Πανεπιστημιο Θεσσαλιας

Διπλωματική Εργασία

Hand gesture recognition with 3D-CNN
Αναγνώριση κινήσεων χεριού με 3D-CNN

Συγγραφέας:
Μυρσίνη Καραφύλλη

Επιβλέποντες:
Gerasimos Potamianos
Nikolaos Bellas
Michael Vassilakopoulos

Μια διατριβή που υποβλήθηκε για την εκπλήρωση των απαιτήσεων της Διπλωματικής Εργασίας για την

Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

3 Ιουλίου, 2019

Δήλωση Πνευματικών Δικαιωμάτων

Εγώ, η Μυρσίνη Καραφύλλη, δηλώνω οτι σε αυτή τη διπλωματική, "Αναγνώριση κινήσεων χεριού με 3D-CNN" και το έργο που παρουσιάζεται είναι δικό μου και αποτελεί πνευματική ιδιοκτησία. Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ' ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Επιβεβαιώνω ότι:

- Όπου έχω συμβουλευθεί δημοσιευμένη δουλειά άλλων γίνεται πάντα αναφορά.
- Έχω αναγνωρίσει όλες τις κύριες πηγές που πήρα βοήθεια.

Υπογράφτηκε:

Ημερομηνία:

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Περίληψη

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Διπλωματική Εργασία

Αναγνώριση κινήσεων χεριού με 3D-CNN

Μυρσίνη Καραφύλλη

Η αναγνώριση χειρονομιών αποτελεί ένα αρκετά διαδεδομένο θέμα σε έρευνες των τελευταίων χρόνων. Οι δυναμικές χειρονομίες έχουν προκαλέσει πολλούς συγγραφείς να εξερευνούν νέες μεθόδους για να πετύχουν υψηλά αποτελέσματα. Αρκετοί από αυτούς προτείνουν λύσεις χρησιμοποιώντας βαθιά νευρωνικά δίκτυα. Ο μεγάλος αριθμός παραμέτρων, τα είδη και η ποσότητα επιπέδων, η αρχικοποίηση των βαρών και πολλά αλλά που θα αναλύσουμε στην διπλωματική αυτή παίζουν πολύ σημαντικό ρόλο στον εντοπισμό της κίνησης και την αναγνώριση χειρονομίας.

Η διπλωματική αυτή παρουσιάζει ένα συνδυασμό μεθόδων βαθιάς μάθησης για το δίκτυο αναγνώρισης χειρονομίας που αποτελείται από ένα μεταποιημένο C3D, ένα κομμάτι από το MobileNet και μια ειδική έκδοση των συνελικτικών LSTM. Ο συνδυασμός των αρχιτεκτονικών συμβάλλει στην εκμάθηση διαφορετικών χαρακτηριστικών από τις ακολουθίες. Για το δίκτυο ανίχνευσης κίνησης έχουμε ως βάση το μοντέλο C3D που έχει δημιουργηθεί για τον σκοπό αυτό και βελτιώνουμε τα αποτελέσματά του σε διάφορα σύνολα δεδομένων συγκρίνοντάς τα με άλλες αρχιτεκτονικές. Ακόμη, για το δίκτυο ανίχνευσης κίνησης αξιοποιούμε όλες τις εικόνες από τα σύνολα δεδομένων μας και περιγράφουμε εκτενώς την μέθοδο που τις εισάγουμε στο δίκτυο συγκρίνοντάς την με τη στρατηγική "jitter". Όσο αναφορά την απεικόνιση των αποτελεσμάτων για το μοντέλο χρονικής κατάτμησης παρουσιάζουμε την πορεία της εκπαίδευσής του και για το μοντέλο αναγνώρισης χειρονομίας υλοποιήσαμε τον πίνακα σύγχυσης κλάσεων.

UNIVERSITY OF THESSALY

Abstract

Department of Electrical And Computer Engineering

Diploma Thesis

Hand gesture recognition with 3D-CNN

Mirsini Karafylli

Gesture recognition is a very popular subject over recent years. Dynamic gestures have prompted many writers to explore new methods to achieve high results. Some of them offer solutions using deep neural networks. The issue of large number of parameters, types and number of layers, weight initialization, and many others that we will analyze in this diploma play an important role in locating movement and in gesture recognition.

This diploma thesis presents a combination of deep learning methods for the gesture recognition network consisting of a different version of C3D, a piece of MobileNet and a special version of a convolutional LSTM convolutive. The combination of the architectures contributes to the learning of different features from the sequences. For the motion detection network, we change the C3D model which was made for this purpose and improve its results in different datasets by comparing it to other architectures. In addition, for the motion detection network, we take advantage of all the images from our datasets and describe extensively the method we use to insert them in the network by comparing it with the "jitter" strategy. About the representation of the results for the time division model we present the course of its training and for the gesture recognition model we have implemented the confusion matrix.

Ευχαριστίες

Πρώτα απ' όλα, είμαι ιδιαίτερα ευγνώμων στον σύμβουλό μου τον Ποταμιάνο Γεράσιμο για όλη τη βοήθειά του, τις παρατηρήσεις του και το χρόνο που πέρασε κατά την ανάπτυξη αυτής της διπλωματικής. Η καθοδήγηση και η υποστήριξή του βοήθησαν να φέρω επιτυχώς αυτό το έργο. Θα ήθελα επίσης να ευχαριστήσω την οικογένειά μου που μου προσέφερε όλα αυτά που είχα ανάγκη για την διεκπεραίωση αυτής της διπλωματικής και την υποστήριξη που μου έδινε καθ' όλη τη διάρκεια της υλοποίησής της. Τέλος, θα ήθελα να ευχαριστήσω όλους τους φίλους μου, παλιούς και νέους, για όλες τις χαρές και εμπειρίες που αποκτήσαμε στα χρόνια φοίτησής μας.

Σας ευχαριστώ όλους!

Στον αγαπημένο μου Γιώργο...

Περιεχόμενα

Λίστα εικόνων

Λίστα Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Πρόλογος

Οι χειρονομίες αποτελούν σημαντικό μέρος της επικοινωνίας μας. Χρησιμοποιούμε κινήσεις για να εκφραστούμε (γλώσσα του σώματος), να δώσουμε οδηγίες, να επικοινωνήσουμε με τους κωφάλαλους και τους κωφούς χρησιμοποιώντας τη νοηματική, να μετρήσουμε αριθμούς κλπ. Αυτό είναι απολύτως φυσικό για τον άνθρωπο, θα ήταν όμως και πολύ χρήσιμο να μπορούσαμε να αλληλεπιδράσουμε με μια έξυπνη συσκευή μέσω μιας χειρονομίας (για παράδειγμα μέσω μιας κάμερας, ενός μικροφώνου ή ακόμα και ενός αισθητήρα όπως το "Project Soli" που δημιουργήθηκε από τη Google). Η Amazon Echo είναι μια κατηγορία συσκευών σχεδιασμένων για να ελέγχεται από τη φωνή ανθρώπων και προσφέρει πολλές δυνατότητες στους αγοραστές της, μουσική, τηλεχειρισμό, έξυπνο σπίτι και ψηφιακό βοηθό. Οι ίδιες δυνατότητες θα μπορούσαν να παρέχονται απο μια έξυπνη συσκευή που θα μπορούσε να αναγνωρίσει χειρονομίες για τα ίδια ή και περισσότερα καθήκοντα.

Δεν υπάρχει αμφιβολία οτι η αναγνώριση χειρονομίας είναι ενα πολύ δύσκολο θέμα έρευνας στον τομέα της τεχνητής νοημοσύνης λόγω του μικρού μεγέθους των χεριών και των δακτύλων, το μήκος και τα όρια κάθε χειρονομίας. Η αναγνώριση χειρονομίας μπορεί να χωριστεί σε δυο τμήματα: συνεχής και μεμονωμένη. Η πρώτη περιλαμβάνει μια ακολουθία δεδομένων με άγνωστα όρια, μήκος και αριθμό χειρονομιών. Η δεύτερη αποτελείται απο μια κίνηση σε κάθε ακολουθία δεδομένων και μια σε κάθε δεδομένο μήκος ορίων. Για την πρώτη κατηγορία έχουν γίνει πολλές μελέτες διότι εντάσσεται περισσότερο στην καθημερινότητά μας σε σχέση με την πρώτη.

Η αναγνώριση χειρονομίας σε συνεχείς ακολουθίες έχει μια ακόμη δυσκολία σε σχέση με τις μεμονωμένες διότι περιέχει και την αναγνώριση της κίνησης. Υπάρχει διαφορά στην αναγνώριση κίνησης με αυτή της χειρονομίας στη δομή και την αρχιτεκτονική των μοντέλων που χρησιμοποιούμε. Μπορούμε να καταλάβουμε απο μια εικόνα της ακολουθίας εαν το άτομο εκτελεί κάποια κίνηση ή είναι σε αδράνεια χωρίς να χρειαζόμαστε κάποιο ιστορικό στιγμιοτύπων ενώ δεν μπορούμε να καταλάβουμε ποιά χειρονομία εκτελεί εαν δεν επεξεργαστούμε κάποιες εικόνες πριν απο αυτή που εξετάζεται. Φυσικά υπάρχουν και χειρονομίες που είναι στατικές των οποίων η αναγνώρισή τους δε χρειάζεται ιστορικό εικόνων. Σε αυτή τη διπλωματική εξετάζονται κυρίως δυναμικές εικόνες των οποίων είναι σημαντικός ο συνδυασμός της θέσης των χεριών μαζί με την κίνηση που εκτελούν. Το φόντο και τα αντικείμενα γύρω απο ένα άτομο που εκτελεί μια κίνηση είναι το περιβάλλον που βρίσκεται και αποτελούν πληροφορίες που δε συνδράμουν σε καμία απο τις δυο κατηγορίες που εξετάζουμε. Για αυτό το λόγο χρησιμοποιούμε ενα μεγάλο

σύνολο δεδομένων το οποίο έχει ως σκοπό να εκπαιδευτεί σε ένα δίκτυο στη μετακίνηση των χεριών σε σχέση με το χρόνο.

Το πρώτο νευρωνικό δίκτυο είναι αυτό της χρονικής κατάτμησης. Αποτελεί το μοντέλο που κατηγοριοποιεί κάθε στιγμιότυπο μιας ακολουθίας ως πρώτη κατηγορία την "οριακή" εικόνα, με την έννοια ότι το άτομο πρόκειται να εκτελέσει ή έχει εκτελέσει ήδη μία χειρονομία και ως δεύτερη κατηγορία τη "μη-οριακή" εικόνα δηλαδή οτι το άτομο εκτελεί μια χειρονομία είτε αυτή ανήκει στις κλάσεις του συνόλου δεδομένου μας είτε όχι. Αρκετές υλοποιήσεις του μοντέλου αυτού έχουν γίνει και με δισδιάστατα επίπεδα που έχουν ικανοποιητικά αποτελέσματα στην εκπαίδευση χωρικών χαρακτηριστικών. Εντούτοις η διάσταση του χρόνου πρέπει να συμπεριληφθεί στην εκμάθηση των χαρακτηριστικών για καλύτερα αποτελέσματα. Τα τρισδιάστατα συνελικτικά επίπεδα (3DCNN) πλέον χρησιμοποιούνται σε πολλές αρχιτεκτονικές για την αναγνώριση κίνησης και είναι μεγάλη πρόκληση για πολλούς ερευνητές ο ορθός εντοπισμός κίνησης ώστε να μη χαθούν δεδομένα στη χρονική διάσταση στην έξοδο του μοντέλου. Θα αναλύσουμε αρκετές μεθόδους για τη διατήρηση αυτών των δεδομένων στα παρακάτω Κεφάλαια.

Φυσικά, το δεύτερο νευρωνικό δίκτυο δε θα μπορούσε να είναι άλλο από το δίκτυο αναγνώρισης χειρονομίας. Μετά την εφαρμογή της χρονικής κατάτμησης στα βίντεο, οι συνεχείς αλληλουχίες μετατρέπονται σε απομονωμένες ακολουθίες χειρονομίας που πρέπει να επισημανθούν. Υπάρχουν πολλοί τρόποι ανίχνευσης μιας χειρονομίας απο μια ακολουθία βίντεο όπως το 3DCNN, το οποίο πέτυχε σπουδαία αποτελέσματα στις προκλήσεις αναγνώρισης χειρονομίας μεγάλης κλίμακας μεμονωμένων κινήσεων. Ωστόσο το μήκος ποικίλει απο τη μία ακολουθία στην άλλη. Αυτός είναι ο λόγος για τον οποίο τα είδη ανατροφοδοτούμενων νευρωνικών δικτύων (ΑΝΔ) προωθούνται για τον έλεγχο των χρονικών χαρακτηριστικών. Εντούτοις είναι δύσκολο να εκπαιδεύονται τα βασικά ΑΝΔ για την επίλυση προβλημάτων που απαιτούν μακρόχρονη εξάρτηση.

Το LSTM είναι μια έκδοση των ΑΝΔ που χρησιμοποιεί ειδικές μονάδες και όχι τις βασικές που χρησιμοποιεί το ΑΝΔ. Μια επέκταση του LSTM που περιλαμβάνει συνελικτικές δομές και στις μεταβάσεις εισόδου-κατάστασης και στις εσωτερικές καταστάσεις ονομάζεται ConvLSTM. Αυτή η αρχιτεκτονική χρησιμοποιείται για την αναγνώριση δράσης, αναγνώριση χειρονομίας αλλά και σε άλλους τομείς. Η είσοδος του ConvLSTM ποικίλλει. Μπορεί να είναι μια εικόνα, χάρτες χαρακτηριστικών απο δισδιάστατο συνελικτικό νευρωνικό δίκτυο ή ακόμα και χάρτες χαρακτηριστικών απο τρισδιάστατο συνελικτικό νευρωνικό δίκτυο. Σε αυτή τη διατριβή, ο τρίτος τύπος χρσιμοποιείται ως είσοδος στο δίκτυο. Λόγω του γεγονότος οτι το μέγεθος παραμέτρων του ConvLSTM είναι μεγάλο, εφαρμόζεται μια μέθοδος προκειμένου να μειωθεί ο αριθμός παραμέτρων και να επικεντρωθεί σε ένα σημαντικό μέρος της εξαγωγής χαρακτηριστικών.

1.2 Δομή και συνεισφορά της Διπλωματικής

Το παρόν έγγραφο αναλύει 2 νευρωνικά δίκτυα:

- Για το δίκτυο ανίχνευσης κίνησης εμπνευστήκαμε αρχικά απο την δομή του άρθρου [16]. Τα αποτελέσματα δεν ήταν καθόλου ικανοποιητικά για τα σύνολα δεδομένων που χρησιμοποιούμε οπότε αλλάξαμε τις τιμές των παραμέτρων, κάποια απο τα είδη και την ποσότητα των επιπέδων με βάση το άρθρο [5].
- Για το δίκτυο αναγνώρισης χειρονομίας χρησιμοποιήσαμε ένα τμήμα απο το δίκτυο ανίχνευσης κίνησης, μια τροποποίηση της αρχιτεκτονικής "ConvLSTM" που προτείνει το άρθρο [6] και ενα κομμάτι του "MobileNet".

Η δομή της Διπλωματικής είναι η ακόλουθη:

Στο Κεφάλαιο 2 αναλύουμε την εσωτερική λειτουργία του νεύρωνα και τις συναρτήσεις ενεργοποίησής του που χρησιμοποιούμε στον κώδικα αυτής της διπλωματικής. Ακόμη ορίζεται ο αλγόριθμος οπισθοδιάδοσης σφάλματος των επιπέδων που χρησιμοποιεί η στοχαστική μέθοδος καθοδικής κλίσης για να ανανεώσει τα βάρη του δικτύου.

Στο Κεφάλαιο 3 ορίζεται η αρχιτεκτονική δομή των συνελικτικών νευρωνικών δικτύων και των διασταλμένων συνελίξεων. Επίσης εξηγούμε την ισορροπημένη τετραγωνική συνάρτηση σφάλματος Hinge που παρουσιάστηκε στο άρθρο [5] και χρησιμοποιείται στο δίκτυο χρονικής κατάτμησης του οποίου η ποσότητα και το είδος των επιπέδων ορίζεται σε Πίνακα.

Στο Κεφάλαιο 4 αναλύουμε την εξέλιξη των ανατροφοδοτούμενων νευρωνικών δικτύων. Οι τροποποιήσεις που έχουν γίνει πάνω στην αρχιτεκτονική τους είναι αρκετές και σε αυτό το Κεφάλαιο τονίζουμε τις περιπτώσεις που χρησιμοποιείται η κάθε έκδοσή τους. Ακόμη, αναλύουμε τα πλεονεκτήματα των διαχωρίσιμων συνελίξεων στη διάσταση του βάθους σε σχέση με τις βασικές πράξεις συνέλιξης.

Στο Κεφάλαιο 5 επικεντρωνόμαστε στα πλεονεκτήματα των MobileNets και γιατί χρησιμοποιούμε κομμάτι απο αυτά στο δίκτυο αναγνώρισης χειρονομίας μας. Στην συνέχεια αναλύεται η δομή του δικτύου αναγνώρισης χειρονομίας και παρουσιάζεται αναλυτικά σε Πίνακες.

Στο Κεφάλαιο 6 παρουσιάζουμε τη δομή του συνόλου δεδομένων που θα χρησιμοποιήσουμε για το δίκτυο χρονικής κατάτμησης. Το σύνολο αυτό περιέχει μεγάλες σε διάρκεια ακολουθίες με αποτέλεσμα να πρέπει να τις επεξεργαστούμε χωρίζοντάς τες κατάλληλα. Έτσι παρουσιάζουμε μέσω ενος παραδείγματος το αποτέλεσμα της στρατηγικής "jitter" και της δικιάς μας μεθόδου για την επεξεργασία των ακολουθιών αυτών.

Στο Κεφάλαιο 7 αναλύουμε κάποια ανεπιθύμητα φαινόμενα που προκύπτουν κατά τη διάρκεια ή μετά την εκπαίδευση. Ακόμη παρουσιάζουμε την επεξεργασία του πίνακα αληθείας των δύο δικτύων για εκπαίδευση και τα αποτελέσματά της συγκρίνοντάς την με άλλες κορυφαίες αρχιτεκτονικές.

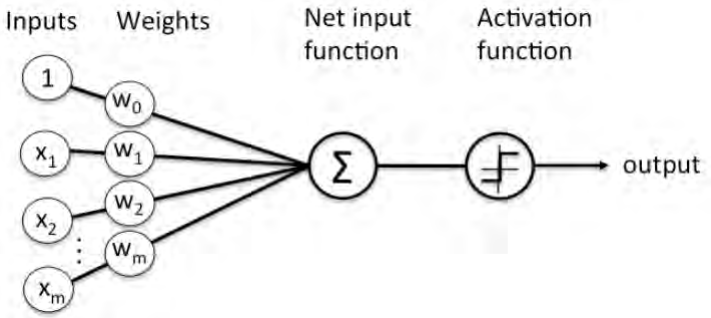
Στο Κεφάλαιο 8 παρουσιάζουμε τα συμπεράσματα της Διπλωματικής αυτής μαζί με μελλοντικά σχέδια για αυτή.

Κεφάλαιο 2

Θεωρία των Νευρωνικών δικτύων

2.1 Εσωτερική λειτουργία του νευρώνα

Τα Νευρωνικά Δίκτυα (Neural Networks ή ΝΔ) χωρίζονται σε δυο ομάδες: στα βιολογικά που αναφέρονται στο νευρικό σύστημα του ανθρώπου και στα τεχνητά (ΤΝΔ) που αναφέρονται σε ένα μαθηματικό μοντέλο εμπνευσμένο από τη δομή του ανθρώπινου εγκεφάλου. Η εκπαίδευση των ΤΝΔ είναι ίδια με αυτή των βιολογικών, δηλαδή μέσα απο ένα σύνολο δεδομένων προσπαθούν να αναλύσουν επαναληπτικά το περιβάλλον τους. Οι δύο αυτές ομάδες απαρτίζονται απο ένα σύνολο νευρώνων (units) το οποίο επεξεργάζεται και αποθηκεύει πληροφορίες. Οι νευρώνες αποτελούνται απο έναν αριθμό δομών διάδοσης σήματος, έναν αθροιστή και μια δομή ενεργοποίησής τους. Ο κάθε νευρώνας έχει τη δυνατότητα να δέχεται πολλές εισόδους αλλά η έξοδός του θα είναι μοναδική. Συνεπακόλουθα, πολλές μοναδικές έξοδοι είναι είσοδοι σε άλλους νευρώνες κι έτσι σχηματίζεται ένα δίκτυο.



Σχημα 2.1: Σχηματική αναπαράσταση του νευρώνα. Εικόνα απο την ιστοσελίδα [1]

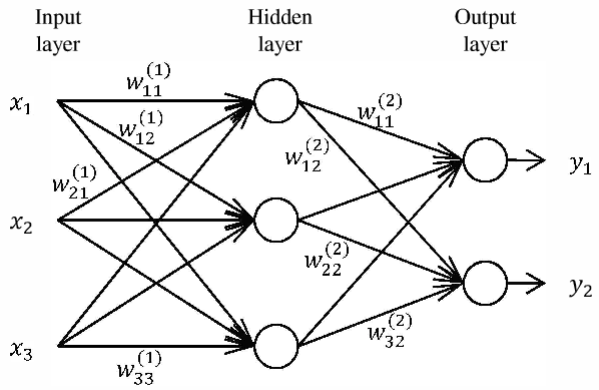
Μεταξύ του στρώματος εισόδου και εξόδου ενος ΝΔ, διακρίνουμε τα κρυφά επίπεδα (hidden layers) τα οποία απαρτίζονται απο ένα σύνολο επιπέδων με νευρώνες που έχουν κοινά χαρακτηριστικά. Οι δομές διάδοσης σήματος αποτελούν τα βάρη (weights) του ΝΔ και ανάλογα με τη σημαντικότητά τους διαφέρουν σε κάθε νευρώνα. Με βάση τις πληροφορίες που δέχεται το στρώμα εισόδου (input layer) και τα αντίστοιχα βάρη του νευρώνα, η συνάρτηση ενεργοποίησης (activation function) υπολογίζει το στρώμα εξόδου (output layer).

Αφού εξηγήσαμε γενικά τη δομή του νευρώνα, οφείλουμε να αναλύσουμε και την εσωτερική λειτουργία του. Το σύνολο των βαρών όπως βλέπουμε και στο Σχήμα 2.1 αναπαριστάται απο τις μεταβλητές w_1, w_2, ... w_m και πολλαπλασιάζεται με τις εισόδους x_0, x_1, x_2, ... x_m που αντιστοιχούν σε κάθε ένα απο τα βάρη. Ειδικότερα, τα βάρη αντικατοπτρίζουν την προσπάθεια του ΝΔ να συνδέσει την προβλεπόμενη τιμή με την τιμή αληθείας. Αρχικά ξεκινούν με τυχαίες τιμές και στη συνέχεια με τη διαδικασία που θα εξηγήσουμε παρακάτω τις ανανεώνουν με βάση τη συνάρτηση σφάλματος σε σχέση με τις τιμές αληθείας.

Ο αθροιστής που περιγράφεται ως "Net input function" στο Σχήμα 2.1 προσθέτει τα m γινόμενα των βαρών με τις εισόδους και προσθέτει ακόμα έναν όρο που ονομάζεται *bias*. Ο όρος αυτός βοηθάει στη μετακίνηση της καμπύλης εκμάθησης ανάλογα με την τιμή του. Ο κάθε νευρώνας έχει από ένα σταθερό όρο *bias* ο οποίος δεν επηρεάζεται από το στρώμα εισόδου επειδή πάντα αυξάνει το άθροισμα γινομένου κατά μία μονάδα και στο Σχήμα 2.1 είναι ο w_0 όρος. Ο μαθηματικός τύπος του αθροιστή για το νευρώνα του Σχήματος ειναι ο παρακάτω:

$$L = \sum_{i=1}^{m} w_i x_i + b \tag{2.1}$$

Στο Σχήμα 2.2 διακρίνουμε τρία επίπεδα όπου L_1, L_2, L_3 το επίπεδο εισόδου, το κρυφό επίπεδο και το επίπεδο εξόδου αντίστοιχα. Επιπροσθέτως, έχουμε τρεις νευρώνες εισόδου, τρεις κρυφούς νευρώνες και δύο νευρώνες εξόδου χωρίς να συμπεριλάβουμε τις τιμές bias.



Σχημα 2.2: Βασικό τεχνητό νευρωνικό δίκτυο με ένα κρυφό επίπεδο. Εικόνα απο την ιστοσελίδα [2]

Κάθε επίπεδο συμβολίζεται με L_i όπου $i \in (1, 3)$. Οι παράμετροι του ΝΔ μας είναι το σύνολο των $(W, b) = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}, W^{(3)}, b^{(3)})$, και αντιστοιχούν δύο πίνακες σε κάθε επίπεδο. Ακόμη, ο συμβολισμός $W_{21}^{(l)}$ περιγράφει την παράμετρο του βάρους που συνδέει το δεύτερο νευρώνα επιπέδου (l) με τον πρώτο νευρώνα επιπέδου $(l+1)$. Παρακάτω φαίνεται η συνάρτηση ενεργοποίησης για τον υπολογισμό του δεύτερου νευρώνα και δεύτερου επιπέδου του Σχήματος 2.2 που συμβολίζουμε με f και $b_1^{(l)}$ το bias επιπέδου $l+1$ του πρώτου νευρώνα.

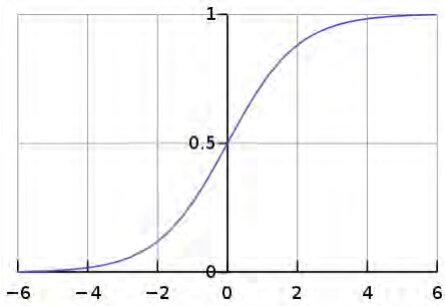
$$f(W_{12}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{32}^{(1)} x_3 + b_2^{(1)}) \tag{2.2}$$

2.2 Συναρτήσεις ενεργοποίησης

Όπως είδαμε στο Σχήμα 2.1 κάθε νευρώνας περιέχει μια συνάρτηση ενεργοποίησης η οποία υπολογίζεται στην έξοδό του με ένα μη-γραμμικό μετασχηματισμό. Γενικότερα, οι γραμμικές συναρτήσεις ειναι πολύ περιορισμένες σε δυνατότητες σε σχέση με τις μη-γραμμικές διότι είναι πιο εύκολο για ένα μοντέλο να προσαρμόσει διαφόρων ειδών δεδομένα σε μία πιο ευέλικτη συνάρτηση. Κάποια ευρέως γνωστά παραδείγματα τέτοιων συναρτήσεων είναι η σιγμοειδής, η υπερβολικής εφαπτομένης και η τόξου εφαπτομένης. Σε αυτή την ενότητα θα αναλύσουμε τη σιγμοειδή, τη μονάδα γραμμικής ανόρθωσης και την τυποποιημένη εκθετική συνάρτηση.

2.2.1 Σιγμοειδής συνάρτηση ενεργοποίησης

Η σιγμοειδής συνάρτηση ενεργοποίησης (sigmoid activation function) έχει μια μορφή S-σχήματος. Ο κύριος λόγος που εφαρμόζουμε τη συνάρτηση αυτή στα επίπεδά μας είναι επειδή έχει ως έξοδο τιμές που κυμαίνονται στο διάστημα $(0, 1)$. Επομένως, χρησιμοποιείται κυρίως σε αρχιτεκτονικές όπως αυτής της διπλωματικής, που αφορά αναγνώριση κίνησης. Τέτοια μοντέλα χρειάζονται ως έξοδο μία πιθανότητα σχετικά με το κατά πόσο μία εικόνα περιέχει ένα άτομο που εκτελεί μια κίνηση ή βρίσκεται σε αδράνεια. Με άλλα λόγια είναι μία συνάρτηση με δύο κλάσεις. Αποτελεί λοιπόν τη συνάρτηση ενεργοποίησης του πρώτου ΝΔ αυτής της διπλωματικής επειδή η πιθανότητα για την πρόβλεψη μιας κλάσης ορίζεται στα όρια εξόδου της σιγμοειδούς. Ακόμη, η συνάρτηση είναι διαφορίσιμη, που σημαίνει οτι μπορεί να βρεθεί η κλίση μέσω δύο δεδομένων σημείων. Ο μαθηματικός τύπος της σιγμοειδούς συνάρτησης είναι ο (2.3) και εικονίζεται στο Σχήμα 2.3 .



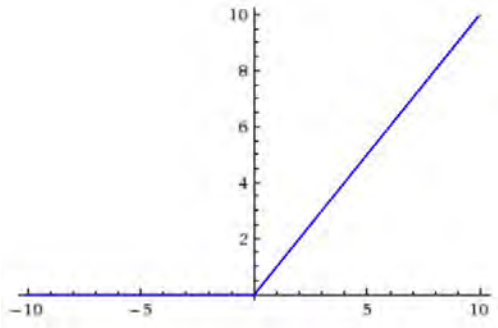
Σχημα 2.3: Σιγμοειδής Συνάρτηση Ενεργοποίησης. Εικόνα απο την ιστοσελίδα [3]

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

2.2.2 Μονάδες γραμμικής ανόρθωσης

Η μονάδα γραμμικής ανόρθωσης (rectified linear unit ή ReLU) αποτελεί την πιο διαδεδομένη συνάρτηση ενεργοποίησης για βαθιά ΝΔ. Αρχικά θεωρείται και αυτή ως μη-γραμμική συνάρτηση όπως διακρίνουμε και απο το Σχήμα 2.4 και η έξοδος της συνάρτησης αυτής κυμαίνεται στο διάστημα $(0, \infty)$. Με την πρώτη ματιά φαίνεται οτι η ReLU δεν είναι τόσο εύελικτη και αυτό προκαλεί σύγχυση σχετικά με αυτά που αναφέραμε στην εισαγωγή της ενότητας αυτής.

Σε ένα βαθύ ΝΔ υπάρχουν νευρώνες πολλαπλών επιπέδων. Χρησιμοποιώντας μια άλλη μη-γραμμική συνάρτηση ενεργοποίησης οπως η σιγμοειδής, προκαλεί σε όλους σχεδόν τους νευρώνες μία ανάλογη ενεργοποίηση. Αυτό σημαίνει ότι το μοντέλο θα επεξεργαστεί όλες αυτές τις ενεργοποιήσεις με αποτέλεσμα μία ανώφελη και χρονοβόρα διαδικασία. Εδώ είναι που χρησιμεύει η ReLU καθώς λόγω της ιδιομορφίας της έχει μηδενική έξοδο για αρνητικές τιμές εισόδου με αποτέλεσμα λιγότερες πράξεις όπως αναφέρει και το άρθρο [4]. Ο μαθηματικός τύπος της συνάρτησης είναι στην σχέση (2.4).



Σχημα 2.4: Μονάδα γραμμικής ανόρθωσης. Εικόνα απο την ιστοσελίδα [4]

$$f(x) = \max\big(0, x\big) \tag{2.4}$$

2.2.3 Τυποποιημένη Εκθετική Συνάρτηση

Η τυποποιημένη εκθετική συνάρτηση (softmax function ή normalized exponential function) δέχεται ως είσοδο ένα διάνυσμα K πραγματικών αριθμών και το ομαλοποιεί σε μια κατανομή πιθανοτήτων που αποτελείται απο K τιμές που αθροίζονται στη μονάδα. Ειδικότερα, η έξοδός της αντιπροσωπεύει την πιθανότητα κάθε κλάσης να είναι αληθής και προφανώς είναι χρήσιμη για μοντέλα που κάνουν προβλέψεις για πολλές κλάσεις όπως το μοντέλο της διπλωματικής αυτής για την αναγνώριση χειρονομίας. Ο μαθηματικός τύπος της softmax παρουσιάζεται στην παρακάτω εξίσωση όπου z το διάνυσμα των εισόδων στο στρώμα εξόδου και όπου j οι νευρώνες εξόδου.

$$\sigma\big(z\big)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{2.5}$$

2.3 Αλγόριθμος οπισθοδιάδοσης σφάλματος

Στην ενότητα αυτή θα αναλύσουμε τον αλγόριθμο οπισθοδιάδοσης σφάλματος (backpropagation algorithm). Γενικότερα, η ροή των δεδομένων έχει μία κατεύθυνση, από το επίπεδο εισόδου προς το επίπεδο εξόδου. Στην περίπτωση της επιβλεπόμενης μάθησης, κατά τη διάρκεια της εκπαίδευσης συγκρίνουμε τις τιμές αληθείας με τις προβλεπόμενες τιμές και με βάση τη συνάρτηση σφάλματος υπολογίζουμε την ποινή της απόκλισης των δύο αυτών τιμών. Συγκεκριμένα, η ελαχιστοποίηση της απόκλισης των δύο τιμών δηλαδή της συνάρτησης λάθους είναι ίδια με την εύρεση της μικρότερης δυνατής τιμής της συνάρτησης σφάλματος.

Η κλίση της συνάρτησης απώλειας μπορεί να υπολογιστεί εάν την παραγωγίσουμε λαμβάνοντας υπόψιν τα βάρη και τα biases. Με αυτή την ενέργεια επιδιώκουμε να υπολογίσουμε πόσο επηρεάζει η κάθε μεταβλητή το ολικό λάθος. Παρόλα αυτά οι τιμές τους αλλάζουν συνεχώς στα διάφορα στρώματα και είναι δύσκολο να υπολογίζονται. Το φαινόμενο αυτό προσπαθεί να εξομαλύνει ο αλγόριθμος οπισθοδιάδοσης σφάλματος.

Ο αλγόριθμος αυτός αποτελεί το εργαλείο που χρησιμοποιεί η μέθοδος που θα αναλύσουμε στην επόμενη ενότητα για να αλλάξει τις τιμές των βαρών. Το κύριο χαρακτηριστικό του είναι η επαναληπτικότητα, αναδρομικότητα και αποτελεσματικότητά του στον υπολογισμό της ενημέρωσης των βαρών με σκοπό να βελτιστοποιήσει το δίκτυο έως ότου είναι σε θέση να εκτελέσει το έργο για το οποίο εκπαιδεύεται. Ακόμη, είναι στενά συνδεδεμένος με τον αλγόριθμο Gauss-Newton.

Στις επόμενες παραγράφους θα αναλύσουμε την εύρεση της παραγώγου του σφάλματος που το αναπαριστούμε με $E = L(t, y)$, όπου t είναι η τιμή στόχος ή αλλιώς η τιμή πρόβλεψης και y είναι η τιμή αληθείας. Οπως βλέπουμε και απο το Σχήμα 2.1 πέρα απο τις εισόδους και τα βάρη, έχουμε τον αθροιστή και την συνάρτηση ενεργοποίησης. Θέτουμε σε αυτή την ενότητα την έξοδο του αθροιστή ως net_j και w_{ij} ως το i-στό βάρος στο j-οστό νευρώνα. Έτσι, με τον κανόνα της αλυσίδας έχουμε την παρακάτω σχέση:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial w_{ij}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial net_j}\frac{\partial net_j}{\partial w_{ij}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial net_j}\frac{\partial\left(\sum_{k=1}^{m} w_{kj} + o_k\right)}{\partial w_{ij}} \tag{2.6}$$

Στο άθροισμα που παραγωγίσαμε όλοι οι όροι μηδενίζονται εκτός απο τον i-οστό όρο καταλήγοντας στην παρακάτω εξίσωση.

$$\frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial net_j}\frac{\partial w_{ij}o_i}{\partial w_{ij}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial net_j}o_i \tag{2.7}$$

Η σχέση $o_j = \sigma(net_j)$ είναι η έξοδος απο τη συνάρτηση ενεργοποίησης κάθε νευρώνα. Απο την εξίσωση (2.7), εξετάζουμε τις περιπτώσεις που ο νευρώνας είναι στην έξοδο ή σε κάποιο κρυφό επίπεδο και παρουσιάζουμε τις δύο αυτές περιπτώσεις στην (2.8).

$$\delta_j = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial net_j} = \begin{cases} \frac{\partial L(o_j,t)}{\partial \varphi(o_j)}\frac{d(\varphi(o_j))}{d(o_j)} \\ \\ \left(\sum_{l\in L}(w_{jl}\delta_l)\right)\frac{d(\varphi(o_j))}{d(o_j)} \end{cases} \tag{2.8}$$

Για να ανανεώσουμε την τιμή ενός βάρους χρησιμοποιώντας τη μέθοδο της επόμενης ενότητας, πρέπει να επιλέξουμε μια τιμή ρυθμού εκμάθησης η και υποθέτουμε οτι είναι μεγαλύτερη του μηδενός. Η αλλαγή στα βάρη πρέπει να επηρεάζει την τιμή της συνάρτησης σφάλματος. Με την παρακάτω εξίσωση επιβεβαιώνουμε ότι κάθε αλλαγή στα βάρη μειώνει τη συνάρτηση σφάλματος.

$$\Delta w_{ij} = -\eta\frac{\partial E}{\partial w_{ij}} = -\eta o_i \delta_j \tag{2.9}$$

2.4 Στοχαστική μέθοδος καθοδικής κλίσης

Υπάρχουν πολλοί αλγόριθμοι που βελτιστοποιούν συναρτήσεις. Χωρίζονται σε κατηγορίες αναλόγα με το εάν βασίζονται σε κλίση ή οχι, με την έννοια οτι αντλούν πληροφορίες όχι μόνο απο τα αποτελέσματα των συναρτήσεων αλλα και από την κλίση τους. Οπως αναφέρει και η ονομασία, η στοχαστική μέθοδος καθοδικής κλίσης (stochastic gradient descent) αποτελεί μια μέθοδο ενημέρωσης του συνόλου των παραμέτρων του νευρωνικού δικτύου και με ένα επαναληπτικό τρόπο ελαχιστοποιεί τη συνάρτηση σφάλματος και κατά συνέπεια ανήκει στην κατηγορία των αλγορίθμων που βασίζονται στην κλίση της συνάρτησης.

Παρόμοια είναι και η μέθοδος καθοδικής κλίσης (gradient descent). Η διαφορά τους είναι οτι η μέθοδος SGD που θα ακολουθήσουμε αλλάζει τα βάρη σε κάθε επανάληψη. Αντιθέτως, η δεύτερη μέθοδος ανανεώνει τα βάρη σε κάθε κύκλο εκπαίδευσης, με την έννοια ότι περνάει όλα τα δείγματα εκπαίδευσης μια φορά για να ανανεώσει τα βάρη. Συνήθως όταν το σύνολο δεδομένων είναι τόσο μεγάλο όσο αυτών της διπλωματικής αυτής, χρησιμοποιούμε την πρώτη μέθοδο διότι θα χρειαστεί πολύς χρόνος για να ολοκληρωθεί ένας κύκλος. Η μέθοδος καθοδικής κλίσης περιγράφεται αναλυτικά στο άρθρο [18] .

Τα βάρη ανανεώνονται με την παρακάτω επαναληπτική διαδικασία και $Q_i(w)$ είναι η τιμή της συνάρτησης σφάλματος στο i-οστό δείγμα δεδομένων:

$$w = w - \eta\sum_{i=1}^{m}\nabla Q_i(w)/n \tag{2.10}$$

Τα βήματα που αναφέραμε είναι : αρχικά εκτελείται η διαδικασία διάδοσης (forward propagation) και βάσει της εισόδου έχουμε την έξοδο που έδωσε το νευρωνικό δίκτυο. Με βάση αυτή την έξοδο υπολογίζουμε το σφάλμα της και τώρα η SGD αρχίζει να ανανεώνει τα βάρη με τη βοήθεια του αλγορίθμου οπισθοδιάδοσης σφάλματος ο οποίος επαναληπτικά υπολογίζει τη σχέση (2.10). Φυσικά στη διαδικασία που μόλις αναφέραμε προκύπτουν αρκετά προβλήματα σχετικά με τη βελτίωση των βαρών τα οποία θα αναλύσουμε σε άλλη ενότητα.

Κεφάλαιο 3

Δίκτυο Χρονικής Κατάτμησης

Σε αυτό το Κεφάλαιο θα αναλύσουμε το νευρωνικό δίκτυο που αφορά την επεξεργασία ενος βίντεο με άγνωστο αριθμό χειρονομιών απο το σύνολο δεδομένων που θα αναφέρουμε στο Κεφάλαιο 5. Στην ενότητα 3.1 παρουσιάζουμε την εξήγηση των επίπεδων που χρησιμοποιούμε στο δίκτυο αυτού του Κεφαλάιου. Στην ενότητα 3.2 περιγράφουμε μια διαφορετική εκδοχή των βασικών συνελίξεων οπου με την βοήθεια μίας παραμέτρου μεγαλώνει τη δεκτική ζώνη των φίλτρων. Στην ενότητα 3.3 αναλύουμε αρκετές μεθόδους απο διάφορα άρθρα σχετικά με την συντήρηση πληροφοριών στον χρόνο διότι οι πράξεις συνέλιξης μειώνουν και την χρονική διάσταση πέρα απο τη χωρική. Τέλος, στην ενότητα 3.4 εξηγούμε τη συνάρτηση κόστους που θα χρησιμοποιήσουμε μόνο για το δίκτυο χρονικής κατάτμησης.

3.1 Συνελικτικά Νευρωνικά Δίκτυα

Τα ΣΝΔ (Convolutional Neural Networks) είναι αντίληπτρο (perceptron) πολλαπλών στρώσεων και αποτελεί τον αλγόριθμο βαθιάς μάθησης (deep learning) που δέχεται ως είσοδο μια εικόνα ή ένα βίντεο και προσπαθεί να προβλέψει αυτό που απεικονίζει με βάση τα δεδομένα που έχει εκπαιδευτεί. Όπως αναφέραμε και προηγουμένως η αρχιτεκτονική τους είναι ανάλογη με αυτή των βιολογικών νευρώνων και παρακάτω θα αναλύσουμε κάποια στοιχεία αυτής της διαδικασίας, κυρίως αυτά που χρησιμοποιούμε σε αυτή τη διπλωματική.

Γενικά, η συνήθης είσοδος στο πρώτο συνελικτικό επίπεδο (convolutional layer) μιας δισδιάστατης συνέλιξης είναι ένας τρισδιάστατος όγκος δεδομένων που οι διαστάσεις του ορίζονται απο το ύψος, το πλάτος και το βάθος της εικόνας. Ως βάθος θεωρούμε τα 3 κανάλια της εικόνας που αποτελούνται απο το κόκκινο, πράσινο και μπλέ κανάλι (RGB). Η συνέλιξη ονομάζεται δισδιάστατη διότι παρότι η εικόνα εισόδου έχει τρία κανάλια, το αποτέλεσμα της συνέλιξης είναι ένας δισδιάστατος χάρτης χαρακτηριστικών. Η τρισδιάστατη συνέλιξη χρησιμοποιείται για να δώσουμε ως είσοδο στο ΝΔ ένα βίντεο και η έξοδος του επιπέδου συνέλιξης περιλαμβάνει και τη διάσταση του χρόνου. Στις επόμενες ενότητες αυτού του κεφαλαίου αφού η διάσταση του χρόνου είναι ένας πολύ σημαντικός παράγοντας για την πρόβλεψη κίνησης σε ένα βίντεο θα αναλύσουμε περαιτέρω την επεξεργασία της διάστασης αυτής.

Ως δεκτική ζώνη (receptive field) ορίζουμε ένα κομμάτι της εισόδου που κοιτάζει ένα συγκεκριμένο χαρακτηριστικό συνελικτικού νευρωνικού δικτύου και ορίζεται από τις διαστάσεις του και το κεντρικό σημείο του μεγέθους του. Επιπλέον δεν είναι όλα τα εικονοστοιχεία σε μία δεκτική ζώνη το ίδιο σημαντικά σχετικά με το χαρακτηριστικό που αναφέρονται. Συγκεκριμένα, σε μια δεκτική

ζώνη, όσο πιο κοντά είναι ένα εικονοστοιχείο στο κέντρο της τόσο πιο πολύ συμβάλει στον υπολογισμό του εξαγώγιμου χαρακτηριστικού. Η διαδικασία της συνέλιξης ξεκινά με ένα τρισδιάστατο φίλτρο το οποίο σαρώνει την εικόνα εισόδου μέσω της δεκτικής ζώνης. Κατά τη διάρκεια της σάρωσης, πραγματοποιούνται τρισδιάστατα εσωτερικά γινόμενα και εξάγουν ένα δισδιάστατο χάρτη χαρακτηριστικών. Στη συνέχεια ο χάρτης αυτός μέσω ενος μη-γραμμικού μετασχηματισμού, δηλαδή της συνάρτησης ενεργοποίησης εξάγει τις ενεργοποιήσεις ανάλογα με τις τιμές εξόδου της.

Δυο σημαντικές παράμετροι είναι το βήμα (stride) και το γέμισμα (padding). Το πρώτο καθορίζει πόσα εικονοστοιχεία θα παραλείψει το φίλτρο σε κάθε βήμα της σάρωσης της εικόνας ή αλλιώς πόσο θα μετακινηθεί το δεκτικό πεδίο σε ύψος και σε πλάτος. Οι τιμές που θέτουμε κυρίως για το βήμα συνέλιξης είναι δύο με τρία, ανάλογα με την αρχιτεκτονική και τις διαστάσεις της εισόδου. Γενικά όσο πιο μεγάλο είναι το βήμα τόσο πιο μικρή σε διαστάσεις θα είναι και η έξοδος του συνελικτικού επιπέδου. Εκτός από το μέγεθος του φίλτρου, η έξοδος του επιπέδου επηρεάζεται και από το βήμα όταν είναι ίσο με τη μονάδα, με αποτέλεσμα να είναι μεγαλύτερη σε διαστάσεις καθώς δεν παραλείπεται ούτε μία χωρική μονάδα. Το γέμισμα αποτελεί την παράμετρο που γεμίζει με μηδενικά το περίγραμμα της εισόδου. Ένας από τους λόγους που εφαρμόζουμε το γέμισμα είναι η διατήρηση των χωρικών διαστάσεων.

Η παρακάτω σχέση περιγράφει τη διάσταση εξόδου (O) με βάση το βήμα (S), το γέμισμα (P), την είσοδο (I) και το φίλτρο (F).

$$O_{w,h} = \frac{I_{w,h} + 2P_{w,h} - F_{w,h}}{S_{w,h}} + 1 \tag{3.1}$$

Ακόμη εκτός απο τα συνελικτικά επίπεδα υπάρχουν και τα επίπεδα εξαγωγής τοπικής μέγιστης τιμής (max pooling layers) που χρησιμοποιούνται επίσης σε βαθιά νευρωνικά δίκτυα. Τα επίπεδα αυτά όπως περιγράφει και η ονομασία τους στηρίζονται στην εξαγωγή του στοιχείου με τη μεγαλύτερη τιμή. Επίσης βοηθάει στην αποφυγή της υπερ-μάθησης και μειώνει το υπολογιστικό κόστος, μειώνοντας τον αριθμό των παραμέτρων που χρησιμοποιούνται για την εκπαίδευση. Αντίστοιχα είναι και τα επίπεδα υποδειγματοληψίας εξαγωγής τοπικής μέσης τιμής (average pooling layers) αλλά αντί για τη μέγιστη τιμή του δεκτικού πεδίου που εξετάζεται τοπικά, υπολογίζεται η μέση τιμή των τιμών που περιέχονται στο πεδίο.

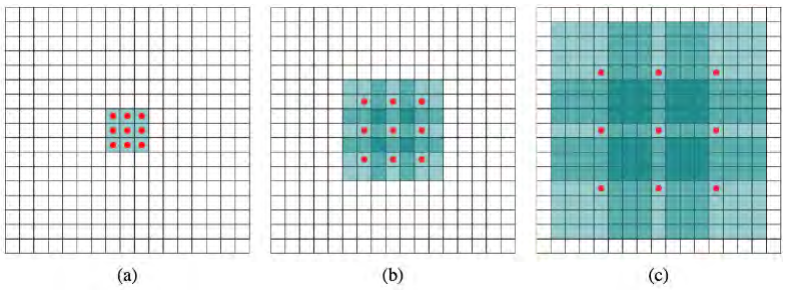
Τέλος, περιγράφουμε τα πλήρως συνδεδεμένα επίπεδα (fully connected layers) που χρησιμοποιούνται ως τελευταία επίπεδα σε ένα ΝΔ και το τελικό επίπεδο εξάγει την κλάση που προέβλεψε το ΝΔ. Αποτελούνται απο μία σύνθεση εσωτερικών γινομένων και μία συνάρτηση ενεργοποίησης.

3.2 Διασταλμένες Συνελίξεις

Οι διασταλμένες συνελίξεις (dilated convolutions) προτείνουν μια ακόμη παράμετρο στις πράξεις συνέλιξης που ονομάζεται ρυθμός διαστολής (dilation rate). Η παράμετρος αυτή καθορίζει την απόσταση μεταξύ των τιμών των φίλτρων που σαρώνουν μία εικόνα. Στην ουσία είναι σαν την απλή συνέλιξη, απλά με κάποια κενά που καθορίζει ο ρυθμός διαστολής ανάμεσα στη δεκτική ζώνη των φίλτρων. Οι πράξεις με διασταλμένες συνελίξεις υλοποιούνται και σε πεδία πέρα απο δισδιάστατες εικόνες. Για παράδειγμα, εφαρμόζονται σε προβλήματα αναγνώρισης από κείμενο σε ομιλία όπως αναφέρει και το άρθρο [20] και έχουν τρία κύρια πλεονεκτήματα.

Πρώτον, χρησιμοποιούνται για την ανίχνευση λεπτομερειών στα δεδομένα επειδή επεξεργάζονται τις εισόδους σε υψηλότερες αναλύσεις. Αυτό είναι πολύ χρήσιμο για τη μελέτη μας και ιδιαίτερα στο σύνολο των δεδομένων μας διότι η κίνηση με τα χέρια δεν αποτελεί τεράστιο τμήμα της εικόνας μας επειδή οι άνθρωποι είναι πολύ μακριά από την κάμερα. Αντιθέτως στο σύνολο δεδομένων με την ονομασία "Jester" που βρίσκεται στην ιστοσελίδα [33], τα άτομα βρίσκονται πολύ κοντά στην κάμερα και είναι πιο ευδιάκριτη η κίνηση. Δεύτερον, μειώνονται οι παράμετροι βαρών με αποτέλεσμα το δίκτυο να χρειάζεται λιγότερη μνήμη. Τρίτο και τελευταίο, η δεκτική ζώνη των επιπέδων αυξάνεται λόγω της ευρύτερης προβολής στην είσοδο.

Στο Σχήμα 3.1(a) μπορούμε να δούμε την δεκτική ζώνη της βασικής συνέλιξης, στο (b) έχουμε διασταλμένη συνέλιξη με ρυθμό διαστολής ίσο με 2 και τέλος στο (c) έχουμε διασταλμένη συνέλιξη με ρυθμό διαστολής ίσο με 4.



Σχημα 3.1: Διασταλμένες συνελίξεις. Εικόνα απο την ιστοσελίδα [21]

3.3 Εντοπισμός Κίνησης

Τα τελευταία χρόνια το θέμα πάνω στην ανίχνευση κίνησης στο χρόνο, αποτελεί ένα κρίσιμο κομμάτι στον τομέα της μηχανικής μάθησης. Έχουν γίνει πολλές προσπάθειες για την εξεύρεση αλγορίθμων που επιτυγχάνουν ακριβή αποτελέσματα. Τις περισσότερες φορές δουλεύουμε με βίντεο που δεν έχει υποστεί επεξεργασία όσον αφορά τον αριθμό χειρονομιών και την διάρκειά τους, διότι προκειμένου να εξάγουμε μεμονωμένες χειρονομίες, πρέπει πρώτα να μετατρέψουμε τα

βιντεο χειροκίνητα ή μέσω ενός νευρωνικού δικτύου. Με βάση τα παραπάνω, δεδομένου ενός μεγάλου βίντεο που περιλαμβάνει διάφορα παραδείγματα κίνησης και πολύπλοκα περιεχόμενα στο φόντο του, δεν πρέπει μόνο να αναγνωρίσουμε τις κατηγορίες χειρονομίας αλλά και να ορίσουμε την έναρξη και τη λήξη τους.

Για να ανιχνεύσουμε τα όρια χειρονομίας σε μια ακολουθία δεδομένων υπάρχουν τρεις μέθοδοι που χρησιμοποιούνται περισσότερο. Ο πρώτος είναι ένας "ολισθαίνων όγκος" ή "παράθυρο πολλαπλών κλιμάκων" για χρονική κατάτμηση. Ωστόσο, απαιτεί τεράστια υπολογιστική ισχύ και δεν συνιστάται για την αναγνώριση χειρονομίας σε βίντεο με ασταθή διάρκεια. Ο δεύτερος αποτελείται από ένα δυαδικό ταξινομητή που αναγνωρίζει αν ένας αριθμός πλαισίων είναι επισημασμένος ως θέση ανάπαυσης ή δράσης, υποθέτοντας ότι τα πλαίσια έναρξης και λήξης είναι παρόμοια. Παρόλα αυτά, στην ενότητα 3.4 αναλύουμε τον λόγο που ο δυαδικός ταξινομητής δεν είναι η κατάλληλη μέθοδος στο δικό μας σύνολο δεδομένων. Το τρίτο χρησιμοποιεί έναν ταξινομητή νευρωνικών δικτύων για να διακρίνει κάθε εικόνα εάν πρόκειται για όριο ή για μέρος μιας χειρονομίας. Σε αυτή τη διπλωματική, η τρίτη μέθοδος χρησιμοποιείται ως το πρώτο νευρωνικό δίκτυο για να ταξινομεί τα όρια χειρονομίας.

Ένας σημαντικός κανόνας για τον ορθό εντοπισμό κίνησης είναι να μη χαθούν δεδομένα στη χρονική διάσταση. Παρόλο που οι αρχιτεκτονικές τρισδιάστατων συνελικτικών νευρωνικών δικτύων στα άρθρα [8] και [9] αποδείχτηκε ότι μαθαίνουν χωρικά και χρονικά χαρακτηριστικά απευθείας από ακατέργαστα βίντεο, δεν καταφέρνουν να χειριστούν την απώλεια λεπτομερειών στο χρόνο. Για παράδειγμα, από την ευρέως γνωστή αρχιτεκτονική C3D που αναλύεται στο άρθρο [8], τα επίπεδα από "conv1a" εως "conv5b" μειώνουν το χρονικό μήκος ενός βίντεο εισόδου κατά 8 φορές. Έγινε μια προσπάθεια από το άρθρο [12] για να ξεπεραστεί αυτή τη δυσκολία με αναδειγματοληψία στην χρονική διάσταση και υποδειγματοληψία στη χωρική διάσταση. Η προτεινόμενη μέθοδος ονομάζεται Convolutional-De-Convolutional (CDC), και είναι ένα φίλτρο το οποίο πραγματοποιεί ταυτόχρονα συνέλιξη στην χωρική διάσταση και απο-συνέλιξη στο χρόνο. Ωστόσο, τα άρθρα [9] και [5] αναφέρουν ότι αυτή η μέθοδος δεν είναι τόσο αποτελεσματική επειδή κατά τη διάρκεια της χρονικής δειγματοληψίας (πριν από την αναδειγματοληψία με τα φίλτρα CDC), τα χρονικά δεδομένα συμπιέζονται από τα φίλτρα που περιλαμβάνουν δύο αντίγραφα των επιπέδων εσωτερικού γινομένου (fully connected layers) της αρχιτεκτονικής C3D και συνεπακόλουθα προκαλείται μεγαλύτερη πιθανότητα υπερ-εκπαίδευσης (overfitting).

Συνεπώς, το ερώτημα που προκύπτει είναι πώς θα μπορούσαμε να διατηρήσουμε τις αρχικές πληροφορίες της σύνδεσης του χρόνου με το χώρο, μειώνοντας παράλληλα το μέγεθος στο δισδιάστατο χώρο. Ο πιο προφανής τρόπος είναι να διατηρήσουμε το βήμα της χρονικής υποδειγματοληψίας (pooling stride) στην τιμή ένα μόνο στη χρονική διάσταση. Παρ' όλα αυτά με τη μέθοδο αυτή, το χρονικό περιεχόμενο κάθε μονάδας για την πρόβλεψη κατηγορίας μειώνεται όπως αναφέρουν και τα άρθρα [13], [9]. Αυτό συμβαίνει διότι καθώς αυξάνεται η έξοδος στην χρονική διάσταση του συνελικτικού επιπέδου παράλληλα μειώνεται το μέγεθος του δεκτικού πεδίου κάθε μονάδας στην έξοδο. Επομένως, το άρθρο [16] προτείνει τα Temporal Preservation Convolutional (TPC) φίλτρα, ή αλλιώς τα συνελικτικά φίλτρα για την διατήρηση στο χρόνο για να αντικαταστήσουν τα αρχικά τρισδιάστατα συνελικτικά φίλτρα. Επίσης υπόσχεται μια μεγάλη αναβάθμιση

του "C3D" με αυτή τη μέθοδο αφήνοντας να χαθούν ελάχιστες πληροφορίες στη διάσταση του χρόνου για τον εντοπισμό της κίνησης. Αυτό επιτυγχάνεται με την προσθήκη του ρυθμού διαστολής στα επίπεδα σε συνδυασμό με τη διατήρηση του βήματος της χρονικής υποδειγματοληψίας στην τιμή ένα. Ο χρονικός ρυθμός διαστολής αναφέρεται στο άρθρο ως temporal atrous rate και αυξάνεται εκθετικά μέσα στα επίπεδα από το δύο εώς και το οκτώ.

Αν και το άρθρο [5] είναι εμπνευσμένο από το άρθρο [16] έχει κάποιες διαφωνίες σχετικά με την τιμή του ρυθμού διαστολής στο τελευταίο επίπεδο. Έτσι, αλλάζει την τιμή του στο "conv5" σε ένα, καθοδηγούμενο από το άρθρο [13] και το άρθρο [14] για την ορθότερη λειτουργία του. Χρησιμοποιεί επίσης το "Res3D" από το άρθρο [15] σε συνδυασμό με το χρονικό ρυθμό διαστολής και το ονομάζει "Temporal Dilated Res3D". Σε αυτή τη διπλωματική εργασία θα χρησιμοποιήσουμε το "C3D" από το άρθρο [9] και εμπνευσμένοι απο τα παραπάνω θα μεταποιήσουμε κάποια μεγέθη των φίλτρων και του ρυθμού διαστολής διαμορφώνοντας ένα άλλο "C3D" μοντέλο που παρουσιάζεται στον Πίνακα 3.1.

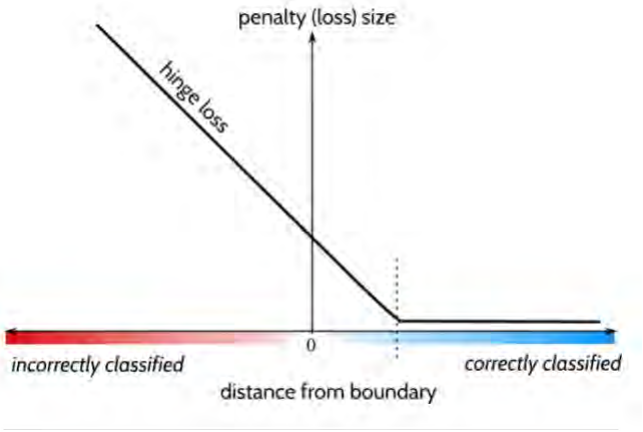
3.4 Ισορροπημένη Τετραγωνική Συνάρτηση Σφάλματος Hinge

Στα νευρωνικά δίκτυα, δυαδική ταξινόμιση (binary classification) ορίζεται η μέθοδος ταξινόμησης των δεδομένων σε δύο κατηγορίες. Συνήθως εφαρμόζεται σε προβλήματα όπως για παράδειγμα εάν ένα άτομο έχει περάσει ένα test, ένα email είναι μια απάτη, κλπ. Αν και φαίνεται να είναι ιδανικό για ένα πρόβλημα όπως η ταξινόμηση των στιγμιοτύπων σε κατηγορίες που θεωρούνται η αρχή ή το τέλος μίας χειρονομίας (δυαδική ταξινόμηση), υπάρχουν δυσκολίες. Στο σύνολο δεδομένων μας, κάθε ακολουθία χειρονομιών αποτελείται από στιγμιότυπα που περιλαμβάνουν κάποιες οριακές εικόνες της χειρονομίας και φυσικά την ίδια. Μόνο δύο ή τρία στιγμιότυπα μπορούν να θεωρηθούν ώς όρια της κίνησης και αυτός είναι ο λόγος για τον οποίο η δυαδική ταξινόμιση δεν είναι η ιδανική επιλογή. Συγκεκριμένα, τα περισσότερα στιγμιότυπα δε θα κατηγοριοποιηθούν ως όρια απο την δυαδική ταξινόμηση και το μοντέλο θα τα θεωρεί αρνητικά, έτσι η ακρίβεια του μοντέλου θα είναι υψηλή επειδή θα έχει υπολογίσει τα περισσότερα σωστά, ακόμα κι αν ορισμένα από τα οριακά δείγματα δεν έχουν προβλεφθεί σωστά. Αυτό είναι ανώφελο για το πρόβλημά μας, επειδή αυτός είναι ο κύριος σκοπός του δικτύου αναγνώρισης κίνησης, να δίνει διαφορετικό βάρος στις εικόνες που προβλέπει ορθά ως μή οριακές, και διαφορετικό σε αυτές που υποδηλώνουν την αρχή και το τέλος μιας χειρονομίας.

Επομένως, χρειαζόμαστε μια συνάρτηση που θα επικεντρωθεί στα λάθος ταξινομημένα στιγμιότυπα και δε θα επηρεαστεί από την πλειοψηφία των σωστά ταξινομημένων. Μια τέτοια συνάρτηση σφάλματος θα μπορούσε να είναι η "Hinge" που συμπεριφέρεται ως μια μηχανή διανυσμάτων υποστήριξης (SVM). Η συνάρτηση αυτή για μία προβλεπόμενη τιμή y' ορίζεται απο τη μαθηματική εξίσωση (3.2) και παρατηρούμε οτι αν η προβλεπόμενη τιμή έχει το ίδιο πρόσημο με την πραγματική τιμή, τότε η τιμή της συνάρτησης σφάλματος απλώς παραμένει μηδενική. Με άλλα λόγια, εάν είμαστε σωστά ταξινομημένοι εκτός περιθωρίου τότε δεν έχει σημασία

πόσο μακριά είμαστε από τη σωστή μεριά. Από την άλλη πλευρά, εάν η προβλεπόμενη και η πραγματική τιμή έχουν διαφορετική ένδειξη τότε η μέγιστη συνάρτηση άρα και η συνάρτηση κόστους θα έχουν ως έξοδο τη λάθος προβλεπόμενη τιμή συν ένα. Ήτοι, εάν είμαστε μέσα στο περιθώριο ή στην εσφαλμένη πλευρά, παίρνουμε μια ποινή άμεσα ανάλογη με το πόσο μακριά βρίσκομαστε στη λάθος πλευρά.

$$l(y, y') = \max(0, 1 - y * y') \tag{3.2}$$



Σχημα 3.2: Ισορροπημένη Τετραγωνική Συνάρτηση Σφάλματος Hinge. Εικόνα απο το άρθρο [22]

Η συνάρτηση σφάλματος στην εξίσωση (3.2) ορίζεται για μία πρόβλεψη. Για μία ακολουθία εικόνων που περιέχει πολλές προβλέψεις σε κάθε μία απο αυτές, ο μαθηματικός τύπος για τη συνάρτηση σφάλματος "Hinge" ορίζεται από την εξίσωση (3.3) όπου N ο αριθμός των εικόνων μιας χειρονομίας.

$$L = \frac{1}{N} * \sum_{i=1}^{\infty} \left(\max(0, 1 - y' * y) \right)^2 \tag{3.3}$$

Επιπρόσθετα, ο αριθμός των μη-οριακών στιγμιοτύπων είναι πολύ μεγαλύτερος από τα οριακά στιγμιότυπα. Έτσι, το άρθρο [5] πρότεινε μια νέα συνάρτηση κόστους που ονομάζεται "Ισορροπημένη Τετραγωνική Συνάρτηση Σφάλματος Hinge" (Balanced Squared Hinge Loss) και περιλαμβάνει συντελεστές αναλογίας για να διαχωρίζει με τη σωστή αναλογία τις δύο κλάσεις των στιγμιοτύπων. Η εξίσωση (3.4) περιγράφει με τον πρώτο όρο του αθροίσματος το μέσο σφάλμα για τις εικόνες που θεωρούνται όρια για την ακολουθία της χειρονομίας και με το δεύτερο όρο το μέσο σφάλμα για τις εικόνες που δε θεωρούνται όρια. Επιπλέον, r_b και r_n είναι οι συντελεστές αναλογίας για τα οριακά και μη-οριακά στιγμιότυπα αντίστοιχα.

$$L = r_b * \frac{1}{N} * \sum_{i=1}^{\infty} \left(\max(0, (1 - y') * y) \right)^2 + r_n * \frac{1}{N} * \sum_{i=1}^{\infty} \left(\max(0, (1 - y) * y') \right)^2 \tag{3.4}$$

Πινακας 3.1: C3D μοντέλο διπλωματικής και άρθρου [16]

Το μέγεθος εξόδου είναι της μορφής (αριθμός καναλιών x χρονικό μήκος x ύψος x μήκος). Τα φίλτρα χρησιμοποιούν τη μορφή (χρονικό μήκος x ύψος x μήκος, ρυθμός διαστολής) για τα επίπεδα συνέλιξης και (χρονικό μήκος x ύψος x μήκος, stride(χρονική επικάλυψη των φίλτρων , επικάλυψη των φίλτρων σε ύψος, επικάλυψη των φίλτρων σε μήκος)) για τα επίπεδα υποδειγματοληψίας.

	Μοντέλο απο το άρθρο [16]		C3D διπλωματικής	
Μέγεθος εισόδου 3 x L x 112 x 112				
Επίπεδα	Παράμετροι Εισόδου	Μέγεθος εξόδου	Παράμετροι Εισόδου	Μέγεθος εξόδου
conv1	3x3x3,1	64xLx112x112	3x7x7,1 stride(1,2,2)	64xLx56x56
pool1	3x2x2 stride(1,2,2)	64xLx56x56	1x2x2 stride(1,2,2)	64xLx28x28
conv2	3x3x3,1	128xLx56x56	1x1x1,1 3x3x3,2	128xLx28x28
pool2	3x3x2 stride(1,2,2)	128xLx28x28	1x2x2 stride(1,2,2)	128xLx14x14
conv3	3x3x3,2 3x3x3,2	256xLx28x28	1x1x1,1 3x3x3,4 3x3x3,4	256xLx14x14
pool3	3x3x2 stride(1,2,2)	256xLx14x14	1x2x2 stride(1,2,2)	256xLx7x7
conv4	3x3x3,4 3x3x3,4	512xLx14x14 512xLx14x14	1x1x1,1 stride(1,2,2) 3x3x3,8 3x3x3,8	512xLx4x4
pool4	3x2x2 stride(1,2,2)	512xLx7x7	1x2x2 stride(1,2,2)	512xLx2x2
conv5	3x3x3,8 3x3x3,8	512xLx7x7	3x3x3,1 3x3x3,1	512xLx2x2
pool5	3x3x2 stride(1,2,2)	512xLx4x4	1x2x2 stride(1,2,2)	512xLx1x1
avgpool	-	-	1x7x7 stride(1,7,7)	512xLx1x1
reshape	-	-	(512,)	Lx512
conv6/dense4	1x4x4,1	4096xLx1x1	1	1
conv7	1x1x1,1	4096xLx1x1	-	-
conv8	1x1x1,1	(K + 1)xLx1x1	-	-

Κεφάλαιο 4

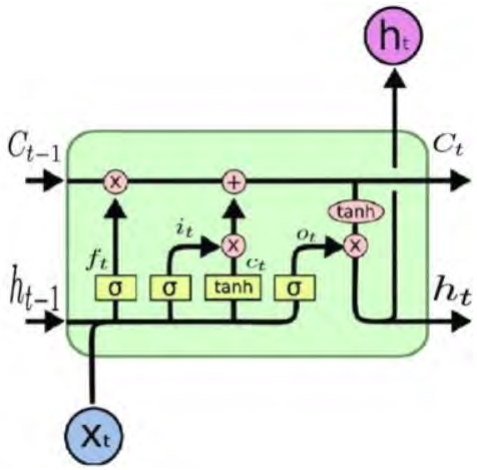
Συνελικτικά LSTM και διαχωρίσιμες συνελίξεις βάθους

Ο στόχος αυτού του κεφαλαίου είναι να παράσχει γενικό υλικό για δίκτυα βασισμένα σε ανατροφοδοτούμενα νευρωνικά δίκτυα (RNN) και την έννοια των εκτεταμένων στη διάσταση του βάθους διεσταλμένων συνελίξεων (depthwise dilated convolutions). Κάθε έκδοση των ανατροφοδοτούμενων νευρωνικών δικτύων έχει πλεονεκτήματα και μειονεκτήματα ανάλογα με το αντικείμενο της μελέτης. Συγκεκριμένα σε αυτή τη διπλωματική, θα χρησιμοποιήσουμε μια ειδική έκδοση των συνελικτικών LSTM (ConvLSTM) που προτείνεται στο άρθρο [6]. Η ενότητα 4.1 αναλύει την αναβάθμιση από τα βασικά ανατροφοδοτούμενα νευρωνικά δίκτυα στα LSTM δίκτυα. Η ενότητα 4.2 περιλαμβάνει τα πλεονεκτήματα των ConvLSTM σε σχέση με τα LSTM δίκτυα και επίσης το κατά πόσο το σχήμα της εισόδου επηρεάζει τις εσωτερικές συνελικτικές λειτουργίες. Η ενότητα 4.3 εξηγεί μέσω ενός παραδείγματος τα πλεονεκτήματα των διαχωρίσιμων συνελίξεων που εφαρμόζονται στο εσωτερικό κελί των ConvLSTM. Τέλος, αναλύουμε στην ενότητα 4.4 την ευρέως διαδεδομένη συνάρτηση κόστους με την ονομασία διασταυρούμενη εντροπία.

4.1 Η εξέλιξη των ανατροφοδοτούμενου νευρωνικού δικτύου

Το κύριο πλεονέκτημα των ανατροφοδοτούμενων νευρωνικών δικτύων (ΑΝΔ) έναντι των τεχνητών νευρωνικών δικτύων (ΤΝΔ) είναι ότι η πρώτη αρχιτεκτονική μπορεί να μοντελοποιήσει ακολουθία δεδομένων, για παράδειγμα χρονικών σειρών, έτσι ώστε κάθε δείγμα να μπορεί να εξαρτάται από τα προηγούμενα. Αντίθετα, τα ΤΝΔ δεν μπορούν να μοντελοποιήσουν μια ακολουθία δεδομένων. Επομένως τα ΤΝΔ είναι χρήσιμα μόνο όταν κάθε δείγμα θεωρείται αυτόνομο απο τα προηγούμενα και τα επόμενά του. Ωστόσο, τα ΑΝΔ αντιμετωπίζουν σοβαρά μειονεκτήματα για τρεις λόγους. Ο πρώτος είναι ότι δεν είναι σε θέση να επεξεργαστούν πολύ μεγάλες ακολουθίες εάν η συνάρτηση ενεργοποίησης είναι η "tanh". Ο δεύτερος ότι δεν μπορούν να χρησιμοποιηθούν σε πολύ βαθιές αρχιτεκτονικές εξαιτίας του κορεσμού των συνάρτησεων ενεργοποίησης που χρησιμοποιούνται σε ΑΝΔ. Τέλος, είναι επισφαλές όταν χρησιμοποιείται η "relu" ως συνάρτηση ενεργοποίησης.

Η αρχιτεκτονική Long Short-Term Memory (LSTM) που θεωρείται ένα ιδιαίτερο είδος ΑΝΔ, διαδόθηκε από τον Hochreiter Schmidhuber (1997) και χρησιμοποιήθηκε από πολλούς προγραμματιστές. Η βασική διαφορά της LSTM με τα ΑΝΔ είναι ότι η LSTM δεν αντιμετωπίζει δυσκολία σε βαθιές αρχιτεκτονικές επειδή εμπεριέχει ενα κομμάτι μνήμης το οποίο της δίνει τη δυνατότητα να κρατάει πληροφορίες για αρκετά βήματα και προφανώς να μοντελοποιεί μεγαλύτερου εύρους χρονικές ακολουθίες με υψηλότερη ακρίβεια από τα ΑΝΔ. Κατά τη διάρκεια της εκπαίδευσης του LSTM ένα σύνολο απο πύλες πραγματοποιεί κάποιες ενέργειες όπως την επιλογή του να θυμάται, να ξεχνάει ή να αγνοεί τις πληροφορίες και αυτές οι ενέργειες υπολογίζονται απο τις εξισώσεις (4.1), (4.2), (4.4) αντίστοιχα. Φυσικά οφείλουμε να αναφέρουμε ότι πρέπει να γίνουν περισσότεροι υπολογισμοί λόγω των σύνθετων συνάρτησεων ενεργοποίησης όπως επίσης να προστεθούν περισσότερα βάρη σε κάθε κόμβο. Στην εικόνα 4.1 παρουσιάζουμε το διάγραμμα της αρχιτεκτονικής LSTM και τις αντίστοιχες εξισώσεις στις (4.1), (4.2), (4.3), (4.4), (4.5).



Σχημα 4.1: Απεικόνιση της αρχιτεκτονικής LSTM. Εικόνα απο την ιστοσελίδα [23]

$$i_t = \sigma\big(W_{xi} \times x_t + W_{hi} \times h_{t-1} + W_{ci} \circ c_{t-1} + b_i\big) \tag{4.1}$$

$$f_t = \sigma\big(W_{xf} \times x_t + W_{hf} \times h_{t-1} + W_{cf} \circ c_{t-1} + b_f\big) \tag{4.2}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ tanh\big(W_{xc} \times x_t + W_{hc} \times h_{t-1} + b_c\big) \tag{4.3}$$

$$o_t = \sigma\big(W_{xo} \times x_t + W_{ho} \times h_{t-1} + W_{co} \circ c_t + b_o\big) \tag{4.4}$$

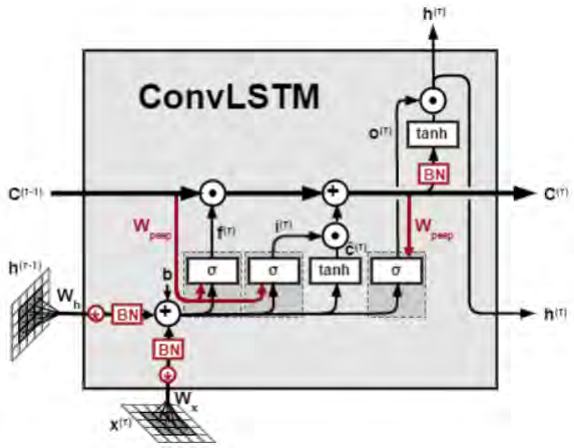
$$h_t = o_t \circ tanh\big(c_t\big) \tag{4.5}$$

4.2 Εισαγωγή στα Συνελικτικά LSTM Δίκτυα

Αν και με την πρώτη ματιά η LSTM αρχιτεκτονική φαίνεται ιδανική, αντιμετωπίζει ορισμένες ελλείψεις. Προκειμένου να επεξεργαστεί τις ακολουθίες εικόνων, τα χωρικά χαρακτηριστικά των δισδιάστατων συνελικτικών νευρωνικών δίκτυων διανυσματοποιούνται σε ένα γενικό πλαίσιο πριν παραδοθούν ως είσοδοι στο LSTM κελί. Ακόμη, οι χάρτες χωρικών χαρακτηριστικών που δημιουργούνται μέσα στο LSTM κελί δεν περιλαμβάνουν πληροφορίες σχετικά με την αλληλεπίδραση των χαρακτηριστικών στη χωρική διάσταση από τις αρχικές εικόνες. Ένα νέο μοντέλο έχει προταθεί από το άρθρο [7] για να επεξεργαστεί αρχικά τα διαδοχικά δεδομένα από βροχοπτώσεις, τα αυτο-οδηγούμενα αυτοκίνητα και ούτω καθεξής. Το όνομα αυτού του μοντέλου είναι "Συνελικτικό LSTM Δίκτυο" (ConvLSTM), και παρουσιάζεται στο Σχήμα 4.2 όπως επίσης και οι αντίστοιχες εξισώσεις του (4.6), (4.7), (4.7), (4.9), (4.10). Πρόκειται για μια παραλλαγή της αρχιτεκτονικής LSTM, καθώς περιέχει μία λειτουργία συνέλιξης στο εσωτερικό του LSTM κελιού. Τα συνελικτικά δίκτυα LSTM σχεδιάστηκαν για τρισδιάστατα δεδομένα στην είσοδο τους σε αντίθεση με την βασική LSTM αρχιτεκτονική η οποία δέχεται μονοδιάστατα δεδομένα. Για παράδειγμα, το AlexNet απο το άρθρο [24] και το VGG-16 απο το άρθρο [25], παράγουν τους χάρτες χωρικών χαρακτηριστικών που πρόκειται να τροφοδοτηθούν στην είσοδο του συνελικτικού δικτύου LSTM. Επίσης, η έξοδος ενός κελιού τρισδιάστατου συνελικτικού νευρωνικού δικτύου είναι ένας χωροχρονικός χάρτης χαρακτηριστικών που μπορεί να χρησιμοποιηθεί ως είσοδος στο συνελικτικό δίκτυο LSTM. Ωστόσο, ο σκοπός της δημιουργίας του ConvLSTM ήταν να ληφθούν οι αρχικές εικόνες ως είσοδοι για να γίνουν οι χωρικές συνελίξεις μέσα στο κελί. Έτσι, όταν η είσοδος έχει ήδη υποστεί επεξεργασία με χωρικές συνελίξεις από τα AlexNet, VGG ή 3DCNN, η ερώτηση που προκύπτει είναι κατά πόσο θα μπορούσαν να επικεντρωθούν οι εσωτερικές συνελικτικές δομές του ConvLSTM στην αρχική σύνδεση των χωροχρονικών χαρακτηριστικών.

Οι εσωτερικές συνελικτικές δομές του ConvLSTM κελιού, συμβάλλουν στην εξαγωγή χαρακτηριστικών ανάλογα με το είδος της εισόδου που δίνεται. Εδώ, παρουσιάζουμε τρεις τρόπους για την παροχή δεδομένων εισόδου στο συνελικτικό δίκτυο LSTM. Το πρώτο είναι να τροφοδοτήσουμε τις αρχικές εικόνες ως είσοδο. Σε αυτή την περίπτωση, το συνελικτικό δίκτυο LSTM αποκτά με επιτυχία τα χωροχρονικά χαρακτηριστικά λόγω της έμφυτης συνελικτικής δομής του, όπως αναφέρει και το άρθρο [7]. Το δεύτερο είναι να τροφοδοτήσει το συνελικτικό LSTM κελί με τον χάρτη χαρακτηριστικών ενός δισδιάστατου συνελικτικού νευρωνικού δικτύου. Όπως αναφέρθηκε προηγουμένως, αυτό δεν είναι τόσο ωφέλιμο, διότι οι συνελικτικές δομές του κελιού δε θα έχουν σημαντική επίδραση στην έξοδο, και αυτό επιφέρει ανώφελη και επιπρόσθετη υπολογιστική ισχύ. Η τρίτη είναι η λήψη του χάρτη χαρακτηριστικών ενός τρισδιάστατου συνελικτικού νευρωνικού δικτύου. Το τρισδιάστατο συνελικτικό νευρωνικό δίκτυο έχει εκπαιδευτεί και έχει μάθει τα χωροχρονικά χαρακτηριστικά. Αυτή η μέθοδος προτείνεται από το άρθρο [6].

$$i_t = \sigma\big(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i\big) \tag{4.6}$$



Σχημα 4.2: Απεικόνιση της αρχιτεκτονικής ConvLSTM. Εικόνα απο την ιστοσελίδα [23]

$$f_t = \sigma\big(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f\big) \tag{4.7}$$

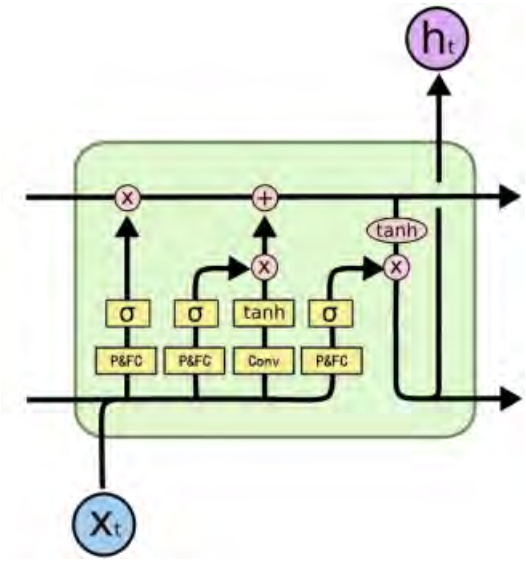
$$c_t = f_t \circ c_{t-1} + i_t \circ tanh\big(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c\big) \tag{4.8}$$

$$o_t = \sigma\big(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ c_t + b_o\big) \tag{4.9}$$

$$h_t = o_t \circ tanh\big(c_t\big) \tag{4.10}$$

Η ιδέα της μείωσης του αριθμού των παραμέτρων και των υπολογισμών είναι πολύ ελκυστική. Στο άρθρο [6] παρουσιάστηκαν τέσσερις μέθοδοι για να επωφεληθούμε από τις συνελικτικές δομές όπου ήταν απαραίτητες. Ορισμένες από αυτές τις μεθόδους χρησιμοποιούν μηχανισμούς προσοχής (attention mechanisms). Οι μηχανισμοί προσοχής επιτρέπουν στο μοντέλο όπως αναφέρει και το όνομα, να εστιάζουν σε κρίσιμα κομμάτια των χαρακτηριστικών. Σε αυτή τη διπλωματική, η μέθοδος των μηχανισμών προσοχής μέσα στο συνελικτικό LSTM κελί δε θα εφαρμοστεί, επειδή οι αλλαγές του μοντέλου μέσω των διεσταλμένων συνελίξεων είναι αρκετές για να παράγουν ικανοποιητικά αποτελέσματα. Συγκεκριμένα, στη μέθοδο που αναφέρεται πρώτη στο άρθρο, οι τιμές των πυλών (4.13),(4.14),(4.15) υπολογίζονται για κάθε χάρτη χαρακτηριστικών και οχι για κάθε στοιχείο. Ακόμη, με την εκτέλεση της υποδειγματοληψίας χαρακτηριστικών εισόδου, έχουμε αποτέλεσμα τη μείωση της χωρικής διάστασης στις κρυφές καταστάσεις του ConvLSTM και την είσοδο. Έτσι, αναφερόμενοι στο υπολογιστικό κόστος, οι λειτουργίες έχουν μετατραπεί από συνελικτικές σε πράξεις εσωτερικού γινομένου, οι οποίες είναι πολύ λιγότερες σε υπολογισμούς. Διατηρούνται μόνο οι συνελικτικές δομές για την είσοδο στη μεταβατική κατάσταση, οι οποίες μετατρέπονται επίσης σε διαχωρίσιμες συνελίξεις στην διάσταση του βάθους μειώνοντας τον αριθμό των παραμέτρων. Η αναπαράσταση του μοντέλου που θα επιλέξουμε ως μέρος του δικτύου για την αναγνώριση χειρονομίας είναι στο Σχήμα 4.3 και οι αντίστοιχες εξισώσεις στις

(4.11), (4.12), (4.13), (4.14), (4.15), (4.16), (4.17), (4.18).



Σχημα 4.3: Τροποποίηση της αρχιτεκτονικής ConvLSTM. Εικόνα απο το άρθρο [6]

$$\overline{X_t} = GlobalAveragePooling\big(X_t\big) \tag{4.11}$$

$$\overline{H}_{t-1} = GlobalAveragePooling\big(H_{t-1}\big) \tag{4.12}$$

$$i_t = \sigma\big(W_{xi} \times \overline{X_t} + W_{hi} \times \overline{H}_{t-1} + b_i\big) \tag{4.13}$$

$$f_t = \sigma\big(W_{xf} \times \overline{X_t} + W_{hf} \times \overline{H}_{t-1} + b_f\big) \tag{4.14}$$

$$o_t = \sigma\big(W_{xo} \times \overline{X_t} + W_{ho} \times \overline{H}_{t-1} + b_o\big) \tag{4.15}$$

$$G_t = tanh\big(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c\big) \tag{4.16}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ G_t \tag{4.17}$$

$$H_t = o_t \circ tanh\big(C_t\big) \tag{4.18}$$

4.3 Διαχωρίσιμες συνελίξεις βάθους

Οι ξεχωριστές συνελίξεις βάθους αποτελούνται απο δύο τμήματα: τη συνέλιξη στη διάσταση του βάθους και τη σημειακή συνέλιξη. Αρχικά, η συνέλιξη στη διάσταση του βάθους εφαρμόζεται στα δεδομένα εισόδου. Σε αντίθεση με τις βασικές συνελίξεις, τρία φίλτρα κινούνται μέσα στην εικόνα. Κάθε φίλτρο αντιστοιχεί σε κάθε κανάλι της εικόνας και κινείται μόνο σε αυτό. Η έξοδος των τριών φίλτρων που συμπίπτουν με τα τρία κανάλια είναι το ίδιο μήκος-ύψος που θα ήταν με τη βασική συνέλιξη, αλλά το βάθος είναι ακόμα το ίδιο, δηλαδή τρισδιάστατο, σε αντίθεση με τη βασική συνέλιξη που θα ήταν μονοδιάστατο. Έτσι, αν έχουμε μια εικόνα μεγέθους 28x28x3 εικονοστοιχείων ή αλλιώς μια εικόνα RBG 28x28 και 128 φίλτρα μεγέθους 5x5x3 στη βασική συνέλιξη με μηδενικό γέμισμα (padding) και βήμα (stride) ίσο με ένα θα πραγματοποιηθούν

$$\big(28-5+1\big)\times\big(28-5+1\big)\times 5\times 5\times 3\times 128 = 5.529.600 \tag{4.19}$$

πολλαπλασιασμοί σε έξοδο 128 καναλιών. Ο γενικός τύπος για τον υπολογισμό των πολλαπλασιασμών φαίνεται στην εξίσωση (4.22). Τις περισσότερες φορές θέλουμε να εξαγάγουμε περισσότερα κανάλια όπως 256, 512, 1024, κλπ. Τι συμβαίνει με τη συνέλιξη στη διάσταση του βάθους; Υπάρχουν τρια διαχωρίσιμα φίλτρα για την εικόνα που κινούνται 24 x 24 φορές. Αυτοί είναι οι πολλαπλασιασμοί του πρώτου μέρους της συνέλιξης στη διάσταση του βάθους:

$$3\times 8\times 8\times 24\times 24 = 110.592 \tag{4.20}$$

και ο συνολικός αριθμός πολλαπλασιασμού είναι στην εξίσωση (4.23). Το δεύτερο βήμα είναι η σημειακή συνέλιξη. Μετά το πρώτο βήμα, έχουμε τρία κανάλια μιας δισδιάστατης εικόνας 24x24. Τώρα, πρέπει να αυξήσουμε τον αριθμό των καναλιών για κάθε εικόνα. Το φίλτρο αυτής της συνέλιξης έχει σχήμα 1 x 1 και για αυτό έχει το όνομα η συνέλιξη ως σημειακή. Το βάθος εξαρτάται από τον αριθμό των καναλιών της εισόδου. Έτσι, δημιουργούμε τον αριθμό των φίλτρων σύμφωνα με τα επιθυμητά κανάλια εξόδου. Οι πολλαπλασιασμοί είναι

$$128\times 1\times 1\times 3\times 24\times 24 = 221.184 \tag{4.21}$$

Ο γενικός τύπος για τον πολλαπλασιασμό της σημειακής συνέλιξης είναι στην εξίσωση (4.24). Το σύνολο των πολλαπλασιασμών είναι 221.184 + 110.592 = 331.776 το οποίο είναι αρκετά μικρότερο σε σύγκριση με τη βασική συνέλιξη που ισούται με 5.529.600.

$$\text{normal} = Nc\times h\times h\times D\times\big(H-h+1\big)\times\big(W-h+1\big) \tag{4.22}$$

$$\text{depthwise} = D\times h\times h\times 1\times\big(H-h+1\big)\times\big(W-h+1\big) \tag{4.23}$$

$$\text{pointwise} = Nc\times 1\times 1\times D\times\big(H-h+1\big)\times\big(W-h+1\big) \tag{4.24}$$

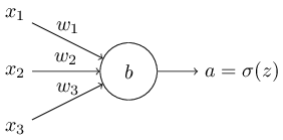
$$\text{depthwise separable} = Nc\times 1\times 1\times D\times\big(H-h+1\big)\times\big(W-h+1\big) \tag{4.25}$$

Γενικά, το παραπάνω παράδειγμα εξηγεί ότι με τη βασική συνέλιξη η αρχική εικόνα ανακατασκευάζεται 128 φορές ενώ με τη συνέλιξη στη διάσταση του βάθους ανασυνθέτουμε την εικόνα εισόδου μια φορά. Στη συνέχεια, η εικόνα που είναι μικρότερη επεκτείνεται σε 128 κανάλια. Φυσικά, θα πρέπει να εξετάσουμε πόσο πολύπλοκο είναι το δίκτυο που θέλουμε να δημιουργήσουμε, επειδή η μείωση των παραμέτρων μετατρέπεται σε ένα δίκτυο πολύ πιο απλό. Στην περίπτωσή μας, η ξεχωριστή συνέλιξη στη διάσταση του βάθους αποτελεί μέρος του δικτύου μας "αναγνώριση χειρονομίας" που εμπεριέχει πολλά στρώματα και επίσης υλοποιείται μέσα στο συνελικτικό LSTM δίκτυο για την απλοποίηση του κελιού και τη μείωση των παραμέτρων.

4.4 Συνάρτηση κόστους διασταυρούμενης εντροπίας

Κατά τη διάρκεια της εκπαίδευσης, στα βαθιά ΝΔ προκύπτουν κάποιες δυσκολίες που αφορούν την απόδοσή τους σε σχέση με τα απλά. Αντί να εγκαταλείψουμε τα βαθιά ΝΔ θα προσπαθήσουμε βάσει του άρθρου [11] να αναλύσουμε τα αίτια της δυσκολίας της απόδοσης του. Γενικότερα, τα επίπεδα σε ένα ΝΔ δεν εκπαιδεύονται με την ίδια ταχύτητα σε όλο το εύρος τους. Συγκεκριμένα, τα επίπεδα που είναι πιο κοντά στην έξοδο μαθαίνουν καλύτερα ενώ τα αρχικά επίπεδα σχεδόν καθόλου. Αυτό το φαινόμενο ονομάζεται επιβράδυνση της μάθησης και οφείλεται σε εξαιρέσεις περιπτώσεων ή αλλιώς πολλές φορές ορίζεται και ώς "κακή τύχη" γιατί εξαρτάται απο την τυχαία αρχικοποίηση των βαρών.

Σε αυτό το πρόβλημα συμβάλλει σε μεγάλο βαθμό και η συνάρτηση σφάλματος. Στην ενότητα αυτή αναλύουμε τη συνάρτηση η οποία αποτρέπει το φαινόμενο της επιβράδυνσης της μάθησης με το παράδειγμα ενός νευρώνα που απεικονίζεται στο Σχήμα 4.4 και ονομάζεται συνάρτηση κόστους διασταυρούμενης εντροπίας (cross-entropy) . Η έξοδος του αθροιστή είναι στην σχέση (2.2) που αναλύσαμε σε προηγούμενη ενότητα. Η έξοδος αυτή εισέρχεται στη συνάρτηση ενεργοποίησης ως έξοδος του νευρώνα με το συμβολισμό α. Η συνάρτηση κόστους διασταυρούμενης εντροπίας είναι στην εξίσωση (4.26).



Σχημα 4.4: Σχηματική αναπαράσταση ενός απλού νευρώνα. Εικόνα απο την ιστοσελίδα [10]

$$C = -\frac{1}{N} * \sum_{x} \big(ylna + \big(1-y\big)ln\big(1-\alpha\big)\big) \tag{4.26}$$

Όπου N είναι ο συνολικός αριθμός των δεδομένων εκπαίδευσης εισόδου x, όπου y τα επιθυμητά αποτελέσματα αυτών. Δεν είναι ξεκάθαρο πως αυτή η συνάρτηση

είναι συνάρτηση κόστους και πως βοηθάει στο φαινόμενο που προαναφέραμε. Πριν προχωρήσουμε παρακάτω ας σταθούμε λίγο στη συνάρτηση αυτή. Δύο είναι τα κριτήρια που κάνουν την εξίσωση (4.26) συνάρτηση κόστους. Πρώτο, δεν είναι αρνητική συνάρτηση επειδή όλοι οι όροι της εξίσωσης είναι αρνητικοί αφού οι τιμές των λογαρίθμων κινούνται στο διάστημα $0, 1$ και υπάρχει αρνητικό πρόσημο στην αρχή του αθροίσματος. Δεύτερο, όταν η τιμή εξόδου του νευρώνα είναι κοντά στην τιμή αληθείας, τότε η συνάρτηση κόστους (4.26) είναι κοντά στο μηδέν. Οι δύο αυτές προϋποθέσεις πληρούνται και απο την τετραγωνική συνάρτηση κόστους, με τη διαφορά ότι η συνάρτηση κόστους διασταυρούμενης εντροπίας αποτρέπει την επιβράδυνση της μάθησης. Για να καταλάβουμε με ποιο τρόπο το επιτυγχάνει αυτό, θα επεξεργαστούμε τη συνάρτηση σε βάθος παραγωγίζοντάς τη σε σχέση με τα βάρη και θα τη συμβολίζουμε με $\sigma(z)$.

$$\begin{aligned}\frac{\partial C}{\partial w_j} &= -\frac{1}{N}\sum_x \left[\frac{y}{\sigma(z)} + \frac{(1-y)}{1-\sigma(z)}\right]\frac{\partial \sigma}{\partial w_j} \\ &= -\frac{1}{N}\sum_x \left[\frac{y}{\sigma(z)} + \frac{(1-y)}{1-\sigma(z)}\right]\sigma'(z)x_j \\ &= -\frac{1}{N}\sum_x \left[\frac{y(1-\sigma(z))}{\sigma(z)(1-\sigma(z))} + \frac{(1-y)\sigma(z)}{(1-\sigma(z))\sigma(z)}\right]\sigma'(z)x_j\end{aligned} \tag{4.27}$$

Θεωρώντας ότι πρώτον, οι όροι σ' και $(1-\sigma(z))\sigma(z)$ ακυρώνονται μεταξύ τους και δεύτερον, η σιγμοειδής συνάρτηση ενεργοποίησης είναι στην εξίσωση (2.3), μετά απο παραγοντοποιήσεις και απλοποιήσεις της (4.27) έχουμε αυτό το αποτέλεσμα:

$$\frac{1}{N}\sum_x x_i(\sigma(z) - y) \tag{4.28}$$

Η συνάρτηση αυτή δηλώνει ότι ο ρυθμός με τον οποίο μαθαίνει το βάρος ελέγχεται από την $(\sigma(z) - y)$ δηλαδή απο την τιμή λάθους στην έξοδο. Όσο μεγαλύτερο είναι το λάθος, τόσο πιο γρήγορα μαθαίνει ο νευρώνας. Σε αυτή την εξίσωση ο όρος σ' έχει απαληφθεί οπότε δε συμβάλλει στο πρόβλημα της επιβράδυνσης μάθησης. Περισσότερες λεπτομέρειες βρίσκονται στο άρθρο [11].

Κεφάλαιο 5

Δίκτυο αναγνώρισης χειρονομίας

Στο προηγούμενο Κεφάλαιο αναλύσαμε ένα κομμάτι από το δίκτυο αναγνώρισης χειρονομίας το ConvLSTM. Σε αυτό το Κεφάλαιο θα παρουσιάσουμε το υπόλοιπο κομμάτι του δικτύου. Συγκεκριμένα ξεκινάμε με την ενότητα 5.1 που περιέχει την αρχιτεκτονική και τη συνεισφορά των Mobilenets στα νευρωνικά δίκτυα και το λόγο που θα χρησιμοποιήσουμε ένα κομμάτι από αυτά στη διπλωματική. Στην ενότητα 5.2 αναφέρουμε το συνδυασμό των αρχιτεκτονικών που θα χρησιμοποιήσουμε στο δίκτυο αναγνώρισης χειρονομίας παρουσιάζοντας σε πίνακες το κομμάτι από το C3D μοντέλο της διπλωματικής στον Πίνακα 5.1 και το κομμάτι από τα Mobilenets στον Πίνακα 5.2.

5.1 Εισαγωγή στα Mobilenets

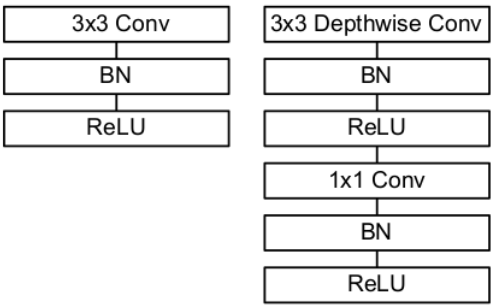
Απο τότε που το AlexNet κέρδισε το διαγωνισμό "ImageNet: ILSVRC 2012" στο άρθρο [26], πολλές μελέτες έχουν γίνει πάνω στην ανάπτυξη βαθύτερων και περιπλοκότερων δικτύων έτσι ώστε να επιτευχθεί όλο και υψηλότερη ακρίβεια σε αυτά. Παρόλα αυτά οι βελτιώσεις αυτές δεν καθιστούν απαραίτητα τα δίκτυα αυτά πιο αποτελεσματικά σε σχέση με το μέγεθος και την ταχύτητα που τρέχουν.

Ο χρόνος απόκρισης του ΝΔ, ο περιορισμένος χώρος εγκατάστασής του (ειδικά στα μοντέλα ρομποτικής όπως αυτο-οδηγούμενα αυτοκίνητα κλπ), οι υψηλές ανάγκες υπολογιστικής ισχύος και η μνήμη που απαιτείται είναι σημαντικοί παράγοντες που πρέπει να ληφθούν υπόψη και καθιστούν την εφαρμογή τους στα μοντέλα εξαιρετικά δύσκολη.

Πολλές μελέτες επικεντρώνονται περισσότερο στο μέγεθος των μοντέλων και όχι τόσο στο χρόνο απόκρισής τους. Το άρθρο [27] που αναλύει τα MobileNets υπόσχεται μια κλάση αρχιτεκτονικής η οποία επιτρέπει σε κάθε προγραμματιστή να προσαρμόζει κατάλληλα το μέγεθος σε σχέση με την ταχύτητα του μοντέλου που θέλει να αναπτύξει. Τα MobileNets επικεντρώνονται κυρίως στη βελτίωση της καθυστέρησης μικρών ΝΔ που απευθύνονται σε εφαρμογές που τρέχουν σε μικρές κινητές (mobile) συσκευές και ενσωματωμένα συστήματα (embedded systems). Αποτελούνται κυρίως απο επίπεδα ξεχωριστών συνελίξεων στη διάσταση του βάθους (που αναπτύξαμε στην ενότητα 4.3) εκτός απο το πρώτο επίπεδο που υλοποιείται με πλήρως συνδεδεμένα επίπεδα. Όλα τα συνελικτικά επίπεδα ακολουθούνται απο ένα επίπεδο κανονικοποίησης παρτίδας (το οποίο θα αναλύσουμε στο Κεφάλαιο 7) και απο ένα επίπεδο μονάδας γραμμικής ανόρθωσης με εξαίρεση το

τελευταίο επίπεδο το οποίο είναι γραμμικό. Το επόμενο είναι ένα επίπεδο τυποποιημένης εκθετικής συνάρτησης για την πρόβλεψη της κλάσης.

Σε σύνολο στο MobileNet υπάρχουν 28 επίπεδα αλλά θα χρησιμοποιήσουμε μόνο κάποια απο αυτά για να τα προσαρμόσουμε στο δίκτυο αναγνώρισης χειρονομίας που αναλύουμε στην επόμενη ενότητα. Στο Σχήμα 5.1 συγκρίνεται στα αριστερά του σχήματος το βασικό συνελικτικό επίπεδο μαζί με το επίπεδο κανονικοποίησης παρτίδας και τη μή-γραμμική συνάρτηση μονάδας γραμμικής ανόρθωσης και στα δεξιά ένα επίπεδο συνέλιξης στη διάσταση του βάθους, 1x1 φίλτρα σημειακής συνέλιξης και επίσης ένα επίπεδο κανονικοποίησης παρτίδας και μονάδα γραμμικής ανόρθωσης μετά απο κάθε ένα απο τα δύο συνελικτικά επίπεδα.



Σχημα 5.1: Κελί βασικής συνέλιξης και κελί ξεχωριστής συνέλιξης στη διάσταση του βάθους. Εικόνα από [27]

5.2 Συνδιασμός αρχιτεκτονικών

Μετά το δίκτυο χρονικής κατάτμησης που αναλύσαμε στο Κεφάλαιο 3 το αντικείμενο της μελέτης μας μετατρέπεται απο αναγνώριση ενός βίντεο με πολλαπλές χειρονομίες σε αναγνώριση πολλών βίντεο με μεμονωμένες χειρονομίες. Συνεπώς μπορεί να εφαρμοστεί οποιοδήποτε ΝΔ που εφαρμόζεται σε αναγνώριση χειρονομιών.

Στόχος των πρώτων επιπέδων του δίκτυου αναγνώρισης αυτής της διπλωματικής είναι να μάθει τα χωροχρονικά χαρακτηριστικά ταυτόχρονα για να είναι πιο αποδοτικό. Το μοντέλο συνελικτικών επιπέδων με την ονομασία "Two-Stream" απο το άρθρο [30] εξάγει χρονικά και χωρικά χαρακτηριστικά ξεχωριστά, με εισόδους έγχρωμα και στοιβαγμένης οπτικής ροής δεδομένα. Το μοντέλο αρχιτεκτονικής ΑΝΔ που αναφέρεται στο άρθρο [31] μαθαίνει πρώτα τα χωρικά χαρακτηριστικά απο κάθε εικόνα και μετά τα χρονικά βάσει των ακολουθιών στη διάσταση του χώρου χρησιμοποιώντας τη βασική δομή των ΑΝΔ. Ενα ακόμη μοντέλο είναι το "VideoLSTM" απο το άρθρο [32] που χρησιμοποιεί συνελικτικά LSTM

για να μάθει χωροχρονικά χαρακτηριστικά απο δισδιάστατους χάρτες χαρακτηριστικών. Αυτές οι τρεις μέθοδοι μαθαίνουν τα χωροχρονικά χαρακτηριστικά ξεχωριστά ή σε διαφορετικά στάδια. Γι' αυτό το λόγο σαν πρώτα επίπεδα στο δίκτυο αναγνώρισης χειρονομίας επιλέγουμε τρισδιάστατα συνελικτικά επίπεδα που μαθαίνουν τα χαρακτηριστικά των δυο διαστάσεων ταυτόχρονα.

Στο δίκτυο αυτό συνδυάζουμε απο κάποιες μεθόδους που αναφέραμε προηγουμένως (στο άρθρο [16], στο Κεφάλαιο 3, στην ενότητα 4.2 και 5.1) διάφορα επίπεδα με σκοπό να αναγνωρίσουμε ορθά την κλάση της χειρονομίας.

Για το πρώτο κομμάτι του δικτύου αναγνώρισης χειρονομίας δανειζόμαστε ένα κομμάτι απο το C3D μοντέλο της διπλωματικής αυτής απο τον Πίνακα 3.1. Αλλάζουμε την τιμή του ρυθμού διαστολής στα κελιά 2,3,4 και την ποσότητα των επιπέδων όπως φαίνεται και στον Πίνακα 5.1, για να εκπαιδεύσουμε το ΝΔ στα βραχυπρόθεσμα χωροχρονικά χαρακτηριστικά.

Πινακας 5.1: Κομμάτι απο τον Πίνακα 3.1

Το μέγεθος εξόδου είναι της μορφής (αριθμός καναλιών x χρονικό μήκος x ύψος x μήκος). Τα φίλτρα χρησιμοποιούν τη μορφή (χρονικό μήκος x ύψος x μήκος, ρυθμός διαστολής) για τα επίπεδα συνέλιξης και (χρονικό μήκος x ύψος x μήκος, stride(χρονική επικάλυψη των φίλτρων , επικάλυψη των φίλτρων σε ύψος, επικάλυψη των φίλτρων σε μήκος)) για τα επίπεδα υποδειγματοληψίας.

Συνδυασμός αρχιτεκτονικών		
Είσοδος	βίντεο μεγέθους 3 x L x 112 x 112	
Επίπεδα	Τιμές παραμέτρων επιπέδων	Μέγεθος εξόδου
conv1	3x7x7,1 stride(1,2,2)	64xLx56x56
conv2	1x1x1,1 4 x [3x3x3,1]	64xLx56x56
pool2	1x2x2 stride(1,1,1)	64xLx56x56
conv3	1x1x1,1 stride(2,2,2) 4 x [3x3x3,1]	128x(L/2)x28x28
pool3	1x2x2 stride(1,1,1)	128x(L/2)x28x28
conv4	1x1x1,1 4 x [3x3x3,1]	256x(L/2)x28x28

Για το δεύτερο κομμάτι του δικτύου αναγνώρισης χειρονομίας χρησιμοποιούμε το συνελικτικό LSTM που αναλύσαμε στην ενότητα 4.2 όπου αναφέραμε ένα εναλλακτικό τρόπο για να χρησιμοποιήσουμε όσο το δυνατόν λιγότερες πράξεις συνέλιξης, η μέθοδος εικονίζεται στο Σχήμα 4.3. Το επίπεδο που χρησιμοποιεί αυτή την εσωτερική δομή είναι σχεδιασμένο έτσι ώστε κατά τη διάρκεια της εκπαίδευσης να μαθαίνει τους βραχυπρόθεσμους χωροχρονικούς χάρτες χαρακτηριστικών. Ο λόγος που χρησιμοποιήσαμε ενα μεταποιημένο και κυρίως συνελικτικό LSTM αντί του βασικού LSTM είναι για να κρατήσουμε τις πληροφορίες χωρικών συσχετίσεων όταν συγχωνεύουμε τα χρονικά χαρακτηριστικά μαζί με τα βήματα της ανατροφοδοτούμενης αρχιτεκτονικής. Θέτουμε την τιμή της επικάλυψης των φίλτρων σε (1,1) που υποδηλώνει ότι δεν αλλάζει τη χωρική διάσταση. Σε κάθε βήμα του ConvLSTM εξάγονται δισδιάστατοι χωροχρονικοί χάρτες χαρακτηριστικών που συγχωνεύονται με τους χάρτες των προηγούμενων εικόνων. Έτσι, τα αρχικά βίντεο μετατρέπονται σε δισδιάστατους χάρτες χαρακτηριστικών για να γίνει μια πιο βαθιά μελέτη σε αυτούς.

Για το τρίτο κομμάτι του δικτύου αναγνώρισης χειρονομίας χρησιμοποιούμε ένα κομμάτι απο τα MobileNets που παρουσιάζουμε στον Πίνακα 5.2. Αφού εφαρμόσουμε το κομμάτι ΝΔ που παρουσιάζεται στον Πίνακα 5.1 και το επίπεδο με εσωτερική δομή του Σχήματος 4.3 παρατηρούμε οτι η έξοδός τους είναι αρκετά μεγάλη στη χωρική διάσταση. Για ένα τόσο περίπλοκο σύνολο δεδομένων με μέγεθος 20 κλάσεων το μοντέλο μας ως τώρα είναι αρκετά απλό. Ανασχηματίζουμε την έξοδο στη διάσταση $28, 28, 256$ και την εισάγουμε σε ένα δισδιάστατο συνελικτικό μοντέλο για να εκπαιδεύσουμε το δίκτυο σε βαθύτερα χωροχρονικά χαρακτηριστικά. Αναλύσαμε ήδη τα πλεονεκτήματα των διαχωρίσιμων συνελίξεων στη διάσταση του βάθους στην ενότητα 4.3 και για αυτό το λόγο χρησιμοποιούμε ενα κομμάτι απο το MobileNet που αναφέραμε στην προηγούμενη ενότητα που πληρεί τις προδιαγραφές βαθύτερης εκμάθησης.

Πινακας 5.2: Επιλεγμένα επίπεδα απο το MobileNet

Το μέγεθος εξόδου είναι της μορφής (αριθμός καναλιών x ύψος x μήκος), τα φίλτρα της συνέλιξης στη διάσταση του βάθους χρησιμοποιούν τη μορφή (αριθμός καναλιών εισόδου x ύψος x μήκος) και τα φίλτρα της σημειακής συνέλιξης χρησιμοποιούν τη μορφή (αριθμός καναλιών εισόδου x αριθμός καναλιών εξόδου x ύψος x μήκος) και η μορφή για την είσοδο της σημειακής συνέλιξης είναι (αριθμός καναλιών εξόδου x ύψος x μήκος)

Είσοδος	256 x 28 x 28		
Επίπεδα	Τύπος Συνέλιξης	Τιμές παραμέτρων επιπέδων	Μέγεθος εξόδου
sepconv1	σ.β. σ.σ.	256x3x3,s1 256x256x1x1,s1	256x28x28
sepconv2	σ.β. σ.σ.	256x3x3,s2 256x512x1x1,s1	256xLx14x14 512xLx14x14
sepconv3	σ.β. σ.σ.	[512x3x3,s1]x5 [512x512x1x1,s1]x5	512x14x14
sepconv4	σ.β. σ.σ.	512x3x3,s2 512x1024x1x1,s1	512x7x7 1024x7x7
sepconv5	σ.β. σ.σ.	1024x3x3,s2 1024x1024x1x1,s1	1024x4x4

Κεφάλαιο 6

Προεπεξεργασία δεδομένων

Σε αυτό το Κεφάλαιο παρουσιάζουμε τα δεδομένα που θα χρησιμοποιήσουμε και την προεπεξεργασία τους για να τα εισάγουμε στα δύο νευρωνικά δίκτυα. Στην ενότητα 6.1 αναλύουμε τη δομή και τις πηγές των συνόλων δεδομένων που θα χρησιμοποιήσουμε. Στην ενότητα 6.2 αναφέρουμε μια εναλλακτική μέθοδο που έχουμε πρόσβαση στα βίντεο χωρίς να τα αποθηκεύουμε σε εικόνες. Στην ενότητα 6.3 παρουσιάζουμε στον Πίνακα 6.3 τα αποτελέσματα της στρατηγικής "jitter" για την εισαγωγή εικόνων στο δίκτυο χρονικής κατάτμησης και το λόγο αποτυχίας της στα σύνολα δεδομένων δοκιμής. Στην ενότητα 6.4 παρουσιάζουμε την μέθοδο που χρησιμοποιούμε στη διπλωματική για να εισάγουμε τις εικόνες στο δίκτυο χρονικής κατάτμησης και παρουσιάζουμε τους αλγορίθμους με ψευδοκώδικα. Αφού αναλύσαμε την μέθοδο επεξεργασίας παρουσιάζουμε στην ενότητα 6.5 την επισήμανση των ορίων στο νέο μέγεθος τους.

6.1 Βάσεις δεδομένων

Το σύνολο δεδομένων που θα χρησιμοποιήσουμε στα δύο δίκτυα είναι το "Multimodal Gesture Recognition: Montalbano V2 (ECCV '14)" και βρίσκεται στην ιστοσελίδα [28] με την ονομασία "ChaLearn Looking at People". Περιέχει πολλές κατηγορίες απο σύνολα δεδομένων που αφορούν τη μελέτη και την ανάλυση των ανθρώπινων κινήσεων σε βίντεο. Κάποια από αυτά ειδικεύονται στην αναγνώριση συγκεκριμένων μελών του ανθρώπινου σώματος. Επίσης, εξάγονται σε διάφορες μορφές όπως ήχου, βίντεο με βάση το μοντέλο RGB που διαθέτει τρία κανάλια πληροφορίας, βίντεο βάθους και καταγραφή της κίνησης του ανθρώπινου σκελετού σε κάθε χρονική περίοδο. Συγκεκριμένα στο δικό μας μοντέλο εκτελούνται 14.000 χειρονομίες από διάφορους ανθρώπους που στέκονται σε διαφορετικό φόντο σε κάθε βίντεο με σκοπό να επικεντρωθεί το δίκτυο στην κίνηση που εκτελείται. Ακόμη, υπάρχουν 20 διαφορετικές κλάσεις και το λεξιλόγιο χειρονομιών που εκτελείται απο τους ανθρώπους στα βίντεο παρουσιάζεται στον Πίνακα 6.1. Θα χρησιμοποιήσουμε μόνο στιγμιότυπα από τα έγχρωμα βίντεο διότι σε αυτή τη διπλωματική επικεντρωνόμαστε περισσότερο στην επεξεργασία των ακολουθιών με κύριο παράγοντα το μέγεθος που εισάγουμε στο δίκτυο, παρά την επεξεργασία πολλαπλών τύπων δεδομένων.

Δύο είναι τα σύνολα που θα χρησιμοποιήσουμε για να εκπαιδεύσουμε μόνο το δίκτυο αναγνώρισης χειρονομίας. Το πρώτο σύνολο δεδομένων ονομάζεται "Jester" και είναι διαθέσιμο στην ιστοσελίδα [33]. Αποτελείται από μια μεγάλη συλλογή μεμονωμένων και επισημασμένων χειρονομιών που εκτελούνται από άτομα μπροστά σε μια κάμερα υπολογιστή. Θεωρείται το μεγαλύτερο σύνολο δεδομένων μεμονωμένων χειρονομιών καθώς περιέχει 148.094 έγχρωμα βίντεο με 27 διαφορετικά είδη όπου η κάθε κατηγορία κλάσης περιέχει περισσότερα από 5.000 δείγματα. Το δεύτερο σύνολο δεδομένων για την αναγνώριση χειρονομίας είναι τα έγχρωμα βίντεο από την ιστοσελίδα [41] με την ονομασία "Isolated Gesture Recognition (ICPR '16)". Αποτελείται από 249 κλάσεις χειρονομιών που εκτελούνται από 21 διαφορετικά άτομα και περιέχει έγχρωμα βίντεο αλλά και βίντεο βάθους. Στην διπλωματική αυτή θα χρησιμοποιήσουμε μόνο τα έγχρωμα βίντεο.

Πινακας 6.1: Λεξιλόγιο απο το σύνολο "Multimodal Gesture Recognition: Montalbano V2 (ECCV '14)"

Συντομογραφία	Ολόκληρη φράση στα Ιταλικά	Μετάφραση
VA	Vattene	Βγείτε
VQ	Vieni qui	Ελάτε εδώ
PF	Perfetto	Τέλεια
FU	E' un furbo	Είναι έξυπνος
CP	Che due palle	Τι δύο μπάλες
CV	Che vuoi	Τι θέλεις
DC	Vanno d'accordo	Συμφωνούν
SP	Sei pazzo	Είσαι τρελός
CM	Cos hai combinato	Τι έχεις συνδυάσει
FN	Non me ne frega niente	Δε με νοιάζει
OK	Ok	Ολα καλά
CF	Cosa ti farei	Τι θα κάνατε
BS	Basta	Αρκετά
PR	Le vuoi prendere	Θέλετε να τα πάρετε
NU	Non ce ne piu	Δεν υπάρχουν άλλα
FM	Ho fame	Πεινάω
TT	Tanto tempo fa	Πριν από πολύ καιρό
BN	Buonissimo	Πολύ γευστικό
MC	Si sono messi d'accordo	Έχουν συμφωνήσει
ST	Sono stufo	Βαριέμαι

6.2 Δυναμική χρήση του συνόλου δεδομένων

Πολλές μέθοδοι αναγνώρισης χειρονομιών αποθηκεύουν τα βίντεο και εξάγουν στιγμιότυπα ανάλογα με τη διάρκεια της χειρονομίας η οποία δίνεται συνήθως σε μορφή αρχείου κειμένου. Στη συνέχεια αυτές οι μέθοδοι επεξεργάζονται τις εικόνες αυτές και τις μεταφέρουν στο νευρωνικό δίκτυο. Όμως η διαδικασία της μετατροπής του βίντεο σε εικόνες δεσμεύει μεγαλύτερο κομμάτι μνήμης και επειδή τα σύνολα δεδομένων που θέλουμε να εκπαιδεύσουμε στο δίκτυο μας είναι πολλά είναι προτιμότερο να κάνουμε οικονομία χώρου όπου είναι εφικτό.

Έτσι σε αυτή τη διπλωματική θα χρησιμοποιήσουμε αρχεία τα οποία περιέχουν συναρτήσεις που μας επιτρέπουν να έχουμε πρόσβαση απευθείας στα βίντεο χωρίς να χρειάζεται να εχουμε αποθηκευμένη την εικόνα και κατ' επέκτασιν επεξεργαζόμαστε τα δεδομένα μας δυναμικά. Λεπτομέρειες για αυτές τις εντολές οι οποίες έχουν πολλές ακόμη λειτουργίες υπάρχουν στην ιστοσελίδα [29].

6.3 Στρατηγική "jitter"

Γενικότερα σε μια χειρονομία υπάρχουν 3 χρονικές φάσεις: η αρχή, ο πυρήνας και το τέλος. Ανάλογα με το άτομο που εκτελεί τη χειρονομία στο σύνολο δεδομένων μας η διάρκειά της διαφέρει κάθε φορά. Ένας τρόπος να χωρίσουμε την ακολουθία είναι να ορίσουμε ένα συγκεκριμένο μήκος παραθύρου οπου για την εκπαίδευση κάθε χειρονομίας θα εισάγουμε τόσες εικόνες όσες το μήκος του παραθύρου. Εντούτοις μπορεί να αποκόψουμε σημαντικές πληροφορίες όπως τα όρια της χειρονομίας τα οποία αποτελούν το σημαντικότερο μέρος της εκπαίδευσης του δικτύου χρονικής κατάτμησης. Ένας άλλος τρόπος που προτείνεται στο άρθρο [17] είναι να μοιράσει το αρχικό μήκος του βίντεο σε ένα ισόποσα κατανεμημένο μήκος. Η μέθοδος αυτή ονομάζεται χρονική στρατηγική "jitter".

Η σχέση (6.1) είναι για τις μεταποιημένες ακολουθίες, όπου i είναι το i-στο δείγμα της ακολουθίας , S είναι το αρχικό μέγεθος της ακολουθίας και L είναι το μέγεθος που θέλουμε να την μεταποιήσουμε.

$$Idx_i = \frac{S}{L} \times (i + \frac{jit}{2}) \tag{6.1}$$

Το αποτέλεσμα της παραπάνω εξίσωσης είναι η παρακάτω σχέση που αναπαριστά τη νεα ακολουθία με το νεο μέγεθος L. Έτσι παίρνουμε ανά ίσα χρονικά διαστήματα στιγμιότυπα του βίντεο χωρίς να είναι απαραίτητο να είναι συνεχόμενα τα δείγματα αυτά.

$$\{Idx_1, Idx_2, ..., Idx_L\} \tag{6.2}$$

Προφανώς και η μέθοδος αυτή για κάθε ακολουθία προσθέτει αρκετές πράξεις στη διάρκεια της εκπαίδευσης και της επαλήθευσης. Στον παρακάτω Πίνακα παρουσιάζουμε ένα παράδειγμα απο το σύνολο δεδομένων μας απο το "ChaLearn" που έχει τη μορφή της πρώτης σειράς του Πίνακα 6.2 το οποίο θα χρησιμοποιήσουμε στο Κεφάλαιο αυτό για να παρουσιάσουμε τις διαφορές στα αποτελέσματα και στην προεπεξεργασία των διαφόρων μεθόδων που αναλύουμε σε αυτή και σε επόμενες ενότητες.

ΠΙΝΑΚΑΣ 6.2: Δείγμα νούμερο 0001 απο το σύνολο της ιστοσελίδας "ChaLearn"

Γραμμές	Αριθμός κλάσης, αρχή ακολουθίας, τέλος ακολουθίας
1	20,107,137
2	5,268,309
3	11,363,392
4	2,1086,1116
5	6,1269,1304
6	8,1345,1375
7	14,1385,1412
8	6,1530,1576
9	17,1585,1630
10	1,1788,1827

Έστω οτι εφαρμόζουμε στο παραπάνω παράδειγμα τη στρατηγική "jitter" με ένα τελικό επιθυμητό μέγεθος ακολουθιών ίσο με 15, τότε προκύπτουν τα αποτελέσματα του Πίνακα 6.3. Παρατηρούμε οτι ενώ έχουμε επιλέξει ένα πολύ μικρό επιθυμητό μήκος ακολουθίας (για να είναι αισθητή η διαφορά στον αναγνώστη από τις αρχικές ακολουθίες), η στρατηγική "jitter" εφαρμόζεται μέσα στα όρια των ακολουθιών και προσαρμόζει με μεγάλη επιτυχία τα διάφορα μεγέθη τους σε 15 ομοιόμορφα κατανεμημένες εικόνες.

Εκπαιδεύσαμε λοιπόν με τη στρατηγική αυτή το σύνολο δεδομένων μας και το δίκτυο είχε απο την αρχή πολύ καλά αποτελέσματα με υψηλή ακρίβεια και μικρό σφάλμα. Λεπτομέρειες για την εκπαίδευση του δικτύου αυτού δε θα αναλύσουμε σε αυτή τη διπλωματική διότι οι δοκιμές σε βίντεο που το δίκτυο δεν είχε εκπαιδευτεί προηγουμένως δεν είχαν τα αντίστοιχα αποτελέσματα. Συγκεκριμένα, όταν εισάγαμε τις ακολουθίες εκπαίδευσης από το σύνολο δεδομένων της ιστοσελίδας "ChaLearn" με την παραπάνω στρατηγική τα αποτελέσματα ήταν ορθά. Με άλλα λόγια το δίκτυο είχε ώς έξοδο μεγάλη πιθανότητα απο τη σιγμοειδή συνάρτηση ενεργοποίησης που αναλύσαμε στην ενότητα 2.2.1 για τις οριακές εικόνες και μικρή για τις μη-οριακές (καθώς αποτελούν μέρος της χειρονομίας).

Πινακας 6.3: Εφαρμογή στρατηγικής "jitter" στο βίντεο με αριθμό 0001 απο το σύνολο της ιστοσελίδας "ChaLearn"

Γραμμές	Αριθμός κλάσης, μεμονωμένα δείγματα απο την ακολουθία
1	20, [108 109 112 113 116 117 120 122 124 126 128 130 132 135 136]
2	5, [269 271 274 277 280 283 285 288 291 293 297 299 302 305 308]
3	11, [364 366 368 369 371 374 376 377 379 382 384 386 388 389 391]
4	2, [1087 1089 1091 1093 1095 1097 1099 1100 1103 1105 1107 1109 1112 1113 1115]
5	6, [1270 1272 1275 1277 1279 1281 1284 1287 1288 1291 1293 1296 1299 1300 1303]
6	8, [1346 1347 1349 1352 1353 1356 1358 1360 1362 1364 1366 1368 1370 1372 1374]
7	14,[1386 1387 1388 1389 1390 1391 1392 1393 1394 1395 1396 1397 1398 1399 1400]
8	6,[1531 1534 1537 1540 1543 1545 1549 1553 1555 1559 1561 1565 1567 1571 1574]
9	17, [1586 1588 1591 1595 1597 1600 1604 1607 1610 1613 1617 1619 1622 1625 1628]
10	1, [1789 1792 1793 1796 1799 1802 1805 1807 1810 1812 1815 1817 1821 1823 1826]

Επειδή τα αποτελέσματα ήταν ιδανικά με μεγάλη ακρίβεια και πολύ χαμηλό σφάλμα από τον πρώτο κιόλας κύκλο εκπαίδευσης, έπρεπε να εξετάσουμε το ενδεχόμενο να δίναμε στο δίκτυο διαφορετικά στιγμιότυπα που δεν ανήκαν σε κάποια κλάση ή περισσότερα στιγμιότυπα από αυτά που ορίζει το κείμενο αρχείου με τις κλάσεις που εκτελούνται στο βίντεο.

Για παράδειγμα τί αποτελέσματα θα είχαμε εάν στον Πίνακα 6.3 εισάγαμε τις εικόνες 20 εως 80 στη γραμμή 1 και στη συνέχεια εφαρμόζαμε τη στρατηγική "jitter". Τα αποτελέσματα προς έκπληξήν μας ήταν όμοια. Όποιοδήποτε σύνολο εικόνων και να εισάγαμε στο δίκτυο είχαμε τα ίδια αποτελέσματα, δηλαδή μεγάλες πιθανότητες ως έξοδο από τη συνάρτηση 2.2.1 στις δυο πρώτες και στις δυο τελευταίες εικόνες που υποδεικνύει ότι το δίκτυο τις αναγνώριζε ως οριακές και πολύ χαμηλές στις υπόλοιπες. Αυτό δεν είναι το ιδανικό αποτέλεσμα αφού ανεξαρτήτως του μεγέθους και της διάρκειας της ακολουθίας θα πρέπει να προβλέπει σωστά για κάθε εικόνα της αν αποτελεί όριο ή όχι.

Εφόσον λοιπόν εισάγαμε εικόνες που δεν ανήκουν σε κάποια κλάση θα έπρεπε όλες οι πιθανότητες απο τη συνάρτηση 2.2.1 να είναι χαμηλές ή έστω να μην έχουν ένα σταθερό μοτίβο όπως αυτό που το εκπαιδεύσαμε, δηλαδή να αναγνωρίζει ανεξάρτητα απο την ακολουθία που του δίνεται μόνο τις δυο πρώτες και τελευταίες εικόνες ως όρια.

Εκτός απο τα προαναφερόμενα θα πρέπει να λάβουμε υπόψη και το γεγονός οτι αρκετές εικόνες χάνονται και ειδικά αυτές που βρίσκονται ανάμεσα στις ακολουθίες που δίνει η ιστοσελίδα "ChaLearn", οι οποίες θεωρούνται μη-οριακές, για παράδειγμα οι εικόνες 137-268. Σε αυτές οι άνθρωποι εκτελούν διάφορες χειρονομίες που είναι εκτός του λεξιλογίου που δίνεται στην "ChaLearn" και με αυτή τη μέθοδο χάνονται αυτά τα στιγμιότυπα. Στις επόμενες ενότητες θα χρησιμοποιήσουμε όλες σχεδόν τις εικόνες για να εκπαιδεύσουμε το δίκτυό μας.

6.4 Μέθοδος επεξεργασίας ακολουθιών της διπλωματικής

Στην προηγούμενη ενότητα δε χρησιμοποιήσαμε όλα τα στιγμιότυπα που μας δίνει το σύνολο δεδομένων μας. Παρατηρήσαμε μια ανοδική πορεία στην ακρίβεια κατά την εκπαίδευση του δικτύου όταν συμπεριλάβαμε τις εικόνες ανάμεσα απο τις ακολουθίες που δίνονται στο αρχείο κειμένου και τις ορίσαμε ως μη-οριακές στην εκπαίδευση του. Τα στιγμιότυπα πριν απο την πρώτη ακολουθία δε λαμβάνονται υπόψη διότι τις περισσότερες φορές το άτομο στα βίντεο αργεί να εκτελέσει κάποια χειρονομία απο το λεξιλόγιο της "ChaLearn" με αποτέλεσμα η νέα ακολουθία να επιβαρύνει άσκοπα τη μνήμη σε μια μόνο επανάληψη. Ωστόσο χρησιμοποιούμε όλες τις υπόλοιπες ανάμεσα απο τα κενά.

Πρέπει να διανέμουμε ομοιόμορφα τα στιγμιότυπα απο τα κενά στις γειτονικές ακολουθίες. Για παράδειγμα στον Πίνακα 6.2 στη γραμμή 1 θα μπορούσαμε να έχουμε μια ακολουθία που να αρχίζει απο την εικόνα 107 και να τελειώνει στην εικόνα 202. Έτσι χρησιμοποιούμε επιπλέον 65 μη-οριακές εικόνες για να εκπαιδεύσουμε το δίκτυό μας. Τη διαδικασία αυτή την επαναλαμβάνουμε σε κάθε γραμμή του αρχείου κειμένου που περιέχει μια λίστα γραμμών που διαχωρίζονται με κόμμα και έχουν τη μορφή κλάση,αρχή,τέλος.

Επειδή η προεπεξεργασία των δεδομένων της διπλωματικής για την εισαγωγή τους στο δίκτυο χρονικής κατάτμησης αποτελεί ένα σημαντικό κομμάτι στην εκπαίδευσή του, παρουσιάζουμε σε ψευδοκώδικα την επεξεργασία των ακολουθιών. Αρχικά έχουμε μια μεταβλητή με την ονομασία $extra$ η οποία είναι τα στιγμιότυπα που θα προσθέσουμε στην αρχή της ακολουθίας που εξετάζουμε στη συγκεκριμένη επανάληψη η οποία κρατάει την τιμή από αυτήν που προσθέσαμε στην προηγούμενη ακολουθία ή μηδενική τιμή εαν η ακολουθία στη συγκεκριμένη επανάληψη είναι η πρώτη. Για να προστεθεί η τιμή αυτή θα πρέπει η τωρινή ακολουθία (αυτής της επανάληψης) να έχει τουλάχιστον 2 εικόνες διαφορά με την επόμενη για να έχει νόημα να προσθέσουμε τουλάχιστον μια εικόνα σε κάθε ακολουθία. Έχουμε μια ακόμη σημαντική μεταβλητή που είναι η $extra_n$ και αντιπροσωπεύει τις εικόνες που θα προστεθούν στο τέλος της τωρινής ακολουθίας και στην αρχή της επόμενης μέσω της μεταβλητής $extra$.

Η συγκεκριμένη μεθοδολογία αξιοποιεί πλήρως τα κενά μεταξύ των ακολουθιών. Το μεταποιημένο σύνολο δεδομένων εισάγεται στο δίκτυο χρονικής κατάτμησης και τα αποτελέσματα της εκπαίδευσής του θα αναλυθούν στο επόμενο Κεφάλαιο. Στην προσπάθειά μας να συγκρίνουμε τα αποτελέσματα και με άλλες αρχιτεκτονικές όπως αυτή του "Temporal Dilated Res3D" που αναφέραμε στην ενότητα 3.3, δεν καταφέραμε να το εκπαιδεύσουμε λόγω υψηλών απαιτήσεων σε μνήμη. Με το συγκεκριμένο αλγόριθμο παρατηρήσαμε οτι οι εικόνες δεν περιορίζονταν σε ποσότητα σε κάθε επανάληψη και για το λόγο αυτό εκπαιδεύσαμε το δίκτυο χρονικής κατάτμησης με ένα ακόμη αλγόριθμο. Έχει τη δομή του δεύτερου αλγόριθμου με τη διαφορά οτι ορίζουμε μια μεταβλητή με την ονομασία $number$ η οποία θέτει ένα όριο στις εικόνες που θα προσθέσουμε.

Algorithm 1 Χρήση όλων των στιγμιοτύπων ανάμεσα απο τα κενά των δοσμένων ακολουθιών

```
extra ← 0
for <όλες τι γραμμές του αρχείου εκτός της τελευταίας> do
  label ← line_f
  start ← line_s
  end ← line_t
  label_n ← line_nf
  start_n ← line_ns
  end_n ← line_nt
  if start_n − end > 2 then
    extra_n ← (start_n − end)/2
    start ← start − extra
    end ← end − extra_n
  else
    start ← start − extra
    end ← end
  end if
  start ← start_n − extra
end for
```

Algorithm 2 Χρήση περιορισμένων στιγμιοτύπων ανάμεσα απο τα κενά των δοσμένων ακολουθιών

```
extra ← 0
for <όλες τι γραμμές του αρχείου εκτός της τελευταίας> do
  label ← line_f
  start ← line_s
  end ← line_t
  label_n ← line_nf
  start_n ← line_ns
  end_n ← line_nt
  if start_n − end > 2 then
    extra_n ← (start_n − end)/2
    if extra_n > number then
    <extra_n = number >
    end if
    start ← start − extra
    end ← end − extra_n
  else
    if extra > number then
      <extra = number>
    end if
    start ← start − extra
    end ← end
  end if
  start ← start_n − extra
end for
```

6.5 Επεξεργασία δυαδικών πινάκων για την επισήμανση των εικόνων

Το σύνολο δεδομένων από την ιστοσελίδα [28] κατηγοριοποιείται σε φακέλους όπου ο κάθε ένας περιέχει βίντεο σε διάφορες μορφές μαζί με τα αρχεία κειμένου που έχουν πληροφορίες σχετικά με τα βίντεο αυτά. Αρχικά αναπτύξαμε ένα αλγόριθμο ο οποίος ψάχνει μέσα σε όλους τους φακέλους για να βρει το αρχείο κειμένου "label.txt". Από αυτό αντλεί τις πληροφορίες ανα γραμμή για την κλάση, την αρχή και το τέλος των βίντεο όλων των μορφών. Στη συνέχεια με βάση τον αλγόριθμο 1 που αναλύσαμε στο Κεφάλαιο 6 τα δεδομένα αποθηκεύονται σε μια λίστα με εξής πληροφορίες: μονοπάτι φακέλου, κλάση, νέα αρχή, νέο τέλος, επιπλέον εικόνες αρχής, επιπλέον εικόνες τέλους. Παρόλο που εκπαιδεύουμε το δίκτυό μας με επιπλέον εικόνες (οι οποίες βρίσκονταν μεταξύ των αρχικών ακολουθιών) θα πρέπει να κρατήσουμε τις δυο τελευταίες μεταβλητές και να τις αποθηκεύσουμε έτσι ώστε να γνωρίζουμε μετέπειτα πού αρχίζει και πού τελειώνει η ακολουθία. Κατά τη διάρκεια της εκπαίδευσης ο αλγόριθμος 1 χρησιμοποιείται μόνο στο δίκτυο χρονικής κατάτμησης ενώ στο δεύτερο δίκτυο χρησιμοποιούμε τη στρατηγική "jitter" για οικονομία χρόνου διότι ο αλγόριθμος 1 περιέχει περισσότερες πράξεις και εικόνες απο αυτές που εξάγει η στρατηγική "jitter". Κάθε επανάληψη καταναλώνει περισσότερη ώρα για να ολοκληρωθεί και επομένως αργεί ο κύκλος εκπαίδευσης.

Για την εκπαίδευση του δικτύου μας χρησιμοποιούμε τη βιβλιοθήκη βαθιάς μάθησης "Keras" η οποία περιλαμβάνει τρεις ξεχωριστές λειτουργίες οι οποίες μπορούν να χρησιμοποιηθούν για την εκπαίδευση διαφόρων μοντέλων. Σε αυτή τη διπλωματική θα χρησιμοποιήσουμε τη λειτουργία "fit-generator" η οποία όπως υποδηλώνει και το όνομά της θεωρεί ότι υπάρχει μια ειδική συνάρτηση που δημιουργεί τα δεδομένα για τη λειτουργία αυτή. Η συνάρτηση της γεννήτριας αποδίδει το μέγεθος παρτίδας (batch size) που έχουμε ορίσει στη λειτουργία του "Keras". Στο ΝΔ χρονικής κατάτμησης χρησιμοποιούμε παρτίδα μεγέθους ένα και σε αυτό της αναγνώρισης χειρονομίας μεγέθους 4. Όταν η λειτουργία δεχθεί την παρτίδα δεδομένων πραγματοποιεί τον αλγόριθμο οπισθοδιάδοσης σφάλματος που αναλύσαμε στο Κεφάλαιο 2. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να φτάσουμε τον επιθυμητό αριθμό κύκλου μάθησης που έχουμε ορίσει.

Η συνάρτηση αυτή διαφέρει στα δύο ΝΔ για δυο λόγους. Πρώτον επεξεργάζεται με διαφορετικό τρόπο τα δεδομένα και δεύτερον διαφοροποιείται στα διανύσματα εξόδου στο κάθε ΝΔ. Το κοινό χαρακτηριστικό των δυο συναρτήσεων είναι ότι επεξεργάζονται δυναμικά τα δεδομένα εκπαίδευσης (training data), επαλήθευσης (validation data) και δοκιμής (test data). Κατά τη διάρκεια της εκπαίδευσης η συνάρτηση που χρησιμοποιεί η λειτουργία "fit-generator" του ΝΔ εύρεσης χρονικών ορίων καλείται σε κάθε επανάληψη και επιστρέφει ενα πίνακα μεγέθους $(1, L, 112, 112)$ και ένα δυαδικό πίνακα που περιέχει τις κλάσεις για κάθε εικόνα. Ο δυαδικός πίνακας έχει μέγεθος L και αρχικοποιούμε όλα τα στοιχεία του με μηδενικά (δηλαδή καθορίζουμε όλες τις εικόνες της ακολουθίας μη-οριακές).

Στη συνέχεια ορίζουμε ως οριακές τις εικόνες αλλάζοντας το μηδέν σε ένα στις θέσεις του πίνακα $labels$ όπως ορίζουμε στην πρώτη γραμμή της εξίσωσης (6.3) λαμβάνοντας υπόψιν τις επιπλέον εικόνες που προσθέσαμε στην αρχή της ακολουθίας. Με άλλα λόγια αν στην αρχική ακολουθία θέταμε στον πίνακα $labels$ στις θέσεις $labels[0] = label[1] = 1$ την τιμή ένα, τότε εάν προσθέσουμε για παράδειγμα δυο επιπλέον εικόνες θα πρέπει να αλλάξει η θέση των ορίων σε $labels[2] = label[3] = 1$. Με την ίδια λογική θέσαμε την τιμή 1 και στα όρια τέλους της ακολουθίας. Με τον όρο $extra_s$ συμβολίζουμε την τιμή που είχαμε αποθηκεύσει στη λίστα ως οι επιπλέον εικόνες στην αρχή της ακολουθίας και με τον όρο $extra_e$ τον αριθμό των επιπλέον εικόνων στο τέλος της ακολουθίας. Τέλος η λειτουργία αυτή προωθεί τους πίνακες που αναφέραμε για εκπαίδευση.

$$\begin{aligned} labels[extra_s] &= labels[extra_s + 1] = 1 \\ labels[end - start - 1 - extra_e] &= labels[end - start - 2 - extra_e] = 1 \end{aligned} \quad (6.3)$$

Η ειδική συνάρτηση που τροφοδοτεί τη γεννήτρια κατά τη διάρκεια της εκπαίδευσης του δικτύου αναγνώρισης χειρονομίας επεξεργάζεται τις κλάσεις των δεδομένων με τη βοήθεια μιας συνάρτησης της βιβλιοθήκης "Keras" που χρησιμοποιείται κυρίως για μελέτες κατηγοριοποίησης δύο ή περισσότερων κλάσεων. Η συνάρτηση αυτή εξάγει ενα πίνακα ανάλογο με το μέγεθος του αριθμού των κλάσεων στο λεξιλόγιο του συνόλου δεδομένων. Με άλλα λόγια οι θέσεις αυτού του πίνακα αντιπροσωπεύουν τις κλάσεις και η συνάρτηση της βιβλιοθήκης "Keras" ενεργοποιεί αυτόματα (θέτει την τιμή ένα) στη θέση που αντιστοιχεί η κλάση της ακολουθίας που εξετάζουμε.

6.6 Αποτελέσματα πιθανοτήτων σε εικόνες

Σε αυτή την ενότητα δε θα παρουσιάσουμε αποτελέσματα απο την εκπαίδευση (ακρίβεια και σφάλμα) του δικτύου χρονικής κατάτμησης με βάση τις μεθόδους στους αλγόριθμους 1 και 2 αλλά τα αποτελέσματα δοκιμής σε μια ακολουθία του συνόλου δεδομένων μας. Για να κατανοήσουμε την επόμενη ενότητα θα πρέπει να αναλύσουμε και να απεικονίσουμε το πώς κατηγοριοποιούμε μια ακολουθία και πώς γίνεται η πρόβλεψη σε κάθε ένα απο τα στιγμιότυπα που την αντιπροσωπεύουν.

Στον Πίνακα 6.3 παρουσιάσαμε τα αποτελέσματα της στρατηγικής "jitter" όπου σε κάθε γραμμή αντιστοιχεί εκτός απο τον αριθμό κλάσης και η κανονικοποιημένη λίστα δειγμάτων όπως δίνεται απο την "ChaLearn". Για να την εισάγουμε στο δίκτυο χρονικής κατάτμησης με ένα απο τους δύο αλγόριθμους που αναφέραμε προηγουμένως θα πρέπει να τη μετατρέψουμε σε δυαδική λίστα στην οποία θα απεικονίζουμε με μηδενικά τις μη-οριακές εικόνες και με μονάδα τις οριακές. Λεπτομέρειες για την εισαγωγή τους στο δίκτυο για εκπαίδευση θα δοθούν στο επόμενο Κεφάλαιο. Εδώ θα παρουσιάσουμε ένα παράδειγμα που χρησιμοποιούμε τον αλγόριθμο 1 στον Πίνακα 6.4.

Στον παραπάνω Πίνακα παρατηρούμε οτι τα όρια των ακολουθιών έχουν επεκταθεί έτσι ώστε να αξιοποιούν πλήρως όλα τα στιμιότυπα για την εκπαίδευση του δικτύου μας εκτός απο αυτά πριν απο την πρώτη ακολουθία. Οι πληροφορίες σχετικά με τις αρχικές ακολουθίες αποθηκεύονται κατάλληλα ώστε να μπορέσουμε

ΠΙΝΑΚΑΣ 6.4: Δείγμα νούμερο 0001 απο το σύνολο στην "ChaLearn" με την εφαρμογή του αλγόριθμου 1

Γραμμές	Αριθμός κλάσης, αρχή ακολουθίας, τέλος ακολουθίας
1	20,107,202
2	5,203,336
3	11,337,739
4	2,740,1192
5	6,1193,1324
6	8,1325,1380
7	14,1381,1471
8	6,1472,1580
9	17,1581,1709
10	1,1710,1827

μετέπειτα να επισημάνουμε ως οριακές εικόνες συγκεκριμένες θέσεις του πίνακα που θα εισάγουμε για την εκπαίδευση (λεπτομέρειες για το πώς επιτυγχάνεται αυτό θα αναλυθούν στο Κεφάλαιο 7).

Στη γραμμή 6 στον παραπάνω Πίνακα έχουμε την κλάση 8 η οποία από το λεξιλόγιό μας είναι η κλάση "Είσαι τρελός" και στην αρχική της μορφή έχει τα όρια $1345, 1375$. Συνεπώς έχει επεκταθεί κατά 20 στιγμιότυπα στην αρχή της και κατά 5 στο τέλος της. Όταν περάσουμε την εκτεταμένη ακολουθία για δοκιμή στο νευρωνικό μας δίκτυο θα πρέπει να προβλέψει τα στιγμιότυπα $1345, 1346$ ως αρχή της ακολουθίας, τα $1374, 1375$ ως τέλος και τα υπόλοιπα ως μη-οριακά.

Σε προηγούμενο Κεφάλαιο αναλύσαμε το κύριο πλεονέκτημα του δικτύου χρονικής κατάτμησης που έχουμε χρησιμοποιήσει σε αυτή τη διπλωματική το οποίο είναι η διατήρηση της διάστασης στο χρόνο με την έννοια οτι διατηρείται η χρονική διάρκεια της ακολουθίας που εισάγεται στο δίκτυο και εξάγεται ένας πίνακας τιμών δισδιάστατος με μια γραμμή και στήλες όσο η χρονική διάρκεια της ακολουθίας. Αυτό σημαίνει οτι εαν εισάγουμε την παραπάνω εκτεταμένη ακολουθία η οποία έχει μήκος $1380 - 1325 = 55$ εικόνες θα έχουμε ως έξοδο απο τη σιγμοειδή συνάρτηση ενα πίνακα με 55 πιθανότητες που αντιστοιχούν σε κάθε μία εικόνα και υποδεικνύουν κατά πόσο μια εικόνα ειναι οριακή ή όχι.

Σχημα 6.1: Δείγμα 0001 και στιγμιότυπα 1328,1364,1374,1379 απο το σύνολο δεδομένων "ChaLearn"

Πινακας 6.5: Αποτελέσματα για τη γραμμή 6 απο το δείγμα 0001

Εικόνα σχήματος 6.1	Αποτέλεσμα
1	0.000077050528
2	0.0016587141
3	0.63958561
4	0.075009301

Έχουμε απομονώσει κάποιες εικόνες απο αυτή την ακολουθία στο Σχήμα 6.1 και παρουσιάζουμε τα αποτελέσματα του δικτύου χρονικής κατάτμησης στον Πίνακα 6.5. Οι κόκκινες ενδείξεις τιμών που είναι σημειωμένες στις εικόνες είναι οι αντίστοιχες τιμές στην περιγραφή του Σχήματος. Στην πρώτη εικόνα το άτομο βρίσκεται σε αδράνεια και η πιθανότητα να είναι οριακή κίνηση είναι πολύ μικρή όπως φαίνεται στον Πίνακα 6.5. Στη δεύτερη ο άνθρωπος της εικόνας εκτελεί τη χειρονομία η οποία βρίσκεται περίπου στη μέση της διάρκειάς της. Στην τρίτη εικόνα αν παρατηρήσουμε προσεκτικά ο άνθρωπος κατεβάζει το δεξί του χέρι και το δίκτυο προβλέπει επιτυχημένα οτι αποτελεί όριο μιας χειρονομίας ή διαφορετικά προβλέπει μια πιθανή κίνηση η οποία θεωρείται το τέλος της χειρονομίας "Είσαι τρελός".

Γενικότερα μετά από μελέτη αρκετών αποτελεσμάτων στις ακολουθίες δοκιμής παρατηρήσαμε οτι στα γειτονικά στιγμιότυπα των ακολουθιών προκύπτουν μεγαλύτερες πιθανότητες σε σχέση με τις υπόλοιπες. Ιδανικά λοιπόν το μοντέλο μας εφόσον έχει εκπαιδευτεί με δύο οριακές εικόνες στην αρχή και δύο στο τέλος του, θα πρεπει να έχει ως έξοδο ένα πίνακα με πιθανότητες που τείνουν προς το μηδέν για τις μη-οριακές εικόνες και κοντά στη μονάδα για τις οριακές.

Κεφάλαιο 7

Εκπαίδευση και πειράματα

Σε αυτό το Κεφάλαιο θα αναλύσουμε την εκπαίδευση των ΝΔ ανίχνευσης κίνησης και αναγνώρισης χειρονομίας. Στην ενότητα 7.1 παρουσιάζουμε κάποια από τα προβλήματα που προέκυψαν κατά τη διαδικασία της εκπαίδευσης και τους τρόπους αντιμετώπισής τους. Στην ενότητα 7.2 απεικονίζουμε την πορεία της ακρίβειας και του σφάλματος στο δίκτυο χρονικής κατάτμησης, συγκρίνουμε το δίκτυο αναγνώρισης χειρονομίας με άλλες αρχιτεκτονικές και αναλύουμε κάποιες πληροφορίες για την υλοποίησή μας στο "Keras". Στην ενότητα 7.3 παρουσιάζουμε τον Πίνακα Σύγχυσης για τις προβλέψεις κλάσεων του δικτύου αναγνώρισης χειρονομίας.

7.1 Τεχνικές βελτίωσης σε ανεπιθύμητα φαινόμενα

Όταν ξεκινάμε την εκπαίδευση ενός οποιουδήποτε ΝΔ αρχικοποιώντας τυχαία τα βάρη του δικτύου είναι προφανές ότι το δίκτυο δε θα εξάγει σωστά αποτελέσματα. Κατά τη διαδικασία της εκπαίδευσης το ΝΔ έχει ως αποτέλεσμα χαμηλή ακρίβεια στην αρχή η οποία αυξάνεται με το χρόνο κλιμακωτά. Μόνο η ανοδική πορεία της ακρίβειας επιβεβαιώνει ότι το μοντέλο μας αναγνωρίζει ορθά τις κατηγορίες των δεδομένων. Όσον αφορά τη συνάρτηση σφάλματος θα ήταν ιδανικό να διατηρεί καθοδική πορεία μέχρι το τέλος της εκπαίδευσης όπου στο τέλος θα συγκλίνει σε μια σταθερή τιμή. Η βελτίωση του δικτύου επιτυγχάνεται όταν προσαρμόζουμε τα βάρη με βάση τη διαφορά της προβλεπόμενης και της πραγματικής τιμής. Οι παράμετροι και οι τιμές τους αντίστοιχα είναι πολύ σημαντικές για να έχουμε σταθερή άνοδο στην ακρίβεια και κάθοδο στη συνάρτηση σφάλματος. Προκύπτουν αρκετά προβλήματα τα οποία προέρχονται από το συνδυασμό του είδους των επιπέδων, των υπερ-παραμέτρων και της αρχιτεκτονικής.

Στη βασική αρχιτεκτονική C3D που αναλύσαμε στο Κεφάλαιο 3 προσθέσαμε μερικά ακόμη επίπεδα κανονικοποίησης παρτίδας (batch normalization). Παρουσιάστηκε για πρώτη φορά στο άρθρο [19], το Φεβρουάριο του 2015 και αναφέρεται στην αντιμετώπιση του φαινομένου αλλαγής στην κατανομή εισόδου κάθε επιπέδου ενός ΝΔ κατά τη διάρκεια της εκπαίδευσης. Το φαινόμενο αυτό ονομάζεται "Internal Covariate Shift". Η είσοδος σε κάθε επίπεδο επηρεάζεται από τις αλλαγές στις παραμέτρους των προηγούμενων επιπέδων το οποίο απαιτεί μικρότερο ρυθμό εκμάθησης επιβραδύνοντας έτσι την εκπαίδευση. Ακόμα και σε μικρές αλλαγές στο δίκτυο επηρεάζεται η κατανομή εισόδου σε εσωτερικά στρώματα και αυτό είναι ένα μεγάλο πρόβλημα στα βαθιά ΝΔ. Το πρόβλημα αυτό αντιμετωπίζεται με

την τεχνική κανονικοποίησης παρτίδας η οποία βελτιώνει την ταχύτητα, την απόδοση και κυρίως τη σταθερότητα των ΝΔ. Ακόμη μειώνει την ανάγκη για χρήση των επιπέδων περιορισμού ενεργοποίησης (dropout layers).

Ο ρυθμός εκμάθησης (learning rate) αποτελεί την παράμετρο που ελέγχει το πόσο μεγάλο βήμα θα κάνουμε σε κάθε επανάληψη για να προσαρμοστούν τα βάρη του δικτύου σε σχέση με την κλίση του σφάλματος. Όσο πιο μικρή τιμή έχει, τόσο πιο αργά κινείται κατά μήκος της κλίσης με καθοδική πορεία. Αυτό θα ηταν μια καλή ιδέα (να χρησιμοποιήσουμε ενα χαμηλό ρυθμό εκμάθησης για να διασφαλίσουμε οτι δε θα χάσουμε κανένα τοπικό ελάχιστο στη συνάρτηση σφάλματος την οποία αναλύσαμε στην ενότητα 3.4), θα μπορούσε επίσης να σημαίνει ότι θα πάρει πολύ χρόνο για να συγκλίνει. Με τη χρήση της τεχνικής κανονικοποίησης παρτίδας έχουμε τη δυνατότητα να χρησιμοποιούμε μεγάλο ρυθμό εκμάθησης.

Το άρθρο [5] χρησιμοποιεί μια συνάρτηση μείωσης του ρυθμού εκμάθησης η οποία κατά τη διάρκεια της εκπαίδευσης μειώνεται κλιμακωτά ανάλογα με τον κύκλο. Παρόλα αυτά στο ΝΔ χρονικής κατάτμησης εφαρμόζοντας μια τέτοια συνάρτηση κατέληγε στα ίδια αποτελέσματα σε σχέση με μια σταθερή τιμή ρυθμού εκμάθησης στο διπλάσιο χρόνο. Για αυτό το λόγο στο δίκτυο χρονικής κατάτμησης (του οποίου η δομή παρουσιάζεται στον Πίνακα 3.1) θα χρησιμοποιήσουμε τη σταθερή τιμή 0.001 για το ρυθμό εκμάθησης ενω στο δίκτυο αναγνώρισης χειρονομίας που αναλύσαμε στην ενότητα 5.2 θα χρησιμοποιήσουμε τη συνάρτηση αυτή.

Μια άλλη σημαντική παράμετρος είναι το μέγεθος παρτίδας (batch size). Ορίζεται ως αριθμός των ακολουθιών του βίντεο που εισάγουμε στο δίκτυο σε κάθε επανάληψη με σκοπό να ανανέωσουμε τα βάρη του. Είναι προφανές οτι όσο μεγαλύτερη τιμή έχει το μέγεθος της παρτίδας τόσο περισσότερη μνήμη απαιτείται κατά τη διάρκεια της εκπαίδευσης σε μια επανάληψη. Συνεπώς επειδή κάθε ακολουθία έχει διαφορετικό αριθμό εικόνων και δεν τις ομαλοποιούμε θα πρέπει να λάβουμε υπόψιν το μέγεθος της μνήμης που καταναλώνουμε. Αυτό το επιτυγχάνουμε με το να θέσουμε την τιμή του μεγέθους παρτίδας ίσο με ένα. Έτσι δεν περιοριζόμαστε και επιβεβαιώνουμε οτι με την εφαρμογή του αλγόριθμου 1 (που αναφέραμε παραπάνω για την προεπεξεργασία δεδομένων) δεν ξεπερνάμε τα όρια της διαθέσιμης μνήμης ανεξαρτήτως πόσο μεγάλη είναι η ακολουθία που εισάγουμε ακόμη και με τις επιπλέον εικόνες που υπολογίσαμε.

Ένα πολύ σημαντικό φαινόμενο αποτελεί η υπερ-εκπαίδευση (overfitting). Ορίζεται ως το φαινόμενο κατά το οποίο ένα ΝΔ έχει εκπαιδευτεί στα δεδομένα εκπαίδευσης αλλά δε μπορεί να προβλέψει με επιτυχία άγνωστα δείγματα και οφείλεται κυρίως στο γεγονός οτι οι παράμετροι εκπαίδευσης είναι πολλές και δημιουργούν ενα αρκετά περίπλοκο μοντέλο. Στα ΝΔ που έχουν πολλά επίπεδα είναι ενα πολύ συχνό φαινόμενο. Ακόμη υπάρχει και το αντίθετο φαινόμενο που ονομάζεται υποεκπαίδευση (underfitting) και υποδηλώνει οτι το μοντέλο που χρησιμοποιούμε είναι πολύ απλό για να κατηγοριοποιήσει τα δεδομένα στις κλάσεις τους.

Το φαινόμενο της υπερ-εκπαίδευσης μπορεί να αντιμετωπιστεί με πολλούς τρόπους, δύο απο αυτούς είναι η χρήση των επιπέδων περιορισμού ενεργοποίησης και η πρόωρη διακοπή (early stopping). Παραπάνω αναφέραμε το λόγο για τον οποίο δε χρησιμοποιούμε τα επίπεδα περιορισμού ενεργοποίησης. Η χρήση της πρόωρης διακοπής είναι αρκετή για την αποφυγή του φαινομένου στο μοντέλο μας. Η λειτουργία αυτή μας δίνει τη δυνατότητα να εντοπίσουμε αυτόματα το τέλος της καθοδικής πορείας του σφάλματος στα δείγματα επαλήθευσης. Το τέλος αυτό σηματοδοτεί την αρχή της υπερ-εκπαίδευσης.

Ενας άλλος τρόπος για την αποφυγή υπερ-εκπαίδευσης είναι η ομαλοποίηση L2 (regularization). Ο τρόπος αυτός ορίζει ποινές στα βάρη που τείνουν να μεγαλώνουν και τα κρατάει μικρά ή μηδενικά. Μια άλλη ονομασία για την ποινή του βάρους είναι η "αποσύνθεση βάρους" (weight decay) και όπως υποδηλώνει και η ονομασία είναι η τάση να αναγκάζουμε τα βάρη να τείνουν στο μηδέν.

Δύο ενέργειες της ομαλοποίησης L2 είναι οι εξής:

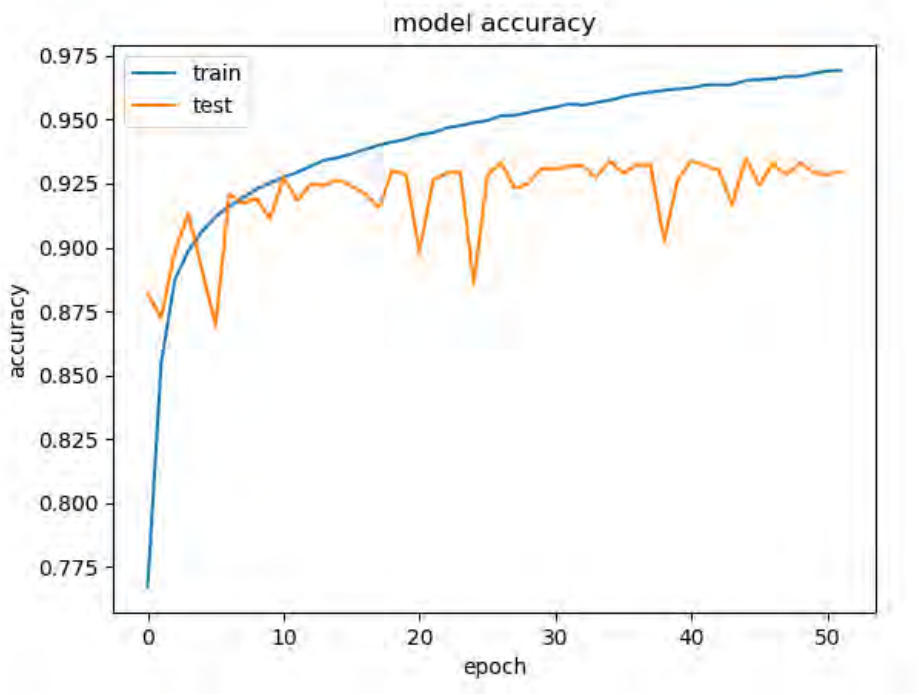
- Στη συνάρτηση κόστους προστίθεται ο όρος $(1/2)\lambda w^2$ για κάθε βάρος.
- Εχει την τάση να οδηγεί τα βάρη σε μικρότερες τιμές.

7.2 Λεπτομέρειες εκτέλεσης

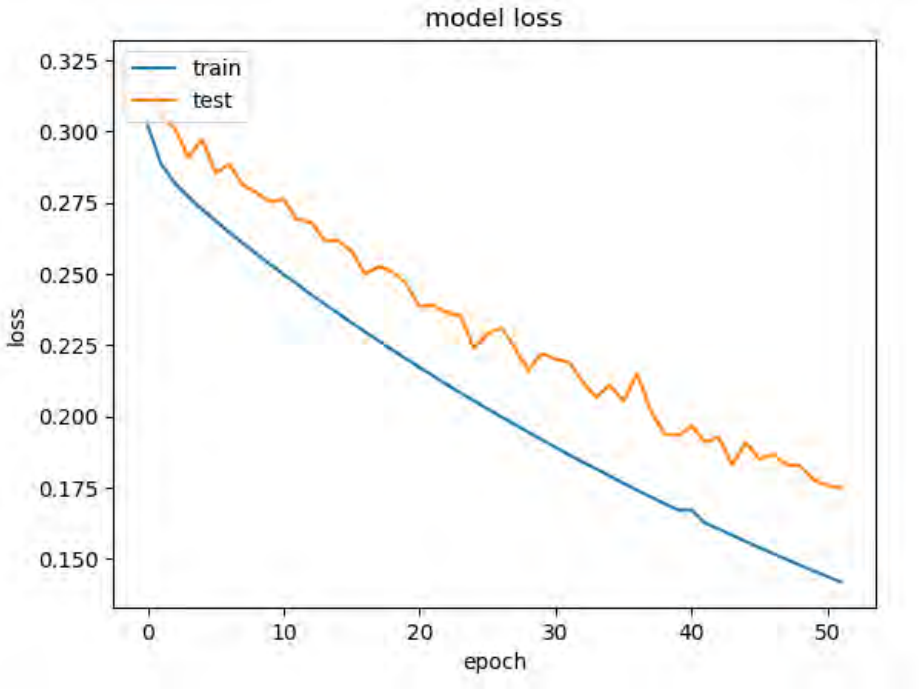
7.2.1 Αποτελέσματα του δικτύου χρονικής κατάτμησης

Ως είσοδο για το ΝΔ χρονικής κατάτμησης έχουμε το μέγεθος $(None, 112, 112, 3)$ και κατά τη διάρκεια της εκπαίδευσης εισάγεται αυτόματα μια επιπλέον μεταβλητή η οποία είναι η τιμή παρτίδας. Πρέπει να ορίσουμε ως πρώτη τιμή τη $None$ (που θεωρούμε ως χρονική διάσταση) ώστε να μπορέσουμε κατά τη διάρκεια της εκπαίδευσης να της δώσουμε διαφορετικές τιμές L σε κάθε επανάληψη.

Το ΝΔ χρονικής κατάτμησης εκπαιδεύεται μόνο με το σύνολο δεδομένων από την ιστοσελίδα του "ChaLearn" που αναφέραμε στην ενότητα 6.2 διότι οι κλάσεις είναι μόνο δύο και τα αποτελέσματα ήταν αρκετά ικανοποιητικά. Η έξοδος του δικτύου χρονικής κατάτμησης εξάγει τις πιθανότητες κάθε εικόνας ερευνώντας κατά πόσο κάθε μια από αυτές απεικονίζουν μια κίνηση. Όπως παρατηρούμε και απο τα σχήματα 7.1, 7.2 το μοντέλο μας για το ΝΔ χρονικής κατάτμησης ολοκλήρωσε 52 κύκλους εκπαίδευσης μέχρι να το σταματήσει η λειτουργία της βιβλιοθήκης "Keras" μέσω της πρόωρης διακοπής. Η πορεία της ακρίβειας και του σφάλματος στα δείγματα εκπαίδευσης (train στο σχήμα με τη μπλε γραμμή) είναι σταθερά ανοδική και καθοδική αντίστοιχα. Η ακρίβεια και το σφάλμα των δειγμάτων επαλήθευσης δεν είναι τόσο ομαλή στους πρώτους κύκλους εκπαίδευσης σε αντίθεση με τους τελευταίους που συγκλίνουν σε ικανοποιητικές τιμές.



Σχημα 7.1: Σχηματική απεικόνιση της ακρίβειας του μοντέλου (model accuracy) κατά τη διάρκεια των κύκλων εκπαίδευσης (epochs)



Σχημα 7.2: Σχηματική απεικόνιση του σφάλματος του μοντέλου (model loss) κατά τη διάρκεια των κύκλων εκπαίδευσης (epochs)

Η σύγκριση των αποτελεσμάτων με την αρχιτεκτονική "ResNet" απο το άρθρο [5] παρουσιάζονται στον Πίνακα 7.1. Για να συγκρίνουμε το μοντέλο μας με άλλες μεθόδους θα πρέπει να χρησιμοποιήσουμε την ίδια μεθοδο επεξεργασίας των βίντεο και για αυτό εκπαιδεύσαμε με τη μέθοδο αυτή την αρχιτεκτονική του "ResNet" του άρθρου που αναφέραμε και χρειάστηκε 30 κύκλους εκπαίδευσης μέχρι να τη σταματήσει η πρόωρη διακοπή.

Πινακας 7.1: Σύγκριση αποτελεσμάτων για το δίκτυο χρονικής κατάτμησης

Μοντέλο	ακρίβεια δοκιμής (%)
C3D διπλωματικής	92.83
ResNet [5]	93.87

7.2.2 Αποτελέσματα του δικτύου αναγνώρισης κίνησης

Στο δίκτυο αναγνώρισης χειρονομίας η είσοδος από τα βίντεο περιορίζεται στις 32 εικόνες (επειδή εφαρμόζουμε τη στρατηγική "jitter") και έχουμε ένα σταθερό κανονικοποιημένο μέγεθος ως είσοδο που είναι το $(32, 112, 112, 3)$ με τιμή παρτίδας ίση με 4. Αρχικά εκπαιδεύουμε το δίκτυο αναγνώρισης χειρονομίας με το σύνολο δεδομένων "Jester" που είναι διαθέσιμο στην ιστοσελίδα [33] με ρυθμό εκμάθησης που μειώνεται κλιμακωτά από 0.001 μέχρι 0.00001 κατά τη διάρκεια των κύκλων εκπαίδευσης. Στην συνέχεια κάνουμε "Fine-tuning" με τα βάρη που είχαμε στην έξοδο χρησιμοποιώντας τα ως αρχικοποίηση στο δίκτυο για να το εκπαιδεύσουμε με το σύνολο δεδομένων "Isolated Gesture Recognition (ICPR '16)" απο την ιστοσελίδα [41] και να μπορέσουμε να συγκρίνουμε με άλλες αρχιτεκτονικές. Στην αρχή όταν εκπαιδεύσαμε το ΝΔ αναγνώρισης χειρονομίας μόνο με το σύνολο δεδομένων από το "ChaLearn" παρατηρήσαμε υπο-εκπαίδευση γι' αυτό και το εκπαιδεύσαμε πρώτα στο σύνολο δεδομένων "Jester" και μετά με το σύνολο δεδομένων από το "ChaLearn".

Το μοντέλο αναγνώρισης χειρονομίας μπορεί να συγκριθεί με οποιαδήποτε αρχιτεκτονική διότι η μέθοδος που εισάγουμε τα δεδομένα στην αναγνώριση μεμονωμένων χειρονομιών είναι ίδια αρκεί να αναφέρεται στο ίδιο σύνολο δεδομένων. Όπως αναφέραμε και προηγουμένως η διπλωματική αυτή επικεντρώνεται περισσότερο στη διατήρηση των χρονικών χαρτών χαρακτηριστικών (δηλαδή την επιτυχημένη υλοποίηση του δικτύου χρονικής κατάτμησης) παρά στο συνδυασμό των διαφόρων μορφών σε σύνολα δεδομένων έτσι επιλέγουμε την μορφή μόνο έγχρωμων βίντεο. Στον Πίνακα 7.2 παρουσιάζουμε κάποιες από τις κορυφαίες αρχιτεκτονικές σε σύγκριση με το δικό μας μοντέλο το οποίο είναι το "C3D του Πίνακα 5.1 + ConvLSTM από Σχήμα 4.3 + MobileNet".

Πινακας 7.2: Σύνολο δεδομένων "IsoGD"

Μοντέλο	ακρίβεια δοκιμής (%) σε RGB δεδομένα
ResNet50 [39]	33.22
Pyramidal C3D [38]	36.58
C3D [40]	37.30
Res3D [42]	45.07
C3D του Πίνακα 5.1 + ConvLSTM από Σχήμα 4.3 + MobileNet	50.88
3DCNN+BiConvLSTM+2DCNN [43]	51.31
Res3D+ConvLSTM+MobileNet [6]	52.01
Res3D+ConvLSTM Variant(a)+MobileNet [6]	55.98

7.2.3 Σχετικά με την βιβλιοθήκη "Keras"

Στη βιβλιοθήκη "Keras" υπάρχουν δυο είδη κατασκευής μοντέλων: τα διαδοχικά και τα λειτουργικά. Η διαδοχική (sequential) έκδοση δομής επιπέδων μας επιτρέπει να δημιουργούμε μοντέλα των οποίων τα επίπεδα εκτελούνται διαδοχικά και εφαρμόζεται στις περισσότερες αρχιτεκτονικές. Έχει περιορισμένες δυνατότητες καθώς δεν επιτρέπει τη δημιουργία μοντέλων που μοιράζονται επίπεδα ή έχουν πολλαπλές εισόδους-εξόδους. Αντιθέτως η λειτουργική (functional) έκδοση είναι πιο ευέλικτη καθώς μας επιτρέπει να δημιουργούμε μοντέλα των οποίων τα στρώματα δεν εκτελούνται μόνο διαδοχικά (δηλαδή δεν εξαρτώνται από το προηγούμενο και το επόμενο επίπεδο). Για την ακρίβεια μπορούμε να συνδέσουμε όσα επίπεδα επιθυμούμε και σε οποιαδήποτε σειρά. Ως αποτέλεσμα, είναι δυνατή η δημιουργία πολύπλοκων δικτύων όπως το δίκτυο "υπολειμματικής μάθησης" (residual learning). Το δικό μας μοντέλο μπορεί να χρησιμοποιήσει και τις δύο εκδόσεις επειδή εκτελεί σειριακά τα επίπεδά του και δεν έχει διακλαδώσεις. Εμείς όμως προτιμήσαμε τη λειτουργική έκδοση για την ευελιξία που προσφέρει σε σύγκριση με τα άλλα μοντέλα.

Στα μοντέλα μας χρησιμοποιήσαμε ένα απο τα εργαλεία που παρέχει η βιβλιοθήκη βαθιάς μάθησης "Keras" η οποία είναι η πρόωρη διακοπή με δείκτη υπομονής (patience) ίσο με δυο. Ο δείκτης αυτός καθορίζει για πόσους κύκλους ακόμη θα συνεχίσει το μοντέλο την εκπαίδευση με βάση μια παράμετρο που θέτουμε. Χρησιμοποιήσαμε ως παράμετρο το σφάλμα των δειγμάτων επαλήθευσης η οποία συγκρίνει την τιμή που εξάγεται στον τελευταίο κύκλο με προηγούμενες τιμές του και φροντίζει το μοντέλο να έχει καθοδική πορεία στις τιμές του σφάλματος αυτού. Για δείκτη ίσο με δύο καθορίζουμε 2 κύκλους εκπαίδευσης χωρίς βελτίωση (μείωση του σφάλματος δειγμάτων) και κατά συνέπεια θα σταματήσει η εκπαίδευση.

Οφείλουμε να αναφέρουμε κάποιες πληροφορίες σχετικά με τα αρχεία και την έκδοση των βιβλιοθηκών που χρησιμοποιούμε. Ενας απο τους συγγραφείς του άρθρου [5] έχει ανεβάσει στην ιστοσελίδα [36] την υλοποίηση της ισορροπημένης τετραγωνικής συνάρτησης σφάλματος "hinge" που αναφέραμε στην ενότητα 3.4. Επειδή η αρχιτεκτονική του σχήματος 4.3 αποτελεί ενα κομμάτι του δικτύου αναγνώρισης χειρονομίας (αντί για ενα απλό συνελικτικό LSTM) πρέπει να χρησιμοποιήσουμε τη συνάρτηση "Gated" απο την ιστοσελίδα [35] που έχει ανεβάσει ο ίδιος συγγραφέας για το άρθρο που αναφέραμε προηγουμένως. Οδηγίες για την αντικατάσταση των αρχικών αρχείων που εκτελούν την εκπαίδευση του μοντέλου της βιβλιοθήκης "Keras" με αυτών που αναφέραμε περιέχει η κάθε ιστοσελίδα αντίστοιχα.

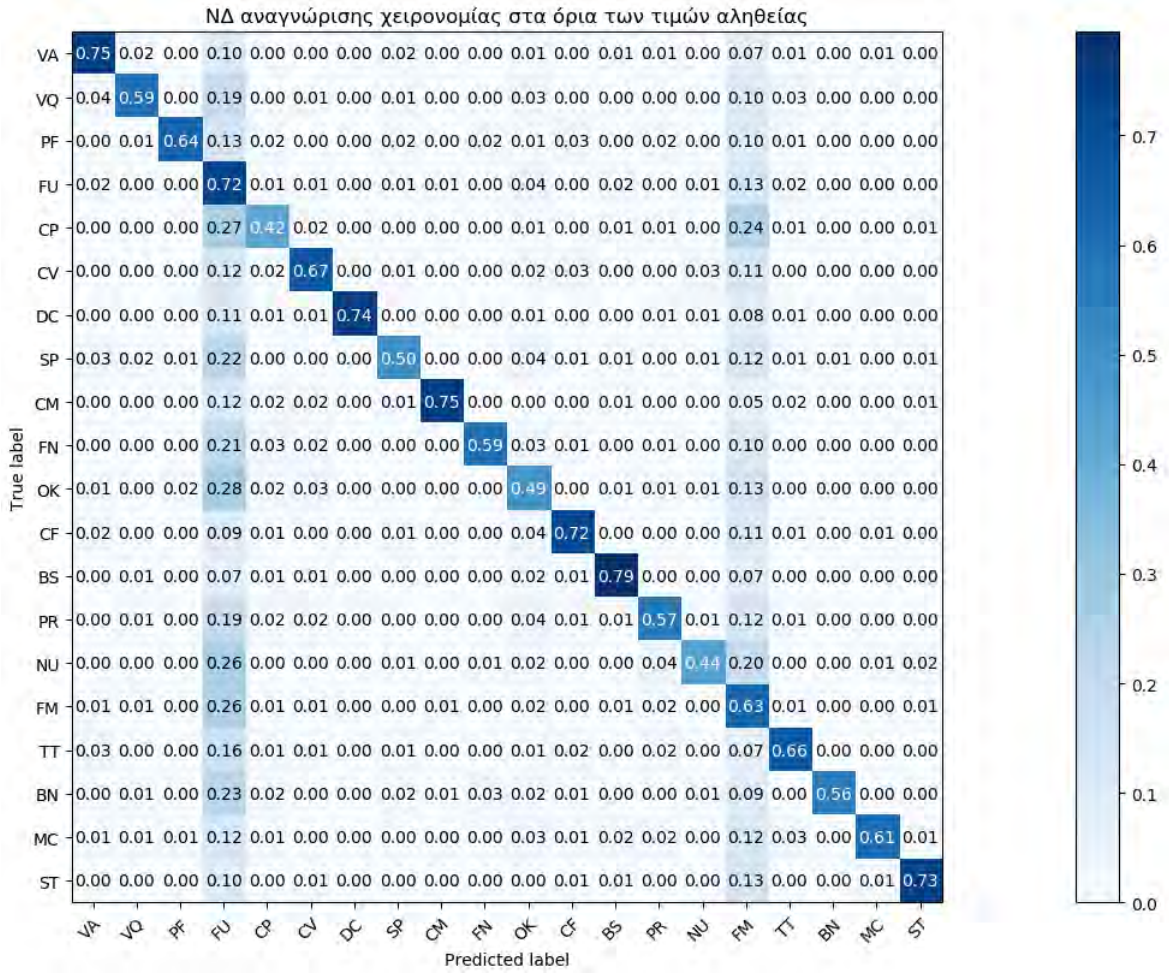
Αφού αντικαταστήσουμε αυτά τα αρχεία τα οποία προσθέτουν κάποιες ακόμη λειτουργίες στα αρχικά (χωρίς να αλλάζει ο σκοπός τους), πρέπει να τρέχουμε το πρόγραμμά μας και με τις κατάλληλες εκδόσεις οι οποίες είναι οι εξής:

- Keras version **2.0.0**
- Python **2.7.15**
- Tensorflow **1.2.0**

7.3 Απεικόνιση των αποτελεσμάτων αναγνώρισης χειρονομίας

Αφού εκπαιδεύσουμε το ΝΔ αναγνώρισης χειρονομίας εξάγουμε τα αποτελέσματα σε μορφή αρχείου έχοντας συμπληρώσει τα νέα όρια ακολουθιών. Η παρουσίαση των αποτελεσμάτων σε αυτή την ενότητα γίνεται με τα αρχικά όρια ακολουθιών. Συγκρίνουμε λοιπόν τα δεδομένα αληθείας με αυτά που εξήγαγε το ΝΔ αναγνώρισης χειρονομίας μέσω του Πίνακα Σύγχυσης Δεδομένων (Confusion Matrix) και στη δική μας περίπτωση πίνακα σύγχυσης κλάσεων. Ο πίνακας αυτός αποτελεί το δείκτη απόδοσης για το πρόβλημα της ταξινόμησης μηχανικής μάθησης δυο ή περισσότερων κατηγοριών.

Οι συντομογραφίες του πίνακα είναι απο το λεξιλόγιο του συνόλου δεδομένων μας που παρουσιάζονται στον Πίνακα 6.1. Παρατηρούμε οτι οι κλάσεις "Είσαι έξυπνος" (FU) και "Πεινάω" (FM) είναι δύσκολο να αναγνωριστούν απο το δίκτυο αναγνώρισης χειρονομίας και συμπίπτουν με αρκετές απο τις άλλες κλάσεις.



Σχημα 7.3: Πίνακας σύγχυσης κλάσεων

Κεφάλαιο 8

Συμπεράσματα

8.1 Περίληψη της Διπλωματικής

Στη διπλωματική αυτή υλοποιήσαμε δυο ΝΔ με σκοπό την αναγνώριση κίνησης και χειρονομίας απο ανεπεξέργαστα βίντεο. Προσπαθήσαμε μέσω της ανάλυσης των όρων, των ανεπιθύμητων φαινομένων αλλα και παραδειγμάτων να εξήγησουμε την υλοποίηση των αλγορίθμων, μεθόδων και λειτουργιών των μοντέλων οσο αναλυτικότερα μπορούμε. Η έρευνα αυτής της μελέτης ήταν μεγάλη πρόκληση και οι παράμετροι βελτίωσής της πολλοί. Η εκπαίδευση του δικτύου αναγνώρισης κίνησης χρειάστηκε πολλές ώρες για να ολοκληρωθεί παρόλο που χρησιμοποιήσαμε μόνο ένα σύνολο δεδομένων. Το δίκτυο αναγνώρισης χειρονομίας χρειάστηκε ακόμη περισσότερο χρόνο για να εκπαιδευτεί αφού χρησιμοποιήσαμε πρώτα το σύνολο δεδομένων "Jester" και μετά το σύνολο απο το "ChaLearn". Συνεπώς οι δοκιμές που εφαρμόζαμε κάθε φορά απαιτούσαν μέρες για να ολοκληρωθούν οι απαραίτητοι κύκλοι εκπαίδευσης ώστε να μπορέσουμε να κρίνουμε αν οι αλλαγές ήταν επιτυχείς ή όχι.

Ο αρχικός σκοπός μας ήταν η αναγνώριση κίνησης και χειρονομιών μέσα απο την εκπαίδευση ακολουθιών πολλών μορφών που μας παρέχει το σύνολο δεδομένων μας. Στην πορεία όμως επικεντρωθήκαμε στην απόδοση του ΝΔ αναγνώρισης κίνησης σχετικά με τη διατήρηση των χρονικών χαρτών χαρακτηριστικών. Για τη διαμόρφωση του δικτύου αναγνώρισης κίνησης ακολουθήσαμε αρχικά το άρθρο [16] το οποίο κατά την εκπαίδευσή του δε μας έδινε ικανοποιητικά αποτελέσματα για το σύνολο δεδομένων που επιλέξαμε. Έτσι αποφασίσαμε μέσω του άρθρου [5] να εφαρμόσουμε τις παρατηρήσεις και τις αλλαγές που προτείνει για τη βελτίωση της απόδοσης διαφόρων μοντέλων. Αναφερθήκαμε σε όλα αυτά τα άρθρα που συμβουλευτήκαμε για τη δημιουργία του δικτύου χρονικής κατάτμησης ειδικά στην ενότητα 3.3. Εκτός απο την παρουσίαση των αποτελεσμάτων ακρίβειας και σφάλματος αναπτύξαμε και τον αλγόριθμο για τον υπολογισμό του μέσου όρου του δείκτη "Jaccard" για να παρουσιάσουμε την απόδοσή του στις ορθώς κατηγοριοποιημένες ακολουθίες και στα όρια που προβλέψαμε.

8.2 Μελλοντικό σχέδιο εργασίας

Η διπλωματική αυτή αποτελεί μια βάση για αναγνώριση απο ζωντανή μετάδοση κάμερας εαν αλλάξει ο τρόπος προεπεξεργασίας των δεδομένων εκπαίδευσης, επαλήθευσης και δοκιμής σε συνδυασμό με κάποιες ακόμη προϋποθέσεις. Το "Tensorflow.js" αποτελεί τη βιβλιοθήκη για την ανάπτυξη και την κατάρτιση μοντέλων τεχνητής νοημοσύνης σε γλώσσα "JavaScript" και την ανάπτυξη του μοντέλου για να αναγνωρίζει κινήσεις μέσω των προγραμμάτων περιήγησης. Αυτό προϋποθέτει πολύ βαθύτερη μελέτη στο μοντέλο που αναπτύξαμε έτσι ώστε να μειώσουμε τις παραμέτρους που το επιβαρύνουν οι οποίες δεν είναι απαραίτητες και να βελτιώσουμε το χρόνο απόκρισής του. Ακόμη ο συνδυασμός και ο τύπος των επιπέδων συμβάλλουν αρκετά στη βελτίωση της ακρίβειας και του σφάλματος.

Ένας ακόμη στόχος θα ήταν να αξιοποιήσουμε όλες τις μορφές που μας δίνει το σύνολο δεδομένων μας ή να χρησιμοποιήσουμε το συνδυασμό του ήχου και των έγχρωμων βίντεο που αποτελούν την πιο διαδεδομένη μορφή δεδομένων σε αντίθεση με τα δεδομένα βάθους που χρειάζονται την κατάλληλη συσκευή.

Βιβλιογραφία

[1] Skymind, "A Beginner's Guide to Neural Networks and Deep Learning", *A.I. Wiki*. Διαθέσιμο: https://skymind.ai/wiki/neural-network

[2] Laurikkala Mikko, Suzuki Satoshi, and Vilkko Matti, "Predicting operator's cognitive and motion skills from joystick inputs", *42nd Annual Conference of the IEEE Industrial Electronics Society (IECON)*, Florence, 2016, pp. 5935-5940. doi: https://doi.org/10.1109/IECON.2016.7792994

[3] "Sigmoid function", *WikiPedia*. Διαθέσιμο: https://en.wikipedia.org/wiki/Sigmoid-function

[4] Avinash Sharma V., "Understanding Activation Functions in Neural Networks", *Medium*. Διαθέσιμο: https://link.medium.com/pnjTlnoHSX

[5] Guangming Zhu, Liang Zhang, Peiyi Shen, Juan Song, Syed Afaq Ali Shah, and Mohammed Bennamoun, "Continuous Gesture Segmentation and Recognition Using 3DCNN and Convolutional LSTM", *IEEE Transactions on Multimedia*, vol. 21, no. 4, April, pp. 1011-1021, 2019. doi: https://doi.org/10.1109/TMM.2018.2869278

[6] Liang Zhang, Guangming Zhu, Lin Mei, Peiyi Shen, Syed Afaq Ali Shah, and Mohammed Bennamoun, "Attention in convolutional LSTM for gesture recognition", *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1957-1966, 2018.

[7] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting", *Advances in neural information processing systems*, pp. 802–810, 2015.

[8] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks", *IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015, pp. 4489-4497. doi: https://doi.org/10.1109/ICCV.2015.510

[9] Zheng Shou, Dongang Wang, and Shih-Fu Chang, "Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 1049-1058. doi: https://doi.org/10.1109/CVPR.2016.119

[10] Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015. Διαθέσιμο: http://neuralnetworksanddeeplearning.com

[11] Milos Gajdos, "Fun With Neural Networks in Go", In *Machine Learning Explorations*. Διαθέσιμο: http://mlexplore.org/2016/07/27/fun-with-neural-networks-in-go/

[12] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang, "CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 1417-1426. doi: https://doi.org/10.1109/CVPR.2017.155

[13] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser, "Dilated Residual Networks", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 636-644. doi: https://doi.org/10.1109/CVPR.2017.75

[14] Bill Triggs, "Empirical filter estimation for subpixel interpolation and matching", *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV)*, Vancouver, BC, Canada, 2001, vol.2, pp. 550-557. doi: https://doi.org/10.1109/ICCV.2001.937674

[15] Tran Du, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri, "ConvNet Architecture Search for Spatiotemporal Feature Learning", arXiv:1708.05038, 2017.

[16] Ke Yang, Peng Qiao, Dongsheng Li, Shaohe Lv, and Yong Dou, "Exploring Temporal Preservation Networks for Precise Temporal Action Localization", arXiv:1708.03280, 2017.

[17] Guangming Zhu, Liang Zhang, Peiyi Shen, and Juan Song, "Multimodal Gesture Recognition Using 3-D Convolution and Convolutional LSTM", *IEEE Access*, vol. 5, pp. 4517-4524, 2017. doi: https://doi.org/10.1109/ACCESS.2017.2684186

[18] Vitaly Bushaev, "How do we 'train' neural networks?", *Medium*, Nov. 2017. Διαθέσιμο: https://medium.com

[19] Sergey Ioffe and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", arXiv:1502.03167, 2015.

[20] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio", *"Proceedings of the 9th ISCA Speech Synthesis Workshop"*, Sunnyvale, 2016, p. 125.

[21] Eren Gölge, "Dilated Convolution", *A Blog From Human-engineer-being*, Feb. 2017. Διαθέσιμο: http://www.erogol.com/dilated-convolution/

[22] "How do you minimize hinge-loss?", *Mathematics Stack Exchange*, Aug. 2014. Διαθέσιμο: https://math.stackexchange.com/q/2899178

[23] Isaias Prestes, "What is the difference between ConvLSTM and CNN LSTM?". Διαθέσιμο: https://quora.com

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep Convolutional neural networks", *Communications of the ACM*, vol. 60, no. 6, May, pp. 84-90, 2017. doi: https://doi.org/10.1145/3065386

[25] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv:1409.1556, 2014.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, vol. 115, pp. 211-252, 2015. doi: https://doi.org/10.1007/s11263-015-0816-y

[27] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, *"MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications"*, arXiv:1704.04861, 2017.

[28] ChaLearn, "Multimodal Gesture Recognition: Montalbano V2 (ECCV '14)". Διαθέσιμο:http://chalearnlap.cvc.uab.es

[29] ChaLearn, "Challenge and Data description". Διαθέσιμο: http://gesture.chalearn.org/

[30] Karen Simonyan and Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", arXiv:1406.2199, 2014.

[31] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 2625-2634. doi: https://doi.org/10.1109/CVPR.2015.7298878

[32] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek, "VideoLSTM Convolves, Attends and Flows for Action Recognition", arXiv:1607.01794, 2018.

[33] TwentyBN, "The 20BN-jester Dataset V1". Διαθέσιμο: https://www.twentybn.com/datasets/jester

[34] Pigou Lionel, Aäron van den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. "Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video", *International Journal of Computer Vision*, vol. 126, pp. 430-439, 2016. doi: https://doi.org/10.1007/s11263-016-0957-7

[35] Guangming Zhu, "Attention in Convolutional LSTM for Gesture Recognition", *Github*. Διαθέσιμο: https://github.com/GuangmingZhu/AttentionConvLSTM

[36] Guangming Zhu, "Continuous Gesture Segmentation and Recognition using 3DCNN and Convolutional LSTM ", *Github*. Διαθέσιμο: https://github.com/GuangmingZhu/ContinuousGR

[37] Scikit-learn. Διαθέσιμο: https://scikit-learn.org/

[38] Guangming Zhu, Liang Zhang, Lin Mei, Jie Shao, Juan Song, and Peiyi Shen, "Large-scale Isolated Gesture Recognition using pyramidal 3D convolutional networks", *23rd International Conference on Pattern Recognition (ICPR)*, Cancun, 2016, pp. 19-24. doi: https://doi.org/10.1109/ICPR.2016.7899601

[39] P. Narayana, J. R. Beveridge, and B. A. Draper, "Gesture Recognition: Focus on the Hands", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 5235-5244. doi: https://doi.org/10.1109/CVPR.2018.00549

[40] Yunan Li et al., "Large-scale gesture recognition with a fusion of RGB-D data based on the C3D model", *23rd International Conference on Pattern Recognition (ICPR)*, Cancun, 2016, pp. 25-30. doi: https://doi.org/10.1109/ICPR.2016.7899602

[41] ChaLearn, "Isolated Gesture Recognition (ICPR '16)". Διαθέσιμο: http://chalearnlap.cvc.uab.es/dataset/21/description/

[42] Q. Miao et al., "Multimodal Gesture Recognition Based on the ResC3D Network", *IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, 2017, pp. 3047-3055. doi: https://doi.org/10.1109/ICCVW.2017.360

[43] Liang Zhang, Guangming Zhu, Peiyi Shen, and Juan Song, "Learning Spatiotemporal Features Using 3DCNN and Convolutional LSTM for Gesture Recognition", *IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, 2017, pp. 3120-3128. doi: https://doi.org/10.1109/ICCVW.2017.369