πως λειτουργεί ένα Αναδρομικό Νευρωνικό Δίκτυο . . . 16
4.3.2 Long Short-Term Memory . . . 18
4.4 Εκπαίδευση Δικτύου . . . 20

5 State of the Art **23**
5.1 Δυναμικός Προγραμματισμός . . . 24
5.2 Προσεγγίσεις από πολλούς συντελεστές . . . 24
5.3 Συνδυαστικές Μέθοδοι . . . 25
5.4 Προσέγγιση με Βαθιά Μηχανική Μάθηση . . . 25

6 Δεδομένα **27**
6.1 Datasets . . . 27
6.1.1 Ballroom . . . 27
6.1.2 Hainsworth . . . 29
6.1.3 SMC . . . 31
6.2 Δυσκολία Εύρεσης Αξιόπιστων Dataset . . . 32

7 Μεθολογία **34**
7.1 Μεθοδολογία . . . 34
7.1.1 Ηχητική Κυμματομορφή . . . 34
7.1.2 Σπεκτρογράφημα Mel . . . 35
7.1.3 Νευρωνικό Δίκτυο . . . 35
7.1.4 Post Processing . . . 36
7.1.4.1 Συνάρτηση Αυτοσυσχέτισης . . . 36

7.1.4.2 Υπολογισμός κύριου παλμού παραθύρου 38
7.1.5 Εκπαίδευση 38
7.1.5.1 Class Imbalance 39

8 Πειραματικό Στάδιο **40**
8.1 Αξιολόγηση 40
8.2 Πειράματα 41
8.2.1 Πειράματα 1-4 41
8.2.1.1 Πείραμα 1 41
8.2.1.2 Πείραμα 2 44
8.2.1.3 Πείραμα 3 46
8.2.1.4 Πείραμα 4 48
8.2.2 Πείραμα 5 50
8.2.3 Πειράματα 6-7 52
8.2.3.1 Πείραμα 6 52
8.2.3.2 Πείραμα 7 52
8.2.4 Πείραμα 8 53
8.2.5 Πείραμα 9 56

9 Συμπεράσματα και Μελλοντική Ανάπτυξη **58**
9.0.1 Συμπεράσματα 58
9.0.2 Μελλοντική Ανάπτυξη 59

References **60**

1 Εισαγωγή

Τις τελευταίες δεκαετίες, παρατηρείται μια σημαντικά μεγάλη επανάσταση στα πλαίσια της μουσικής δημιουργίας. Η αύξηση της δημοτικότητας παρόχων μουσικής όπως το Spotify, το Apple Music, το Tidal και το SoundCloud, προξένησαν αλλαγές στις μουσικές συνήθειες των καταναλωτών, από την αγορά φυσικών μέσων(όπως δίσκους μουσικής) σε κατανάλωση από υπηρεσίες streaming. Αυτό είχε ως αποτέλεσμα την αύξηση της διαθέσιμης μουσικής. Μεγάλη ποσότητα δεδομένων πρέπει να επεξεργαστούν ώστε να παρέχονται σωστές προτάσεις στους χρήστες των παραπάνω υπηρεσιών ανάλογα με τις μουσικές τους προτιμήσεις και ανάλογα με τη μουσική που έχουν ήδη ακούσει χρησιμοποιώντας την υπηρεσία. Η μουσική που παρέχεται ταξινομείται με βάση τα χαρακτηριστικά της(όπως είδος, διάθεση και άλλα). Είναι αδύνατο να γίνει η ταξινόμηση αυτή χειροκίνητα, αφού η ποσότητα των δεδομένων αυξάνεται εκθετικά. Επίσης, η μουσική αντίληψη διαφέρει από άνθρωπο σε άνθρωπο, πράγμα που οδηγεί σε διφορούμενους χαρακτηρισμούς. Όλα αυτά τα ανοιχτά ερωτήματα οδήγησαν στη δημιουργία του κλάδου **Ανάκτησης και Ανακατασκευής Πληροφορίας από Μουσική**(Music Information Retrieval, MIR). Σκοπός του MIR είναι η αυτόματη εξαγωγή χρήσιμων πληροφοριών από μουσική, χρησιμοποιώντας τεχνικές που προέρχονται από τους τομείς της επεξεργασίας σήματος, της μηχανικής μάθησης, της ψυχοακουστικής θεωρίας κ.α. Η πληροφορία κωδικοποιείται σε χαρακτηριστικά που υπολογίστηκαν για την περιγραφή της μουσικής. Αυτά τα χαρακτηριστικά μπορούν να χωριστούν σε τρία επίπεδα, ανάλογα με το πόσο κοντά στην ανθρώπινη αντίληψη βρίσκονται. Τα χαρακτηριστικά χαμηλού επιπέδου προέρχονται άμεσα από το ηχητικό σήμα και δεν είναι αντιληπτά από τον άνθρωπο αλλά χρησιμοποιούνται για να κωδικοποιήσουν τα χαρακτηριστικά του σήματος. Τα χαρακτηριστικά μεσαίου επιπέδου υπολογίζονται από αυτά το χαμηλού επιπέδου. Με τη χρήση μουσικής θεωρίας, αυτά τα χαρακτηριστικά περιγράφουν με μεγαλύτερη ακρίβεια τη μουσική όπως την αντιλαμβάνεται ο άνθρωπος(ρυθμός, αρμονία). Τέλος, τα χαρακτηριστικά υψηλού επιπέδου αντιπροσωπεύουν πιο αφηρημένες έννοιες, όπως το συναίσθημα. Σε αυτή τη διατριβή επικεντρωνόμαστε στον ρυθμό και ειδικότερα, στην ανίχνευση ενός από τα θεμελιώδη χαρακτηριστικά του: τον χτύπο(beat). Το beat είναι αυτό που κάνει εμάς, τους ανθρώπους, να κουνάμε το κεφάλι μας ή να χτυπάμε παλαμάκια ακούγοντας μουσική. Η δυνατότητα του ακροατή να το αντιληφθεί συνδέεται με την ικανότητα του ανθρώπου να μπορεί να αντιλαμβάνεται τις περιοδικότητες σε μια σειρά γεγονότων μέσα σε ένα μουσικό κομμάτι.

Η αντίληψη των γεγονότων σχετίζεται με τα ηχητικά χαρακτηριστικά του σήματος όπως το transient. Το transient ορίζεται σαν μια μεγάλη αύξηση ακολουθούμενη από μία μείωση του πλάτους του σήματος σε μικρό χρόνο. Η σειρά μουσικών γεγονότων γίνεται αντιληπτή σαν μια σειρά από transients που βρίσκονται στο ηχητικό σήμα. Επιπρόσθετα, το πως αντιλαμβανόμαστε ένα μουσικό κομμάτι, δεν έχει να κάνει μόνο με τη συγκεκριμένη θέση του κάθε transient, αλλά και με τη γενικότερη δομή του κομματιού. Η ρυθμική δομή παρέχει μια οργανωμένη και περιοδική παράταξη των μουσικών γεγονότων. Έτσι, βασιζόμενοι σε αυτή τη δομή, οι ακροατές μπορούν να καταλάβουν την εμφάνιση του τοπικού beat, αλλά και να προβλέψουν την τα επόμενα beats λόγω της περιοδικής τους προδιάθεσης.

Το **Beat Tracking** είναι ένα από τα σημαντικότερα πεδία του κλάδου του MIR και έχει σαν στόχο την αυτόματη ανάκτηση της ακολουθίας των χτύπων από ένα μουσικό κομμάτι. Αυτό είναι πολύ σημαντικό ώστε να αναλυθεί η ρυθμική δομή ενός ηχητικού σήματος, που είναι χρήσιμο χαρακτηριστικό για την ομαδοποίηση τραγουδιών. Επιπρόσθετα, το beat, ως χαρακτηριστικό μεσαίου επιπέδου, μπορεί να χρησιμοποιηθεί για ανάλυση υψηλότερου επιπέδου. Επίσης, η αυτόματη ανάκτηση της ακολουθίας των χτύπων μπορεί να χρησιμοποιηθεί σαν εργαλείο για την αυτόματη διόρθωση ηχογραφήσεων ή για τη δημιουργία αυτόματα προσαρμοζόμενων ηχητικών εφέ(για χρήση σε ζωντανές παραστάσεις).

Πολλές μέθοδοι έχουν δημιουργηθεί για να επιλύσουν το πρόβλημα του beat tracking. Οι περισσότερες βασίζονται σε πιθανοτικά μοντέλα για να μιμηθούν την ανθρώπινη αντίληψη της μουσικής δομής. Η **βαθιά μηχανική μάθηση**(deep learning) είναι μία από τις πιο πολλά υποσχόμενες μεθόδους.

Στον κλάδο του MIR, η βαθιά μηχανική μάθηση έχει χρησιμοποιηθεί για την επίλυση αρκετών προβλημάτων, όπως, ομαδοποίηση μουσικής σε είδη, ανάλυση δομής της μουσικής και beat tracking. Σε ό,τι αφορά το beat tracking, οι αρχιτεκτονικές που προτείνονται συνήθως είναι αυτές των **Αναδρομικών Νευρωνικών Δικτύων**(Recurrent Neural Networks, RNN), τα οποία είναι ειδικά φτιαγμένα για να χρησιμοποιούνται σε προβλήματα χρονοσειρών. Σε αυτή τη διατριβή κατασκευάζουμε, μελετάμε και συγκρίνουμε ορισμένες αρκετά δημοφιλείς αρχιτεκτονικές στον χώρο του beat tracking και εξετάζουμε τη συμπεριφορά τους σε διαφορετικά δεδομένα εκπαίδευσης.

1.1 Ορισμός προβλήματος

Στο σημείο αυτό, είναι σημαντικό να ορίσουμε ακριβώς το πρόβλημα που θα αντιμετωπίσουμε σε αυτή την εργασία.

Έστω N ο αριθμός των μουσικών χτύπων σε ένα κομμάτι μουσικής. Αριθμώντας, όλους τους χτύπους, ορίζουμε το σετ των μουσικών χτύπων ως $\mathcal{B} = [1 : N] := \{1, 2, \ldots, N\}$. Δεδομένου ενός αρχείου μουσικής με την εκτέλεση του κομματιού, τα μουσικά beats αντιστοιχούν σε συγκεκριμένες χρονικές θέσεις μέσα στο αρχείο αυτό. Έστω $\pi : \mathcal{B} \to \mathbb{R}$ η αντιστοιχία που ορίζει τη χρονική θέση $\pi(b)$ μέσα στο έργο μουσικής για κάθε χτύπο $b \in \mathcal{B}$. Σκοπός του beat tracking είναι η ανάκτηση του σετ $\{\pi(b)|b \in \mathcal{B}\}$.

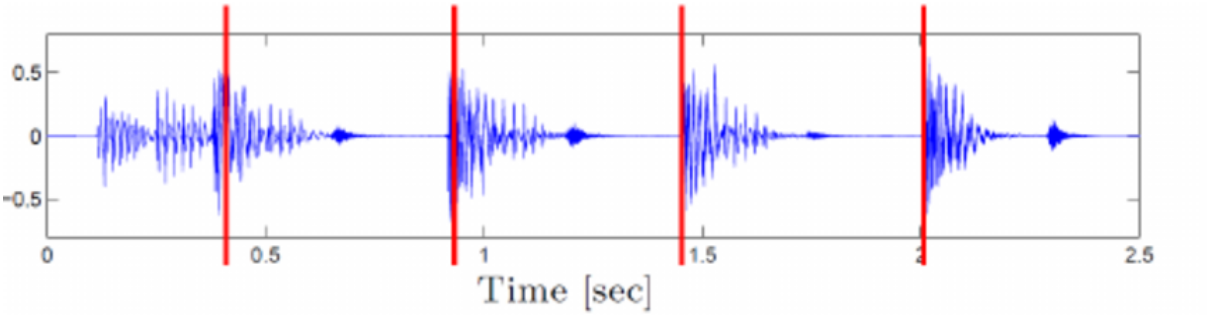

Figure. 1.1: Χρονικές θέσεις των χτύπων σε κυμματομορφή από αρχείο μουσικής

1.2 Σκοπός

Σκοπός της παρούσας εργασίας είναι η κατασκευή και μελέτη συστημάτων για beat tracking. Σκοπός δεν είναι η δημιουργία ενός state of the art συστήματος. Στόχος είναι η σύγκριση των αποδόσεων των διαφόρων νευρωνικών δικτύων που δημιουργήθηκαν. Εξετάζουμε, επίσης την επίδραση διαφόρων υπερπαραμέτρων των συστημάτων και τις αλλαγές που επιφέρουν στο τελικό αποτέλεσμα.

1.3 Δομή της εργασίας

Η εργασία αυτή χωρίζεται σε 7 κεφάλαια.

Στα κεφάλαια 2,3 και 4, περιγράφονται τα στοιχεία που χρησιμοποιήθηκαν για την ανάπτυξη αυτής της εργασίας.

Στο Κεφάλαιο 2, δίνεται ο ορισμός της Μουσικής και περιγράφονται οι έννοιες του Ρυθμού και του Μουσικού Γεγονότος.

Στο Κεφάλαιο 3, περιγράφονται οι τεχνικές που χρησιμοποιήθηκαν από το πεδίο της Επεξεργασίας Σήματος

Στο κεφάλαιο 4, παρουσιάζονται τα διάφορα είδη αρχιτεκτονικών Νευρωνικών Δικτύων που χρησιμοποιήθηκαν, καθώς και οι τεχνικές εκπαίδευσής τους.

Στο Κεφάλαιο 5, δίνονται οι διάφορες τεχνικές και αλγόριθμοι που έχουν χρησιμοποιηθεί για την επίλυση του προβλήματος του Beat Tracking.

Στο Κεφάλαιο 6, παρουσιάζονται τα διαφορετικά datasets που χρησιμοποιήθηκαν στο πλαίσιο αυτής της εργασίας.

Στο Κεφάλαιο 7, περιγράφεται η μεθοδος που ακολουθήθηκε σε αυτή την εργασία για την επίλυση του προβλήματος του Beat Tracking.

Το Κεφάλαιο 8, περιλαμβάνει τα πειράματα που πραγματοποιήθηκαν.

Τέλος, στο Κεφάλαιο 9, δίνονται τα τελικά συμπεράσματα, καθώς και ένα πιθανό μελλοντικό πλάνο εργασίας.

2 Μουσική Θεωρία

2.1 Ορισμός Μουσικής

Ως μουσική ορίζεται η τέχνη που βασίζεται στην οργάνωση ήχων με σκοπό τη σύνθεση, εκτέλεση και ακρόαση/λήψη ενός έργου. Με τον όρο μουσική εννοείται επίσης και το σύνολο ήχων από το οποίο απαρτίζεται ένα μουσικό κομμάτι. Τόσο ο ορισμός της μουσικής, όσο και σχετικά με τη μουσική θέματα όπως η εκτέλεση, η σύνθεση και η σπουδαιότητά της, διαφέρουν από πολιτισμό σε πολιτισμό και ανάλογα με το κοινωνικό πλαίσιο. Η ερώτηση 'τι είναι μουσική;' έχει γίνει θέμα συζητήσεων - μεταξύ λογίων και μη -, έχει δεχτεί πληθώρα απαντήσεων, όμως καμία δεν ερμηνεύει το φαινόμενο της εν λόγω τέχνης σε καθολικό, διαπολιτισμικό επίπεδο. Οι περισσότεροι ορισμοί, που επιχειρούν να περιγράψουν τη μουσική, περιέχουν κοινά στοιχεία όπως ο τόνος(που περιγράφει τη μελωδία και την αρμονία), ο ρυθμός(που συνδέεται με έννοιες όπως το tempo, το μέτρο και η ερμηνεία), η δυναμική και ηχητικές ιδιότητες όπως η ποιότητα του τόνου(timbre). Διαφορετικά είδη μουσικής πιθανόν να τονίζουν ή να παραλείπουν κάποια από τα παραπάνω στοιχεία. Η μουσική εκτελείται από ένα ευρύ φάσμα μουσικών οργάνων και φωνητικών τεχνικών. Επίσης, υπάρχουν μουσικά έργα στα οποία δεν υπάρχει τραγούδι(instrumental) και άλλα που περιέχουν αποκλειστικά τραγούδι(a capella).Γνωστή και ως Απολλώνια Τέχνη, η μουσική παίρνει το όνομά της από τις εννέα Μούσες της αρχαίας ελληνικής μυθολογίας.

2.2 Ρυθμός

Ο ρυθμός παράγεται από τη διαδοχική διάταξη ήχων και σιωπών στον χρόνο. Το μέτρο ομαδοποιεί τους μουσικούς παλμούς ανάλογα με τα επαναλαμβανόμενα μοτίβα που παρατηρούνται σε ένα έργο μουσικής. Ο βαθμός του μέτρου(time signature) ορίζει τον αριθμό των χτύπων μέσα σε ένα μέτρο. Ο βαθμός αυτός μπορεί να αλλάζει σε όλη τη διάρκεια του κομματιού. Συγκεκριμένες αξίες, κατά τη διάρκεια ενός μουσικού κομματιού, μπορούν να υπερτονιστούν με χρήση υψηλότερων δυναμικών και αλλαγής της διάρκειάς τους. Στις περισσότερες μουσικές παραδόσεις, υπάρχουν συμβάσεις που ορίζουν την ιεραρχία των χτύπων με βάση τον τονισμό τους, ώστε να ορίζεται με μεγαλύτερη σαφήνεια το μέτρο που

χρησιμοποιείται. Οι συγκεκομμένοι(syncopated) ρυθμοί αντιτίθενται σε αυτές τις συμβάσεις τονίζοντας απρόσμενα σημεία των χτύπων. Η ταυτόχρονη χρήση δύο ή και περισσότερων ρυθμών, που ο ένας δεν προέρχεται από τον δεύτερο, ονομάζεται πολυρυθμία(polyrhythm).

2.3 Μουσικά Γεγονότα

Τα μουσικά γεγονότα μπορούν να χωριστούν σε δύο κατηγορίες: κρουστικά γεγονότα(percussive events) και αρμονικά γεγονότα(harmonic events). Το πρώτα περιγράφονται ως μια απότομη αλλαγή του πλάτους του σήματος. Τα δεύτερα χαρακτηρίζονται από αλλαγή στα συχνοτικά στοιχεία του σήματος, αλλά η ενέργεια του σήματος μπορεί να παραμείνει ίδια.

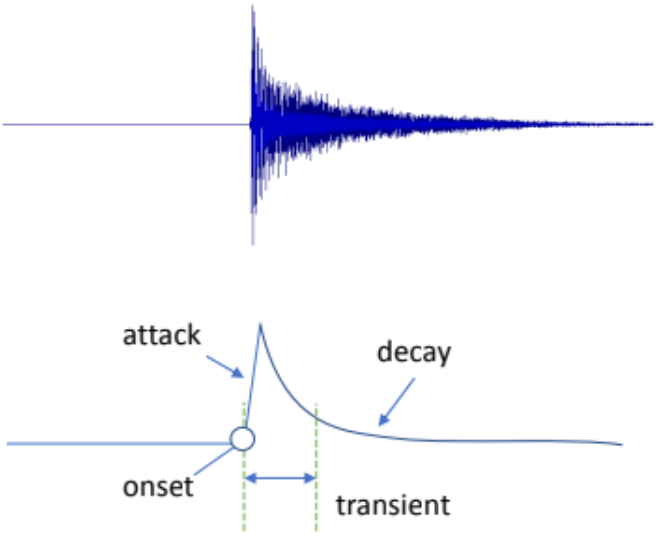

Figure. 2.1: Ένα μουσικό γεγονός και τα χαρακτηριστικά του.

Κάθε γεγονός χαρακτηρίζεται από τέσσερις ιδιότητες: το onset, το attack, το transient και το decay. Το onset προσδιορίζει τη στιγμή, την οποία ξεκινάει το γεγονός. Το attack είναι η χρονικό διάστημα κατά το οποίο, η ένταση του σήματος αυξάνεται. Το transient είναι οι χρονικές στιγμές του γεγονότος όπου το σήμα αλλάζει με γρήγορο ρυθμό και απρόβλεπτα. Τέλος, το decay είναι οι χρονικές στιγμές μετά το attack, όταν το σήμα έχει μια πιο σταθερή εξέλιξη που τελικά οδηγεί στον μηδενισμό του πλάτους του.

3 Στοιχεία Επεξεργασίας Σήματος

Με δεδομένο ένα ηχητικό σήμα, πρέπει να πραγματοποιηθεί μια προεπεξεργασία ώστε να εξάγουμε κάποιες χρήσιμες πληροφορίες από αυτό. Ο **Ο μετασχηματισμός Fourier Μικρής Διάρκειας**(Short-Time Fourier Transform, STFT) αναπαριστά το αρχικό σήμα, και την εξέλιξή του, στους άξονες του χρόνου και της συχνότητας. Είναι, συνήθως, ο ακρογωνιαίος λίθος των συστημάτων που επεξεργάζονται ηχητικά σήματα.

3.1 Short-Time Fourier Transform

Ο Short-Time Fourier Transform είναι ένας μετασχηματισμός που χρησιμοποιείται για να προσδιορίσει τη συχνότικό και φασικό περιεχόμενο σε μικρά μέρη του σήματος $x(t)$, καθώς αυτό αλλάζει στον χρόνο t.

Το διακριτό σήμα, $x(l)$, χωρίζεται σε επικαλυπτόμενα κομμάτια(frames) μήκους W, καθώς πολλαπλασιάζεται με ένα ολισθαίνων παράθυρο(sliding window), $w(l)$, για το οποίο ισχύει: $w(l) \neq 0$ για $-W/2 \leq l \leq W/2 - 1$. Για να υπολογιστεί ο STFT, τα δείγματα του παραθυροποιημένου σήματος πολλαπλασιάζονται με έναν φασιθέτη. Έτσι προκύπτει ο παρακάτω τύπος:

$$X(n,k) = \sum_{l=-W/2}^{W/2-1} w(l) \cdot x(l+nh) \cdot e^{-2\pi jlk/W}$$

Στον παραπάνω τύπο, ως n ορίζουμε το frame index, ως h τον αριθμό των αλμάτων σε δείγματα(hop size) στον άξονα του χρόνου μεταξύ γειτονικών frames και ως k το index της ομάδας συχνοτήτων που υπολογίζουμε.

Όσο μεγαλύτερο είναι το W, τόσο περισσότερα δείγματα του σήματος συμπεριλαμβάνονται σε ένα frame. Έτσι, έχουμε μεγαλύτερη ευκρίνεια στον άξονα των συχνοτήτων. Όταν θέλουμε να εξετάσουμε τα harmonic events του σήματος προτιμάμε μεγαλύτερες τιμές του παραθύρου, W. Αντίθετα, όσο μικρότερο είναι το παράθυρο, τόσο μεγαλύτερη ευκρίνεια έχουμε στον άξονα του χρόνου. Αυτό βοηθάει στην ανάλυση των percussive events. Τα percussive events, όπως αναφέραμε, χαρακτηρίζονται από απότομες αλλαγές της έντασης του σήματος, η οποία κατανέμεται σε όλο το μήκους του παραθύρου. Έτσι, αν το παράθυρο είναι αρκετά μεγάλο ώστε να χωρέσει περισσότερα από ένα percussive events, τότε δε θα μπορέσουμε να

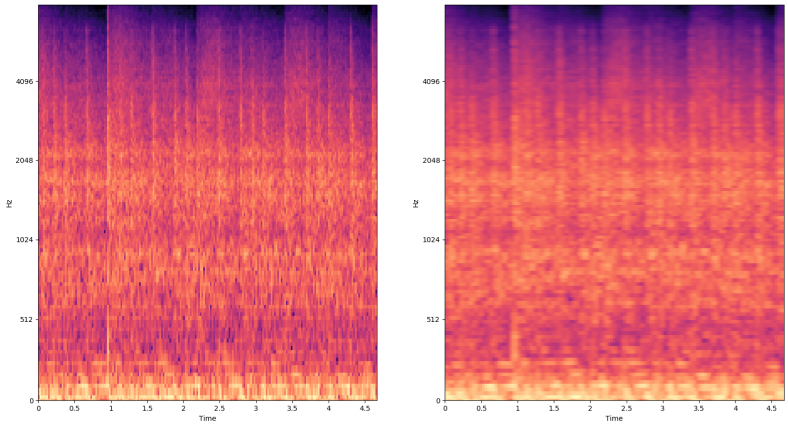

Figure. 3.1: Αριστερά, STFT spectrogram με $W = 1024$. Δεξιά, STFT με $W = 4096$. Όλες οι άλλες παράμετροι είναι ίδιες και στα δύο σχήματα.

τα διαχωρίσουμε. Αυτό το φαινόμενο ονομάζεται **συμβιβασμός χρονου-συχνότητας**(time-frequency compromise).

3.2 Filter-bank

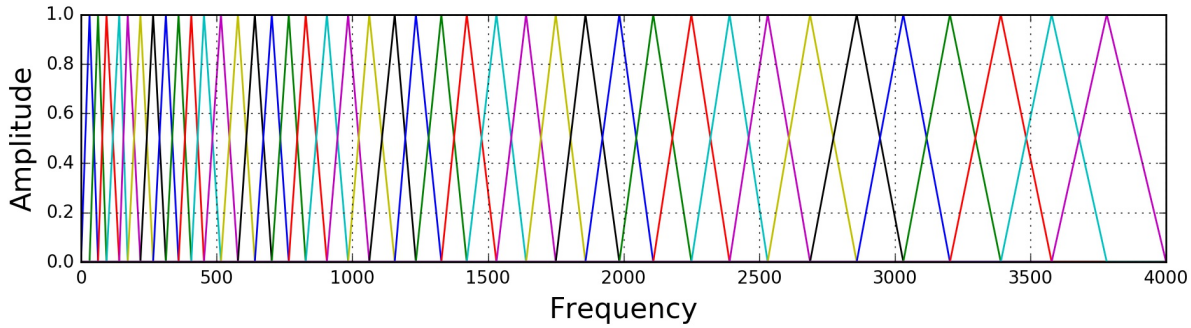

Figure. 3.2: Τριγωνικό filter-bank 40 φίλτρων κατανεμημένα λογαριθμικά στον άξονα των συχνοτήτων.

Με σκοπό να προσεγγίσουμε με μεγαλύτερη ακρίβεια την ανθρώπινη ακοή, μπορούμε να φιλτράρουμε κάθε frame του STFT με ένα filter-bank. Το filter-bank αποτελείται από μία σειρά από επικαλυπτόμενα φίλτρα. Κάθε φίλτρο αναλύει μόνο ένα συχνοτικό μέρος του frame και από αυτό παράγεται μια νέα παράμετρος. Με αυτόν τον τρόπο, και υπολογίζοντας όλες τις παραμέτρους που μπορεί να παράξει το filter-bank, μειώνουμε αισθητά τις παραμέτρους του STFT στον άξονα της συχνότητας. Για παράδειγμα, στην περίπτωση που χρησιμοποιήσουμε ένα filter-bank με 20 φίλτρα. Οι k παράμετροι του STFT στον άξονα της συχνότητας γίνονται 20.

Ένα πολύ διαδεδομένο filter-bank, που χρησιμοποιείται για να προσεγγίσει την ανθρώπινη ακοή, είναι το Mel filter-bank. Στο συγκεκριμένο filter-bank, φίλτρα τοποθετούνται σε ίσες αποστάσεις μεταξύ τους στη λογαριθμική κλίμακα Mel. Στο πλαίσιο αυτής της εργασίας, χρησιμοποιούμε filter-banks τέτοιου τύπου.

4 Νευρωνικά Δίκτυα

Τα **Τεχνητά Νευρωνικά Δίκτυα**(Artificial Neural Networks, ANN) είναι υπολογιστικά συστήματα εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα, που σχηματίζουν τον εγκέφαλο. Τα συστήματα αυτά ”μαθαίνουν” να εκτελούν διεργασίες μετά από εκπαίδευση, χωρίς να τους δίνονται συγκεκριμένες εντολές για αυτή τη διεργασία. Για παράδειγμα, στην αναγνώριση εικόνων, ένα δίκτυο μπορεί να εκπαιδευτεί να αναγνωρίζει εικόνες που περιέχουν γάτες, αναλύοντας παραδείγματα εικόνων τα οποία έχουν ονομαστεί από τον ”εκπαιδευτή” ως ”γάτα” ή ”όχι γάτα”. Μετά την εκπαίδευση, χρησιμοποιεί τη γνώση που απέκτησε, ώστε να αναγνωρίζει τις γάτες σε άλλες εικόνες.

Ειδικότερα, τα Νευρωνικά Δίκτυα είναι ισχυρά μη γραμμικά μοντέλα τα οποία αντιστοιχίζουν την είσοδο που τους δίνεται, $\mathbf{x}$ σε μια έξοδο, $\mathbf{y}$. Ουσιαστικά, το νευρωνικό δίκτυο μπορεί να περιγραφεί από τον τύπο, $\mathbf{y} = f(\mathbf{x}; \mathbf{\Theta})$, όπου f η συνάρτηση αντιστοίχισης και $\mathbf{\Theta}$ οι εκπαιδεύσιμες παράμετροι. Η αντιστοίχιση μαθαίνεται κατά τη διαρκεί της εκπαίδευσης, όπου οι παράμετροι $\mathbf{\Theta}$ μεταβάλλονται. Αλλάζοντας τις $\mathbf{\Theta}$ κατά την εκπαίδευση, αλλάζουν τα βάρη των συνδέσεων του δικτύου και έτσι, τελικά το δίκτυο μπορεί να μάθει τη βέλτιστη αναπαράσταση της εισόδου. Βέβαια, οι δυνατότητες του δικτύου εξαρτώνται από την τοπολογία του.

4.1 Τοπολογία

Η αρχιτεκτονική ενός νευρωνικού δικτύου ορίζεται από τις επεξεργαστικές του μονάδες(units) και από τον τρόπο που αυτές συνδέονται. Στη συνέχεια εξετάζουμε τη δομή των δικτύων ξεκινώντας από τον απλό νευρώνα και συνεχίζοντας με πιο περίπλοκες αρχιτεκτονικές.

4.1.1 Τεχνητός Νευρώνας

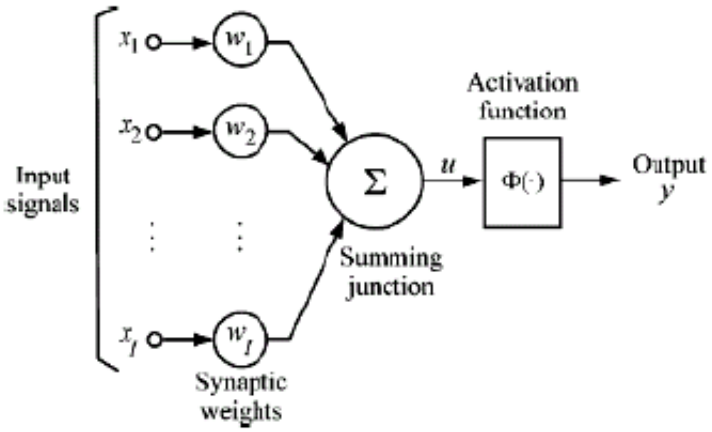

Figure. 4.1: Γραφική αναπαράσταση ενός τεχνητού νευρώνα

Η βασική επεξεργαστική μονάδα των ANN είναι ο τεχνητός νευρώνας. Η έξοδος του νευρώνα, h, υπολογίζεται σαν αποτέλεσμα μιας μη γραμμικής συνάρτησης, g, που ονομάζεται συνάρτηση ενεργοποίησης(activation function), σύμφωνα με το σταθμισμένο άθροισμα των εισόδων, **x**. Ο υπολογισμός της εξόδου δίνεται από τον παρακάτω τύπο:

$$h = g(\mathbf{x}^T\mathbf{w} + b) = g(\mathbf{x}; \boldsymbol{\theta})$$

όπου **w** είναι το διάνυσμα των βαρών του νευρώνα και b η προκατάληψή του(bias). Για να διευκολύνουμε την αναπαράσταση ενσωματώνουμε το bias στο διάνυσμα **w** και προκύπτει το διάνυσμα $\boldsymbol{\theta}$.

Η συνάρτηση ενεργοποίησης είναι μια μη γραμμική συνάρτηση, η επιλογή της οποίας παίζει σημαντικό ρόλο στη διαδικασία της μάθησης. Κάποιες από τις πιο δημοφιλείς activation functions είναι: η σιγμοειδής(sigmoid ή logistic function), η υπερβολική εφαπτομένη και η softmax.

Χρησιμοποιώντας έναν μοναδικό νευρώνα, μπορούν να επιλυθούν γραμμικά διαχωρίσιμα προβλήματα. Η υπολογιστική ισχύς ενός νευρώνα είναι αρκετά περιορισμένη, όμως, η δύναμη των νευρωνικών δικτύων έγκειται στις συνδέσεις μεταξύ των νευρώνων.

4.1.2 Perceptron πολλών επιπέδων

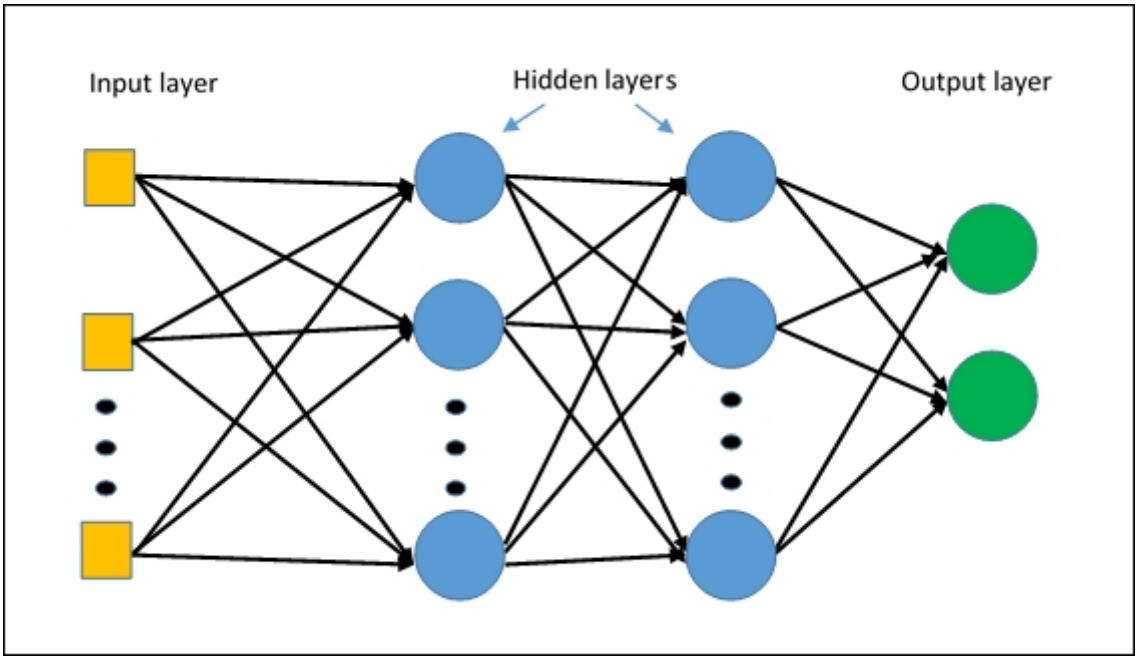

Figure. 4.2: Γραφική αναπαράσταση multiple layer perceptron με δύο κρυφά επίπεδα

Σύνδεση μεταξύ δύο νευρώνων υφίσταται, όταν η έξοδος του ενός χρησιμοποιείται σαν είσοδος του άλλου. Η σύνδεση πολλών νευρώνων μεταξύ τους μπορεί να δημιουργήσει μια συνάρτηση που να έχει τη δυνατότητα να λύση μη γραμμικά διαχωρίσιμα προβλήματα. Μια από τις πιο

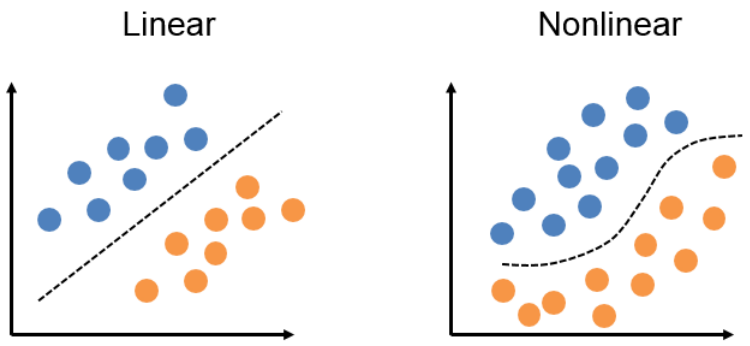

Figure. 4.3: Αριστερά, γραμμικά διαχωρίσιμο και δεξιά, μη γραμμικά διαχωρίσιμο πρόβλημα. Παρατηρούμε ότι το όριο απόφασης στο γραμμικά διαχωρίσιμο πρόβλημα είναι μια ευθεία γραμμή, ενώ στο μη γραμμικά διαχωρίσιμο όχι

συνηθισμένες κατασκευές είναι το **Perceptron Πολλών Επιπέδων**(Multiple Layer Perceptron, MLP). Στην τοπολογία αυτού του τύπου, οι νευρώνες διαμερίζονται σε επίπεδα(layers), και τα γειτονικά επίπεδα συνδέονται με συνδέσεις μονής κατεύθυνσης, χωρίς feedback. Ουσιαστικά, κάθε νευρώνας του i επιπέδου δέχεται σαν είσοδο τις εξόδους από κάθε νευρώνα του $i-1$ επιπέδου. Η έξοδός του τροφοδοτείται στους νευρώνες του επόμενου επιπέδου, ή στην έξοδο του δικτύου αν ο νευρώνας ανήκει στο τελευταίο κρυφό επίπεδο.

Για ένα δίκτυο με m επίπεδα, για την έξοδο του iοστού επιπέδου, όπου $0 < i < m$, ισχύει:

$$\mathbf{h}^{(i)} = g^{(i)}(\mathbf{W}^{(i)}\mathbf{h}^{(i-1)} + \mathbf{b}^{(i)}) = g^{(i)}(\mathbf{h}^{(i-1)}; \boldsymbol{\theta}^{(i)})$$

όπου $\mathbf{h}^{(i)}$ η έξοδος του επιπέδου, $g^{(i)}$ η συνάρτηση ενεργοποίησης των νευρώνων του επιπέδου, $\mathbf{W}^{(i)}$ ο πίνακας των βαρών των συνδέσεων και $\mathbf{b}^{(i)}$ το διάνυσμα της προκατάληψης. Όμοια,

$$\mathbf{h}^{(0)} = g^{(0)}(\mathbf{x}; \boldsymbol{\theta}^{(0)})$$

και

$$\mathbf{h}^{(m-1)} = g^{(0)}(\mathbf{h}^{(m-2)}; \boldsymbol{\theta}^{(m-1)})$$

Τα επίπεδα $0 < i < m-1$ ονομάζονται **κρυφά επίπεδα**(hidden layers) αφού η συμβολή τους στο δίκτυο δεν είναι εμφανής εκτός του μοντέλου. Τα επίπεδα $i = 0$ και $i = m-1$ ονομάζονται επίπεδο εισόδου και εξόδου, αντίστοιχα. Γενικά, ένα δίκτυο ονομάζεται **βαθύ**(deep) αν $m > 3$, δηλαδή έχει τουλάχιστον ένα κρυφό επίπεδο.

4.2 Συνελικτικά Νευρωνικά Δίκτυα

Στη βαθιά μηχανική μάθηση, τα **Συνελικτικά Νευρωνικά Δίκτυα**(Convolutional Neural Networks, CNN) είναι μια κατηγορία νευρωνικών δικτύων που χρησιμοποιούνται συνήθως στον κλάδο της ανάλυσης εικόνας.

Τα CNN είναι κανονικοποιημένες εκδόσεις των MLP. Τα MLP είναι πλήρως συνδεδεμένα δίκτυα. Αυτή η συνδεσιμότητα τα καθιστά επιρρεπή σε υπερεκπαίδευση. Τα CNNs αξιοποιούν τα ιεραρχικά μοτίβα που υπάρχουν στα δεδομένα και κατασκευάζουν περίπλοκα πρότυπα χρησιμοποιώντας μικρότερα και απλούστερα. Έτσι, τα CNNs είναι λιγότερο περίπλοκα δίκτυα

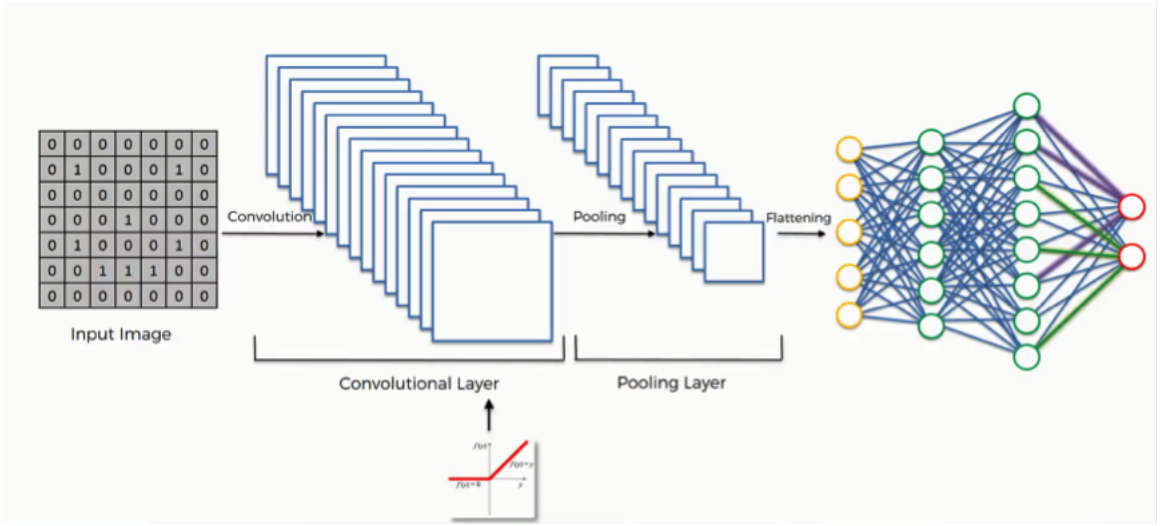

Figure. 4.4: Γραφική αναπαράσταση ενός CNN που συνδέεται με ένα MLP

από τα MLP. Αυτός ο τύπος δικτύου είναι εμπνευσμένος από τον οπτικό φλοιό(visual cortex) του εγκεφάλου.

Η λέξη ”συνελικτικό” στην ονομασία του δικτύου υποδεικνύει ότι αξιοποιεί τη μαθηματική πράξη της συνέλιξης. Η συνέλιξη είναι ένας ειδικός τύπος γραμμικού υπολογισμού. Τα CNN χρησιμοποιούν τη συνέλιξη στη θέση του πολλαπλασιασμού πινάκων, σε τουλάχιστον ένα από τα επίπεδά τους.

Τα κρυφά επίπεδα ενός CNN, συνήθως, αποτελούνται από μια σειρά από **συνελικτικά επίπεδα**(convolutional layers) που πραγματοποιούν συνελίξεις. Συνήθως, η συνάρτηση ενεργοποίησης αυτών των επιπέδων είναι ένα ReLu layer που ακολουθείται από επιπλέον επίπεδα, που ονομάζονται **επίπεδα υπερδειγματοληψίας**(pooling layers)

4.2.1 Convolutional layer

Κατά τον προγραμματισμό ενός CNN, κάθε convolutional layer πρέπει να έχει τα παρακάτω γνωρίσματα:

- Η είσοδός του πρέπει να είναι ένα tensor με διαστάσεις (αριθμός εικόνων) x (πλάτος εικόνας) x (ύψος εικόνας) x(αριθμός χρωμάτων(image depth))
- Συνελικτικούς πυρήνες(convolutional kernels) το πλάτος και το ύψος των οποίων είναι υπερπαράμετροι και το βάθος τους ίσο με το image depth των εικόνων που επεξεργάζεται. Τα συνελικτικά επίπεδα διασχίζουν την είσοδο σε μέρη και περνάνε το αποτέλεσμα

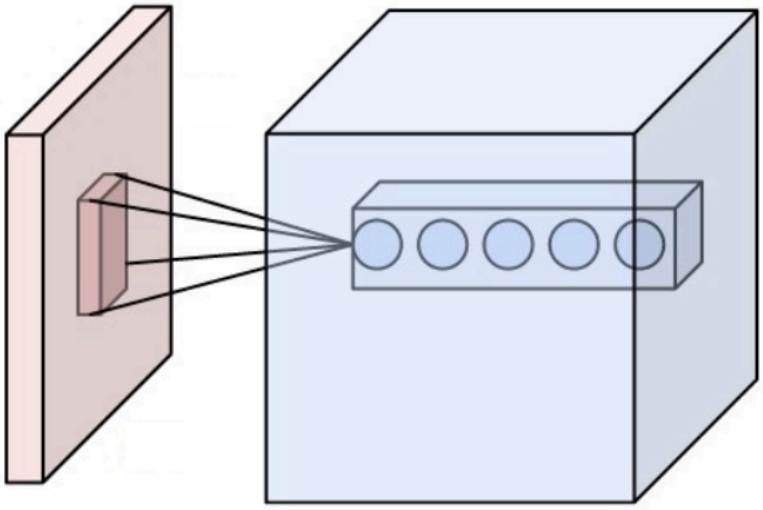

Figure. 4.5: Οι νευρώνες του συνελικτικού επιπέδου συνδέονται με το δεκτικό τους πεδίο

τους στο επόμενο επίπεδο. Αυτή η λειτουργία είναι παρόμοια με τον τρόπο που ανταποκρίνονται οι νευρώνες στον οπτικό φλοιό για ένα συγκεκριμένο ερέθισμα.

Κάθε συνελικτικός νευρώνας επεξεργάζεται δεδομένα που βρίσκονται αποκλειστικά στο δεκτικό του πεδίο(receptive field). Αν και για τις ίδιες λειτουργίες θα μπορούσε να χρησιμοποιηθεί ένα fully connected MLP, δεν είναι πρακτικό να χρησιμοποιηθεί αυτή η αρχιτεκτονική για την επεξεργασία εικόνων. Θα έπρεπε να χρησιμοποιηθεί μεγάλο πλήθος νευρώνων, ακόμα και σε αρχιτεκτονικές λίγων επιπέδων, εξαιτίας του μεγάλου μεγέθους της εισόδου. Για παράδειγμα, ένα MLP, για μια μικρή εικόνα διαστάσεων 100 x 100, έχει 10000 παραμέτρους για κάθε νευρώνα του δευτέρου επιπέδου. Η συνέλιξη δίνει λύσει σε αυτό το πρόβλημα, αφού μειώνει τον αριθμό των εκπαιδεύσιμων παραμέτρων, επιτρέποντας στο δίκτυο να έχει περισσότερα επίπεδα και λιγότερες παραμέτρους. Παραδείγματος χάρην, ανεξάρτητα από το μέγεθος της εικόνας, η χρήση πυρήνων 5 x 5, που μοιράζονται τα ίδια βάρη, χρειάζεται μόνο 25 εκπαιδεύσιμες παραμέτρους.

4.2.2 Pooling Layer

Τα CNN ορισμένες φορές περιλαμβάνουν τοπικά ή μην επίπεδα για τον εξορθολογισμό του υπολογισμού. Τα pooling layers μειώνουν τις διαστάσεις των δεδομένων συγκεντρώνοντας τις εξόδους των νευρώνων σε ομάδες. Το **τοπικό pooling**(local pooling) κάνει μικρές ομαδοποιήσεις, συνήθως 2 x 2. Αντίθετα, το **καθολικό pooling**(global pooling) επιδρά σε

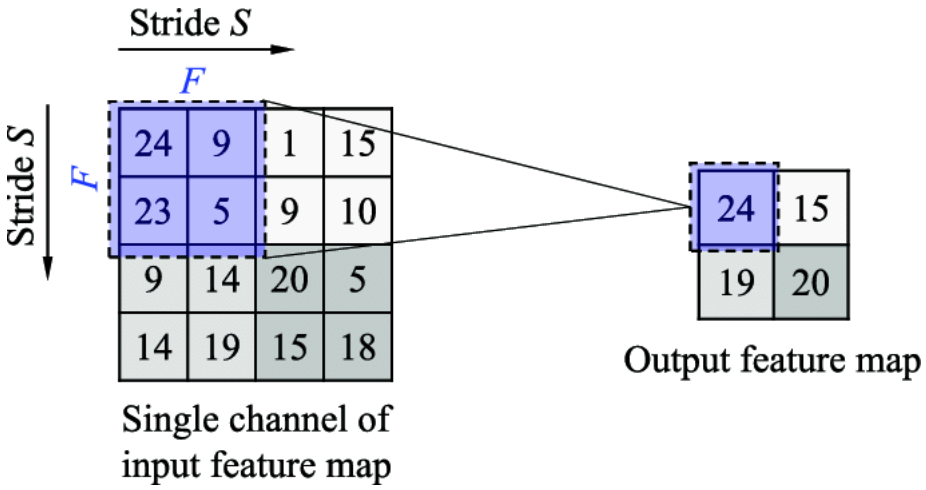

Figure. 4.6: Γραφική αναπαράσταση max pooling διασκελισμό(stride) $S = 2$

όλους τους νευρώνες του convolutional layer. Η τεχνική του pooling, μπορεί να υπολογίζει τη μέγιστη τιμή(max pooling) ή τον μέσο όρο(average pooling) της κάθε ομάδας από νευρώνες.

4.3 Αναδρομικά Νευρωνικά Δίκτυα

Τα **Αναδρομικά Νευρωνικά Δίκτυα**(Recurrent Neural Networks, RNN) είναι ένας τύπος ANN στα οποία οι συνδέσεις τους, δημιουργούν έναν κατευθυνόμενο γράφο, κατά μήκος μιας χρονοσειράς. Σε αντίθεση με τα MLP, τα RNN χρησιμοποιούν την εσωτερική τους κατάσταση(μνήμη) για να επεξεργαστούν ακολουθίες από εισόδους. Έτσι, χρησιμοποιούνται για επίλυση προβλημάτων όπως, η αναγνώριση φωνής, η αναγνώριση κειμένου, το beat tracking, όπου υπάρχει χρονική εξάρτηση μεταξύ των δεδομένων εισόδου.

4.3.1 Πως λειτουργεί ένα Αναδρομικό Νευρωνικό Δίκτυο

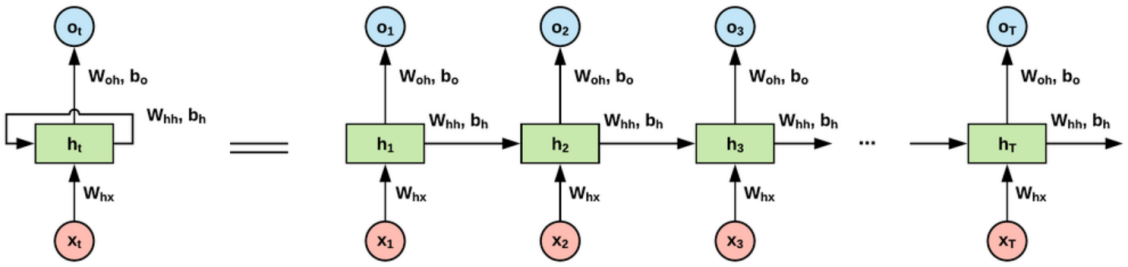

Figure. 4.7: Ξεδίπλωμα ενός RNN

Η παραπάνω εικόνα παρουσιάζει το ξεδίπλωμα ενός RNN, δηλαδή, την αναπαράσταση της

κατάστασής του για κάθε βήμα της ακολουθίας που παρέχεται ως είσοδος. Εδώ φαίνεται η ιδιότητα της μνήμης ενός RNN. Σε κάθε βήμα(time step), t, πραγματοποιούνται οι ακόλουθοι υπολογισμοί:

Για την κρυφή κατάσταση(hidden state) του RNN, h_t, ισχύει:

$$h_t = f(W_{hh}h_{t-1} + b_h + W_{hx}x_t)$$

όπου:

- x_t είναι η είσοδος για το time step t. Η είσοδος αυτή μπορεί να είναι και διάνυσμα εκτός από αριθμός.
- h_t είναι η κρυφή κατάσταση για το time step t. Λειτουργεί ως η ”μνήμη” του δικτύου και υπολογίζεται με βάση την προηγούμενη κρυφή κατάσταση, h_{t-1}, και την είσοδο του συγκεκριμένου βήματος.
- W_{hx} και W_{hh} είναι εκπαιδευσιμες παράμετροι του συστήματος.
- b_h είναι το bias που επιδρά πάνω στο προηγούμενο hidden state, h_{t-1}, για τον υπολογισμό του h_t.
- Η συναρτήσεις, f, που χρησιμοποιούνται συνήθως σαν activation function του hidden state είναι η $tanh$ και η $ReLU$.

Για την έξοδο, o_t:

$$o_t = g(W_{oh}h_t + b_o)$$

όπου:

- o_t είναι η έξοδος για το time step t. Η έξοδος αυτή μπορεί να είναι και διάνυσμα εκτός από αριθμός.
- W_{ox} είναι και αυτή μια εκπαιδεύσιμη παράμετρος του συστήματος, όπως οι W_{hx} και W_{hh}.
- b είναι το bias που επιδρά πάνω στο hidden state, h_t, για τον υπολογισμό του της εξόδου, o_t.
- Η συναρτήσεις, g, που χρησιμοποιούνται συνήθως σαν activation function, για τον υπολογισμό του o_t, είναι η $sigmoid$, η $softmax$, η $tanh$ και η $ReLU$.

Παρατηρούμε ότι, αντίθετα με τα παραδοσιακά ANNs, το δίκτυο χρησιμοποιεί τις ίδιες εκπαιδεύσιμες παραμέτρους για όλα τα time steps. Η παραπάνω εικόνα, παρουσιάζει ένα RNN με εξόδους σε κάθε time step. Αυτό δεν είναι πάντα απαραίτητο και εξαρτάται από τη διεργασία που εκτελεί το δίκτυο. Για παράδειγμα, όταν θέλουμε να προβλέψουμε το συναίσθημα μιας πρότασης, ίσως να ενδιαφερόμαστε μόνο για την τελευταία έξοδο δηλαδή το συναίσθημα αφού το σύστημα διάβασε ολόκληρη την πρόταση. Αντίστοιχα, είναι δυνατόν να μη χρησιμοποιήσουμε είσοδο σε κάθε time step.

4.3.2 Long Short-Term Memory

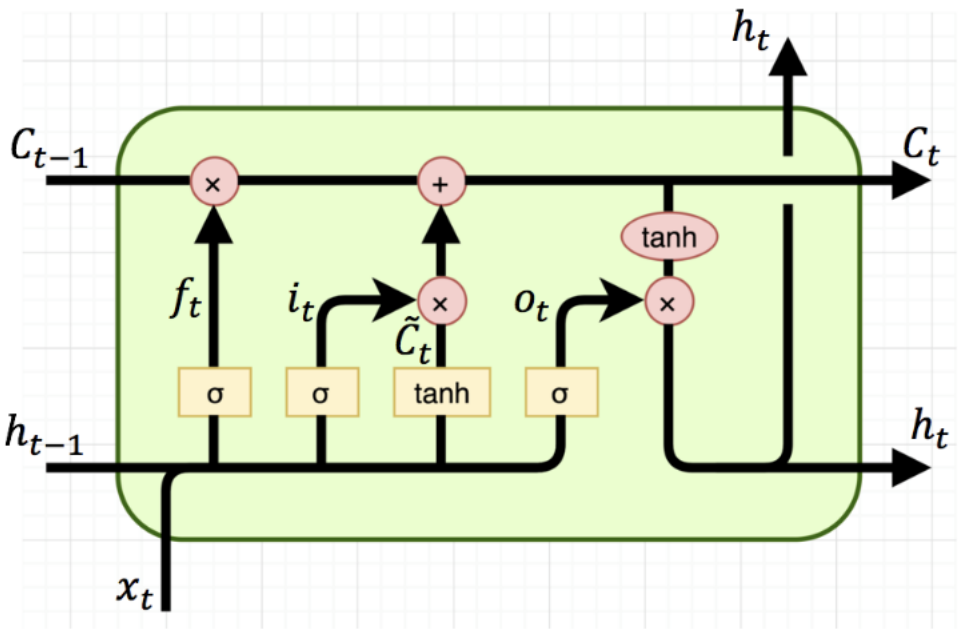

Figure. 4.8: Γραφική αναπαράσταση μιας μονάδας LSTM.

Στην προηγούμενη υποενότητα, περιγράψαμε τη λειτουργία των κλασσικών(ή vanilla) RNN. Θεωρητικά, αυτά τα δίκτυα μπορούν να ”θυμούνται” μεγάλες ακολουθίες από δεδομένα. Όμως, αυτό υπολογιστικά είναι αδύνατο. Κατά την εκπαίδευση του vanilla RNN, χρησιμοποιώντας τη μέθοδο back-propagation, τα διαφορικά, τα οποία διαδίδονται ”προς τα πίσω” στον χρόνο, μπορούν να μηδενιστούν ή να απειριστούν(Vanishing/Exploding Gradient Problem). Έτσι, καθίσταται αδύνατη η εκπαίδευση του δικτύου.

Μια λύση σε αυτό το πρόβλημα, δίνει μια πιο σύνθετη αρχιτεκτονική αναδρομικού δικτύου, το **Long Short-Term Memory**(LSTM). Αυτός ο τύπος δικτύου, μπορεί να ”συγκρατεί”

ακολουθίες εισόδου μεγάλου μήκους. Μια συνήθης αρχιτεκτονική των μονάδων του LSTM περιλαμβάνει: μια πύλη εισόδου(input gate), μια πύλη εξόδου(output gate) και μια πύλη forget(forget gate). Το input gate ελέγχει τη ροή ενός νέου στοιχείου, το forget gate ελέγχει αν θα παραμείνει το στοιχείο αυτό στη μονάδα του LSTM και το output gate ελέγχει τον βαθμό με τον οποίο, το στοιχείο που βρίσκεται στη μονάδα χρησιμοποιείται για να υπολογιστεί το αποτέλεσμα της συνάρτησης ενεργοποίησης της μονάδας. Τα activation functions του πυλών μιας μονάδας LSTM είναι sigmoid functions. Οι υπολογισμοί που πραγματοποιεί ένα LSTM unit κατά το εμπρόσθιο πέρασμα, είναι οι παρακάτω:

$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$

Οι μεταβλητές που χρησιμοποιούνται είναι:

- $x_t \in \mathbb{R}^d$: διάνυσμα εισόδου του LSTM unit
- $f_t \in \mathbb{R}^h$: διάνυσμα ενεργοποίησης του forget gate
- $i_t \in \mathbb{R}^h$: διάνυσμα ενεργοποίησης του input gate
- $o_t \in \mathbb{R}^h$: διάνυσμα ενεργοποίησης του output gate
- $h_t \in \mathbb{R}^h$: το διάνυσμα των κρυφών καταστάσεων, γνωστό και ως διάνυσμα εξόδου της μονάδας του LSTM
- $c_t \in \mathbb{R}^h$: cell state vector
- $W \in \mathbb{R}^{h \times d}, U \in \mathbb{R}^{h \times h}, b \in \mathbb{R}^h$: πίνακες με βάρη και διάνυσμα bias, εκπαιδευσιμες παράμετροι

και τα activation functions:

- σ_g: σιγμοειδής συνάρτηση
- σ_c: συνάρτηση υπερβολικής εφαπτομένης

- σ_h: συνάρτηση υπερβολικής εφαπτομένης ή ταυτοτική συνάρτηση

4.4 Εκπαίδευση Δικτύου

Κατά την εκπαίδευση, το μοντέλο προσαρμόζει επαναληπτικά τις παραμέτρους του, ώστε να προσεγγίσει τη συνάρτηση αντιστοίχησης που θέλουμε(με την προϋπόθεση ότι τα δεδομένα εκπαίδευσης είναι αξιόπιστα). Σε κάθε επανάληψη, το μοντέλο κάνει μια πρόβλεψη με βάση τα δεδομένα εισόδου που του δίνονται. Ο αλγόριθμος εκπαίδευσης μεταβάλλει τις παραμέτρους, $\mathbf{\Theta}$, ώστε να ελαχιστοποιήσει τη διαφορά μεταξύ της πρόβλεψης, $\hat{\mathbf{y}}$, και της επιθυμητής τιμής, $\mathbf{y}$. Η σχέση μεταξύ της διαφοράς αυτής με τις παραμέτρους του μοντέλου περιγράφεται από τη **συνάρτηση κόστους**(loss function). Κατά την εκπαιδευτική διαδικασία, προσπαθούμε να προσεγγίσουμε το ολικό ελάχιστον της συνάρτησης κόστους, ψάχνοντας επαναληπτικά για το σημείο όπου το διαφορικό της ως προς τις εκπαιδεύσιμες παραμέτρους είναι 0.

$$\frac{\mathrm{d}(J(\mathbf{\Theta}))}{\mathrm{d}\mathbf{\Theta}} = 0$$

Ένας από τους πιο διαδεδομένους επαναληπτικούς αλγορίθμους, που χρησιμοποιείται για την ελαχιστοποίηση του loss function, είναι ο **Stochastic Gradient Descent**(SGD). Σε κάθε επανάληψη, ο SGD υπολογίζει το σφάλμα για ένα συγκεκριμένο στοιχείο του training set και προσαρμόζει τις εκπαιδεύσιμες παραμέτρους με βάση τον παρακάτω τύπο:

$$\mathbf{\Theta}_{i+1} = \mathbf{\Theta}_i - \alpha \frac{\mathrm{d}(J(\mathbf{\Theta}_i))}{\mathrm{d}\mathbf{\Theta}}$$

όπου α ο **ρυθμός μάθησης**(learning rate). Το learning rate καθορίζει τον βαθμό μεταβολής των παραμέτρων για την επόμενη επανάληψη του αλγορίθμου. Η επιλογή του βέλτιστου learning rate είναι σημαντική για τη σωστή λειτουργία του αλγορίθμου. Επιλογή, πολύ μικρού learning rate καθιστά τον αλγόριθμο πολύ αργό και απαιτούνται πολλές επαναλήψεις ώστε να προσεγγίσουμε το ολικό ελάχιστο. Πολύ μεγάλο learning rate οδηγεί σε απρόβλεπτες ”συμπεριφορές”. Η πιο συνήθεις από αυτές είναι η αποτυχία σύγκλησης προς το ολικό ελάχιστο(overshooting). Για καλύτερη πρόβλεψη του loss function, μπορούμε να υπολογίσουμε την κλίση χρησιμοποιώντας ένα υποσύνολο από τα δεδομένα του training set και όχι μόνο ένα στοιχείο. Αυτή η παραλλαγή του SGD ονομάζεται **Mini-batch Gradient Descent**. Ο τύπος που χρησιμοποιείται για τη μεταβολή των παραμέτρων είναι ο ίδιος, όμως ο

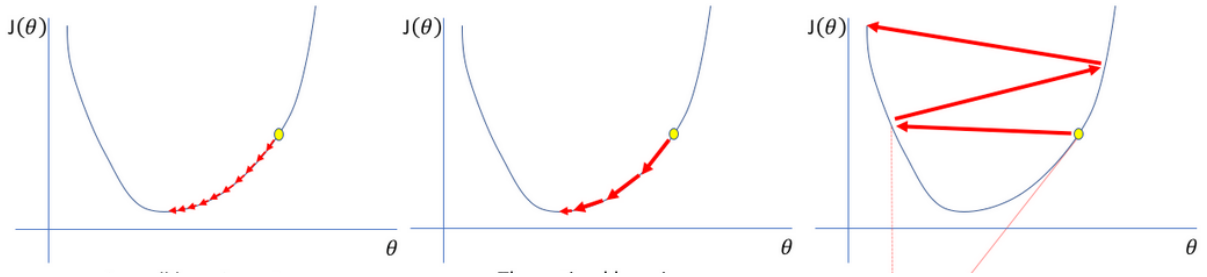

Figure. 4.9: Γραφική αναπαράσταση της σύγκλησης προς το τοπικό ελάχιστο ανάλογα με το learning rate που έχει επιλεχθεί. Αριστερά, παρατηρούμε ότι ο αλγόριθμος αργεί να συγκλίνει προς το ολικό ελάχιστο(πολύ μικρό learning rate). Δεξιά παρατηρείται το φαινόμενο του overshooting(πολύ μεγάλο learning rate). Τέλος, στο μεσαίο σχήμα, έχει επιλεχθεί το σωστό learning rate.

υπολογισμός του loss function περιλαμβάνει όλα τα στοιχεία του υποσυνόλου.

Η μεταβολή των παραμέτρων εξαρτάται από το σφάλμα στην έξοδο. Αυτό δεν μπορεί να χρησιμοποιηθεί ευθέως στην περίπτωση των δικτύων βαθιάς μηχανικής μάθησης, αφού η είσοδος και η έξοδος των νευρώνων στα κρυφά επίπεδα είναι άγνωστες. Έτσι, χρησιμοποιείται συνήθως μια διαδικασία που ονομάζεται **Back Propagation**(BK). Όπως έχει αναφερθεί, οι έξοδοι των προηγούμενων επιπέδων χρησιμοποιούνται σαν είσοδοι στο επόμενο επίπεδο και η κρυφή έξοδος γίνεται γνωστή με την οπίσθια διάδοση του σφάλματος.

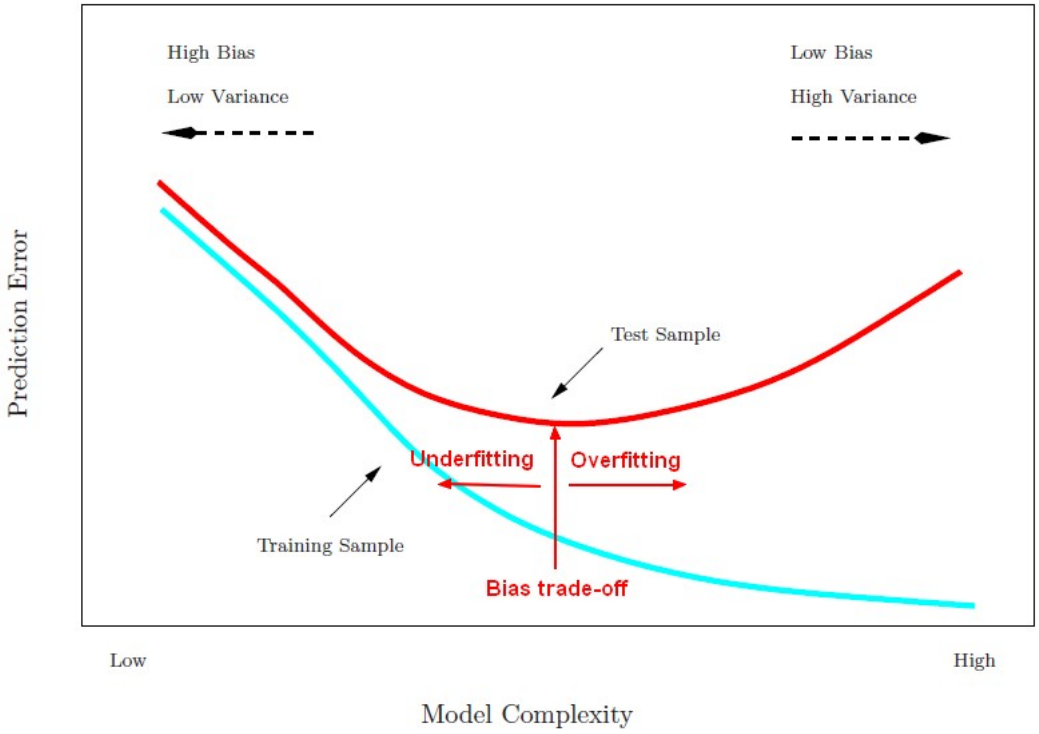

Figure. 4.10: Γραφική αναπαράσταση των εννοιών underfitting και overfitting

Ένα μοντέλο μηχανικής μάθησης είναι χρήσιμο όταν μπορεί να γενικεύσει ένα πρόβλημα, δηλαδή, να λειτουργεί με επιτυχία για άγνωστα σε αυτό δεδομένα. Κατά την εκπαίδευση,

διαχωρίζουμε τα διαθέσιμα δεδομένα σε δύο ομάδες: το training set και το validation set. Για να ελέγξουμε τη δυνατότητα του μοντέλου να γενικεύει, υπολογίζουμε το loss function για τα δυο αυτά set ξεχωριστά, εκπαιδεύοντας το σύστημά μας μόνο με τα δεδομένα του training set. Το μοντέλο μας χάνει τη δυνατότητα να γενικεύει, όταν validation loss αρχίζει να αυξάνεται(overfitting). Για να αποφύγουμε το overfitting, χρησιμοποιείται μια τεχνική που ονομάζεται **early stopping**. Με αυτή την τεχνική, η εκπαίδευση του μοντέλου σταματάει όταν παρατηρηθεί αύξηση του loss function για το validation set.

5 State of the Art

Σε αυτή την ενότητα παρουσιάζονται οι μέθοδοι που έχουν χρησιμοποιηθεί μέχρι σήμερα στο πεδίο του Beat Tracking, πέρα από το Deep Learning. Το beat, όπως αναφέρθηκε, γίνεται αντιληπτό μέσω μιας σειράς γεγονότων μέσα σε ένα μουσικό κομμάτι. Αυτά τα γεγονότα είναι με τέτοιο τρόπο τοποθετημένα στον χρόνο, ώστε ο άνθρωπος να αντιλαμβάνεται και να προβλέπει τη χρονική στιγμή του επόμενου beat. Ένα σύστημα για beat tracking περιλαμβάνει τρία κύρια μέρη, το καθένα από τα οποία μιμείται ένα συγκεκριμένο κομμάτι της διαδικασίας που εκτελεί ο άνθρωπος για την πρόβλεψη των beats.

Το πρώτο μέρος δέχεται σαν είσοδο ένα ηχητικό σήμα σαν είσοδο και έχει ως στόχο να περιγράψει τη θέση των μουσικών γεγονότων στον χρόνο, χρησιμοποιώντας τεχνικές επεξεργασίας σήματος. Με βάση τον τύπο των γεγονότων που προσπαθούμε να εξετάσουμε, χρησιμοποιούνται διαφορετικές τεχνικές.

Κάθε μουσικό γεγονός αρχίζει με ένα onset. Η συνάρτηση που περιγράφει τη θέση των onsets ονομάζεται **Onset Detection Function**(ODF). Η ODF είναι μια συνάρτηση με ρυθμό δειγματοληψίας χαμηλότερο από αυτόν του αρχείου μουσικής, που παίρνει μεγάλες τιμές κοντά στις θέσεις των onsets. Σε ένα σύστημα για Beat Tracking, το πρώτο μέρος του συστήματος καλείται να δημιουργήσει τη συνάρτηση ODF.

Όταν οι θέσεις των μουσικών onsets έχουν εξακριβωθεί, ο άνθρωπος προσπαθεί να διακρίνει μια επαναληψιμότητα στη σειρά γεγονότων. Αυτή τη διαδικασία καλείται να μιμηθεί το δεύτερο μέρος του συστήματος, η οποία αναλύει και προβλέπει τις περιοδικότητες στη συνάρτηση ODF. Η έξοδος της δεύτερης αυτής φάσης του συστήματος, είναι μία αρχική εκτίμηση της περιόδου των beat, δηλαδή της απόστασης μεταξύ δύο διαδοχικών beat.

Το τρίτο και τελευταίο μέρος εκμεταλλεύεται τις πληροφορίες που προέρχονται από το δεύτερο μέρος για τον υπολογισμό της πιθανότερης διάταξης των beat. Σκοπός του είναι να προβλέψει της θέσεις των beat, οι οποίες πρέπει να βρίσκονται όσο πιο κοντά γίνεται στις θέσεις των onsets, διατηρώντας, παράλληλα, την απόσταση μεταξύ των beats όσο πιο σταθερή γίνεται.

5.1 Δυναμικός Προγραμματισμός

Μια σκοπιά από την οποία μπορεί να προσεγγιστεί το Beat Tracking, είναι αυτή του **Δυναμικού Προγραμματισμού**(Dynamic Programming, DP). Ο Δυναμικός Προγραμματισμός είναι μια μέθοδος επίλυσης περίπλοκων προβλημάτων, υποδιαιρώντας τα σε ένα σύνολο μικρότερων προβλημάτων.

Ο Δυναμικός Προγραμματισμός χρησιμοποιείται ώστε να εξαχθούν οι θέσεις των beat, αναζητώντας την αλληλουχία που ταιριάζει με τις υψηλότερες τιμές της ODF, και το προβλεπόμενο tempo. Ένας άπληστος αλγόριθμος, που θα εξέταζε όλες τις πιθανές ακολουθίες της αλληλουχίας, θα είχε απαγορευτική υπολογιστική πολυπλοκότητα. Αντίθετα, η αξιοποίηση του δυναμικού προγραμματισμού μειώνει την πολυπλοκότητα, εξετάζοντας μικρότερα μέρη της ODF. Όταν ολόκληρη η αλληλουχία έχει αξιολογηθεί, οι θέσεις των beat υπολογίζονται ως τα τοπικά μέγιστα των υποακολουθιών.

5.2 Προσεγγίσεις από πολλούς συντελεστές

Ως συντελεστής ορίζεται ένα σύστημα που έχει προγραμματιστεί να εκτελέσει μια συγκεκριμένη διεργασία. Για τη διεργασία του Beat Tracking, οι αρχιτεκτονικές πολλών συντελεστών χρησιμοποιούνται ώστε να εξετάσουν το πρόβλημα παράλληλα. Σε περίπτωση που κάποιος συντελεστής, δεν ανταποκριθεί με επιτυχία στη διαδικασία, το σύστημα θα είναι σε θέση να προβλέψει τη σωστή αλληλουχία υπό την προϋπόθεση ότι κάποιος άλλος συντελεστής ”τα κατάφερε”.

Ένα μοντέλο πολλών συντελεστών που προτείνεται, χρησιμοποιεί πολλά εύρη ζώνης συχνοτήτων για να εξάγει τα onsets, τα οποία ανατίθενται σε ζευγάρια από συντελεστές, που προβλέπουν την αλληλουχία των beat με δεδομένο τη σειρά των onsets. Κάθε συντελεστής συνεργάζεται με το ζευγάρι του και υποχρεώνει ο ένας τον άλλο να εξερευνήσει διαφορετικές στρατηγικές ανίχνευσης. Επιπρόσθετα, κάθε συντελεστής κάνει αυτοαξιολόγηση, χρησιμοποιώντας μουσική γνώση με βάση το σήμα εισόδου. Σε περίπτωση που η υπόθεση είναι αξιόπιστη, ο συντελεστής προσαρμόζει τις παραμέτρους του, ώστε να διατηρήσει τη συγκεκριμένη υπόθεση. Στο τέλος, ο διαχειριστής των συντελεστών επιλέγει την αλληλουχία των beat από τον πιο αξιόπιστο συντελεστή, ως έξοδο του συστήματος.

5.3 Συνδυαστικές Μέθοδοι

Στις συνδυαστικές μεθόδους, αξιοποιούνται περισσότεροι από έναν αλγόριθμοι για Beat Tracking. Ένας δείκτης αμοιβαίας συμφωνίας(Mutual Agreement, MA) χρησιμοποιείται για να καθορίσει τη συσχέτιση μεταξύ των τελικών ακολουθιών. Μια μεθοδολογία που συναντάμε στη βιβλιογραφία, είναι η σύγκριση της εξόδου πολλών διαφορετικών αλγορίθμων μετρώντας το MA των προβλέψεων τους. Ο υπολογισμός του MA γίνεται με τη σύγκριση των διαφόρων εξόδων. Όποιος αλγόριθμος ”συμφωνεί” με τους περισσότερους άλλους αλγορίθμους βαθμολογείται με το μεγαλύτερο MA και, τελικά, η πρόβλεψή του επιλέγεται ως έξοδος.

5.4 Προσέγγιση με Βαθιά Μηχανική Μάθηση

Τα τελευταία χρόνια, η **Βαθιά Μηχανική Μάθηση**(Deep Learning) χρησιμοποιείται για να βελτιώσει την απόδοση των συστημάτων Beat Tracking. Η μουσική είναι ανθρώπινο δημιούργημα και έτσι, η βαθιά μηχανική μάθηση χρησιμοποιείται λόγω της ικανότητάς της να μιμηθεί την ανθρώπινη αντίληψη. Το Deep Learning έχει αποδειχθεί αρκετά χρήσιμο στα πεδία της ανάλυσης της μουσικής δομής(music structural analysis) και της μουσικής κατηγοριοποίησης(music classification).

Σε αυτή την περίπτωση, δε χρειάζεται να υπολογιστεί η συνάρτηση ODF. Χρήσιμα χαρακτηριστικά εξάγονται από τη μουσική κυμματομορφή, με τη χρήση μεθόδων επεξεργασίας σήματος. Το μοντέλο βαθιάς μάθησης προσπαθεί να μιμηθεί την ανθρώπινη σκέψη. Έτσι, δε χρειάζεται η εξαγωγή ”τεχνητών” χαρακτηριστικών, όπως η ODF, αφού δεν είναι η πιο αξιόπιστη αναπαράσταση του σήματος. Σε αυτή την περίπτωση, τα, χρονοσυχνοτικά συνήθως, χαρακτηριστικά που εξάγονται από το ηχητικό σήμα, αντιπροσωπεύουν την ανθρώπινη ακοή, η οποία, σε συνδυασμό με τον ανθρώπινο εγκέφαλο οδηγεί σε πιο αξιόπιστα αποτελέσματα.

Η συνηθέστερη αρχιτεκτονική Νευρωνικού Δικτύου που χρησιμοποιείται για το Beat Tracking είναι τα RNN, ώστε να εξάγουν χρήσιμες πληροφορίες για την περιοδικότητα του σήματος εισόδου. Φασματικά χαρακτηριστικά εξάγονται από την κυμματομορφή και δίνονται σαν είσοδος στο δίκτυο και στη συνέχεια αυτό δίνει σαν έξοδο την πιθανότητα ύπαρξης beat. Στη συνέχεια, και αφού έχει γίνει η πρόβλεψη για ολόκληρο το μουσικό κομμάτι, δημιουργείται

η **Συνάρτηση Ενεργοποίησης των Beats**(Beat Activation Function, BAF). Για την εξαγωγή των τελικών θέσεων των beats, χρησιμοποιείται μια μέθοδος επιλογής κορυφών(peak picking method). Η επιλογή των κορυφών πραγματοποιείται, συνήθως, χρησιμοποιώντας τη **Συνάρτηση Αυτοσυσχέτισης**(Autocorrelation Function, ACF) της BAF.

Algorithm	Fscore
Dynamic Programming	0.500
Multiple Agents	–
Ensemble approach	0.666
Deep Learning	0.938

Table 5.1: Επιδόσεις παραπάνω αλγορίθμων

6 Δεδομένα

Σε αυτή την ενότητα παρουσιάζονται τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων καθώς και τις δυσκολίες που αντιμετωπίστηκαν κατά την αναζήτηση αξιόπιστων dataset.

Στα πλαίσια αυτής της εργασίας, χρησιμοποιήθηκαν τρία διαφορετικά datasets: το **Ballroom**, το **Haisworth** και το **SMC**. Συνολικά, χρησιμοποιήθηκαν 1124 αρχεία μουσικών κομματιών και 11 ώρες και 25 λεπτά μουσικής.

6.1 Datasets

6.1.1 Ballroom

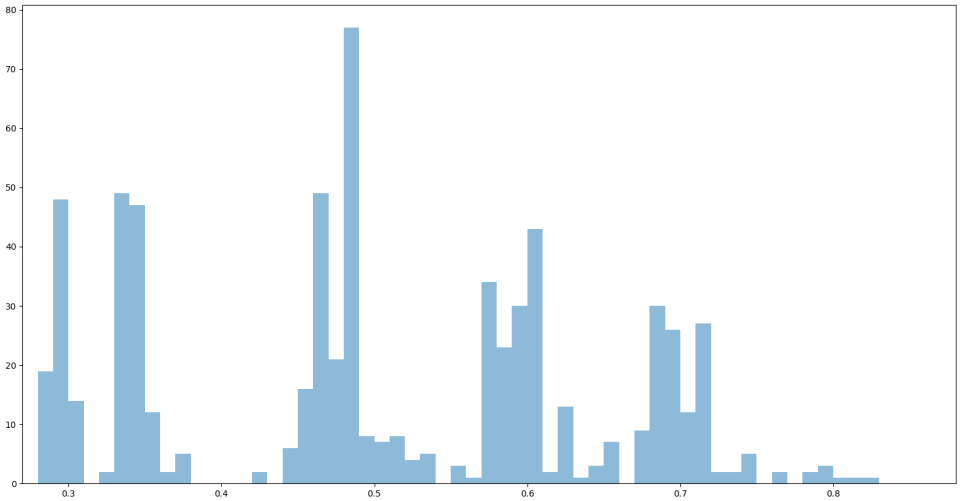

Figure. 6.1: Η κατανομή των intervals για το Ballroom Dataset

Το dataset Ballroom περιέχει 685 αρχεία μουσικής μέσου μεγέθους 30 δευτερολέπτων ενώ η συνολική του χρονική διάρκεια είναι 5 ώρες και 57 λεπτά. Περιλαμβάνει μουσική που κατηγοριοποιείται σε 8 είδη, με βάση το είδος χορού που αντιπροσωπεύει(ChaChaCha, Jive, Quickstep,...). Τα περισσότερα κομμάτια σε αυτό το dataset διαθέτουν κρουστά.

Συνολικά, το dataset αυτό περιλαμβάνει 43838 χρονικές στιγμές που διαθέτουν beat. Στο παραπάνω σχήμα φαίνεται η κατανομή της μέσης τιμής των intervals από το ένα beat στο επόμενο ανά μουσικό κομμάτι σύμφωνα με τα annotations του dataset. Παρατηρούμε ότι περιλαμβάνει μουσικά κομμάτια από αργό έως αρκετά γρήγορο tempo, ενώ, όπως φαίνεται στην παρακάτω εικόνα, παρατηρούνται πολύ μικρά tempo fluctuations. Χρησιμοποιείται ευρέως για την εκπαίδευση Δικτύων σε πεδία όλου του φάσματος του MIR.

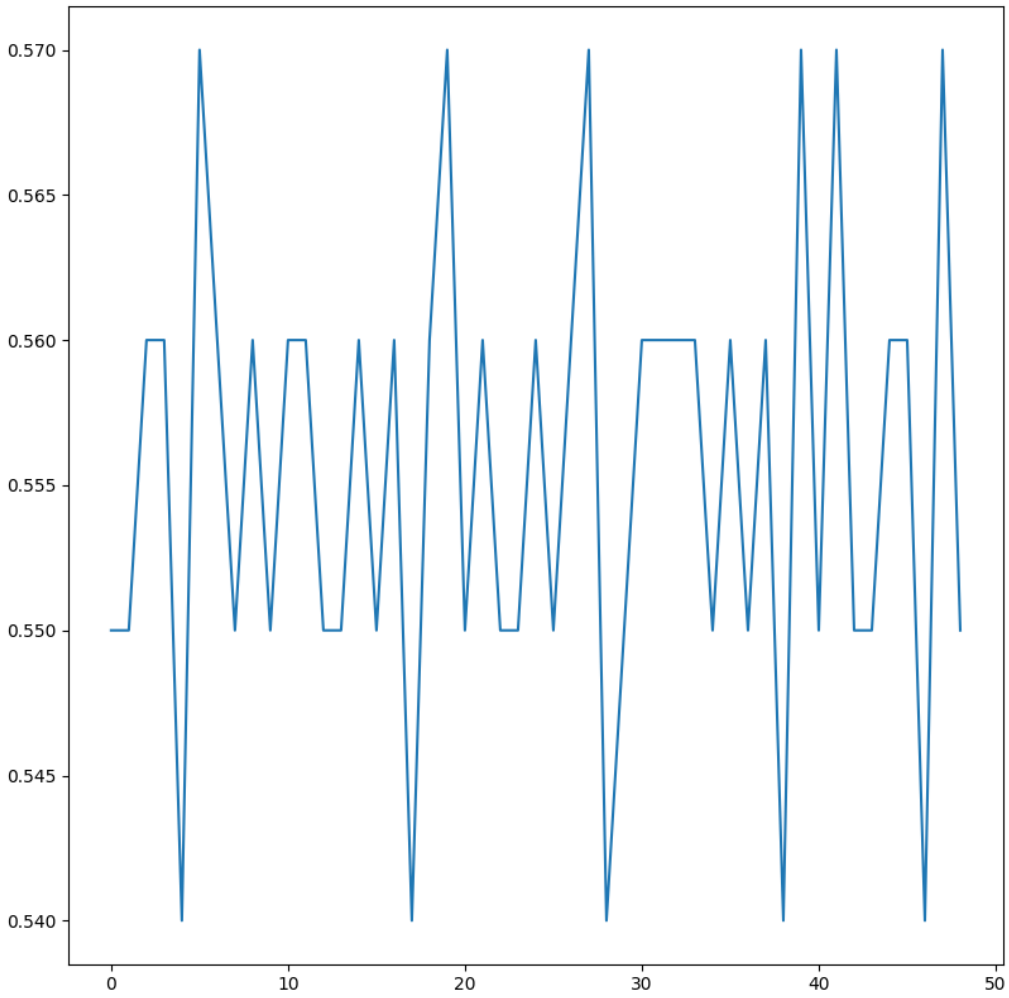

Figure. 6.2: Μεταβολή της ταχύτητας ενός τυχαίου μουσικού κομματιού από το Ballroom Dataset

6.1.2 Hainsworth

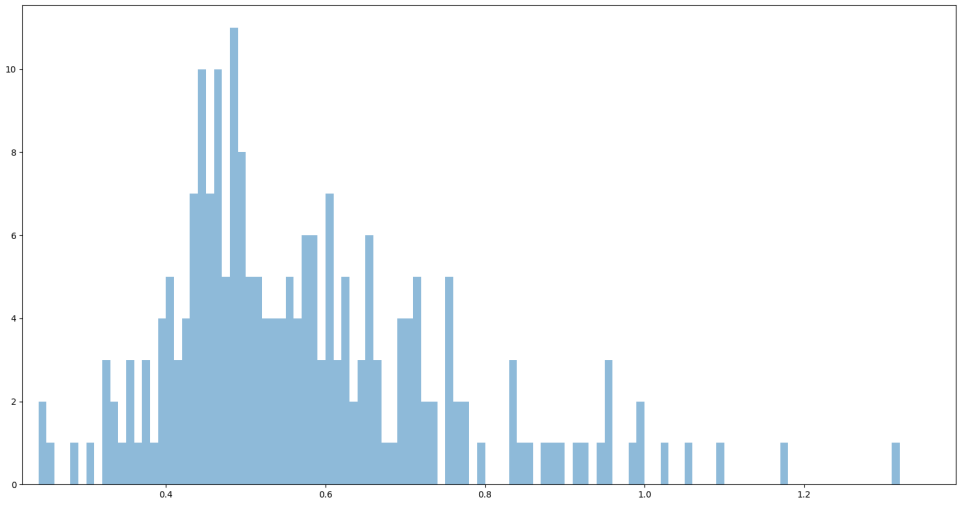

Figure. 6.3: Η κατανομή των intervals για το Hainsworth Dataset

Το dataset Hainsworth περιέχει 222 αρχεία μουσικής μέσου μεγέθους 60 δευτερολέπτων ενώ η συνολική του χρονική διάρκεια είναι 3 ώρες και 19 λεπτά. Περιλαμβάνει μουσική που μπορεί να κατηγοριοποιηθεί ως μουσική του δυτικού κόσμου. Τα περισσότερα κομμάτια σε αυτό το dataset διαθέτουν κρουστά.

Συνολικά, το dataset αυτό περιλαμβάνει 22540 χρονικές στιγμές που διαθέτουν beat. Στο παραπάνω σχήμα φαίνεται η κατανομή της μέσης τιμής των intervals από το ένα beat στο επόμενο ανά μουσικό κομμάτι σύμφωνα με τα annotations του dataset. Παρατηρούμε ότι το dataset αυτό περιλαμβάνει, ως επί το πλείστον, κομμάτια μέσης ταχύτητας, ενώ, όπως φαίνεται στην εικόνα 6.4, παρατηρούνται πολύ μικρά tempo fluctuations. Χρησιμοποιείται ευρέως για την εκπαίδευση Δικτύων σε πεδία όλου του φάσματος του MIR.

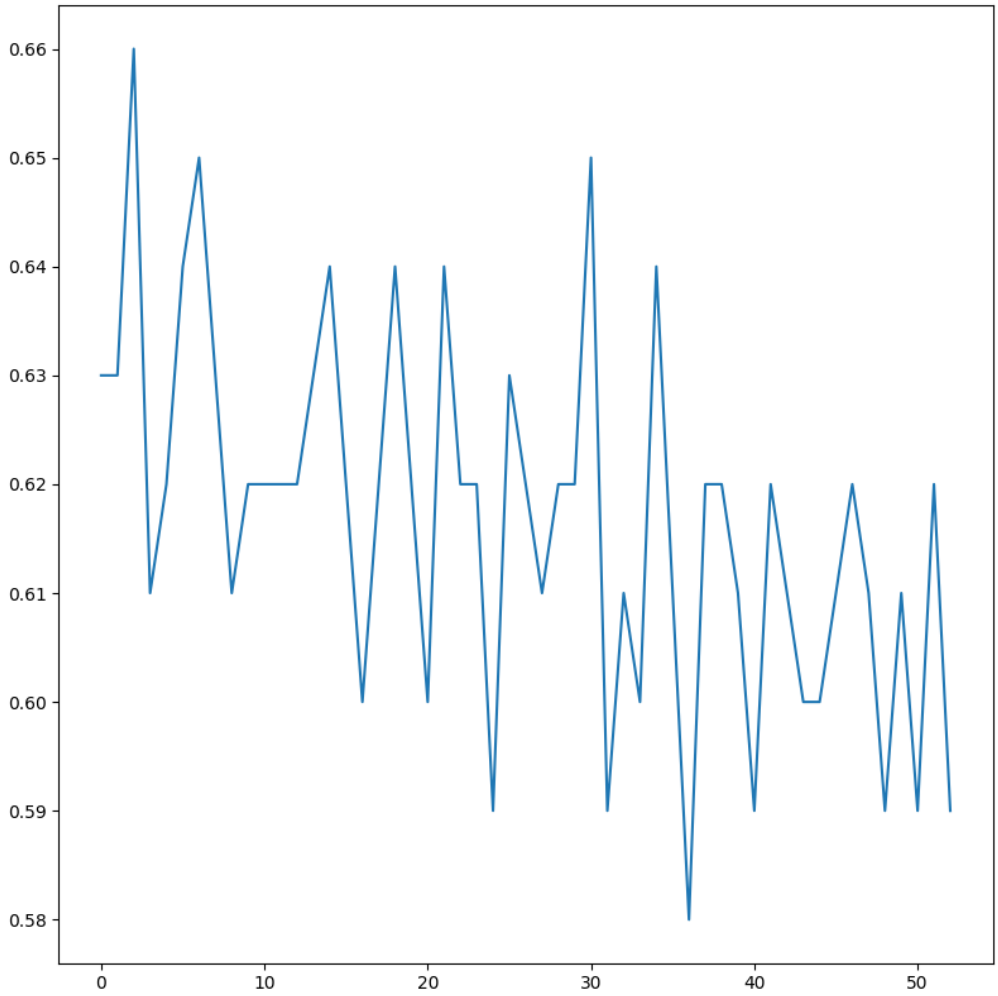

Figure. 6.4: Μεταβολή της ταχύτητας ενός τυχαίου μουσικού κομματιού από το Hainsworth Dataset

6.1.3 SMC

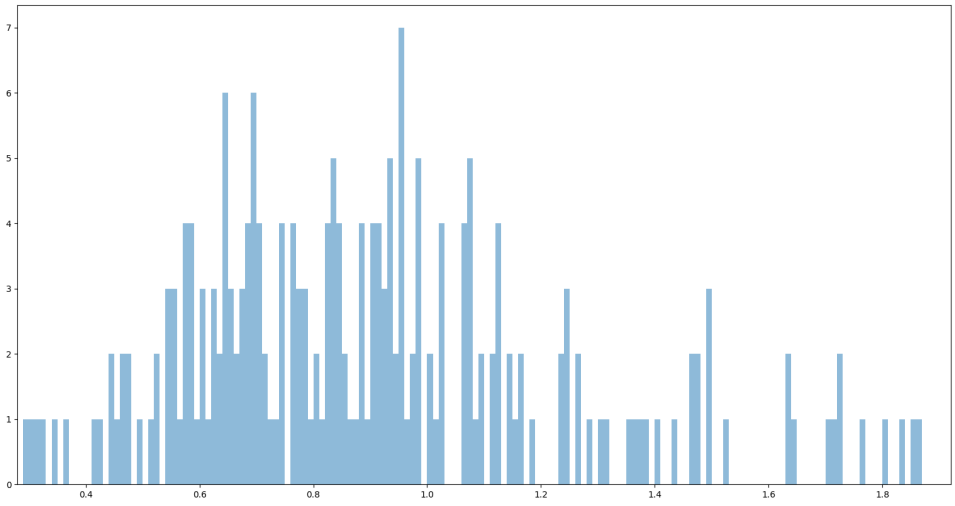

Figure. 6.5: Η κατανομή των intervals για το SMC Dataset

Το dataset SMC περιέχει 217 αρχεία μουσικής μεγέθους 40 δευτερολέπτων ενώ η συνολική του χρονική διάρκεια είναι 2 ώρες και 25 λεπτά. Περιλαμβάνει μουσικά κομμάτια η ταχύτητα των οποίων μεταβάλλεται σε όλη τη διάρκεια τους. Η μουσική που περιλαμβάνει βασίζεται περισσότερο στην καλή ερμηνεία παρά στο ακριβές παίξιμο των αξιών. Τα περισσότερα κομμάτια σε αυτό το dataset δε διαθέτουν κρουστά.

Συνολικά, το dataset αυτό περιλαμβάνει 1070 χρονικές στιγμές που διαθέτουν beat και είναι το μικρότερο σε μέγεθος από τα τρία datasets που χρησιμοποιήθηκαν. Στο παραπάνω σχήμα φαίνεται η κατανομή της μέσης τιμής των intervals από το ένα beat στο επόμενο ανά μουσικό κομμάτι σύμφωνα με τα annotations του dataset. Παρατηρούμε ότι το dataset αυτό περιλαμβάνει, κομμάτια σε ολόκληρο το φάσμα των ταχυτήτων ενώ, όπως φαίνεται στην εικόνα 6.6, τα tempo fluctuations είναι αρκετά μεγάλα. Χρησιμοποιείται ευρέως στο evaluation phase, για Δίκτυα που αφορούν το Beat Tracking.

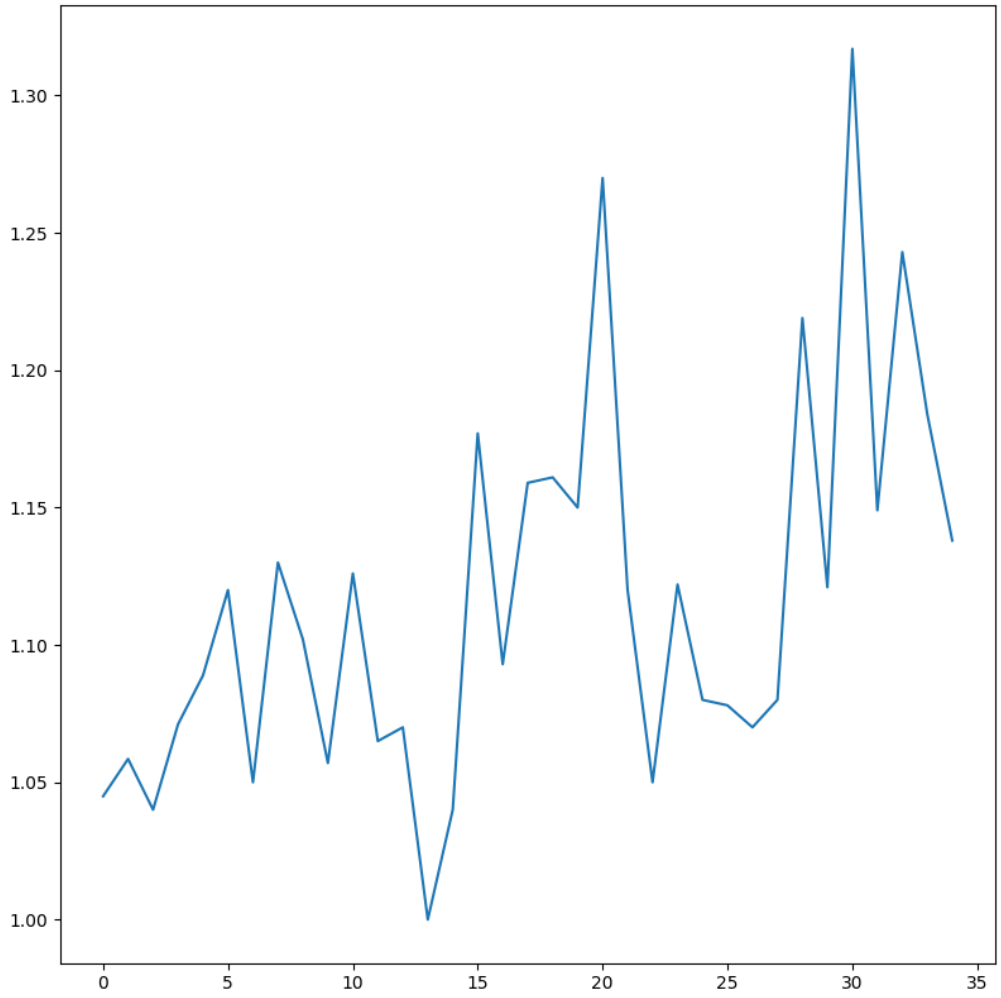

Figure. 6.6: Μεταβολή της ταχύτητας ενός τυχαίου μουσικού κομματιού από το SMC Dataset

6.2 Δυσκολία Εύρεσης Αξιόπιστων Dataset

Η αναζήτηση αξιόπιστων dataset για Beat Tracking είναι μια δύσκολη διαδικασία. Τα περισσότερα datasets περιλαμβάνουν μόνο annotations και όχι αρχεία μουσικής, λόγω πνευματικών δικαιωμάτων. Τα αρχεία μουσικής που χρησιμοποιήθηκαν κατά την κατασκευή του dataset αναφέρονται, όμως, ο εντοπισμός των συγκεκριμένων αρχείων είναι δύσκολος έως αδύνατος.

Στα αρχικά στάδια της εργασίας, η αποτυχία εύρεσης dataset οδήγησε στην προσπάθεια

δημιουργίας ενός νέου dataset. Τα annotations ήταν αρκετά ανακριβή, ενώ τα δεδομένα μας χαρακτηρίζονταν από μικρή διασπορά(variance), αφού χρησιμοποιήθηκαν κομμάτια από παρόμοια είδη μουσικής.

Μια ακόμα δυσκολία που αντιμετωπίζεται, είναι ότι τα annotations στα dataset, που, τελικά, βρήκαμε, έχουν δημιουργηθεί από διαφορετικούς ανθρώπους, με διαφορετική αντίληψη για τη θέση των beats. Όπως αναφέρθηκε παραπάνω, ο τρόπος που αντιλαμβανόμαστε τη μουσική διαφέρει από άνθρωπο σε άνθρωπο. Έτσι, είναι αναμενόμενη η διαφορά αυτή στα anntations.

7 Μεθολογία

Σε αυτή την ενότητα παρουσιάζεται η μεθοδολογία που ακολουθήθηκε, στο πλαίσιο αυτής της εργασίας, για την επίλυση του προβλήματος του Beat Tracking. Για τον προγραμματισμό χρησιμοποιήσαμε τη γλώσσα Python η οποία διαθέτει αρκετά χρήσιμα εργαλεία για την εξαγωγή χρήσιμων δεδομένων από αρχεία μουσικής, εκπαίδευση Νευρωνικών Δικτύων και αριθμητικές πράξεις πινάκων και διανυσμάτων.

Για την επεξεργασία του μουσικού σήματος και την εξαγωγή χρήσιμων χαρακτηριστικών της ηχητικής κυμματομορφής χρησιμοποιήθηκε τη βιβλιοθήκη **librosa**. Για τη δημιουργία και την εκπαίδευση των Νευρωνικών Δικτύων, χρησιμοποιήθηκε η βιβλιοθήκη **keras**, η οποία έχει έτοιμες συναρτήσεις των βασικών επιπέδων Νευρωνικών Δικτύων προγραμματισμένες σε **tensorflow**. Για τη διευκόλυνση των υπολογισμών και για μεγαλύτερη ταχύτητα στην εκτέλεση των προγραμμάτων, χρησιμοποιήθηκε η βιβλιοθήκη **numpy**. Τέλος, για τον σχεδιασμό γραφικών παραστάσεων, χρησιμοποιήθηκε η **matplotlib**.

7.1 Μεθοδολογία

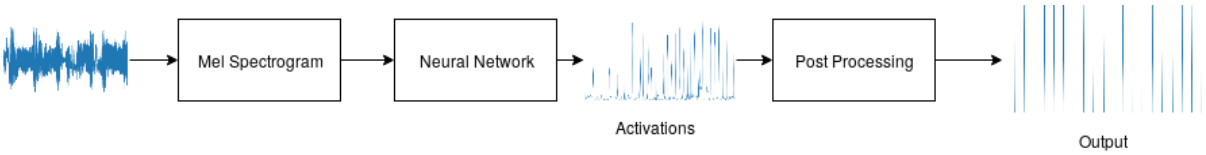

Figure. 7.1: Γραφική αναπαράσταση των βημάτων που πραγματοποιήθηκαν σε αυτή την εργασία.

7.1.1 Ηχητική Κυμματομορφή

Το πρώτο στάδιο περιλαμβάνει την εξαγωγή της ηχητικής κυμματομορφής από το αρχείο μουσικής.

Αρχικά, εξάγεται η στέρεο ηχητική κυμμτομορφή από το αρχείο μουσικής, με ρυθμό δειγματοληψίας $44.1kHz$. Στη συνέχεια, μετατρέπεται σε μονοφωνική παίρνοντας τον μέσο όρο των δύο κυμματομορφών.

7.1.2 Σπεκτρογράφημα Mel

Στο δεύτερο στάδιο, εξάγονται τα συχνοτικά χαρακτηριστικά του σήματος σε μορφή σπεκτρογραφήματος.

Ο υπολογισμός γίνεται χρησιμοποιώντας τον **Βραχυπρόθεσμο Μετασχηματισμό Fourier**(Short-Time Fourier Transform, STFT) που αναπαριστά τη συχνοτική εξέλιξη το σήματος στον χρόνο. Στα περισσότερα πειράματα(Πειράματα 1-6), για τον STFT χρησιμοποιήθηκαν οι παρακάτω παράμετροι:

- Αριθμό αλμάτων σε δείγματα(hop length), $h = 441$. Όπως αναφέρθηκε πριν, το ηχητικό σήμα δειγματοληπτείται με ρυθμό $44.1kHz$, δηλαδή, σε κάθε δευτερόλεπτο υπάρχουν 44100 δείγματα. Για να επιτευχθεί αριθμός frames ανά δευτερόλεπτο, $frames = 100$, δηλαδή, κάθε frame να αντιστοιχεί σε χρονική διάρκεια $10ms$ χρησιμοποιείται αυτό το hop length.
- Μήκος παραθύρου, $W = 2048$. Σύμφωνα με τη βιβλιογραφία, το μήκος αυτό είναι το πιο αξιόπιστο για τον διαχωρισμό harmonic και percussive events, όταν στο σύστημα δίνεται ως είσοδος μόνο ένα spectrogram.

Στη συνέχεια, το αποτέλεσμα του STFT περνάει από ένα **Mel filter-bank**. Στα περισσότερα πειράματα(Πείραμα 1-6), ο αριθμός φίλτρων(bins) που χρησιμοποιείται είναι, $bins = 128$. Έτσι, τελικά, προκύπτει το τελικό σπεκτρογράφημα που θα χρησιμοποιηθεί σαν είσοδος του Νευρωνικού Δικτύου.

7.1.3 Νευρωνικό Δίκτυο

Στο πλαίσιο αυτής της διατριβής, το πρόβλημα του Beat Tracking αντιμετωπίζεται ως ένα **πρόβλημα κατηγοριοποίησης**(classification problem). Σε κάθε πείραμα, δημιουργείται ένας classifier.

Σε όλα τα πειράματα, ως είσοδος στο Νευρωνικό Δίκτυο δίνεται ένα κομμάτι σπεκτρογραφήματος, μήκους 11 frames(συνολική διάρκεια $110ms$). Το Νευρωνικό Δίκτυο καλείται να ”απαντήσει” αν το μεσαίο frame της εισόδου αντιστοιχεί σε beat ή όχι. Δίνει σαν έξοδο, δηλαδή, την πιθανότητα το συγκεκριμένο frame να έχει beat.

Οι θέσεις των frames που αντιστοιχούν σε ”beat”, παίρνουν την τιμή 1, ενώ οι θέσεις που αντιστοιχούν σε ”no beat”, την τιμή 0.

7.1.4 Post Processing

Αφού το δίκτυο κατηγοριοποιήσει όλα τα frames του μουσικού κομματιού, δημιουργείται το **Beat Activation Function**(BAF), παρατάσσοντας όλες τις κατηγοριοποιήσεις στον άξονα του χρόνου.

Ένα παράθυρο 3 δευτερολέπτων(300 frames) χρησιμοποιείται για τους παρακάτω υπολογισμούς.

7.1.4.1 Συνάρτηση Αυτοσυσχέτισης

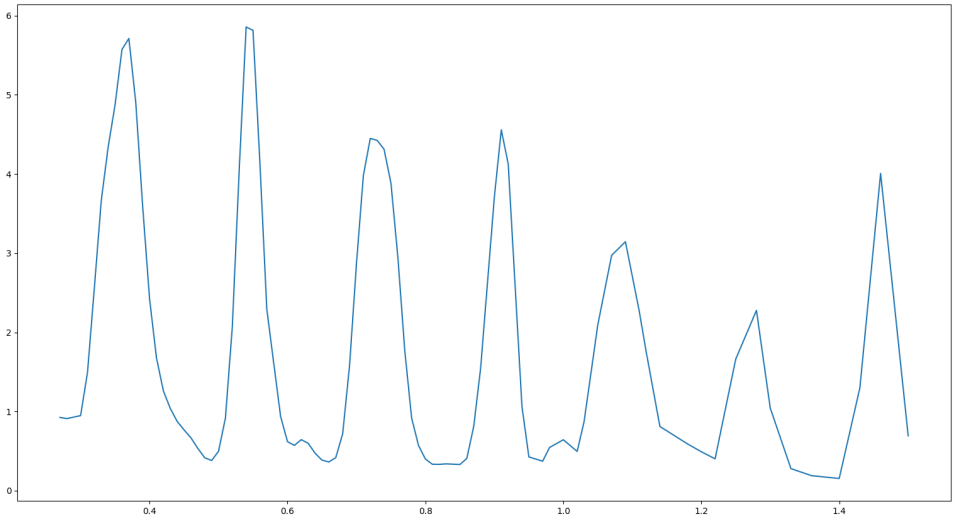

Figure. 7.2: Συνάρτηση Αυτοσυσχέτισης για ένα μουσικό κομμάτι 164 bpm. Παρατηρείται ότι η μέγιστη τιμή της ACF δε βρίσκεται στη σωστή θέση, δηλαδή στην τιμή της ομάδας interval που ανήκουν τα 164 bpm($\tau = 0.37\ sec$).

Υπολογίζεται η **Συνάρτηση Αυτοσυσχέτισης**(Autocorrelation Function, ACF), $A(\tau)$, της BAF, $a(n)$, ώστε να εκτιμηθεί ο κύριος παλμός του κομματιού(dominant beat interval), στο συγκεκριμένο παράθυρο. Ο τύπος του autocorrelation function είναι ο παρακάτω:

$$A(\tau) = \sum_{n=0}^{299} a(n+\tau) \cdot a(n)$$

Όπου τ, το beat interval. Στο πλαίσιο αυτής της εργασίας, εξετάζονται intervals από $0.24\ sec$(το interval που αντιστοιχεί στα $250\ bpm$) έως $1.5\ sec$(το interval που αντιστοιχεί στα $40\ bpm$).

Το μουσικό tempo ορίζεται beats per minute(bpm). Κάθε βαθμίδα bpm διαχωρίζεται από τις γειτονικές του με ένα interval ακρίβειας $1\ ms$. Στην BAF, όμως, τα frames διαχωρίζονται με ακρίβεια $10\ ms$. Έτσι, σε αυτό το σημείο, δεν είναι δυνατό να γίνει πλήρης διαχωρισμός των bpm. Στην επόμενη φάση αυτό το πρόβλημα αντιμετωπίζεται.

Στη μουσική παρατηρούνται, συχνά, μικρά tempo fluctuations, ακόμα και σε μουσικά κομμάτια που είναι ηχογραφημένα με τη βοήθεια μετρονόμου. Έτσι, ορισμένα beats συμβαίνουν λίγο πριν ή λίγο μετά από την πραγματική τους θέση. Για τον λόγο αυτό, χρησιμοποιείται ένα **παράθυρο Hamming**(Hamming window), $s(\tau)$ για την εξομάλυνση της ACF. Το μήκος του παραθύρου δεν έχει μεγάλη σημασία. Πρέπει, όμως να είναι αρκετά μεγάλο ώστε να καλύπτει τα πιθανά μικρά tempo fluctuations και μικρότερο από το ελάχιστο interval που εξετάζουμε, τ_{min}. Στα πλαίσια της εργασίας αυτής, το ελάχιστο interval είναι $\tau_{min} = 240\ msec$. Το παράθυρο που χρησιμοποιείται έχει μήκος, $\tau_l = 210\ msec$.

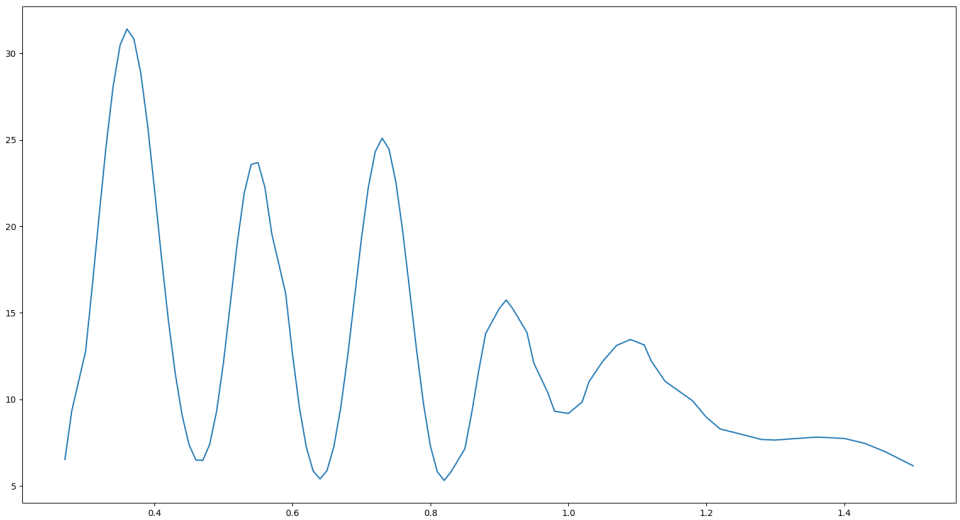

Figure. 7.3: Εξομαλισμένη Συνάρτηση Αυτοσυσχέτισης για το ίδιο μουσικό κομμάτι $164\ bpm$, της εικόνας 7.2. Η μέγιστη τιμή βρίσκεται, πλέον στη σωστή θέση, $\tau = 0.37\ sec$.

Η θέση της μέγιστης τιμής της εξομαλισμένης ACF δίνει την ομάδα από bpm που κυριαρχεί σε αυτό το παράθυρο.

7.1.4.2 Υπολογισμός κύριου παλμού παραθύρου

Τα bpm της ομάδας διαχωρίζονται με ακρίβεια τριών δεκαδικών ψηφίων σε intervals. Για κάθε interval, i, υπολογίζεται το μέγιστο άθροισμα της BAF, με βάση τον τύπο:

$$p^* = position(\max_{p=1,\dots,i} \sum_k a(p + k \cdot i))$$

Ο αριθμός $p*$, αντιστοιχεί στη θέση που ξεκινάει το μέγιστο άθροισμα για κάθε i, δηλαδή τη θέση του πρώτου beat του παραθύρου, σε περίπτωση που το κύριο interval ήταν το i. Η θέση με το μέγιστο άθροισμα, αντιστοιχεί στον κύριο παλμό του παραθύρου.

Έχοντας βρει τη θέση του επόμενου beat, μια νέα επανάληψη του αλγορίθμου ξεκινάει. Το νέο παράθυρο 3 δευτερολέπτων ξεκινάει από τη θέση του beat που μόλις βρέθηκε.

7.1.5 Εκπαίδευση

Δημιουργούνται τα spectrograms των μουσικών κομματιών. Σύμφωνα με το ground truth, συγκεντρώνονται όλες οι 11-άδες από frames που στο μεσαίο frame βρίσκεται beat και ονοματίζονται με το class label 1 που αντιστοιχεί στην απάντηση ”beat”. Όλες οι άλλες 11-άδες ονοματίζονται με το class label 0 που αντιστοιχεί σε ”no beat”.

Το 20% των 1 και το 20% των 0, αξιοποιούνται για τον σχηματισμό του **σετ εξακρίβωσης**(validation set), που χρησιμοποιείται για τον έλεγχο του overfitting. Τα υπόλοιπα instances αξιοποιούνται σαν **σετ εκπαίδευσης**(training set).

Πριν την εκπαίδευση, τα δεδομένα κανονικοποιούνται ώστε κάθε feature στο training set να έχει μέση τιμή 0 και τυπική απόκλιση 1 σύμφωνα με τον τύπο:

$$\textbf{x\_norm} = \frac{\textbf{x} - \textbf{mean}}{\textbf{std}}$$

Όπου **x_norm** και **x** το κανονικοποιημένο και το αρχικό διάνυσμα των features ενός frame, αντίστοιχα. **mean** και **std**, τα διανύσματα μέσων τιμών και των τυπικών αποκλίσεων, του training set.

Κανονικοποίση πραγματοποιείται και στο validation set, χρησιμοποιώντας τις τιμές του training

set.

Κατά την εκπαίδευση του δικτύου, χρησιμοποιείται η τεχνική **Mini-Batch Gradinet Descent** με $batch\_size = 32$. Με τη μέθοδο **Early Stopping** σε περίπτωση που παρατηρηθεί αύξηση της συνάρτησης κόστους(loss function) στο validation set, η εκπαίδευση του δικτύου σταματάει. Ο αλγόριθμος ”περιμένει” 10 εποχές για τυχόν βελτίωση του loss function.

7.1.5.1 Class Imbalance

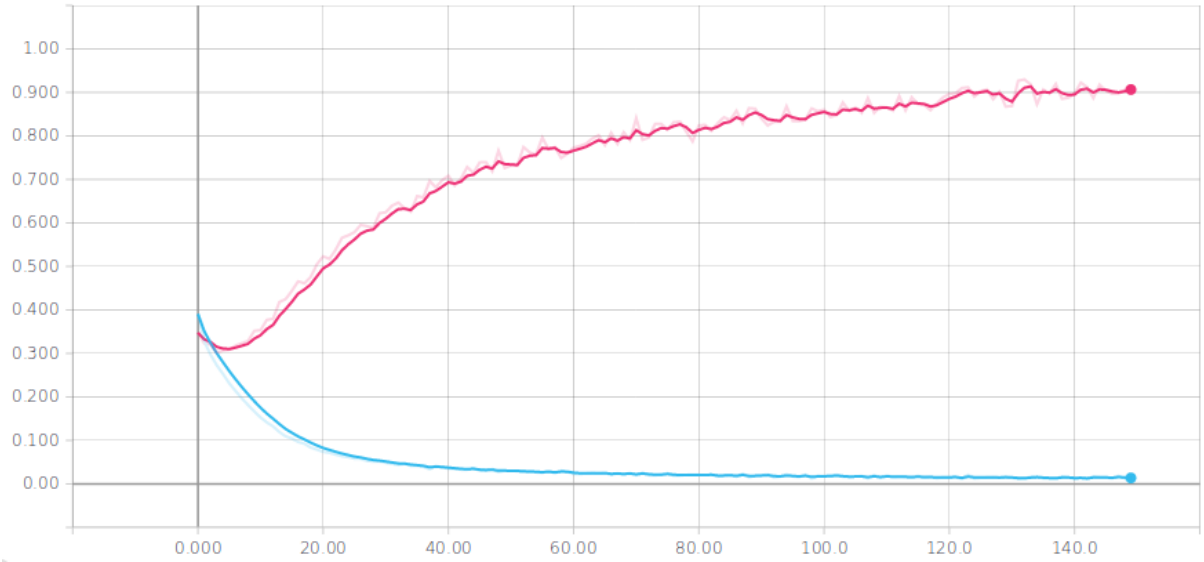

Figure. 7.4: Καμπύλες συνάρτησης κόστους για εκπαίδευση με ανεξέλεγκτο class imbalance και χωρίς εφαρμογή early stopping. Γαλάζιο, training set και Ροζ, validation set. Παρατηρείται ότι από την τέταρτη εποχή και έπειτα, το σφάλμα του validation set αυξάνεται.

Σε ένα κομμάτι $60\ bpm$, για κάθε ένα frame που περιέχει beat(1), υπάρχουν 99 frames που δεν περιέχουν beat(0). Παρατηρείται **ανισορροπία κλάσεων**(class imbalance) στα δεδομένα. Αυτή η ανισορροπία, δημιουργεί πρόβλημα στην εκπαίδευση του δικτύου, το οποίο γίνεται υπερβολικά προκατειλημμένο προς τη μηδενική κλάση.

Το πρόβλημα λύνεται με την υποδειγματοληψία(undersampling) της κλάσης 0 πριν τον διαχωρισμό των δεδομένων σε training και validation set. Το imbalance διατηρείται στα περισσότερα πειράματα, χρησιμοποιώντας $zero\_to\_one\_ratio = 15$(για κάθε θέση beat, υπάρχουν 15 θέσεις χωρίς beat). Στο πειραματικό στάδιο, εξερευνούνται τα αποτελέσματα που δίνουν διάφορα ποσοστά imbalance.

8 Πειραματικό Στάδιο

Σε αυτή την ενότητα, παρουσιάζονται τα πειράματα που πραγματοποιήθηκαν στο πλαίσιο αυτής της εργασίας. Εξετάζονται και συγκρίνονται ορισμένες αρχιτεκτονικές δικτύων που περιγράφονται στη βιβλιογραφία, με διαφορετική(απλούστερη), όμως, είσοδο, καθώς και ορισμένα άλλα δίκτυα.

8.1 Αξιολόγηση

Όπως αναφέρθηκε, τα annotations για beat tracking δεν είναι απόλυτα ακριβή. Το πρόβλημα αυτό γίνεται μεγαλύτερο, αν αναλογιστούμε ότι κάθε dataset έχει δημιουργηθεί από διαφορετικό άνθρωπο, με διαφορετική αντίληψη για τη θέση των beats. Έτσι, είναι θεμιτό να δοθεί ένα διάστημα απόκλισης από την annotated θέση του beat. Η το διάστημα αυτό είναι αρκετά μικρό ώστε να μην είναι αντιληπτό από ανθρώπινο αυτί.

Για την αξιολόγηση της απόδοσης κάθε αρχιτεκτονικής, χρησιμοποιήθηκαν οι παρακάτω μετρικές.

- Relevant instances: Τα instances του dataset στα οποία υπάρχει beat.
- True Positives: Ο αριθμός των Relevant Instances που βρήκε ο αλγόριθμος.
- False Positives: Ο αριθμός των Instances που ο αλγόριθμος κατηγοριοποίησε ως Positive αλλά είναι Negative.
- Precision: Το ποσοστό από τα Instances που επέλεξε ο αλγόριθμος και είναι Relevant

$$Precision = \frac{true\_positives}{selected\_items}$$

- Recall: Το ποσοστό από τα Relevant Instances που επέλεξε ο αλγόριθμος

$$Recall = \frac{true\_positives}{relevant\_items}$$

- F-score: Μετρική που συνδυάζει τις μετρικές Precision και Recall για να εξετάσει την ακρίβεια του αλγορίθμου. Είναι η βασική μετρική που χρησιμοποιήθηκε σε αυτή την

εργασία.

$$Fscore = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

8.2 Πειράματα

8.2.1 Πειράματα 1-4

Στα πειράματα αυτά, εξετάζονται οι δημοφιλέστερες αρχιτεκτονικές που χρησιμοποιούνται για Beat Tracking. Κατά την εκπαίδευση, τα βάρη του δικτύων αρχικοποιούνται με βάση μια ομοιόμορφη κατανομή με διάστημα $[-0.1, 0.1]$. Όλα τα επίπεδα των δικτύων αυτών είναι διπλής κατεύθυνσης και αποτελούνται από μονάδες LSTM. Η συνάρτηση ενεργοποίησης που χρησιμοποιείται για τα hidden layers είναι η $tanh$, ενώ, για το output layer, η $sigmoid$.

Για κάθε αρχιτεκτονική, εκπαιδεύονται 3 δίκτυα. Κάθε δίκτυο εκπαιδεύεται με δύο από τα τρία σετ που χρησιμοποιήθηκαν.

8.2.1.1 Πείραμα 1

Χρησιμοποιούνται 2 hidden layers με 20 units Bidirectional LSTM units η κάθε μια, δηλαδή, συνολικά 80 units. Παρακάτω, παρουσιάζονται οι καμπύλες κόστους για κάθε δίκτυο.

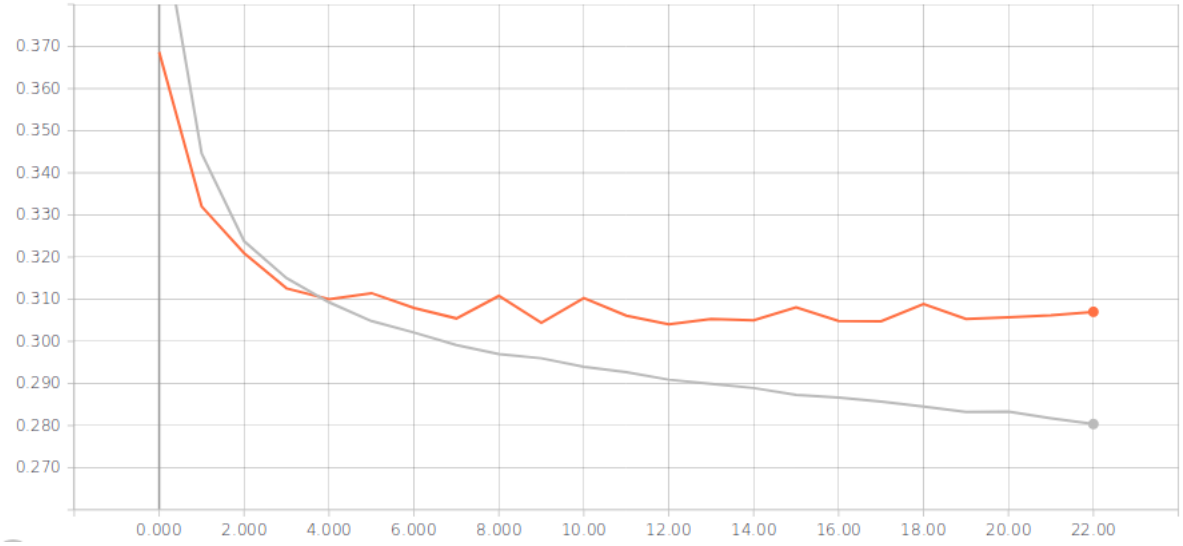

Figure. 8.1: Εκπαίδευση με Ballroom και Hainsworth

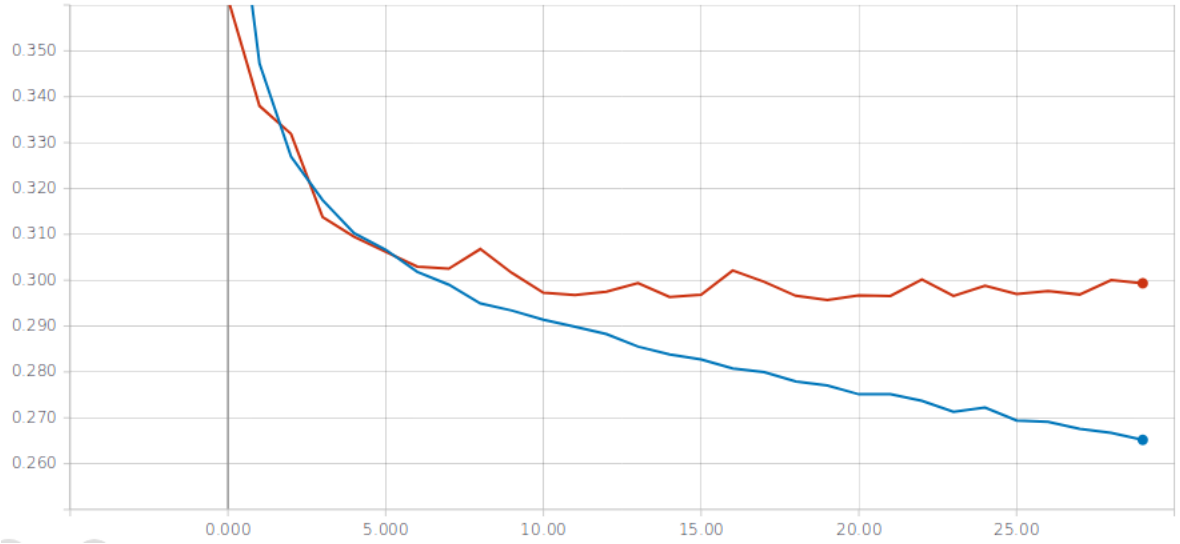

Figure. 8.2: Εκπαίδευση με Ballroom και SMC

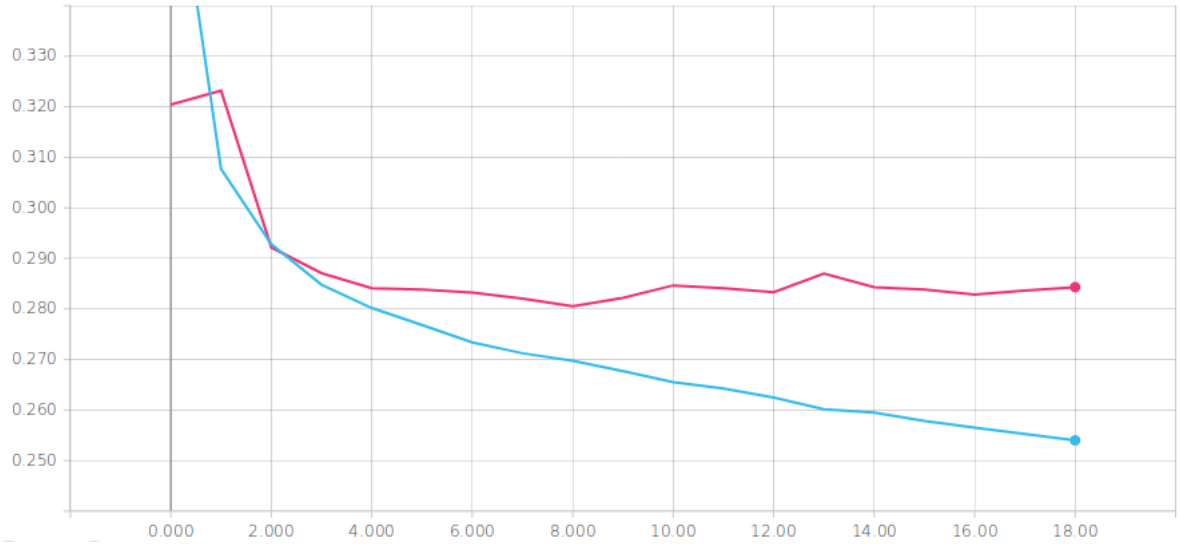

Figure. 8.3: Εκπαίδευση με Hainsworth και SMC

Τα αποτελέσματα του Πειράματος 1 είναι τα παρακάτω

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H	10701	7721	21212	0.2669	0.7215	0.3896	0.516
B, S		7753	20340	0.2760	0.7245	0.3997	
H, S		8133	19739	0.2918	0.7600	0.4217	

Table 8.1: Αποτελέσματα για το SMC dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H	22640	19923	15182	0.5675	0.8800	0.6900	0.867
B, S		19101	15814	0.5471	0.8437	0.6637	
H, S		18815	13154	0.6010	0.8752	0.7127	

Table 8.2: Αποτελέσματα για το Hainsworth dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H	43838	37344	25028	0.5987	0.8519	0.7032	0.938
B, S		37131	23912	0.6083	0.8470	0.7081	
H, S		34538	27224	0.5592	0.7879	0.6541	

Table 8.3: Αποτελέσματα για το Ballroom dataset

Όπως και σε όλα τα επόμενα περάματα, παρατηρείται ότι η απόδοση για το SMC dataset είναι η χαμηλότερη. Ακόμα και στην περίπτωση του State of the Art συστήματος, η τιμή του Fscore είναι αρκετά χαμηλή. Αυτό συμβαίνει διότι το dataset αυτό περιλαμβάνει μουσική με αρκετά μεγάλα tempo fluctuations και ίσως, σε κάποιο βαθμό, ανακριβή annotations.

Το σύστημα βρίσκει τα περισσότερα relevant instances, όμως παρουσιάζονται αρκετά false positives.

8.2.1.2 Πείραμα 2

Χρησιμοποιούνται 2 hidden layers με 25 units Bidirectional LSTM units η κάθε μια, δηλαδή, συνολικά 100 units. Παρακάτω, παρουσιάζονται οι καμπύλες κόστους για κάθε δίκτυο.

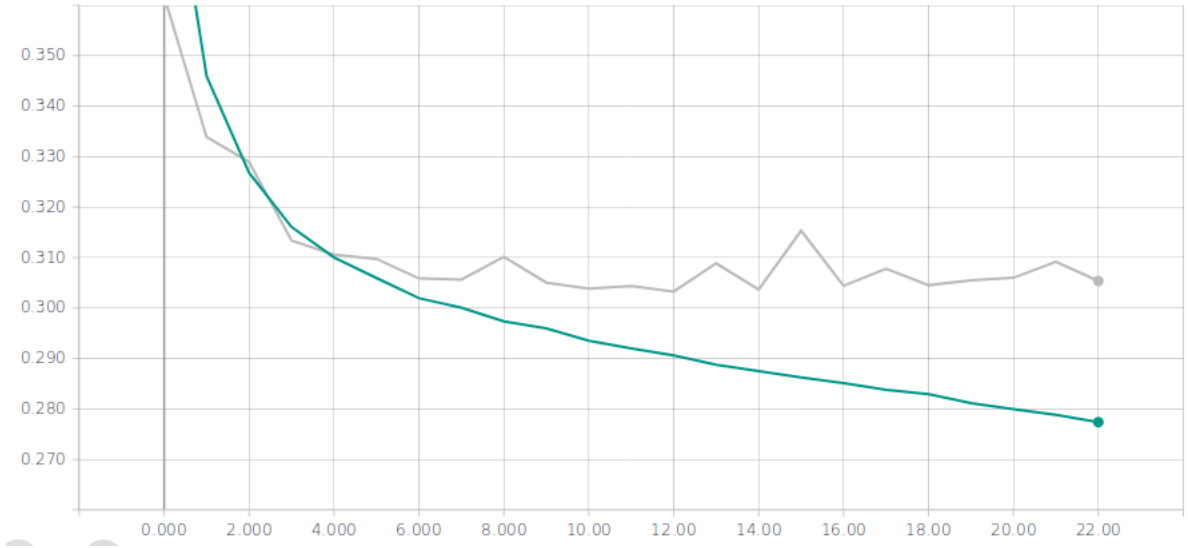

Figure. 8.4: Εκπαίδευση με Ballroom και Hainsworth

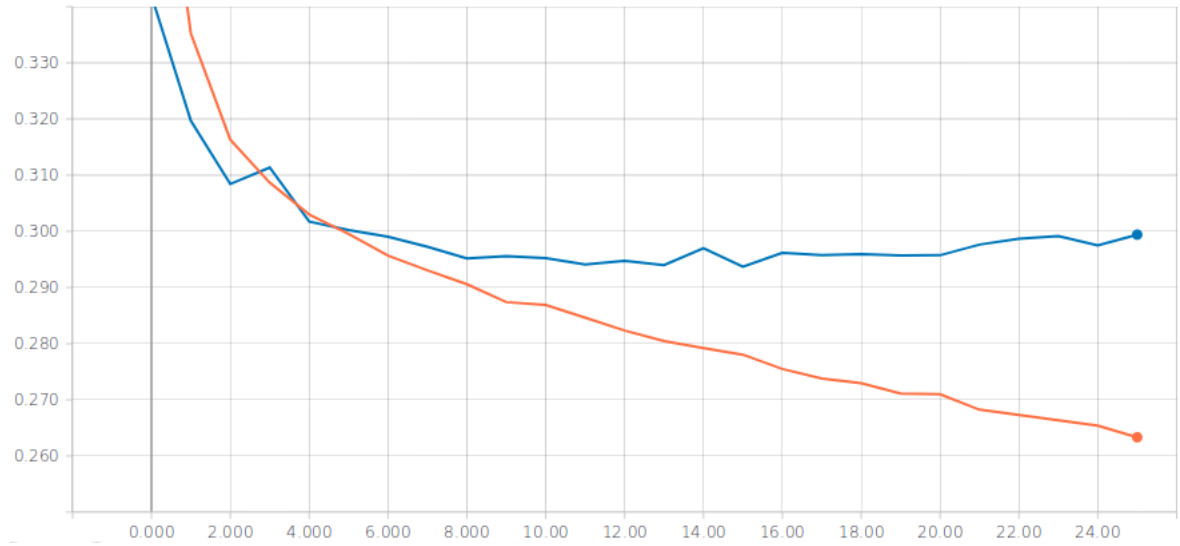

Figure. 8.5: Εκπαίδευση με Ballroom και SMC

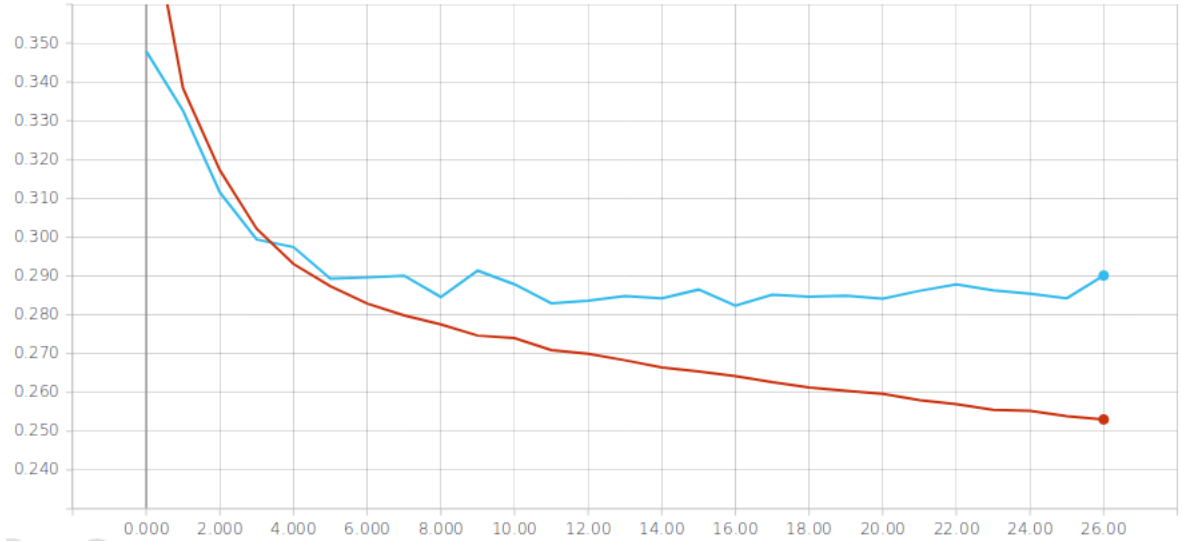

Figure. 8.6: Εκπαίδευση με Hainsworth και SMC

Τα αποτελέσματα του Πειράματος 2 είναι τα παρακάτω

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
B, H		7528	20618	0.2675	0.7035	0.3876	
B, S	10701	7931	20213	0.2818	0.7411	0.4083	0.516
H, S		7898	19513	0.2881	0.7381	0.4145	

Table 8.4: Αποτελέσματα για το SMC dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
B, H		19895	14483	0.5787	0.8788	0.6978	
B, S	22640	199143	16857	0.5318	0.8455	0.6529	0.867
H, S		19946	12802	0.6091	0.8810	0.7202	

Table 8.5: Αποτελέσματα για το Hainsworth dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
B, H		37232	23523	0.6128	0.8493	0.7119	
B, S	43838	37515	24563	0.6043	0.8558	0.7084	0.938
H, S		35025	25903	0.5749	0.7990	0.6686	

Table 8.6: Αποτελέσματα για το Ballroom dataset

Το πείραμα αυτό παρουσιάζει λίγο καλύτερα αποτελέσματα από το προηγούμενο. Τα False Positives μειώνονται, ενώ τα true Positives αυξάνονται. Οι διαφορά των training loss και validation loss, παραμένει περίπου ίδια με το Πείραμα 1, σε όλες τις περιπτώσεις.

8.2.1.3 Πείραμα 3

Χρησιμοποιούνται 3 hidden layers με 20 units Bidirectional LSTM units η κάθε μια, δηλαδή, συνολικά 100 units. Παρακάτω, παρουσιάζονται οι καμπύλες κόστους για κάθε δίκτυο.

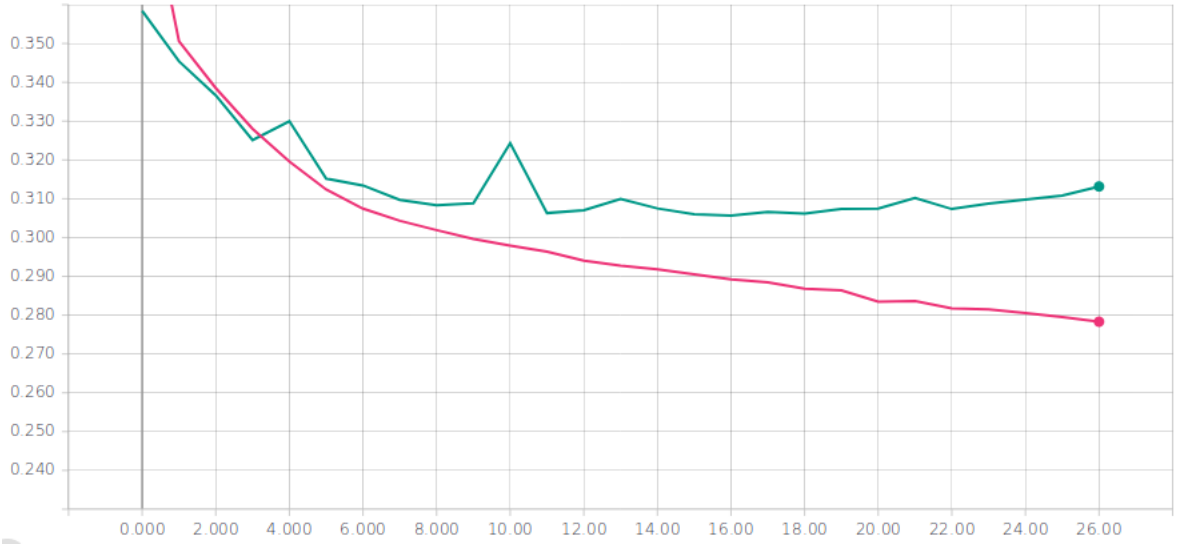

Figure. 8.7: Εκπαίδευση με Ballroom και Hainsworth

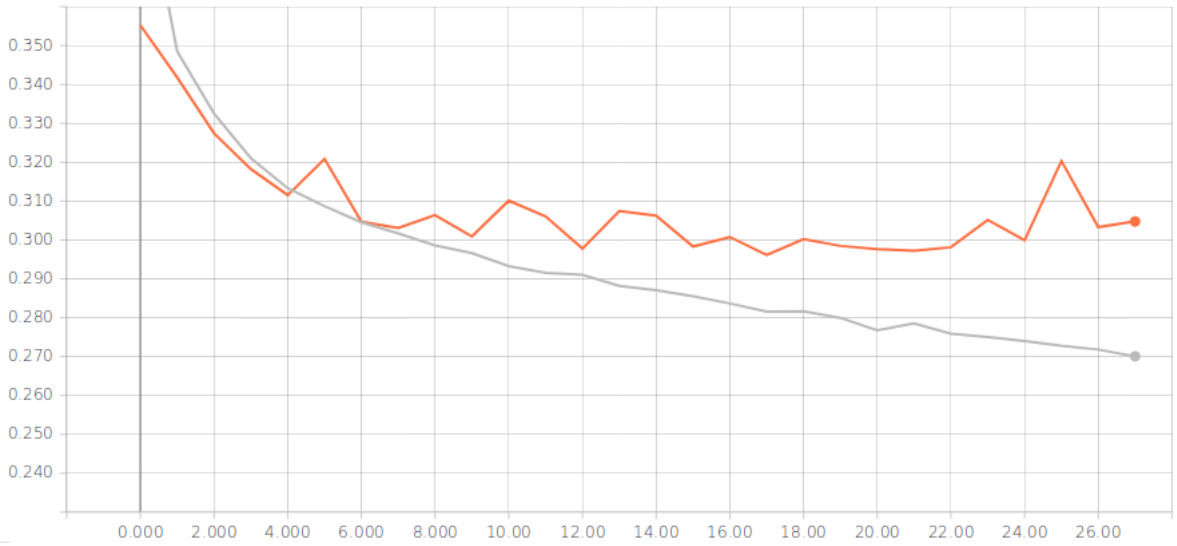

Figure. 8.8: Εκπαίδευση με Ballroom και SMC

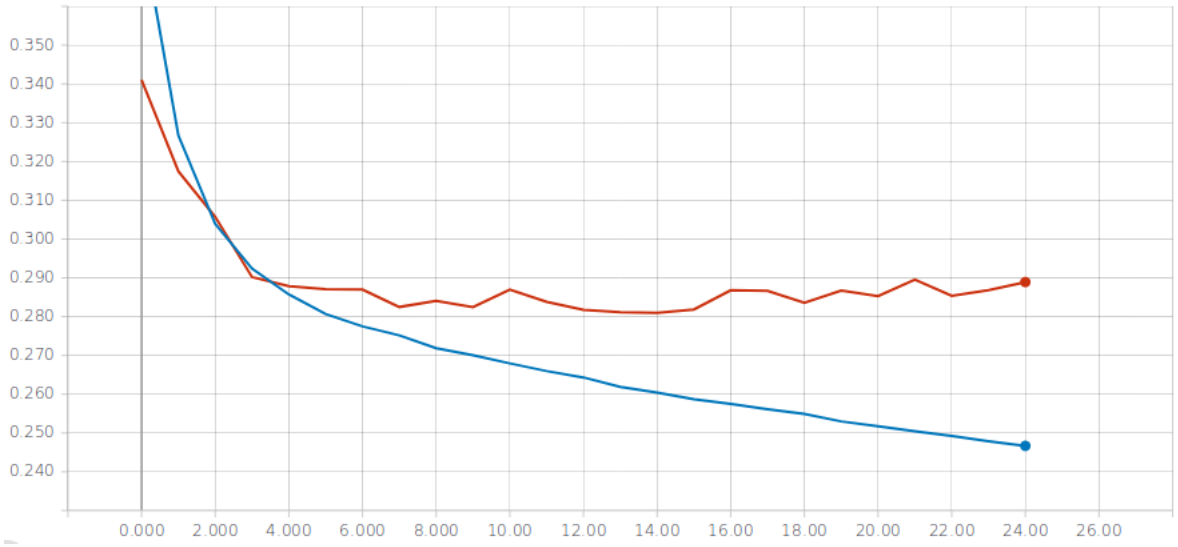

Figure. 8.9: Εκπαίδευση με Hainsworth και SMC

Τα αποτελέσματα του Πειράματος 3 είναι τα παρακάτω

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H		7635	21545	0.2617	0.7135	0.3829	
B, S	10701	7831	19685	0.2846	0.7318	0.4098	0.516
H, S		7909	19171	0.2921	0.7391	0.4187	

Table 8.7: Αποτελέσματα για το SMC dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H		20075	15838	0.5589	0.8865	0.6856	
B, S	22640	18928	15994	0.5420	0.8360	0.6577	0.867
H, S		19593	12543	0.6097	0.8654	0.7154	

Table 8.8: Αποτελέσματα για το Hainsworth dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H		37427	25422	0.5955	0.8538	0.7016	
B, S	43838	37183	23409	0.6137	0.8482	0.7121	0.938
H, S		34078	25519	0.5718	0.7779	0.6583	

Table 8.9: Αποτελέσματα για το Ballroom dataset

Παρατηρούνται περίπου ίδια αποτελέσματα με τις προηγούμενες αρχιτεκτονικές.

8.2.1.4 Πείραμα 4

Χρησιμοποιούνται 4 hidden layers με 25 units Bidirectional LSTM units η κάθε μια, δηλαδή, συνολικά 100 units. Παρακάτω, παρουσιάζονται οι καμπύλες κόστους για κάθε δίκτυο.

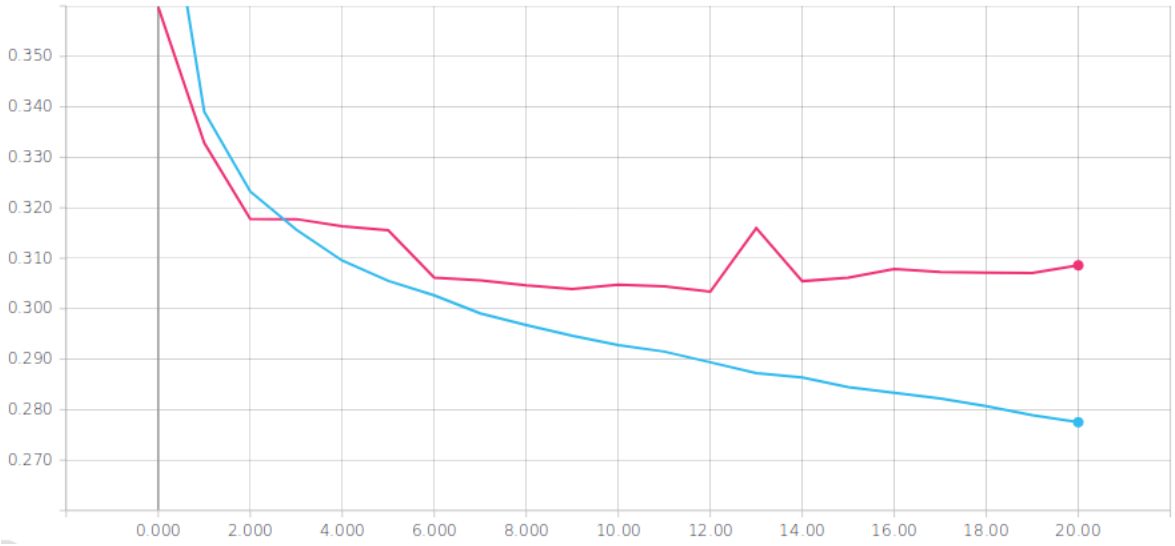

Figure. 8.10: Εκπαίδευση με Ballroom και Hainsworth

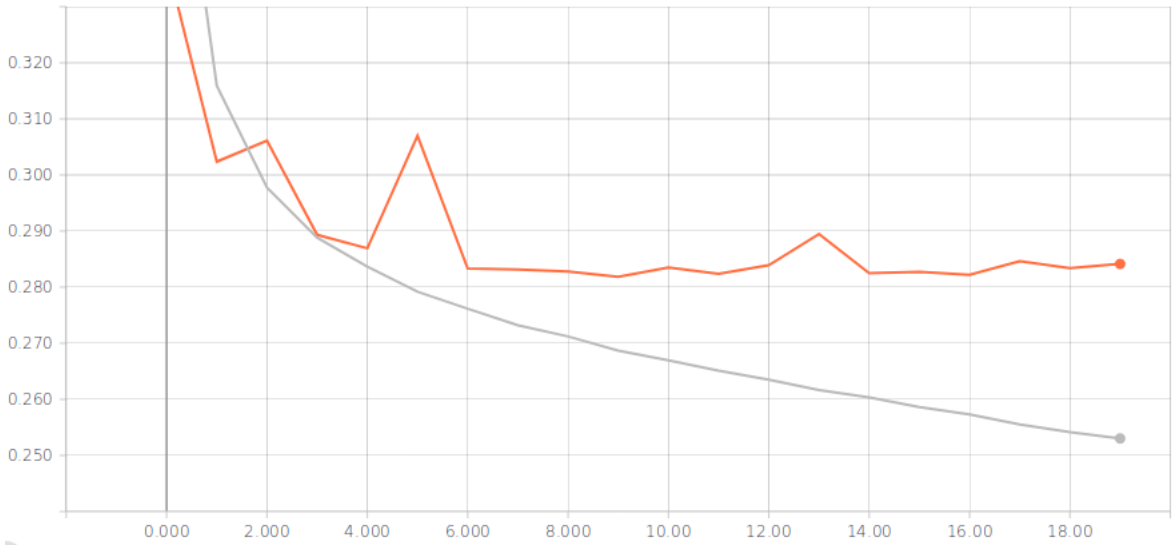

Figure. 8.11: Εκπαίδευση με Ballroom και SMC

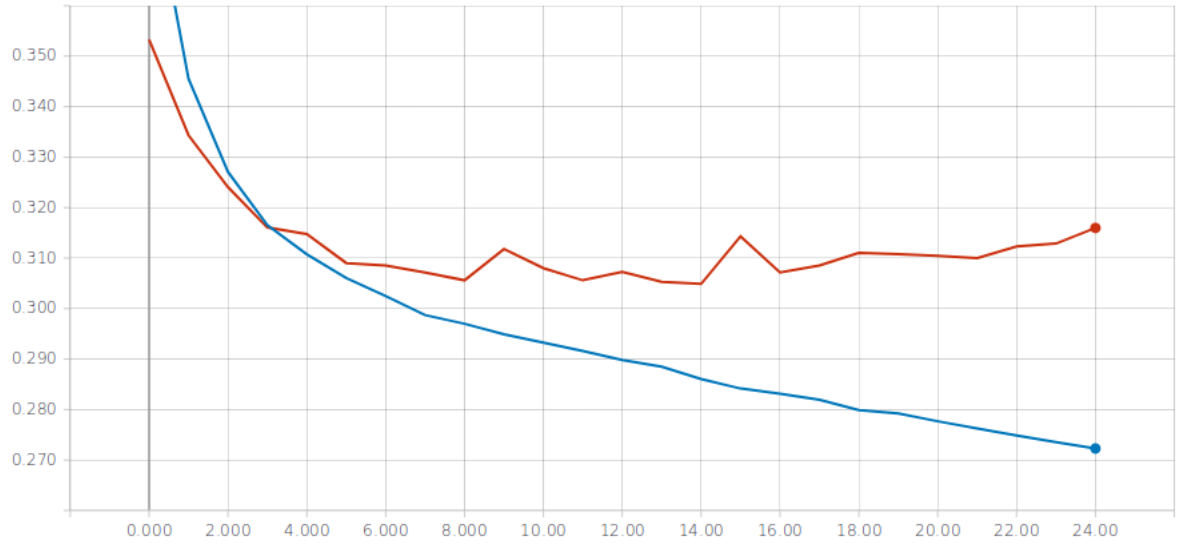

Figure. 8.12: Εκπαίδευση με Hainsworth και SMC

Τα αποτελέσματα του Πειράματος 4 είναι τα παρακάτω

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
B, H		7511	20401	0.2691	0.7019	0.3890	
B, S	10701	7615	20814	0.2679	0.7116	0.3892	0.516
H, S		7917	19559	0.2881	0.7398	0.4148	

Table 8.10: Αποτελέσματα για το SMC dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
B, H		19782	14642	0.5747	0.8738	0.6933	
B, S	22640	19935	15453	0.5633	0.88805	0.6871	0.867
H, S		18979	16341	0.5373	0.8383	0.6549	

Table 8.11: Αποτελέσματα για το Hainsworth dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
B, H		37237	23768	0.6110	0.8494	0.7107	
B, S	43838	37373	24276	0.6062	0.8525	0.7686	0.938
H, S		37188	23356	0.6138	0.8483	0.7123	

Table 8.12: Αποτελέσματα για το Ballroom dataset

Παρατηρούνται περίπου ίδια αποτελέσματα με τις προηγούμενες αρχιτεκτονικές.

8.2.2 Πείραμα 5

Σε αυτό το πείραμα χρησιμοποιήθηκε αρχιτεκτονική CNN για την επίλυση του προβλήματος. Το input του δικτύου είναι παρόμοιο με αυτό των προηγούμενων πειραμάτων. Το δίκτυο δέχεται σαν είσοδο ένα tensor με διαστάσεις $(32)x(300)x(1)x(128)$. Η πρώτη διάσταση αντιστοιχεί στο batch size, η δεύτερη στον αριθμό των frames της εισόδου, η τρίτη αντιστοιχεί στο ύψος της εικόνας, που στην περίπτωση μιας χρονοσειράς είναι 1, και η τέταρτη αντιστοιχεί στα bins συχνοτήτων του Mel Spectrogram. Το CNN αποτελείται από 3 Convolutional Layers, με αριθμό φίλτρων 20 μεγέθους 3. Ως activation function χρησιμοποιείται η $ReLU$.

Παρακάτω, παρουσιάζονται οι καμπύλες κόστους για κάθε δίκτυο.

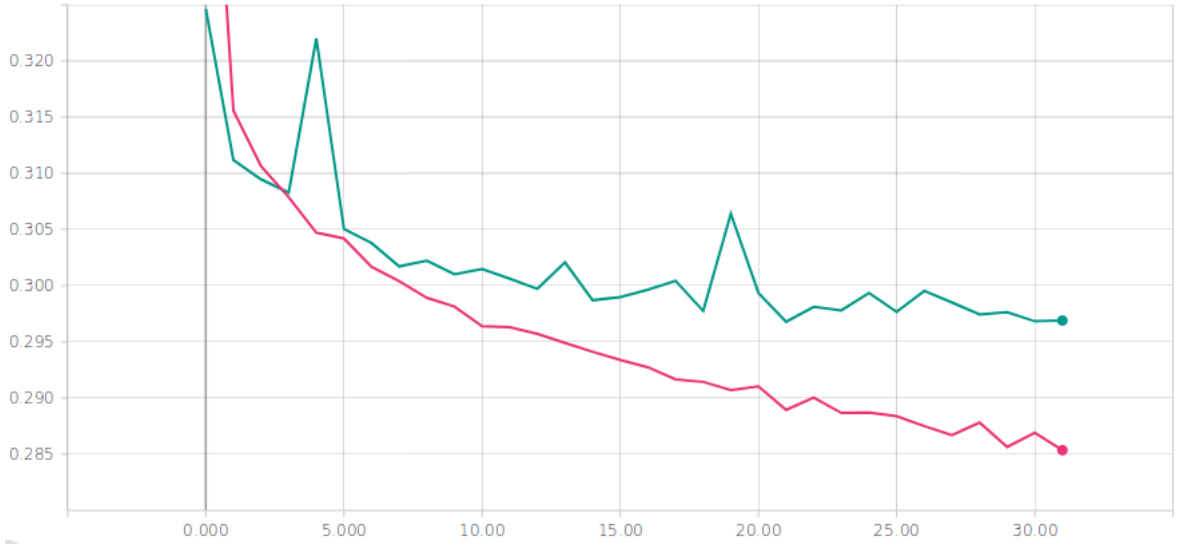

Figure. 8.13: Εκπαίδευση με Ballroom και Hainsworth

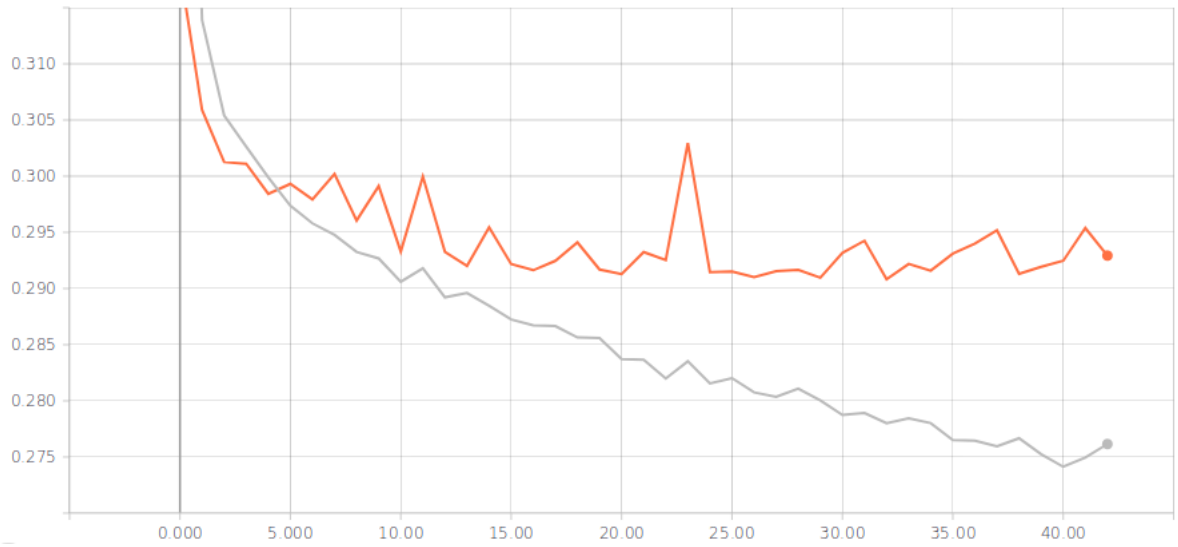

Figure. 8.14: Εκπαίδευση με Ballroom και SMC

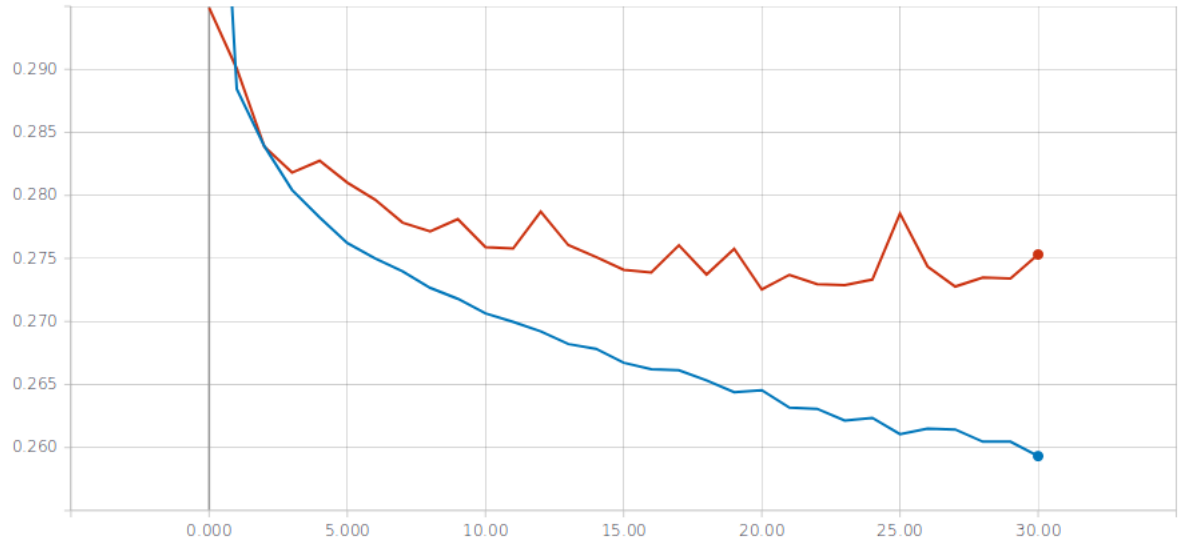

Figure. 8.15: Εκπαίδευση με Hainsworth και SMC

Τα αποτελέσματα του Πειράματος 5 είναι τα παρακάτω:

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H		7593	20202	0.2732	0.7098	0.3945	
B, S	10701	7918	20795	0.2758	0.7399	0.4018	0.516
H, S		7942	20929	0.2751	0.7422	0.4014	

Table 8.13: Αποτελέσματα για το SMC dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H		19717	14482	0.5765	0.8709	0.6938	
B, S	22640	19397	16857	0.5350	0.8568	0.6587	0.867
H, S		19918	14660	0.5760	0.8798	0.6962	

Table 8.14: Αποτελέσματα για το Hainsworth dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H		37136	23455	0.6129	0.8471	0.7112	
B, S	43838	37134	24739	0.6002	0.8471	0.7026	0.938
H, S		35682	27666	0.5633	0.8140	0.6658	

Table 8.15: Αποτελέσματα για το Ballroom dataset

Σε σχέση με την εκπαίδευση των προηγούμενων δικτύων, η εκπαίδευση των CNN δικτύων κράτησε περισσότερο χρόνο για το ίδιο μέγεθος δεδομένων. Τα αποτεέσματα των πειραμάτω δεν είναι καλύτερα από αυτά των LSTM αρχιτεκτονικών.

8.2.3 Πειράματα 6-7

Στα πειράματα 6 και 7, χρησιμοποιείται η αρχιτεκτονική LSTM και τα training data που έδωσαν τα καλύτερα αποτελέσματα στα προηγούμενα πειράματα. Συγκεκριμένα, παρατηρείται ότι η αρχιτεκτονική 2 layer με 25 Bidirectional units, στην οποία χρησιμοποιήθηκαν ως training data τα datasets Hainsworth και SMC, δίνει μέσο Fscore 0.6011, που είναι το μεγαλύτερο. Με βάση την απόδοση αυτή, χρησιμοποιείται η παραπάνω αρχιτεκτονική. Σκοπός των πειραμάτων 6 και 7, είναι να διαπιστωθεί αν η συνεργασία περισσότερων από ένα δικτύων δίνει καλύτερα αποτελέσματα.

Το τελικό BAF δημιουργείται από τον μέσο όρο των BAF που δημιουργεί κάθε δίκτυο ξεχωριστά.

8.2.3.1 Πείραμα 6

Στο πείραμα 6, συνδυάζεται η αρχιτεκτονική LSTM με την αντίστοιχή της από το Πείραμα 5 (CNN)

Τα αποτελέσματα του Πειράματος 6 είναι τα παρακάτω:

Testing Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
S	10701	8167	20355	0.2863	0.7632	0.4164	0.516
H	22640	20177	13443	0.6001	0.8912	0.7172	0.867
B	43838	35997	26101	0.5791	0.8193	0.6786	0.938

Table 8.16: Αποτελέσματα Πειράματος 6

Παρατηρείται ότι, για το dataset στο οποίο δεν εκπαιδεύτηαν τα δίκτυα, υπάρχει μικρή αύξηση της απόδοσης του συστήματος.

8.2.3.2 Πείραμα 7

Στο πείραμα 7, εκπαιδεύουμε 5 ίδια δίκτυα της καλύτερης αρχιτεκτονικής. Η αρχικοποίηση των παραμέτρων πριν την εκπαίδευση, γίνεται τυχαία. Έτσι, τελικά κάθε δίκτυο θα παράγει διαφορετικό BAF.

Τα αποτελέσματα του Πειράματος 7 είναι τα παρακάτω:

Testing Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
S	10701	8161	19973	0.2901	0.7626	0.4203	0.516
H	22640	20234	12706	0.6143	0.8957	0.7281	0.867
B	43838	35881	26355	0.5763	0.8185	0.6765	0.938

Table 8.17: Αποτελέσματα Πειράματος 7

Παρατηρείται ότι, για το dataset στο οποίο δεν εκπαιδεύτηκαν τα δίκτυα, η απόδοση είναι μικρότερη σε σχέση με το Πείραμα 6.

8.2.4 Πείραμα 8

Σε αυτό το πείραμα δίνουμε μια πιο σύνθετη είσοδο στο νευρωνικό μας δίκτυο. Χρησιμοποιείται η αρχιτεκτονική του πειράματος 4, δηλαδή, 3 LSTM επίπεδα με 25 bidirectional units το καθένα. Η είσοδος που τροφοδοτείται το δίκτυο σχηματίζεται από τον συνδιασμό τριών διαφορετικών Mel Spectrogram. Τα παράθυρα του STFT που χρησιμοποιούνται είναι 1024, 2048, 4096 ενώ τα filter banks που χρησιμοποιούνται έχουν μέγεθος 22, 45 και 90, αντίστοιχα. Επιπρόσθετα, σε συνδιασμό με την προηγούμενη είσοδο, υπολογίζεται και η θετική διάμεση διαφορά πρώτης τάξης(positive first order median difference), για τα προηγούμενα σπεκτρογραφήματα σύμφωνα με τον τύπο:

$$D^{+}(n,m) = H(M(n,m) - M_{median}(n,m))$$

όπου:

- $M(n,m)$ το Mel spectrogram, με n το frame index και m το mel index.
- $M_{median}(n,m) = median\{M(n-l^{*}),\ldots,M(n,m)\}$, όπου l^{*}, το μήκος στο οποίο υπολογίζεται η διάμεση τιμή.
- $H(x) = \frac{x+|x|}{2}$, ώστε να συγκρατούνται μόνο τα θετικά αποτελέσματα.

Παρακάτω, παρουσιάζονται οι καμπύλες κόστους για κάθε δίκτυο.

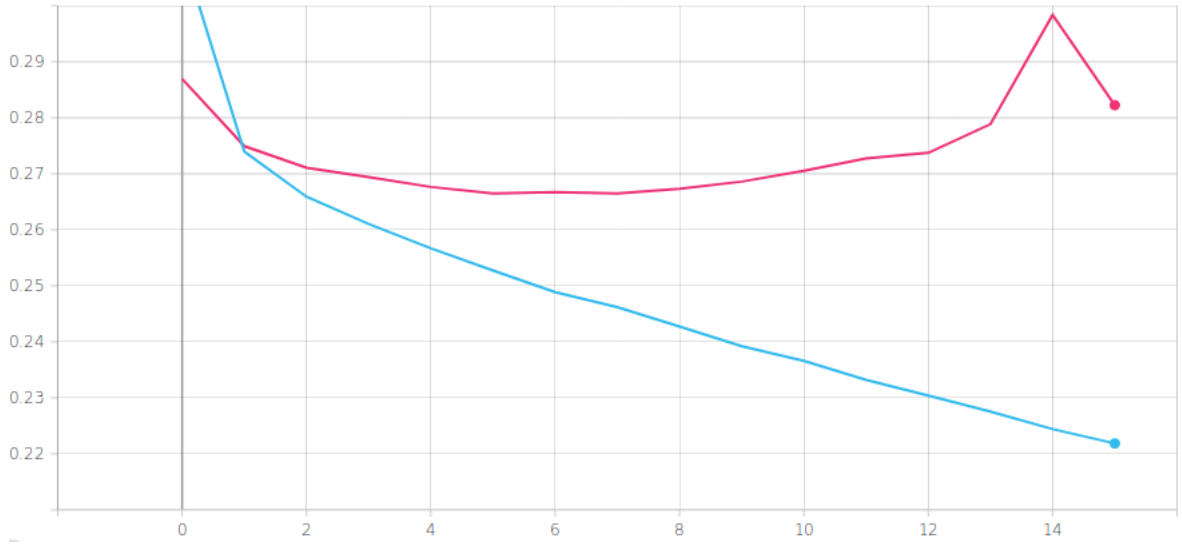

Figure. 8.16: Εκπαίδευση με Ballroom και Hainsworth

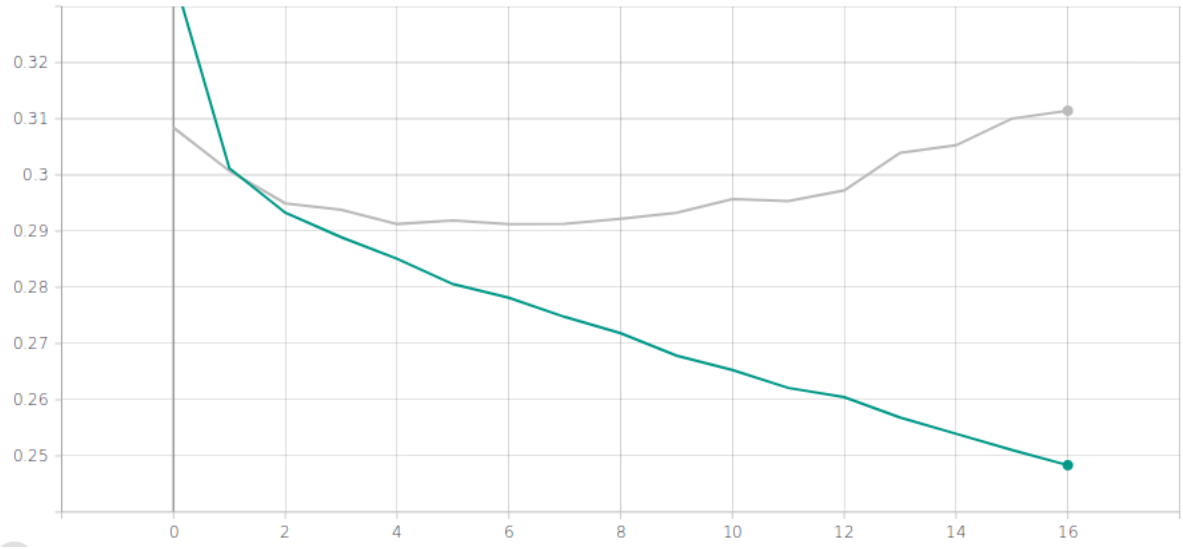

Figure. 8.17: Εκπαίδευση με Ballroom και SMC

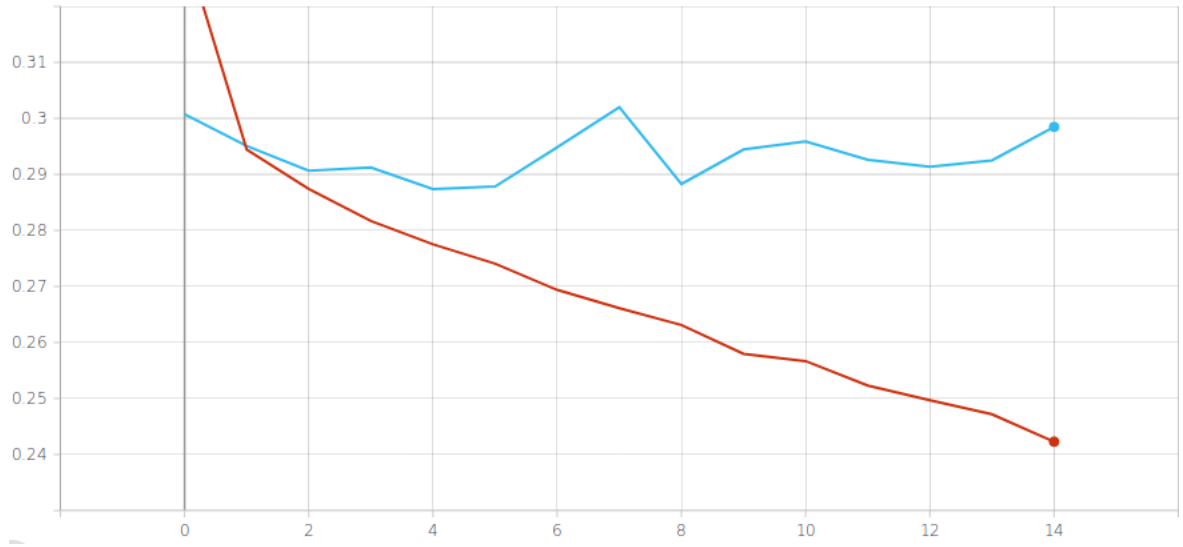

Figure. 8.18: Εκπαίδευση με Hainsworth και SMC

Τα αποτελέσματα του Πειράματος 8 είναι τα παρακάτω:

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H		7560	19215	0.2824	0.7065	0.4035	
B, S	10701	7940	18088	0.3051	0.7420	0.4324	0.516
H, S		7929	17323	0.3140	0.7410	0.4411	

Table 8.18: Αποτελέσματα για το SMC dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H		20127	12649	0.6141	0.8890	0.7264	
B, S	22640	18939	15173	0.5552	0.8365	0.6674	0.867
H, S		20219	11539	0.6367	0.8931	0.7434	

Table 8.19: Αποτελέσματα για το Hainsworth dataset

Training Data	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore_{SOA}
B, H		37624	21623	0.6350	0.8583	0.7200	
B, S	43838	37834	20933	0.6438	0.8630	0.7375	0.938
H, S		35175	24793	0.5866	0.8024	0.6777	

Table 8.20: Αποτελέσματα για το Ballroom dataset

Παρατηρείται καλύτερη απόδοση για όλα τα dataset που χρησιμοποιήθηκαν. Η διαφορετική είσοδος, η οποία διαχωρίζει περισσότερο τα percussive events, φαίνεται ότι βοηθάει την απόδοση του δικτύου.

8.2.5 Πείραμα 9

Σε αυτό το πείραμα εξερευνούμε πώς επηρεάζει το Class Imbalance τα αποτελέσματα. Χρησιμοποιείται το δίκτυο και το training set που έδωσε τα καλύτερα αποτελέσματα στα πειράματα 1-4, δηλαδή η αρχιτεκτονική 2 LSTM επιπέδων με 25 bidirectional units το καθένα. Εξετάστηκαν τρεις περιπτώσεις imbalance, με $zero\ to\ one\ ratio = 1, 15$ και 30.

Παρακάτω, παρουσιάζονται οι καμπύλες κόστους για κάθε $zero\ to\ one\ ratio$.

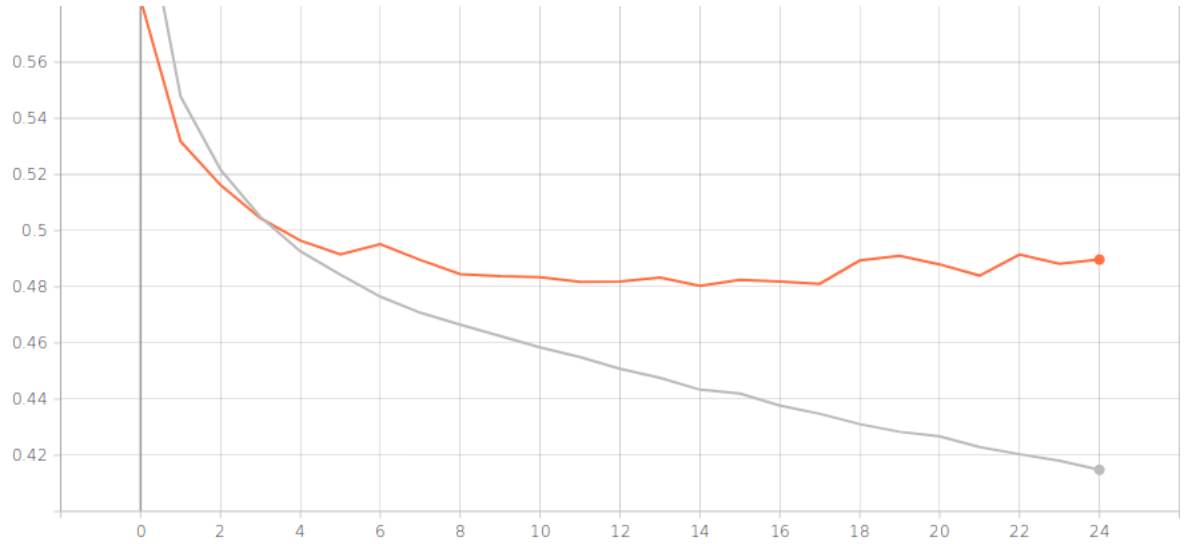

Figure. 8.19: Εκπαίδευση με $zero\ to\ one\ ratio = 1$

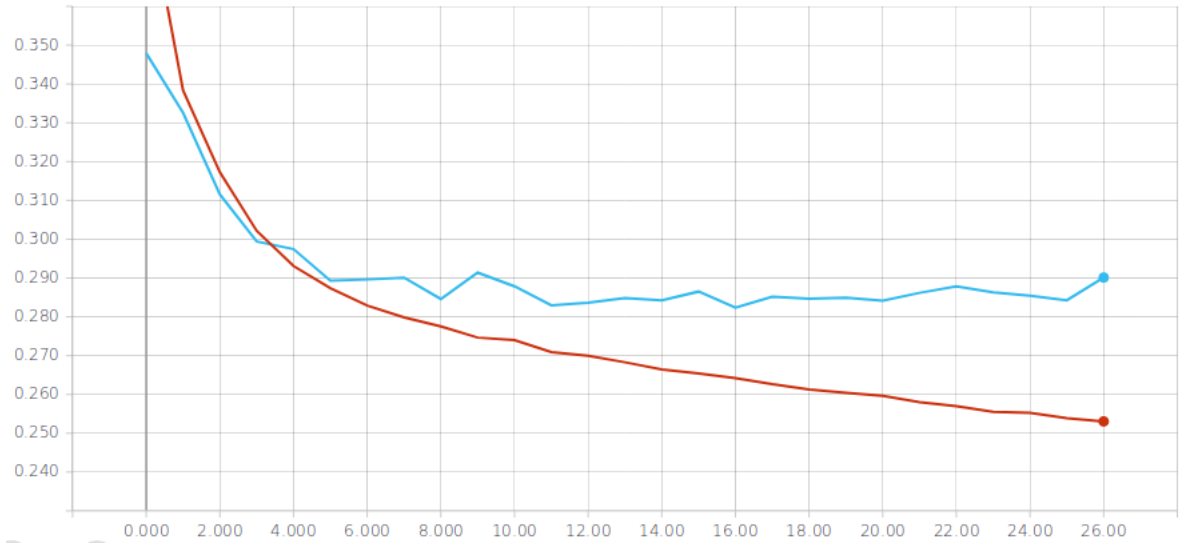

Figure. 8.20: Εκπαίδευση με $zero\ to\ one\ ratio = 15$

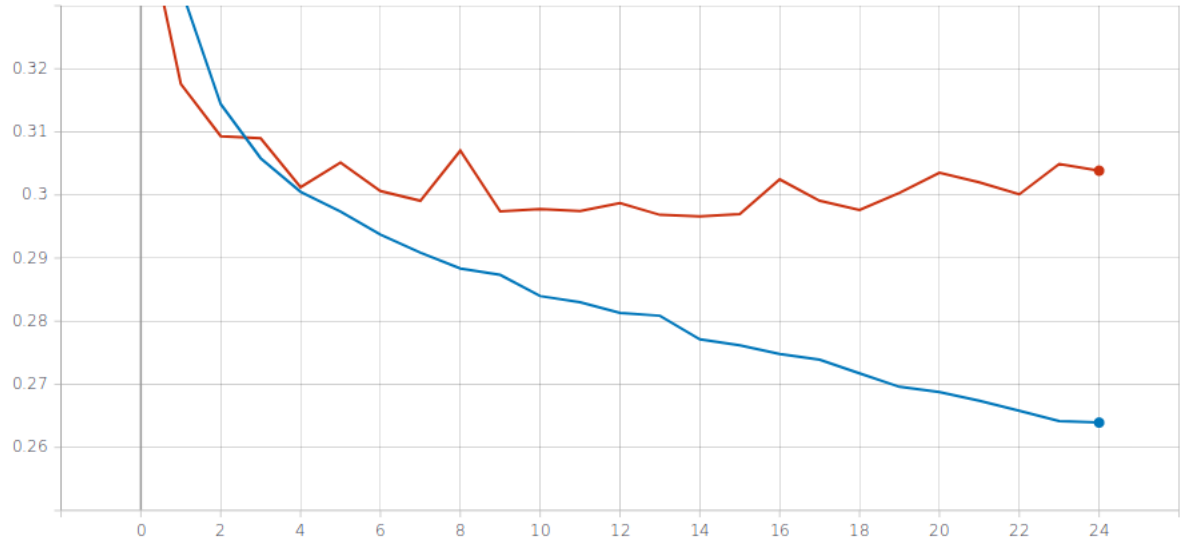

Figure. 8.21: Εκπαίδευση με $zero\ to\ one\ ratio = 30$

Τα αποτελέσματα του Πειράματος 9 είναι τα παρακάτω:

zero to one ratio	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
1		8487	24026	0.2610	0.7931	0.3982	
15	10701	7898	19513	0.2881	0.7381	0.4145	0.516
30		8001	19956	0.2862	0.7477	0.4139	

Table 8.21: Αποτελέσματα για το SMC dataset

zero to one ratio	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
1		20454	19913	0.5067	0.9034	0.6493	
15	22640	19946	12802	0.6091	0.8810	0.7202	0.867
30		20100	13464	0.5989	0.8878	0.7153	

Table 8.22: Αποτελέσματα για το Hainsworth dataset

zero to one ratio	Relevant	T_P	F_P	Precision	Recall	Fscore	Fscore$_{SOA}$
1		36685	35876	0.5056	0.8368	0.6303	
15	43838	35025	25903	0.5749	0.7990	0.6686	0.938
30		34781	27358	0.5597	0.7934	0.6564	

Table 8.23: Αποτελέσματα για το Ballroom dataset

Σε όλες τις περιπτώσεις παρατηρείται καλύτερη απόδοση για $zero\ to\ one\ ratio = 15$. Μεγαλύτερο class imbalance φαίνεται να μην βοηθάει το δίκτυο.

9 Συμπεράσματα και Μελλοντική Ανάπτυξη

9.0.1 Συμπεράσματα

Από την παραπάνω εργασία και τα πειράματα, φαίνεται η σημαντικότητα των training data που χρησιμοποιούνται στο εκπαιδευτικό στάδιο. Ανακριβή datasets δίνουν ανακριβή αποτελέσματα και ”μπερδεύουν” το δίκτυο.

Στην περίπτωση του SMC dataset, παρατηρούνται τα χαμηλότερα αποτελέσματα. Αυτό συμβαίνει και στην βιβλιογραφία. Το dataset αυτό περιλαμβάνει μυσική με μεγάλα tempo fluctuations και ίσως έχει ανακριβή annotations.

Το Πείραμα 9 έδειξε ότι πολύ μικρό ή πολύ μεγάλο Class Imbalance στο training set δε δίνει καλά αποτελέσματα. Ένα καλό Class Imbalance, πιθανότατα είναι αυτό που χρησιμοποιήθηκε στα Πειράματα 1-8.

Τα δίκτυα που εκπαιδεύτηκαν στο Πείραμα 8, με την πιο σύνθετη είσοδο, έδωσαν λίγο καλύτερα αποτελέσματα από τα υπόλοιπα. Παρατηρείται μείωση των False Positives. Η πιο σύνθετη είσοδος φαίνεται να βοηθάει το δίκτυο να ξεχωρίζει καλύτερα τις θέσεις των beats.

Η συνεργασία των δικτύων δεν δίνει πολύ καλύτερα αποτελέσματα για το dataset στο οποίο δεν εκπαιδεύτηκαν τα δίκτυα(Ballroom). Για τα άλλα δύο datasets, η απόδοση του συστήματος βελτιώνεται.

Τα συστήματα που δημιουργήθηκαν απέτυχαν να δώσουν καλύτερη απόδοση από αυτή του State of the Art. Αυτό μπορεί να οφείλεται σε πολλούς παράγοντες. Το State of the Art σύστημα χρησιμοποιεί διαφορετικού τύπου post processing με Μαρκοβιανά Μοντέλα Κρυφών Καταστάσεων(Hidden Markov Models). Επίσης, ίσως οι υπαρπαράμετροι και ο τρόπος εκπαίδευσης των συστημάτων που αναπτύχθηκαν, διαφέρουν από αυτές του State of the Art.

Όπως φαίνεται και στα ακουστικά αποτελέσματα που δημιουργήθηκαν ώστε να εξακριβωθεί η επιτυχία των συστημάτων, αρκετές φορές, το σύστημα επιλέγει την διπλάσια ταχύτητα από την επιθυμητή ή την υποδιπλάσια. Έτσι δικαιολογούνται τα πολλά False Positives. Σε περίπτωση μεγάλων tempo fluctuations, το σύστημα δεν έχει τόσο καλή απόδοση.

9.0.2 Μελλοντική Ανάπτυξη

Μελλοντικά, μπορεί να χρησιμοποιηθεί διαφορετική μέθοδος post processing και να χρησιμοποιηθούν περισσότερα datasets, ώστε να επιτευχθούν καλύτερα αποτελέσματα.

Επίσης, θα δοκιμαστεί να δημιουργηθεί ένα δίκτυο, το οποίο θα παίρνει ως έισοδο 3 δευτερόλεπτα μουσικής(300 frames) και θα δίνει σαν έξοδο ποια frames αντιστοιχούν σε beats και ποια όχι.

Βιβλιογραφία

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F., and Widmer, G. (2016a). Madmom: A new python audio and music signal processing library. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1174–1178. ACM.

Böck, S., Krebs, F., and Widmer, G. (2016b). Joint beat and downbeat tracking with recurrent neural networks. In *ISMIR*, pages 255–261.

Böck, S. and Schedl, M. (2011). Enhanced beat tracking with context-aware neural networks. In *Proc. Int. Conf. Digital Audio Effects*, pages 135–139.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

Chollet, F. et al. (2015). Keras. https://keras.io.

Davies, M. E., Degara, N., and Plumbley, M. D. (2009). Evaluation methods for musical audio beat tracking algorithms. *Queen Mary University of London, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06*.

Ellis, D. P. (2007). Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60.

Goto, M. and Muraoka, Y. (1996). Beat tracking based on multiple-agent architecture-a real-time beat tracking system for audio signals. In *Proceedings of the Second International Conference on Multiagent Systems*, pages 103–110.

Holzapfel, A., Davies, M. E., Zapata, J. R., Oliveira, J. L., and Gouyon, F. (2012). Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2539–2548.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python.

Juslin, P. N. and Sloboda, J. A. (2001). *Music and emotion: Theory and research.* Oxford University Press.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep

convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Walt, S. v. d., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30.

Zahray, L., Nakamura, E., and Yoshii, K. Beat and downbeat detection with chord recognition based on multi-task learning of recurrent neural networks.