ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Αναγνώριση σκηνών βίας και εκτάκτων αναγκών με Deep Learning

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΣΤΑΣΙΟΣ ΔΙΑΜΑΝΤΗΣ

Επιβλέπουσα: Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π

Αθήνα, Μάιος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Αναγνώριση σκηνών βίας και εκτάκτων αναγκών με Deep Learning

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΣΤΑΣΙΟΣ ΔΙΑΜΑΝΤΗΣ

Επιβλέπουσα: Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 5 Ιουλίου 2023.

.....	.....	.....
Θεοδώρα Βαρβαρίγου Καθηγήτρια Ε.Μ.Π	Εμμανουήλ Βαρβαρίγος Καθηγήτρια Ε.Μ.Π	Συμεών Παπαβασιλείου Καθηγήτρια Ε.Μ.Π

Αθήνα, Μαΐος 2023.

.....

Αναστάσιος Π. Διαμάντης
Ηλεκτρολόγος Μηχανικός & Μηχανικός Υπολογιστών

Περίληψη

Τα φαινόμενα βίαιων συμβάντων σε δημόσιους χώρους είναι ένα σημαντικό θέμα για την κοινωνία, που πέρα απο κίνδυνο σωματικής βλάβης προκαλεί σοβαρές οικονομικές ζημιές σε επιχειρήσεις και δημόσιους χώρους.

Η καταστολή τους αποτελούσε πάντα δύσκολο πρόβλημα, καθώς οι παρόντες των συμβάντων καθυστερούν να καλέσουν τις αρμόδιες αρχές. Επίσης, με την μέση απόκριση της αστυνομίας να είναι γύρω στα 13 λεπτά για σοβαρά συμβάντα [1] και το μέσο βίαιο συμβάν να διαρκεί κάτω απο 1 λεπτό, η έγκαιρη ενημέρωση των αρχών είναι κρίσιμη.

Η ραγδαία αύξηση των κάμερων ασφαλείας σε δημόσιους χώρους αποτελεί μια σημαντική ευκαιρία για τον εντοπισμό τους και την έγκαιρη ενημέρωση των αρχών. Ωστόσο, η παρακολούθηση χιλιάδων καμερών ταυτόχρονα απο ανθρώπους απαιτεί τεράστιο ποσό ανθρωπίνων πόρων και καθιστάται οικονομικά ασύμφορο.

Στόχος της παρούσας διπλωματικής εργασίας είναι να διερευνήσει τις προκλήσεις που υπάρχουν στον τομέα και να παρουσιάσει τεχνικές βαθιάς εκμάθησης για τον εντοπισμό σκηνών βίας σε βίντεο ή/και εικόνες.

Abstract

The phenomenon of violent events in public places is an important issue for public safety, which in addition to the risk of physical harm, also causes significant economic damages to businesses and public places.

Suppressing violent events has always been a difficult problem, as those present at the events take too long to call the relevant authorities. Also, with the average police response being around 13 minutes for serious incidents [1] and the average violent incident lasting less than 1 minute, timely notification of authorities is of crucial importance.

The rapid increase in surveillance cameras in public places presents an important opportunity to detect these events and inform the authorities in time. However, monitoring thousands of cameras simultaneously by humans requires a huge amount of human resources and becomes economically unviable.

The aim of this thesis is to investigate the challenges that exist in the field and to present deep learning techniques for the detection of violent scenes in videos and/or images.

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω την κα. Θεοδώρα Βαρβαρίγου για την ευκαιρία που μου έδωσε να εκπονήσω τη συγκεκριμένη εργασία καθώς και τον κ. Τάσο Νικολακόπουλο για την πολύ χρήσιμη καθοδήγηση που μου έδωσε κατά τη διεξαγωγή της διπλωματικής μου.

Επιπλέον, θα ήθελα να ευχαριστήσω τον φίλο μου Ιωάννη Ζαρόγιαννη για την γλωσσική επιμέλεια που μου παρείχε όπως επίσης και τον Παύλο Ψειμάδα που μου παρείχε τον υπολογιστή του καθώς και το γραφείο του που μου επιτρέψαν να φέρω την παρούσα εργασία εις πέρας.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου και την οικογένειά μου για την στήριξή και ενθάρρυνσή τους κατά την διάρκεια των σπουδών μου.

Disclaimer

Η παρούσα διατριβή εστιάζει στο κρίσιμο θέμα της ανίχνευσης βίας με τη χρήση νευρωνικών δικτύων, με στόχο να συμβάλει στον επιτακτικό στόχο της μείωσης της βίας και της προώθησης μιας ασφαλέστερης κοινωνίας. Είναι σημαντικό να σημειωθεί ότι η διατριβή εμπεριέχει έναν περιορισμένο αριθμό σκηνών (εικόνων) ήπιας βίας, επιλεγμένες προσεκτικά από δημόσια διαθέσιμα σύνολα δεδομένων. Ωστόσο, σε συμμόρφωση με τους ηθικούς λόγους και τις ανησυχίες περί απορρήτου, όλα τα αναγνωρίσιμα πρόσωπα στις εικόνες έχουν σκόπιμα επισκιαστεί ή θολώσει. Η συμπερίληψη τέτοιων εικόνων είναι απαραίτητη για την ακριβή εκπαίδευση και αξιολόγηση του προτεινόμενου συστήματος ανίχνευσης βίας, καθώς αντικατοπτρίζει τα σενάρια του πραγματικού κόσμου που συναντώνται στην πράξη. Η πρόθεση πίσω από αυτήν την ένταξη είναι αποκλειστικά η προώθηση της κατανόησης και της ανάπτυξης αποτελεσματικών μεθοδολογιών ανίχνευσης βίας. Προσπαθούμε να διασφαλίσουμε ότι αυτές οι εικόνες χρησιμοποιούνται υπεύθυνα και με ευαισθησία, έχοντας κατά νου τον απώτερο σκοπό της μόχλευσης της τεχνητής νοημοσύνης για το γενικότερο καλό, την προώθηση ενός ασφαλέστερου περιβάλλοντος και την προστασία των ατόμων από βλάβη.

Περιεχόμενα

Κεφάλαιο 1

Εκτεταμένη Περίληψη

1.1 Εισαγωγή

Η βία είναι ένα φαινόμενο το οποίο χαρακτηρίζεται απο την χρήση σωματικής δύναμης και επιθετικότητας με αποτέλεσμα την ψυχολογική ή / και σωματική βλάβη του εκλαμβάνοντος. Μπορεί να εκδηλωθεί με πολλές διαφορετικές μορφές, συμπεριλαμβανομένης της σωματικής βίας (όπως χτυπήματα, γροθιές ή κλωτσιές), σεξουαλική βία, συναισθηματική ή ψυχολογική βία (όπως απειλές, εκφοβισμός ή λεκτική κακοποίηση), ακόμη και συστημική βία (όπως διακρίσεις ή θεσμοθετημένη καταπίεση). Η βία μπορεί να προκαλέσει τόσο συναισθηματική όσο και σωματική βλάβη και μπορεί να έχει μακροχρόνιες επιπτώσεις τόσο σε άτομα όσο και σε κοινότητες. Πρόκειται για μια αρνητική και καταστροφική συμπεριφορά που πρέπει να αποφεύγεται και να καταστέλλεται όποτε είναι δυνατόν.

Η πιο εμφανής και ευδιάκριτη (απο τρίτους) εκδήλωσή της εμφανίζεται στη μορφή της σωματικής βίας, η οποία μπορεί να εκδηλώνεται είτε εντός προσωπικών χώρων (π.χ.: Ενδοοικογενειακή βία), είτε σε δημόσιους χώρους (π.χ.: Δημόσιες πλατείες ή ιδιωτικές επιχειρήσεις). Η δημόσια εκδηλούμενη βία συγκεκριμένα, βρίσκεται σε έξαρση στη νεολαία σε όλο τον κόσμο και θεωρείται ζήτημα δημόσιας υγείας [2]. Ως επί το πλείστον, η δημόσια έκφανσή της συνήθως περιλαμβάνει παραπάνω απο δύο εμπλεκόμενους και δημιουργεί χάος και αναταραχή. Πέραν της σωματικής και ψυχολογικής βλάβης που δέχονται οι εμπλεκόμενοι, δημιουργούνται σοβαρές υλικές και οικονομικές ζημιές που συμβαίνουν άμεσα αλλά και έμμεσα (μείωση της επιχειρηματικής δραστηριότητας μετά απο βίαιο συμβάν) στους δημόσιους χώρους ή / και ιδιωτικές επιχειρήσεις.

Η πρόβλεψη και καταστολή τέτοιων συμβάντων συνήθως απαιτεί την συνεχή παρουσία ενός αρμοδίου οργάνου (π.χ.: Αστυνομία ή Προσωπικό Ασφαλείας), των οποίων η πανταχού παρών παρουσία είναι λογιστικά ασύμφορη αλλά και άβολη για τους πολίτες. Μια ενδεχόμενη ευκαιρία αντιμετώπισης του προβλήματος εμφανίζεται με την ραγδαία άνοδο των καμερών ασφαλείας σε δημόσιους αλλά και ιδιωτικούς χώρους. Η ταυτόχρονη παρακολούθηση όλων των καμερών ασφαλείας όμως, είναι πάλι οικονομικά και λογιστικά μη βιώσιμη καθώς απαιτεί τεράστιο ανθρώπινο δυναμικό να παρακολουθεί ενδελεχώς. Επιπλέον, εγείρονται θέματα προστασίας προσωπικών δεδομένων από δόλιους χειριστές που παρακολουθούν διαρκώς δημόσιους χώρους. Εδώ εμφανίζεται η ανάγκη για την ανάπτυξη μιας μεθόδου για τον αυτόματο, μη επανδρωμένο εντοπισμό της βίας.

Αξίζει να σημειωθεί ότι πέραν των προαναφερθέντων περιπτώσεων χρήσης, υπάρχει τεράστια ζήτηση για την ανάπτυξη μιας γενικευμένης λύσης για αυτό το πρόβλημα για τον αυτόματο έλεγχο περιεχομένου στο διαδίκτυο. Δηλαδή, βίντεο και εικόνες που περιέχουν βίαιες σκηνές σε μέσα κοινωνικής δικτύωσης πρέπει να επισημαίνονται κατευθείαν, χωρίς παρέμβαση τρίτων, προλαβαίνοντας την θέαση τους απο ευάλωτο ανήλικο κοινό.

1.2 Αντικείμενο της Διπλωματικής

1.2.1 Στόχος

Η ανάπτυξη ενός συστήματος γενικευμένης αναγνώρισης βίαιων σκηνών σε βίντεο, όπως προαναφέρθηκε, λόγω των πολλών περιπτώσεων χρήσης έχει και πολλά ενδιαφερόμενα μέρη (π.χ.: Κρατικούς φορείς και ιδιωτικές επιχειρήσεις). Ενδεχομένως σημαντικότερες όμως, είναι οι γνώσεις και τεχνικές που αντλούνται απο την διερεύνηση επίλυσης αυτού του προβλήματος. Η αναγνώριση βίαιων συμβάντων πρόκειται για πολυεπίπεδο και πολύπλοκο πρόβλημα, όπου απαιτεί μεταξύ άλλων αναγνώριση ανθρώπων και της επαφής μεταξύ τους, έντονων κινήσεων και πρόθεσης σε διαφορετικές συνθήκες και φωτισμό. Μια λογική εικασία είναι λοιπόν ότι οι τεχνικές αυτές έχουν εφαρμογή και πέραν αυτού του προβλήματος, σε διάφορους τομείς της αναγνώρισης ανθρωπίνων δράσεων.

Την τελευταία δεκαετία υπήρξε μεγάλη προσπάθεια για την ανάπτυξη μοντέλων που βοηθούν στον εντοπισμό βίαιων συμβάντων με διάφορες μεθόδους και ποικίλο βαθμό επιτυχίας. Αυτές διαφοροποιούνται σε διαφορετικές μεθόδους, με την κλασσική αντιμετώπιση να είναι χρησιμοποιώντας είτε χειροποίητα χαρακτηριστικά που αντιπροσωπεύουν ρητά την τροχιά κίνησης, τον προσανατολισμό των άκρων, την τοπική εμφάνιση, τις αλλαγές μεταξύ των καρέ του βίντεο κτλ. (Nievas et al. [3], Hassner et al. [4], Gao et al. [5] και λοιποί [6, 7]). Ωστόσο, οι τεχνικές που χρησιμοποιούν χειροποίητα χαρακτηριστικά και κλασσικές μεθόδους Όρασης Υπολογιστών είναι σε γενικές γραμμές ακατάλληλες για την ανάπτυξη εφαρμογών στον πραγματικό κόσμο λόγω της αδυναμίας τους να γενικοποίησουν την αναγνώριση των συμβάντων σε διαφορετικές συνθήκες.

Η μεγάλη ανάπτυξη που έχει έρθει στον τομέα της Μηχανικής μάθησης (Machine Learning) και συγκεκριμένα στα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks) έχει στρέψει την προσπάθεια για ανάπτυξη μοντέλων εξ'ολοκλήρου σε αυτές τις μεθόδους (ή και με συνδυασμό χειροποίητων χαρακτηριστικών), λόγω της δυνατότητας τους να γενικοποιήσουν αλλά και να εκμεταλλευτούν υπάρχοντα μοντέλα που έχουν προπονηθεί για παρεμφερή προβλήματα (Transfer Learning). Αυτές κυμαίνονται σε λύσεις αναγνώρισης εικόνας καρέ προς καρέ [8], αναγνωρίσης καρέ προς καρέ και χρονικών εξαρτήσεων (CNN-LSTM) [9], δίκτυο δύο διαφορετικών ροών με συνδυασμό οπτικών ροών και RGB καρέ σε 3D Συνελικτικό Νευρωνικό Δίκτυο[10] και λοιπά.

Στόχος της παρούσας διπλωματικής εργασίας είναι η σφαιρική ανάλυση του προβλήματος στα εκάστοτε μέρη του και των προκλήσεων που εμφανίζονται, συμβουλεύοντας την υπάρχουσα επιστημονική βιβλιογραφία συγκεκριμένα πάνω στο θέμα αλλά και ευρύτερα. Πιο συγκεκριμένα επί του θέματος είναι η εύρεση δεδομένων για την εκπαίδευση και αξιολόγηση των μεθόδων που θα υλοποιηθουν, όπως και η ανάλυση βιβλιοθηκών (αλλά και δημιουργία νέων) για την επεξεργασία και χειραγώγηση των ακατέργαστων δεδομένων. Κεντρικού ενδιαφέροντος της εργασίας όμως είναι η ανάπτυξη, καθώς και ποιοτική και ποσοτική αξιολόγηση, των υπάρχων μοντέλων και τεχνικών για την πρόβλεψη βίαιων δράσεων, με στόχο την κατανόηση και βελτίωσή τους. Τέλος, ένα μέρος θα αποτελέσει και η μελέτη περί της μεταφοράς του μοντέλου απο πειραματικό στάδιο σε εφαρμογές πραγματικού κόσμου, ικανές να τρέχουν σε διαφορετικές πλατφόρμες.

1.2.2 Προκλήσεις

Στην παρούσα εργασία, στην προσπάθεια εύρεσης λύσης για το συγκεκριμένο πρόβλημα, καθώς και για κάθε παρόμοιο πρόβλημα ταξινόμησης βίντεο εμφανίζονται κάποια συγκεκριμένα εμπόδια. Πρόκειται για γενικές προκλήσεις που πρέπει να αντιμετωπιστούν πριν βρεθεί ειδική λύση επί τους προβλήματος. Αυτά θα αποτελέσουν το πρώτο κομμάτι της εργασίας. Αναφορικά μερικά παρακάτω.

Μέγεθος Δεδομένων

Το βάθος ενός τεχνητού νευρωνικού δικτύου (ANN) και το πλήθος των βαρών του, καθορίζει ένα άνω όριο για το πόσο πολύπλοκα είναι τα μοτίβα που μπορεί να εξάγει το δίκτυο. Δηλαδή ένα μεγαλύτερο νευρωνικό δίκτυο έχει ενδεχομένως την δυνατότητα να βρίσκει και να ταξινομεί πιο πολύπλοκες σχέσεις μεταξύ των δεδομένων[11]. Ενα εξίσου λογικό συμπέρασμα είναι ότι όσο πιο πολύπλοκο ένα πρόβλημα ταξινόμησης, τόσο πιο μεγάλα πρέπει να είναι και τα δείγματα δεδομένων που το αναπαριστούν. Στην κοινότητα των τεχνητών νευρωνικών δικτύων (ANNs), οι μελετητές και επαγγελματίες χρησιμοποιούν διάφορους εμπειρικούς κανόνες για την εύρεση του απαιτούμενου πλήθους των δεδομένων. Ένας απο αυτούς είναι ότι τα δείγματα πρέπει να είναι 10 με 100 φορές μεγαλύτερα απο τον αριθμό των χαρακτηριστικών[12, 13]. Ωστόσο, ο πιο διαδεδομένος ευριστικός κανόνας είναι ότι τα δείγματα πρέπει να είναι τουλάχιστον 10 φορές μεγαλύτερα απο το πλήθος των βαρών του δικτύου [14, 15]. Συνεπώς λόγω της μεγάλης πολυπλοκότητας του προβλήματος, η γενικευμένη λύση του απαιτεί μεγάλο πλήθος δεδομένων, το οποίο αποτελεί μεγάλο εμπόδιο στην προκειμένη περίπτωση καθώς η βία σε δημόσιους χώρους είναι σχετικά σπάνιο φαινόμενο.

Αμφισημία στα δεδομένα

Η δυσκολία ταξινόμησης μίας εικόνας σε κάποια κατηγορία βασίζεται αρκετά στο πόσο κοινά είναι τα χαρακτηριστικά των διαφορετικών κατηγοριών. Όσο πιο όμοιες είναι στο ανθρώπινο μάτι οι διαφορετικές κατηγορίες, τόσο πιο λεπτομερή χαρακτηριστικά πρέπει να εντοπίσει το μοντέλο για την ταξινόμησή τους. Για παράδειγμα ένα πρόβλημα ταξινόμησης εικόνων στις κατηγορίες "αιλουροειδές" ή "πτηνό" είναι πιο εύκολη απο την ταξινόμηση στις κατηγορίες "Περιστέρι" ή "Δεκαοχτούρα" καθώς οι διαφορές τους είναι πολύ πιο λεπτές.

(a) Αιλουροειδές (b) Πτηνό

(c) Περιστέρι (d) Δεκαοχτούρα

Σχήμα 1.1: **Κατηγορίες με διαφορετική δυσκολία ταξινόμησης.** Οι κατηγορίες στην πρώτη γραμμή (Αιλουροειδές / Πτηνό) είναι πιο "εύκολες" απο την δεύτερη (Περιστέρι / Δεκαοχτούρα).

Το παραπάνω παράδειγμα μπορεί να φαίνεται ανάξιο αναφοράς, ωστόσο όσον αφορά την αναγνώριση ανθρωπίνων δράσεων, η ταξινόμηση τους δεν είναι τόσο προφανής. Αυτό ευθύνεται στο γεγονός ότι οι ανθρώπινες δράσεις δεν εντάσσονται σε διακριτές κατηγορίες με τόσο αυστηρούς κανόνες (όσο π.χ. η ταξινόμηση ενός φυτού ή ζώου). Κάθε ανθρώπινη δράση έχει ιεραρχική φύση αναλόγως με την οπτική γωνία απο την οποία εξετάζεται. Παραδείγματος χάρη, ένας μπορεί να θεωρήσει μια διακριτή ανθρώπινη πράξη (π.χ. : "Κατεβαίνω τα σκαλιά") με βάση τον στόχο και πρόθεση της πράξης (π.χ. : "Πηγαίνω στο σχολείο").

(a) (b) (c)

Σχήμα 1.2: **Αμφισημία δεδομένων.** Κάθε πράξη έχει διαφορετική σημασία και ταξινόμηση ανάλογα με την πρόθεση και την κατηγορία που ταξινομείται. Στο (α) τα λιοντάρια παίζουν ή κυνηγάνε θήραμα, (β) ο παίκτης σουτάρει ή δίνει πάσα, (γ) οι άνθρωποι διασκεδάζουν ή τσακώνονται;

Χρονικές εξαρτήσεις στα δεδομένα

Οποιαδήποτε δράση, ανθρώπινη και μη, δεν πρόκειται για σημειακό στο χρόνο γεγονός αλλά αιτιατό. Δηλαδή οποιαδήποτε δράση, πρόκειται για ένα συμβάν το οποίο έχει κάποιο αποτέλεσμα που βρίσκεται στο μέλλον, με την σειρά των γεγονότων να παίζει κομβικό ρόλο για την κατηγοριοποίηση της.

Σχήμα 1.3: **Παράδειγμα χρονικής εξάρτησης δεδομένων** Η πόρτα ήταν ανοιχτή και κλείνει ή κλειστή και ανοίγει;

Στην παραπάνω εικόνα, αν θεωρήσουμε ότι πρόκειται για ένα μια ακίνητο στιγμιότυπο ενός βίντεο, τότε ο βαθμός χρονικής εξάρτησης των δεδομένων εξαρτάται άμεσα απο την δράση που

θέλουμε να εντοπίσουμε. Για παράδειγμα, αν ο σκοπός είναι η ταξινόμηση του βίντεο (μέσω της εικόνας) στις κατηγορίες "ανοίγει πόρτα" / "κλείνει πόρτα" τότε η ταξινόμηση της εικόνας είναι αδύνατη χωρίς να γνωρίζουμε αν η πόρτα πριν ήταν κλειστή ή ανοικτή.

Η συντονισμένη προσπάθεια στον χώρο έχει οδηγήσει την αναγνώριση μοτίβων σε εικόνες να έχει ραγδαία βελτίωση την τελευταία δεκαετία. Τέτοιες προσπάθειες, όπως για παράδειγμα ο ετήσιος διαγωνισμός ImageNet Large Scale Visual Recognition Challenge (ILSVRC) καθώς και η δημιουργία τεράστιων συνόλων δεδομένων (datasets), όπως το ImageNet, το MNIST κ.α., έχουν το απρόοπτο πλεονέκτημα ότι βαθιά νευρωνικά δίκτυα που έχουν προπονηθεί σε χιλιάδες εικόνες ανάμεσα σε χιλιάδες κατηγορίες μπορούν να χρησιμοποιηθούν και για προβλήματα σε άλλους τομείς (Μεταφορά Μάθησης)[16].

Ωστόσο, η βελτίωση που έχει επέλθει στην αναγνώριση εικόνων δεν είχε την ίδια πρόοδο και στην αναγνώριση βίντεο, με τις καλύτερες αρχιτεκτονικές να μην αποκλίνουν σημαντικά απο την ανάλυση κάθε καρέ μεμονωμένα. Στον τομέα των βίντεο είναι ανοιχτό ερώτημα αν η εκπαίδευση ενός δικτύου σε ένα επαρκώς μεγάλο σύνολο δεδομένων ταξινόμησης ενεργειών, θα δώσει παρόμοια ώθηση στην απόδοση όταν εφαρμόζεται σε διαφορετική χρονική εργασία ή σύνολο δεδομένων[16]. Εμφανίζεται λοιπόν, σοβαρό εμπόδιο στο να μπορέσουμε να αναπτύξουμε τεχνικές οι οποίες να μας επιτρέπουν να εντοπίσουμε έντονες χρονικές εξαρτήσεις στα δεδομένα μας.

Απόδοση

Η αύξηση της πολυπλοκότητας των προβλημάτων, όπως προαναφέρθηκε, οδηγεί άμεσα στην αύξηση του μεγέθους του δικτύου και συνεπώς των βαρών. Ειδικά στην αναγνώριση σε βίντεο, τα νευρωνικά δίκτυα τείνουν να έχουν εκθετικά περισσότερα βάρη, ανάλογα βέβαια με το πόσα συνεχόμενα καρέ μπορεί να επεξεργαστεί το δίκτυο καθώς προφανώς και την αρχιτεκτονική που χρησιμοποιείται. Εγείρονται λοιπόν, σοβαρά προβλήματα δυνατότητας εκπαίδευσης αυτών των βαθιών νευρωνικών δίκτυων, απαιτώντας μονάδες επεξεργασίας γραφικών (GPUs) ή μονάδες επεξεργασίας τανυστών (TPUs) με τεράστια επεξεργαστική ισχύ που βρίσκεται μόνο σε υπερυπολογιστές φτιαγμένους ειδικά για προπόνηση βαθιών νευρωνικών δικτύων. Ακόμα πιο σημαντικό πέρα της επεξεργαστικής ισχύς, που επηρεάζει μόνο τον χρόνο εκπαίδευσης, είναι η μεγάλη μνήμη (VRAM) που απαιτείται να έχουν έτσι ώστε να μπορούν να αποθηκευτούν όλα τα καρέ απο πολλά βίντεο ταυτόχρονα που χωρίς αυτή δε μπορούν να εκπαιδευτούν γενικά, ανεξαρτήτου χρόνου.

Το παραπάνω μπορεί να μην φαίνεται ως τόσο σοβαρό εμπόδιο, καθώς τα δίκτυα μπορούν ενδεχομένως να εκπαιδευτούν σε απομακρυσμένους υπερυπολογιστές και έπειτα να δοθούν στους καταναλωτές. Πρέπει ωστόσο, το νευρωνικό δίκτυο να είναι έστω επαρκώς αποδοτικό ώστε να μπορούν τρέξουν δεδομένα για αξιολόγηση μόνο τοπικά, σε υπολογιστές με ρεαλιστικούς πόρους. Επίσης, η χρονική απόκριση του δικτύου είναι κομβικής σημασίας ειδικά αν πρόκειται για εφαρμογή πραγματικού χρόνου. Παραδείγματος χάρη, σε ένα νευρωνικό δίκτυο αναγνώρισης οχημάτων που παραβαίνουν τον κώδικα οδικής κυκλοφορίας με σκοπό την άμεση κινητοποίηση των αρχών, η χρονική καθυστέρηση της εφαρμογής αναιρεί την όλη αρχική σκοπιμότητα της ιδέας. Επιβάλλεται λοιπόν, σε κάθε προσπάθεια ανάπτυξης ενός μοντέλου με βλέψεις υλοποίησης στον πραγματικό κόσμο, πέραν απο καθαρά ερευνητικό / ακαδημαϊκό σκοπό, να εκληφθεί εξ'αρχής υπ'όψιν η αρχιτεκτονική του μοντέλου με σκοπό την εύρεση όσο πιο αποδοτικής λύσης, και όχι εκείνης που έχει απλά το αποτέλεσμα με την καλύτερη επίδοση.

1.3 Δομή της Εργασίας

Στο κεφάλαιο 2 της παρούσας εργασίες αναλύεται ένα μεγάλο κομμάτι του θεωρητικού υπόβαθρου, το οποίο είναι αρκετό (ή έστω καλύπτει σε σημαντικό βαθμό) την κατανόηση της παρούσας εργασίας. Η πληροφορία που περιέχεται σε αυτό το κεφάλαιο χρησιμοποιείται ως "δεξαμενή γνώσεων" για την ποιοτική και ποσοτική αξιολόγηση των υπάρχουσων τεχνολογιών αλλά και την ανάπτυξη νέων.

Στο κεφάλαιο 3 αναλύονται όλα τα διαθέσιμα δημόσια σύνολα δεδομένων στην μορφή βίντεο με τον σκοπό αναγνώρισης σκηνών βίας με βάση την ποιότητα τους και την χρησιμότητά τους στην εκπαίδευση νευρωνικών δικτύων στην παρούσα εργασία. Επίσης, αναλύονται (έως την στιγμή εκπόνησης της παρούσης εργασίας) όλες οι υπάρχουσες σχετικές εργασίες αναγνώρισης *πραγματικών* σκηνών βίας με χρήση τεχνικών μηχανικής μάθησης.

Στο κεφάλαιο 4 παρουσιάζεται η ανάπτυξη δύο διαφορετικών μεθόδων αντιμετώπισης του προβλήματος της αναγνώρισης πραγματικών σκηνών βίας, αντλώντας έμπνευση από υπάρχουσες τεχνολογίες, με χωρικές και χωροχρονικές μεθόδους αντίστοιχα. Οι εκάστοτε μέθοδοι παρουσιάζονται μαζί με τα σχετικά πειράματα αλλά και εξετάζονται με βάση την πρακτικότητα και χρηστικότητα τους στην δυνατότητα εφαρμογής τους στον πραγματικό κόσμο με διάφορες μετρικές.

Στο κεφάλαιο 5 καταγράφονται τα βασικά συμπεράσματα της εργασίας καθώς και παρουσιάζονται ενδεχόμενα μελλοντικά ερευνητικά σχέδια στο θέμα της αναγνώρισης βίας.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Μηχανική Μάθηση

2.1.1 Ορισμός Μηχανικής Μάθησης

Η μηχανική μάθηση είναι ένα πεδίο μελέτης και πρακτικής που περιλαμβάνει την ανάπτυξη και χρήση αλγορίθμων και στατιστικών μοντέλων που επιτρέπουν στα συστήματα υπολογιστών να βελτιώσουν την απόδοσή τους σε μια συγκεκριμένη εργασία ή σύνολο εργασιών με βάση μοτίβα και γνώσεις που προέρχονται από δεδομένα, χωρίς να προγραμματίζονται ρητά. Με άλλα λόγια, η μηχανική μάθηση είναι η διαδικασία εκπαίδευσης και βελτιστοποίησης μοντέλων σε δεδομένα για τη λήψη ακριβών προβλέψεων ή αποφάσεων χωρίς να είναι ρητά προγραμματισμένος να το κάνει. Είναι ένα υποπεδίο της τεχνητής νοημοσύνης που εστιάζει στη δημιουργία ευφυών συστημάτων που μπορούν να μάθουν και να προσαρμοστούν από δεδομένα.

Ο τομέας της μηχανικής μάθησης και των νευρωνικών δικτύων έχει εμπνευστεί άμεσα απο την ανθρώπινη ευφυΐα και τον ανθρώπινο εγκέφαλο. Η ιδέα ότι οι μηχανές μπορούν να μάθουν και να προσαρμοστούν είχε πρώτα προταθεί απο τον Arthur Samuel το 1959 ο οποίος ανέπτυξε ένα πρόγραμμα που μπορούσε να παίξει το παιχνίδι checkers και να βελτιώσει την απόδοσή του με την πάροδο του χρόνου μέσω δοκιμής και λάθους[17]. Η ανάπτυξη του τομέα της μηχανικής μάθησης, και ειδικότερα των νευρωνικών δικτύων, έχει προέλθει άμεσα απο την μελέτη του ανθρώπινου εγκεφάλου και των διεργασιών του. Άξιο αναφοράς είναι ότι το πρώτο τεχνητό νευρωνικό δίκτυο δεν αναπτύχθηκε απο επιστήμονα υπολογιστών, αλλά απο τον ψυχολόγο Frank Rosenblatt γνωστό ως νευρώνα perceptron[18].

2.1.2 Είδη Μηχανικής Μάθησης

Η μηχανική μάθηση διακρίνεται σε τρεις βασικές διακριτές συνιστώσες, κάθε μια απο τις οποίες χρησιμοποιείται για υλοποίηση διαφορετικού στόχου και διαχωρίζονται με βάση τον βαθμό της παρέμβασης του ανθρώπου στην εκπαίδευση του.

Επιβλεπόμενη Μάθηση

Στην επιβλεπόμενη μάθηση, ο αλγόριθμος μηχανικής μάθησης εκπαιδεύεται σε δεδομένα με ετικέτες, πράγμα που σημαίνει ότι τα δεδομένα εισόδου συσχετίζονται με τη σωστή έξοδο ή οποιαδήποτε μεταβλητή στόχου. Ο αλγόριθμος μαθαίνει να αντιστοιχίζει τα δεδομένα εισόδου στη σωστή έξοδο ελαχιστοποιώντας μια συνάρτηση κόστους (loss function). Ο στόχος της επιβλεπόμενης μάθησης είναι να δημιουργήσει ένα μοντέλο που μπορεί να προβλέψει με ακρίβεια την έξοδο για νέα δεδομένα εισόδου που δεν έχει δει ξανά ο αλγόριθμος.

Παραδείγματα προβλημάτων στα οποία χρησιμοποιείται επιβλεπόμενη μάθηση περιλαμβάνουν ταξινόμηση εικόνων, αναγνώριση ομιλίας και ανάλυση παλινδρόμησης.

Μη Επιβλεπόμενη Μάθηση

Στην μη επιβλεπόμενη μάθηση, ο αλγόριθμος μηχανικής εκμάθησης εκπαιδεύεται σε δεδομένα χωρίς ετικέτες, πράγμα που σημαίνει ότι τα δεδομένα εισόδου δεν συσχετίζονται με καμία εξόδου ή μεταβλητή στόχου. Ο αλγόριθμος μαθαίνει να αναγνωρίζει μοτίβα και δομή μέσα στα δεδομένα ελαχιστοποιώντας μια συνάρτηση κόστους (loss function). Ο στόχος της μη επιβλεπόμενης μάθησης είναι να βρει μη εμφανή μοτίβα και σχέσεις στα δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν για οπτικοποίηση δεδομένων, ομαδοποίηση ή εξαγωγή χαρακτηριστικών. Παραδείγματα προβλημάτων στα οποία χρησιμοποείται μη επιβλεπόμενη μάθηση περιλαμβάνουν ανίχνευση ανωμαλιών, ομαδοποίηση και μείωση διαστάσεων.

Ενισχυτική Μάθηση

Στην ενισχυτική μάθηση, ο αλγόριθμος μηχανικής μάθησης αλληλεπιδρά με ένα περιβάλλον για να μάθει πώς να κάνει ενέργειες προσπαθώντας ταυτόχρονα να μεγιστοποιήσει διάφορες μεταβλητές "ανταμοιβής" αλλά και να ελαχιστοποιήσει μετρικές "ποινών". Ο αλγόριθμος μαθαίνει μέσω του πειραματισμού και των λαθών του λαμβάνοντας ανατροφοδότηση από το περιβάλλον σχετικά με την ποιότητα των ενεργειών του. Ο στόχος της ενισχυτικής μάθησης είναι η εκμάθηση μιας βέλτιστης πολιτικής δράσης που μεγιστοποιεί την αναμενόμενη σωρευτική ανταμοιβή με την πάροδο του χρόνου. Παραδείγματα προβλημάτων στα οποία χρησιμοποιείται ενισχυτική μάθηση είναι η ρομποτική, η αυτόνομη οδήγηση καθώς και τα παιχνίδια .

Αξίζει να αναφερθεί ότι υπάρχουν και άλλα είδη μηχανικής μάθησης, συνδυάζοντας διάφορα απο τα προαναφερόμενα είδη (π.χ. : Ημι-Επιβλεπόμενη Μάθηση) καθώς και άλλες πιο ειδικές κατηγορίες εκμάθησης που θα μελετήσουμε παρακάτω όπως βαθιά μάθηση (deep learning), μεταφορά μάθησης (transfer learning) κ.α.

2.1.3 Ταξινόμηση

Η ταξινόμηση είναι ένας τύπος εποπτευόμενης μάθησης στη μηχανική μάθηση που περιλαμβάνει εκπαίδευση ενός μοντέλου σε δεδομένα με ετικέτα για την πρόβλεψη της κατηγορίας που ανήκει μια νέα εισόδος (απο τις κατηγορίες που έχει εκπαιδευτεί το μοντέλο). Ο στόχος της ταξινόμησης είναι να μάθουμε το όριο που διαχωρίζει διαφορετικές κλάσεις στον διανυσματικό χώρο των χαρακτηριστικών. Τα δεδομένα εισόδου για προβλήματα ταξινόμησης αποτελούνται από δυο μέρη : ένα σύνολο δεδομένων ή χαρακτηριστικών που περιγράφουν την είσοδο και μια ετικέτα κλάσης που καθορίζει την κατηγορία ή την κλάση στην οποία ανήκει η είσοδος.

Για παράδειγμα, σε ένα πρόβλημα δυαδικής ταξινόμησης, ο στόχος είναι να ταξινομηθούν τα δεδομένα εισόδου σε μία από τις δύο πιθανές κατηγορίες, όπως *spam* ή *μη spam* ή θετικά / αρνητικά συναισθήματα. Στην ταξινόμηση πολλαπλών τάξεων, ο στόχος είναι να ταξινομηθούν τα δεδομένα εισόδου σε μία από πολλές πιθανές κατηγορίες, όπως διαφορετικοί τύποι φρούτων ή ζώων. Τα χαρακτηριστικά εισόδου θα μπορούσαν να είναι αριθμητικά ή κατηγορικά και ο ταξινομητής θα μπορούσε να είναι μια ποικιλία αλγορίθμων όπως η λογιστική παλινδρόμηση, τα δέντρα αποφάσεων, τα τυχαία δάση, οι μηχανές διανυσμάτων υποστήριξης, τα νευρωνικά δίκτυα και άλλα.

2.1.4 Μετρικές Αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων των αλγορίθμων μηχανικής μάθησης χρησιμοποιούνται διάφορες μετρικές κάθε μια με διαφορετική χρήση. Παρόλο που μια ολοκληρωμένη εικόνα της επίδοσης με μία μόνο μεμονωμένη μετρική θα ήταν βολική, δεν είναι εφικτό να καταγραφούν τα διαφορετικά στατιστικά χαρακτηριστικά της κατανομής των προβλέψεων σε μία μόνο μεταβλητή [19]. Ανάλογα με τους στόχους της εκάστοτε εφαρμογής υπάρχουν διαφορετικές μετρικές που αναπαριστούν τις στατιστικές ιδιομορφίες των προβλέψεων του μοντέλου. Αναφορικά μερικές απο αυτές είναι:

Accuracy

Η μετρική αξιολόγησης "accuracy" (ή ακρίβεια) είναι η πιο συχνά χρησιμοποιούμενη μέτρηση αξιολόγησης για μοντέλα ταξινόμησης. Μετρά το ποσοστό των σωστών προβλέψεων που έγιναν από το μοντέλο. Η μετρική accuracy υπολογίζεται ως ο λόγος του αριθμού των σωστών προβλέψεων προς τον συνολικό αριθμό των προβλέψεων. Είναι μια απλή και διαισθητική μέτρηση, αλλά μπορεί να μην είναι κατάλληλη για μη ισορροπημένα σύνολα δεδομένων ή όταν το κόστος των ψευδώς θετικών και των ψευδώς αρνητικών είναι διαφορετικό.

$$Accuracy = \frac{Αριθμός\ δεδομένων\ που\ έχουν\ προβλεφθεί\ σωστά}{Συνολικός\ αριθμός\ δεδομένων} \times 100$$

Precision & Recall

Οι μετρικές precision και recall είναι δύο μετρήσεις που χρησιμοποιούνται για την αξιολόγηση μοντέλων ταξινόμησης όταν οι κλάσεις είναι ανισορροπημένες. Το precision είναι το κλάσμα των αληθινών θετικών από όλες τις θετικές προβλέψεις, ενώ το recall είναι το κλάσμα των αληθινών θετικών από όλες τις πραγματικές θετικές περιπτώσεις. Η σημασία αυτών των μετρικών, σε αντίθεση με την απλή μετρική accuracy είναι όταν το κόστος ενός ψευδούς αρνητικού είναι πολύ μεγαλύτερο απο εκείνο ενός ψευδούς θετικού. Για παράδειγμα, στην διάγνωση μιας θανατηφόρας ασθένειας η ψευδής αρνητική διάγνωση θα κόστιζε την ζωή του ασθενούς (άρα χρειαζόμαστε μεγάλο recall). Το precision και το recall μπορούν να συνδυαστούν σε ένα ενιαίο σκορ που ονομάζεται F1-score, το οποίο είναι ένα αρμονικό μέσο και των δύο.

$$Precision = \frac{Αληθώς\ θετικά}{Αληθώς\ θετικά + Ψευδή\ θετικά}$$

$$Recall = \frac{Αληθώς\ θετικά}{Αληθώς\ θετικά + Ψευδή\ αρνητικά}$$

F-1 Score

Η βαθμολογία F1 είναι μια μέτρηση που συνδυάζει το precision και το recall σε μια ενιαία βαθμολογία. Είναι ο *αρμονικός μέσος* του precision και recall. Η βαθμολογία F1 είναι μια καλύτερη μέτρηση αξιολόγησης από την μετρική accuracy όταν αντιμετωπίζουμε μη ισορροπημένα σύνολα δεδομένων ή όταν το κόστος των ψευδώς θετικών και των ψευδώς αρνητικών είναι διαφορετικό. Μια υψηλότερη βαθμολογία F1 υποδηλώνει καλύτερη ισορροπία μεταξύ precision και recall.

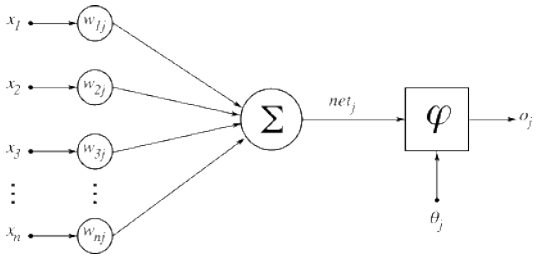
$$F_1\ Score = \frac{2}{Precision^{-1} + Recall^{-1}} = \frac{Precision \times Recall}{Precision + Recall} \times 2$$

2.2 Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα είναι μια κατηγορία μοντέλων μηχανικής μάθησης, όπως προαναφέρθηκε, εμπνευσμένα άμεσα απο τον ανθρώπινο εγκέφαλο. Αποτελούνται, απο διάφορους διασυνδεδεμένους "κόμβους", που ονομάζονται νευρώνες, οι οποιοί είναι διατεταγμένοι σε διαφορετικά επίπεδα. Τα νευρωνικά δίκτυα αποτελούνται απο ένα επίπεδο εισόδου, ένα (ή περισσότερα) κρυφό επίπεδο και ένα επίπεδο εξόδου. Το επίπεδο εισόδου λαμβάνει το διάνυσμα εισόδου που περιέχει την πληροφορία που θα επεξεργαστεί, το κρυφό επίπεδο εξάγει τα πολύπλοκα μοτίβα και επεξεργάζεται την πληροφορία και το επίπεδο εξόδου εξάγει το διάνυσμα του αποτελέσματος ανάλογα με το εκάστοτε πρόβλημα. Πιο ειδικά, στα Νευρωνικά Δίκτυα Πρόσθιας Ανατροφοδότησης (Feedforward Neural Networks) δεν δημιουργούνται κύκλοι μεταξύ των κόμβων και κάθε επίπεδο συνδέεται μόνο με το επόμενο.

Κάθε νευρώνας λαμβάνει ως είσοδο την έξοδο όλων των νευρώνων απο το προηγούμενο επίπεδο πολλαπλασιασμένη με κάποιο σχετικό βάρος. Αθροίζοντας όλα τα βάρη των νευρώνων απο το προηγούμενο επίπεδο μαζί με μια επιπλέον είσοδο που ονομάζεται πόλωση (bias), στη συνέχεια ο νευρώνας περνάει το αποτέλεσμα απο μια συνάρτηση ενεργοποίησης, με σκοπό την απογραμμικοποίηση της πληροφορίας, και έπειτα μεταβιβάζει στο επόμενο επίπεδο την πληροφορία με τον ίδιο τρόπο. Οι νευρώνες με αυτόν τον τρόπο μπορούν να βρίσκουν αφηρημένα μοτίβα και μη γραμμικές σχέσεις στα δεδομένα.

$$O_j = \varphi(\sum_{j=1}^{n} w_j x_j + \theta_j) \tag{2.1}$$



Σχήμα 2.1: **Τεχνητός νευρώνας**

Η ουσία της δυνατότητας των νευρωνικών να βρίσκουν μοτίβα στα δεδομένα βρίσκεται στη ικανότητα τους να προσαρμόζουν τα βάρη τους. Αυτό το καταφέρνουν περνώντας τα δεδομένα μέσα απο το δίκτυο μέσω μιας διαδικασίας που ονομάζεται πρόσθια διάδοση (forward propagation) όπου μετά με μια συνάρτηση κόστους υπολογίζουν πόσο αποκλίνουν απο τις αναμενόμενες τιμές μιας σωστής πρόβλεψης ή εξόδου. Έπειτα το σφάλμα πηγαίνει προς τα πίσω στο δίκτυο μέσω μιας διαδικασίας που λέγεται οπίσθια διάδοση (backward propagation) και τα βάρη και οι πολώσεις των νευρώνων προσαρμόζονται σε σχέση με τη συνεισφορά τους στο συνολικό σφάλμα. Τα βάρη προσαρμόζονται με βάση κάποιον αλγόριθμο βελτιστοποίησης ο οποίος προσπαθεί να ελαχιστοποιήσει την συνάρτηση κόστους όπως θα δούμε παρακάτω.

2.2.1 Συναρτήσεις Ενεργοποίησης

Οι συναρτήσεις ενεργοποίησης είναι ένα κρίσιμο στοιχείο των νευρωνικών δικτύων, καθώς εισάγουν μη γραμμικότητα στο μοντέλο και επιτρέπουν την αναγνώριση πολυπλοκότερων σχέσεων. Όπως προαναφέρθηκε, οι νευρώνες αθροίζουν την έξοδο των προηγούμενων νευρώνων,

πολλαπλασιασμένη με το σχετικό βάρος, και την πόλωση μέσω της εξίσωσης

$$\sum_{j=1}^{n} w_j x_j + \theta_j$$

. Η παραπάνω εξίσωση, όντας τελείως γραμμική δε μπορεί να μοντελοποιήσει μη γραμμικές σχέσεις μεταξύ των δεδομένων. Χωρίς την ύπαρξη συναρτήσεων ενεργοποίησης το συνολικό νευρωνικό δίκτυο θα μπορούσε να αναχθεί σε μια απλή πολυωνυμική εξίσωση πρώτου βαθμού, ανεξαρτήτως του πλήθους κρυφών επιπέδων. Αναφορικά τέσσερα απο τα πιο διαδεδομένα παραδείγματα συναρτήσεων ενεργοποιήσης είναι η ReLU, η Softmax, η Σιγμοειδής (Sigmoid) και η Υπερβολική Εφαπτομένη (Tanh).

ReLU

$$ReLU(x) = max(x, 0)$$

Η ReLU (Rectified Linear Unit) πρόκειται για μια μη γραμμική συνάρτηση που επιστρέφει την είσοδο αν είναι θετική και 0 αν είναι αρνητική. Πρόκειται για την πιο διαδεδομένη συνάρτηση ενεργοποίησης στα κρυφά επίπεδα, καθώς παρέχει πολλά πλεονεκτήματα σε αντίθεση με τις Sigmoid και Tanh.

Το πρώτο εμφανές πλεονέκτημά της είναι η αποδοτικότητα, αφενώς λόγω της υπολογιστικής της ευκολίας αλλά και του ότι όλοι οι νευρώνες δεν ενεργοποιούνται ταυτόχρονα αλλά μόνο ένα υποσύνολό τους. Ο αληθινός λόγος που χρησιμοποιείται όμως ως επί το πλείστον στα κρυφά επίπεδα, σε αντίθεση με την Σιγμοειδή ή την Υπερβολική Εφαπτομένη, είναι ότι αντιμετωπίζει το πρόβλημα των εξαφανιζόμενων κλίσεων (Vanishing Gradient Problem) καθώς έχει σταθερή κλίση 1 για θετικές εισόδους[20].

Ωστόσο, η ReLU έχει επίσης ορισμένα μειονεκτήματα. Ένα από τα κύρια μειονεκτήματα είναι ότι μπορεί να υποφέρει από το πρόβλημα «dying ReLU», όπου ορισμένοι νευρώνες μπορεί να γίνουν ανενεργοί και να σταματήσουν να ανταποκρίνονται σε οποιαδήποτε είσοδο και να έχουν μόνιμα μηδενική έξοδο. Αυτό μπορεί να συμβεί εάν τα βάρη αρχικοποιηθούν έτσι ώστε να παράγει πάντα αρνητική έξοδο ή εάν ο ρυθμός εκμάθησης είναι πολύ υψηλός, με αποτέλεσμα τα βάρη να ενημερώνονται με τρόπο ώστε ο νευρώνας να είναι πάντα αρνητικός.[21] Υπάρχουν τρόποι αντιμετώπισης του παραπάνω προβλήματος ένας εκ των οποίων είναι μια παραλλαγή της ReLU που ονομάζεται "Leaky ReLU".

Sigmoid

$$Sigmoid(x) = \frac{1}{1 + e^{-x}}$$

Η Σιγμοειδής (Sigmoid) πρόκειται για μια μη γραμμική συνάρτηση που αντιστοιχεί κάθε αριθμό $x \in \mathbb{R}$ σε μια τιμή μέσα στο $(0, 1)$. Είναι καταλληλότερη για χρήση στο επίπεδο εξόδου για προβλήματα δυαδικής ταξινόμησης αλλά και κανονικοποίησης εξόδων σε προβλήματα ταξινόμησης πολλαπλών κλάσεων [20].

Το πλεονέκτημα της Σιγμοειδούς στην χρήση τους στο επίπεδο εξόδου βρίσκεται στο γεγονός ότι αντιστοιχεί την είσοδο στις τιμές στο $(0, 1)$ που μπορεί να ερμηνευτεί ως πιθανότητα της εξόδου να ανήκει σε συγκεκριμένη κλάση ή κατηγορία. Επίσης η σιγμοειδής προσφέρει πλεονεκτήματα, λόγω της ομαλότητας και του ότι είναι συνεχώς διαφορίσιμη, στο να χρησιμοποιηθεί σε διάφορους αλγορίθμους βελτιστοποίησης βασισμένους σε μεθόδους υπολογισμού κλίσης.

Ένα απο τα μειονεκτήματα ωστόσο είναι το πρόβλημα των εξαφανιζόμενων κλίσεων, το οποίο εμφανίζεται όταν η κλίση της σιγμοειδούς συνάρτησης γίνεται πολύ μικρή για μεγάλες ή μικρές τιμές εισόδου. Αυτό μπορεί να δυσκολέψει την ενημέρωση των βαρών και των πολώσεων των νευρώνων στο δίκτυο κατά τη διάρκεια της εκπαίδευσης. Επιπλέον, η σιγμοειδής συνάρτηση δεν είναι μηδενική στο κέντρο, γεγονός που μπορεί να κάνει πιο δύσκολη τη σύγκλιση του αλγόριθμου βελτιστοποίησης.

Tanh

$$Tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Η Υπερβολική Εφαπτομένη (Tanh) πρόκειται για μια γραμμική συνάρτηση ενεργοποίησης, παραλλαγή της Σιγμοειδούς που αντιστοιχεί κάθε είσοδο στο στο σύνολο $(-1, 1)$. Η βασική διαφορά της με την Σιγμοειδή είναι ότι είναι κεντραρισμένη στο 0 και συμμετρική ως προς την αρχή την αρχή των αξόνων γεγονός που διευκολύνει την σύγκλιση του αλγορίθμου βελτιστοποίησης. Σε αντίθεση με την σιγμοειδή που χρησιμοποιείται στο επίπεδο εξόδου, η υπερβολική εφαπτομένη χρησιμοποιείται στο κρυφό επίπεδο και κυρίως σε δίκτυα Βραχυπρόθεσμης Μακροπρόθεσμης Μνήμης (LSTM). Ωστόσο, όπως και η Σιγμοειδής είναι εξίσου ευάλωτη στο πρόβλημα των εξαφανιζόμενων κλίσεων, ειδικά σε μεγάλες τιμές εισόδου, αλλά είναι και υπολογιστικά πιο απαιτητική απο την ReLU.

Softmax

$$Softmax(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \text{ για } j = 1, ..., K$$

Η Softmax πρόκειται για μια συνάρτηση ενεργοποίησης που χρησιμοποιείται συχνά σε προβλήματα ταξινόμησης πολλαπλών κλάσεων. Χρησιμοποιείται στο επίπεδο εξόδου και λαμβάνει ως είσοδο ένα διάνυσμα αληθινών αριθμών και το αντιστοιχεί σε μια κανονικοποιημένη πιθανοτική κατανομή με πλήθος τιμών ίσο με τον αριθμό των κλάσεων. Δηλαδή, λαμβάνει τις τελικές εξόδους απο το επίπεδο εξόδου ως αληθινούς αριθμούς και δίνει ως έξοδο για κάθε κλάση την πιθανότητα απο $(0, 1)$ να ανήκει στη συγκεκριμένη κλάση με όλες τις πιθανότητες να έχουν άθροισμα ίσο με 1. Ωστόσο, η softmax είναι ευάλωτη σε θορυβώδη δεδομένα και δεδομένα "outliers" [22]. Για παράδειγμα, έστω σε ένα πρόβλημα ταξινόμησης εικόνων στις κατηγορίες : Γάτα / Σκύλος, με την εισαγωγή μιας εικόνας "outlier" (π.χ. λιοντάρι), η λάθος ταξινόμηση του στοιχείου στην κατηγορία "Σκύλος" επηρεάζει την συνολική πιθανοτική κατανομή, αυξάνοντας ενδεχομένως την πιθανότητα μια τυχαία εικόνα να ταξινομηθεί ως σκύλος παρόλο που η εικόνα προφανώς αναπαριστά μια γάτα.

2.2.2 Συναρτήσεις Κόστους

Οι συναρτήσεις κόστους (loss functions) πρόκειται για μαθηματικές συναρτήσεις που έχουν στόχο την μοντελοποίηση και ποσοτικοποίηση της απόκλισης των τιμών προβλέψεων του μοντέλου απο τις αναμενόμενες τιμές. Πρόκειται για ένα ουσιαστικό μέρος των νευρωνικών δικτύων, καθώς παρέχουν μια μετρική της επίδοσης του νευρωνικού δικτύου και καθοδηγούν τον αλγόριθμο βελτιστοποίησης για την εύρεση του βέλτιστου συνόλου βαρών και πολώσεων που ελαχιστοποιούν την συνάρτηση κόστους. Υπάρχουν διάφορες συναρτήσεις κόστους ανάλογα με το είδος του προβλήματος που εξετάζεται και το είδος εξόδου που παράγει το μοντέλο. Αναφορικά μερικές απο αυτές είναι:

Mean-Squared Error

$$MSE(x) = \frac{1}{N}\sum_{i=1}^{n}(Y_{\text{true}} - Y_{\text{pred}})^2$$

Η συνάρτηση κόστους Μέσου Τετραγωνικού Λάθους (Mean-Squared Error) πρόκειται για μια ευρέως χρησιμοποιούμενη συνάρτηση κόστους στα νευρωνικά δίκτυα, ιδιαίτερα σε προβλήματα ανάλυσης παλινδρόμησης όπου η έξοδος είναι συνεχής τιμή. Η τιμή της ισούται με την τετραγωνική διαφορά της τιμής πρόβλεψης με την αναμενόμενη κανονικοποιημένη ως προς τον αριθμό δεδομένων. Προτέρημα της αποτελεί το γεγονός ότι παρέχει διαισθητική κατανόηση και γεωμετρική σημασία, ωστόσο είναι ευαίσθητη σε ακραίες τιμές[23]. Επίσης, η συνάρτηση Μέσου Τετραγωνικού Λάθους δεν έχει κάποια πιθανολογική ερμηνεία και ως εκ τούτου δεν είναι κατάλληλη για την χρήση της σε προβλήματα όπου η έξοδος είναι μια κατανομή πιθανότητας σε πολλές πιθανές κλάσεις.

Binary Cross-Entropy

$$BCE(x) = -y_{\text{true}} * log(y_{\text{pred}})$$

Η συνάρτηση Binary Cross-Entropy πρόκειται για μια συνάρτηση κόστους ευρέως χρησιμοποιούμενη σε προβλήματα δυαδικής ταξινόμησης. Προσφέρει μια μετρική της διαφοράς μεταξύ της εξόδου που έχει προβλέψει το μοντέλο και της αναμενόμενης εξόδου, που είναι μια δυαδική τιμή (0 ή 1). Έχει πιθανολογική ερμηνεία και δεν είναι τόσο ευάλωτη σε ακραίες τιμές όσο η MSE, ωστόσο δεν είναι κατάλληλη για προβλήματα πολλαπλών κλάσεων ή περιπτώσεις όπου τα δεδομένα δεν είναι ισορροπημένα.

Categorical Cross-Entropy

$$CCE(x) = -\sum_{i=1}^{N} y_{\text{true}_i} * log(y_{\text{pred}_i})$$

Η συνάρτηση Categorical Cross-Entropy πρόκειται για μια συνάρτηση κόστους ευρέως χρησιμοποιούμενη σε προβλήματα ταξινόμησης πολλαπλών κλάσεων. Παρέχει ακριβώς τα ίδια πλεονεκτήματα με την Binary Cross-Entropy για προβλήματα πολλαπλών κλάσεων. Απαιτεί ωστόσο, η μορφή των ετικετών των δεδομένων να είναι σε μορφή "one-hot encoding" όπου κάθε ετικέτα αναπαριστάται ως ένα διάνυσμα μήκους ίσο με τον αριθμό κλάσεων, με τιμή 1 στον δείκτη της κλάσης που ανήκει και 0 αλλού.

Sparse Cross-Entropy

$$SCE(x) = -\sum_{i=1}^{N} y_{\text{true}_i} * log(y_{\text{pred}_i})$$

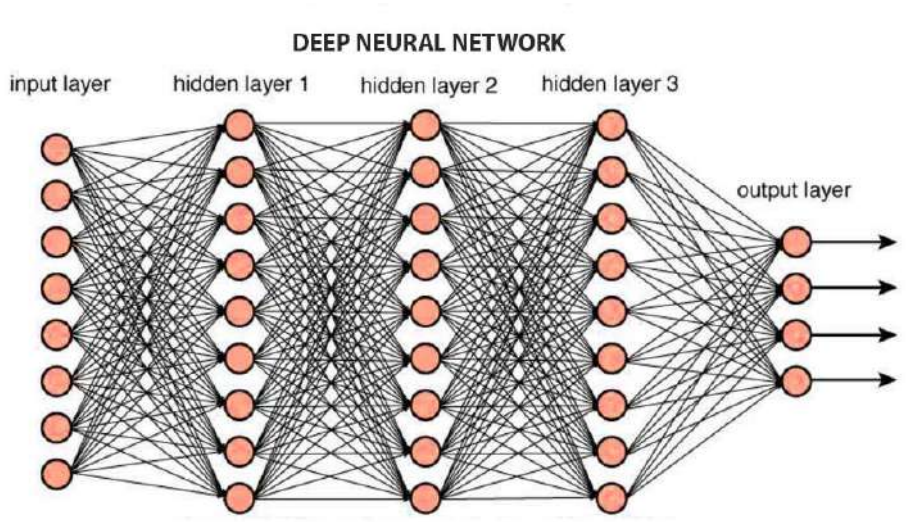
Η συνάρτηση κόστους Sparse Cross-Entropy είναι ακριβώς ίδια με την Categorical Cross-Entropy με την μόνη διαφορά να είναι στο τρόπο κωδικοποιήσης των ετικετών. Σε αντίθεση με την παραπάνω, δεν απαιτεί "one-hot encoding" αλλά μια οι ετικέτες κωδικοποιούνται ως ένας ακέραιος που αναπαριστά τον δείκτη της κλάσης που ανήκει. Ένα ενδεχόμενο πλεονέκτημα σε σχέση με την παραπάνω, είναι η μειωμένη απαίτηση μνήμης ειδικά σε υπερβολικά μεγάλα σύνολα δεδομένων με πολλές κλάσεις.

2.2.3 Βαθιά Μάθηση / Πολυεπίπεδα Νευρωνικά Δίκτυα

Η Βαθιά Μάθηση (Deep Learning) πρόκειται για ένα υποσύνολο της Μηχανικής μάθησης που προσπαθεί να ενσωματώσει υπολογιστικά μοντέλα και αλγορίθμους που προσπαθούν να προσωμοιώσουν όσο το δυνατόν καλύτερα τα βιολογικά νευρωνικά δίκτυα και διεργασίες που βρίσκονται στον ανθρώπινο εγκέφαλο (Τεχνητά Νευρωνικά Δίκτυα (ANN)) [24, 25]. Όποτε ο εγκέφαλος λαμβάνει νέες πληροφορίες, προσπαθεί να τις συγκρίνει με πληροφορίες που ήδη γνωρίζει για να προσπαθήσει να τις κατανοήσει. Ο εγκέφαλος προσπαθεί να αποκρυπτογραφήσει τις πληροφορίες μέσω της επισήμανσης και της τοποθέτησης των στοιχείων σε διάφορες κατηγορίες και η βαθιά μάθηση χρησιμοποιεί την ίδια νοοτροπία.

Ο όρος "βαθιά" στην βαθιά μάθηση είναι τεχνικός όρος και αναφέρεται στο πλήθος των επιπέδων (layers) που χρησιμοποιεί. Όπως θα δούμε και παρακάτω, υπάρχουν τρία είδη επιπέδων: το επίπεδο εισόδου που λαμβάνει τα δεδομένα εισόδου, το επίπεδο εξόδου που παράγει τα δεδομένα εξόδου (π.χ. την κλάση της ταξινόμησης των δεδομένων εισόδου) και τα κρυφά επίπεδα (hidden layers) που εξάγουν τα μοτίβα μέσα απο τα δεδομένα. Ένα βαθύ νευρωνικό δίκτυο (DNN) διαφέρει απο ένα κοινό νευρωνικό δίκτυο (ένα κρυφό επίπεδο μόνο) απο το γεγονός ότι έχει μεγάλο αριθμό κρυφών νευρώνων που του δίνουν την δυνατότητα να εξάγει πολύ πιο πολύπλοκα μοτίβα από τα δεδομένα[25]. Όσο τα δεδομένα προχωράνε απο ένα κρυφό επίπεδο σε άλλο, τα πιο απλά χαρακτηριστικά ανασυνδυάζονται και ανασυντίθενται σε πιο πολύπλοκα χαρακτηριστικά.

Η βαθιά μάθηση έχει επιτύχει σημαντικές ανακαλύψεις σε πολλούς τομείς, συμπεριλαμβανομένης της αναγνώρισης εικόνων, της επεξεργασίας φυσικής γλώσσας και της αναγνώρισης ομιλίας. Είναι ένα ισχυρό εργαλείο για την επίλυση σύνθετων προβλημάτων και έχει τη δυνατότητα να επαναστατήσει σε πολλούς κλάδους, συμπεριλαμβανομένης της υγειονομικής περίθαλψης, της χρηματοδότησης και των μεταφορών. Αυτό βασίζεται στο γεγονός ότι η βαθιά μάθηση έχει την δυνατότητα να εξάγει πολύ πιο λεπτομερή χαρακτηριστικά απο τις απλές μεθόδους Μηχανικής Μάθησης. Αυτό συμβαίνει βέβαια, με κόστος την απαίτηση πολύ μεγαλύτερου όγκου δεδομένων και επεξεργαστικής ισχύς κάτι που αποκαλούσε εμπόδιο στο παρελθόν για την επαρκή έρευνα και οικοδόμησή τους.



Σχήμα 2.2: **Παράδειγμα βαθιού νευρωνικού δικτύου**

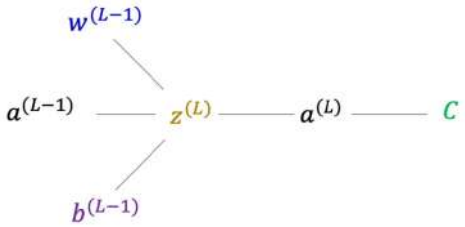
2.3 Εκπαίδευση Νευρωνικών Δικτύων

Όπως προαναφέρθηκε, με χρήση κατάλληλων βαρών, πολώσεων, πλήθος νευρώνων και επιπέδων και συναρτήσεις ενεργοποίησης τα τεχνητά νευρωνικά δίκτυα έχουν την δυνατότητα να εξάγουν πολύπλοκα μοτίβα απο διαφορετικά είδη δεδομένων. Ο στόχος της εκπαίδευσης είναι η εύρεση των κατάλληλων βαρών και πολώσεων για το εκάστοτε πρόβλημα. Έστω ότι η επίδοση ενός νευρωνικού δικτύου μοντελοποιείται μέσω μιας συνάρτησης κόστους, με την βέλτιστη επίδοση του δικτύου να βρίσκεται στο ολικό ελάχιστο της συνάρτησης κόστους. Τότε, η βελτίωση της επίδοσης επιτυγχάνεται με την εύρεση βαρών και πολώσεων που μειώνουν τη συνάρτηση κόστους. Η πραγματική καινοτομία στην επιστήμη των νευρωνικών δικτύων προήλθε απο την ανάπτυξη του αλγορίθμου οπίσθιας διάδοσης τη δεκαετία του 1960, που πρόσφερε έναν απλό τρόπο για την εύρεση και ανανέωση των βαρών και πολώσεων ενός δικτύου επαναληπτικά.

2.3.1 Οπίσθια Διάδοση Σφάλματος (Backpropagation)

Ο αλγόριθμος οπίσθιας διάδοσης (backpropagation) του σφάλματος είναι η πιο διαδεδομένη μορφή εκπαίδευσης νευρωνικών δικτύων πρόσθιας ανατροφοδότησης. Αυτό οφείλεται στο γεγονός ότι, αντίθετα με τους προκάτοχούς του, όπως τον κανόνα μάθησης του perceptron ή τον κανόνα μάθησης Widrow-Hoff, μπορεί να χρησιμοποιηθεί για την εκπαίδευση μη γραμμικών στοιχείων αυθαίρετης συνδεσιμότητας [26]. Δεδομένου ότι τέτοια δίκτυα απαιτούνται συχνά για εφαρμογές πραγματικού κόσμου, μια τέτοια διαδικασία εκμάθησης είναι κρίσιμη. Σχεδόν τόσο σημαντική όσο η δύναμή του, στο να εξηγήσει τη δημοτικότητά του, είναι η απλότητά του. Η βασική ιδέα είναι παλιά και απλή. Ο αλγόριθμος περιλαμβάνει τον ορισμό μιας συνάρτησης σφάλματος και την προσαρμογή κάθε βάρους και πόλωσης στο δίκτυο, με τον επαναληπτικό υπολογισμό προς τα πίσω της επίδρασης κάθε βάρους και πόλωσης στο συνολικό κόστος. Η απλότητα του αλγορίθμου κρύβεται στον τρόπο υπολογισμού της επίδρασης του κάθε βάρους / πόλωσης και χρήση απλών ιδιοτήτων διαφορικού λογισμού, όπως τον κανόνα αλυσίδας. Η αρχική ιδέα ξεκίνησε απο την δεκαετία του 1960, ωστόσο η διαδεδομένη του αποδοχή για την εκπαίδευση νευρωνικών δικτύων προήλθε το απο τους Rumelhart, Hinton, Williams που συνέβαλαν στην ανάπτυξη και βελτίωσή του[27].

Ο αλγόριθμος



Σχήμα 2.3: **Παράδειγμα ενός νευρώνα**

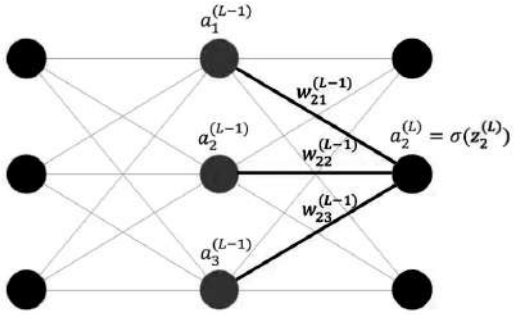
Στο σχήμα 2.3 φαίνεται πως ο κάθε όρος μεμονωμένα συμβάλει στον υπολογισμό του συνολικού κόστους C για μια συνδεσμολογία 1-1 νευρώνα. Για τον υπολογισμό του αθροίσματος $z^{(L)}$ υπολογίζουμε το άθροισμα του γινομένου του βάρους $w^{(L-1)}$ και της ενεργοποιήσης του προηγούμενου νευρώνα $a^{(L-1)}$ με την πόλωση $b^{(L-1)}$. Η ενεργοποίηση του νευρώνα $a^{(L)}$ θεωρούμε ότι γίνεται με κάποια μη γραμμική συνάρτηση $\sigma()$ και για λόγους απλότητας θεωρούμε ότι η συνάρτηση κόστους είναι εκείνη του μέσου τετραγωνικού σφάλματος.

$$z^{(L)} = w^{(L-1)}a^{(L-1)} + b^{(L-1)} \tag{2.2}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2.3}$$

$$C = (a^{(L)} - y)^2 \tag{2.4}$$

Επεκτείνοντας το πρόβλημα στη γενική περίπτωση ενός πολυεπίπεδου δικτύου έχουμε :



Σχήμα 2.4: **Παράδειγμα πολυεπίπεδου δικτύου**

$$z^{(L)} = \sum_{k=1}^{N_L-1} (w_{jk}^{(L-1)} a_k^{(L-1)}) + b^{(L-1)} \tag{2.5}$$

$$C = \sum_{j=1}^{N_L} (a_j^{(L)} - y_j)^2 \tag{2.6}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2.7}$$

Για τον υπολογισμό της συνεισφοράς ενός συγκεκριμένου βάρους ή πόλωσης, στη συνάρτηση κόστους ο υπολογισμός απλοποιείται με τη χρήση του κανόνα αλυσίδας.

$$\frac{\partial C}{\partial w_{jk}^{(L-1)}} = \frac{\partial z_j^{(L)}}{\partial w_{jk}^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C}{\partial a_j^{(L)}} = a_k^{(L-1)} \sigma'(z_j^{(L)}) 2(a_j^{(L)} - y_j) \tag{2.8}$$

$$\frac{\partial C}{\partial b_j^{(L-1)}} = \frac{\partial z_j^{(L)}}{\partial b_j^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C}{\partial a_j^{(L)}} = \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C}{\partial a_j^{(L)}} = \sigma'(z_j^{(L)}) 2(a_j^{(L)} - y_j) \tag{2.9}$$

$$\frac{\partial C}{\partial a_k^{(L-1)}} = \sum_{j=1}^{N_L} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_k^{(L-1)}}{\partial z_j^{(L)}} \frac{\partial C}{\partial a_k^{(L-1)}} \tag{2.10}$$

Αυτό που αλλάζει ωστόσο, στην γενίκευση του προβλήματος σε δίκτυο με αυθαίρετο αριθμό νευρώνων σε κάθε επίπεδο είναι ότι, όπως φαίνεται στο σχήμα 2.4, η ενεργοποίηση ενός νευρώνα στο προηγούμενο επίπεδο επηρεάζει το συνολικό κόστος μέσω κάθε νευρώνα στο τελευταίο επίπεδο (όπως βλέπουμε στην τελευταία εξίσωση). Η οπίσθια διάδοση του σφάλματος εμφανίζεται όταν αναζητείται η συνεισφορά των βαρών και πολώσεων, σε προηγούμενο επίπεδο, στο συνολικό κόστος.

$$\frac{\partial C}{\partial w_{jk}^{(L-2)}} = \frac{\partial z_j^{(L-1)}}{\partial w_{jk}^{(L-2)}} \frac{\partial a_j^{(L-1)}}{\partial z_j^{(L-1)}} \frac{\partial C}{\partial a_j^{(L-1)}} = \frac{\partial z_j^{(L-1)}}{\partial w_{jk}^{(L-2)}} \frac{\partial a_j^{(L-1)}}{\partial z_j^{(L-1)}} \sum_{j=1}^{N_L} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_k^{(L-1)}}{\partial z_j^{(L)}} \frac{\partial C}{\partial a_k^{(L-1)}} \tag{2.11}$$

$$\frac{\partial C}{\partial b_j^{(L-2)}} = \frac{\partial z_j^{(L-1)}}{\partial b_j^{(L-2)}} \frac{\partial a_j^{(L-1)}}{\partial z_j^{(L-1)}} \frac{\partial C}{\partial a_j^{(L-1)}} = \frac{\partial z_j^{(L-1)}}{\partial b_j^{(L-2)}} \frac{\partial a_j^{(L-1)}}{\partial z_j^{(L-1)}} \sum_{j=1}^{N_L} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_k^{(L-1)}}{\partial z_j^{(L)}} \frac{\partial C}{\partial a_k^{(L-1)}} \quad (2.12)$$

Με την παραπάνω μεθοδολογία, υπολογίζεται η συνεισφορά κάθε βάρους και πόλωσης στο συνολικό κόστος, η διαδικασία επαναλαμβάνεται για κάθε νευρώνα απο το τελικό επίπεδο προς τα πίσω.

2.3.2 Αλγόριθμοι Κατάβασης Κλίσης (Gradient Descent)

Στην προηγούμενη ενότητα παρουσιάστηκε η μέθοδος με την οποία διαδίδεται προς τα πίσω το σφάλμα για τον υπολογισμό της συνεισφοράς κάθε βάρους και πόλωσης στο συνολικό κόστος. Τα βάρη και οι πολώσεις ανανεώνονται με τιμή αναλογική με το αντίθετο της κλίσης της συνάρτησης κόστους.

$$w_{jk}^{(L-n)}(t+1) = w_{jk}^{(L-n)}(t) - \eta \frac{\partial C}{\partial w_{jk}^{(L-n)}(t)} \quad (2.13)$$

$$b_j^{(L-n)}(t+1) = b_j^{(L-n)}(t) - \eta \frac{\partial C}{\partial b_j^{(L-n)}(t)} \quad (2.14)$$

Στις παραπάνω εξισώσεις η τιμή η πρόκειται για μια υπερπαράμετρο που ονομάζεται ρυθμός μάθησης (learning rate). Στόχος της είναι να βοηθήσει στη σύγκλιση της συνάρτησης κόστους σε ένα ελάχιστο για το σύνολο των δεδομένων χωρίς να γίνει overshoot ή ταλάντωση γύρω απο την βέλτιστη λύση.

Ο αλγόριθμος κατάβασης κλίσης είναι ο αλγόριθμος βελτιστοποίησης που προσπαθεί να βρεί το ελάχιστο της συνάρτησης κόστους παίρνοντας βήματα στην αντίθετη κατεύθυνση απο εκείνη των κλίσεων.

Υπάρχουν τρία είδη του αλγόριθμου κατάβασης κλίσης με διαφορετικά προτερήματα ανάλογα με το μέγεθος του συνόλου δεδομένων, την πολυπλοκότητα του προβλήματος που εξετάζεται και την φύση των δεδομένων. Κάθε μία απο τις μεθόδους κάνει έναν συμβιβασμό μεταξύ την ακρίβεια της ενημέρωσης των παραμέτρων και τον χρόνο που απαιτείται για να πραγματοποιηθεί μια ενημέρωση[28].

Batch Gradient Descent

Το Batch Gradient Descent υπολογίζει την κλίση της συνάρτησης κόστους ως προς τις παραμέτρους θ ως προς όλο το σύνολο δεδομένων :

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta)$$

Καθώς απαιτείται να υπολογίσουμε τις κλίσεις για κάθε παράμετρο για όλο το σύνολο δεδομένων για να κάνουμε μόνο *μια* ενημέρωση, το Batch Gradient Descent είναι αργό και δύσχρηστο για μεγάλα σύνολα δεδομένων που δεν χωράνε στην μνήμη. Επίσης, δε μας δίνεται η δυνατότητα να ανανεώσουμε το μοντέλο παρουσιάζοντάς του νέα δεδομένα "εν-κινήσει". Ο αλγόριθμος ωστόσο είναι εγγυημένος ότι θα συγκλίνει στο ολικό ελάχιστο για κυρτές επιφάνειες σφάλματος και σε τοπικό ελάχιστο για μη κυρτές επιφάνειες [28].

Stochastic Gradient Descent

Σε αντίθεση η Στοχαστική Κατάβαση Κλίσης (Stochastic Gradient Descent / SGD) ανανεώνει τις παραμέτρους με *κάθε* δεδομένο εκπαίδευσης x^i και ετικέτα y^i.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^i; y^i)$$

Η Batch Gradient Descent, εκτελεί πλεονάζοντες υπολογισμούς για μεγάλα σύνολα δεδομένων, καθώς αναγκάζεται να ξαναυπολογίσει τις κλίσεις για παρόμοια παραδείγματα πριν την ανανέωση. Η SGD όμως ξεπερνάει αυτόν τον πλεονασμό εκτελώντας μία ανανέωση την φορά. Αυτό οδηγεί ωστόσο την SGD στο να έχει μεγάλη διακύμανση κατά την εκπαίδευση, με τον κίνδυνο να δυσκολέψει τη σύγκλιση οδηγώντας σε overshoot γύρω απο την βέλτιστη λύση. Ωστόσο, έχει φανεί ότι με την σταδιακή μείωση του ρυθμού μάθησης η SGD είναι σχεδόν σίγουρο να συγκλίνει σε ολικό ελάχιστο ή τοπικό για κυρτές και μη κυρτές επιφάνειες σφάλματος αντίστοιχα [28].

Mini-Batch Gradient Descent

Η Mini-Batch Gradient Descent πρόκειται για μια ένωση των δύο προηγούμενων αλγορίθμων, εκτελώντας ενημέρωση των παραμέτρων θ κάθε n δείγματα.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{i:i+n}; y^{i:i+n})$$

Με αυτό τον τρόπο μειώνεται η διακύμανση που προκαλείται απο την απλή SGD , οδηγώντας σε καλύτερη σύγκλιση. Αλλά ταυτόχρονα βελτιώνεται και η απόδοση και ταχύτητα της σύγκλισης σε σχέση με την Batch Gradient Descent. Πρόκειται για την πιο βασική μέθοδο εκπαίδευσης τεχνητών νευρωνικών δικτύων, και ο όρος SGD χρησιμοποιείται σε γενικές γραμμές όταν υλοποιούμε Mini-Batch Gradient Descent.

2.3.3 Αλγόριθμοι Βελτιστοποίησης Κατάβασης Κλίσης

Η Mini-Batch Gradient Descent, παρόλο που σε γενικές γραμμές προσφέρει καλή σύγκλιση, εμφανίζει αρκετές προκλήσεις. Αναφορικά

- Η επιλογή κατάλληλου ρυθμού μάθησης είναι δύσκολος. Λάθος επιλογή μεγάλου ρυθμού μάθησης οδηγεί ενδεχομένως σε αδυναμία σύγκλισης και ταλάντωση γύρω απο κάποιο τοπικό ελάχιστο και λάθος επιλογή μεγάλου ρυθμού μάθησης καθυστερεί πάρα πολύ τον χρόνο σύγκλισης.

- Ο ρυθμός μάθησης παραμένει ίδιος για όλες τις ενημερώσεις παραμέτρων, ανεξαρτήτως απο το ενδεχόμενο ότι οι συχνότητες εμφάνισης μερικών χαρακτηριστικών μπορεί να είναι διαφορετικές.

- Η εύρεση ενός ολικού ελάχιστου της συνάρτησης κόστους μπορεί να δυσκολευτεί απο το ενδεχόμενο να παγιδευτεί σε ένα απο τα πολυάριθμα τοπικά ελάχιστα. Εξίσου, μεγάλο εμπόδιο στην εύρεση ολικού ελάχιστου αποτελούν τα λεγόμενα "saddle points" που δεν αποτελούν ούτε τοπικό ελάχιστο ούτε ολικό, αλλά σημεία όπου η κλίση είναι μηδενική προς όλες τις κατευθύνσεις παγιδεύοντας την SGD κάνοντάς το δύσκολο να ξεφύγει.

Για την αντιμετώπιση των παραπάνω προβλημάτων έχουν επινοηθεί διάφοροι αλγόριθμοι και τεχνικές βελτιστοποίησης του αλγορίθμου κατάβασης κλίσης. Αναφορικά μερικοί:

Momentum

Η SGD έχει πρόβλημα στην πλοήγηση σε σημεία όπου η καμπύλες της επιφάνειας είναι πιο απότομες προς τα μία διάσταση σε σχέση με μια άλλη [29], που συμβαίνει συνήθως σε τοπικά ελάχιστα. Για την αντιμετώπιση του προβλήματος όπου η SGD ταλαντεύεται στις πλαγιές ενός τοπικού ελάχιστου έχει επινοηθεί η Momentum (Φόρα) που βοηθάει να επιταχύνει την SGD προς την σχετική κατεύθυνση και αποσβένει τις ταλαντώσεις. Μπορούμε να φανταστούμε ότι βοηθάει την SGD να αναπτύξει "φόρα" όσο κατεβαίνει τις πλαγιές της επιφάνειας σφάλματος και

να ξεπεράσει τα τοπικά ελάχιστα με ένα άνω όριο στην "ταχύτητα" που ορίζεται ως η μεταβλητή γ στην παρακάτω εξίσωση:

$$u_t = \gamma u_{t-1} + \eta \cdot \nabla_\theta J(\theta)$$
$$\theta = \theta - u_t$$

Nesterov Accelerated Gradient

Με την χρήση του "momentum" ωστόσο, η SGD επιταχύνει "τυφλά" κατεβαίνοντας μια πλαγιά της επιφάνειας της συνάρτησης κόστους, χωρίς να γνωρίζει τι έρχεται μετά. Θα ήταν επιθυμητό με κάποιον τρόπο να γνώριζε ότι μετά έρχεται άλλη πλαγιά για να επιβραδύνει. Αυτό ακριβώς κάνει η Nesterov Accelerated Gradient (NAG). Μέσω του $\theta - \gamma u_{t-1}$ εκτιμά την επόμενη θέση των παραμέτρων, μια ιδέα για το που θα βρίσκονται στην επόμενη επανάληψη. Έτσι, ο αλγόριθμος υπολογίζει την κλίση με βάση την επόμενη αναμενόμενη τιμή των παραμέτρων (προβλέποντας π.χ. ότι έρχεται λόφος) και βοηθάει στην επιτάχυνση της σύγκλισης και μειώνει την πιθανότητα "overshoot".

$$u_t = \gamma u_{t-1} + \eta \cdot \nabla_\theta J(\theta - \gamma u_{t-1})$$
$$\theta = \theta - u_t$$

Adagrad

Ο αλγόριθμος Adagrad [30] πρόκειται για μια μέθοδο προσαρμογής του ρυθμού μάθησης, εκτελώντας μεγαλύτερες ενημερώσεις για δεδομένα χαμηλότερης συχνότητας και μικρότερες ενημερώσεις για πιο υψίσυχνα δεδομένα. Ουσιαστικά, ο αλγόριθμος Adagrad διαιρεί τον ρυθμό μάθησης με την τετραγωνική ρίζα του αθροίσματος των τετραγώνων των προηγούμενων κλίσεων (μιας συγκεκριμένης παραμέτρου). Ο αλγόριθμος Adagrad είναι χρήσιμος για "αραιά" δεδομένα όπου ορισμένες παράμετροι μπορεί να έχουν πολλές ενημερώσεις ενώ άλλες έχουν λίγες ή καθόλου. Ωστόσο, ένα πιθανό μειονέκτημα του Adagrad είναι ότι ο ρυθμός εκμάθησης μπορεί να γίνει πολύ μικρός με την πάροδο του χρόνου, ιδιαίτερα για παραμέτρους με μικρές κλίσεις.

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \varepsilon}} \cdot g_{t,i}$$

Το $G_t \in \mathbb{R}^{d \times d}$ είναι ένας διαγώνιος πίνακας, όπου κάθε διαγώνιο στοιχείο είναι το άθροισμα των τετραγώνων των κλίσεων έως την χρονική στιγμή t, ενώ το ε είναι ένας όρος (της τάξης του $1e-8$) για την αποφυγή της διαίρεσης με το μηδέν.

Adadelta

Ο αλγόριθμος Adadelta [31] πρόκειται για μια επέκταση του αλγορίθμου Adagrad, που έχει στόχο να αντιμετωπίσει την μονοτονική φθίνουσα φύση του. Αντί να αθροίζει όλα τα προηγούμενα τετράγωνα των κλίσεων, ο Adadelta υπολογίζει μόνο ένα κινούμενο παράθυρο w κλίσεων.

Ο Adadelta προσαρμόζει τον ρυθμό εκμάθησης κάθε παραμέτρου χρησιμοποιώντας δύο πρόσθετες μεταβλητές: έναν τρέχοντα μέσο όρο των τετραγωνικών ενημερώσεων και έναν τρέχοντα μέσο όρο των τετραγωνικών κλίσεων. Αντί να αποθηκεύει αναποτελεσματικά w προηγούμενες τετραγωνισμένες κλίσεις, το άθροισμα των κλίσεων αναδρομικά ορίζεται ως ένας φθίνων μέσος όρος όλων των προηγούμενων κλίσεων. Ο κινούμενος μέσος $E[g^2]_t$ την χρονική στιγμή t εξαρτάται μόνο στον προηγούμενο μέσο (πολλαπλασιασμένο με έναν όρο γ όπως και στο "Momentum") και στην τωρινή κλίση :

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma) g_t^2$$

$$\Delta\theta_t = -\frac{\eta}{\sqrt{E[g^2]_t + \varepsilon}} g_t$$

Adam

Ο αλγόριθμος Adam (Adaptive Moment Estimation) [32] είναι μια άλλη μέθοδος που υπολογίζει προσαρμοσμένους ρυθμούς μάθησης για κάθε παράμετρο. Πέρα απο την αποθήκευση των φθίνων προηγούμενων τετραγωνισμένων κλίσεων u_t όπως ο αλγόριθμος Adadelta, ο Adam αποθηκεύει και τον φθίνων μέσο όρο των προηγούμενων κλίσεων m_t, όμοια με το momentum:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$u_t = \beta_2 u_{t-1} + (1 - \beta_2) g_t$$

Το m_t και το u_t είναι εκτιμήσεις του μέσου και της μη κεντραρισμένης διακύμανσης της κλίσης. Καθώς το m_t και το u_t έχουν αρχικοποιηθεί ως μηδενικά διανύσματα, τείνουν να είναι πολωμένα προς το 0, ειδικά όταν οι όροι β_1, β_2 είναι κοντά στο 1. Για την αντιστάθμιση αυτής της πόλωσης υπάρχουν οι εξής όροι :

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{u}_t = \frac{u_t}{1 - \beta_2^t}$$

Έπειτα οι μεταβλητές προσαρμόζονται με παρόμοιο τρόπο όπως και στους άλλους αλγορίθμους βελτιστοποίσης όπως ο Adagrad :

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{u}_t} + \varepsilon} \cdot \hat{m}_t$$

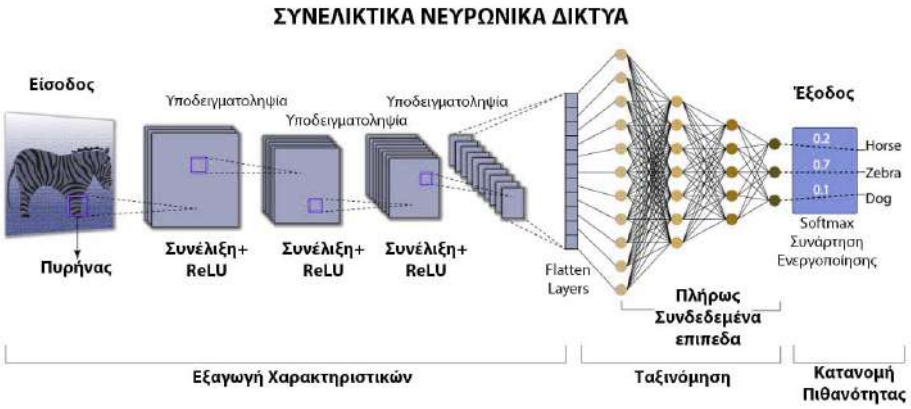
Ο αλγόριθμος Adam είναι απο τους πιο σύνηθης αλγορίθμους βελτιστοποίησης καθώς παρέχει γρηγορότερη σύγκλιση και ανθεκτικότητα στις θορυβώδεις κλίσεις σε πολλά διαφορετικά προβλήματα.

2.4 Ταξινόμηση Εικόνων

2.4.1 Δισδιάστατα Συνελικτικά Νευρωνικά Δίκτυα

Τα Συνελικτικά Νευρωνικά Δίκτυα [33, 34] έχουν πρωτοποριακά αποτελέσματα την τελευταία δεκαετία σε διάφορους τομείς αναγνώρισης μοτίβων, από επεξεργασία φωνής σε αναγνώριση μοτίβων σε εικόνες. Η πιο ωφέλιμη πτυχή των Συνελικτικών Νευρωνικών Δικτύων είναι ότι μπορούν να μειώσουν σε μεγάλο βαθμό τον απαιτούμενο αριθμό των παραμέτρων σε σχέση με κλασσικές μορφές τεχνητών νευρωνικών δικτύων ταξινόμησης. Αυτό έχει επιτρέψει σε ενδιαφερόμενους ερευνητές να αναπτύξουν μεγαλύτερα μοντέλα για πολύπλοκα προβλήματα, που δε θα ήταν εφικτό με κλασσικά τεχνητά νευρωνικά δίκτυα. Ένα άλλο ενδεχόμενο πλεονέκτημά τους είναι η δυνατότητα εξαγωγής χωρικών χαρακτηριστικών απο τις εικόνες, ανεξαρτήτως της θέσης τους πάνω στην εικόνα. Μια ακόμα σημαντική πτυχή των Συνελικτικών Νευρωνικών Δικτύων είναι η δυνατότητα να εξάγουν αφηρημένα χαρακτηρηστικά και μοτίβα όσο είσοδος κάνει οπίσθια διάδοση προς τα βαθύτερα επίπεδα. Για παράδειγμα, στην ταξινόμηση εικόνων, μπορεί τα πρώτα επίπεδα να εντοπίζουν χαρακτηριστικά όπως ακμές, μετά στο δεύτερο επίπεδο κάποια απλά χαρακτηριστικά, και στα τελευταία επίπεδα χαρακτηριστικά όπως πρόσωπα.

Η σημασία της συνέλιξης γίνεται κατανοήτη αν θεωρήσει κανείς τον αριθμό των βαρών που θα απαιτούνταν για την ένωση μια εικόνας με έναν μόνο νευρώνα. Μια εικόνα , έστω μεγέθους $64 \times 64 \times 3$, θα απαιτούσε $64 \times 64 \times 3 = 12,288$ βάρη για την ένωσή της με μόνο έναν νευρώνα.



Σχήμα 2.5: **Παράδειγμα Συνελικτικού Νευρωνικού Δικτύου**

Ένα κρυφό επίπεδο με 16×16 νευρώνες θα απαιτούσε $3,145,728$ βάρη, χωρίς να έχει την δυνατότητα να ταξινομήσει τίποτα ουσιώδες.

Αναζητώντας μια πιο αποδοτική λύση, μια ιδέα για την μείωση των ενώσεων θα ήταν κάθε νευρώνας να "κοιτάει" τοπικές περιοχές στην εικόνα αντί για ολόκληρη την εικόνα. Δηλαδή, έστω ότι υπάρχουν 16×16 νευρώνες στο επόμενο επίπεδο, θα μπορούσε κάθε νευρώνας να κοιτάει μια περιοχή 5×5 και λοιπόν να έχουμε $5 \times 5 \times 3$ με 16×16 ενώσεις, δηλαδή $19,200$ βάρη [35, 36]. Παρόλο που μειώθηκε σημαντικά το πλήθος των παραμέτρων, παραμένουν ακόμα πολλές παράμετροι για τον υπολογισμό τους.

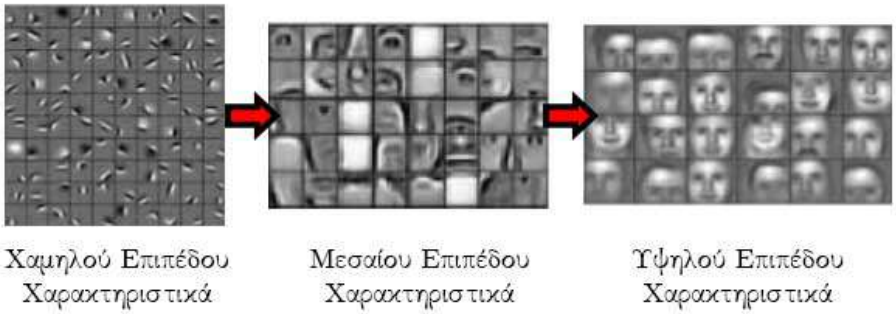
Μια ακόμη ιδέα θα ήταν τα τοπικά βάρη να παραμείνουν σταθερά για κάθε νευρώνα του επόμενου επιπέδου το οποίο θα μείωνε τις απαιτούμενες ενώσεις σε $16 \times 16 = 256$ (από $3,145,728$ αρχικά). Αυτή η ιδέα ισοδυναμεί με την μετακίνηση ενός παραθύρου $5 \times 5 \times 3$ πάνω στην εικόνα και αντιστοίχηση της παραγώμενης εξόδου στην αντίστοιχη θέση. Αυτό παρουσιάζει την ευκαιρία για την αναγνώριση μοτίβων (και επεξεργασία της εικόνας) ανεξαρτήτως της θέσης τους, και για αυτό ονομάζεται συνέλιξη.

Επίπεδο Συνέλιξης

Τα κινούμενα παράθυρα ονομάζονται φίλτρα καθώς έχουν την ίδια μορφή με τα κλασσικά φίλτρα στην επεξεργασία εικόνων. Ωστόσο, σε αντίθεση με τα κλασσικά φίλτρα, τα μοτίβα που είναι ικανά να εντοπίζουν καθορίζονται απο την διαδικασία της εκπαίδευσης και όχι χειροκίνητα. Με την τοποθέτηση πολλών επιπέδων νευρώνων παράλληλα, στο επίπεδο μετά το επίπεδο εισόδου, δίνεται η δυνατότητα το δίκτυο να περιέχει πολλά φίλτρα ανά επίπεδο για την εύρεση διαφορετικών χαρακτηριστικών ανά επίπεδο.

Κάθε φίλτρο, κατά την συνέλιξη με την εικόνα, περνάει πάνω απο την εικόνα με ένα βήμα, που ονομάζεται **stride**, και το αποτέλεσμα κάθε συνέλιξης με ένα από τα φίλτρα δημιουργεί έναν χάρτη χαρακτηριστικών (**feature map**). Για τον εντοπισμό χαρακτηριστικών που ενδεχομένως να βρίσκονται στα όρια της εικόνας, τα όρια της εικόνας συμπληρώνονται με μηδενικά, διαδικασία που ονομάζεται **zero-padding**.

Καθώς οι πράξεις συνέλιξης είναι γραμμμικές είναι σημαντικό, όπως έχει αναφερθεί σε προηγούμενες ενότητες, να εισάγουμε μη γραμμικότητα στο μοντέλο αλλιώς όλα τα επίπεδα συνελίξεων θα αντιστοιχούσαν με ένα μόνο ισοδύναμο επίπεδο. Για αυτό το λόγο, χρησιμοποιούνται συναρτήσεις ενεργοποίησης μετά την συνέλιξη για την απογραμμικοποίηση, με την πιο σύνηθη να είναι η ReLU λόγω του γεγονότος ότι αντιμετωπίζει το πρόβλημα των εξαφανιζόμενων κλίσεων.



Σχήμα 2.6: **Ενεργοποιήσεις Επίπεδων Συνέλιξης**. Τα πρώτα επίπεδα συνέλιξης βρίσκουν πιο χαμηλού επιπέδου μοτίβα και τα επόμενα επίπεδα μπορούν να εντοπίσουν πιο υψηλού επιπέδου χαρακτηριστικά.

Επίπεδο Υποδειγματοληψίας

Έπειτα, τα επίπεδα συνελίξεων υποδειγματοληπτούνται με μια διαδικασία που ονομάζεται **Pooling**. Τα επίπεδα υποδειγματοληψίας έχουν στόχο την μείωση των διαστάσεων της αναπαράστασης και έτσι να μειώσει το πλήθος παραμέτρων και συνεπώς την υπολογιστική πολυπλοκότητα του μοντέλου. Στη διαδικασία Pooling περνάνε παράθυρα (συνήθως μεγέθους 2×2) πάνω απο την εικόνα (με βήμα συνήθως 2) τα οποία υποδειγματοληπτούν την εικόνα κρατώντας είτε την μέγιστη τιμή (**Max-Pooling**) ή την μέση τιμή (**Average-Pooling**).

Πέραν της μείωσης των διαστάσεων του μοντέλου η υποδειγματοληψία προσφέρει ανθεκτικότητα σε μικρές μεταθέσεις της εισόδου[37]. Με τον υπολογισμό της μέσης ή της μέγιστης τιμής σε μια περιοχή, το Pooling διασφαλίζει ότι οι μικρές διακυμάνσεις στην είσοδο δεν επηρεάζουν την έξοδο του δικτύου.

Παρόλο που λόγω των επιπέδων υποδειγματοληψίας χάνεται ένα ποσοστό των τοπικών δεδομένων [38], έχει ωστόσο το αποτέλεσμα το μοντέλο να μπορεί να κρατήσει υψηλότερου επιπέδου χαρακτηριστικά και μοτίβα. Χωρίς τα Pooling επίπεδα, τα συνελικτικά νευρωνικά δίκτυα θα ήταν πιο ευάλωτα στην υπερβολική προσαρμογή (overfitting) του συνόλου των δεδομένων καθώς και θα ήταν υπολογιστικά πιο πολύπλοκη η εκπαίδευση του δικτύου.

Πλήρως Συνδεδεμένα Επίπεδα

Ο συνδυασμός επιπέδων συνελίξεων για την εξαγωγή χαρακτηρηστικών, συναρτήσεων ενεργοποίησης για την απογραμμικοποίηση και υποδειγματοληψίας επαναλαμβάνεται ξάνα (συνήθως 1-2 φορές) για την εξαγωγή πιο λεπτομερών χαρακτηριστικών. Για την ταξινόμηση των εικόνων σε κατηγορίες οι χάρτες χαρακτηριστικών απο τρισδιάστατους τένσορες (Ύψος, Πλάτος, Βάθος) μετατρέπονται σε μονοσδιάστατους τένσορες μεγέθους Ύψος $\times$ Πλάτος $\times$ Βάθος. Η χωρική πληροφορία διατηρείται και μπορεί να μοντελοποιηθεί απο τα πλήρως συνδεδεμένα επίπεδα καθώς διατηρείται η σειρά των στοιχείων στον τένσορα. Στη συνέχεια ο μονοσδιάστατος τένσορας περνάει ως είσοδο σε ένα κλασσικό νευρωνικό δίκτυο πρόσθιας ανατροφοδότησης, όπως έχουμε αναφέρει σε προηγούμενες ενότητες, με έξοδο το πλήθος κλάσεων ταξινόμησης των εικόνων.

2.4.2 Προεπεξεργασία Δεδομένων

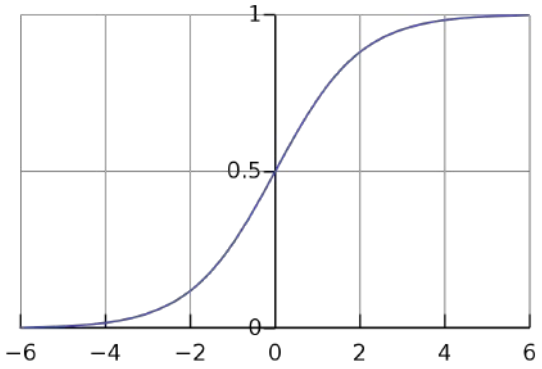
Η προεπεξεργασία δεδομένων, ειδικά εικόνων, είναι ουσιαστικό βήμα κατά την εκπαίδευση των Συνελικτικών Νευρωνικών Δικτύων και επηρεάζει σημαντικά την επίδοση του μοντέλου.

Αφενός, ακατέργαστα δεδομένα δεν παράγουν μοντέλα με καλή ακρίβεια, αλλά η επεξεργασία των δεδομένων μπορεί να είναι και ένας τρόπος να αυξήσουμε την ποικιλομορφία του συνόλου δεδομένων.

Κανονικοποίηση Δεδομένων Εισόδου

Η κανονικοποίηση των τιμών εισόδου, ειδικά όταν πρόκειται για εικόνες, είναι ένα σημαντικό βήμα για την βελτιστοποίηση της επίδοσης των νευρωνικών δικτύων για διάφορους λόγους. Αρχικά, το κάθε pixel μπορεί να έχει ευρύ φάσμα τιμών, το οποίο μπορεί να οδηγήσει σε αριθμητική αστάθεια αν δεν κανονικοποιηθεί. Με την κανονικοποίηση των τιμών των pixels, μπορούμε να διασφαλιστεί ότι οι παράμετροι παραμένουν εντός λογικών ορίων και αποτρέπονται αριθμητικά προβλήματα, όπως το συσσωρευμένο σφάλμα στρογγυλοποίησης και οδηγεί σε πιο σταθερή διαδικασία εκπαίδευσης [39].

Επιπλέον, η κανονικοποίηση των τιμών βοηθάει την σύγκλιση των αλγορίθμων κατάβασης κλίσης, μειώνοντας την αριθμητική διακύμανση μεταξύ των κλίσεων, το οποίο οδηγεί σε γρηγορότερη σύγκλιση. Τέλος, η κανονικοποίηση μπορεί επίσης να βοηθήσει στην απόδοση των ενεργοποιήσεων του νευρωνικού δικτύου.



Σχήμα 2.7: **Σιγμοειδής Συνάρτηση Ενεργοποίησης**

Όπως βλέπουμε παραπάνω, η Σιγμοειδής μπορεί να κορεστεί όταν τα δεδομένα εισόδου έχουν πολύ μεγάλες ή πολύ μικρές τιμές, καθώς η ενεργοποιήσεις των νευρώνων οδηγούνται στα άκρα της σιγμοειδούς που η κλίση είναι πολύ μικρή, επιβραδύνοντας την εκπαίδευση.

$$x^{'} = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

Παραπάνω βλέπουμε ένα παράδειγμα κανονικοποίησης γύρω απο το $[0, 1]$, ωστόσο υπάρχουν πολλοί τρόποι κανονικοποίησης των δεδομένων εκλαμβάνοντας στατιστικά χαρακτηριστικά της κατανομής των τιμών εισόδου. Ο πιο διαδεδομένος από αυτούς ονομάζεται *Z-Score Normalisation* και έχει στόχο την κλιμάκωση της εισόδου έτσι ώστε να έχει μέση τιμή $\mu = 0$ και τυπική απόκλιση $std = 1$.

$$x^{'} = \frac{(x - \mu)}{\sigma}$$

Κανονικοποιήση Παρτίδας / Batch Normalisation

Όπως προαναφέρθηκε, η κανονικοποίηση των δεδομένων εισόδου οδηγούν σε μοντέλα με καλύτερη επίδοση, γρηγορότερη σύγκλιση και εξάλειψη του προβλήματος κορεσμού των συναρτήσεων ενεργοποίησης. Ωστόσο, το πρόβλημα εκπαίδευσης των νευρωνικών δικτύων περιπλέκεται απο

το γεγονός ότι η κατανομή της εισόδου κάθε επιπέδου αλλάζει, όσο οι παράμετροι του προηγούμενου επιπέδου αλλάζουν. Αυτό το φαινόμενο αναφέρεται ως *Internal Covariate Shift*, και αντιμετωπίζεται με την κανονικοποιήση των εισόδων κάθε επιπέδου [40]. Το παραπάνω φαινόμενο αυξάνεται όσο πιο βαθύ είναι το νευρωνικό δίκτυο.

Παρόλο που στην πράξη το πρόβλημα κορεσμού των συναρτήσεων ενεργοποίησης και εξαφανιζόμενων κλίσεων αντιμετωπίζεται με χρήση συναρτήσεων ενεργοποίησης ReLU, προσεκτική αρχικοποιήση παρεμέτρων και μικρών ρυθμών εκμάθησης, αν ωστόσο ήταν δυνατή η εξασφάλιση ότι η κατανομή των μη γραμμικών εισόδων παραμένει πιο σταθερή κατά την διάρκεια της εκπαίδευσης και ο αλγόριθμος βελτιστοποίησης θα ήταν λιγότερο πιθανό να επηρεάζεται απο τον κορεσμό και η εκπαίδευση θα επιταχυνόταν [40].

Μια λύση στο παραπάνω πρόβλημα είναι η χρήση επίπεδων κανονικοποίησης παρτίδας (batch normalisation) που βρίσκονται πριν την είσοδο κάθε επιπέδου και κανονικοποιεί της εισόδους διορθώνοντας την μέση τιμή και τυπική απόκλισή τους. Η κανονικοποίηση παρτίδας έχει επίσης θετική επίδραση στη ροή της κλίσης στο δίκτυο, μειώνοντας την εξάρτηση των κλίσεων στην κλίμακα των παραμέτρων ή των αρχικών τους τιμών. Αυτό επιτρέπει την χρήση μεγαλύτερων ρυθμών εκπαίδευσης χωρίς κίνδυνο απόκλισης. Επίσης, η κανονικοποίηση παρτίδας κανονικοποιεί το μοντέλο και μειώνει την ανάγκη για χρήση επιπέδων "Dropout". Τέλος, η κανονικοποίηση παρτίδας μας επιτρέπει την χρήση συναρτήσεων ενεργοποίησης ευάλωτων στον κορεσμό, όπως η σιγμοειδής, εξαλείφοντας τον κίνδυνο εξαφανιζόμενων κλίσεων και χαμηλού ρυθμού εκπαίδευσης [40].

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{2.15}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{2.16}$$

Οι παραπάνω τιμές μ_B και σ_B αντιπροσωπεύουν την μέση τιμή και τυπική απόκλιση των τιμών εισόδου ενός m-διάστατου επιπέδου $B = x_{1...m}$.

$$\hat{x_i} = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{2.17}$$

$$y_i = \gamma \hat{x_i} + \beta \tag{2.18}$$

Στις παραπάνω εξισώσεις οι τιμές x_i κανονικοποιούνται σε $\hat{x_i}$ και έπειτα μέσω των μεταβλητών γ, β, οι οποίες μαθαίνονται κατά την διάρκεια της εκπαίδευσης, οι κανονικοποιημένες είσοδοι $\hat{x_i}$ μετατοπίζονται και κλιμακώνονται και κάθε επίπεδο λαμβάνει ως είσοδο την τιμή y_i.

Αύξηση Δεδομένων / Data Augmentation

Η πολυπλοκότητα του προβλήματος, που καλείται να μοντελοποιήσει το εκάστοτε δίκτυο, συσχετίζεται άμεσα με την ποσότητα δεδομένων που απαιτούνται. Αρκετά πολύπλοκα προβλήματα δηλαδή, απαιτούν εκατοντάδες χιλιάδες δείγματα εισόδου τα οποία πολλές φορές περιέχουν έντονη πληροφορία μη σχετική με τα μοτίβα που αναζητούνται. Αν τα δεδομένα δεν αναπαριστούν καλά το πρόβλημα που εξετάζεται, τότε το μοντέλο είναι ευάλωτο σε υπερβολική προσαρμογή (overfitting). Η αύξηση δεδομένων (data augmentation) είναι ένας τρόπος εισαγωγής μικρών διακυμάνσεων και διαστρεβλώσεων των δεδομένων εισόδου με στόχο την άυξηση του πλήθους των δεδομένων εισόδου και την βελτίωση της αναπαράστασης του προβλήματος στα δεδομένα.

Τα δεδομένα εισόδου διαστρεβλώνονται με διάφορους τρόπους, για παράδειγμα αν πρόκειται για αρχείο ήχου με την εισαγωγή λευκού θορύβου ή αν πρόκειται για εικόνες με την τυχαία στροφή, αλλαγή χρώματος, ζουμάρισμα, κόψιμο της εικόνας κ.α. . Με αυτές τις μετατροπές

αυξάνεται το πλήθος των δεδομένων και αυξάνουν την ανθεκτικότητα του μοντέλου σε διαφορετικές συνθήκες (όπως π.χ. : προσανατολισμός, φωτισμός, θέση κ.α.). Ένα άλλο πλεονέκτημα είναι ότι με την εισαγωγή θορύβου και μεταβλητότας στα δεδομένα είναι πολύ δύσκολο το μοντέλο να απομνημονευτεί όλο το σύνολο δεδομένων, μειώνοντας την πιθανότητα υπερβολικής προσαρμογής και βοηθαέι στην γενίκευση. Παρακάτω βλέπουμε μερικά παραδείγματα αύξησης δεδομένων σε εικόνες.

Σχήμα 2.8: **Παραδείγματα αύξησης δεδομένων**. Ξεκινώντας απο πάνω αριστερά έχουμε την αρχική εικόνα, εισαγωγή θόρυβου στο χρώμα, τυχαία στροφή, τυχαία αλλαγή φωτεινότητας

2.4.3 Μεταφορά Μάθησης (Transfer Learning)

Η μεταφορά μάθησης είναι μια τεχνική στη μηχανική και την βαθιά μάθηση κατά την οποία ένα προεκπαιδευμένο μοντέλο σε ένα σχετικό πρόβλημα χρησιμοποιείται ως αφετηρία για ένα νέο πρόβλημα. Η αρχική ιδέα για την μεταφορά μάθησης μπορεί να προέρχεται, όπως πολλοί τομείς της μηχανικής μάθησης, απο την ψυχολογία. Σύμφωνα με τη θεωρία γενίκευσης της μεταφοράς, όπως προτείνεται από τον ψυχολόγο C.H. Judd, η δυνατότητα μεταφοράς της γνώσης (από έναν τομέα σε έναν άλλο) είναι το αποτέλεσμα της γενίκευσης της εμπειρίας. Είναι δυνατό να μεταφερθεί η γνώση απο έναν τομέα σε έναν άλλο δηλαδή, αν το άτομο προσπαθήσει να γενικεύσει την εμπειρία του. Σύμφωνα με αυτή την θεωρία, προαπαιτούμενο είναι να υπάρχει κάποια σύνδεση μεταξύ των δύο διαφορετικών δραστηριοτήτων μάθησης. Για παράδειγμα, ένας

άνθρωπος που έχει μάθει να παίζει το βιολί μπορεί να εκπαιδευτεί πιο γρήγορα στο να παίζει το πιάνο (Positive Transfer), καθώς και τα δύο είναι μουσικά όργανα και μπορεί να έχουν κοινή απαιτούμενη γνώση. Ωστόσο, απο την άλλη υπάρχει η περίπτωση να έχουμε το αντίθετο αποτέλεσμα (Negative Transfer). Για παράδειγμα, η γνώση της ισπανικής γλώσσας μπορεί να επηρεάσει αρνητικά την εκμάθηση της γαλλικής λόγω έντονων διαφορών στη σύνταξη της γραμματικής, ρημάτων κτλ.

Στον τομέα της επιβλεπόμενης μάθησης στα τεχνητά νευρωνικά δίκτυα, η μεταφορά μάθησης έχει οδηγήσει πολλούς τομείς προβλημάτων ταξινόμησης, ειδικά σε προβλήματα ταξινόμησης εικόνων, σε ραγδαία πρόοδο. Προσπάθειες, όπως ο διαγωνισμός ILSVRC αλλά και η δημιουργία τεράστιων συνόλων δεδομένων όπως το ImageNet, MNIST και δεκάδες άλλα, έχουν δείξει ότι η εκπαίδευση ενός μοντέλου σε ένα τεράστιο σύνολο δεδομένων με εκατοντάδες κατηγορίες, βοηθάει και διευκολύνει αρκετά την γενίκευση του μοντέλου κατά την εκπαίδευσή του σε μια άλλη κατηγορία. Για πολλά απο τα πιο πετυχημένα μοντέλα του διαγωνισμού ILSVRC μάλιστα (όπως InceptionNet, AlexNet, ResNet, MobileNet κ.α.) δίνεται δυνατότητα χρήσης τους απο το κοινό μέσω βιβλιοθηκών, προεκπαιδευμένα σε τεράστια σύνολα δεδομένων (όπως ImageNet).

Η μεταφορά μάθησης για την κατηγοριοποίηση εικόνων στα συνελικτικά νευρωνικά δίκτυα διαχωρίζεται σε δύο κατηγορίες:

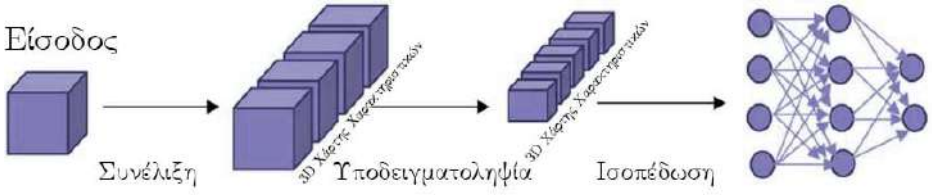
Εξαγωγή χαρακτηριστικών (Feature Extraction): Στην εξαγωγή χαρακτηριστικών, το προεκπαιδευμένο μοντέλο CNN χρησιμοποιείται ως εξαγωγή σταθερών χαρακτηριστικών και τα τελικά στρώματα του μοντέλου αντικαθίστανται με νέα επίπεδα που είναι εκπαιδευμένα για τη νέα εργασία. Δηλαδή, από το μοντέλο κρατούνται μόνο τα επίπεδα συνέλιξης και απορρίπτονται τα υπάρχων πλήρως συνδεδεμένα επίπεδα. Το προεκπαιδευμένο μοντέλο CNN συνήθως εκπαιδεύεται σε ένα μεγάλο σύνολο δεδομένων, όπως το ImageNet, και έχει μάθει να αναγνωρίζει χαρακτηριστικά χαμηλού και υψηλού επιπέδου που μπορούν να είναι χρήσιμα για άλλες εργασίες. Χρησιμοποιώντας το προεκπαιδευμένο μοντέλο ως εργαλείο εξαγωγής χαρακτηριστικών, το νέο μοντέλο μπορεί να αξιοποιήσει τις εκμαθημένες λειτουργίες και να επικεντρωθεί σε χαρακτηριστικά εκμάθησης που αφορούν συγκεκριμένες εργασίες.

Fine-Tuning: Στο "fine-tuning", το προεκπαιδευμένο μοντέλο CNN χρησιμοποιείται ως προετοιμασία για ένα νέο μοντέλο και ολόκληρο το μοντέλο εκπαιδεύεται στη νέα εργασία. Το προεκπαιδευμένο μοντέλο μένει ως έχει, πέραν του επίπεδου εξόδου στα πλήρως συνδεδεμένα επίπεδα (ή μερικά παραπάνω), το οποίο προσαρμόζεται στο πλήθος των κλάσεων του νέου προβλήματος. Το προεκπαιδευμένο μοντέλο τυπικά ρυθμίζεται με ακρίβεια χρησιμοποιώντας μικρό ρυθμό εκμάθησης για τα αρχικά επίπεδα (ή παγώνεται γενικά η εκπαίδευση για αυτά τα επίπεδα), έτσι ώστε να αποφευχθεί η υπερβολική προσαρμογή στο νέο σύνολο δεδομένων και να επηρεαστούν αρκετά οι ήδη προσαρμοσμένοι παράμετροι.

Παρόμοια προσπάθεια έχει γίνει και στον τομέα του βίντεο, με τεράστια σύνολα δεδομένων όπως το Kinetics ή το UCF-101, τα οποία έχουν χιλιάδες κατηγορίες ανθρωπίνων δράσεων. Ωστόσο, παραμένει ακόμα ανοιχτό ερώτημα το αν οι χρονικές εξαρτήσεις στα δεδομένα βίντεο μπορούν να εκμεταλλευτούν την μεταφορά μάθησης σε ένα διαφορετικό χρονικά εξαρτώμενο σύνολο δεδομένων, σε παρόμοιο βαθμό με τις χωρικές εξαρτήσεις στην μεταφορά μάθησης σε εικόνες[16].

2.5 Τρισδιάστατα Συνελικτικά Νευρωνικά Δίκτυα

Η πρόσφατη επιτυχία των Συνελικτικών Νευρωνικών δικτύων, έχει οδηγήσει την επιστημονική κοινότητα να τα υιοθετήσει ως βασικό τρόπο αναγνώρισης αντικειμένων απο εικόνες, μοτίβων από ήχο κ.α. Για την αναγνώριση ενεργειών σε βίντεο, σε αντίθεση με την αναγνώριση εικόνας, η αναγνώριση ενεργειών λειτουργεί σε καρέ βίντεο, που σημαίνει ότι πρέπει επίσης να το μοντέλο να έχει την δυνατότητα να εξάγει χρονικές πληροφορίες. Ωστόσο, η ανάπτυξη μοντέλων ικανά στην αναγνώριση χρονικών εξαρτήσεων απο τα δεδομένα δεν έχει δείξει την ίδια πρόοδο όσο εκείνη των χωρικών εξαρτήσεων.[16] Μια ιδέα για την αναγνώριση δράσεων σε βίντεο και άλλων χωρο-χρονικών εξαρτήσεων θα ήταν να χρησιμοποιήσουμε την ίδια αρχιτεκτονική με τα ΣΝΔ δύο διαστάσεων και απλά να προσθέσουμε μια διάσταση για την επιπλέον διάσταση του χρόνου (2 χωρικές διαστάσεις και 1 χωρική). Αυτή είναι η λογική πίσω απο τα Τρισδιάστατα Συνελικτικά Νευρωνικά Δίκτυα.



Σχήμα 2.9: **Παράδειγμα 3D Συνελικτικού Νευρωνικού Δικτύου**.

Τα Τρισδιάστατα Συνελικτικά Νευρωνικά Δίκτυα (3D-CNN) έχουν ακριβώς τις ίδιες αρχές με τα δισδιάστατα, αλλά χρησιμοποιούν συνελίξεις και φίλτρα τριών διαστάσεων. Στην περίπτωση βίντεο για παράδειγμα, η είσοδος είναι της μορφής $W \times H \times D$, όπου W,H είναι οι χωρικές διαστάσεις της εικόνας και η τρίτη διάσταση D είναι πολλά καρέ βίντεο τοποθετημένα. Στην συνέχεια, τρισδιάστατα φίλτρα, δηλαδή πυρήνες (kernels) συνελλίσονται με την είσοδο και εξάγουν χωρο-χρονικά μοτίβα από τα δεδομένα. Ο υπολογισμός της 3D συνέλιξης στην θέση (x, y, z) στον j "κύβο χαρακτηριστικών" στο i επίπεδο συνέλιξης έχει την παρακάτω μορφή.

$$V_{i,j}^{x,y,z} = f(\sum_{m} \sum_{h=0}^{H_i-1} \sum_{w=0}^{W_i-1} \sum_{c=0}^{C_i-1} k_{i,j,m}^{h,w,c} V_{(i-1),m}^{(x+h),(y+w),(z+c)} + b_{i,j})$$

Στην παραπάνω εξίσωση $k_{i,j,m}^{h,w,c}$ αναπαριστά την τιμή στην θέση h, w, c στο j πυρήνα συνέλιξης στο i επίπεδο στον m κύβο χαρακτηριστικών του προηγούμενου επιπέδου. Το H_i, W_i, C_i συμβολίζει το μέγεθος του πυρήνα. Το $V_{(i-1),m}^{(x+h),(y+w),(z+c)}$ την τιμή στην θέση $(x + h, y + w, z + c)$ του m κύβου χαρακτηριστικών του προηγούμενου επιπέδου, το $V_{i,j}^{x,y,z}$ την έξοδο στη θέση (x, y, z) του j χάρτη χαρακτηριστικών στο i επίπεδο, το $b_{i,j}$ την πόλωση και τέλος η $f(\cdot)$ είναι η συνάρτηση ενεργοποίησης.

Στη συνέχεια, οι χάρτες χαρακτηριστικών περνάνε απο Επίπεδα Υποδειγματοληψίας (Pooling) τα οποία τις χρονικές και χωρικές διαστάσεις κρατώντας την μέση ή μέγιστη τιμή απο περιοχές. Για παράδειγμα, ένα επίπεδο υποδειγματοληψίας με μέγεθος υποδειγματοληψίας $(2, 2, 2)$ και βήμα $(2, 2, 2)$, χωρίζει τον κύβο χαρακτηριστικών σε $2 \times 2 \times 2$ περιοχές και υπολογίζει την μέγιστη / μέση τιμή. Μετά απο την επανάληψη αυτής της διαδικασίας φίλτρων - υποδειγματοληψίας, η έξοδος τους "ισοπεδώνεται" σε μία διάσταση όπου περνάει στα πλήρως συνδεδεμένα επίπεδα για την ταξινόμηση τους.

Παίρνοντας ένα παράδειγμα ένα 2D-CNN ταξινόμησης εικόνων του ImageNet διαγωνισμού 2014 [41], η ταξινόμηση μιας εικόνας μεγέθους $224 \times 224 \times 3$ θα απαιτούσε 53 δισεκατομμύρια

υπολογισμούς. Στο πρώτο επίπεδο συνέλιξης υπάρχουν 96 φίλτρα μεγέθους 7×7 με βήμα 2. Οπότε κάθε φίλτρο εφαρμόζεται $(\frac{224}{2})^2$ φορές. Αν σκεφτούμε ότι η προσαρμογή αυτού του μοντέλου για να δέχεται 3D-είσοδο, έστω να δέχεται είσοδο σειρά εικόνων με μεγέθος δηλαδή $224 \times 224 \times 224 \times 3$. Τότε η εφαρμογή ενός φίλτρου μεγέθους $7 \times 7 \times 7$ με το ίδιο βήμα θα χρειαζόταν να εφαρμοστεί 112 περισσότερες φορες απο την 2D περίπτωση, με κάθε εφαρμογή να απαιτεί 7 φορές περισσότερους υπολογισμούς. Δηλαδή, για την επέκταση όλου αυτού του δικτύου σε 3D θα απαιτούσε 6.1 τρισσεκατομύρια υπολογισμούς για την ταξινόμηση ενός μόνο βίντεο.

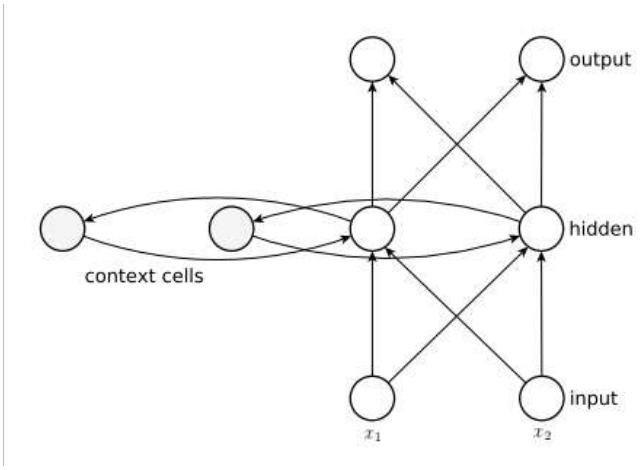
Το παραπάνω γεγονός, οδηγεί στο συμπέρασμα ότι τα 3D-CNN να φαίνονται ως κακή ιδέα. Επίσης, λόγω της επιπλέον χρονικής διάστασης στα δεδομένα εισόδου βίντεο, ο όγκος ενός απαιτούμενου συνόλου δεδομένων για την ταξινόμηση δράσεων σε βίντεο είναι τεράστιος και συνεπώς και η προεπεξεργασία των δεδομένων επίσης πολύ περισσότερο υπολογιστικά απαιτητική. Η περιπλοκότητα και απαιτητικότητα των μοντέλων είχαν παρουσιάσει εμπόδιο στο παρελθόν και η προσπάθεια είχε στραφεί σχεδόν εξ'ολοκλήρου στην αναγνώριση αντικειμένων και ταξινόμηση εικόνων με ανάπτυξη τεράστιων συνόλων δεδομένων και προσπάθειες όπως το ImageNet [42]. Ωστόσο με διάφορες τεχνικές, υπάρχουν τρόποι να μειώσουμε την υπολογιστική απαιτητικότητα των 3D-CNN, όπως την εφαρμογή των συνελίξεων σε χώρο Fourier [43], την χρήση διαχωρίσιμων φίλτρων [44] , την χρήση sparse 3D-CNN[45] κ.α., καθώς και άλλους συνδυασμούς αρχιτεκτονικών που θα δούμε σε παρακάτω κεφάλαια.

2.6 Αναδρομικά Νευρωνικά Δίκτυα

Τα Αναδρομικά Νευρωνικά Δίκτυα [46] (Recurrent Neural Network - RNN) είναι γενικός όρος που αναφέρεται σε οποιοδήποτε δίκτυο του οποίου οι νευρώνες στέλνουν αναδρομικά σήματα μεταξύ τους, δηλαδή οι ενώσεις μεταξύ των νευρώνων δημιουργούν κύκλο, επιτρέποντας την έξοδο ενός νευρώνα να επηρεάσει την μεταγενέστερη είσοδο του ίδιου νευρώνα. Έτσι, οι βρόχοι ανάδρασής που υπάρχουν στο δίκτυο δημιουργούν μια μορφή μνήμης μέσα στο δίκτυο, η οποία επιτρέπει στο δίκτυο να διατηρεί πληροφορίες απο προηγούμενα βήματα της ακολουθίας και να χρησιμοποιεί αυτές τις πληροφορίες για να ενημερώσει τις προβλέψεις του για το τρέχον βήμα. Αυτό τους επιτρέπει να έχουν δυνατότητα για δυναμική συμπεριφορά και πρόκειται για μοντέλα τα οποία έχουν δημιουργηθεί για να χειρίζονται σειριακά δεδομένα. Παραδείγματα εφαρμογών τους είναι η αναγνώριση φωνής [47], η αναγνώριση χειρόγραφων κειμένων [48] κ.α.

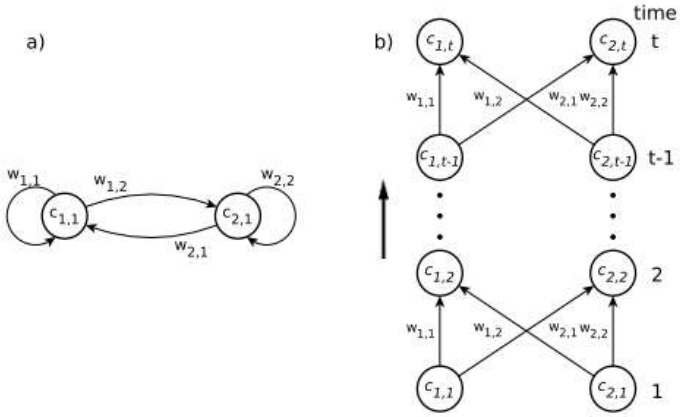
Τα Αναδρομικά Νευρωνικά Δίκτυα κυμαίνονται μεταξύ διάφορων αρχιτεκτονικών από μερικώς έως πλήρως συνδεδεμένες, με μία απλή αρχιτεκτονική να είναι το δίκτυο Elman [49]. Το δίκτυο Elman μοιάζει με ένα νευρωνικό δίκτυο τριών επιπέδων, αλλα επιπροσθέτως οι έξοδοι του κρυφού επιπέδου αποθηκεύονται στις λεγόμενες "μονάδες συμφραζόμενων" (context cells). Οι έξοδοι των μονάδων συμφραζόμενων κυκλικά τροφοδοτούνται στους νευρώνες του κρυφού επιπέδου μαζί με το αρχικό σήμα. Κάθε νευρώνας έχει την δική του μονάδα συμφραζόμενων και δέχεται είσοδο και απο το επίπεδο εισόδου αλλά και απο τις μονάδες συμφραζόμενων. Τα δίκτυα Elman μπορούν να εκπαιδευτούν με τυπική οπίσθια διάδοση του σφάλματος, με την έξοδο των μονάδων συμφραζόμενων να θεωρούνται απλά ως επιπλέον είσοδο. Παρακάτω βλέπουμε ένα παράδειγμα ενός δικτύου Elman.

Τα Αναδρομικά Νευρωνικά δίκτυα εκπαιδεύονται διαφορετικά απο τα κλασσικά Νευρωνικά Δίκτυα Πρόσθιας Διάδοσης, με τους πιο διαδεδομένους αλγόριθμους για την εκπαίδευση RNN σε προβλήματα επιβλεπόμενης μάθησης να είναι η χρονική οπίσθια διάδοση (backpropagation through time / BPTT) και η αναδρομική μάθηση πραγματικού χρόνου (real-time recurrent learning / RTRL). Ο αλγόριθμος BPTT [50] λειτουργεί με την ιδέα ότι για κάθε χρονική στιγμή, υπάρχει ΝΔ πρόσθιας διάδοσης με ακριβώς ίδια συμπεριφορά με κάθε RNN. Για την εύρεση του αντίστοιχου ΝΔΠΔ πρέπει να ξεδιπλώσουμε χρονικά το αντίστοιχο αναδρομικό ΝΔ. Δηλαδή



Σχήμα 2.10: **Παράδειγμα Δικτύου Elman**

στο τέλος εκπαιδευτικής ακολουθίας, το δίκτυο ξεδιπλώνεται χρονικά (σαν να ήταν ΝΔΠΔ) και η εκπαίδευση γίνεται με τον τυπικό αλγόριθμο οπίσθιας διάδοσης. Στον αλγόριθμο πραγματικού χρόνου οι κλίσεις υπολογίζονται όσο η είσοδος παρουσιάζεται στον δίκτυο, χωρίς να χρειάζεται οπίσθια διάδοση του σφάλματος. Ωστόσο, ο αλγόριθμος έχει σημαντικό υπολογιστικό κόστος ανά κύκλο ενημερώσεων [50].



Σχήμα 2.11: **Παράδειγμα Ξεδίπλωσης Αναδρομικού Νευρωνικού Δικτύου** Στην αριστερή εικόνα (α) βλέπουμε ένας πλήρες αναδρομικό ΝΔ. Το ίδιο ΝΔ στην δεξιά εικόνα (β) ξεδιπλώνεται χρονικά με ένα ξεχωριστό επίπεδο για κάθε χρονικό βήμα και μετατρέπεται σε ένα ΝΔΠΔ.

Παρ'όλο το γεγονός ότι αποτελούν μια πολύ καλή μέθοδο για την εξαγωγή χρονικών εξαρτήσεων εμφανίζονται διάφορα εμπόδια με την εκπαιδευσή τους στην κανονική τους έκδοση. Ενώ ο αλγόριθμος οπίσθιας διάδοσης μετατρέπεται εύκολα για την εφαρμογή στα Αναδρομικά Νευρωνικά Δίκτυα, παρουσιάζονται προβλήματα όπως η δυσκολία αποφυγής παγίδευσης σε τοπικά ελάχιστα [46]. Ωστόσο, το μεγαλύτερο πρόβλημα αποτελεί το πρόβλημα των εξαφανιζόμενων κλίσεων το οποίο, ενώ παρουσιάζεται και στα μη Αναδρομικά ΝΔ, παρουσιάζει μεγαλύτερο εμπόδιο στην προκειμένη περίπτωση. Ο λόγος βρίσκεται στο γεγονός ότι οι κλίσεις πρέπει επιπροσθέτως να διαδοθούν προς τα πίσω και στον χρόνο, και σε κάθε πολλαπλασιασμό των κλίσεων κατά την οπίσθια διάδοση γίνονται ολοένα μικρότερες ή μεγαλύτερες οδηγώντας στο πρόβλημα των εξαφανιζόμενων ή των "εκραγόμενων" κλίσεων αντίστοιχα. Ως αποτέλεσμα, τα

παραδοσιακά Αναδρομικά Νευρωνικά Δίκτυα δεν μπορούν να μάθουν αποδοτικά "μακροχρόνιες" χρονικές εξαρτήσεις που διαχωρίζονται απο αρκετά χρονικά βήματα (περίπου 10 χρονικά βήματα [51, 52]), επειδή οι κλίσεις γίνονται πολύ μικρές για να ενημερώσουν αποτελεσματικά τα βάρη στο δίκτυο. Η εξήγηση της αιτίας φαίνεται παρακάτω. Η ανανέωση των βαρών από την χρονική στιγμή t' στη χρονική στιγμή t βρίσκεται με την εξίσωση :

$$\Delta W_{[u,v]} = -\eta \frac{\partial E_{\text{total}}(t', t)}{\partial W_{[u,v]}},$$

όπου :

$$\frac{\partial E_{\text{total}}(t', t)}{\partial W_{[u,v]}} = \sum_{\tau=t'}^{t} \theta_u(\tau) X_{[u,v]}(\tau),$$

όπου το οπισθιοδιαδομένο σφάλμα την χρονική στιγμή τ (με $t' <= \tau <= t$) της μονάδας u είναι:

$$\theta_u(\tau) = f'_u(z_u(\tau)) \left(\sum_{u \in U} W_{uv} \theta_u(\tau + 1) \right) \tag{2.19}$$

Συνεπώς, αν έχουμε ένα πλήρες αναδρομικό ΝΔ με ένα σύνολο μονάδων (όχι εισόδου) U, τότε το σήμα σφάλματος που συμβαίνει σε οποιοδήποτε νευρώνα στο επίπεδο εξόδου $o \in O$, στο χρονικό βήμα τ, διαδίδεται οπίσθια για $t - t'$ χρονικά βήματα, με $t' < t$ για ένα αυθαίρετο νευρώνα u. Αυτό οδηγεί το σφάλμα στο να κλιμακωθεί με τον παρακάτω παράγοντα:

$$\frac{\partial \theta_u(t')}{\partial \theta_o(t)} = \begin{cases} f'_u(z_u(t')) W_{[o,u]} & \text{αν } t - t' = 1 \\ f'_u(z_u(t')) W_{[o,u]} (\sum_{u \in U} \frac{\partial \theta_u(t'+1)}{\partial \theta_o(t)} W_{[u,v]}) & \text{αν } t - t' > 1 \end{cases} \tag{2.20}$$

Για την επίλυση της παραπάνω εξίσωσης, πρέπει να την ξεδιπλώσουμε χρονικά. Για $t' <= \tau <= t$, έστω u_τ να είναι ένας νευρώνας, όχι απο το επίπεδο εισόδου, σε ένα από τα ξεδιπλωμένα δίκτυα την χρονική στιγμή τ. Θέτωντας $u_t = v, u_{t'} = o$, έχουμε την εξίσωση:

$$\frac{\partial \theta_u(t')}{\partial \theta_o(t)} = \sum_{u_{t'} \in U} ... \sum_{u_{t-1} \in U} \left(\prod_{\tau=t'+1}^{t} f'_{u_\tau}(z'_{u_\tau}(t - \tau - t')) W_{[u_\tau, u_{\tau-1}]} \right) \tag{2.21}$$

Παρατηρούμε ότι αν :

$$|f'_{u_\tau}(z'_{u_\tau}(t - \tau - t')) W_{[u_\tau, u_{\tau-1}]}| > 1 \tag{2.22}$$

για όλα τα τ, τότε το γινόμενό τους θα αυξηθεί εκθετικά, κάνοντας το σφάλμα να "εκραγεί". Επίσης, σφάλματα αντίθετα στον νευρώνα v οδηγεί σε βάρη που ταλαντεύονται και ασταθής εκπαίδευση. Αν απο την άλλη έχουμε :

$$|f'_{u_\tau}(z'_{u_\tau}(t - \tau - t')) W_{[u_\tau, u_{\tau-1}]}| < 1 \tag{2.23}$$

για όλα τα τ, τότε το γινόμενό τους μειώνεται εκθετικά, κάνοντας το σφάλμα να εξαφανιστεί και αποτρέποντας το δίκτυο απο το να εκπαιδευτεί σε οποιοδήποτε επιθυμητό χρονικό διάστημα.

Έχουν προταθεί πολλές λύσεις για το παραπάνω πρόβλημα όπως η χρήση πυλών (Gated Recurrent Unit) ή και δικτύων Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory / LSTM) που θα μελετήσουμε παρακάτω.

2.6.1 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (LSTM)

Ένας τρόπος αντιμετώπισης του προβλήματος εξαφανιζόμενων / εκραγόμενων κλίσεων είναι τα Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory / LSTM). Τα LSTM μπορούν να εντοπίσουν χρονικές εξαρτήσεις από δεδομένα που διαχωρίζονται απο 1000 χρονικά βήματα [50], σε αντίθεση με τα τυπικά RNN που μπορούν 5-10 [53]. Τα LSTM πετυχαίνουν το παραπάνω με την χρήση "καρουσέλ συνεχούς σφάλματος" (Constant Error Carousels / CEC). Απο την εξίσωση 2.19 βλέπουμε ότι το τοπικό σφάλμα στην μονάδα u (απο την ίδια) την χρονική στιγμή τ δίνεται απο την εξίσωση :

$$\theta_u(\tau) = f'_u(z_u(\tau))W_{[u,u]}\theta_u(\tau+1)$$

Απο τις εξισώσεις 2.22 και 2.23 βρίσκουμε ότι προκειμένου να έχουμε σταθερή ροή σφάλματος στην μονάδα u πρέπει να ισχύει :

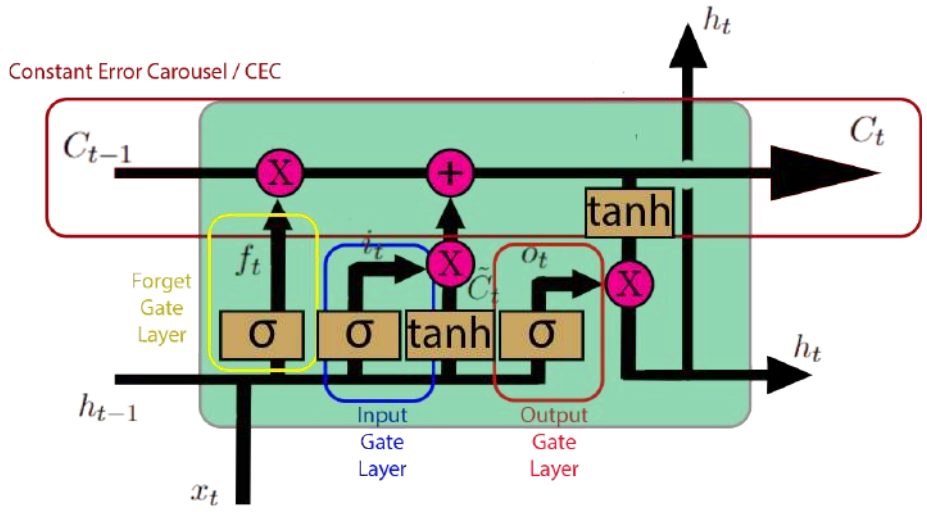
$$f'_u(z_u(\tau))W_{[u,u]} = 1.0$$

και συνεπώς ολοκληρώνοντας την παραπάνω εξίσωση έχουμε :

$$f_u(z_u(\tau)) = \frac{z_u(\tau)}{W_{[u,u]}}$$

Που σημαίνει ότι η f_u πρέπει να είναι γραμμική και η ενεργοποίηση της u πρέπει να παραμένει χρονικά σταθερή. Δηλαδή:

$$y_u(\tau+1) = f_u(z_u(\tau+1)) = f_u(y_u(\tau)W_{[u,u]}) = y_u(\tau)$$



Σχήμα 2.12: **Παράδειγμα Δικτύου LSTM**

Το παραπάνω διασφαλίζεται με την χρήση της ιδιοσυνάρτησης $f_u = id$ και θέτωντας $W_{[u,u]} = 1.0$. Αυτή η διατήρηση του σφάλματος ονομάζεται "καρουσέλ συνεχούς σφάλματος" (Constant Error Carousel / CEC) και είναι ο βασικός τρόπος που τα LSTM αντιμετωπίζουν το πρόβλημα εξαφανιζόμενων / εκραγόμενων κλίσεων και μπορούν συγκρατήσουν βραχυπρόθεσμη μνήμη για μεγάλο χρονικό διάστημα [50].

Το καρουσέλ συνεχούς σφάλματος, παρόλου του πολύπλοκου ονόματός του, είναι ουσιαστικά ένας απλός τρόπος τα LSTM να μπορούν να συγκρατούν την κατάσταση της μονάδας με την δυνατότητα να μην επηρεαστεί καθόλου. Ωστόσο, η δυνατότητα να προσθέτουν ή να αφαιρέσουν πληροφορίες στην εσωτερική κατάσταση τους επιτυγχάνεται με την χρήση πυλών.

Η δυνατότητα να αφαιρέσουν μη σχετικές πληροφορίες γίνεται μέσω ενός επιπέδου πύλης που ονομάζεται "Forget Gate Layer". Το επίπεδο Forget Gate Layer παίρνει ως είσοδο την προηγούμενη έξοδο h_{t-1} και την νέα είσοδο x_t και μέσω μια σιγμοειδής συνάρτησης ενεργοποίησης βγάζει ένα αποτέλεσμα f_t με τιμή 0 έως 1 για κάθε τιμή που περιέχεται στην κατάσταση της μονάδας, συμβολίζοντας το ποσοστό της πληροφορίας που θέλουμε να συγκρατήσουμε.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2.24}$$

Η ενημέρωση της εσωτερικής κατάστασης προσθέτοντας νέες τιμές γίνεται σε δύο μέρη. Το πρώτο μέρος ονομάζεται "Input Gate Layer" και βγάζει μέσω μιας σιγμοειδούς συνάρτησης έξοδο ένα διάνυσμα i_t, με τιμές απο 0 έως 1, το οποίο αποφασίζει ποιές τιμές θα ενημερώσουμε. Έπειτα ένα επίπεδο υπερβολικής εφαπτομένης παράγει ως έξοδο ένα διάνυσμα νέων υποψήφιων τιμών $\tilde{C}_t$ που θα μπορούσαν προστεθούν στην εσωτερική κατάσταση.

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{2.25}$$

$$\tilde{C}_t = tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{2.26}$$

Άρα η νέα τιμή του επόμενου χρονικού βήματος της εσωτερικής κατάστασης δίνεται απο την εξίσωση:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.27}$$

Η έξοδος της μονάδας h_t υπολογίζεται περνώντας έναν γραμμικό συνδυασμό της προηγούμενης έξοδου h_{t-1} και της εισόδου x_t μέσω απο μια σιγμοειδή βγάζοντας έξοδο ένα διάνυσμα ενεργοποίησης εξόδου o_t. Έπειτα το διάνυσμα o_t πολλαπλασιάζεται (για να εξάγουμε επιλεκτικά τμήματα της εσωτερικής κατάστασης) στοιχειωδώς με την τωρινή εσωτερική κατάσταση $\tilde{C}_t$ η οποία έχει περάσει απο μια συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης (για να ωθήσουμε τις τιμές μεταξύ -1 και 1).

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{2.28}$$

$$h_t = o_t * tanh(\tilde{C}_t) \tag{2.29}$$

Υπάρχουν πολλές παραλλαγές της κλασσικής αρχιτεκτονικής ενός LSTM, όπως Depth-Gated RNN [54] με την πιο διάσημη παραλλαγή [55] όμως να χρησιμοποιεί ενώσεις "peephole" όπου οι πύλες (Input, Forget, Output) έχουν πρόσβαση στην εσωτερική κατάσταση της μονάδας. Ωστόσο, σύμφωνα με Greff, et al. (2015) [56], η κλασσική έκδοση του LSTM έχει ικανοποιητική απόδοση σε διάφορα προβλήματα και δεν υπάρχουν σημαντικές διαφορές μεταξύ τους. Μια πιο δραματική παραλλαγή που γίνεται ολοένα και πιο διάσημη είναι το Gated Recurrent Unit (GRU) [57] όπου απλοποιεί την αρχιτεκτονική του LSTM συνδυάζοντας τα επίπεδα πυλών "Input" και "Forget" σε μία μόνο πύλη που ονομάζεται "Update", μαζί με τον συνδυασμό της εσωτερικής κατάστασης και της εξόδου σε μία.

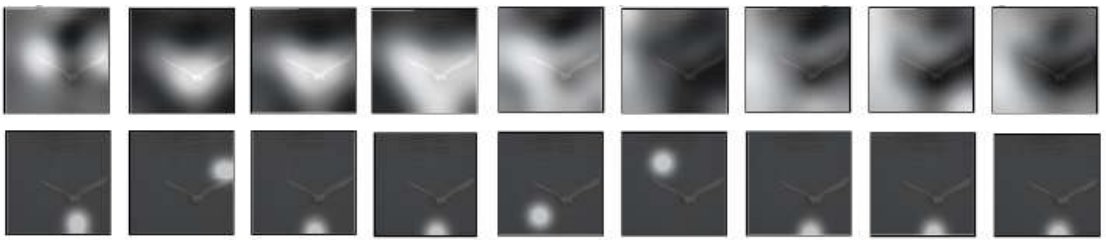
Χρησιμοποιούνται πολλοί μέθοδοι για την εκπαίδευση των LSTM όπως "Backpropagation Through Time" και "Real Time Recurrent Learning", ωστόσο πολλοί αλγόριθμοι όπως ο "Rprop" [58] και άλλοι έχουν δείξει βελτιωμένη επίδοση σε σχέση με τις κλασσικές μεθόδους εκπαίδευσης.

2.7 Attention

Η έννοια του Attention (Προσοχή) είναι ένας μηχανισμός με τον οποίο μπορεί να βελτιωθεί η επίδοση των μοντέλων εστιάζοντας μόνο σε συγκεκριμένα κομμάτια της εισόδου. Οι μηχανισμοί προσοχής επιτρέπουν στα μοντέλα να παρακολουθούν επιλεκτικά διαφορετικά μέρη της εισόδου, καθιστώντας τα πιο αποτελεσματικά στην αναγνώριση μακροχρόνιων εξαρτήσεων και βελτιώνοντας τη συνολική ποιότητα της εξόδου.

Η ιδέα της προσοχής στα τεχνητά νευρωνικά δίκτυα έχει άμεσα εμπνευστεί απο την ανθρώπινη ικανότητα να εστιάσει επιλεκτικά σε συγκεκριμένα ερεθίσματα απ'το περιβάλλον φιλτράροντας τους περισπασμούς και κατανέμοντας αποτελεσματικά τους νοητικούς πόρους. Εισήχθη πρώτα από Bahdanau, Dzmitry et. al [59] ως μια νέα προσέγγιση στα νευρωνικά δίκτυα μετάφρασης κειμένου, όπου επέτρεπε στο μοντέλο να εστιάζει επιλεκτικά σε διαφορετικά μέρη του κειμένου κατά την διαδικασία της μετάφρασης βελτιώνοντας δραστικά την επίδοση. Έκτοτε έχει αποτελέσει μια βασική έννοια στην επεξεργασία φυσικής γλώσσας δίνοντας έμπνευση για καινοτόμες νέες αρχιτεκτονικές όπως τον Μετατροπέα / Transformer [60], ωστόσο και σε άλλους τομείς όπως στην αναγνώριση ενεργειών ως τρόπος "έμφασης" συγκεκριμένων περιοχών στην εικόνα [61, 62].

Η προσοχή διαχωρίζεται σε δύο διαφορετικές έννοιες στοχαστικής / ντετερμινιστικής προσοχής Hard-Attention / Soft-Attention, ωστόσο υπάρχουν πάρα πολλές διαφορετικές υλοποιήσεις των δύο. [63] Κατά κύριο λόγο, η υλοποίηση της προσοχής σε οποιοδήποτε μοντέλο περιλαμβάνει την δημιουργία ενός διανύσματος συνάφειας $\hat{z}_t$ το οποίο αποτελεί μια δυναμική αναπαράσταση του σχετικού τμήματος της εισόδου την χρονική στιγμή t. Ο μηχανισμός φ υπολογίζει το $\hat{z}_t$ λαμβάνοντας ως είσοδο ένα σύνολο διανυσμάτων $a_i, i = 1, ..., L$ που αντιστοιχεί σε διαφορετικά χαρακτηριστικά που έχουν εξαχθεί σε διαφορετικές τοποθεσίες της εικόνας. Για κάθε θέση i ο μηχανισμός παράγει ένα θετικό βάρος α_i που μπορεί να ερμηνευθεί ως η πιθανότητα ότι η θέση i είναι η σωστή τοποθεσία να εστιάσει το μοντέλο (στοχαστική προσοχή) ή ως την σχετική σημασία της περιοχής i για την ένωση των a_i μαζί (ντετερμινιστική προσοχή) [62]. Ο υπολογισμός των α_i για κάθε διάνυσμα a_i υλοποιείται από ένα μοντέλο προσοχής f_{att} με διαφορετικές υλοποιήσεις ανάλογα το πρόβλημα.



Σχήμα 2.13: Παράδειγμα **Χαλαρής Προσοχής** (Πάνω) και **Αυστηρής Προσοχής** (Κάτω) κατά την εύρεση σχετικών περιοχών για την παραγωγή λεζάντας με είσοδο εικόνα. (Xu 2015) [62]

Η "Αυστηρή Προσοχή" (Hard-Attention) πρόκειται για έναν μηχανισμό προσοχής που επιτρέπει στα μοντέλα να εστιάσουν σε συγκεκριμένο τμημά της εισόδου. Αναπαριστώντας την μεταβλητή θέσης s_t ως την θέση την οποία πρέπει να εστιάσει ο μηχανισμός προσοχής προκειμένου να παράγει την t έξοδο, τότε το $s_{t,i}$ είναι ένας δείκτης που έχει τιμή ίση με 1 στην θέση στην οποία εστιάζει ο μηχανισμός για την εξαγωγή των χαρακτηριστικών. Αντιμετωπίζοντας τις θέσεις προσοχής ως ενδιάμεσες λανθάνουσες μεταβλητές, τότε μπορεί να οριστεί μία κατανομή παραμετροποιημένη απο το α_i και να δούμε το $\hat{z}_t$ ως τυχαία μεταβλητή (εξού και ο

λόγος που είναι στοχαστική μέθοδος).

$$p(s_{t,i} = 1 | s_{j<t}, \mathbf{a}) = \alpha_{t,i}$$

$$\hat{z}_t = \sum_i s_{t,i} \mathbf{a}_i$$

Η "Χαλαρή Προσοχή" (Soft-Attention) πρόκειται για έναν μηχανισμό προσοχής που επιτρέπει στα μοντέλα να εστιάσουν σε διαφορετικά τμήματα της εισόδου ξεχωριστά. Αντί για την δειγματοληψία της θέσης της προσοχής s_t κάθε χρονική στιγμή, η χαλαρή προσοχή μπορεί να λάβει το διάνυσμα συνάφειας $\hat{z}_t$ απευθείας και να ορίσει ένα ντετερμινιστικό μοντέλο προσοχής υπολογίζοντας ένα διάνυσμα χαλαρής προσοχής $\varphi(\mathbf{a}_i, \alpha_i) = \sum_i^L \alpha_i a_i$. Όλο το μοντέλο είναι συνεχές και συνεχώς διαφορίσιμο, οπότε η εκπαίδευσή του είναι εύκολη με χρήση κλασσικών αλγορίθμων οπίσθιας διάδοσης.

Κεφάλαιο 3

Αναγνώριση Ανθρωπίνων Δράσεων σε Βίντεο

3.1 Σύνολο Δεδομένων

Ένα "καλό" σύνολο δεδομένων είναι μέγιστης σημασίας στην βαθιά μάθηση και η επίδοση ενός μοντέλου συνδέεται άμεσα με το μέγεθος και την ποιότητα των δεδομένων στα οποία εκπαιδεύεται. Το σύνολο δεδομένων πρέπει να είναι ποικίλο και αντιπροσωπευτικό του τομέα του προβλήματος, καθώς αυτό επιτρέπει στο μοντέλο να μάθει να αναγνωρίζει διαφορετικές παραλλαγές της ίδιας έννοιας ή αντικειμένου. Τα δεδομένα πρέπει ουσιαστικά να μπορούν να παρέχουν την δυνατότητα στο μοντέλο να *γενικεύσει* και να αποφύγει την υπερβολική εκμάθηση (overfitting).

Αφενώς, το πλήθος των δεδομένων πρέπει να είναι επαρκώς μεγάλο, συγκριτικά με το πλήθος παραμέτρων και την πολυπλοκότητα του προβλήματος, καθώς τα δεδομένα είναι πηγή όλων των απαραίτητων ερεθισμάτων προκειμένου το μοντέλο να ενημερώσει τα βάρη του. Όσο πιο πολύπλοκο το πρόβλημα που στοχεύεται να μοντελοποιηθεί, τόσο πιο μεγάλο το δίκτυο απαιτείται ώστε να είναι ικανό να εντοπίζει πολύπλοκα μοτίβα και συνεπώς τόσο μεγαλύτερο πλήθος δεδομένων απαιτείται. Αρκετοί εμπειρικοί κανόνες σχετίζουν το απαραίτητο μέγεθος του σύνολου δεδομένων με το πλήθος των παραμέτρων, με έναν διάσημο κανόνα να είναι ότι χρειάζονται μια τάξης μεγέθους παραπάνω δείγματα όσο παράμετροι προκειμένου να εκπαιδευτεί επαρκώς ένα δίκτυο. Δηλαδή, μικρότερα μοντέλα με πολλά δεδομένα ενδεχομένως να έχουν καλύτερη επίδοση απο μεγάλα μοντέλα με λιγότερα δεδομένα. Ωστόσο, στην πράξη υπάρχουν πολύ πιο πολύπλοκες σχέσεις μεταξύ δεδομένων και παραμέτρων. Για παράδειγμα, η αύξηση του πλήθους των δεδομένων σε μερικές περιπτώσεις μπορεί παραδόξως αρχικά να μειώσει σημαντικά την επίδοση του μοντέλου πριν την βελτιώσει ξανά (Deep Double Descent) [64].

Αφετέρου, η πληθικότητα των δεδομένων δεν έχει σημασία αν τα δεδομένα δεν είναι *ποιοτικά*. Η εκτίμηση της ποιότητας των δεδομένων, είναι εμπειρική προσέγγιση και αφορά την δυνατότητα των δεδομένων να εκπαιδεύσουν επαρκώς ένα μοντέλο, που είναι γνωστό ότι είναι ικανό, πάνω σε ένα συγκεκριμένο πρόβλημα. Αρχικά, ένα ποιοτικό σύνολο δεδομένων πρέπει να αναπαριστά ικανοποιητικά τα χαρακτηριστικά τα οποία καλείται το μοντέλο να εξάγει. Πρέπει να διασφαλίζει αξιοπιστία στο ότι τα δεδομένα έχουν κατάλληλες ετικέτες και ότι δεν υπάρχουν δείγματα που δεν περιέχουν τα χαρακτηριστικά που αναζητούνται. Υπερβολικά θορυβώδη δείγματα επίσης, μετά απο μία αναλογία σήματος προς θόρυβο (Signal-Noise Ratio / SNR), δυσκολεύουν την συνάρτηση κατάβασης κλίσης να βρεί ολικό ελάχιστο αλλά και μειώνουν την ταχύτητα σύγκλισης.

3.1.1 Datasets

Movie Fights Dataset

Το Movie Fights Dataset [65] πρόκειται για ένα σύνολο δεδομένων βίαων σκηνών κυρίως απο ταινίες δράσης. Τα βίντεο είναι χωρισμένα με ετικέτες ίσα μεταξύ Μη-Βίαων / Βίαων σκηνών. Είναι σύνολο 200 βίντεο (100 σε κάθε κλάση) μεγέθους 720×480 ή 720×576 διάρκειας 2 δευτερολέπτων και καρέ ανά λεπτό 25/29 FPS (Frames Per Second). Συνολικά δηλαδή το σύνολο δεδομένων περιέχει περίπου 10,000 καρέ βίντεο.

Όπως φαίνεται και απο τις εικόνες παρακάτω, καθώς τα βίντεο είναι παρμένα από ταινίες, δεν αναπαριστούν ρεαλιστικές συνθήκες βίας ή επιθετικής συμπεριφοράς. Τα βίντεο είναι επίσης τραβηγμένα σε τέλειες συνθήκες φωτισμού με την κάμερα να βρίσκεται κοντά στα σημεία ενδιαφέροντος στις 90 μοίρες (σε αντίθεση με τις κάμερες ασφαλείας που συνήθως βρίσκονται ψηλά τραβώντας απο γωνία).

Συνεπώς, δεν είναι ικανό dataset για το πρόβλημα αναγνώρισης *πραγματικών* σκηνών βίας απο κάμερες ασφαλείας και η εκπαίδευση ενός νευρωνικού δικτύου ικανού να ταξινομεί βίντεο από το συγκεκριμένο dataset γίνεται τετριμμένη με την χρήση έτοιμων μοντέλων [8].

(a) (b) (c)

Σχήμα 3.1: **Καρέ βίντεο από Movie Fights Dataset**

Hockey Fights Dataset

Το Hockey Fights Dataset πρόκειται για ένα σύνολο δεδομένων βίαων σκηνών παρμένων απο βίαιων συμβάντων σε αγώνες hockey. Τα βίντεο είναι χωρισμένα με ετικέτες ίσα μεταξύ Μη-Βίαων / Βίαων σκηνών. Είναι σύνολο 1000 βίντεο (500 σε κάθε κλάση) μεγέθους 360×288 διάρκειας 1 δευτερολέπτου και καρέ ανά λεπτό 25 FPS (Frames Per Second). Συνολικά δηλαδή το σύνολο δεδομένων περιέχει 25,000 καρέ βίντεο.

Παρ'όλο που τα βίντεο είναι παρμένα απο αγώνες hockey και αναπαριστούν πραγματικές συνθήκες βίας και επιθετικής συμπεριφοράς, είναι παρ'όλα αυτά περιορισμένα σε αγώνες hockey και μόνο. Όπως φαίνεται και απο τις εικόνες παρακάτω, καθώς τα βίντεο είναι παρμένα από τηλεοπτικούς αγώνες hockey, έχουν γυριστεί σε τέλειες συνθήκες φωτισμού με τον χειριστή της κάμερας να ακολουθεί συνεχώς τα σημεία ενδιαφέροντος (σε αντίθεση με τις κάμερες ασφαλείας που συνήθως βρίσκονται ψηλά στατικά).

Συνεπώς, όπως και το Movie Fights Dataset, δεν είναι ικανό dataset για το πρόβλημα αναγνώρισης *πραγματικών* σκηνών βίας απο κάμερες ασφαλείας καθώς είναι αδύνατη η εύρεση κάποιας γενικευμένης λύσης στο πρόβλημα της βίας καθώς οι σκηνές είναι περιορισμένες σε αγώνες hockey μόνο. Η εκπαίδευση ενός νευρωνικού δικτύου ικανού να ταξινομεί βίντεο από το συγκεκριμένο dataset, όπως και στο Movie Fights Dataset γίνεται τετριμμένη με την χρήση των ίδιων έτοιμων μοντέλων[8].

(a) (b) (c)

Σχήμα 3.2: **Καρέ βίντεο από Hockey Fights Dataset**

UCF-Crime

Το UCF-Crime [66] πρόκειται για τη παραλλαγή του γνωστού UCF-101 dataset (που περιέχει 101 κατηγορίες αναγνώρισης ανθρωπίνων δράσεων) σε 13 κατηγορίες εγκληματικών δράσεων Κατάχρηση, Σύλληψη, Εμπρησμός, Επίθεση, Τροχαία ατυχήματα, Διάρρηξη, Έκρηξη, Τσακωμός, Ληστεία, Πυροβολισμοί, Κλοπές, Κλοπές καταστημάτων και Βανδαλισμοί. Πρόκειται για 1900 μη-περικομμένα βίντεο πραγματικού κόσμου από κάμερες ασφαλείας συνολικού χρόνου 128 ωρών. Όλες οι εικόνες είναι μεγέθους 64×64 και έχει σύνολο 1,377,653 καρέ βίντεο χωρισμένα μεταξύ train / test.

Το UCF-Crime πρόκειται για ένα άκρως ρεαλιστικό και επαρκώς μεγάλο σύνολο δεδομένων ικανό για την αναγνώριση σκηνών βίας και άλλων παρόμοιων κατηγοριών εκτάκτων αναγκών απο κάμερες ασφαλείας. Ωστόσο, το γεγονός ότι τα βίντεο δεν είναι περικομμένα, καθώς και η ύπαρξη τόσων πολλών κατηγοριών, κάνει το σύνολο δεδομένων πιο κατάλληλο για την χρήση σε προβλήματα αναγνώρισης ανωμαλιών και όχι σε επιβλεπόμενης μάθησης.

(a) (b) (c)

Σχήμα 3.3: **Καρέ βίντεο από 3 κατηγορίες του UCF-Crime dataset.** Η εικόνα (α) αναπαριστά βανδαλισμό, η (β) τροχαίο ατύχημα και η (γ) επίθεση

RWF-2000

Το RWF-2000 [67] πρόκειται για ένα σύνολο δεδομένων βίαιων συμβάντων απο κάμερες ασφαλείας από βίντεο στο YouTube. Περιέχει 2000 βίντεο χωρισμένα μεταξύ Βίαιων / Μη-Βίαων συμβάντων διάρκειας 5 δευτερολέπτων με καρέ ανα δευτερόλεπτο 30 FPS (Frames Per Second). Οι εικόνες είναι ακανόνιστου μεγέθους και περιέχει σύνολο 300,000 καρέ βίντεο.

Το RWF-2000 πρόκειται για ένα πολύ ρεαλιστικό dataset σκηνών βίας όπου σε μερικές περιπτώσεις προκαλούν ανησυχία. Καθώς τα βίντεο έχουν καταγραφεί απο κάμερες ασφαλείας σε δημόσιους χώρους περιέχει σκηνές με πληθώρα συνθηκών με διαφορετικές γωνίες, διαφορετικό φωτισμό, γρήγορη κίνηση αντικειμένων και ατόμων, ύπαρξη πλήθους κ.α. Είναι λοιπόν,

ικανό σύνολο δεδομένων στο πλαίσιο της παρούσας εργασίας και θα χρησιμοποιηθεί ως σημείο αναφοράς. ’

(a) (b) (c)

Σχήμα 3.4: **Καρέ βίντεο από RWF-2000 dataset.** Στην εικόνα (α) βλέπουμε βίαο συμβάν απο μακρινή απόσταση, στην (β) πλήθος και πανικό και στην (γ) χαμηλής ανάλυσης εικόνα.

3.1.2 Χαρακτηριστικά βίαιων συμβάντων

Η βία είναι ένα σύνθετο και πολύπλευρο φαινόμενο, με πολλές μορφές εκδήλωσης όπως κοινωνική, πολιτική, ψυχολογική, λεκτική και σωματική. Παρ’όλο που υπάρχει ένας βαθμός επικάλυψης μεταξύ των διαφορετικών μορφών εκδηλώσεων της βίας, στην παρούσα εργασία κέντρο ενδιαφέροντος παρουσιάζει μόνο η σωματική βία. Ωστόσο, η βία πρόκειται για μια αφηρημένη έννοια και η αναγνώρισή της απαιτεί αναγνώριση χαρακτηριστικών και κατανόηση έννοιων με διαφορετικά επίπεδα ασάφειας. Στην προσπάθεια εύρεσης ενός αφηρημένου ορισμού των διάφορων χαρακτηριστικών / εννοιών βίαιων συμβάντων, διαχωρίζουμε σε τρία διαφορετικά μέρη τις έννοιες / χαρακτηριστικά. Σε εκείνα που αφορούν την πρόκληση βίας, εκείνα που αφορούν την αποδοχή της και της επιπτώσεις της στο γύρω περιβάλλον.

(a) (b) (c)

Σχήμα 3.5: **Παραδείγματα χαρακτηριστικών αποδοχής βίας.** Στην εικόνα (α) το κεφάλι πάει πίσω, στην (β) το άτομο βρίσκεται λιπόθυμο και στην (γ) άτομο να πέφτει

Η αποδοχή της βίας δεν είναι σαφώς ορισμένη σαν έννοια αλλά ταυτόχρονα είναι και η πιο επαρκής για την αναγνώριση βίαιων συμβάντων. Το κοινό στοιχείο όλων των συμβάντων βίας είναι ο τραυματισμός των εμπλεκόμενων. Ένας ανθρώπινος παρατηρητής έχοντας έντονο το συναίσθημα της ενσυναίσθησης μπορεί κατευθείαν να αναγνωρίσει ποιές κινήσεις προκαλούν σωματικό πόνο και βλάβη, συγκρίνοντας με τις δικές του προσωπικές εμπειρίες. Γίνεται άμεσα κατανοητό με την παρακολούθηση των βίντεο σε πραγματικές βίαιες σκηνές ότι το μεγαλύτερο

κομμάτι της οπτικής προσοχής πηγαίνει προς τον δεχόμενο της βίας. Ωστόσο, οποιοδήποτε μοντέλο το οποίο αναγνωρίζει βίαια συμβάντα σε ικανοποιητικό βαθμό πρέπει να μάθει να "γνωρίζει" τι καθιστά φυσιολογική ανθρώπινη κατάσταση και πότε κάποιος άνθρωπος δέχεται σωματική βλάβη / πόνο. Παραδείγματα χαρακτηριστικών που υποδεικνύουν σωματική βλάβη του ατόμου μπορεί να είναι η πτώση του ατόμου στο έδαφος ή η έντονη κίνηση της κεφαλής προς τα πίσω (π.χ. μετά απο κάποιο χτύπημα).

(a) Αντικείμενο / Όπλο (b) Απότομη Κίνηση Χεριών (c) Πάλη

(d) Απότομη Κίνηση Ποδιών (e) Λεκτική Διαμάχη (f) Επιθετική Συμπεριφορά

Σχήμα 3.6: **Παραδείγματα χαρακτηριστικών πρόκλησης βίας**

Η πρόκληση της βίας από την άλλη έχει μερικά πιο σαφώς ορισμένα χαρακτηριστικά που ενδεχομένως να είναι ευκολότερα να εντοπιστούν. Αυτά μπορεί να είναι απότομες κινήσεις των χεριών κατευθυνόμενες προς άλλο άτομο (π.χ. γροθιές), απότομες κινήσεις των ποδιών κατευθυνόμενες προς άλλο άτομο (π.χ. λακτίσματα), χρήση κάποιου αντικειμένου / όπλου ή η απότομη κίνηση των σωμάτων μεταξύ δύο ατόμων που πιάνονται μεταξύ τους πέφτοντας προς το έδαφος (π.χ. πάλη) κ.α. Ωστόσο υπάρχουν και άλλα λιγότερο σαφή χαρακτηριστικά όπως επιθετική συμπεριφορά, πρόθεση για χτύπημα κ.α. Η πρόθεση για τραυματισμό του αντιπάλου και το επιθετικό συναίσθημα είναι αρκετά δύσκολες να εντοπιστούν χωρίς την κατανόηση πιο αφηρημένων εννοιών, ωστόσο είναι κατα μεγάλο ποσοστό ικανές για την αναγνώριση σχεδόν όλων των βίαιων συμβάντων.

Τέλος, χαρακτηριστικά που υποδεικνύουν σκηνή βίας τα οποία είναι εμφανή στο γύρω περιβάλλον της μπορεί να είναι υλικές ζημιές στο γύρω περιβάλλον, δημόσια αναταραχή των ατόμων που είναι παρόντα, στραμμένα βλέμματα προς το σημείο ενδιαφέροντος κ.α. Οι πληροφορίες που δεχόμαστε απο το περιβάλλον μπορεί είτε να ενισχύσουν την πεποίθηση ότι ένα σημείο ενδιαφέροντος καθιστά βία ή να παρέχει ενδείξεις προς του τι συμβαίνει εκτός οπτικού πεδίου της εικόνας. Ωστόσο, είναι σαφές ότι αυτά είναι πιο δύσκολο να εντοπιστούν και ταυτόχρονα περιέχουν μικρότερο ποσοστό πληροφορίας επί του προβλήματος.

(a) Πανικός στο πλήθος

(b) Άτομα χωρίζουν εμπλεκόμενους

(c) Στραμμένη προσοχή

Σχήμα 3.7: **Παραδείγματα χαρακτηριστικών στο περιβάλλον.**

3.2 Ανάλυση Βιβλιογραφίας

3.2.1 Transfer Learning

Έχουν υπάρξει διάφορες προσπάθειες για την χρήση μεταφοράς μάθησης στην αναγνώριση σκηνών βίας όπως στο άρθρο "Violence Detection in Surveillance Videos with Deep Network Using Transfer Learning" [8] όπου οι συγγραφείς ερευνούν την χρήση μεθόδων βαθιάς μάθησης για την αναγνώριση σκηνών βίας σε αντίθεση με κλασσικές μεθόδους που χρησιμοποιούν χειροποίητα χαρακτηριστικά. Οι συγγραφείς χρησιμοποιούν ένα έτοιμο μοντέλο 2D-CNN GoogleNet, βασισμένο στην αρχιτεκτονική Inception, προεκπαιδευμένο σε 1000 κατηγορίες εικόνων στο σύνολο δεδομένων ImageNet και στη συνέχεια "fine-tuned" σε σύνολα δεδομένων σκηνών βίας. Δείχνουν ότι μπορούν να πετύχουν ικανοποιητικότερα αποτελέσματα σε σχέση με τεχνικές χειροποίητων χαρακτηριστικών, ωστόσο το μοντέλο αξιολογείται σε μη ρεαλιστικά σύνολα δεδομένων όπως το Hockey και Movies Dataset που έχουν προαναφερθεί. Επίσης, καθώς πρόκειται για μοντέλο 2D-CNN δεν έχει την δυνατότητα να εξάγει χρονικές εξαρτήσεις απο τα δεδομένα παρα μόνο χωρικές και συνεπώς είναι λογικό ότι δεν καθιστά ικανή αντιμετώπιση για την αναγνώριση βίας σε πραγματικές συνθήκες.

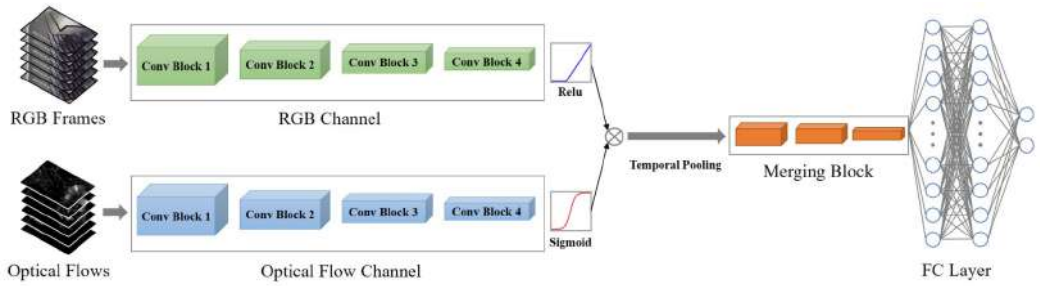
Η παραπάνω υλοποίηση ωστόσο, δίνει ενδείξεις για το πόσο ικανό εργαλείο είναι η μεταφορά μάθησης στην εξαγωγή χωρικών πληροφοριών. Ωστόσο, το ίδιο δεν ισχύει για την εξαγωγή χρονικών πληροφοριών και είναι ανοικτό ερώτημα στον τομέα του βίντεο αν γίνεται η προεκπαίδευση το μοντέλου με ένα τεράστιο σύνολο δεδομένων βίντεο αναγνώρισης ενεργειών σε εκατοντάδες κατηγορίες να προσφέρει παρόμοια βελτίωση της επίδοσης όπως στον τομέα της εικόνας [16]. Η προεκπαίδευση σε σύνολα δεδομένων βίντεο (π.χ. : Kinetics) ωστόσο προσφέρουν σημαντική βελτίωση της επίδοσης, ωστόσο εξαρτάται κυριώς απο την αρχιτεκτονική του μοντέλου που χρησιμοποιείται [16].

3.2.2 Flow Gated Network

Στο επιστημονικό άρθρο "RWF-2000: An Open Large Scale Video Database for Violence Detection" [10] οι συγγραφείς εισάγουν την δημιουργία του συνόλου δεδομένων βίαιων βίντεο "RWF-2000" (που προαναφέρθηκε) το οποίο το χρησιμοποιούν για την αξιολόγηση όλων των τελευταίας τεχνολογίας μοντέλων αναγνώρισης σκηνών βίας. Στην συνέχεια, παρουσιάζουν ένα νέο μοντέλο για την αναγνώριση σκηνών βίας ονόματι Flow Gated Network.

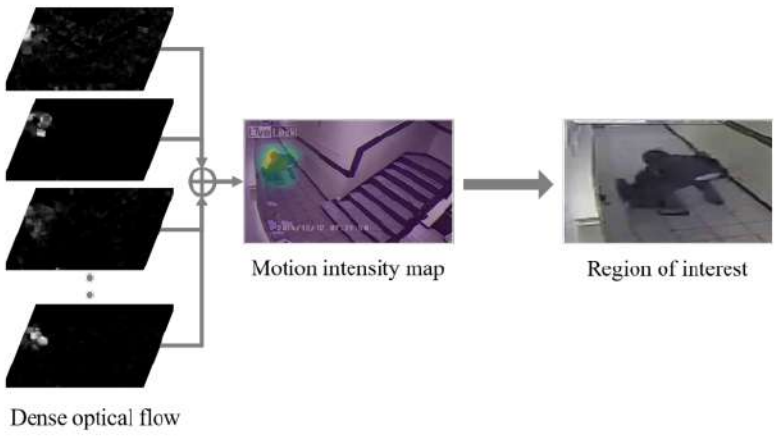
Στο Σχήμα 4.10 βλέπουμε την βασική δομή του Flow Gated Network που αποτελείται απο 4 μέρη : το κανάλι RGB, το κανάλι Οπτικών Ροών, το Μπλοκ Συγχώνευσης και τα

Πλήρως Συνδεδεμένα Επίπεδα. Το κανάλι RGB και το κανάλι Οπτικών Ροών αποτελούνται απο σειριακά 3D Συνελικτικά Νευρωνικά Δίκτυα με συνεπής δομή έτσι ώστε οι έξοδοί τους να μπορούν να συγχωνευτούν. Το Μπλοκ Συγχώνευσης αποτελείται επίσης απο 3D ΣΝΔ, τα οποία επεξεργάζονται την πληροφορία μετά από χρονική υποδειγματοληψία. Τέλος, τα πλήρως συνδεδεμένα επίπεδα χρησιμοποιούνται για την παραγωγή εξόδου.



Σχήμα 3.8: **Αρχιτεκτονική Flow-Gated Network.** (Cheng 2021 [10])

Η βασική ιδέα πίσω από το συγκεκριμένο μοντέλο είναι η χρήση του καναλιού Οπτικών Ροών σε συνδυασμό με Max Pooling προκειμένου το μοντέλο να μπορεί να δώσει μεγαλύτερο βάρος στα σημεία έντονου ενδιαφέροντος. Η συνάρτηση ενεργοποίησης ReLU χρησιμοποιείται στο τέλος του καναλιού RGB, ενώ στο κανάλι Οπτικών Ροών χρησιμοποιείται η Σιγμοειδής. Έπειτα τα αποτελέσματα πολλαπλασιάζονται και περνάνε από ένα επίπεδο Max Pooling. Καθώς η Σιγμοειδής δίνει μία έξοδο μεταξύ 0 και 1, μπορεί να ερμηνευτεί ο ρόλος της ως ένας συντελεστής κλιμάκωσης του RGB καναλιού. Αφού στη συνέχεια χρησιμοποιείται Max Pooling και συνεπώς κρατούνται μόνο τα τοπικά μέγιστα, το κομμάτι του RGB καναλιού που έχει πολλαπλασιαστεί με 1 είναι πιο πιθανό να παραμείνει, ενώ το κομμάτι που έχει πολλαπλασιαστεί με 0 είναι πιο πιθανό να κοπεί. Αυτός ο μηχανισμός είναι ένας τρόπος το μοντέλο να χρησιμοποιεί το κανάλι Οπτικών Ροών για να κρατήσει τα σημεία ενδιαφέροντος (έντονης κίνησης) και να μην εκλάβει υπ'όψιν τα υπόλοιπα. [10]



Σχήμα 3.9: **Στρατηγική περικοπής εικόνας χρησιμοποιώντας Οπτική Ροή.** (Cheng 2021 [10])

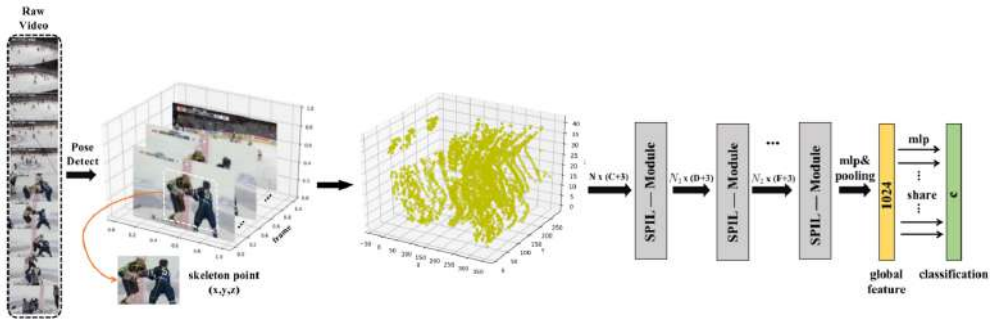
Ωστόσο, η παραπάνω υλοποίηση εμφανίζει και αυτή μερικά προβλήματα. Αρχικά, η χρήση οπτικών ροών ως έναν τρόπο περιορισμού του παρασκηνίου και περικοπής της περιοχής ενδιαφέροντος προσθέτει μεγάλο υπολογιστικό κόστος στο μοντέλο. Υπάρχουν πολλοί αλγόριθμοι

υπολογισμού οπτικής ροής στην εικόνα, με τη συγκεκριμένη υλοποίηση να χρησιμοποιεί τον αλγόριθμο του Gunner Farneback [68] για τον υπολογισμό της σε γειτονικά καρέ. Ο υπολογισμός οπτικών ροών προσθέτει βαθμό υπολογιστικής πολυπλοκότητας στο μοντέλο ανάλογο της υπολογιστικής πολυπλοκότητας του υπόλοιπου δικτύου. Πέρα από επιπλέον υπολογιστικό κόστος, παρέχεται προφανώς και μια χρονική καθυστέρηση καθώς απαιτείται προεπεξεργασία της εισόδου πριν την είσοδό της στο δίκτυο, το οποίο καθιστά την συγκεκριμένη υλοποίηση λιγότερο ικανή για την χρήση της σε εφαρμογές πραγματικού κόσμου. Για την επιτάχυνση του υπολογισμού της απαιτείται είτε ειδικό υλικό (hardware) ή χρήση άλλων μεθόδων εκτίμησης της οπτικής ροής.

Επιπλέον, η χρήση τρισδιάστατων συνελίξεων, για την εξαγωγή χωρο-χρονικών εξαρτήσεων, είναι υπολογιστικά ακριβή επίσης. Υπάρχουν διάφοροι μέθοδοι για την επιτάχυνση και εκτίμηση του υπολογισμού τους όπως η χρήση χωριζόμενων συνελίξεων κατά βάθος (Depth-Wise Separable Convolutions) από την αρχιτεκτονική του MobileNet [69] ή Ψευδό-τρισδιάστατων Υπολειπόμενων Δικτύων (Pseudo-3D Residual Network) [70]. Παρ'όλο που οι συγγραφείς αναφέρουν ότι μπορεί να χρησιμοποιηθεί ένας συνδυασμός των παραπάνω δύο μεθόδων για την επιτάχυνση του υπολογισμού των συνελίξεων, δεν αναφέρουν λεπτομερώς τον τρόπο.

3.2.3 SPIL Convolution

Στο άρθρο ονόματι "Human Interaction Learning on 3D Skeleton Point Clouds for Video Violence Recognition" [71] οι Su et al. παρουσιάζουν μια μέθοδο αναγνώρισης σκηνών βίας μαθαίνοντας τις σχέσεις μεταξύ των ατόμων από σημεία του σκελετού. Στην μέθοδο αυτή πρώτα υπολογίζονται τρισδιάστατα σημειακά νέφοι (3D Point Clouds) που εξάγονται απο ανθρώπινους σκελετούς σε βίντεο και στην συνέχεια γίνεται μάθηση στη σχέση μεταξύ αυτών των σημείων. Η παραπάνω μέθοδος παρουσιάζει μια νέα μονάδα ονόματι Skeleton Points Interaction Module (SPIL) για την μοντελοποίηση της αλληλεπίδρασης μεταξύ των σημείων των σκελετών. Το SPIL χρησιμοποιώντας μια στρατηγική κατανομής βαρών μεταξύ τοπικών συγγενικών σημείων, έχει στόχο να εστιάσει επιλεκτικά στα πιο σχετικά τμήματά τους με βάση τα χαρακτηριστικά τους και πληροφορίες της χωρο-χρονικής τους θέσης. Για την κωδικοποίηση διαφορετικών ειδών σχεσιακών πληροφοριών, χρησιμοποιείται ένας μηχανισμός πολυμετωπικής προσοχής (multi-head attention) για την συγκέντρωση διαφορετικών χαρακτηριστικών απο ξεχωριστά μέτωπα για την διαχείριση διαφορετικών ειδών σχέσεων μεταξύ των σημείων.



Σχήμα 3.10: **Επισκόπηση Μοντέλου SPIL Convolution.** Το μοντέλο χρησιμοποιεί τη μέθοδο ανίχνευσης στάσης για να εξάγει συντεταγμένες σκελετού από κάθε καρέ του βίντεο. Αυτά τα σημεία ανθρωπίνων σκελετών παρέχονται στην μορφή σημειακών νέφων ως είσοδοι στις μονάδες SPIL που εκτελούν διάδοση πληροφοριών αναθέτοντας διαφορετικά βάρη σε διαφορετικά σημεία του σκελετού. Τελικά, εξάγεται ένα καθολικό χαρακτηριστικό για την εκτέλεση ταξινόμησης. (Su 2020 [71])

Υπολογισμός 3D Point Clouds

Το πρόβλημα το οποίο προσπαθεί να αντιμετωπίσει η παραπάνω αντιμετώπιση είναι ότι κλασσικές μέθοδοι αναγνώρισης εστιάζουν σε χαρακτηριστικά τα οποία εξάγονται απο ολόκληρη την σκηνή, οδηγώντας σε παρεμβολή μεταξύ των σχετικών πληροφοριών και άσχετων πληροφοριών στο παρασκήνιο. Καθώς η βία πρόκειται για ένα καθαρά ανθρώπινο φαινόμενο, οδηγείται η προσπάθεια στην εξαγωγή ανθρωπίνων κινήσεων και αλληλεπιδράσεων, αντί για μεθόδους όπως π.χ. : οπτική ροή, μέσω αναγνώρισης ανθρώπινων σκελετών. Οι κινήσεις και η στάση των ανθρώπων στα βίντεο αντικατοπτρίζονται σε αλληλουχίες ανθρώπινων σκελετικών σημείων, τα οποία μετατρέπονται σε τρισδιάστατα σημειακά νέφοι, που κατ'ουσίαν πρόκειται απλά για ένα μαθηματικό σύνολο ή μια αταξινόμητη συλλογή σημείων. Ο αλγόριθμος που χρησιμοποείται για τον εντοπισμό των σκελετικών σημείων [72] απλοποιημένος είναι:

- **Αναγνώριση Ανθρώπων**

  Το πρώτο βήμα του αλγορίθμου είναι να ανιχνεύσει όλα τα άτομα που υπάρχουν στην εικόνα. Αυτό γίνεται χρησιμοποιώντας έναν έτοιμο ανιχνευτή ανθρώπων, όπως το Faster R-CNN [73] ή το YOLO [74]. Η έξοδος αυτού του βήματος είναι ένα σύνολο πλαισίων οριοθέτησης που περικλείουν κάθε άτομο στην εικόνα.

- **Εκτίμηση Στάσης**

  Στη συνέχεια, ο αλγόριθμος υπολογίζει τη στάση κάθε ατόμου στην εικόνα. Αυτό γίνεται χρησιμοποιώντας ένα συνελικτικό νευρωνικό δίκτυο (CNN) που λαμβάνει ως είσοδο την εικόνα και τα πλαίσια οριοθέτησης από το βήμα 1. Το δίκτυο εξάγει ένα σύνολο σημείων κλειδιών, τα οποία αντιστοιχούν σε διαφορετικά μέρη του ανθρώπινου σώματος (π.χ. ώμους, αγκώνες, γόνατα κ.λπ.). Το δίκτυο εκπαιδεύεται χρησιμοποιώντας ένα μεγάλο σύνολο δεδομένων ετικετών εικόνων με αντίστοιχα σημεία-κλειδιά.

- **Βελτίωση Στάσης**

  Η έξοδος από το βήμα 2 μπορεί να είναι θορυβώδης και ανακριβής, ειδικά όταν πολλά άτομα βρίσκονται κοντά το ένα στο άλλο στην εικόνα. Επομένως, ο αλγόριθμος εκτελεί ένα βήμα βελτίωσης για να βελτιώσει την ακρίβεια των εκτιμώμενων στάσεων. Αυτό το βήμα χρησιμοποιεί μια τεχνική που ονομάζεται επαναληπτική ανάδραση σφαλμάτων, όπου τα εκτιμώμενα σημεία-κλειδιά βελτιώνονται σε πολλαπλές επαναλήψεις. Σε κάθε επανάληψη, τα εκτιμώμενα σημεία-κλειδιά συγκρίνονται με τα βασικά σημεία-κλειδιά και το σφάλμα χρησιμοποιείται για την προσαρμογή των εκτιμώμενων σημείων κλειδιών. Αυτή η διαδικασία συνεχίζεται έως ότου τα εκτιμώμενα σημεία-κλειδιά συγκλίνουν σε μια σταθερή λύση.

- **Συσχέτιση Στάσεων**

  Το τελευταίο βήμα του αλγορίθμου είναι να συσχετίσει τις εκτιμώμενες στάσεις με το σωστό άτομο στην εικόνα. Αυτό γίνεται συγκρίνοντας τα εκτιμώμενα βασικά σημεία-κλειδιά με τα οριοθετημένα πλαίσια από το βήμα 1 και χρησιμοποιώντας έναν μέγιστο διμερή αλγόριθμο αντιστοίχισης για την αντιστοίχηση κάθε στάσης στο σωστό άτομο.

Με την επανάληψη της ίδιας διαδικασίας για κάθε καρέ χρονικά λαμβάνουμε ως έξοδο ένα Τρισδιάστατο Σημειακό Νέφος, με συντεταγμένες (x, y, z) όπου x, y αναπαριστούν τις χωρικές συντεταγμένες και z το z σε σειρά καρέ. Το θέμα είναι ωστόσο ότι καθώς ο παραπάνω αλγόριθμος αφορά διακριτά σημεία συντεταγμένων, δε γίνεται να χρησιμοποιηθούν κλασσικές συνελικτικές μέθοδοι εντοπισμού χαρακτηριστικών.

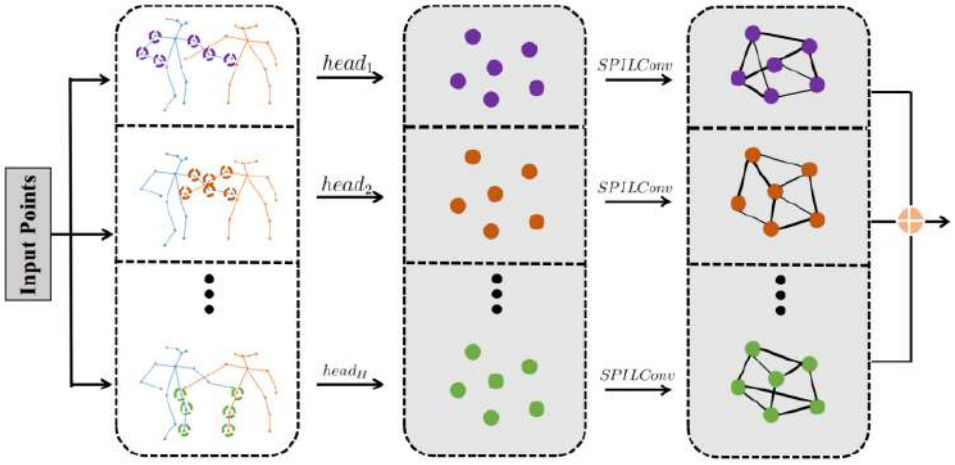
SPIL / Εξαγωγή Χαρακτηριστικών από 3D Point Clouds

Για τον υπολογισμό των βαρών αλληλεπίδρασης μεταξύ των σημείων, η κλασσική αντιμετώπιση ήταν η χρήση του αλγορίθμου K-γειτόνων. Ωστόσο δεν έχουν επίδραση όλα τα κοντινά σημεία στο τωρινό σημείο και ο υπολογισμός της αλληλεπίδρασης άσχετων σκελετικών σημείων οδηγεί σε συγκεχυμένα αποτελέσματα. Για την αντιμετώπιση αυτού του προβλήματος, στη SPIL, όπως φαίνεται στο Σχήμα 3.11 διαισθητικά, τα βάρη αλληλεπίδρασης μεταξύ των σημείων σε διαφορετικά σκελετικά σημεία κατανέμονται με βάση τις σχέσεις μεταξύ των σημείων. Η μονάδα μαθαίνει να καλύπτει ή να αποδυναμώνει μέρος των βαρών συνέλιξης ανάλογα με τα χαρακτηριστικά γνωρίσματα των γειτόνων.

Έστω ένα σύνολο σημείων $p_1, p_2, ..., p_k \in \rho^3$ με βάση τους K-γείτονες ενός κέντρου βάρους, όπου κάθε τοπική περιοχή ομαδοποιείται εντός μίας ακτίνας. Θέτωντας την ακτίνα σε $(r \times T_{frame})$ εγγυάται ότι η τοπική περιοχή θα καλύψει μεταξύ και εντός των καρέ στον χώρο αλλά και στον χρόνο. Για τον υπολογισμό των βαρών W_{ij}, όπου W_{ij} η ένωση μεταξύ του σημείου i και του σημείου j, λαμβάνονται υπόψη και τα δύο χαρακτηριστικά ομοιότητας και σχέσεις χαρακτηριστικών θέσης. Για το σκοπό αυτό, διερευνούνται ξεχωριστά οι πληροφορίες χαρακτηριστικών και θέσης και στη συνέχεια εκτελούνται μοντελοποιήσεις υψηλού επιπέδου για την προσαρμογή δυναμικά στη δομή των αντικειμένων. Ο υπολογισμός των βαρών μεταξύ κάθε γειτονικού σημείου υπολογίζεται ως :

$$W_{ij} = \Phi(R^F(p_i^f, p_j^f), R^L(p_i^l, p_j^l)) \tag{3.1}$$

Όπου $R^F(p_i^f, p_j^f)$ αναπαριστά την σχέση χαρακτηριστικών μεταξύ των σημείων, $R^L(p_i^l, p_j^l)$ αναπαριστά την σχέση θέσης. Η συνάρτηση Φ παίζει τον ρόλο του συνδυασμού πληροφοριών χαρακτηριστικών και θέσης. Ο υπολογισμός των $R^F(p_i^f, p_j^f)$, $R^L(p_i^l, p_j^l)$ αναλύονται στο άρθρο με τον υπολογισμό των πληροφοριών θέσης $R^L(p_i^l, p_j^l)$ να χρησιμοποιεί τρεις διαφορετικές μεθόδους βασισμένους στην Ευκλείδια απόσταση με διαφορετικούς βαθμούς επιτυχίας.



Σχήμα 3.11: **Απεικόνηση πολυ-μετωπικής μονάδας SPIL.** Ένα μόνο μέτωπο κωδικοποιεί το σημειακά σκελετικά νέφοι από την είσοδο ανεξάρτητα, και πολλά ανεξάρτητα μέτωπα είναι υπεύθυνα για την επεξεργασία διαφορετικώνς τύπων πληροφοριών από τα σημεία και στο τέλος να τα συγκεντρώσουν μαζί. Κάθε κόμβος υποδηλώνει ένα σημείο σκελετικής άρθρωσης και κάθε άκρη είναι ένα βαθμωτό βάρος, που υπολογίζεται σύμφωνα με τα χαρακτηριστικά δύο σημείων και τη σχετική θέση τους (Su 2020 [71]).

Πολυμετωπική προσοχή

Η κάθε μονάδα SPIL είναι ικανή να εξάγει χαρακτηριστικά αλληλεπίδρασης μεταξύ των σκελετικών σημείων, ωστόσο υπάρχουν παρα πολλές διαφορετικές εννοιολογικά αλληλεπιδράσεις μεταξύ των σημείων.Συγκεκριμένα, όπως φαίνεται στο Σχήμα 3.11, για ένα ορισμένο σκελετικό σημείο όπως η άρθρωση του αγκώνα, το πρώτο μέτωπο προσοχής μπορεί να είναι ευαίσθητο στις πληροφορίες της στάσης της ίδιας της ανθρώπινης άρθρωσης του αγκώνα, π.χ. : για να κρίνουμε αν πρόκειται για στάση «γροθιάς», ενώ το δεύτερο μέτωπο προσοχής ασχολείται περισσότερο με τη σύνδεση των πληροφοριών κίνησης της άρθρωσης του αγκώνα μεταξύ των ανθρώπων για εξαγωγή δυναμικών χαρακτηριστικών. Με τον ίδιο τρόπο τα υπόλοιπα μέτωπα εξάγουν διαφορετικά χαρακτηριστικά ανάλογα με τη συσχέτιση που μπορεί να έχουν.

Έτσι, ο πολυμετωπικός μηχανισμός επιτρέπει τη μονάδα SPIL να εργαστεί παράλληλα για να συλλάβει διάφορους τύπους σχέσεων σημείων. Για κάθε μέτωπο υπολογίζεται το βάρος W_i, όπου $i \in H$ το πλήθος μετώπων, και τα ανεξάρτητα μέτωπα δεν μοιράζονται βάρη κατά τον υπολογισμό. Με αυτόν τον τρόπο το μοντέλο επιχειρεί να συλλάβει πιο ισχυρό σχεσιακό συλλογισμό μεταξύ των σημείων.

Τέλος, για τον συνδυασμό των βαρών μεταξύ πολλών μετώπων για την τελική έξοδο στα πλήρως συνδεμένα επίπεδα ταξινόμησης, κάθε μέτωπο περνάει απο προσαρμοσμένα επίπεδα συνέλιξης και έπειτα ενώνονται μεταξύ τους όλα τα μέτωπα για την τροφοδότηση στα επίπεδα ταξινόμησης.

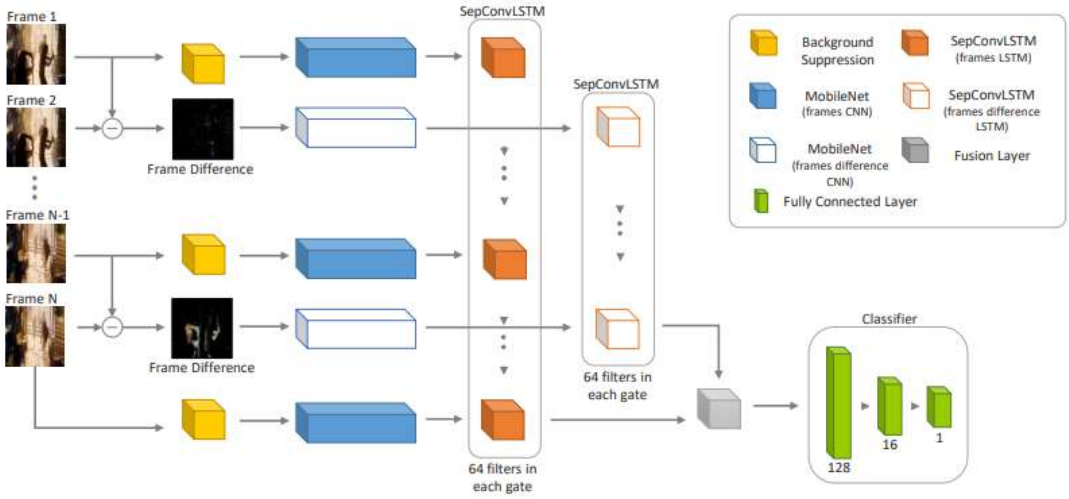
Συνολικά, πρόκειται για μια επιτυχημένη προσπάθεια στην χρήση αναγνώρισης σκελετών στην αναγνώριση ενεργειών, ή πιο συγκεκριμένα στην αναγνώριση βίαιων ενεργειών, που πετύχαινε τελευταίας τεχνολογίας επίδοση την στιγμή της έκδοσής του. Απο άποψης απόδοσης δεν είναι παρείχε βέλτιστη λύση ωστόσο και επίσης έχουν εκδοθεί πιο πετυχημένες μέθοδοι που χρησιμοποιούν αναγνώριση σκελετών έκτοτε [75].

3.2.4 Separable Convolutional LSTM

Στο επιστημονικό άρθρο ονόματι "Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM" [9] οι συγγραφείς παρουσιάζουν μια αρχιτεκτονική για την αναγνώριση σκηνών βίας που είχε χρησιμοποιηθεί μόνο σε προβλήματα τμηματοποίησης βίντεο [76] στο παρελθόν που ονομάζεται Separable Convolutional LSTM. Σε συνδυασμό με άλλες τεχνικές καταφέρνουν να βελτιώσουν την απόδοση του δικτύου σημαντικά καταφέρνοντας ταυτόχρονα τελευταίας τεχνολογίας απόδοση (την στιγμή έκδοσης του άρθρου).

Στην παραπάνω εικόνα βλέπουμε την δομή του δικτύου. Η προτεινόμενη αρχιτεκτονική αποτελείται απο δύο διαφορετικές ροές με παρόμοια αρχιτεκτονική. Κάθε ροή περιέχει ένα Δισδιάστατο Συνελικτικό Νευρωνικό Δίκτυο το οποίο εξάγει τα χωρικά χαρακτηριστικά απο κάθε χρονικό βήμα του βίντεο. Ένα επίπεδο LSTM μαθαίνει να κωδικοποιεί τα παραγόμενα χωρικά χαρακτηριστικά για να δημιουργήσει *χωρό-χρονικούς* χάρτες χαρακτηριστικών οι οποίοι στη συνέχεια προχωράνε στα πλήρως συνδεδεμένα επίπεδα ταξινόμησης. Οι δύο σημαντικές ιδέες πίσω απο την συγκεκριμένη υλοποίηση που καταφέρνουν την μείωση των υπολογιστικών απαιτήσεων του μοντέλου είναι η χρήση Depthwise Separable Convolutions αντί για κανονικές συνελίξεις και διαφορών μεταξύ των συνεχόμενων καρέ του βίντεο ως προσέγγιση της οπτικής ροής, με σημαντική μείωση στο υπολογιστικό κόστος.

Στην πρώτη ροή, τα καρέ του βίντεο περνάνε από προεπεξεργασία για την απόκρυψη πληροφοριών του παρασκηνίου και τροφοδοτούνται σειριακά στο μοντέλο. Μετά απο όταν όλα τα καρέ έχουν περάσει, εξάγονται τα χωρο-χρονικά χαρακτηριστικά απο την κρυφή εσωτερική κατάσταση του τελευταίου χρονικού βήματος του LSTM. Η ίδια διαδικασία ακολουθείται για την δεύτερη ροή, χρησιμοποιώντας ωστόσο τις διαφορές μεταξύ συνεχόμενων καρέ, ως προσέγγιση της οπτικής ροής. Η δεύτερη ροή, που χρησιμοποιεί τις διαφορές μεταξύ των συνεχόμενων καρέ, μαθαίνει να εξάγει τις χρονικές εξαρτήσεις των δεδομένων κρατώντας τις κινήσεις που προκύπτουν



Σχήμα 3.12: **Αρχιτεκτονική Μοντέλου.** (Islam 2021 [9])

απο τις διαφορές των ενδιάμεσων καρέ, ενώ η άλλη ροή εξάγει όλες τις χωρικές πληροφορίες. Τα χωροχρονικά χαρακτηριστικά που παράγονται από τις δύο ροές είναι ικανά να ξεχωρίσουν μεταξύ βίαιων και μη βίντεο.

Depth-Wise Separable Convolutions

Οι χωριζόμενες συνελίξεις [77] είναι μια από τις βασικές τεχνικές για βελτίωση της επίδοσης πολλών αρχιτεκτονικών. Η βασική ιδέα είναι η αντικατάσταση της πλήρης συνέλιξης με μια παραγοντοποιημένη έκδοση που χωρίζει την συνέλιξη σε δύο ξεχωριστά επίπεδα. Το πρώτο επίπεδο ονομάζεται συνέλιξη κατά βάθος (Depth-Wise Convolution), όπου περνάει ένα συνελικτικό φίλτρο ανά κανάλι. Το δεύτερο επίπεδο είναι μια 1×1 συνέλιξη, που ονομάζεται *pointwise* συνέλιξη, και ευθύνεται για το χτίσιμο νέων χαρακτηριστικών υπολογίζοντας γραμμικούς συνδυασμούς των καναλιών εισόδου.

Η κλασσική συνέλιξη λαμβάνει έναν τένσορα εισόδου L_i μεγέθους $h_i \times w_i \times d_i$ και εφαρμόζει ένα φίλτρο συνέλιξης $K \in R^{k \times k \times d_i \times d_j}$ για την παραγωγή ενός τένσορα εξόδου L_j μεγέθους $h_i \times w_i \times d_j$. Οι κλασσικές συνελίξεις έχουν ένα υπολογιστικό κόστος του $h_i \cdot w_i \cdot d_i \cdot d_j \cdot k \cdot k$.

Απ'την άλλη η συνέλιξη κατά βάθος πρόκειται για μια αντικατάσταση των κλασσικών επιπέδων συνελίξεων. Εμπειρικά έχουν το ακριβώς ίδιο αποτέλεσμα με τις κλασσικές συνελίξεις ωστόσο με κόστος:

$$h_i \cdot w_i \cdot d_i(k^2 + d_j)$$

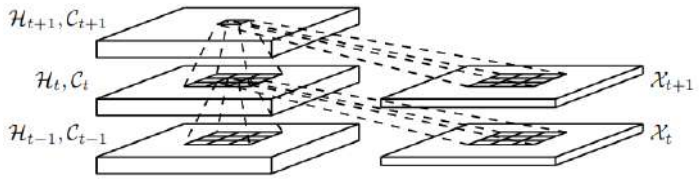
Άρα η συνέλιξη κατά βάθος προσφέρει μια βελτιώση της υπολογιστικής πολυπλοκότητας κατά παράγοντα:

$$\frac{1}{N} + \frac{1}{K^2}$$

όπου N είναι το πλήθος καναλιών εισόδου και K το μεγέθος του πυρήνα της συνέλιξης.

ConvLSTM

Τα κλασσικά LSTM με πλήρως συνδεδεμένα επίπεδα ενώ είναι αρκετά ικανά στην διαχείριση και ταξινόμηση σειριακών δεδομένων, όπως κείμενο ή χρονοσειρές, δεν είναι ικανά να διαχειριστούν χωρικές πληροφορίες. Οι ενώσεις μεταξύ εισόδου-κατάστασης και κατάστασης-κατάστασης δεν περιέχουν χωρικές πληροφορίες [78]. Για την αντιμετώπιση αυτού του προβλήματος, τα



Σχήμα 3.13: **Εσωτερική Δομή ConvLSTM.** (Shi 2015 [78])

Συνελικτικά LSTM (Convolutional LSTM / ConvLSTM) διαφέρουν στο ότι όλοι οι είσοδοι $X_1, ..., X_t$, εξόδοι μονάδων $C_1, ..., C_t$, εσωτερικές καταστάσεις $H_1, ..., H_t$ και πύλες i_t, f_t, o_t είναι τρισδιάστατοι τένσορες που οι τελευταίες διαστάσεις είναι χωρικές διαστάσεις (γραμμές και στήλες) [78]. Για διευκόλυνση μπορούμε να φανταστούμε τις εισόδους και εσωτερικές καταστάσεις ως διανύσματα πάνω σε ένα χωρικό πλέγμα. Το ConvLSTM καθορίζει τη μελλοντική κατάσταση ενός συγκεκριμένου κελιού στο πλέγμα από τις εισόδους και τις προηγούμενες καταστάσεις των τοπικών γειτόνων του [78]. Αυτό μπορεί εύκολα να επιτευχθεί χρησιμοποιώντας έναν τελεστή συνέλιξης στις μεταβάσεις κατάστασης-κατάστασης και εισόδου-κατάστασης (Σχήμα 3.11).

Για την υπολογιστική βελτίωση των ConvLSTM οι συγγραφείς προτείνουν την χρήση χωριζόμενων συνελίξεων κατα βάθος μειώνοντας δραστικά τον αριθμό των παραμέτρων και κάνοντας την κάθε μονάδα συμπαγής και υπολογιστικά ελαφρύα. Παρακάτω βλέπουμε τους υπολογισμούς μέσα σε μια μονάδα SepConvLSTM.

$$i_t = \sigma({}_{1\times1}W_i^x * (W_i^x \circledast x_t) +_{1\times1} W_i^h * (W_i^h \circledast h_{t-1}) + b_i) \tag{3.2}$$

$$f_t = \sigma({}_{1\times1}W_f^x * (W_f^x \circledast x_t) +_{1\times1} W_f^h * (W_f^h \circledast h_{t-1}) + b_f) \tag{3.3}$$

$$\hat{c_t} = \tau({}_{1\times1}W_c^x * (W_c^x \circledast x_t) +_{1\times1} W_c^h * (W_c^h \circledast h_{t-1}) + b_c) \tag{3.4}$$

$$o_t = \sigma({}_{1\times1}W_o^x * (W_o^x \circledast x_t) +_{1\times1} W_o^h * (W_o^h \circledast h_{t-1}) + b_o) \tag{3.5}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \hat{c_t} \tag{3.6}$$

$$h_t = o_t \otimes \tau(c_t) \tag{3.7}$$

Στις παραπάνω εξισώσεις ο τελεστής $*$ αναπαριστά συνέλιξη, $\otimes$ αναπαριστά το γινόμενο Hadamard, σ αναπαριστά σιγμοειδή, τ αναπαριστά υπερβολική εφαπτομένη και $\circledast$ αναπαριστά χωριζόμενη συνέλιξη κατά βάθος. ${}_{1\times1}W$ και W είναι πυρήνες σημειακοί (pointwise) και κατά βάθος αντίστοιχα. Το κελί μνήμης c_t, η εσωτερική κατάσταση h_t και οι ενεργοποιήσεις των πυλών f_t, i_t και o_t είναι όλα 3-διάστατοι τένσορες. Αποτελεί έναν αποτελεσματικό τρόπο κωδικοποίησης τοπικών χωρο-χρονικών χαρτών χαρακτηριστικών ικανών για διαχωρισμό μεταξύ βίαων και μη βίντεο.

Προεπεξεργασία εισόδου

Στην συγκεκριμένη υλοποίηση, στην μία ροή του δικτύου, περνάνε οι διαφορές μεταξύ γειτονικών καρέ της εισόδου, ωθώντας το μοντέλο να κωδικοποιήσει τις χρονικές αλλαγές μεταξύ γειτονικών καρέ και τονώνοντας την πληροφορία κίνησης στο σημείο ενδιαφέροντος. Οι διαφορές των γειτονικών καρέ πρόκειται για μια πιο αποδοτική εναλλακτική στην υπολογιστικά απαιτητική οπτική ροή.

$$fd_i = frame_{i+1} - frame_i$$

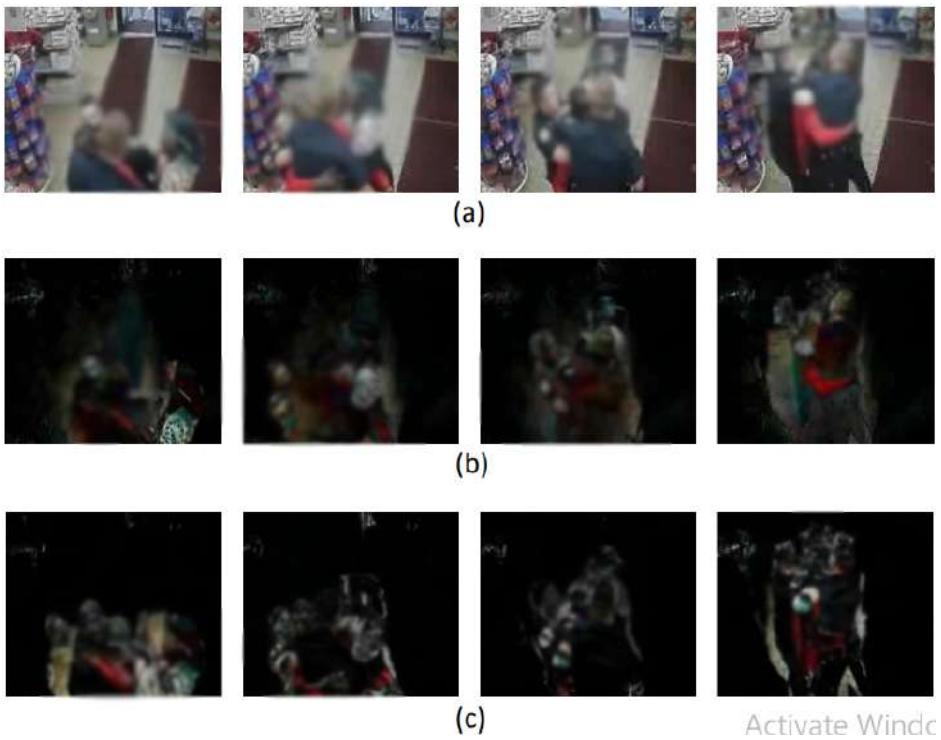
Στην παραπάνω εξίσωση το fd_i αναπαριστά την διαφορά μεταξύ των καρέ και το $frame_i$ το καρέ του βίντεο την χρονική στιγμή i.

Στην άλλη ροή, αντί για την χρήση των καρέ κατευθείαν, οι συγγραφείς προτείνουν προεπεξεργασία των καρέ του βίντεο κάνοντας καταστολή του παρασκηνίου της εικόνας. Για τον υπολογισμό του παρασκηνίου, χωρίς επιπλέον υπολογιστικό κόστος, υπολογίζεται πρώτα η μέση τιμή όλων των καρέ. Το μέσο καρέ περιέχει κυρίως πληροφορίες παρασκηνίου οι οποίες δεν αλλάζουν αρκετά μεταξύ των καρέ. Αφαιρώντας την μέση τιμή απο κάθε καρέ, τα κινούμενα αντικείμενα ενισχύονται καταπιέζοντας τις πληροφορίες του παρασκηνίου. Καθώς η περισσότερη πληροφορία των βίαιων ενεργειών, όπως προαναφέραμε, χαρακτηρίζεται απο κινήσεις του σώματος, και όχι του ακίνητου παρασκηνίου, αυτό κάνει το μοντέλο να εστιάσει περισσότερο στις σχετικές πληροφορίες. Παρακάτω βλέπουμε τις εξισώσεις που εξηγούν αυτή την διαδικασία.

$$avg = \sum_{i=0}^{N} \frac{frame_i}{N}$$

$$bsf_i = |frame_i - avg|$$

Στην παραπάνω εξίσωση, $frame_i$ είναι το i καρέ, το avg είναι η μέση τιμή όλων των καρέ και το bsf_i είναι το αποτέλεσμα της καταπίεσης του παρασκηνίου την στιγμή i που χρησιμοποιείται ως είσοδος στο μοντέλο.



Σχήμα 3.14: **Προεπεξεργασία εισόδου του μοντέλου.** Το (α) δείχνει μη επεξεργασμένα καρέ του βίντεο, το (β) δείχνει το αποτέλεσμα της καταστολής του βίντεο του (α) και το (c) δείχνει τις διαφορές των καρέ του βίντεο του (α) (Islam 2021 [9])

Stream Fusion

Για τον συνδυασμό των ροών μεταξύ τους και την τροφοδότηση στα πλήρως συνδεδεμένα επίπεδα ταξινόμησης οι συγγραφείς προτείνουν τρείς διαφορετικούς τρόπους ένωσης των δύο

ροών που οδηγούν σε τρεις παραλλαγές της αρχιτεκτονικής που φαίνονται παρακάτω.

$$f_{fused} = LeakyReLU(F_{frames}) \otimes Sigmoid(F_{diff}) \tag{3.8}$$

$$f_{fused} = Concat(F_{frames}, F_{diff}) \tag{3.9}$$

$$f_{fused} = F_{frames} \oplus F_{diff} \tag{3.10}$$

Γενικά, πρόκειται για ένα μοντέλο το οποίο πετυχαίνει ικανοποιητικά αποτελέσματα και είναι υπολογιστικά ελαφρύ καθιστώντας το ικανό για την χρήση σε εφαρμογές πραγματικού κόσμου όπου η χρονική απόκριση και η μη-απαίτηση ειδικού εξοπλισμού είναι υψίστης σημασίας. Επιδέχεται ενδεχομένως κάποια βελτίωση της επίδοσης με προ-εκπαίδευση σε κάποιο μεγάλο σύνολο δεδομένων ή χρήση περισσότερων επιπέδων LSTM.

3.2.5 Semi-Supervised Hard Attention

Η χρήση μεθόδων προσοχής, όπως έχει προαναφερθεί σε προηγούμενο κεφάλαιο, για αρκετό διάστημα περιοριζόταν σε προβλήματα επεξεργασίας φυσικής γλώσσας (NLP / Natural Language Processing). Ωστόσο με διάφορες προσπάθειες, όπως την ανάπτυξη των Δικτύων Χωρικών Μετατροπέων (STNs / Spatial Transformer Networks) [79], νέες αρχιτεκτονικές και μέθοδοι αναγνώρισης χωρο-χρονικών μοτίβων στα δεδομένα προσπαθούν να συμπεριλάβουν την χρήση μοντέλων προσοχής στην αρχιτεκτονική τους. Στην συγκεκριμένη υλοποιήση, οι συγγραφείς προτείνουν ένα μοντέλο Ημι-Επιβλεπόμενης μάθησης χρησιμοποιώντας "Αυστηρή" (Στοχαστική) Προσοχή [80].

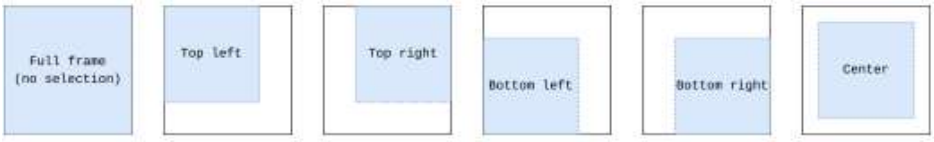
Η βασική έννοια απο την οποία πηγάζει η ιδέα της συγκεκριμένης υλοποίησης είναι ότι αφαιρώντας πλεονάζουσες πληροφορίες απο τα δεδομένα οδηγούν σε μοντέλο είτε μικρότερου μεγέθους (πλήθους παραμέτρων) με την ίδια επίδοση ή μοντέλα ίδιου μεγέθους με βελτιωμένη επίδοση [80]. Η αφαίρεση πλεοναζουσών πληροφοριών γεμίζουν τα δεδομένα εισόδου με πληροφορίες που θα μπορούσαν να χρησιμοποιηθούν για να αναπαριστήσουν πολύτιμα χαρακτηριστικά. Για την συγκράτηση των σημαντικών κομματιών από τα δεδομένα πολλές μέθοδοι χρησιμοποιούν βοηθητικά χαρακτηριστικά όπως εκτίμηση σκελετού [71] κ.α., ωστόσο τα χαρακτηριστικά διαφέρουν μεταξύ κάθε εφαρμογής και δεν μπορούν να χρησιμοποηθούν σε μεγάλο εύρος προβλημάτων Όρασης Υπολογιστών. Η συγκεκριμένη υλοποίηση χρησιμοποιεί στοχαστική προσοχή αντί για βοηθητικά χαρακτηριστικά βελτιώνοντας την δυνατότητα γενίκευσης και αποδοτικότητας σε μεγαλύτερο έυρος προβλημάτων [80].

Η έννοια της αυστηρής προσοχής μπορεί να ερμηνευτεί ως διαδικασία περικοπής πλεονασουσών πληροφοριών από την εικόνα κάθε καρέ βίντεο. Ως εκ τούτου, η αυστηρή προσοχή μπορεί να διατυπωθεί ως μια μέθοδος για τον ορισμό συντεταγμένων μιας συνάρτησης. Η παρακάτω εξίσωση περιγράφει αυτή την διαδικασία.

$$c = f_{crop}(arg_j max(f_{score}(h_j)), h) \tag{3.11}$$

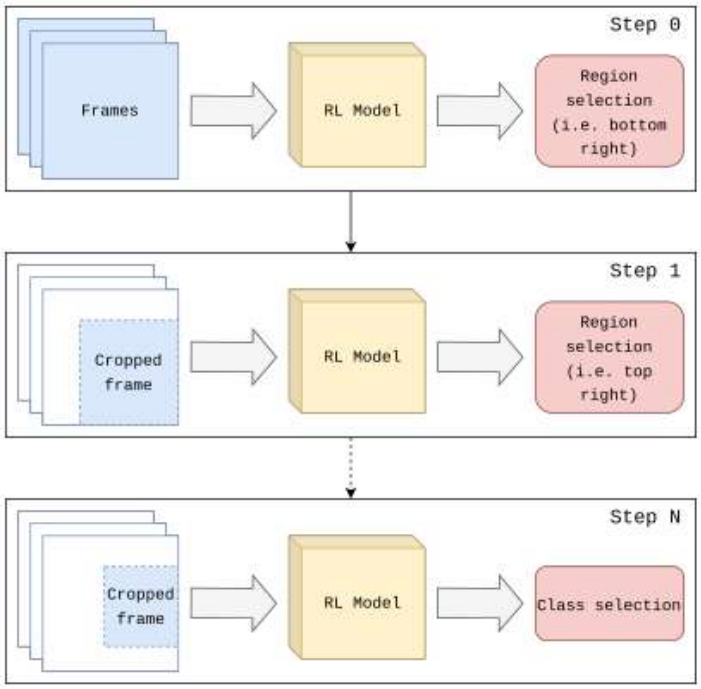
Στην παραπάνω εξίσωση, το c αναπαριστά την έξοδο της μονάδας προσοχής, f_{crop} την συνάρτηση περικοπής, h_j είναι η είσοδος στην περιοχή j και f_{score} είναι η συνάρτηση σκορ προσοχής. Η παραπάνω εξίσωση βασίζεται κυρίως στον υπολογισμό του "σκόρ προσοχής" f_{score} μέσω βαθιάς ενισχυόμενης μάθησης (DRL / Deep Reinforcement Learning). Η συνάρτηση f_{score} λαμβάνει ως είσοδο μια περιοχή h_j και βγάζει ως αποτέλεσμα μια βαθμολογία ανάλογα με την σημασία της στον εντοπισμό της βίας και η περιοχή με την μέγιστη βαθμολογία περικόπτεται χρησιμοποιώντας την συνάρτηση f_{score} και θεωρείται η έξοδος.

Η παραπάνω διαδικασία υλοποιείται σε πολλά στάδια. Το περιβάλλον ενισχυόμενης μάθησης χωρίζει την εικόνα σε προκαθορισμένα "κουτιά", όπως φαίνεται στην παραπάνω εικόνα. Το μοντέλο σε κάθε στάδιο, διαλέγει μια περιοχή απο την εικόνα και αντικαθιστά το συγκεκριμένο καρέ



Σχήμα 3.15: **Προκαθορισμένοι χώροι προσοχής.** (Mohammadi 2023 [80])

με την περιοχή αυτή. Στη συνέχεια ο μηχανισμός προσοχής μπορεί να συνεχίζει επαναληπτικά αυτή την διαδικασία εστιάζοντας σε ολοένα και μικρότερο σημείο.



Σχήμα 3.16: **Μοντέλο προσοχής λαμβάνοντας ως είσοδο βίντεο.** (Mohammadi 2023 [80])

Ο ορισμός στατικών περιοχών αντί για ελευθέρως κινούμενων αντικειμένων βοηθάει στην σύγκλιση του μηχανισμού ενισχυόμενης μάθησης. Ο αλγόριθμος ενισχύομενης μάθησης στον μηχανισμό προσοχής έχει εμπνευστεί απο (A. Manchin 2019 [81]). Στο μοντέλο RL (Επιβλεπόμενης μάθησης) δίνονται δύο ειδών επιβραβεύσεις. Οι επιβραβεύσεις ορίζονται ως +1 για σωστή ταξινόμηση του βίντεο και -1 για λάθος ταξινόμιση του βίντεο. Επίσης, δίνεται μια +0.5 επιβράβευση για να προωθήσει το μοντέλο προσοχής να πειραματιστεί με την επιλογή περιοχών. Το μοντέλο χρησιμοποιεί την μέθοδο Q-learning για την εκπαίδευσή του, επιλέγοντας την περιοχή που αποδίδει τη μεγαλύτερη τιμή Q με βάση την παρακάτω εξίσωση:

$$Q(s,a) = R(s,a) + \gamma max_{\alpha'} Q'(s',\alpha')$$

Στην παραπάνω εξίσωση $Q(s,a)$ αναπαριστά την νέα τιμή Q με βάση την τωρινή κατάσταση s και την δράση a , $R(s,a)$ είναι η ανταμοιβή, $max_{\alpha'} Q'(s',\alpha')$ είναι η μέγιστη αναμενόμενη τιμή Q που λαμβάνεται με βάση την αναμενόμενη κατάσταση s' και μελλοντική δράση a' .

Για το υπόλοιπο μοντέλο οι συγγραφείς χρησιμοποιούν ένα προεκπαιδευμένο I3D μοντέλο [16] στο σύνολο δεδομένων Kinetics [82]. Εξετάζονται πολλές παραλλαγές χρησιμοποιώντας

είτε οπτική ροή ή RGB καρέ μόνο ή συνδυασμό και των δύο. Ωστόσο λόγω του σχετικά μικρού μεγέθους του dataset, πετυχαίνονται καλύτερα αποτελέσματα χρησιμοποιώντας μόνο RGB καρέ.

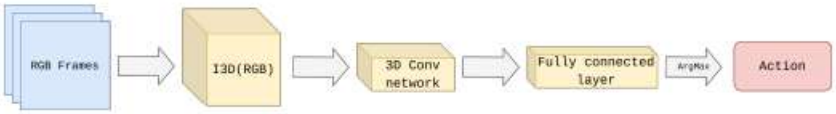


Figure 3: SSHA model architecture (RBG only).

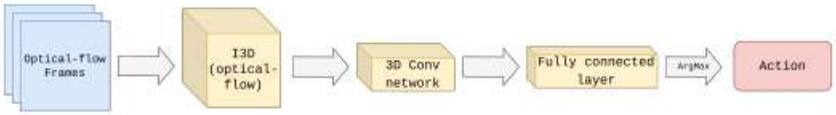


Figure 4: SSHA model architecture (Optical-flow only).

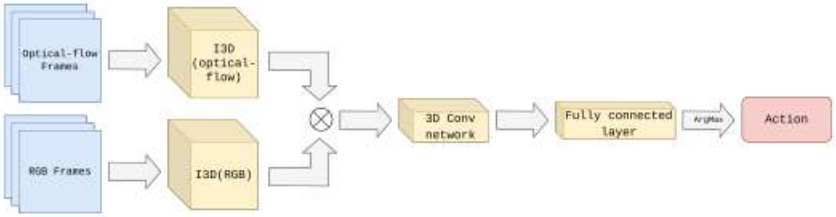


Figure 5: SSHA model architecture (Two-stream fusion).

Σχήμα 3.17: **Διαφορετικές παραλλαγές αρχιτεκτονικής SSHA.** (Mohammadi 2023 [80])

Το συγκεκριμένο μοντέλο πετύχαινε, την στιγμή της έκδοσής του, τελευταίας τεχνολογίας επίδοση. Ωστόσο, η βελτίωση στην επίδοση σε σχέση με την αμέσως επόμενη καλύτερη αρχιτεκτονική (Separable Convolutional LSTM [9]) ήταν σχεδόν αμελητέα (89.75% έναντι 90.4%) και το μοντέλο χρησιμοποιεί πολύ πιο πολύπλοκη αρχιτεκτονική και μεθόδους εκπαίδευσης. Δυστυχώς, οι συγγραφείς δε μας παρέχουν πολλές πληροφορίες για την ακριβής αρχιτεκτονική οπότε δεν μπορούμε να συνάγουμε συμπεράσματα για την αποδοτικότητά του σε σχέση με τα άλλα μοντέλα. Πρόκειται παρ'όλα αυτά μια πολύ καλή και πρωτοποριακή χρήση αυστηρής προσοχής για την αναγνώριση σκηνών βίας και ενεργειών γενικότερα, που ενδεχομένως επιδέχεται βελτίωση στο μέλλον με την χρήση πολλών μετώπων προσοχής ταυτόχρονα χρησιμοποιώντας πολλούς "agents" προσοχής.

Κεφάλαιο 4

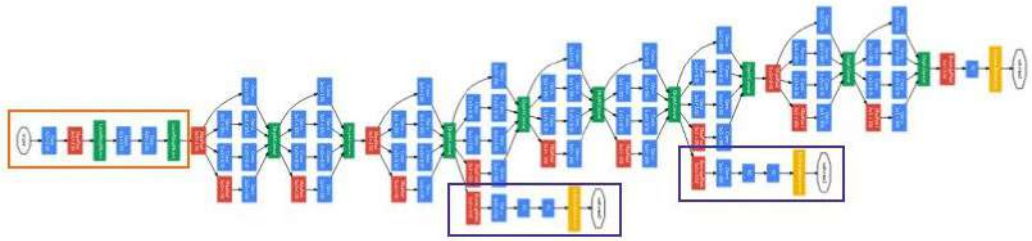
Πειράματα & Αποτελέσματα

4.1 Μεταφορά Μάθησης με ΣΝΔ

Στην παρακάτω υλοποίηση, και πρώτη προσπάθεια στην αντιμετώπιση του προβλήματος αναγνώρισης βίαιων σκηνών απο βίντεο, η προσπάθεια στρέφεται προς την χρήση ενός βαθιού δισδιάστατου συνελικτικού νευρωνικού δικτύου. Θα χρησιμοποιηθούν βαθιά δισδιάστατα συνελικτικά νευρωνικά δίκτυα για την ταξινόμηση των βίντεο καρέ προς καρέ. Για την βέλτιστη δυνατότητα εξαγωγής βίαιων χαρακτηριστικών απο εικόνες, θα γίνει χρήση μεταφοράς μάθησης από επαρκώς μεγάλα δίκτυα. Τα δίκτυα είναι προεκπαιδευμένα σε ένα τεράστιο σύνολο δεδομένων αναγνώρισης εικόνων σε εκατοντάδες κατηγορίες, όπως το ImageNet [36]. Έπειτα, προσαρμόζονται για την χρήση σε προβλήματα αναγνώρισης βίαιων χαρακτηριστικών με την χρήση fine-tuning. Τα μοντέλα γίνονται διαθέσιμα για χρήση μέσω της βιβλιοθήκης ανοιχτού κώδικα ανάπτυξης νευρωνικών δικτύων Keras [83] που εντάσσεται στην βιβλιοθήκη Tensorflow [84].

4.1.1 Επισκόπηση Αρχιτεκτονικής

InceptionNet-V3

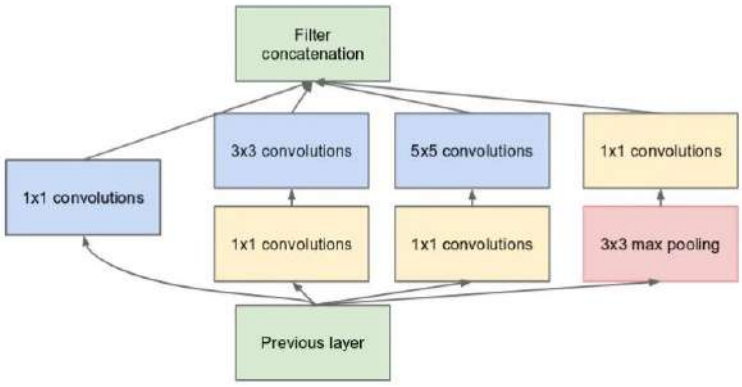


Σχήμα 4.1: **Αρχιτεκτονική InceptionNet-V3** (Szegedy et al. 2015 [85])

Η αρχιτεκτονική Inception [85] έφερε επανάσταση στον τομέα των ταξινομητών CNN στον τομέα της επίδοσης αλλά και ταυτόχρονα στον τομέα της απόδοσης και ταχύτητας των μοντέλων. Πριν την εισαγωγή της αρχιτεκτονικής Inception η κλασσική αντιμετώπιση για την ανοικοδόμηση ταξινομητών CNN εικόνων, ξεκινώντας απο το Le-Net [86], ήταν η στοίβαξη

ολοένα και περισσότερων συνελικτικών επιπέδων οδηγώντας σε βαθύτερα και μεγαλύτερα δίκτυα. Η *αφελής* στοίβαξη περισσότερων επιπέδων ωστόσο, αυξάνει το υπολογιστικό κόστος των δικτύων που οδηγεί σε μείωση της απόδοσής τους αλλά ταυτόχρονα εισάγει και έναν σημαντικό βαθμό δυσκολίας κατα την εκπαίδευσή τους. Πιο σημαντικό όμως είναι το γεγονός ότι πιο βαθιά δίκτυα είναι πιο ευάλωτα στην υπερβολική προσαρμογή και είναι πιο δύσκολη η οπίσθια διάδοση των κλίσεων κάνοντας την εκπαίδευσή τους ασύμφορη.

Στο άρθρο **Going deeper with convolutions** [85] οι συγγραφείς παρουσίασαν διάφορες μεθόδους και αρχιτεκτονικές για την αντιμετώπιση πολλαπλών προβλημάτων που παρουσιάζονταν με τις παραδοσιακές μεθόδους ταξινόμησης εικόνων. Το πρώτο πρόβλημα που προσπάθησαν να αντιμετωπίσουν είναι το γεγονός ότι τα αντικείμενα ενδιαφέροντος που καλούνται να αναγνωρίσουν έχουν υπερβολικά μεγάλη διακύμανση στο μέγεθός τους μεταξύ διαφορετικών εικόνων. Η λύση στο παραπάνω πρόβλημα είναι το λεγόμενο "Inception Module" όπου διαφορετικού μεγέθους φίλτρα στοιβάζονται μεταξύ τους και έπειτα συνενώνονται οδηγώντας σε "πλατύτερο" αντί για βαθύτερο δίκτυο. Έπειτα για την μείωση των καναλιών εισόδου τα φίλτρα περνάνε απο 1×1 συνελίξεις και μετά απο max-pooling.



(b) Inception module with dimension reductions

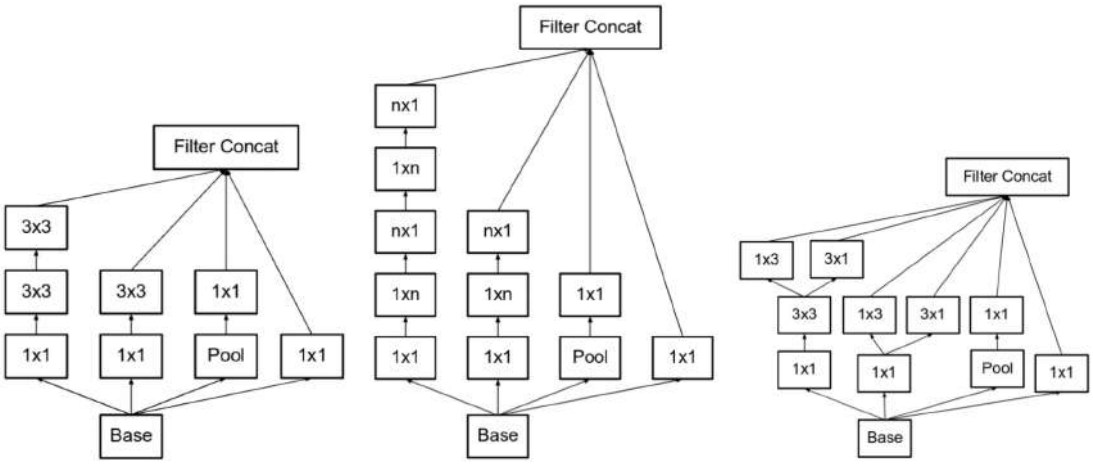
Σχήμα 4.2: **Inception Module με 1x1 συνέλιξη** (Szegedy et al. 2015 [85])

Το InceptionNet-V1 στοιβάζει 9 τέτοια modules στο ίδιο δίκτυο και έχει σύνολο 27 επίπεδα (μαζί με τα επίπεδα υποδειγματοληψίας). Συνεπώς πρόκειται για αρκετά βαθύ δίκτυο το οποίο όπως όλα τα άλλα δίκτυα υπόκειται στα ίδια προβλήματα υπερβολικής προσαρμογής και εξαφανιζόμενων κλίσεων. Η λύση που προτείνουν οι συγγραφείς είναι η χρήση "περιφερειακών ταξινομητών" (auxiliary classifiers) τα οποία είναι τα μωβ τετράγωνα στο Σχήμα 4.1. Με την χρήση ενός "περιφερειακού κόστους" (auxiliary loss) καταφέρνουν να αποτρέψουν το πρόβλημα των εξαφανιζόμενων κλίσεων στο μεσαίο κομμάτι κατά την διάρκεια της εκπαίδευσης και εφαρμόζεται softmax στην έξοδό τους. Η συνολική συνάρτηση κόστους υπολογίζεται από την παρακάτω εξίσωση:

$$Loss_{total} = Loss_{real} + 0.3 \times Loss_{aux_1} + 0.3 \times Loss_{aux_2}$$

Στο ίδιο άρθρο οι εκδότες παρουσιάσαν τα InceptionNet-V2 και InceptionNet-V3 εισάγοντας αλλαγές που βελτιώνουν την επίδοση και υπολογιστική πολυπλοκότητα του μοντέλου. Οι εκδότες είχαν στόχο να μειώσουν το λεγόμενο "representational bottleneck" όπου συνελίξεις που μειώνουν δραστικά την διάσταση της εισόδου μπορεί να οδηγήσουν σε χάσιμο της πληροφορίας. Οι αλλαγές που εισάγαν ήταν η παραγοντοποίηση των 5×5 συνελίξεων σε δύο 3×3 αλλά

και η παραγοντοποίηση συνελίξεων $n \times n$ σε δύο συνελίξεις $1 \times n$ και $n \times 1$ δραστικά μειώνοντας το υπολογιστικό κόστος υπολογισμού των συνελίξεων. Έπειτα, το επόμενο βήμα ήταν το "σπάσιμο" των φίλτρων κατά πλάτος για την αποφυγή μείωσης της διάστασης της εισόδου και συνεπώς της πληροφορίας.



(a) Παραγοντοποίηση συνελίξεων (b) Σπάσιμο συνελίξεων (c) Σπάσιμο κατά πλάτος

Σχήμα 4.3: **Βελτιώσεις InceptionNet-V2** Στην εικόνα (a) βλέπουμε την παραγοντοποίηση της συνέλιξης 5x5 σε δύο συνελίξεις 3x3, στην εικόνα (b) την παραγοντοποίηση των συνελίξων nxn σε 1xn & nx1 και στην εικόνα (c) το σπάσιμο των συνελίξεων κατά πλάτος αντί για κατά βάθος. (Szegedy et al. 2015 [85])

Τέλος, στην τρίτη επανάληψη InceptionNet-V3 εισήχθαν αλλαγές προκειμένου να λυθεί το πρόβλημα ότι οι περιφερειακοί ταξινομητές δεν συμβάλαν τόσο προς το τέλος της εκπαίδευσης όπου η επιδόσεις πλησιάζαν τον κορεσμό. Εισάγαν αλλαγές περί την επίλυση αυτού του προβλήματος και την βελτίωση του InceptionNet-V2 χωρίς την δραστική αλλαγή της αρχιτεκτονικής.

4.1.2 Υλοποίηση Πειράματος

Εκπαίδευση Δικτύου

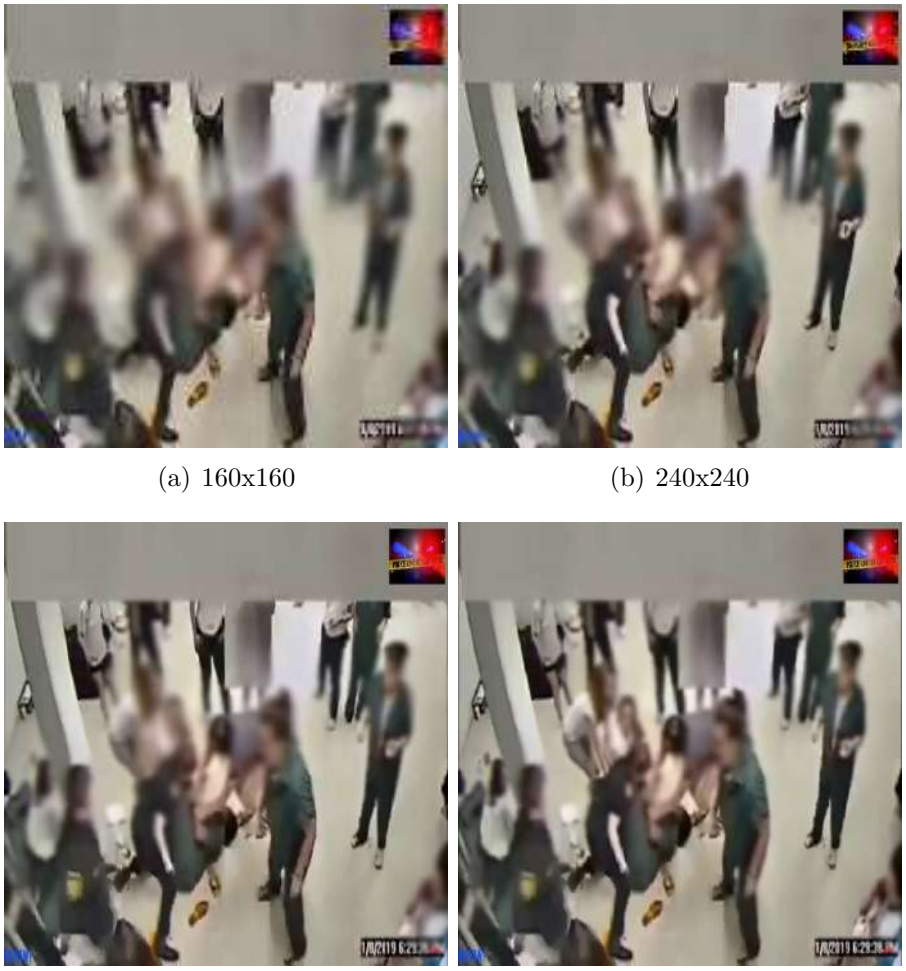
Για την εκπαίδευση του δικτύου και αξιολόγηση των αποτελεσμάτων χρησιμοποιείται αποκλειστικά το σύνολο δεδομένων RWF-2000 [67], καθώς όπως έχει αποδειχτεί και σε προηγούμενο κεφάλαιο παρέχει τις πιο ρεαλιστικές σκηνές πραγματικής βίας, αλλά και ταυτόχρονα το μεγαλύτερο όγκο δεδομένων. Για την διευκόλυνση της επεξεργαστικής δυσκολίας της εκπαίδευσης του μοντέλου τα δεδομένα βίντεο αποθηκεύονται απο πριν καρέ προς καρέ μαζί με τις ετικέτες τους. Για την προεπεξεργασία τους και την τροφοδότηση του μοντέλου χρησιμοποιείται η βιβλιοθήκη ImageDataGenerator [87] που αποτελεί μέρος της βιβλιοθήκης Keras [83].

Αρχικά, το μοντέλο InceptionNet-V3 εισάγεται μέσω της βιβλιοθήκης Keras [83] μαζί με τα βάρη του εκπαιδευμένα στο ImageNet [36]. Το τελευταίο επίπεδο εξόδου, που μας δίνει το αποτέλεσμα ταξινόμησης, έχει αρχικά 1000 νευρώνες εξόδου (όσες και οι κατηγορίες του ImageNet) όπου απορρίπτεται εντελώς για την προσαρμογή του στο πρόβλημά που εξετάζεται. Έπειτα, προστίθεται ένα επίπεδο "Global Average Pooling 2D" που εφαρμόζει average pooling (υποδειγματοληψία μέσου) έως όλες οι χωρικές διαστάσεις να ισούνται με 1, αφήνωντας τις

υπόλοιπες μη επηρεασμένες. Στη συνέχεια προστίθενται δύο νευρώνες εξόδου, όσες και οι κατηγορίες ταξινόμησης (Βία / Μη-Βία).

Το δίκτυο -πλην μερικών επιπέδων προς το επίπεδο εξόδου- "παγώνεται" με τα βάρη να είναι ανεπηρέαστα απο την εκπαίδευση. Αυτό γίνεται καθώς η προεκπαίδευση έχει οδηγήσει το δίκτυο στην εκμάθηση κάποιων χρήσιμων χαρακτηριστικών απο τα δεδομένα, με τα αρχικά επίπεδα να έχουν μάθει βασικά χαρακτηριστικά όπως ακμές γωνίες και σχήματα τα οποία είναι επιθυμητό να συγκρατήσουμε και να προσαρμόσουμε μόνο τα εξωτερικά επίπεδα στο πρόβλημά μας. Καθώς έχουμε δύο νευρώνες εξόδου και η κατηγορίες είναι αμοιβαία αποκλειόμενες, χρησιμοποιείται η συνάρτηση ενεργοποίησης "softmax" και η συνάρτηση κόστους Sparse Categorical Cross Entropy.

Στην αρχή εκπαιδεύεται το δίκτυο στο σύνολο δεδομένων RWF-2000 [67], χωρίς data augmentation με χρήση υπεραραμέτρων ρυθμού μάθησης και momentum με τιμές 0.0001 και 0.9 αντίστοιχα. Το δίκτυο εκπαιδεύεται για 5 εποχές αρχικά με διαφορετικά μεγέθη εισόδου για την εύρεση της βέλτιστης σχέσης απόδοσης/επίδοσης. Όσο μεγαλώνει το μέγεθος εισόδου, τόσο μεγαλώνει το πλήθος πληροφοριών το οποίο είναι διαθέσιμο στα δεδομένα, ωστόσο τόσο μεγαλώνει και η υπολογιστική πολυπλοκότητα εκπαίδευσης του δικτύου φτάνοντας ένα σημείο που είναι πλέον ασύμφορη η χρήση μεγαλύτερης εισόδου. Στόχος είναι η εύρεση της "χρυσής τομής" μεταξύ των προαναφερθέντων μετρικών.

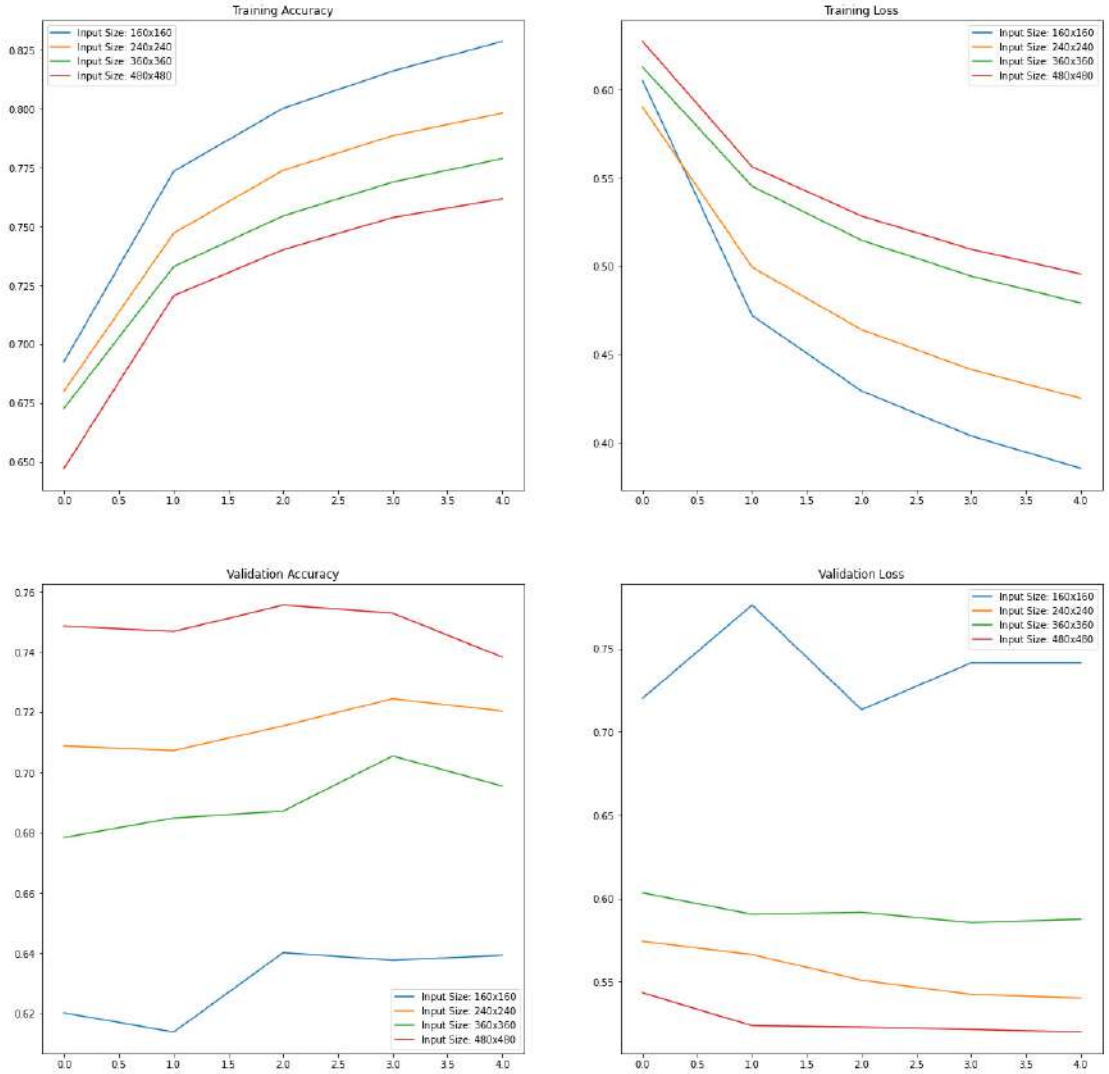


(a) 160x160 (b) 240x240

(c) 360x360 (d) 480x480

Σχήμα 4.4: **Διαφορετικά μεγέθη εικόνων εισόδου**

Παρ΄όλο που η διαφορά της ποιότητας μεταξύ τους δεν είναι τόσο εμφανής στην μορφή που παρουσιάζονται στην παρούσα εργασία, παρατηρείται μια σημαντική διαφορά της ποιότητας μεταξύ των μεγέθων $(160, 160)$ και $(240, 240)$, ενώ μεταξύ των $(360, 360)$ και $(480, 480)$ η διαφορά τους δεν είναι τόσο εμφανής σε έναν ανθρώπινο παρατηρητή. Είναι αξιοσημείωτο να σημειωθεί ωστόσο ότι το πλήθος των pixels ανεβαίνει εκθετικά από το ένα επίπεδο στο άλλο. Το μέγεθος της εισόδου καθορίζει το μέγεθος των χαρτών χαρακτηριστικών και την υπολογιστική πολυπλοκότητα της εφαρμογής των συνελίξεων των φίλτρων σε αυτά.



Σχήμα 4.5: **Σύγκριση αποτελεσμάτων μεταξύ διαφορετικών μεγεθών εισόδου**

Στο Σχήμα 4.5 βλέπουμε σύγκριση μεταξύ των διαφορετικών μεγέθων εισόδου σε 4 διαφορετικές μετρικές Training Accuracy, Training Loss, Validation Accuracy, Validation Loss με τις δύο πρώτες να αντιπροσωπεύουν την επίδοση και σφάλμα στα δεδομένα εκπαίδευσης και τα δύο τελευταία στα δεδομένα επαλήθευσης αντίστοιχα. Από τα δύο πρώτα τεταρτημόρια φαίνεται, παραδόξως, η επίδοση του δικτύου να εξαρτάται αντίστροφα απο το μέγεθος εισόδου, δηλαδή τα δίκτυα με μικρότερο μέγεθος εισόδου έχουν καλύτερη επίδοση. Ωστόσο, λαμβάνοντας υπ'όψιν και το 3ο και 4ο τεταρτημόριο του σχήματος 4.5 στην πραγματικότητα η επίδοση δεν βελτιώνεται, αλλά τα δίκτυα είναι πιο ευάλωτα στην υπερ-προσαρμογή (over-fitting) όσο μικρότερο είναι

το μέγεθος εισόδου. Ωστόσο, το μέγεθος εισόδου $(240, 240)$ αποδείχτηκε ως η βέλτιστη τομή μεταξύ επίδοσης / απόδοσης και θα είναι το μέγεθος που θα χρησιμοποιηθεί.

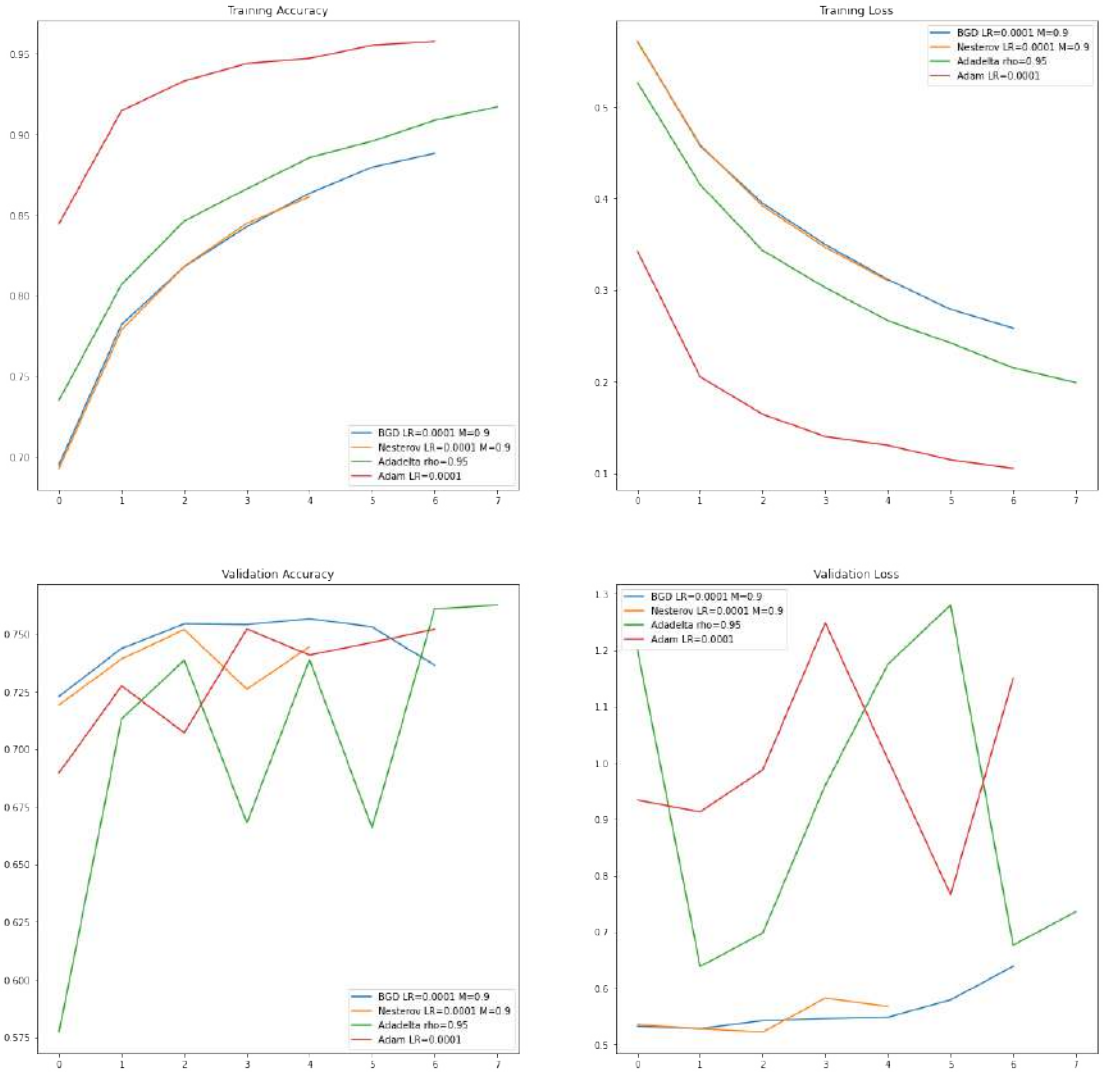
Απ'ότι φαίνεται απο τα δεδομένα στο Σχήμα 4.5, τα δίκτυα φαίνεται να φτάνουν στον κορεσμό σχεδόν άμεσα, μη βελτιώνοντας παραπάνω την επίδοσή τους μετά την πρώτη εποχή εκπαίδευσης και οδηγούνται στην υπερβολική προσαρμογή για να βελτιώσουν περαιτέρω την συνάρτηση κόστους. Για την απόδειξη του παραπάνω κάνουμε χρήση Data Augmentation, επεξεργάζοντας τα δεδομένα με τυχαίο ζουμ, φωτεινότητα, αλλαγή των αξόνων κ.α. έτσι ώστε το δίκτυο να αναγκαστεί να εξάγει τις πιο σημαντικές πληροφορίες βίας απο τα δεδομένα και να γίνει πιο δύσκολο το δίκτυο να απομνημονευτεί τα δεδομένα.

Σχήμα 4.6: **Σύγκριση αποτελεσμάτων μεταξύ διαφορετικών μεγεθών εισόδου**

Η υπερβολική προσαρμογή έχει μειωθεί με την χρήση Data Augmentation όπως φαίνεται απο το Σχήμα 4.6, καθώς η διαφορά μεταξύ σφάλματος εκπαίδευσης και επαλήθευσης έχει έρθει πιο κοντά. Παρ'όλαυτα η επίδοση επαλήθευσης παραμένει σχεδόν σταθερή κατά την διάρκεια της εκπαίδευσης, αλλά και σε ίδιες τιμές επίδοσης σε σχέση με εκείνες χωρίς την χρήση Data Augmentation που υποδηλώνει ότι η υπερβολική προσαρμογή είναι αναπόφευκτη καθώς η επίδοση επαλήθευσης έχει κορεστεί. Ωστόσο, οι παράμετροι του δικτύου μένουν ανεπηρέαστες από την εκπαίδευση εκτός απο το επίπεδο εξόδου που ενδεχομένως να οδηγεί στον κορεσμό. Για την ανάλυση του παραπάνω γεγονότος διερευνείται η περίπτωση της εκπαίδευσης ολόκληρου του δικτύου και όχι μόνο των τελευταίων επιπέδων.

Στη συνέχεια η προσπάθεια στρέφεται στην λεπτή ρύθμιση (fine-tuning) του συνολικού δικτύου, δηλαδή το σύνολο όλων των παραμέτρων του δικτύου "ξε-παγώνονται", επιτρέπωντας την δυνατότητα στους αλγόριθμους βελτιστοποίησης να προσαρμόσουν συνολικά τα βάρη του δικτύου. Τυπικά, κατά την διαδικασία της λεπτής ρύθμισης στην μεταφορά μάθησης αρκετά επίπεδα παγώνονται και το δίκτυο εκπαιδεύεται σταδιακά ξεπαγώνοντας περισσότερα επίπεδα και μειώνοντας τον ρυθμό μάθησης προκειμένου το δίκτυο να μην προσαρμοστεί υπερβολικά στα δεδομένα και χάσει τις χρήσιμες πληροφορίες που έχει αντλήσει κατά τη μεταφορά μάθησης. Ωστόσο, για την επιλογή των επιπέδων που θα δεχτούν λεπτή ρύθμιση δεν υπάρχει κάποιος γενικός κανόνας και έχουν αναπτυχθεί διάφορες ευριστικές τεχνικές [88, 89] με διαφορετικά επίπεδα επιτυχίας.

Στην συγκεκριμένη υλοποίηση, οποιαδήποτε απο τις προαναφερθέντες τεχνικές επιλογής επιπέδων για λεπτή ρύθμιση δεν απέδωσε καρπούς και στράφηκε η προσπάθεια στην λεπτή ρύθμιση όλων των παραμέτρων του δικτύου. Για την εκπαίδευση του συνολικού δικτύου διερευνούνται 4 διαφορετικοί αλγόριθμοι βελτιστοποιήσης με στόχο την εύρεση της βέλτιστης τεχνικής προσαρμογής βαρών. Για την κατάλληλη επιλογή λαμβάνονται υπ'όψιν η αρχιτεκτονική του μοντέλου, ο όγκος του συνόλου δεδομένων αλλά και η ταχύτητα σύγκλισης. Τα δίκτυα εκπαιδεύονται με τον κάθε ξεχωριστό αλγόριθμο έως ότου να αρχίσουν να προσαρμόζουν υπερβολικά. Παρακάτω φαίνονται τα αποτελέσματα του πειράματος.



Σχήμα 4.7: **Σύγκριση αποτελεσμάτων μεταξύ διαφορετικών αλγορίθμων βελτιστοποίησης**

Στο σχήμα 4.7 βλέπουμε την λεπτή ρύθμιση του δικτύου χρησιμοποιώντας 4 διαφορετικούς αλγορίθμους βελτιστοποίησης. Θεωρητικά, η πιο λογική επιλογή αλγορίθμου βελτιστοποιήσης δεδομένης της αρχιτεκτονικής του InceptionNet-V3 αποτελεί ο αλγόριθμος Adam καθώς περιέχει όλα τα πλεονεκτήματα του αλγορίθμου RMSProp ο οποίος είναι η προκαθορισμένη επιλογή για το δίκτυο, αλλά χρησιμοποιεί και momentum. Δεδομένης της φύσης των δεδομένων στην θεωρία καλύτερη είναι η χρήση του αλγορίθμου Adadelta καθώς δουλεύει καλύτερα με μικρότερα σύνολοα δεδομένων [90]. Ωστόσο, απ'ότι φαίνεται απο το Σχήμα 4.7 στην πράξη ο αλγόριθμος Adadelta -αν και σχεδόν ίσο με τη χρήση momentum- πετυχαίνει το καλύτερο αποτέλεσμα αλλά ταλαντεύεται αρκετά γύρω απο την εύρεση τοπικού ελάχιστου. Το γεγονός αυτό οφείλεται στο ότι οι ταλαντώσεις που υπάρχουν κοντά σε ένα τοπικό ελάχιστο εξομαλύνονται με την χρήση momentum αλλά με τον Adadelta αθροίζονται στον αριθμητή [31], για το οποίο η χαμηλότερη τιμή rho ίσως να οδηγούσε σε μικρότερη ταλάντωση. Ο αλγόριθμος Adadelta καταλήγει σε ένα καλύτερο τοπικό ελάχιστο, αλλά με πολύ μικρή διαφορά μεταξύ της απλής χρήσης momentum και προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης, οπότε η επιλογή του αλγορίθμου Batch Gradient Descent με χρήση momentum είναι η λογική επιλογή.

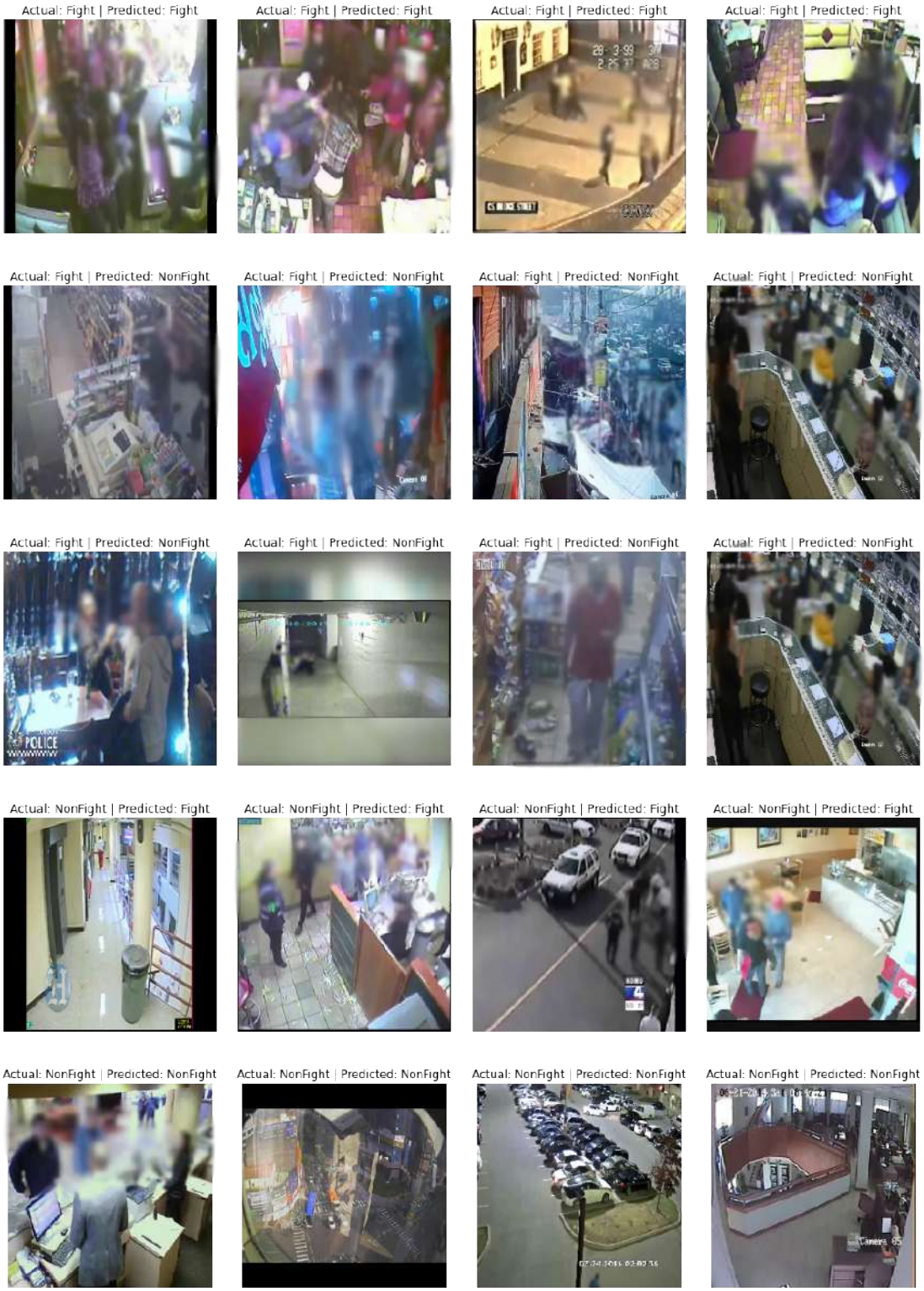
Αξιολόγηση αποτελεσμάτων

Όπως φαίνεται απο το σχήμα 4.8, το μοντέλο είναι ικανό να εντοπίσει σκηνές βίας στις οποίες υπάρχει ξεκάθαρη ανθρώπινη επαφή μεταξύ ενός η πολλών ατόμων ή μη βίαιες σκηνές στις οποίες δεν υπάρχουν άνθρωποι ή βρίσκονται σε μεγάλη απόσταση μεταξύ τους. Ένα σημαντικό πρόβλημα στην αξιολόγηση (και συνεπώς στην εκπαίδευση του δικτύου) είναι ότι σε ένα ολόκληρο βίντεο, υπάρχουν αρκετά μεμονωμένα αποσπάσματα καρέ χωρίς εμφανή βίαια χαρακτηριστικά. Έτσι εισάγεται στο δίκτυο θόρυβος και οδηγούνται οι αλγόριθμοι βελτιστοποιήσης να προσαρμόσουν τις κλίσεις ως προς τυχαία χαρακτηριστικά που υπάρχουν στις εικόνες, μειώνοντας την "στιβαρότητα" του δικτύου.

		Precision	Recall	F-1	Accuracy
RWF-2000	**Βία**	77%	72%	75%	**76%**
	Μη-Βία	74%	79%	76%	
Hockey Fights	**Βία**	67%	51%	58%	**68%**
	Μη-Βία	60%	75%	67%	
Movie Fights	**Βία**	67%	55%	61%	**67%**
	Μη-Βία	63%	74%	68%	

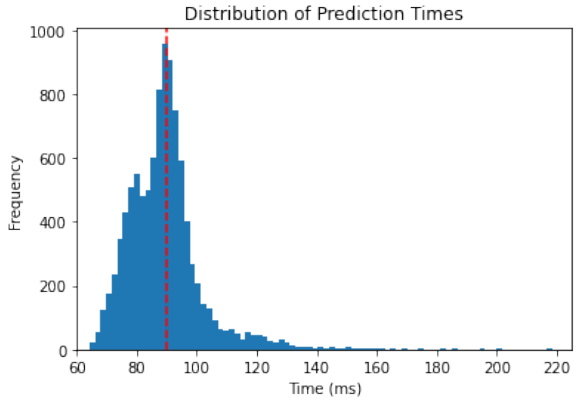
Πίνακας 4.1: **Επιδόσεις μοντέλου σε διαφορετικά σύνολα δεδομένων**

Στον Πίνακα 4.1 βλέπουμε τις επιδόσεις του μοντέλου σε κάθε σύνολο δεδομένων ξεχωριστά. Η χειρότερη απόδοση στα σύνολα δεδομένων Hockey και Movies υποδηλώνει ότι το μοντέλο δεν έχει γενικεύσει την αναγνώριση χαρακτηριστικών βίας και ενδεχομένως επηρεάζεται από συνθήκες στο περιβάλλον όπως κατεύθυνση της κάμερας, σωματική επαφή με παραπάνω απο 2 άτομα και άλλα. Επίσης, παρατηρείται μεγαλύτερη τιμή Precision και μικρότερη Recall στην κατηγορία βίαιων σκηνών και μεγαλύτερη τιμή Recall και χαμηλότερη Precision στην κατηγορία μη βίαιων. Το πρώτο μέρος της προηγούμενης πρότασης οφείλεται στο γεγονός ότι όταν μια σκηνή ταξινομείται ως βίαια είναι γιατί περιέχει ενδεχομένως βίαια χαρακτηριστικά (τα οποία δεν υπάρχουν κατά κύριο λόγο στα μη βίαια βίντεο) και άρα πιο πιθανό να ταξινομήθηκε σωστά (υψηλότερο precision), ωστόσο πολλά μεμονωμένα καρέ βίαιων βίντεο δεν περιέχουν βίαια χαρακτηριστικά οδηγώντας τα να ταξιμομηθούν λανθασμένα ως μη βίαια (χαμηλότερο recall). Το δέυτερο μέρος της πρότασης εξηγείται με τον αντίθετο τρόπο απο εκείνον της πρώτης.



Σχήμα 4.8: **Ποιοτική αξιολόγηση αποτελεσμάτων.** Στην πρώτη σειρά βλέπουμε σωστές προβλέψεις βίαιων σκηνών, στην δεύτερη και τρίτη λάθος πρόβλεψη βίαιων σκηνών, στην τέταρτη λάθος πρόβλεψη μη βίαιων σκηνών και στην πέμπτη σωστή πρόβλεψη μη βίαιων σκηνών.

Τέλος, στο σχήμα 4.9 φαίνεται η απόδοση του δικτύου στην μορφή της χρονικής απόκρισης κατά την αξιολόγηση ενός μεμονωμένου καρέ. Ένω, η μέση τιμή των 90 millisecond μπορεί να φαίνεται ως μικρή, στην πράξη αυτό σημαίνει ότι το μοντέλο έχει άνω όριο ταξινόμησης έως 11 καρέ ανά δευτερόλεπτο. Ωστόσο, η αξιολόγηση εικόνων είναι ως έναν βαθμό φραγμένη απο την είσοδο-έξοδο εικόνων (I/O Bound) για το οποίο υπάρχουν τρόποι αντιμετώπισης κατά την ανάπτυξή εφαρμογών πραγματικού κόσμου.



Σχήμα 4.9: **Χρονική απόκριση μοντέλου για την αξιολόγηση 1 εικόνας.**

4.1.3 Συμπεράσματα Υλοποίησης

Η χρήση μεταφοράς μάθησης και της αρχιτεκτονικής InceptionNet-V3 ενώ παράγει αξιοπρεπή αποτελέσματα δεν είναι ικανοποιητικά για την χρήση τους σε εφαρμογές πραγματικού κόσμου. Αυτό οφείλεται κυρίως στο γεγονός ότι το μοντέλο δεν μπορεί να εξάγει χωροχρονικές εξαρτήσεις απο τα δεδομένα και ταξινομεί τις εισόδους καρέ προς καρέ. Αυτός ο περιορισμός οδηγεί σε μη ικανοποιητικά αποτέσματα, όχι μόνο διότι τα δεδομένα έχουν έντονες χρονικές εξαρτήσεις, αλλά και διότι σε ένα τμήμα βίντεο υπάρχουν πολλές μεμονωμένες χρονικές στιγμές (που αντιπροσωπεύονται απο μεμονωμένα καρέ βίντεο) στις οποίες δεν υπάρχουν βίαια χαρακτηριστικά είτε λόγω παύσης ή για άλλους λόγους. Έτσι το δίκτυο αναγκάζεται να ταξινομήσει το συγκεκριμένο καρέ λανθασμένα οδηγώντας στην φαινόμενη χειρότερη επίδοση και συνεπώς δυνατότητα εκπαίδευσης. Ωστόσο, ανεξαιρέτως των προαναφερθέντων είναι εντυπωσιακή η επίδοση του δικτύου, δεδομένης της αδυναμίας του να εξάγει χρονικές εξαρτήσεις, αλλά και της ευκολίας εκπαίδευσής του, που αποτελεί απόδειξη του καλού σχεδιασμού της αρχιτεκτονικής InceptionNet-V3 αλλά και της πολλά υποσχόμενης δυνατότητας της χρήσης μεταφοράς μάθησης στο πρόβλημα αναγνώρισης βίαιων συμβάντων.

4.2 3D ΣΝΔ-Δύο Ροών με χρήση διαφορών καρέ

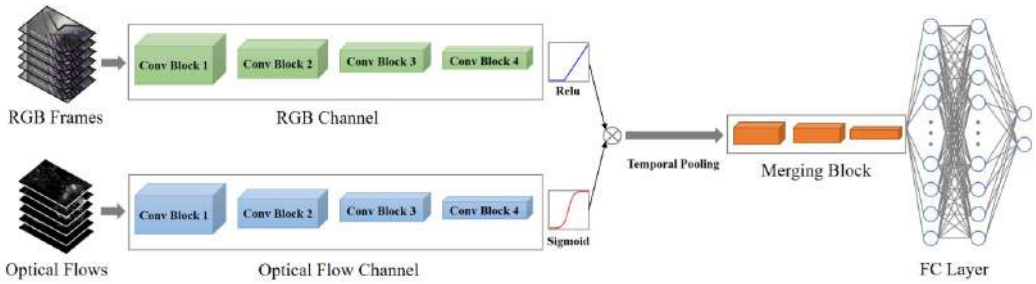
Στην παρακάτω υλοποίηση η προσπάθεια στρέφεται, σε αντίθεση με την πρώτη, στην δυνατότητα εντοπισμού χρονικών εξαρτήσεων απο τα δεδομένα. Η έμπνευση για την προτεινόμενη λύση προέρχεται από την αρχιτεκτονική Flow Gated Network [10] παραλλαγμένη έτσι ώστε να μπορεί να εκπαιδευτεί και να εκτελείται σε συνθήκες πραγματικού κόσμου όπου η δυνατότητες του υλικού είναι περιορισμένες και η χρονική απόκριση του μοντέλου είναι υψίστης σημασίας. Το μοντέλο Flow Gated [10] θα προσαρμοστεί, μειώνοντας το μέγεθός του κατά έναν παράγοντα μεγαλύτερο από 4 και θα χρησιμοποιηθεί μια εναλλακτική στην χρήση οπτικών ροών σημαντικά μικρότερου υπολογιστικού κόστους.

4.2.1 Αρχιτεκτονική μοντέλου

Πριν την εφαρμογή του Flow Gated Network [10], οι μέθοδοι επικεντρωνόντουσαν στην εξαγωγή χωρικών χαρακτηριστικών από μεμονωμένα καρέ και μετά τον συνδυασμό τους για την μοντελοποιήση χρονικών πληροφοριών. Οι εκδότες Ng et al. [91] συνοψίζουν όλες τις μεθόδους για χρονική συγκέντρωση χαρακτηριστικών και οι περισσότερες είναι κατασκευασμένες απο ανθρώπους και δοκιμασμένες έκαστες. Καθώς οι πληροφορίες κίνησης μπορεί να μην είναι χρήσιμες απο έναν τέτοιο χονδροειδές μηχανισμό εξαγωγής χρονικών χαρακτηριστικών (σαν εκείνους που αναφέρονται [91]) το δίκτυο [10] απο το οποίο αντλείται η έμπνευση για την συγκεκριμένη υλοποίηση προσπαθεί να δημιουργήσει "αυτοδίδακτους" μηχανισμούς εξαγωγής χρονικών πληροφοριών.

Η βασική ιδέα πίσω από το από την προτεινόμενη υλοποίηση είναι η χρήση ενός δεύτερου καναλιού το οποίο πληροί τον ρόλο ενός μηχανισμού συσσωμάτωσης των πληροφοριών από τον πρώτο. Συγκεκριμένα, στο δεύτερο κανάλι χρησιμοποιούνται διαφορές καρέ ως μια "πύλη" δίνοντας πληροφορίες στο μοντέλο όσον αφορά ποιές πληροφορίες να συγκρατήσει ή να απορρίψει. Συγκεκριμένα, οι διαφορές μεταξύ των καρέ, που καθιστούν μια υπολογιστικά φθηνή εναλλακτική στην χρήση οπτικών ροών, έχουν τον ρόλο μοντελοποίησης της πληροφορίας κίνησης (ειδικά καθώς οι κάμερες ασφαλείας είναι ακίνητες). Στο κανάλι RGB χρησιμοποείται η συνάρτηση ενεργοποίησης ReLU ενώ στο κανάλι διαφορών καρέ χρησιμοποιείται η σιγμοειδής. Καθώς η έξοδος της σιγμοειδής είναι μεταξύ 0 και 1 είναι ένας παράγοντας κλιμάκωσης της εξόδου του καναλιού RGB. Καθώς στη συνέχεια το max-pooling συγκρατεί τα τοπικά μέγιστα, το σημείο του καναλιού RGB που πολλαπλασιάστηκε με 1 (που αντιπροσωπεύει σημεία έντονης κίνησης) είναι πιο πιθανό να διατηρηθεί ενώ τα σημεία πολλαπλασιασμένα με μηδέν είναι πιο πιθανό να απορριφθούν.

Η δομή του συγκεκριμένου μοντέλου, όπως φαίνεται παρακάτω, αποτελείται από 4 μέρη : το κανάλι RGB, το κανάλι διαφορών καρέ, το τμήμα ένωσης και τα πλήρως συνδεδεμένα επίπεδα. Το κανάλι RGB και το κανάλι διαφορών καρέ αποτελούνται απο σειριακά 3D ΣΝΔ και έχουν συνεπείς δομές έτσι ώστε η έξοδός τους να μπορεί να ενωθεί. Το τμήμα ένωσης αποτελείται επίσης από 3D ΣΝΔ τα οποία επεξεργάζονται τις πληροφορίες μετά από την χρονική συσσωμάτωση πληροφορίων. Τέλος, τα πλήρως συνδεμένα επίπεδα παράγουν έξοδο. Η χρήση οπτικών ροών φαίνεται, εκ πρώτης όψης, να είναι σαφώς καλύτερη μέθοδος για την εκτίμηση κίνησης αντί για την χρήση διαφορών καρέ, καθώς παράγει ένα διανυσματικό πεδίο συνεχούς κίνησης, ακρίβεια ανεξαρτήτως μεμονωμένων pixel και ανθεκτικότητα στις αλλαγές φωτισμού. Ωστόσο, στην συγκεκριμένη εφαρμογή, όπου οι κάμερες είναι ακίνητες και η εκτίμηση κίνησης παίρνει το ρόλο της επισήμανσης μια περιοχής ενδιαφέροντος για το "φιλτράρισμα" των άσχετων περιοχών, σε συνδυασμό με το γεγονός ότι ο υπολογισμός των διαφορών μεταξύ των καρέ είναι μια υπολογιστικά τετριμμένη διαδικασία σε σχέση με την εκτίμηση οπτικών ροών, την καθιστούν μια ελκυστική εναλλακτική.

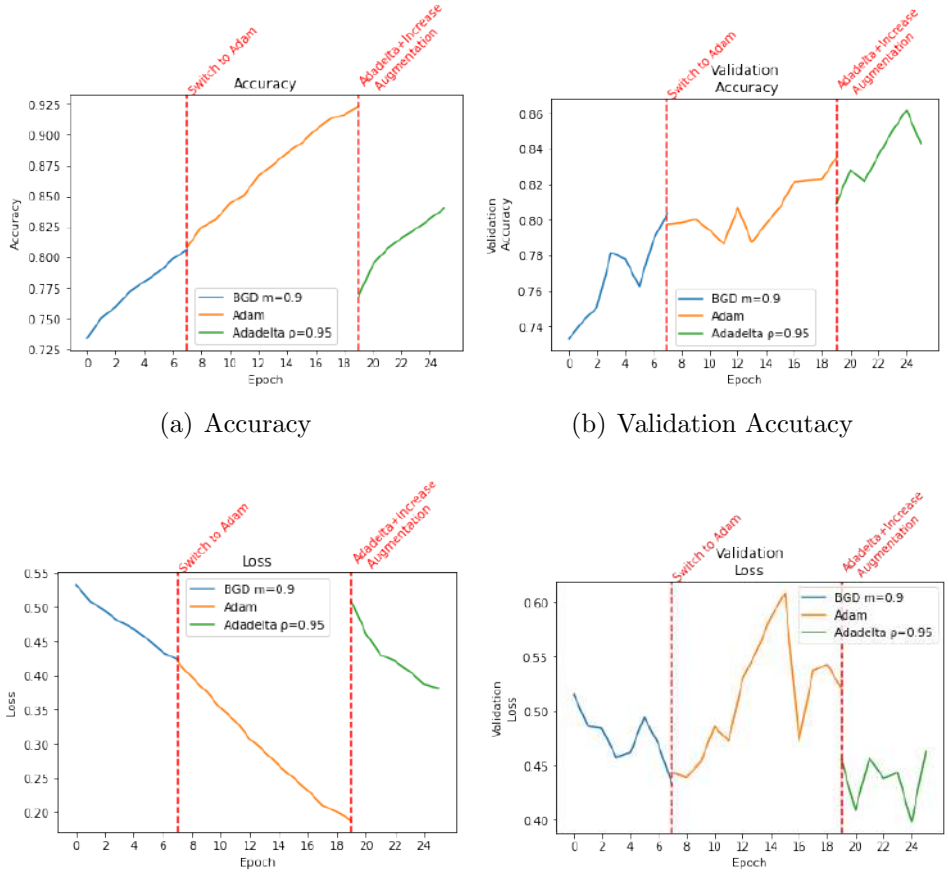


Σχήμα 4.10: **Αρχιτεκτονική Flow Gated Network.** Εικόνα παρμένη από [10]. Η διαφορά με την προτεινόμενη έκδοση του δικτύου είναι η χρήση διαφορών καρέ αντί για οπτική ροή και η μείωση του μεγέθους της εισόδου και των χαρτών χαρακτηριστικών απο 64 καρέ σε 16. (Cheng et al. 2021 [10])

Η χρήση τρισδιάστατων ΣΝΔ έχει αποδειχτεί ισχυρό εργαλείο για την εξαγωγή χωροχρονικών εξαρτήσεων, ωστόσο η εκπαίδευσή τους σε δεδομένα βίντεο δεν είναι τετριμμένη διαδικασία. Το βασικό πρόβλημα με την χρήση τρισδιάστατων ΣΝΔ για την εξαγωγή χωροχρονικών εξαρτήσεων από τα δεδομένα είναι η αυξημένη απαίτηση σε ανάγκες μνήμης αλλά και σημαντική υπολογιστική πολυπλοκότητα που εισάγουν οι τρισδιάστατες συνελίξεις. Μία $3\times3\times3$ συνέλιξη μπορεί να απαιτεί 3 φορές περισσότερους υπολογισμούς από μία 3×3 συνέλιξη αλλά παράγει και 3 φορές μεγαλύτερο χάρτη χαρακτηριστικών και ούτω καθ'εξής. Μια λύση στο παραπάνω πρόβλημα εμφανίζεται στην μορφή των χωριζόμενων συνελίξεων κατά βάθος.

Οι χωριζόμενες συνελίξεις είχαν αρχικά εισαχθεί στο [92] αλλά στη συνέχεια χρησιμοποιήθηκαν και στην αρχιτεκτονική Inception [85] αλλά κατά κύριο λόγο στην αρχιτεκτονική MobileNet [69] για την μείωση του υπολογιστικού κόστους με στόχο την δυνατότητα εκτέλεσης σε κινητές συσκευές. Ουσιαστικά οι συνελίξεις σπάνε σε δύο μέρη μία συνέλιξη κατά βάθος, αποτυπώνοντας τις χωρικές πληροφορίες, και μια σημειακή συνέλιξη $1x1$ για τον συνδυασμό των εξόδων από την συνέλιξη κατά βάθος. Η αρχιτεκτονική Pseudo 3D Residual Network (ResNet) [70] βασίζεται στις ιδέες που αναφέρθηκαν για την βελτίωση των 3D ΣΝΔ αντικαθιστώντας τις $3\times3\times3$ συνελίξεις με δύο συνελίξεις, μια χωρική $1\times3\times3$ και μια χρονική $3\times1\times1$. Αυτό έχει το επιπλέον ενδεχόμενο όφελος ότι μπορούν να χρησιμοποιηθούν προϋπάρχοντα φίλτρα (3×3) απο 2D ΣΝΔ.

Αρχικά προσαρμόζουμε το δίκτυο έτσι ώστε να δέχεται μια αλληλουχία 16 καρέ, μεγέθους $(224, 224, 4)$ όπου τα πρώτα 3 κανάλια είναι τα 3 κανάλια RGB και το τελευταίο είναι η διαφορά μεταξύ του συγκεκριμένου καρέ και του προηγούμενου (με την πρώτη διαφορά καρέ να είναι κενή). Το δίκτυο εκπαιδεύεται αρχικά με Στοχαστική Κατάβαση Κλίσης με momentum=0.9 που εγγυάται την σύγκλιση μέχρι να παρατηρήσουμε ότι το δίκτυο αρχίζει την υπερβολική προσαρμογή, όπου προσαρμόζονται οι μέθοδοι εκπαίδευσης όπως θα αναφερθούν. Χρησιμοποιούνται επίσης μια πληθώρα τεχνικών data augmentation όπως τυχαία στροφή, αλλαγή αξόνων, αλλαγή χρωμάτων, αλλαγή φωτεινότητας, θόρυβος κτλ. Τέλος το μέγεθος παρτίδας παραμένει σταθερό στα 4 λόγω περιορισμού μνήμης του υλικού.

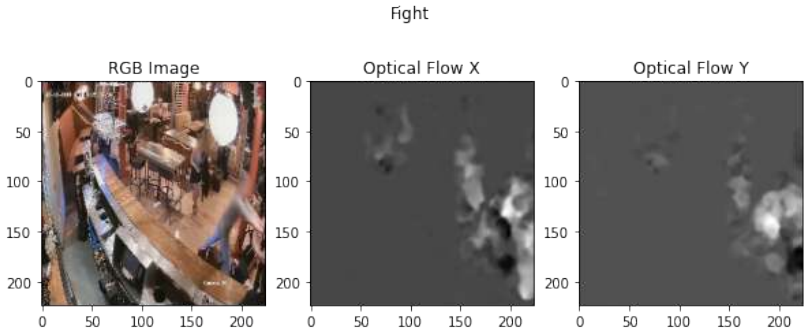


(a) Accuracy (b) Validation Accutacy

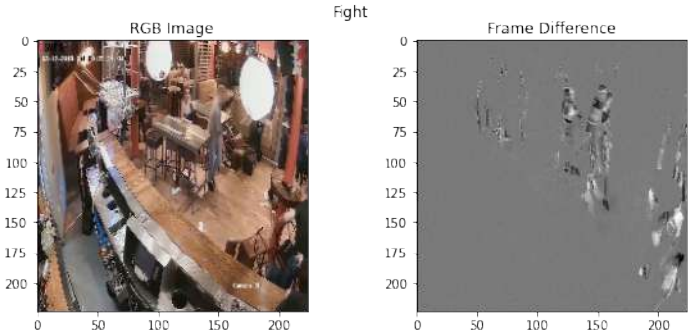
(c) Loss (d) Validation Loss

Σχήμα 4.11: **Ιστορικό εκπαίδευσης.** Στις γραφικές βλέπουμε το ιστορικό εκπαίδευσης σε διάφορες μετρικές αξιολόγησης. Οι κάθετες διακεκομμένες κόκκινες γραμμές υποδηλώνουν τα σημεία που άλλαξε ο αλγόριθμος βελτιστοποίησης και αυξήθηκε η χρήση data augmentation.

Στο Σχήμα 4.11 βλέπουμε το ιστορικό εκπαίδευσης του προτεινόμενου δικτύου. Το δίκτυο εκπαιδεύτηκε σε τρεις ξεχωριστές φάσεις οι οποίες υποδηλώνονται απο τις κάθετες κόκκινες διακεκομμένες γραμμές. Στην αρχή το δίκτυο εκπαιδεύεται με απλή στοχαστική κατάβασης κλίσης με momentum=0.9 έως το δίκτυο αρχίσει να προσαρμόζει υπερβολικά όπου στην συνέχεια ο αλγόριθμος βελτιστοποίησης αλλάζεται στον αλγόριθμο Adam. Η σκεπτική πίσω από αυτήν την επιλογή ήταν η εκμετάλλευση του προσαρμοστικού ρυθμού μάθησης και της βελτιστοποίησης momentum που παρέχει ο αλγόριθμος Adam, χωρίς να χρειάζεται η λεπτή ρύθμιση αυτών των υπερπαραμέτρων χειροκίνητα. Ωστόσο, ο αλγόριθμος Adam παρ'όλο που βελτιώνει την επίδοση σε σχέση με τον προηγούμενο αλγόριθμο, δε παρέχει γρήγορη σύγκλιση και αρχίζει να προσαρμόζει υπερβολικά επίσης, ενδεχομένως λόγω του ότι ο Adam ενδείκνυται για χρήση με μεγαλύτερα σύνολα δεδομένων. Τέλος, για την διαφυγή από το τοπικό ελάχιστο στο οποίο είχε παγιδευτεί το μοντέλο, χρησιμοποιείται ο αλγόριθμος Adadelta, μαζί με εντονότερη χρήση data augmentation προκειμένου να μειωθεί η υπερβολική προσαρμογή (που φαίνεται απο την απότομη μείωση της επίδοσης εκπαίδευσης). Οι αλλαγές παρ'όλης της λογικής που της ακολουθούν, ήταν στα πλαίσια πειραματισμού και θα επιλεγόταν ο αλγόριθμος Adadelta εξ'αρχής αν δεν υπήρχε περιορισμός υλικού.



(a) Optical Flows



(b) Frame Differences

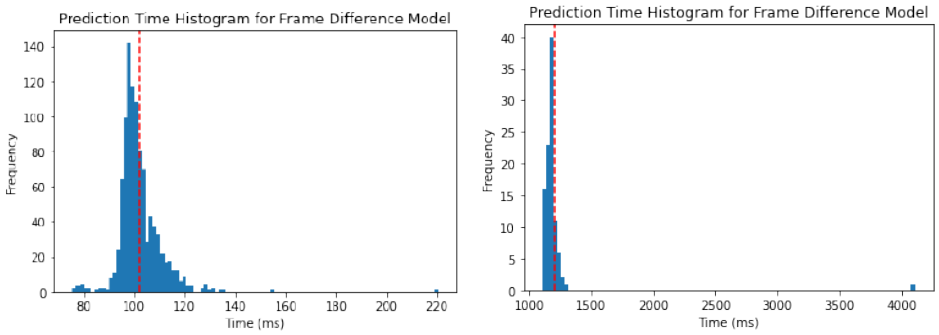
Σχήμα 4.12: **Σύγκριση μεθόδων εκτίμησης κίνησης.** Στην εικόνα (a) βλέπουμε την αρχική εικόνα και οπτική ροή μεταξύ των καρέ και στην εικόνα (b) την αρχική εικόνα και τις διαφορές μεταξύ των καρέ. Στην διαφορά των καρέ το περίγραμμα των αντικειμένων κίνησης είναι πιο καθαρή αλλά υπάρχει θόρυβος στη μορφή μεμονωμένων pixel.

Απ'ότι παρουσιάζεται στον Πίνακα 4.2 το προτεινόμενο μοντέλο πετυχαίνει παρόμοια επίδοση με το αρχικό μοντέλο, με σημαντική υπολογιστική βελτίωση. Στον παρακάτω πίνακα βλέπουμε την χρονική απόκριση του προτεινόμενου μοντέλου σε σχέση με το αρχικό Flow Gated.

		Precision	Recall	F-1	Accuracy
Προτεινόμενο Μοντέλο	**Βία**	86%	85%	85%	**86.2%**
	Μη-Βία	85%	86%	86%	
Αρχικό Flow Gated Μοντέλο	**Βία**	87%	86%	86%	**87.25%**
	Μη-Βία	86%	87%	87%	

Πίνακας 4.2: **Επιδόσεις προτεινόμενου μοντέλου συγκριτικά με αρχικό μοντέλο Flow Gated [10]**

Στο σχήμα 4.13 βλέπουμε την διαφορά στον χρόνο απόκρισης μεταξύ των δύο μοντέλων. Το αρχικό μοντέλο έχει μέση απόκριση πάνω από μια τάξης μεγέθους παραπάνω (1200ms) με μερικά μεμονωμένα δεδομένα αξιολόγησης να φτάνουν στα 4000ms ενώ το προτεινόμενο έχει 100ms. Αύτο σημαίνει ότι το μοντέλο μας μπορεί να κάνει κατά μέσο όρο 12 μετρήσεις για κάθε μία του αρχικού καθιστώντας το πιο ικανό για την χρήση του σε πραγματικές συνθήκες.



(a) Προτεινόμενο Μοντέλο (b) Flow Gated Μοντέλο

Σχήμα 4.13: **Χρονική απόκριση μοντέλων.**

4.2.2 Συμπεράσματα Υλοποίησης

Η αρχιτεκτονική αναγνώρισης ανθρωπίνων δράσεων με δύο ροές, όπου η μία ροή αντιπροσωπεύει την εκτίμηση κίνησης (όπως οπτική ροή ή στην συγκεκριμένη περίπτωση διαφορά καρέ) έχει αποδείξει την δυνατότητα της σε πολλές εφαρμογές [93, 94, 95]. Το βασικό πρόβλημα στην αρχική έκδοση του δικτύου [10] ήταν η μεγάλη απαίτηση σε υπολογιστικούς πόρους, λόγω της δυσκολίας υπολογισμού της οπτικής ροής αλλά και του μεγάλου πλήθους εικόνων εισόδου γεγονός που απαιτούσε ειδικό υλικό με τεράστιο κόστος (που δε δικαιολογείται από την επίδοση του δικτύου). Καταφέραμε να μειώσουμε σε σημαντικό βαθμό την απόκριση του μοντέλου αλλά και την υπολογιστική απαίτηση, χωρίς διαφορά στην επίδοση, κάνοντάς το ικανό για την εφαρμογή του στον πραγματικό κόσμο. Η επίδοση του παρ'όλο του ότι είναι ικανοποιητική, δεν έχει φτάσει ακόμα τα επίπεδα που απαιτούνται για την δημόσια χρήση του. Ενδεχομένως να επιδέχεται βελτίωση με πειραματισμό του μεγέθους παρτίδας καθώς και τεχνικές βελτιστοποίησης για την διαφυγή απο το τοπικό ελάχιστο όπως επίσης και εισαγωγή χαρακτηριστικών από προεκπαιδευμένα 2D ΣΝΔ στα $1 \times 3 \times 3$ φίλτρα από 3×3 φίλτρα άλλων δικτύων.

Κεφάλαιο 5

Συμπεράσματα εργασίας

Η αναγνώριση σκηνών βίας είναι ένας τομέας έρευνας με σημαντική εκδήλωση ενδιαφέροντος από πολυάριθμους δημόσιους φορείς και ερευνητές. Πρόκειται για ένα αντικείμενο με μεγάλη δυνατότητα εφαρμογής σε δημόσιους χώρους που βελτιώνει την δυνατότητα άμεσης ανταπόκρισης, και συνεπώς καταστολής τους, από τις αρμόδιες αρχές, μειώνοντας ενδεχομένως τα συμβάντα βίας και προστατεύοντας τα θύματά της. Αποτελεί ένα παράδειγμα χρήσης των νευρωνικών δικτύων και της μηχανικής μάθησης για το γενικότερο καλό της κοινωνίας.

Πέραν όμως από τα οφέλη που προέρχονται από την άμεση χρήση της τεχνολογίας, αξιοσημείωτη είναι και η έμμεση επίδραση της ανάπτυξης αυτών των τεχνολογιών στον γενικότερο τομέα της αναγνώρισης ανθρωπίνων δράσεων. Η αναγνώρισης σκηνών βίας είναι ένα σύνθετο πρόβλημα το οποίο απαιτεί αναγνώριση πολυεπίπεδων χαρακτηριστικών και αφηρημένων έννοιων από τα δεδομένα. Τέτοιες έννοιες και χαρακτηριστικά για παράδειγμα αποτελούν η ανθρώπινη πρόθεση, επιθετική συμπεριφορά, πανικός στο πλήθος, ανθρώπινη επαφή κτλ. Είναι λογικό λοιπόν, να υποθέσουμε ότι η ανάπτυξη τεχνολογιών που εντοπίζουν ικανοποιητικά τα παραπάνω έχουν εφαρμογή και σε άλλους τομείς επιστημονικής έρευνας.

Στην παρούσα εργασία παρουσιάσαμε δύο διαφορετικές μεθόδους, με διαφορετικό βαθμό επιτυχίας, με αναγνώριση μόνο χωρικών χαρακτηριστικών και χωροχρονικών εξαρτήσεων αντίστοιχα. Στην πρώτη μέθοδο κάναμε χρήση 2D ΣΝΔ και μεταφορά μάθησης από το δίκτυο InceptionNet-V3 προεκπαιδευμένο στο ImageNet [36] πετυχαίνοντας επίδοση 76%. Παρά της μη ικανοποιητικής επίδοσής αύτης της μεθόδου, η επίδοση παραμένει εντυπωσιακή δεδομένης της ευκολίας εκπαίδευσης και του γεγονότος ότι το δίκτυο εξάγει μόνο χωρικές πληροφορίες και δεν είναι ικανό να εντοπίσει χωροχρονικές σχέσεις, όπου ακόμη και ένας ανθρώπινος παρατηρητής θα εντόπιζε δυσκολία στην ταξινόμηση καρέ προς καρέ. Στην δεύτερη μέθοδο κάναμε χρήση διαφορών μεταξύ συνεχόμενων καρέ ως μια μέθοδο εκτίμησης της ανθρώπινης κίνησης η οποία χρησιμοποιήθηκε από το δίκτυο ως μια "πύλη" συγκράτησης των σχετικών πληροφοριών απο τις εικόνες. Το δίκτυο πετυχαίνει επίδοση 86.2% που είναι ικανοποιητική δεδομένης της ευκολίας εκπαίδευσης και ταχύτητάς του που ευθύνεται κατα κύριο λόγο στη χρήση "Pseudo-Residual" 3D ΣΝΔ εμπνευσμένα από το ResNet[70].

5.1 Μελλοντικές επεκτάσεις

Υπάρχουν αρκετές μελλοντικές κατευθύνσεις στις οποίες μπορεί να στραφεί η προσπάθεια στον συγκεκριμένο κλάδο έρευνας. Μια πολλά υποσχόμενη έννοια η οποία έχει αρχίσει να κερδίζει έδαφος σε διάφορους τομείς της μηχανικής μάθησης είναι η έννοια της προσοχής (attention). Ενδεχομένως μια κατεύθυνση θα μπορούσε να ήταν ένας μηχανισμός πολυμετωπικής προσοχής ο οποίος προσπαθούσε να μιμηθεί τους μηχανισμούς που χρησιμοποιούμε εμείς οι

άνθρωποι για να αναγνωρίσουμε βίαια συμβάντα. Μια ιδέα θα ήταν να χωριστούν σε ξεχωριστά εννοιολογικά μέρη τα μέτωπα της προσοχής, το καθένα εκπαιδευμένο σε διαφορετικά σύνολα δεδομένων ανάλογα της έννοιας που προσπαθούν να συλλάβουν. Για παράδειγμα, ένα μέτωπο θα μπορούσε να αφορά την επιθετική βίαια συμπεριφορά χρησιμοποιώντας ταυτόχρονα μηχανισμούς ανίχνευσης σκελετού και να είχε εκπαιδευτεί σε ένα τεράστιο σύνολο δεδομένων αγώνων σε μαχητικά αθλήματα (μαθαίνοντας να διαχωρίζει μεταξύ γροθιάς, λακτίσματος, πάλης κτλ.), ένα άλλο μέτωπο θα μπορούσε να εστιάζει στον τραυματισμό ανθρώπων εκπαιδευμένο σε σύνολα δεδομένων πτώσεων ανθρώπων και ένα τρίτο σε πανικό πλήθους στρέφοντας τους άλλους δύο μηχανισμούς προσοχής προς το ενδεχόμενο σημείο ενδιαφέροντος ανάλογως με το που κοιτάει το πλήθος κτλ. Οι πιθανότητες είναι απεριόριστες, αλλά σίγουρα απαιτείται η ανάπτυξη μεγαλύτερων συνόλων δεδομένων ή σύνθετος συνδυασμός άλλων.

Σύνδεσμος κώδικα εργασίας

Github Repository: https://github.com/jugeekuz/Violence-Detection-Thesis

Βιβλιογραφία

[1] "Nypd response time trends," https://www.nyc.gov/site/911reporting/reports/response-time-trends.page?kbid=148048, accessed: 2023-01-23.

[2] W. H. Foege, M. L. Rosenberg, and J. A. Mercy, "Public health and violence prevention," *Current Issues in Public Health*, vol. 1, pp. 2–9, 1995.

[3] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14*. Springer, 2011, pp. 332–339.

[4] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2012, pp. 1–6.

[5] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image and vision computing*, vol. 48, pp. 37–41, 2016.

[6] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast violence detection in video," in *2014 international conference on computer vision theory and applications (VISAPP)*, vol. 2. IEEE, 2014, pp. 478–485.

[7] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, "Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation," *IEEE transactions on information forensics and security*, vol. 12, no. 12, pp. 2945–2956, 2017.

[8] A. Mumtaz, A. B. Sargano, and Z. Habib, "Violence detection in surveillance videos with deep network using transfer learning," in *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*. IEEE, 2018, pp. 558–563.

[9] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, and M. Farazi, "Efficient two-stream network for violence detection using separable convolutional lstm," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[10] M. Cheng, K. Cai, and M. Li, "Rwf-2000: an open large scale video database for violence detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4183–4190.

[11] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/215a71a12769b056c3c32e7299f1c5ed-Paper.pdf

[12] T. Kavzoglu and P. M. Mather, "The use of backpropagating artificial neural networks in land cover classification," *International journal of remote sensing*, vol. 24, no. 23, pp. 4907–4938, 2003.

[13] A. K. Jain and B. Chandrasekaran, "39 dimensionality and sample size considerations in pattern recognition practice," *Handbook of statistics*, vol. 2, pp. 835–855, 1982.

[14] E. Baum and D. Haussler, "What size net gives valid generalization?" *Advances in neural information processing systems*, vol. 1, 1988.

[15] S. Haykin, *Neural networks and learning machines, 3/E.* Pearson Education India, 2009.

[16] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[17] S. Al, "Some studies in machine learning using the game of checkers," *IBM journal on Research and Development*, vol. 3, pp. 210–229, 1959.

[18] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[19] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27.* Springer, 2005, pp. 345–359.

[20] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *Towards Data Sci*, vol. 6, no. 12, pp. 310–316, 2017.

[21] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, "Dying relu and initialization: Theory and numerical examples," *arXiv preprint arXiv:1903.06733*, 2019.

[22] Y. Ren, P. Zhao, Y. Sheng, D. Yao, and Z. Xu, "Robust softmax regression for multi-class classification with self-paced learning," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 2641–2647.

[23] K. Liano, "Robust error measure for supervised neural network learning with outliers," *IEEE Transactions on Neural Networks*, vol. 7, no. 1, pp. 246–250, 1996.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] F. Pesapane, M. Codari, and F. Sardanelli, "Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine," *European radiology experimental*, vol. 2, pp. 1–10, 2018.

[26] Y. Chauvin and D. E. Rumelhart, *Backpropagation: theory, architectures, and applications.* Psychology press, 2013.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[28] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[29] R. S. Sutton, "Two problems with backpropagation and other steepest-descent learning procedures for networks," in *Proc. of Eightth Annual Conference of the Cognitive Science Society*, 1986, pp. 823–831.

[30] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research*, vol. 12, no. 7, 2011.

[31] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[34] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[35] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition." in *Interspeech*, vol. 2013, 2013, pp. 1173–5.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[37] O. S. Kayhan and J. C. v. Gemert, "On translation invariance in cnns: Convolutional layers can exploit absolute spatial location," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 274–14 285.

[38] A. Hassan and A. Mahmood, "Convolutional recurrent deep learning model for sentence classification," *IEEE Access*, vol. 6, pp. 13 949–13 957, 2018.

[39] S. Malladi and I. Sharapov, "Fastnorm: improving numerical stability of deep network training with efficient normalization," 2018.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[42] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[43] M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through ffts," *arXiv preprint arXiv:1312.5851*, 2013.

[44] R. Rigamonti, A. Sironi, V. Lepetit, and P. Fua, "Learning separable filters," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2754–2761.

[45] B. Graham, "Sparse 3d convolutional neural networks," *arXiv preprint arXiv:1505.02890*, 2015.

[46] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, pp. 64–67, 2001.

[47] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[48] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 855–868, 2008.

[49] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[50] R. C. Staudemeyer and E. R. Morris, "Understanding lstm–a tutorial into long short-term memory recurrent neural networks," *arXiv preprint arXiv:1909.09586*, 2019.

[51] M. C. Mozer, "Induction of multiscale temporal structure," *Advances in neural information processing systems*, vol. 4, 1991.

[52] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen netzen," *Diploma, Technische Universität München*, vol. 91, no. 1, 1991.

[53] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[54] K. Yao, T. Cohn, K. Vylomova, K. Duh, and C. Dyer, "Depth-gated recurrent neural networks," *arXiv preprint arXiv:1508.03790*, vol. 9, p. 98, 2015.

[55] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 3. IEEE, 2000, pp. 189–194.

[56] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[57] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[58] I. Liu, B. Ramakrishnan *et al.*, "Bach in 2014: Music composition with recurrent neural network," *arXiv preprint arXiv:1412.3191*, 2014.

[59] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[61] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," *Advances in neural information processing systems*, vol. 30, 2017.

[62] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[63] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," *arXiv preprint arXiv:2204.07756*, 2022.

[64] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.

[65] E. B. Nievas, O. D. Suarez, G. B. Garcia, and R. Sukthankar, "Movies fight detection dataset," in *Computer Analysis of Images and Patterns*. Springer, 2011, pp. 332–339. [Online]. Available: http://visilab.etsii.uclm.es/personas/oscar/FightDetection/

[66] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

[67] M. Cheng, K. Cai, and M. Li, "Rwf-2000: An open large scale video database for violence detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4183–4190.

[68] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 2003, pp. 363–370.

[69] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[70] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[71] Y. Su, G. Lin, J. Zhu, and Q. Wu, "Human interaction learning on 3d skeleton point clouds for video violence recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 74–90.

[72] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

[73] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[74] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[75] R. Hachiuma, F. Sato, and T. Sekii, "Unified keypoint-based action recognition framework via structured keypoint pooling," *arXiv preprint arXiv:2303.15270*, 2023.

[76] A. Pfeuffer and K. Dietmayer, "Separable convolutional lstms for faster video segmentation," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1072–1078.

[77] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[78] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[79] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[80] H. Mohammadi and E. Nazerfard, "Video violence recognition and localization using a semi-supervised hard attention model," *Expert Systems with Applications*, vol. 212, p. 118791, 2023.

[81] A. Manchin, E. Abbasnejad, and A. Van Den Hengel, "Reinforcement learning with attention that works: A self-supervised approach," in *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part V 26*. Springer, 2019, pp. 223–230.

[82] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the kinetics-700-2020 human action dataset," *arXiv preprint arXiv:2010.10864*, 2020.

[83] F. Chollet *et al.*, "Keras: Deep learning library for theano and tensorflow," *URL: https://keras. io/k*, vol. 7, no. 8, p. T1, 2015.

[84] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[85] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[86] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[87] A. Bhandari, "Image augmentation on the fly using keras imagedatagenerator," *Analytics Vidya*, 2020.

[88] G. Vrbančič and V. Podgorelec, "Transfer learning with adaptive fine-tuning," *IEEE Access*, vol. 8, pp. 196 197–196 211, 2020.

[89] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "Spottune: transfer learning through adaptive fine-tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4805–4814.

[90] S. H. Haji and A. M. Abdulazeez, "Comparison of optimization techniques based on gradient descent algorithm: A review," *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 18, no. 4, pp. 2715–2743, 2021.

[91] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[92] L. Sifre and S. Mallat, "Rigid-motion scattering for texture classification," *arXiv preprint arXiv:1403.1687*, 2014.

[93] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.

[94] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[95] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.