Εθνικο Μετσοβιο Πολυτεχνειο
ΔΠΜΣ Επιστημη Δεδομενων και Μηχανικη Μαθηση
Εργαστηριο Τηλεπισκοπησης

Ταξινόμηση Δράσεων σε Βίντεο Προ-κλινικών Πειραμάτων με Βαθιές Αρχιτεκτονικές Αυτοκωδικοποιητών και 3Δ Συνελίξεων

Μεταπτυχιακη Εργασια

Αλεξανδρος Βυθουλκας

Αθήνα, Ιανουάριος 2023

National Technical University of Athens
IPSP Data Science and Machine Learning
Remote Sensing Laboratory

Action Recognition in Pre-clinical Experiment Videos with Deep Autoencoder Architectures and 3D Convolutions

Master Thesis

Alexandros Vythoulkas

Athens, January 2023

Εθνικο Μετσοβιο Πολυτεχνειο

ΔΠΜΣ Επιστημη Δεδομενων και Μηχανικη Μαθηση

Εργαστηριο Τηλεπισκοπησης

Ταξινόμηση Δράσεων σε Βίντεο Προ-κλινικών Πειραμάτων με Βαθιές Αρχιτεκτονικές Αυτοκωδικοποιητών και 3Δ Συνελίξεων

Μεταπτυχιακη Εργασια

Αλεξανδρος Βυθουλκας

Επιβλέπων: Κωνσταντίνος Καράντζαλος
Καθηγητής Ε.Μ.Π.

Εκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Ιανουαρίου 2023.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
Κωνσταντίνος Καράντζαλος Καθηγητής Ε.Μ.Π.	Maria Vakalopoulou Assistant Professor CentraleSupelec	Αθανάσιος Βουλόδημος Επίκουρος Καθηγητής ΕΜΠ

Αθήνα, Ιανουάριος 2023

Υπεύθυνη Δήλωση
Βεβαιώνω ότι είμαι συγγραφέας αυτής της μεταπτυχιακής εργασίας, και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην πτυχιακή εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης, βεβαιώνω ότι αυτή η πτυχιακή εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τις απαιτήσεις του προγράμματος σπουδών του ΔΠΜΣ *Επιστήμη Δεδομένων και Μηχανική Μάθηση* του Εθνικού Μετσόβιου Πολυτεχνείου.

(Υπογραφή)

Αλέξανδρος Βυθούλκας
Διπλωματούχος Αγρονόμος Και Τοπογράφος Μηχανικός Ε.Μ.Π.

Περίληψη

Στην παρούσα διπλωματική, σκοπός είναι η αυτοματοποίηση των παρατηρήσεων της συμπεριφοράς των επιμύων, κατά την δοκιμασία εξαναγκασμένης κολύμβησης. Το πείραμα της εξαναγκασμένης κολύμβησης αποτελεί ένα σύνηθες μέσο για τη μελέτη της επίδρασης αντικαταθλιπτικών φαρμάκων. Κατά την διάρκεια του πειράματος το υποκείμενο παρουσιάζει διαφορετικές συμπεριφορές οι οποίες αποτελούν αντικείμενο ενδιαφέροντος για τους παρατηρητές, καθώς με τη μέτρηση τους πραγματοποιείται μελέτη των επιδράσεων αντικαταθλιπτικών ουσιών. Για την λύση του προβλήματος αυτού αξιοποιήθηκαν τεχνικές ταξινόμησης δράσεων σε βίντεο.

Το σύνολο δεδομένων παραχωρήθηκε από εργαστήριο ιατροφαρμακευτικής της Ιατρικής Σχολής του ΕΚΠΑ και περιλαμβάνει περίπου 8 ώρες ταξινομημένων βίντεο από 2 διαφορετικούς παρατηρητές, καθώς και εκατοντάδες ακόμη ώρες οι οποίες δεν έχουν ταξινομηθεί. Έπειτα από διόρθωση και επεξεργασία του συνόλου δεδομένων, υλοποιήθηκαν μοντέλα εκτίμησης της συμπεριφοράς από την ανάλυση δεδομένων των βίντεο. Η μεθοδολογία που ακολουθήθηκε είναι ο διαχωρισμός των βίντεο σε μικρά τμήματα, τα οποία στη συνέχεια μπορούν να ταξινομηθούν ως ανεξάρτητα βίντεο. Αρχικά πραγματοποιήθηκε εκπαίδευση μοντέλων με τη χρήση νευρωνικών δικτύων με τεχνολογίες αιχμής στην ταξινόμηση βίντεο, το οποίο είναι το Resnet 2+1D και η αρχιτεκτινοική MViT με χρήση προεκπαιδευμένων βαρών. Οι εκπαιδεύσεις αυτές έδωσαν ευστοχία περίπου 81% και 83% αντίστοιχα στο υποσύνολο επαλήθευσης, έπειτα από προσαρμογή των κατάλληλων υπερπαραμέτρων. Τα μοντέλα αυτά αξιοποιήθηκαν ως μέσα σύγκρισης των επόμενων πειραμάτων.

Για την αξιοποίηση του τεράστιου όγκου μη ταξινομημένων δεδομένων, έγινε προσπάθεια να αξιοποιηθούν με την εκπαίδευση ενός αυτοκωδικοποιητή (autoencoder). Έτσι επιτεύχθηκε η μείωση των διαστάσεων του βίντεο που αποτελεί τα δεδομένα εισόδου των δικτύων. Η κωδικοποίηση αυτή χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου με μειωμένες πλέον διαστάσεις, γεγονός που απλοποίησε ιδιαίτερα την ταξινόμηση και μείωσε κατά πολύ τον χρόνο εκπαίδευσης και πρόβλεψης του μοντέλου. Πραγματοποιήθηκαν δοκιμές με διαφορετικούς αυτοκωδικοποιητές, έτσι ώστε να ελεγχθεί η διαδικασία και να υπάρξει κατανόηση των αποτελεσμάτων.

Τα αποτελέσματα έδειξαν ότι η μείωση των διαστάσεων του βίντεο, τόσο στα κανάλια όσο και στον χρόνο και τον χώρο, αποτύπωσε ικανοποιητικά το βίντεο χωρίς να υπάρχει έλλειψη σημαντικής πληροφορίας. Οι ταξινομήσεις των διανυσμάτων μειωμένης διάστασης, έφεραν αποτελέσματα έως και 73% γεγονός που δεν τα καθιστά κοντά στην ευστοχία των δικτύων με προεκπαιδευμένα βάρη, καθώς υστέρησαν κατά 10% στην ευστοχία.

Ως συμπέρασμα προκύπτει ότι οι αρχιτεκτονικές για την αναγνώριση βίντεο έχουν εξελιχθεί δραματικά τα τελευταία χρόνια, και η χρήση προεκπαιδευμένων βαρών προσδίδει σημαντική γνώση στην ταξινόμηση βίντεο διαφορετικού είδους, όπως το πείραμα εξαναγκασμένης κολύμβησης.

Λέξεις κλειδιά: ταξινόμηση βίντεο, αυτοκωδικοποιητής, εξαναγκασμένη κολύμβηση, βίντεο, ταξινόμηση, βαθιά μάθηση, ταξινόμηση συμπεριφοράς, αναγνώριση συμπεριφοράς, νευροεπιστήμη, φαρμακολογία

Abstract

The purpose of this master thesis, is to automate the observations of rat behavior during forced swim test. The experiment of force swim test is a common way of analysing the effect of antidepressant drugs. During forced swim test, the subject has different behaviors which are significant for the observers in order to analyze antidepressant. To address this problem, video classification methods were used.

The dataset was given by a medical laboratory and contains 8 hours of labeled videos from 2 different observers. It also contains hundreds of hours which are unlabeled. After adjusting and preprocessing the dataset appropriately, action recognition models were designed. The methodology that was used was to cut small pieces of videos and classify them as independent videos, by video classification deep learning networks. As first, the use of a state-of-the-art deep learning networks for video classification was used. These were Resnet 2+1D and MViTv2 with pretrained weights. This training resulted in 81% and 83% accuracy of the validation subset respectively, after adjusting the right hyperparameters. These models were used as baselines to the rest of the experiments.

To exploit the massive unlabeled data, an autoencoder was trained. This way the dimensionality of the input of the network, namely a video, was reduced. The encoding was used to classify the videos used. This yielded in reducing the number of the networks parameters by simplifying the architecture. Moreover, the time needed to train and predict the classification of the FST experiment was reduced by far.

Different autoencoders for the encoded data were trained to test and understand the process.

The experiments emerged that reducing video dimenstionality, in channels, time and space, could represent the videos decently, without a lot of detail deficiency. The classification of those encodings resulted in 73% accuracy, so they underperformed compared to the state of the art architectures with pretrained weights.

Action recognition architectures have been improved in the past years and the use of pretrained weights adds a lot of value to solving different domains like deep forced swim test.

Reducing the dimensionality of the video by training and using the autoencoder by 1/3 of the initial dimensions on video channels, time and space, resulted in a decent representation of the video without sacrificing useful information. The classification of the encoded data resulted in 78% accuracy of the validation subset, which is close to the baseline of the experiment. The procedure was considered as an interesting alternative to Resnet 2+1D because it was by far faster and simplified the problem of automating forced swim test.

Keywords: video classification, autoencoder, forced swim test, deep learning, machine learning, action classification, behavior recognition, neuroscience, medical

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Καράντζαλο για την επίβλεψη αυτής της εργασίας και για την ευκαιρία που μου έδωσε να εκπονήσω το θέμα της παρούσας διπλωματικής. Ακόμη θα ήθελα να ευχαριστήσω ιδιαίτερα τον προπτυχιακό φοιτητή Τηλέμαχο Μουρμούρη για την εξαιρετική συνεργασία που είχαμε κατά την συμπόρευση μας στο πρόβλημα της ταξινόμησης των βίντεο. Ιδιαίτερα ευχαριστώ την Αθανασία Τρανού για την ψυχολογική υποστήριξη και βοήθεια που μου έδωσε, την Ιωάννα Αχμέτη για τη βοήθεια στις διορθώσεις του τεύχους, καθώς και την οικογένεια και τους φίλους μου για την στήριξη και συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Η βαθιά μάθηση είναι ένα πεδίο της μηχανικής μάθησης το οποίο έχει γνωρίσει ταχύτατη εκπαίδευση τα τελευταία χρόνια. Περιλαμβάνει την χρήση τεχνητών νευρωνικών δικτύων τα οποία αποτελούν μοντελοποίηση των βιολογικών νευρώνων του ανθρώπινου οργανισμού, για την μάθηση μοτίβων και την λήψη αποφάσεων βασισμένα σε μεγάλο όγκο δεδομένων. Με την ύπαρξη τεράστιου όγκου δεδομένων καθώς και την αύξηση της υπολογιστικής ισχύς, η βαθιά μάθηση έγινε το κλειδί σε πολλούς τομείς όπως η ταξινόμηση εικόνων και βίντεο.

Η ιστορία της βαθιάς μάθησης ξεκινάει την δεκαετία του 1940, όπου ο Warren McCulloch και Walter Pitts πρότειναν την ιδέα ενός νευρωνικού δικτύου. Ωστόσο, μόλις στις αρχές της δεκαετίας του 2000, όπου δόθηκε προσοχή στα νευρωνικά δίκτυα χάρη στην ανάπτυξη των ικανοτήτων του hardware και του software. Το 2012, τεχνικές βαθιάς μάθησης χρησιμοποιήθηκαν για την κατάκτηση του καλύτερου μοντέλου στον διαγωνισμό του ImageNet, το οποίο κίνησε το ενδιαφέρον των ερευνητών στον τομέα αυτό.

Έκτοτε, η τεχνολογία της βαθιάς μάθησης συνεχίζει να εξελίσσεται και να βελτιώνεται, γεγονός που οδηγεί στην πολυάριθμη εφαρμογή τέτοιων αλγορίθμων σε διαφορετικούς τομείς παραγωγής. Στον τομέα του video classification, οι αλγόριθμοι βαθιάς μάθησης έχουν χρησιμοποιηθεί για την ταξινόμηση και κατηγοριοποίηση βίντεο μέσω του περιεχομένου τους, επιτρέποντας την αποτελεσματική και ακριβή αναζήτηση και ανάλυση των βίντεο. Λόγω του αμέτρητου όγκου δεδομένων βίντεο που υπάρχει και και παράγεται στην σημερινή εποχή, η χρήση νευρωνικών δικτύων για την κατανόηση, ανάλυση και αξιοποίηση τους είναι ιδιαίτερα σημαντική.

1.1 Αντικείμενο Διπλωματικής

Η διπλωματική αυτή πραγματεύεται την αυτοματοποίηση των παρατηρήσεων που είναι αναγκαίες για το πείραμα της εξαναγκασμένης κολύμβησης που πραγματοποιείται σε επιμύες. Το πρόβλημα αυτό είναι ιδιαίτερα σημαντικό, καθώς συμβατικά πραγματοποιείται με ανθρώπινη εργασία η οποία είναι πολύ χρονοβόρα και κουραστική. Αυτό συμβαίνει γιατί ο παρατηρητής για να ταξινομήσει το πείραμα, είναι αναγκαίο να βλέπει το βίντεο σε πραγματική ταχύτητα, και με προσεκτική παρατήρηση τον κινήσεων του επιμύ να επιλέγει ανά πάσα στιγμή την σωστή κατηγορία που προσδιορίζει την συμπεριφορά του υποκειμένου. Επιπλέον, η διαδικασία αυτή οδηγεί σε καθυστερήσεις και ασυνέπειες από τον παρατηρητή λόγω της φύσης της. Έτσι λοιπόν, η σημασία της αυτοματοποίησης της γίνεται εύκολα κατανοητή.

Εκτός των παραπάνω, πρόκειται για ένα ενδιαφέρον θέμα το οποίο μπορεί να αντιμετωπιστεί με την ταξινόμηση βίντεο, όπως και συμβαίνει στην παρούσα διπλωματική. Έτσι γίνεται δυνατή η κατανόηση και εξέταση τεχνολογιών αιχμής για την ταξινόμηση βίντεο οι οποίες παρουσιάζουν ιδιαίτερο ενδιαφέρον, καθώς αποτελεί ένα πρόβλημα για το οποίο δεν υπάρχει τόση γνώση όση στην ταξινόμηση εικόνων. Επίσης, η ταξινόμηση των βίντεο είναι πιο απαιτητική λόγω της προσθήκης της διάστασης του χρόνου που αυξάνει την περιπλοκότητα των δικτύων καθώς και τον αριθμό τον παραμέτρων εκθετικά. Οι συνήθεις πρακτικές που χρησιμοποιούνται για την εξαγωγή

χαρακτηριστικών από βίντεο, είναι νευρωνικά δίκτυα βαθιάς μάθησης με LSTM ή 3D συνελικτικούς νευρώνες.

Το θέμα αυτό έχει εξεταστεί από εμένα και στην προπτυχιακή μου διπλωματική εργασία, όπου εξετάστηκαν συμβατικές μέθοδοι όρασης υπολογιστών για την εξαγωγή και ταξινόμηση ακολουθιών βίντεο, καθώς και deep learning μοντέλων που αποτελούσαν της τεχνολογία αιχμής της συγκεκριμένης εποχής στο πρόβλημα της αναγνώρισης δράσης. Η νέα ιδέα η οποία αποτέλεσε και αφορμή για την εκπόνηση της διπλωματικής αυτής στο ίδιο πρόβλημα είναι η αξιοποίηση των εκατοντάδων αταξινόμητων δεδομένων που είχαμε στη διάθεση μας, μέσω μοντέλων αυτοκωδικοποιητών (autoencoder) για την ενδεχόμενη βελτίωση των αποτελεσμάτων. Η προϋπάρχουσα γνώση πάνω στο πρόβλημα αποτέλεσε καθοριστικό παράγοντα καθώς θεωρήθηκε ότι θα βοηθήσει στην κατανόηση και την μελέτη των autoencoders από τον συγγραφέα.

Αποφασίστηκε ακόμη να υπάρξει νέα υλοποίηση του αναγκαίου κώδικα για όλες τις ανάγκες που προέκυψαν κατά την εκπόνηση της διπλωματικής. Ο λόγος είναι ότι υπήρχε η επιθυμία για πειραματισμό με διαφορετικές τεχνολογίες και τον σχεδιασμό μια καλύτερης αρχιτεκτονικής λογισμικού με δυνατότητα επαναχρησιμοποίησης και εύκολης συντήρησης και κατανόησης. Την βασικότερη αλλαγή στο λογισμικό αποτέλεσε η επιλογή του framework 'pytorch', σε σχέση με το 'keras' που χρησιμοποιήθηκε στην προπτυχιακή εργασία μου. Το σύνολο του κώδικα που χρησιμοποιήθηκε υπάρχει διαθέσιμο στην τοποθεσία github.com/alexVyth/video-classification-trainer.

1.2 Συνεισφορά διπλωματικής

Από την πραγμάτευση του θέματος της διπλωματικής, όπως περιγράφτηκε στην ενότητα 1.1, η παρούσα διπλωματική θα συνεισφέρει με τα εξής στοιχεία:

- Εφαρμογή και προσαρμογή των αρχιτεκτονικών που αποτελούν την τεχνολογίας αιχμής στην αναγνώριση και ταξινόμηση βίντεο για το 2022, στο πείραμα εξαναγκασμένης κολύμβησης.
- Εξέταση της κωδικοποίησης και αποκωδικοποίησης των βίντεο με τη χρήση αυτοκωδικοποιητών.
- Μελέτη της εκπαίδευσης του πειράματος εξαναγκασμένης κολύμβησης με τη χρήση συμπιεσμένων κωδικοποιήσεων από αυτοκωδικοποιητές.
- Δημιουργία τεχνολογίας αιχμής για την ταξινόμηση του πειράματος εξαναγκασμένης κολύμβησης.

1.3 Οργάνωση Κειμένου

Η διπλωματική οργανώνεται ως εξής:

- Το δεύτερο κεφάλαιο αφορά την βιβλιογραφία γύρω από την ταξινόμηση βίντεο και ιδιαίτερα με χρήση autoencoders σε ακολουθίες βίντεο.
- Στο τρίτο κεφάλαιο παρουσιάζεται το σύνολο δεδομένων που χρησιμοποιήθηκε και η μεθοδολογία που ακολουθήθηκε για την πραγματοποίηση των πειραμάτων που διεξήχθησαν.
- Στο τέταρτο κεφάλαιο γίνεται παρουσίαση και ανάλυση των αποτελεσμάτων που προέκυψαν από τα πειράματα.

- Τέλος, στο πέμπτο κεφάλαιο γίνεται ανάλυση των συμπερασμάτων, καθώς και σκέψεις για μελλοντικές επεκτάσεις γύρω απ' το συγκεκριμένο πρόβλημα.

Κεφάλαιο 2

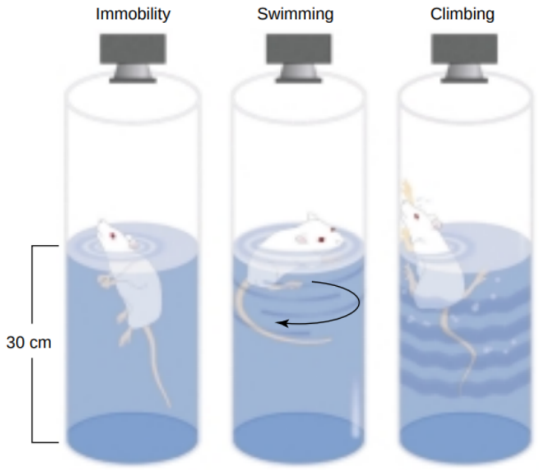
Θεωρητικό Υπόβαθρο και Επισκόπηση Βιβλιογραφίας

2.1 Πείραμα Εξαναγκασμένης Κολύμβησης

Η μελέτη της κατάθλιψης και της θεραπείας αυτής, αποτελεί ένα δύσκολο πρόβλημα καθώς η εύρεση μοντέλων που να περιγράφουν την καταθλιπτική συμπεριφορά είναι ένα σύνθετο πρόβλημα. Μία προσέγγιση για την κατασκευή ενός τέτοιου μοντέλου είναι η δοκιμασία εξαναγκασμένης κολύμβησης και χρησιμοποιείται συχνά για την μελέτη αντικαταθλιπτικών ουσιών.

2.1.1 Περιγραφή του Πειράματος

Το πείραμα εξαναγκασμένης κολύμβησης ορίστηκε απ' τον Porsolt [1] το 1977. Περιλαμβάνει την τοποθέτηση ενός μυ [2] ή επίμυ [3] σε κύλινδρο από πλεξιγκλάς, με νερό ύψους $15cm$ στους $25°C$ απ' την οποία είναι αδύνατο να εξέλθει. Συνηθίζεται οι επιμύες να τοποθετούνται σε πρώτη φάση για 15 λεπτά στον κύλινδρο, ώστε να εξοικειωθούν με τη διαδικασία. Παρατηρείται ότι προσπαθούν για περίπου 2–3 λεπτά να εξέλθουν απ' τον κύλινδρο, και για τα υπόλοιπα λεπτά στέκονται ακίνητα, κάνοντας τις ελάχιστες κινήσεις για να διατηρούν το κεφάλι τους πάνω απ' το νερό. Ο Porsolt ερμήνευσε αυτή την συμπεριφορά ως απελπισία των επιμυών και εγκατάλειψη της προσπάθειας που καταβάλουν να εξέλθουν απ' τον κύλινδρο, λόγω της κατανόησης του γεγονότος ότι δεν είναι δυνατόν να εξέλθουν. Τα ίδια ποντίκια τοποθετήθηκαν για δεύτερη φορά στον κύλινδρο, μια μέρα μετά. Στο σημείο αυτό έγινε η παρατήρηση ότι αν έχουν χορηγηθεί συγκεκριμένες αντικαταθλιπτικές ουσίες, μειώνεται αρκετά ο χρόνος της ακινησίας. Έτσι με την παρατήρηση της συμπεριφοράς του επιμύ, προσδιορίστηκε ένα μοντέλο της επίδρασης και ενδεχομένως της επιτυχίας των αντικαταθλιπτικών φαρμάκων.



Σχήμα 2.1: Αναπαράσταση των 3 σημαντικότερων κατηγοριών του πειράματος εξαναγκασμένης κολύμβησης.

Οι κατηγορίες ενδιαφέροντος για το πείραμα εξαναγκασμένης κολύμβησης, σύμφωνα με το πρωτόκολλο που αναφέρθηκε, είναι:

- Ακινησία (Immobility): Η κατάσταση κατά την οποία ο επίμυς είναι απολύτως ακίνητος ή κάνει τις ελάχιστες κινήσεις ώστε να επιπλέει στο νερό.
- Κολύμβηση (Swimming): Η κατάσταση όπου ο επίμυς κινείται, συνήθως οριζοντίως, κινούμενος κατά πλάτος του κυλίνδρου.
- Αναρρίχηση (Climbing): Η κατάσταση κατά την οποία ο επίμυς, με κινήσεις των εμπρόσθιων ποδιών, κινείται κατά μήκος του κυλίνδρου, προσπαθώντας να εξέλθει απ' τον κύλινδρο.
- Τίναγμα Κεφαλής (Head Shaking): Χαρακτηριστική κίνηση των επιμύων, για την αποβολή του νερού απ το κεφάλι τους.
- Κατάδυση (Diving): Βουτιά του επιμύος στο νερό.

2.1.2 Αυτοματοποίηση των μετρήσεων του Πειράματος

Η σχετική βιβλιογραφία για την αυτοματοποίηση του πειράματος εξαναγκασμένης κολύμβησης είναι σημαντικά περιορισμένη. Επιπλέον η σύγκριση μεταξύ των τεχνικών που έχουν υλοποιηθεί είναι αδύνατη, καθώς δεν υπάρχει κάποιο σύνολο δεδομένων που να αποτελεί πρότυπο για τη σύγκριση και αξιολόγηση των αλγορίθμων ταξινόμησης του πειράματος εξαναγκασμένης κολύμβησης.

Το 2001 δημοσιεύτηκε το άρθρο των Gaël Hédou et al. [4] στο οποίο έγινε χρήση του λογισμικού αναγνώρισης κίνησης EthoVision και ο βαθμός της κίνησης των επιμυών καθορίστηκε σε σχέση με την απόσταση της συνολικής του μετακίνησης κατά τη διάρκεια του πειράματος.

Το 2008 δημοσιεύτηκε το άρθρο των Holger Fröhlich et al. [5] το οποίο αντιμετώπισε το ίδιο πρόβλημα με τη χρήση Μηχανών Διανυσματικής Υποστήριξης. Συγκεκριμένα η μεθοδολογία που χρησιμοποιήθηκε είναι η αποκοπή του βίντεο σε τμήματα των 5 frames που αντιστοιχούν σε 0.2 sec και η εξαγωγή χαρακτηριστικών κίνησης με τεχνικές optical flow για κάθε δείγμα. Η κίνηση ποσοτικοποιήθηκε για κάθε διάστημα 0.2sec και έτσι προέκυψε ιστόγραμμα κίνησης για κάθε πείραμα. Στη συνέχεια δημιουργήθηκαν profiles για κάθε κατηγορία με βάση την ποσότητα της κίνησης, καθώς δεν υπήρχαν αληθή δεδομένα του πειράματος. Τέλος με βάση τα profiles αυτά, τα χαρακτηριστικά ταξινομήθηκαν στις 3 βασικές κατηγορίες του πειράματος (Ακινησία, Κολύμβηση, Αναρρίχηση).

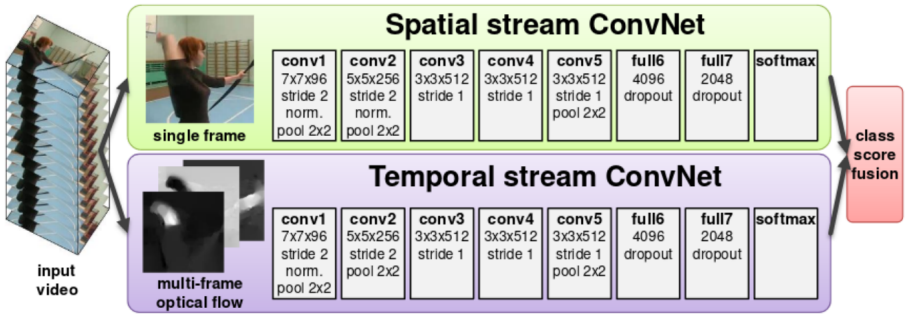
Τέλος, το 2021 δημοσιεύτηκε η εργασία των Arnab Nandi et al. [6] για το λογισμικό DBscorer. Το λογισμικό αυτό αποτελεί ένα γραφικό περιβάλλον ανεπτυγμένο στο προγραμματιστικό περιβάλλον Matlab, το οποίο κατόπιν υπόδειξης και επιβεβαίωσης από τον χρήστη του σημείου ενδιαφέροντος του βίντεο καθώς και τον εντοπισμό του υποβάθρου, εντοπίζει αυτόματα σε 2 μόνο κατηγορίες, δηλαδή κίνηση και μη κίνηση για το πείραμα της εξαναγκασμένης κολύμβησης. Η μεθοδολογία που χρησιμοποιείται στην έρευνα αυτή είναι παρόμοια με εκείνης των Holger Fröhlich et al. [5], δηλαδή με αφαίρεση του υποβάθρου και έπειτα αφαίρεση των διαδοχικών στιγμιοτύπων. Έτσι προκύπτει μια εικόνα με τις διαφορές μεταξύ δυο διαδοχικών frames, τύπου οπτικής κίνησης. Ο μέσος όρος αυτής της εικόνας οπτικής ροής υπολογίζεται, και αποτελεί τον δείκτη για την ταξινόμηση στην κατηγορία της κινητικότητας ή την κατηγορία της ακινησίας, σύμφωνα με κατώφλι (threshold) που ορίστηκε έπειτα από παρατήρηση και μελέτη.

Παρατηρούμε ότι σε όλες τις παραπάνω έρευνες, λείπουν βασικές από τις κατηγορίες του πειράματος εξαναγκασμένης κολύμβησης και έχουν απλοποιήσει σημαντικά το πρόβλημα. Ταυτόχρονα αξιοποιούν μεθόδους όρασης υπολογιστών που κυριαρχούσαν τις προηγούμενες δεκατιές, πριν την επανάσταση της βαθιάς μάθησης.

2.2 Ταξινόμηση Βίντεο

2.2.1 Συμβατικές μέθοδοι και αξιοποίηση του Deep Learning

Πριν την εποχή της επανάστασης της βαθιάς μάθησης, τεχνολογία αιχμής στην ταξινόμηση βίντεο αποτελούσε η μέθοδος των πυκνών τροχιών, [7], η οποία πετυχαίνει ακρίβεια 87.9% στο dataset UCF-101. Δουλεία ορόσημο για την λύση του προβλήματος αυτού με deep learning αποτέλεσε η δημοσίευση των Karen Simonyan και Andrew Zisserman το 2014 [8], με τα συνελικτικά δίκτυα 2 ροών. Η τεχνική αυτή είναι υλοποίηση της ιδέας ότι ο άνθρωπος για να αποκωδικοποιήσει την πληροφορία της όρασης και να αναγνωρίσει μια ανθρώπινη δράση, επεξεργάζεται την σύνθεση των αντικειμένων που βλέπει και της κίνησης αυτών. Έτσι αξιοποιήθηκαν 2 διαφορετικά δίκτυα, ένα για την ταξινόμηση εικόνων, και ένα για την ταξινόμηση της κίνησης, το οποίο τροφοδοτήθηκε με εικόνες οπτικής ροής που εξήχθησαν απ' το βίντεο. Οι εικόνες οπτικής ροής χρησιμοποιήθηκαν σε μεγάλο βαθμό τα επόμενα χρόνια σε πλήθος δημοσιεύσεων για την επίλυση αυτού του προβλήματος. Η αρχιτεκτονική αυτή πέτυχε ευστοχία 88%, και αποτέλεσε την τεχνολογία αιχμής της εποχής, ξεπερνώντας τις ρηχές προσεγγίσεις.



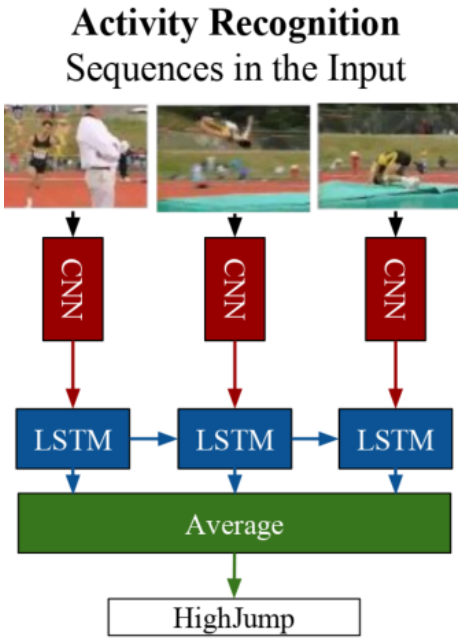
Σχήμα 2.2: Γραφική αναπαράσταση της αρχιτεκτονικής 2 ροών. [8]

2.2.2 Αρχιτεκτονικές LSTM

Επόμενη σημαντική απόπειρα αποτέλεσε η αξιοποίηση των δικτύων Long short-term memory (LSTM) στο πρόβλημα της αναγνώρισης δράσης. Η αρχιτεκτονική LRCN [9] αξιοποίηση δίκτυο CNN σε συνδυασμό με αρχιτεκτονική LSTM και πέτυχε ευστοχία 82.3% στο dataset UCF-101.

2.2.3 Αρχιτεκτονικές 3D Συνελίξεων

Μία άλλη μέθοδος, ίσως η δημοφιλέστερη σήμερα, για την αναγνώριση δράσης είναι οι τρισδιάστατες συνελίξεις. Η καινοτόμα δουλειά 'Learning Spatiotemporal Features with 3D Convolutional Networks' [10] μελέτησε τις 3D συνελίξεις για την αναγνώριση δράσης. Το μοντέλο αυτό εξήγαγε χαρακτηριστικά με τρισδιάστατες συνελίξεις τα



Σχήμα 2.3: Γραφική αναπαράσταση της αρχιτεκτονικής LRCN. [9]

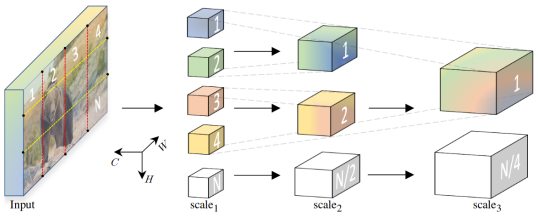
οποία τελικά ταξινομούσε με SVM ταξινομητή. Η χρήση των τρισδιάστατων συνελίξεων, παρότι ήταν ιδιαίτερα ακριβή επεξεργαστικά, αποτελεί μέχρι και σήμερα το στοιχείο που αξιοποιούν οι τεχνολογίες αιχμής για την αναγνώριση δράσης. Τέτοιες αρχιτεκτονικές είναι παραδείγματος χάρη η Inflated 3D [11] που, αποτελεί πιστή μίμηση της αρχιτεκτονικής Inception [12] αλλά με τρισδιάστατες συνελίξεις. Η έρευνα αυτή έδωσε ιδιαίτερη έμφαση στην αξιοποίηση της υπάρχουσας γνώσης για την ταξινόμηση εικόνων στον τομέα της ταξινόμησης βίντεο καθώς και στην αξιοποίηση προεκπαιδευμένων βαρών από ταξινόμηση εικόνας στην ταξινόμηση βίντεο. Το πρόβλημα που παρέμενε ήταν ο τεράστιος αριθμός παραμέτρων που δημιουργούσε πρακτικές δυσκολίες στην εκπαίδευση των δικτύων αυτών. Με το ζητούμενο αυτό ασχολήθηκε η έρευνα 'A Closer Look at Spatiotemporal Convolutions for Action Recognition' [13] που εισήγαγε το δίκτυο Resnet 2+1D. Το δίκτυο αυτό αξιοποιεί την μάθηση με residuals όπως ένα συνηθισμένο Resnet, αλλά ταυτόχρονα εκφράζει τις 3D συνελίξεις ως ένα παράγωγο της χωρικής συνέλιξης ($1x3x3$) και της χρονικής ($3x1x1$).

Μία διαφορετική σύγχρονη προσέγγιση είναι το δίκτυο SlowFast [14] η οποία περιλαμβάνει 2 ροές. Η πρώτη ροή αφορά την αναγνώριση τον ποιοτικών χαρακτηριστικών του βίντεο και τροφοδοτείται με χαμηλή χρονική ανάλυση (frames per second), ενώ η δεύτερη ροή δέχεται ως είσοδο βίντεο με υψηλή χρονική ανάλυση και στοχεύει στην αναγνώριση των στοιχείων κίνησης του δείγματος.

2.2.4 Αρχιτεκτονικές Μετασχηματιστών

Τέλος πρέπει να αναφερθεί η τεχνολογία αιχμής στην ταξινόμηση βίντεο, η οποία είναι οι Μετασχηματιστές Πολλαπλών Κλιμάκων MViT (Multiscale Video Transformers) [15]. Οι αρχιτεκτονικές αυτές έχουν υιοθετήσει την τεχνική Attention [16] η οποία χρησιμοποιείται πολλά χρόνια τώρα στην επεξεργασία φυσικής γλώσσας, στον τομέα της αναγνώρισης βίντεο και γενικότερα της όρασης υπολογιστών. Η αρχιτεκτονική MViT διαφοροποιείται με τα συμβατικά συνελικτικά δίκτυα με σταθερό μέγεθος καναλιών και μοναδική κλίμακα στις εικόνες σε όλο το δίκτυο. Αντιθέτως χρησιμοποιεί με-

τασχηματιστές πολλαπλών κλιμάκων με πολλά στάδια διαφορετικών καναλιών και αναλύσεων. Ξεκινώντας απ' την αρχική ανάλυση της εικόνας, καθώς και τα κανάλια της, τα στάδια ιεραρχικά αυξάνουν το εύρος των καναλιών ενώ μειώνουν την χωρική ανάλυση. Έτσι δημιουργείται μια πυραμίδα διαφορετικών κλιμάκων με χαρακτηριστικά των εικόνων εντός του μετασχηματιστή. Βελτιώση αποτέλεσε η αρχιτεκτονική MViTv2 [17] που δημοσιεύτηκε το 2022 πετυχαίνοντας Accuracy 86.1% στο dataset Kinetics-400.



Σχήμα 2.4: Γραφική αναπαράσταση της αρχιτεκτονικής μετασχηματιστών πολλαπλών κλιμάκων. [9]

2.3 Εκπαίδευση με Χρήση Συμπιεσμένης Πληροφορίας Βίντεο

Μια σταθερή δυσκολία στην ανάπτυξη αποδοτικών μοντέλων ταξινόμησης δράσης είναι η ύπαρξη μεγάλου όγκου δεδομένων. Το γεγονός αυτό φέρνει περιορισμούς τόσο λόγω υπολογιστικών πόρων, όσο και λόγω δυσκολίας στην εξαγωγή αντιπροσωπευτικών χαρακτηριστικών προς ταξινόμηση. Τα τελευταία χρόνια, παρατηρείται η τάση χρήσης συμπιεσμένων αναπαραστάσεων βίντεο για την εκπαίδευση. Η συμπίεση των βίντεο στηρίζεται στο γεγονός ότι οι αλλαγές από καρέ σε καρέ είναι μικρές και μπορούν να εντοπισθούν σχετικά εύκολα με μικρή απώλεια πληροφορίας. Συνήθως αποθηκεύεται το αρχικό frame ($I-frame$) και οι διαφορές των διαδοχικών καρέ σχετικά με την κίνηση και την διαφορά ($P-frame, B-frame$).

Αρκετές δημοσιεύσεις πραγματεύονται το θέμα της εκπαίδευσης με χρήση τέτοιων διανυσμάτων κίνησης [18], [19], [20]. Μάλιστα η τεχνολογία αιχμής για το 2021 για την ταξινόμηση βίντεο η οποία ονομάζεται 'SCSampler' [21] χρησιμοποιεί τέτοιου είδους κωδικοποίηση βίντεο.

Κεφάλαιο 3

Μεθοδολογία

Στο κεφάλαιο αυτό παρουσιάζεται η μεθοδολογία που ακολουθήθηκε σύμφωνα με τον κύκλο ζωής της μηχανικής μάθησης. Πρόκειται για μια σειρά από διαδικασίες η οποία θεωρείται κοινή για τα περισσότερα προβλήματα που καλείται να λύσει η επιστήμη της μηχανικής μάθησης. Ο κύκλος περιλαμβάνει τα ακόλουθα 7 στάδια:

1. Ορισμός του προβλήματος: Το πρώτο στάδιο αφορά τον σαφή ορισμό του προβλήματος και την κατανόηση της αξίας της επίλυσης του καθώς και τις εφαρμογές που μπορεί να έχει.

2. Συλλογή Δεδομένων: Εφόσον το πρόβλημα έχει οριστεί ξεκάθαρα, είναι αναγκαία η συλλογή των κατάλληλων δεδομένων από πηγές δεδομένων. Στη φάση αυτή, πρέπει να καθοριστεί το είδος των δεδομένων, η πηγή και ο τρόπος που μπορούν να βρεθούν, καθώς και λειτουργικοί τρόποι για την αποθήκευση και πρόσβαση σε αυτά.

3. Διαμόρφωση Δεδομένων: Κατά το στάδιο αυτό, τα δεδομένα μετασχηματίζονται απ' την αρχική τους, ωμή μορφή σε μορφή που καθορίζεται έτσι ώστε να είναι χρήσιμα για την επίλυση του προβλήματος που έχει καθοριστεί. Πρόκειται για χρονοβόρα διαδικασία που αφορά το καθάρισμα και τη μετατροπή των δεδομένων. Οι διαδικασίες που εμπλέκονται σε αυτή τη φάση είναι η εξερεύνηση, προ-επεξεργασία, μετασχηματισμοί, συμπλήρωση ή διαγραφή ελλιπών δεδομένων και θορύβου και μετατροπή τους στη κατάλληλη μορφή.

4. Οπτικοποίηση και Εξερεύνηση των Δεδομένων: Το στάδιο αυτό είναι απαραίτητο για την κατανόηση των δεδομένων. Η οπτικοποίηση βοηθάει στην εύρεση μοτίβων και τάσεων των δεδομένων και αναδεικνύει συσχετίσεις μεταξύ μεταβλητών του προβλήματος. Η δημιουργία γραφημάτων επιτρέπει την εμβάθυνση στην κατανόηση των δεδομένων, που τελικά θα καθορίσει τα επόμενα στάδια για την επίλυση του προβλήματος.

5. Ανάπτυξη Μοντέλων: Το στάδιο αυτό περιλαμβάνει την επιλογή και την εκπαίδευση μοντέλων μηχανικής μάθησης. Η μαθηματική γνώση και η γνώση επιστήμης υπολογιστών συνδυάζονται για την εκπαίδευση αλγορίθμων μηχανικής μάθησης, που θα προβλέψουν και θα εκτιμήσουν σύμφωνα με τα παρεχόμενα δεδομένα. Οι αρχιτεκτονικές επιλέγονται με κριτήριο την φύση των δεδομένων, τον διαθέσιμο χρόνο για εκπαίδευση και εφαρμογή του αλγορίθμου κλπ.

6. Αξιολόγηση Μοντέλων: Πρόκειται για το σημαντικό βήμα που καθορίζει την ποιότητα και ακρίβεια των μοντέλων που προέκυψαν. Τα μοντέλα αξιολογούνται σύμφωνα με δείκτες αξιολόγησης (ευστοχία, ακρίβεια, F1-Score, πίνακας σύγχυσης κλπ. για την ταξινόμηση, μέσο τετραγωνικό σφάλμα, μέσο απόλυτο σφάλμα κλπ για την παλινδρόμησή) σε τμήματα των δεδομένων τα οποία δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση των δεδομένων.

7. Deployment και επίβλεψη του μοντέλου: Το τελευταίο στάδιο αφορά την παράταξη (deployment) του μοντέλου σε περιβάλλον παραγωγής για την λήψη αποφάσεων ή τη δημιουργία προβλέψεων σύμφωνα με τα παρεχόμενα δεδομένα. Η

συνεχής παρατήρηση και επίβλεψη της εφαρμογής του μοντέλου είναι απαραίτητη και γι' αυτό συνηθίζεται η κατασκευή ροών εργασιών για την εξασφάλιση της ορθής λειτουργίας του. Στην παρούσα εργασία δεν υπάρχει deployment του μοντέλου με την πραγματική έννοια, εφόσον το λογισμικό δεν έχει χρησιμοποιηθεί στον τομέα της παραγωγής. Αντ' αυτού θα περιγραφεί το υλικό και το λογισμικό που χρειάστηκε για την εκπόνηση των πειραμάτων.

Τα στάδια θα εξειδικευτούν στις επόμενες ενότητες για τη συγκεκριμένη διπλωματική και το πρόβλημα της εξαναγκασμένης κολύμβησης.

3.1 Ορισμός του Προβλήματος

Το πρόβλημα που καλείται να λυθεί στη παρούσα εργασία είναι η αυτόματη ταξινόμηση του πειράματος εξαναγκασμένης κολύμβησης. Πρόκειται για πείραμα κατά το οποίο οι επιμύες τοποθετούνται στο νερό για 5 λεπτά και παρατηρείται η συμπεριφορά τους. Η συμπεριφορά τους κατηγοριοποιείται σε 5 κατηγορίες οι οποίες είναι:

- Αναρρίχηση
- Κολύμβηση
- Ακινησία
- Κατάδυση
- Τίναγμα κεφαλιού

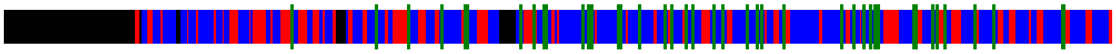
Για την ταξινόμηση του συνηθίζεται η παρατήρηση του βίντεο σε κανονική ταχύτητα, με τον παρατηρητή να επιλέγει για κάθε στιγμή την συμπεριφορά των επιμυών, πατώντας 5 διαφορετικά κουμπιά. Όπως είναι ολοφάνερο, για κάθε βίντεο απαιτούνται 5 λεπτά εργασίας, αρκετά κουραστικής, και πιθανότατα προκύπτουν καθυστερήσεις από τον χρόνο απόκρισης του παρατηρητή.

Για το λόγο αυτό η αυτοματοποίηση της ταξινόμησης της συμπεριφοράς των επιμυών κατά το πείραμα εξαναγκασμένης κολύμβησης είναι ιδιαίτερα σημαντική, καθώς θα γλιτώσει πολλές ώρες εργασίας από τους ερευνητές.

Τα αποτελέσματα αυτού του πειράματος αξιοποιούνται για την έρευνα και ανάλυση της επίδρασης φαρμακευτικών ουσιών στην επίλυση της κατάθλιψης και άλλων ψυχικών ασθενειών.

3.2 Συλλογή δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στη διπλωματική προέρχονται από το εργαστήριο Νευροψυχοφαρμακολογίας της Ιατρικής σχολής του ΕΚΠΑ, από τους καθηγητές Χριστίνα Δάλλα και Νίκο Κόκρα. Τα πειράματα έχουν διεξαχθεί και ταξινομηθεί από αυτούς. Τα δεδομένα που παραλάβαμε ήταν 100 αρχεία βίντεο διάρκειας 5 λεπτών, και εικόνες με την ταξινόμηση της συμπεριφοράς του επιμύ ανά χρονική στιγμή, με μορφή όπως φαίνεται στην εικόνα 3.1.



Σχήμα 3.1: Παράδειγμα εικόνα ταξινόμησης της συμπεριφοράς των επιμυών. Τα χρώματα αντιστοιχίζονται σε διαφορετικές κατηγορίες συμπεριφοράς του επιμύ.

Εκτός από τα 100 ταξινομημένα βίντεο που αναφέρθηκαν διάρκειας περίπου 8 ωρών, παραδόθηκαν και περίπου 80 επιπλέον ώρες αταξινόμητων βίντεο.

3.3 Διαμόρφωσή δεδομένων

Το dataset που χρησιμοποιήθηκε υπέστη μια σειρά τροποποιήσεων ώστε να έρθει σε μορφή κατάλληλη για αξιοποίηση του σε διαδικασίες επιβλεπόμενης ταξινόμησης. Οι τροποποιήσεις αυτές παρουσιάζονται παρακάτω με την αντίστοιχη σειρά.

1. Διαχωρισμός πειραμάτων: Μια συνήθης τακτική για την διεξαγωγή του πειράματος εξαναγκασμένης κολύμβησης είναι η καταγραφή των αντιδράσεων σε 2 επιμύες ταυτόχρονα. Για το λόγο αυτό πολλά από τα βίντεο περιέχουν δύο πειράματα ταυτόχρονα. Τα βίντεο αυτά κόπηκαν στα 2 ώστε να διαχωριστούν.

2. Για τη μείωση του όγκου των δεδομένων, όλα τα βίντεο κόπηκαν εκ' νέου, ώστε ο χώρος που καταγράφεται να συμπίπτει με την δεξαμενή στην οποία διεξάγεται το πείραμα.

3. Τα βίντεο μετατράπηκαν σε πολλαπλές εικόνες μορφής JPEG, ώστε να είναι γρήγορη η πρόσβαση σε αυτές, καθώς και βέλτιστος ο τρόπος αποθήκευσης τους.

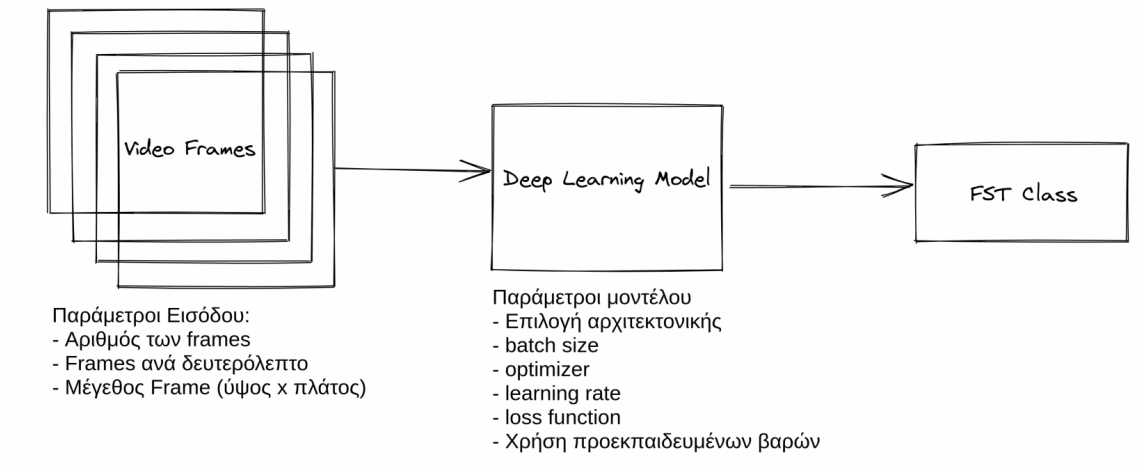
Η ανάπτυξη του λογισμικού που απαιτήθηκε, έγινε με τρόπο ώστε να είναι δυνατή η παραμετροποίηση όλων των στοιχείων που ενδέχεται να επηρεάσουν τα αποτελέσματα του παραγόμενου μοντέλου. Το dataset μετά την βασική προεπεξεργασία αποτελείται από τις εισόδους των μοντέλων, δηλαδή τα frames του κάθε βίντεο, καθώς και τις εικόνες που αναπαριστούν τις παρατηρήσεις των πειραμάτων.

Υπενθυμίζεται ότι τα δεδομένα εκπαίδευσης είναι πεντάλεπτα βίντεο, και αναγκαία είναι η ταξινόμηση τουλάχιστον κάθε δευτερολέπτου αυτών. Για το λόγο αυτό γίνεται η βασική παραδοχή ότι η ταξινόμηση θα διενεργηθεί σε τμήματα του βίντεο με σταθερό χρονικό διάστημα μεταξύ τους, τα οποία είναι απολύτως ανεξάρτητα. Οι αληθείς τιμές βρίσκονται σε αρχεία εικόνων, δυο για κάθε βίντεο λόγω των δυο διαφορετικών παρατηρητών. Υλοποιήθηκε λογισμικό που επιτρέπει την παραμετροποίηση των εξής υπερπαραμέτρων των δεδομένων:

- Τα frames ανά second (fps): Πρόκειται για την χρονική ακρίβεια των βίντεο. Ανάλογα με τα fps που επιλέγονται για την εκπαίδευση κάθε μοντέλου, υπολογίζονται αυτομάτως και οι αντίστοιχες αληθείς τιμές, πχ για 25 fps παραλείπεται κάθε δεύτερη τιμή των αληθών τιμών. Εκτιμάται ότι με χαμηλά fps, δεν θα είναι δυνατή η αναγνώριση των γρήγορων κινήσεων από το μοντέλο, ενώ αντιθέτως με πολύ ψηλά fps , θα δυσχεραίνεται η χρονική συσχέτιση των frames του βίντεο, λόγω της υπέρογκης πληροφορίας, καθώς και θα καθυστερεί ιδιαίτερα η διαδικασία της εκπαίδευσης. Χρησιμοποιήθηκαν $16fps$ καθώς η χρήση $25fps$, που αποτελεί τη συνήθη χρονική ανάλυση των βίντεο δεν προσδίδει ιδιαίτερη πληροφορία, και ανεβάζει σημαντικά την ανάγκη για υπολογιστική ισχύ για την εκπαίδευση των μοντέλων, καθώς διπλασιάζει το μέγεθος εισόδου των δικτύων.

- Το μέγεθος της εικόνας: Το αρχικό μέγεθος της εικόνας πρόκειται να επηρεάσει και πάλι την ταχύτητα εκπαίδευσης και εκτίμησης. Ωστόσο με πολύ μικρό μέγεθος ενδέχεται να δυσχεραίνεται η αναγνώριση λεπτομερών χωρικών προτύπων από το μοντέλο. Η τιμή που χρησιμοποιήθηκε είναι $112x112$, καθώς κατά σύμβαση αποτελεί το μέγεθος που έχουν τα πιο συνηθισμένα datasets όπως το Kinetics, και άρα και τα μοντέλα που εκπαιδεύονται με αυτό.

- Το χρονικό διάστημα κάθε βίντεο προς εκτίμηση: Πρόκειται για το πόσα δευτερόλεπτα ή πόσα frames θα αποτελούν το κάθε βίντεο που θα τροφοδοτεί την είσοδο των μοντέλων. Με μικρές τιμές οι δράσεις των επιμυών δεν θα είναι αναγνωρίσιμες, ωστόσο με μεγάλες τιμές ενδέχεται να συμπεριλαμβάνονται πάνω από μία κατηγορίες συμπεριφοράς στο ίδιο βίντεο. Επιλέχθηκε κάθε βίντεο να αποτελείται από $1.3sec$, χρονικό διάστημα που κρίθηκε ότι είναι αρκετό για την εξαγωγή συμπεράσματος για την συμπεριφορά του επιμύ.



Σχήμα 3.2: Διαγραμματική αναπαράσταση των βασικών υπερπαραμέτρων της εκπαίδευσης

Συνολικά, κατασκευάστηκε αλγόριθμος που δέχεται όλες τις παραπάνω παραμέτρους, και αυτομάτως επιστρέφει τις εισόδους που ζητήθηκαν, καθώς και τις κατάλληλες εξόδους για τα δίκτυα κατά την εκπαίδευση. Όλες οι παραπάνω παράμετροι πρόκειται να βελτιστοποιηθούν. Πραγματοποιήθηκε επιπλέον augmentation κατά τη στιγμή της φόρτωσης των βίντεο στο μοντέλο. Συγκεκριμένα:

- Πιθανότητα 50% καθρεφτισμού κατά των άξονα x, για κάθε βίντεο που εισάγεται στο μοντέλο,

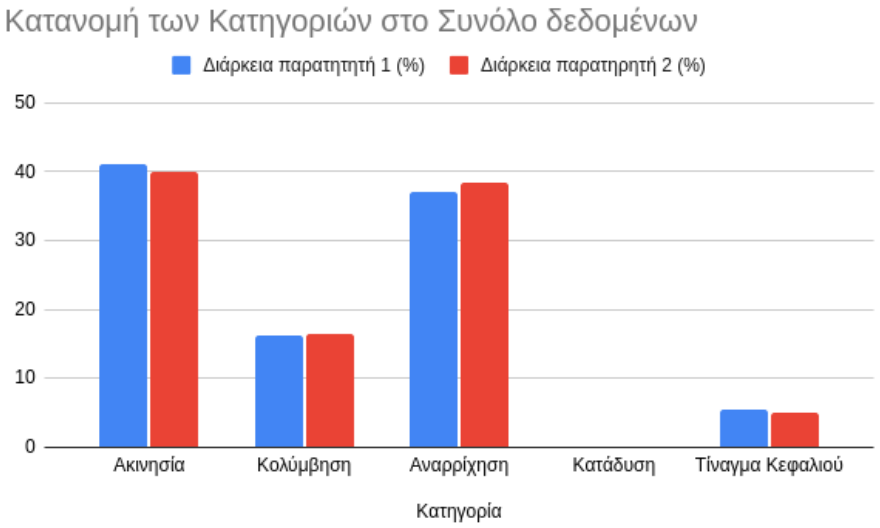
- Τυχαία χωρική αποκοπή (crop) της τάξης του 5% του βίντεο στις διαστάσεις x, y των εικόνων

3.4 Εξερεύνηση δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν περιλαμβάνουν **79.2 ώρες αταξινόμητων** βίντεο τα οποία αφορούν τη πρώτη φάση του πειράματος εξαναγκασμένης κολύμβησης. Κατά την φάση αυτή οι επιμύες τοποθετούνται στο νερό για 15 λεπτά ώστε να εξοικειωθούν με τη διαδικασία και δεν έχουν κατά τ άλλα κάποιο πειραματικό ενδιαφέρον,

γι' αυτό και είναι αταξινόμητα. Τα δεδομένα αυτά προέρχονται από 4 διαφορετικά σετ πειραμάτων που διεξήχθησαν απ' το τμήμα της Ιατρικής του Πανεπιστημίου Αθηνών και διαθέτουν εκατοντάδες διαφορετικούς επιμύες, καθώς και αρκετά διαφορετικές λήψεις μεταξύ τους.

Το σύνολο των δεδομένων περιλαμβάνει ακόμη **8.2 ώρες ταξινομημένων** βίντεο τα οποία αποτελούν και το βασικότερο μέρος του, παρόλο που υστερούν κατά πολύ σε όγκο. Τα βίντεο αυτά περιλαμβάνουν ταξινομήσεις από 2 διαφορετικούς παρατηρητές. Οι παρατηρήσεις στο συγκεκριμένο πείραμα καταγράφουν την συμπεριφορά του επιμύ για κάθε χρονική στιγμή κατά την διεξαγωγή του πειράματος, το οποίο διαρκεί 5 λεπτά. Αποτελείται από 100 βίντεο με διαφορετικούς επιμύες. Τεχνικά η ακρίβεια των παρατηρήσεων είναι 0.3 sec, στην πραγματικότητα ωστόσο δεν μπορεί να θεωρηθεί μεγαλύτερη από 1 sec καθώς επηρεάζεται από τον χρόνο αντίδρασης του παρατηρητή.



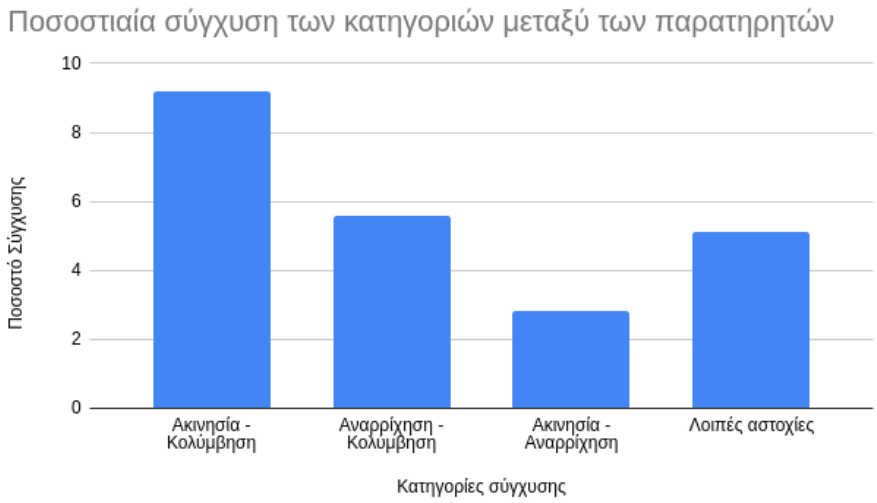
Σχήμα 3.3: Διαγραμματική αναπαράσταση της κατανομής των κατηγοριών στο σύνολο δεδομένων του FST.

Όπως προκύπτει παρουσιάζεται στο σχήμα 3.3 του σχήματος υπάρχει μεγάλη ανισορροπία μεταξύ των κλάσεων, καθώς η κατηγορίες climbing και immobility καταλαμβάνουν το 78% του dataset, ενώ οι 3 υπόλοιπες το 22%. Για την αποφυγή προβλημάτων κατά την εκπαίδευση του dataset λόγω των λάθος τιμών των παρατηρήσεων σε πολλές περιπτώσεις, αποφασίστηκε να διατηρηθούν μόνο τα χρονικά διαστήματα για τα οποία συμφωνούν οι παρατηρητές.

Το γεγονός ότι υπάρχουν διαθέσιμες παρατηρήσεις από 2 διαφορετικούς παρατηρητές, κάνει δυνατό τον υπολογισμό της συμφωνίας μεταξύ τους καθώς και την σύγκριση τους για τον υπολογισμό σύγχυση μεταξύ των κατηγοριών του πειράματος.

Απ' τη σύγκριση που παρουσιάζεται στο σχήμα 3.4 προκύπτει ότι υπάρχει σύγχυση της τάξης του 5-10% μεταξύ των κατηγοριών Ακινησία - Κολύμβηση, καθώς και Αναρρίχηση - Κολύμβηση. Αυτό δικαιολογείται από την δυσκολία στην παρατήρηση του πειράματος. Συνολικά οι παρατηρητές συμφωνούν στο 77.31% της συνολικής διάρκειας των δεδομένων.

Σημειώνεται ακόμη, ότι η κατηγορία 'Τίναγμα Κολύμβησης' διαρκεί περίπου 0.5 - 1 sec, γεγονός που καθιστά δύσκολο το συγχρονισμό της στο βίντεο λόγω του χρόνου



Σχήμα 3.4: Διαγραμματική αναπαράσταση της σύγχυσης των κατηγοριών μεταξύ των παρατηρητών στο σύνολο δεδομένων του FST.

αντίδρασης των παρατηρητών.

3.5 Ανάπτυξη Μοντέλων

Τα πειράματα που διεξήχθησαν είχαν ως σκοπό την ταξινόμηση των συμπεριφορών κατά το πείραμα της εξαναγκασμένης κολύμβησης. Η κεντρική ιδέα του σχεδιασμού των πειραμάτων ήταν τόσο η χρήση τεχνολογίας αιχμής για την ταξινόμηση βίντεο όσο και η αξιοποίηση των χιλιάδων μη ταξινομημένων ωρών με διεξαγωγές πειραμάτων.

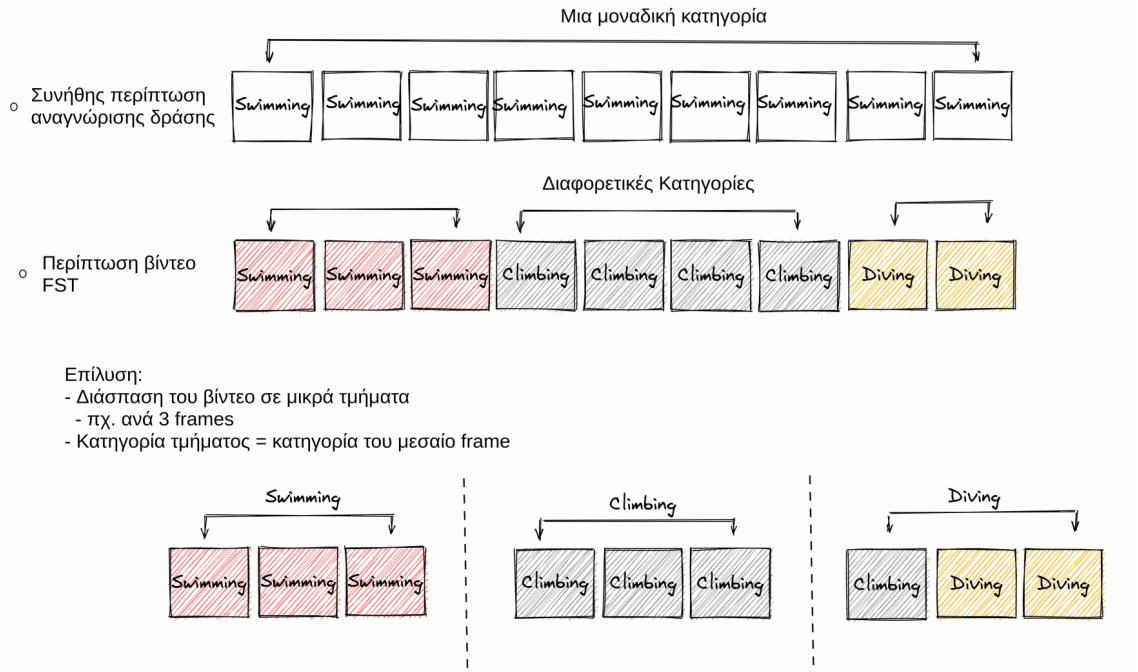
Τα διαφορετικά μοντέλα που εκπαιδεύτηκαν είναι τα εξής:

- Εκπαίδευση μοντέλου με χρήση προ-εκπαιδευμένων βαρών με χρήση τεχνολογιών αιχμής για την ταξινόμησης video.
- Εκπαίδευση αυτοκωδικοποιητών (autoencoders) για την συμπίεση των βίντεο.
- Εκπαίδευση των συμπιεσμένων χαρακτηριστικών των βίντεο.

Για την χρήση όλων των δικτύων που ακολουθούν γίνεται μια βασική σύμβαση. Αυτή είναι ότι κάθε βίντεο θα χωριστεί σε μικρότερα τμήματα της τάξης του ενός δευτερολέπτου, τα οποία στη συνέχεια θα αποτελέσουν την είσοδο των μοντέλων. Ως έξοδος θα επιλεχθεί η τιμή παρατήρησης του μεσαίου frame όπως φαίνεται στην εικόνα 3.5.

3.5.1 Ανάπτυξη μοντέλων με Προ-εκπαιδευμένα Βάρη

Αρχικά δημιουργήθηκαν μοντέλα ταξινόμησης τα οποία θα αποτελέσουν και τη βάση για την σύγκριση με τις επόμενες δοκιμές σύνθεσης αρχιτεκτονικών. Για το σκοπό αυτό χρησιμοποιήθηκαν δυο απ' τα πιο διαδεδομένα δίκτυα αναγνώρισης δράσης, το Resnet 2+1D [13] και το MViTv2 [17]. Τα δίκτυα τροφοδοτήθηκαν με τις ακολουθίες των βίντεο στο έγρωμο σύνθετο RGB, χωρίς την αξιοποίηση του ήχου.



Σχήμα 3.5: Διαγραμματική αναπαράσταση της διάσπασης βίντεο του πειράματος FST για την ταξινόμηση κατηγοριών.

Μοντέλο Resnet 2+1D

Η πρώτη εκπαίδευση επιλέχθηκε να γίνει με τις εξής παραμέτρους:

- Χρήση προ-εκπαιδευμένων βαρών του δικτύου, και πάγωμα όλων εκτός του τελευταίου layer του μοντέλου και χωρίς το πάγωμα κανενός layer (πολλαπλές δοκιμές)
- batch size = 64
- frames per second = 25
- sample frames = 8 και 16 (πολλαπλές δοκιμές)
- optimizer = Adam
- learning rate = 0.0001
- loss = cross entropy
- augmentation εικόνων = 5% και 10% (πολλαπλές δοκιμές)

Μοντέλο MViTv2

Επιπλέον έγινε εκπαίδευση με τη χρήση του μοντέλου MViTv2 [17] το οποίο αποτελεί αρχιτεκτονική που ανήκει στην κατηγορία των transformers και αποτελεί την τεχνολογία αιχμής. Οι παράμετροι που χρησιμοποιήθηκαν είναι αυτοί που προέκυψαν απ' την εκπαίδευση του προηγούμενου μοντέλου και συγκεκριμένα:

- Χρήση προεκπαιδευμένων βαρών του δικτύου, χωρίς πάγωμα τους

- batch size = 64
- frames per second = 25
- sample frames = 16
- optimizer = Adam
- learning rate = 0.0001
- loss = cross entropy
- augmentation εικόνων = 5%

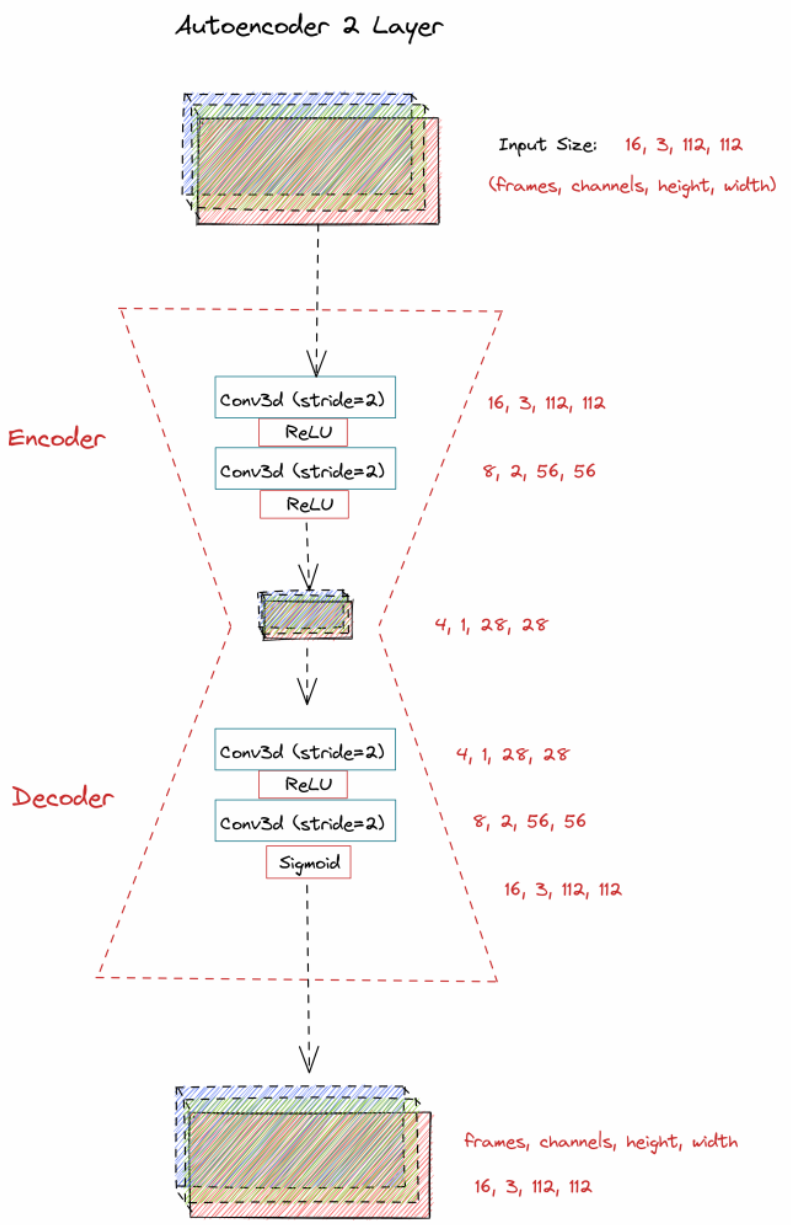
3.5.2 Εκπαίδευση Αυτοκωδικοποιητών

Για την αξιοποίηση των εκατοντάδων ωρών μη ταξινομημένων βίντεο, επιλέχθηκε η χρήση αυτοκωδικοποιητή (autoencoder). Η ιδέα που οδήγησε σε αυτή την απόφαση εκτός από την αξιοποίηση όλων των δεδομένων, είναι ότι το βίντεο περιέχει πάρα πολλή πληροφορία επαναλαμβανόμενη η οποία είναι δυνατόν να συμπιεστεί με κάποιου είδους κωδικοποίηση. Τον ρόλο αυτό μπορεί να αναλάβει ένας αυτοκωδικοποιητής. Πραγματοποιήθηκαν διαφόρων ειδών δοκιμές οι οποίες αναλύονται στα παρακάτω κεφάλαια.

Σε όλα τα πειράματα, ως τελική συνάρτηση ενεργοποίησης, επιλέχθηκε η σιγμοειδής συνάρτηση. Η επιλογή αυτή έγινε καθώς οι εικόνες εισόδου έχουν τιμές απ' το 0 έως το 1, και έτσι η επιλογή της συνάρτησης αυτής διευκολύνει την σύγκριση μεταξύ του βίντεο εισόδου και του βίντεο εξόδου του δικτύου, και οδηγεί σε αντιπροσωπευτική loss function της διαδικασίας της κωδικοποίησης.

2-Layer Autoencoder

Στην εκπαίδευση αυτή επιλέχθηκε να εφαρμοστεί η αρχιτεκτονική που φαίνεται στο σχήμα 3.6. Η αρχιτεκτονική αυτή έχει ως είσοδο:

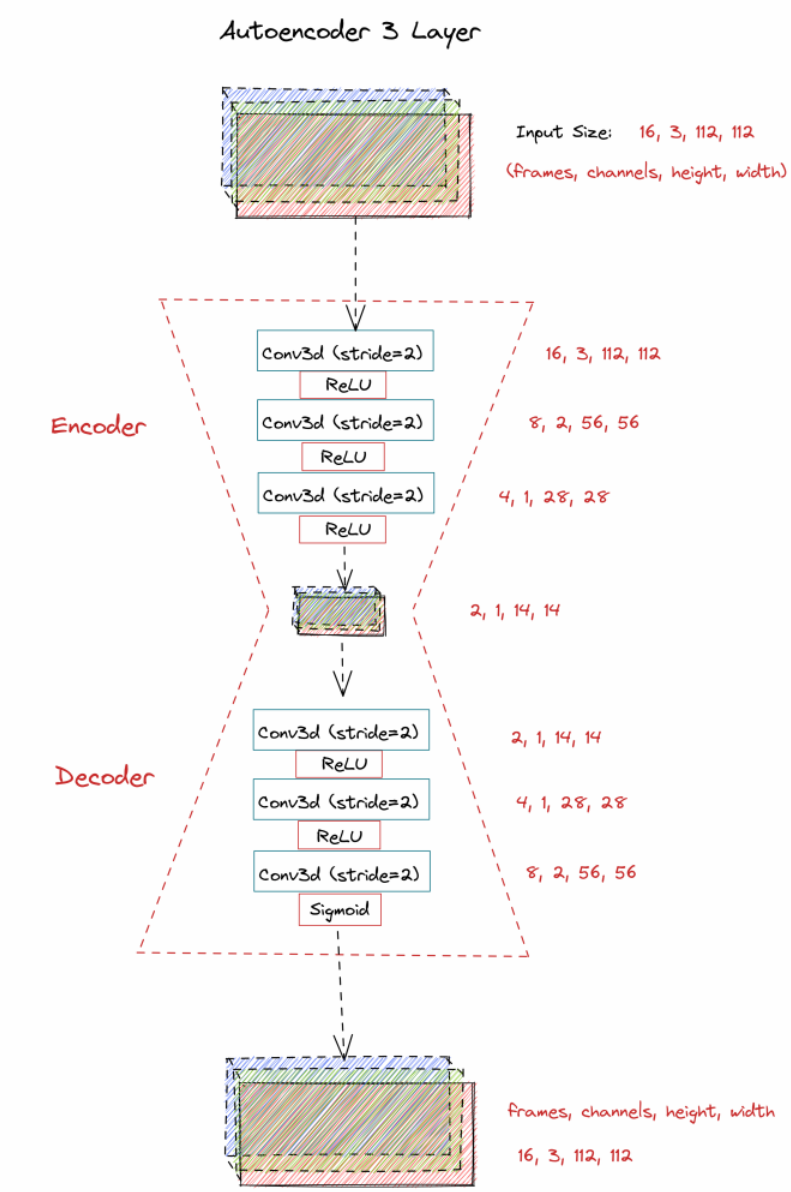


Σχήμα 3.6: Γραφική αναπαράσταση της αρχιτεκτονικής '3-Layer Autoencoder'

Ο encoder εκτελεί 2 φορές 3D συνελίξεις με stride 2, επομένως για 2 αλλεπάλληλες φορές το μέγεθος θα υποδιπλασιαστεί. Με τη συγκεκριμένη αρχιτεκτονική λοιπόν, από $16 * 3 * 112 * 112 = 602112$ που αποτελεί τον αρχικό όγκο της πληροφορίας του βίντεο, μετατρέπεται σε 3136 δηλαδή περίπου 200 φορές μικρότερο.

3-Layer Autoencoder

Στην $2^η$ προσπάθεια, επιλέχθηκε να εφαρμοστεί η αρχιτεκτονική που φαίνεται στο σχήμα 3.7. Στην αρχιτεκτονική αυτή η μόνη διαφορά με την προηγούμενη περίπτωση είναι το γεγονός ότι υπάρχουν 3 layer αντί για 2. Επομένως, συμβαίνει ένας παραπάνω υποδιπλασιασμός στις διαστάσεις της εισόδου του δικτύου.



Σχήμα 3.7: Γραφική αναπαράσταση της αρχιτεκτονικής '3 Layer Autoencoder'

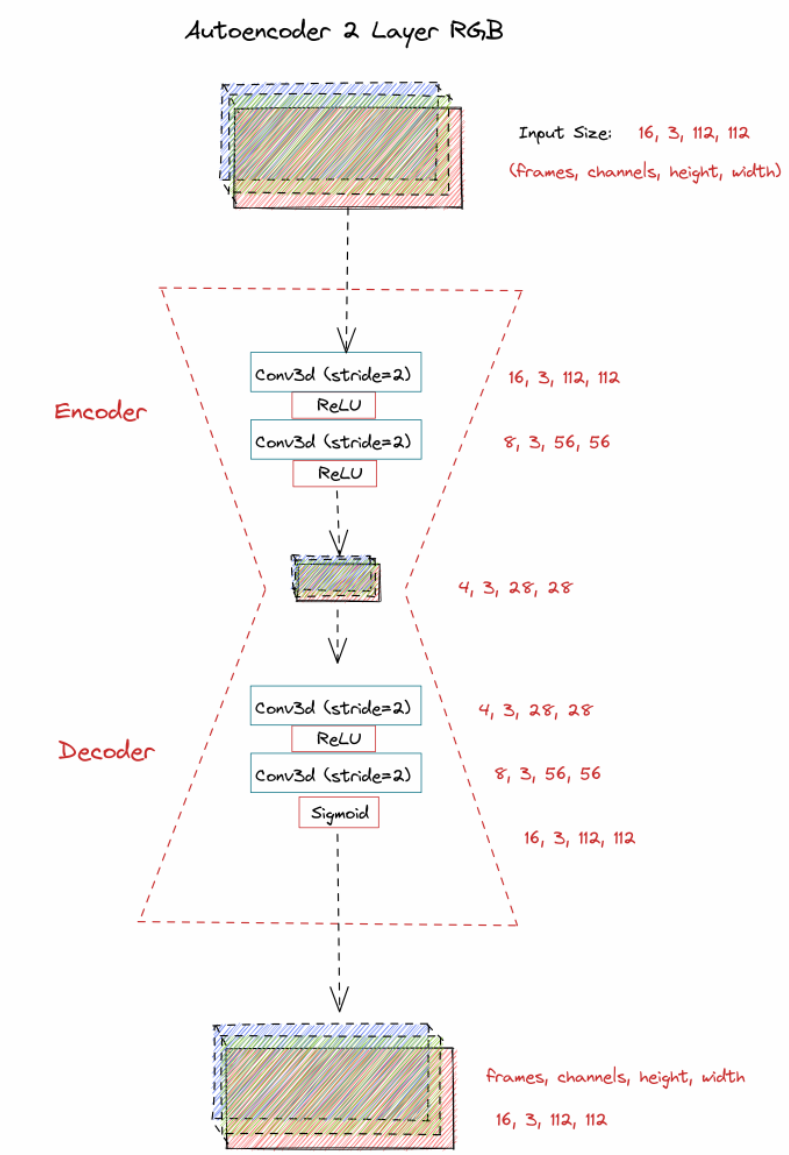
Ο encoder εκτελεί 3 φορές 3D συνελίξεις με stride 2, επομένως για 3 αλλεπάλληλες φορές το μέγεθος θα υποδιπλασιαστεί. Άρα το μέγεθος της κωδικοποιημένης πληροφορίας είναι:

Με τη συγκεκριμένη αρχιτεκτονική λοιπόν, από 602112 που αποτελεί τον αρχικό όγκο της πληροφορίας του βίντεο, μετατρέπεται σε 1536 δηλαδή περίπου 400 φορές μικρότερο.

2-Layer Autoencoder RGB

Στην $3^{η}$ προσπάθεια, επιλέχθηκε να εφαρμοστεί η αρχιτεκτονική που φαίνεται στο σχήμα 3.8. Στην αρχιτεκτονική αυτή η μόνη διαφορά με το πρώτο πείραμα είναι η διατήρηση των 3 καναλιών RGB στην κωδικοποιημένη πληροφορία. Επομένως, αναμένεται τριπλάσιος όγκος της κωδικοποίησης αλλά με διατήρηση των χρωμάτων του

βίντεο.



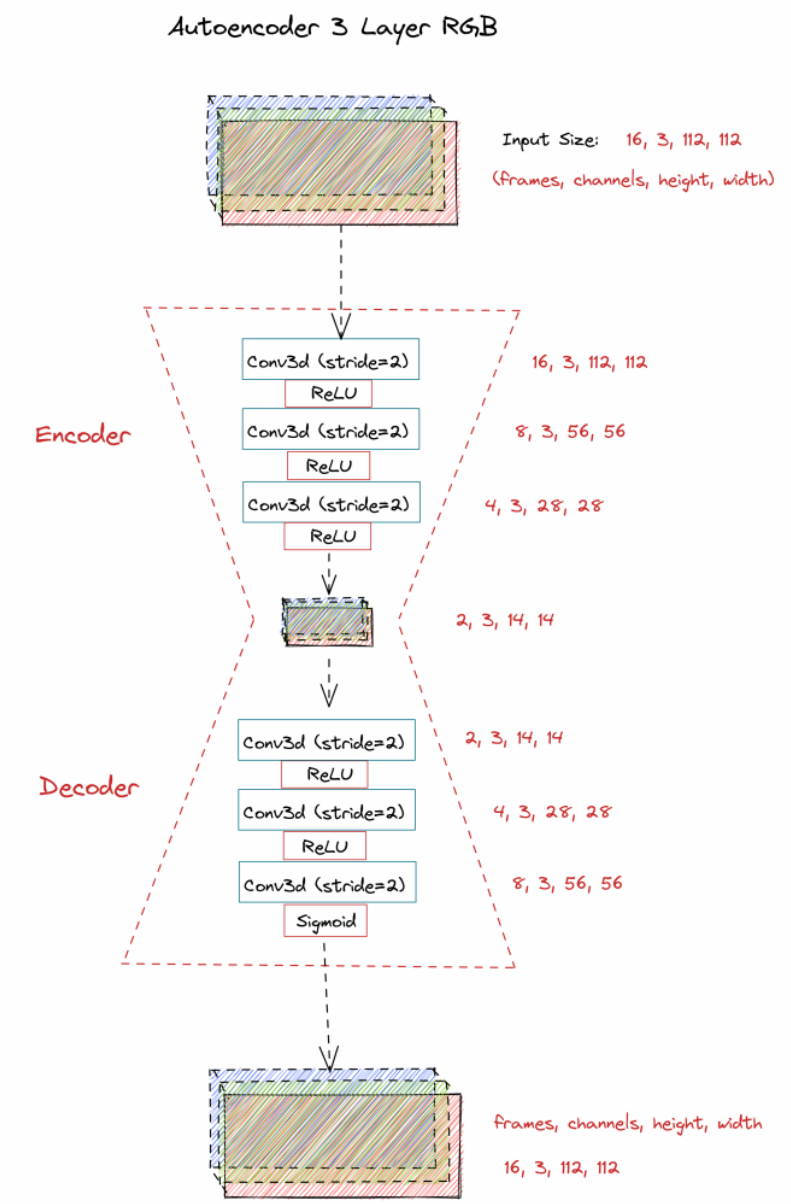
Σχήμα 3.8: Γραφική αναπαράσταση της αρχιτεκτονικής '2 Layer Autoencoder RGB'

Ο encoder εκτελεί 3 φορές 3D συνελίξεις με stride 2, επομένως για 3 αλλεπάλληλες φορές το μέγεθος θα υποδιπλασιαστεί. Άρα το μέγεθος της κωδικοποιημένης πληροφορίας είναι:

Με τη συγκεκριμένη αρχιτεκτονική λοιπόν, από 602112 που αποτελεί τον αρχικό όγκο της πληροφορίας του βίντεο, μετατρέπεται σε 9408 δηλαδή περίπου 64 φορές μικρότερο.

3-Layer Autoencoder RGB

Στην $3^{η}$ προσπάθεια, επιλέχθηκε να εφαρμοστεί η αρχιτεκτονική που φαίνεται στο σχήμα 3.9. Αυτή την φορά επιλέχθηκε η διατήρηση των 3 καναλιών στην κωδικοποιημένη πληροφορία, αλλά με 3 αλλεπάλληλα 3D συνελικτικά επίπεδα.



Σχήμα 3.9: Γραφική αναπαράσταση της αρχιτεκτονικής '3 Layer Autoencoder RGB'

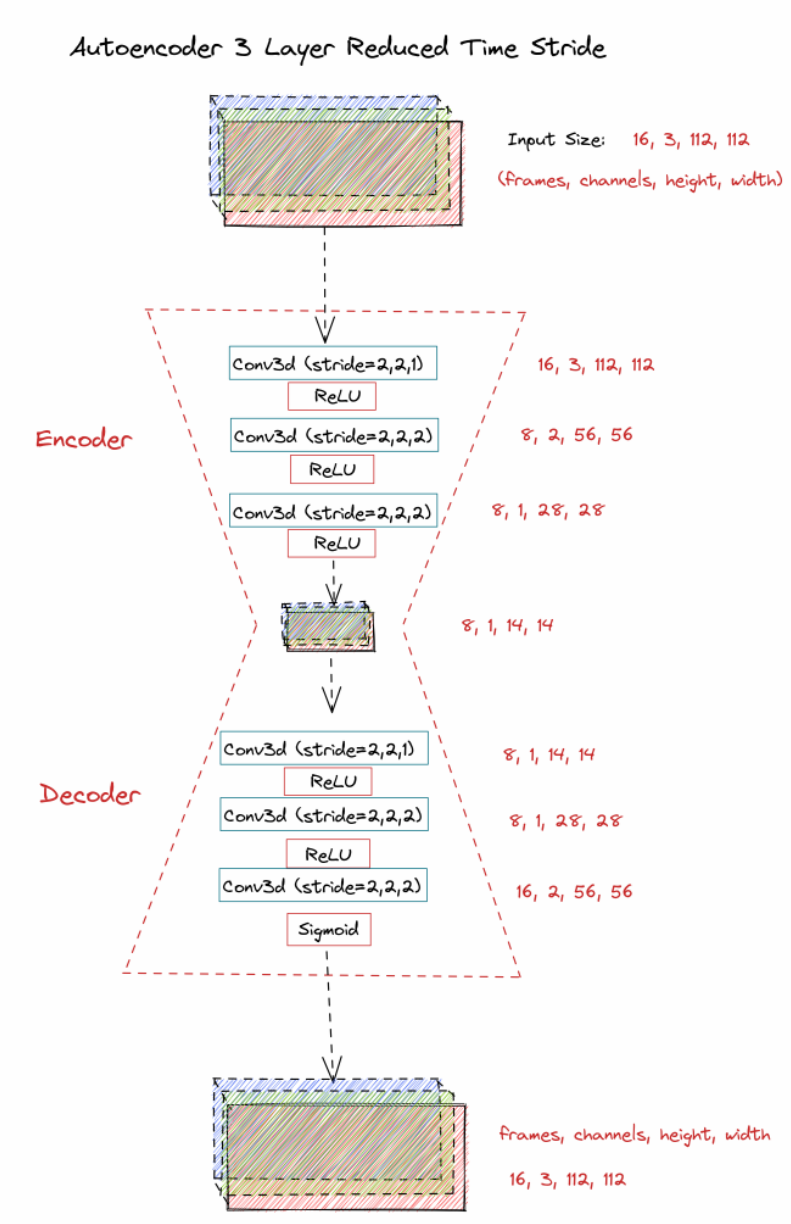
Ο encoder εκτελεί 3 φορές 3D συνελίξεις με stride 2, επομένως για 3 αλλεπάλληλες φορές το μέγεθος θα υποδιπλασιαστεί. Άρα το μέγεθος της κωδικοποιημένης πληροφορίας είναι:

Με τη συγκεκριμένη αρχιτεκτονική λοιπόν, από 602112 που αποτελεί τον αρχικό όγκο της πληροφορίας του βίντεο, μετατρέπεται σε 1176 δηλαδή 512 φορές μικρότερο.

3-Layer Autoencoder Reduced Time Stride

Ως $5^{η}$ αρχιτεκτονική, επιλέχθηκε να δοκιμαστεί διαφορετικός βαθμός συμπίεσης στη χρονική και τη χωρική διάσταση. Συγκεκριμένα δοκιμάστηκε να μειωθεί η συμπίεση στη χρονική διάσταση και να εφαρμοστεί αυξημένη συμπίεση στη χωρική διάσταση. Η ιδέα αυτή πραγματοποιήθηκε γιατί θεωρήθηκε ότι η χρονική διάσταση είναι πιο σημαντική και διαθέτει μικρότερο μέγεθος, επομένως η διατήρηση της μπορεί να

έχει αξία. Η αρχιτεκτονική φαίνεται στο σχήμα 3.10. Η χρονική διάσταση υποδιπλασιάζεται 2 φορές κατά τη συμπίεση των δεδομένων, ενώ η χωρική υποδιπλασιάζεται 3 φορές.

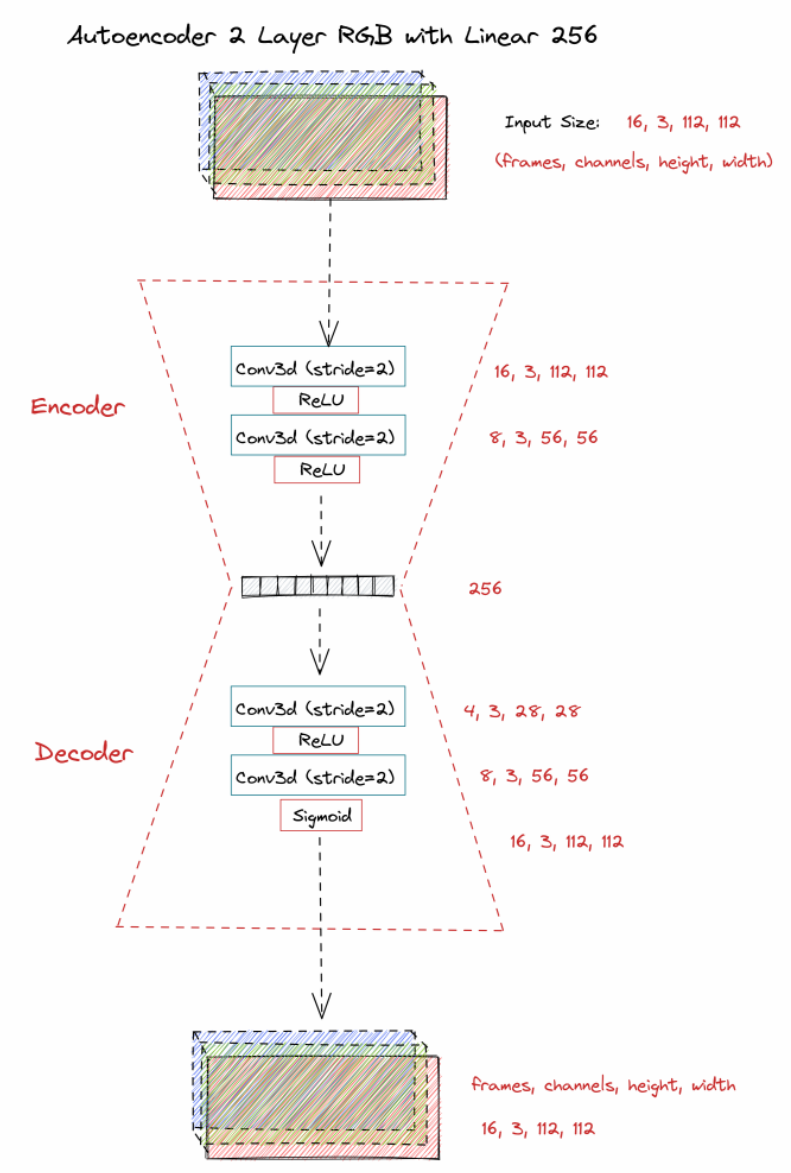


Σχήμα 3.10: Γραφική αναπαράσταση της αρχιτεκτονικής '3 Layer Autoencoder RGB'

Ο encoder εκτελεί 3 φορές 3D συνελίξεις με stride 2 στο ύψος και το πλάτος σε όλες τα επίπεδα ενώ στη διάσταση του χρόνου μόνο στο πρώτο επίπεδο.

2-Layer Autoencoder RGB - 256 Linear

Στη συνέχεια πραγματοποιήθηκε ένα διαφορετικό πείραμα, το οποίο κωδικοποιεί την πληροφορία σε μια διάσταση, και συγκεκριμένα σε μόλις 256 στοιχεία, με τη χρήση γραμμικού επιπέδου. Η αρχιτεκτονική φαίνεται στο σχήμα 3.11.

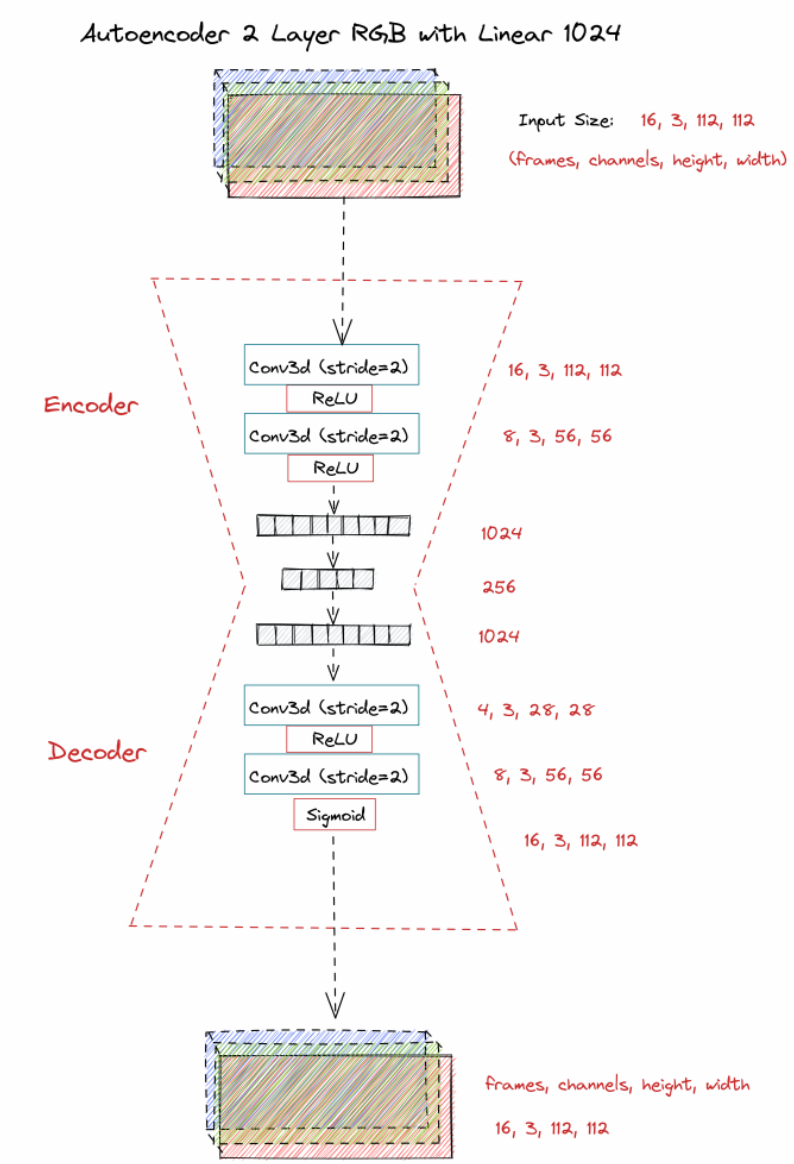


Σχήμα 3.11: Γραφική αναπαράσταση της αρχιτεκτονικής '2-Layer Autoencoder RGB - 256 Linear'

Ο encoder εκτελεί 2 φορές 3D συνελίξεις με stride 2, ενώ τελικά κωδικοποιεί την πληροφορία αυτή σε ένα μονοδιάστατο διάνυσμα 265 διαστάσεων.

2-Layer Autoencoder RGB - 1024-256 Linear

Τέλος πραγματοποιήθηκε ένα παρόμοιο πείραμα με διαφορετική χρήση γραμμικών layers. Συγκεκριμένα, είναι ίδιο με το προηγούμενο αλλά κωδικοποιεί πρώτα την πληροφορία σε μονοδιάστατο διάνυσμα με 1024 διαστάσεις, και έπειτα ξανά σε 256. Η αρχιτεκτονική φαίνεται στο σχήμα 3.12.



Σχήμα 3.12: Γραφική αναπαράσταση της αρχιτεκτονικής '2-Layer Autoencoder RGB - 1024-256 Linear'

3.5.3 Μοντέλα Ταξινόμησης των κωδικοποιημένων βίντεο

Οι συμπιεσμένεας κωδικοποιήσεις των βίντεο που προκύπτουν απ' τους αυτοκωδικοποιητές, θεωρείται ότι έχουν τα πιο σημαντικά χαρακτηριστικά των βίντεο, άρα στην περίπτωση αυτή οι αυτοκωδικοποιητές λειτουργούν και ως σαν εξαγωγείς χαρακτηριστικών. Για το λόγο αυτό, κατασκευάστηκαν μοντέλα τα οποία αρχικά κωδικοποιούν τις εισόδους τους με τη χρήση των βαρών που προέκυψαν κατά την εκπαίδευση των αυτοκωδικοποιητών.

Τα μοντέλα αυτά λοιπόν χωρίζονται σε 3 μέρη:

1. Το τμήμα το οποίο εξάγει τα κωδικοποιημένα χαρακτηριστικά των βίντεο.
2. Εφόσον αυτό είναι αναγκαίο, μετασχηματισμός των δεδομένων ώστε να παραμείνει μία μόνο διάσταση (flattening).

3. Το τελευταίο τμήμα ταξινομεί το διάνυσμα των χαρακτηριστικών με έναν γραμμικό νευρώνα στις 5 κατηγορίες του πειράματος FST.

Για λόγους απλότητας, για όλους τους αυτο-κωδικοποιητές που εκπαιδεύτηκαν, χρησιμοποιήθηκε αυτός ο απλός ταξινομητής που περιγράφεται παραπάνω.

3.6 Αξιολόγηση Μοντέλων

Όπως και σε κάθε διαδικασία επιβλεπόμενης ταξινόμησης, το dataset χωρίστηκε σε 2 επιμέρους sets, τα λεγόμενα training και validation set. Έτσι, το dataset χωρίστηκε σε 2 υποσύνολα, με απολύτως τυχαίο τρόπο. Ο διαχωρισμός έγινε σε επίπεδο βίντεο. Από τα 100 βίντεο τα 20 επιλέχτηκαν για το υποσύνολο validation και τα υπόλοιπα 80 ως train. Δεν θεωρήθηκε αναγκαίο για τις ανάγκες τις διπλωματικής, η ύπαρξη ενός τρίτου υποσυνόλου test, καθώς δεν έγινε λεπτομερή προσαρμογή των υπερπαραμέτρων, επομένως δεν κρίθηκε αναγκαίο.

Οι δείκτες που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων είναι:

- Πίνακας σύγχυσης (Confusion Matrix): Πρόκειται για πίνακα διαστάσεων $k * k$, όπου k ο αριθμός των κλάσεων της ταξινόμησης. Σε κάθε στοιχείο του πίνακα f_{ij}, συμπληρώνεται το πλήθος των δειγμάτων που εκτιμήθηκαν απ΄ το μοντέλο ως κατηγορία i, με πραγματική κατηγορία του δείγματος j. Η διαγώνιος $i = j$ αναφέρει τις πετυχημένες εκτιμήσεις του μοντέλου, ενώ πάνω απ' τη διαγώνιο αναφέρονται οι False Negative προβλέψεις, και κάτω οι False Positive.
- Ευστοχία(Accuracy): $\frac{TP+TN}{TP+TN+FP+FN}$
- Ανάκληση (Recall): $\frac{TP}{TP+FN}$
- Ακρίβεια (Precision): $\frac{TP}{TP+FP}$
- F-Score: $\frac{2*TP}{2*TP+FP+FN}$

Ως κριτήρια για την ποιότητα των autoencoders που εκπαιδεύτηκαν τέθηκαν τα εξής:

- Η τιμή της loss function.
- Η οπτική και εμπειρική αξιολόγηση των βίντεο που προκύπτουν απ' τον decoder.
- Το αποτέλεσμα του encoder στην ταξινόμηση του FST.

Όλοι οι αυτοκωδικοποιητές χρησιμοποίησαν σαν συνάρτηση κόστους το άθροισμα του μέσου τετραγωνικού σφάλματος για κάθε pixel.

3.7 Υλικό και λογισμικό για την εκπαίδευση των δικτύων

3.7.1 Υλικό

Για την εκπαίδευση νευρωνικών δικτύων, και ιδιαίτερα για την ταξινόμηση βίντεο, απαιτούνται εξειδικευμένες κάρτες γραφικών. Για τις ανάγκες τις συγκεκριμένης εργασίας χρησιμοποιήθηκε μηχάνημα με την κάρτα γραφικών 'NVIDIA Tesla V100' με χωρητικότητά $32GB$. Ο επεξεργαστής του μηχανήματος ήταν ο 'Intel Xeon' με 32

πυρήνες και χρονισμό στα $4.4GHz$ και η RAM χωρητικότητας $64GB$. Τα χαρακτηριστικά αυτά έκαναν εφικτή την εκπαίδευση δικτύων με τεράστιο αριθμό παραμέτρων της τάξης πολλών και με κανονικά μεγέθη εισόδου τροφοδότησης τους. Το μηχάνημα αυτό χρησιμοποιήθηκε με απομακρυσμένη σύνδεση με το πρωτόκολλο SSH.

3.7.2 Λογισμικό

Η ανάπτυξη και η χρήση του λογισμικού για την εκπαίδευση των νευρωνικών δικτύων με σκοπό την ταξινόμηση των βίντεο, δημιούργησε μια σειρά αναγκών που λύθηκαν με τη χρήση βιβλιοθηκών και εργαλείων διαφόρων ειδών. Τα εργαλεία αυτά παρουσιάζονται παρακάτω χωρισμένα στις εξής κατηγορίες:

Δομικά στοιχεία ανάπτυξης λογισμικού

- python: Χρησιμοποιήθηκε η γλώσσα python και συγκεκριμένα η έκδοση 3.9.
- git: Έλεγχος εκδόσεων και συνεργατική ανάπτυξη του κώδικα με τη χρήση του git. Αποθήκευση σε απομακρυσμένο server στο github. Το λογισμικό μπορεί να βρεθεί στο link `https://github.com/alexVyth/video-classification-trainer`.

Εργαλεία παραγωγής περιβάλλοντος

Για την δυνατότητα εκπαίδευση σε διαφορετικά περιβάλλοντα με διαφορετικά λειτουργικά συστήματα και συμβατότητα λογισμικών, χρησιμοποιήθηκαν τα εξής εργαλεία:

- pyenv: Καθιστά δυνατή την εγκατάσταση και συντήρηση πολλαπλών εκδόσεων python ανεξαρτήτως του συστήματος.
- poetry: Διευκολύνει την καταγραφή και την αναβάθμιση των εξαρτήσεων του λογισμικού σε python packages.
- Docker: με τη χρήση του λογισμικού Docker έγινε δυνατή η δημιουργία ανεξάρτητων απομονομένων περιβάλλοντων σε διαφορετικά μηχανήματα και ελαχιστοποίησε τον χρόνο για την εγκατάσταση και χρήση του project.

Εργαλεία για την εκπαίδευση νευρωνικών δικτύων

- pytorch: Μια απ' τις δημοφιλέστερες επιλογές για εκπαίδευση νευρωνικών δικτύων σε CPUs και GPUs.
- torchvision: Συμπληρωματική βιβλιοθήκη της pytorch για εργασίες όρασης υπολογιστών.
- pytorch-lightning: high level διεπαφή για την απλούστευση του λογισμικού pytorch και την ανεξαρτητοποίηση των βημάτων της εκπαίδευσης νευρωνικών δικτύων.
- ffmpeg, av, opencv, Pillow: Επεξεργασία εικόνων και βίντεο στην φάση της προεπεξεργασίας των δεδομένων καθώς και σε μετασχηματισμούς πριν την τροφοδότηση τους στο νευρωνικό δίκτυο
- Docker: με τη χρήση του λογισμικού Docker έγινε δυνατή η δημιουργία ανεξάρτητων απομονομένων περιβάλλοντων σε διαφορετικά μηχανήματα και ελαχιστοποίησε τον χρόνο για την εγκατάσταση και χρήση του project.

Εργαλεία για την καταγραφή των πειραμάτων και των αποτελεσμάτων τους

- mlflow: Για την παρακολούθηση και καταγραφή των πειραμάτων και της πορείας τους, χρησιμοποιήθηκε η βιβλιοθήκη mlflow.
- torchmetrics: Χρησιμοποιήθηκε για τον ορισμό και υπολογισμό των μετρικών που χρησιμοποιήθηκαν κατά την εκπαίδευση μοντέλων ανά εποχή εκπαίδευσης.

Κεφάλαιο 4

Αποτελέσματα

4.1 Πειράματα τεχνολογιών αιχμής με προ-εκπαιδευμένα βάρη

4.1.1 Μοντέλο ResNet 2+1D

Το δίκτυο αυτό ήταν το πρώτο που εκπαιδεύτηκε. Αφού εντοπίστηκαν και διορθώθηκαν ορισμένα λάθη στην λογική του κώδικα για την εκπαίδευση, πραγματοποιήθηκαν και καταγράφηκαν 4 πειράματα με αυτό το μοντέλο. Οι δοκιμές στις υπερπαραμέτρους του μοντέλου που πραγματοποιήθηκαν ήταν:

- Με χρήση παγωμένων βαρών εκτός του τελευταίου layer και χωρίς το πάγωμα κανενός layer.
- Με χρονική διάρκεια των δειγμάτων $0.65sec$ και $1.3sec$

Αρχικά εκπαιδεύτηκε το μοντέλο με:

- παγωμένα βάρη
- χρονική διάρκεια των δειγμάτων: $0.65sec$

Η εκπαίδευση αυτή είχε διάρκεια περίπου 6 ώρες για 50 εποχές, και έφερε accuracy 70%. Ωστόσο, παρατηρήθηκε ότι το training accuracy στο μεγαλύτερο μέρος της διαδικασίας ήταν μικρότερο απ' το validation accuracy και για το λόγο αυτό το δεύτερο πείραμα πραγματοποιήθηκε χωρίς το πάγωμα των βαρών δηλαδή:

- χωρίς παγωμένα βάρη
- χρονική διάρκεια των δειγμάτων: $0.65sec$

Το αποτέλεσμα που έφερε αυτή το μοντέλο είχε accuracy 74% δηλαδή αρκετά σημαντική βελτίωση. Προφανώς η μεταβολή των βαρών του τελευταίου layer, δηλαδή του ταξινομητή του μοντέλου, δεν ήταν ικανή ώστε το συνολικό μοντέλο να προσπαρμοστεί στο πρόβλημα του FST.

Η τελευταία δοκιμή αφορούσε την αύξηση της διάρκειας του χρονικού διαστήματος κάθε δείγματος, καθώς υπήρξε η σκέψη ότι τα 0.65 δευτερόλεπτα, είναι πολύ μικρό ακόμη και για τον άνθρωπο ώστε να εκτιμήσει την συμπεριφορά των επιμυών κατά το πείραμα FST. Επομένως αυτή τη φορά οι υπερπαράμετροι ήταν:

- χωρίς παγωμένα βάρη
- χρονική διάρκεια των δειγμάτων: $1.3sec$

Πράγματι αυτή η αλλαγή βοήθησε και έφερε ως αποτέλεσμα accuracy της τάξης του 79%, που αποτελεί και το καλύτερο μοντέλο που επιτεύχθη για την αρχιτεκτινική ResNet 2+1D.

4.1.2 Μοντέλο MViTv2

Το δίκτυο MViTv2 απ' την πρώτη κιόλας εποχή χάρη στα προ-εκπαιδευμένα βάρη πέτυχε accuracy 83% το οποίο ισοδυναμεί με το καλύτερο αποτέλεσμα της προηγούμενης διπλωματικής, το οποίο όμως είχε προκύψει με εικόνες οπτικής ροής, άρα θεωρείται επιτυχία το γεγονός ότι επιτεύχθηκε η ίδια ευστοχία, προσπερνώντας το χρονοβόρο βήμα της εξαγωγής εικόνων οπτικής ροής.

Έγινε δοκιμή και για εκπαίδευση με παγωμένα βάρη, με εξαίρεση το τελευταίο fully-connected layer του δικτύου, αλλά αυτό έφερε accuracy μόλις 75% οπότε δεν βοήθησε.

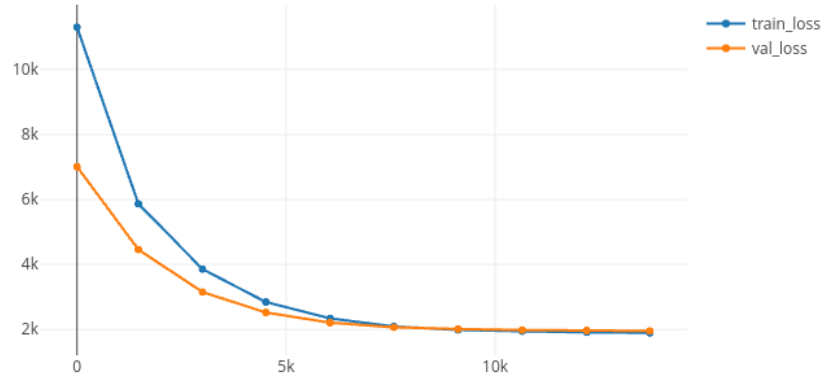
Το μοντέλο αυτό λοιπόν, με accuracy 83% αποτέλεσε το μοντέλο βάσης για τη σύγκριση με τα χειροκίνητα κατασκευασμένα δίκτυα τα οποία δοκιμάστηκαν στην πορεία στα οποία αξιοποιήθηκαν τεχνικές με χρήση αυτοκωδικοποιητών.

4.2 Πειράματα Αυτο-κωδικοποιητών

Εκτελέστηκε σειρά από εκπαιδεύσεις μοντέλων για την κωδικοποίηση των βίντεο, που έχει ως αποτέλεσμα την εξαγωγή των σημαντικών χαρακτηριστικών τους. Η εκτέλεση των πειραμάτων αυτών έγινε στο σύνολο των δεδομένων και όχι μόνο στα ταξινομημένα βίντεο, και η διάρκεια λόγου του πολύ μεγαλύτερου όγκου των δεδομένων για κάθε εκπαίδευση ήταν περίπου 1 μέρα.

Εκπαίδευση 2-Layer Autoencoder

Η αρχιτεκτονική αυτή έφερε ως αποτέλεσμα loss της τάξης του 1900. Όπως φαίνεται και στο διάγραμμα 4.1. Η εκπαίδευση ήταν αρκετά εύκολη, καθώς σε λίγες μόλις εποχές, περίπου 5, είχε επιτευχθεί το ελάχιστο loss και στα 2 υποσύνολα. Σημειώνεται ότι το validation loss είναι μικρότερο του training loss και εκτιμήθηκε ότι αυτό συμβαίνει λόγω του augmentation που συμβαίνει στο training και δημιουργεί αυτή τη μικρή διαφορά δυσκολεύοντας ελάχιστα την μάθηση.



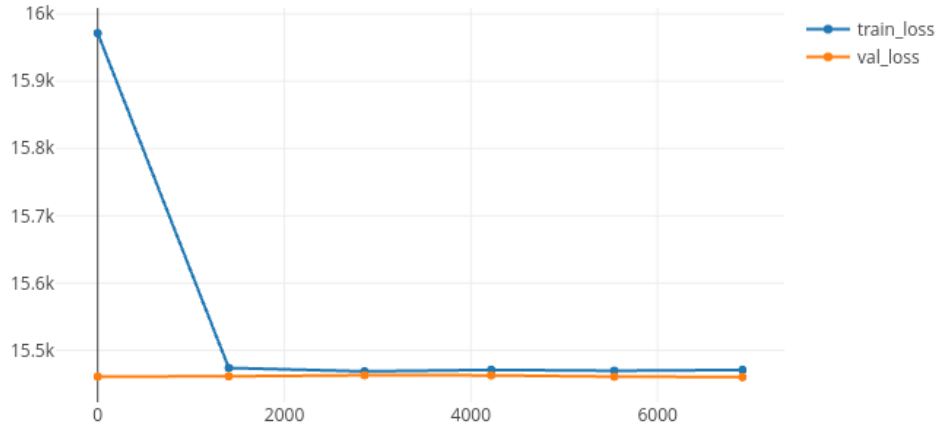
Σχήμα 4.1: Διαγράμματα της συνάρτησης κόστους για τα υποσύνολα train και validation του 2-layer autoencoder. Ως σημεία αποτυπώνονται οι εποχές, καθώς στον άξονα y βρίσκονται τα βήματα (batches) της εκπαίδευσης.

Για την καλύτερη κατανόηση της διαδικασίας, έγινε εξαγωγή και του βίντεο που εξάγεται απ' τον autoencoder, αφού δηλαδή πραγματοποιηθεί και το decoding. Το αποτέλεσμα παρουσιάζεται στην εικόνα 4.2

Σχήμα 4.2: Εικόνα αποτελεσμάτων κωδικοποίησης του 2-layer autoencoder. Αριστερά ένα τυχαίο στιγμιότυπο του βίντεο και δεξιά η εικόνα μετά την κωδικοποίηση και αποκωδικοποίηση.

Εκπαίδευση 3-Layer Autoencoder

Η αρχιτεκτονική αυτή έφερε ως αποτέλεσμα loss της τάξης του 15400. Στο διάγραμμα 4.3 παρουσιάζεται η πορεία της εκπαίδευσης. Το loss είναι μεγαλύτερο κατά $13k$ απ τον 2-layer autoencoder, γεγονός που δικαιολογείται καθώς η συμπίεση είναι πολύ μεγαλύτερη σε αυτή την περίπτωση.



Σχήμα 4.3: Διαγράμματα της συνάρτησης κόστους για τα υποσύνολα train και validation του $2^{ου}$ autoencoder. Ως σημεία αποτυπώνονται οι εποχές, καθώς στον άξονα y βρίσκονται τα βήματα (batches) της εκπαίδευσης.

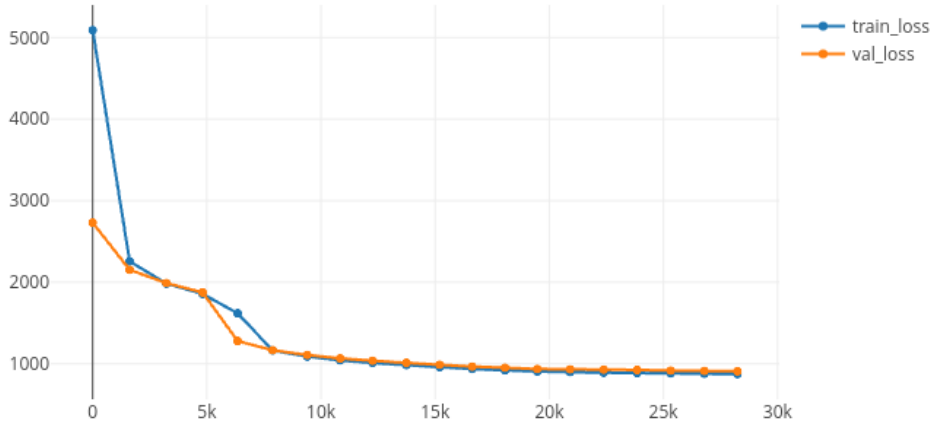
Ωστόσο βλέποντας το αποτέλεσμα στην εικόνα 4.4, η έξοδος του δικτύου δεν κατάφερε να εκφράσει την είσοδο ικανοποιητικά, στη περίπτωση αυτή. Και αυτό μάλλον συμβαίνει λόγω της μεγάλης συμπίεσης των δεδομένων.

Σχήμα 4.4: Εικόνα αποτελεσμάτων κωδικοποίησης του 3-layer autoencoder. Αριστερά ένα τυχαίο στιγμιότυπο του βίντεο και δεξιά η εικόνα μετά την κωδικοποίηση και αποκωδικοποίηση.

Εκπαίδευση 2-Layer Autoencoder RGB

Η αρχιτεκτονική αυτή έφερε ως αποτέλεσμα loss της τάξης του 1000. Στο διάγραμμα 4.5 παρουσιάζεται η πορεία της εκπαίδευσης. Το loss είναι 2 φορές μικρότερο

απ' τον 2-layer autoencoder, γεγονός που δείχνει ότι η διατήρηση των χρωμάτων είναι σημαντική για την σωστή συμπίεση του βίντεο χωρίς μεγάλη απώλεια πληροφορίας.



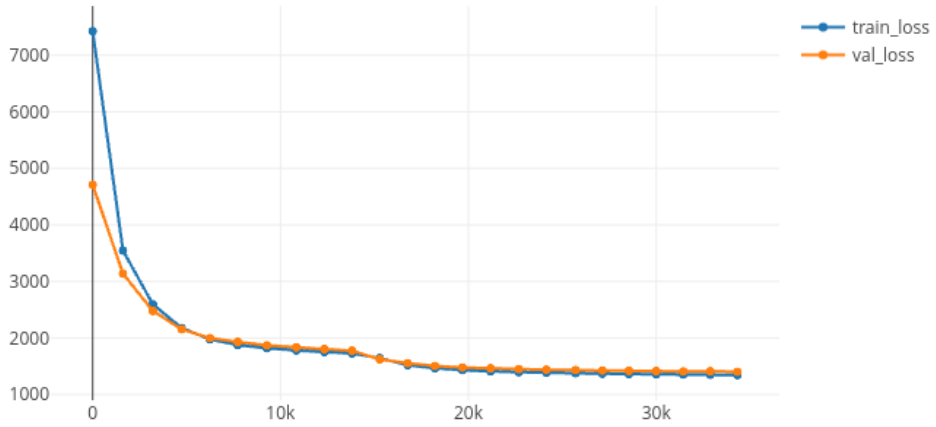
Σχήμα 4.5: Διαγράμματα της συνάρτησης κόστους για τα υποσύνολα train και validation του $3^{ου}$ autoencoder. Ως σημεία αποτυπώνονται οι εποχές, καθώς στον άξονα y βρίσκονται τα βήματα (batches) της εκπαίδευσης.

Παρατηρώντας το αποτέλεσμα, η έξοδος του δικτύου κατάφερε να εκφράσει την είσοδο ικανοποιητικά. Η διαφορά λόγω της διατήρησης των 3 καναλιών είναι προφανής στο αποτέλεσμα, καθώς τα χρώματα είναι ρεαλιστικά σε αυτή την περίπτωση, πολύ παρόμοια με τα αρχικά δείγματα.

Σχήμα 4.6: Εικόνα αποτελεσμάτων κωδικοποίησης του '2-layer RGB Autoencoder'. Αριστερά ένα τυχαίο στιγμιότυπο του βίντεο και δεξιά η εικόνα μετά την κωδικοποίηση και αποκωδικοποίηση.

Εκπαίδευση 3-Layer Autoencoder RGB

Η αρχιτεκτονική αυτή έφερε ως αποτέλεσμα loss της τάξης του 1400. Στο διάγραμμα 4.7 παρουσιάζεται η πορεία της εκπαίδευσης. Το loss είναι λίγο μεγαλύτερο απ' τον 2-layer autoencoder RGB. Αυτό δείχνει ότι η διατήρηση των 3 καναλιών στη συμπίεση των δεδομένων επηρεάζει ιδιαίτερα το loss της εκπαίδευσης, περισσότερο από την συμπίεση των χρονικών και χαρακτηριστικών. Η αξιολόγηση θα γίνει περαιτέρω κατανοητή με την σύγκριση της ευστοχίας της ταξινόμησης των κατηγοριών με τη χρήση των διαφορετικών κωδικοποιήσεων που δημιουργήθηκαν.



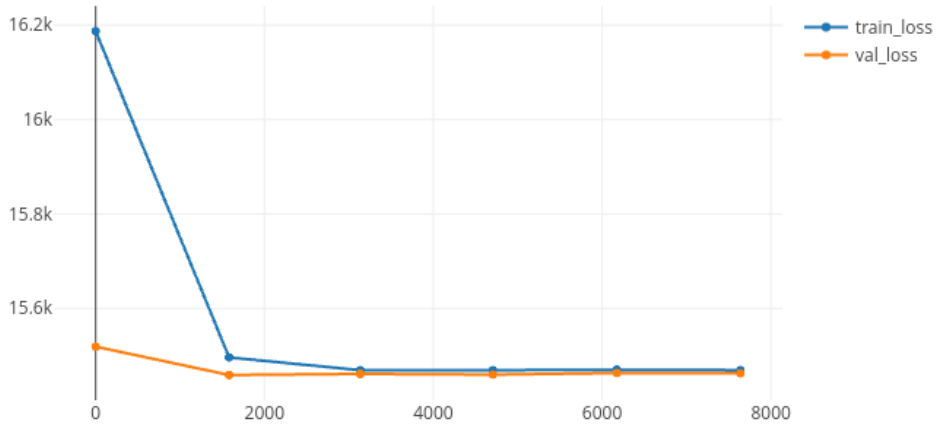
Σχήμα 4.7: Διαγράμματα της συνάρτησης κόστους για τα υποσύνολα train και validation του $4^{ου}$ autoencoder. Ως σημεία αποτυπώνονται οι εποχές, καθώς στον άξονα y βρίσκονται τα βήματα (batches) της εκπαίδευσης.

Παρατηρώντας το αποτέλεσμα, η έξοδος του δικτύου κατάφερε να εκφράσει την είσοδο ικανοποιητικά. Οι εικόνες εξόδου όπως ήταν αναμενόμενο είναι λίγο πιο θολές σε σχέση με αυτές του δικτύου '2 Layer RGB'. Στιγμιότυπο της σύγκρισης παρουσιάζεται στην εικόνα 4.8

Σχήμα 4.8: Εικόνα αποτελεσμάτων κωδικοποίησης του '3-layer RGB Autoencoder'. Αριστερά ένα τυχαίο στιγμιότυπο του βίντεο και δεξιά η εικόνα μετά την κωδικοποίηση και αποκωδικοποίηση.

Εκπαίδευση 3-Layer Reduced Time Stride

Η αρχιτεκτονική αυτή έφερε ως αποτέλεσμα loss της τάξης του 15000. Στο διάγραμμα 4.9 παρουσιάζεται η πορεία της εκπαίδευσης. Το loss είναι λίγο αντίστοιχο με την εκπαίδευση του τον 3-layer autoencoder. Ως συμπέρασμα προέκυψε ότι η ιδέα της διατήρησης του χρόνου δεν ήταν πετυχημένη, καθώς πρόκειται για το δεύτερο χειρότερο loss.



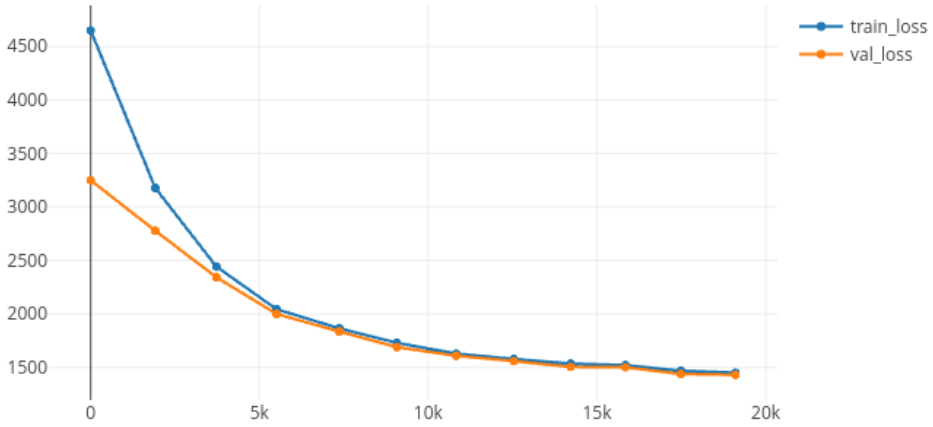
Σχήμα 4.9: Διαγράμματα της συνάρτησης κόστους για τα υποσύνολα train και validation του $5^{ου}$ autoencoder. Ως σημεία αποτυπώνονται οι εποχές, καθώς στον άξονα y βρίσκονται τα βήματα (batches) της εκπαίδευσης.

Παρατηρώντας το αποτέλεσμα, η μετάβαση στην κίνηση είναι πιο ομαλή κατά το πέρας του χρόνου ωστόσο η ευκρίνεια είναι μικρή λόγω της μεγάλης συμπίεσης των χωρικών δεδομένων. Το αποτέλεσμα παρουσιάζεται στην εικόνα 4.10

Σχήμα 4.10: Εικόνα αποτελεσμάτων κωδικοποίησης του '3-layer Reduced Time Stride Autoencoder'. Αριστερά ένα τυχαίο στιγμιότυπο του βίντεο και δεξιά η εικόνα μετά την κωδικοποίηση και αποκωδικοποίηση.

Εκπαίδευση 2 Layer RGB Linear 256

Το loss που επιτεύχθη είναι της τάξης του 1200 το οποίο φάνηκε ενθαρρυντικό, καθώς είναι κοντά στο καλύτερο, ενώ έχει δημιουργηθεί πολύ μεγαλύτερη κωδικοποίηση σε σχέση με τα υπόλοιπα μοντέλα.



/

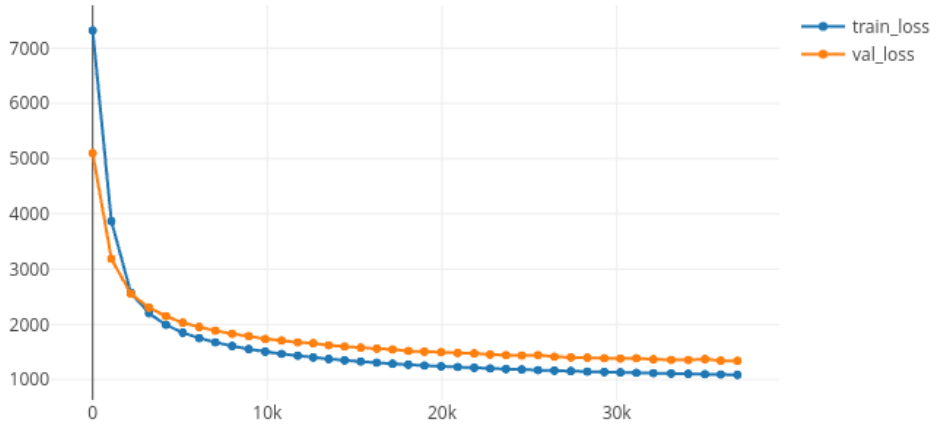
Σχήμα 4.11: Διαγράμματα της συνάρτησης κόστους για τα υποσύνολα train και validation του $6^{ου}$ autoencoder. Ως σημεία αποτυπώνονται οι εποχές, καθώς στον άξονα y βρίσκονται τα βήματα (batches) της εκπαίδευσης.

Παρατηρώντας το αποτέλεσμα, φαίνεται αρκετά ικανοποιητικό, ιδιαίτερα για το μέγεθος της συμπίεσης που υπέστη το δείγμα. Παράδειγμα της εξόδου παρουσιάζεται στο σχήμα 4.12

Σχήμα 4.12: Εικόνα αποτελεσμάτων κωδικοποίησης του '2-Layer RGB - 256 Linear'. Αριστερά ένα τυχαίο στιγμιότυπο του βίντεο και δεξιά η εικόνα μετά την κωδικοποίηση και αποκωδικοποίηση.

Εκπαίδευση 2 Layer RGB Linear 1024-256

Το loss που επιτεύχθη για την αρχιτεκτοινκή αυτή είναι της τάξης του 1000, επομένως είναι ισοδύναμο με την καλύτερη ως τώρα αρχιτεκτονική, την '2-Layer RGB Autoencoder'. Η διαδικασία της εκπαίδευσης παρουσιάζεται στο σχήμα 4.13



Σχήμα 4.13: Διαγράμματα της συνάρτησης κόστους για τα υποσύνολα train και validation του 7^{ου} autoencoder. Ως σημεία αποτυπώνονται οι εποχές, καθώς στον άξονα y βρίσκονται τα βήματα (batches) της εκπαίδευσης.

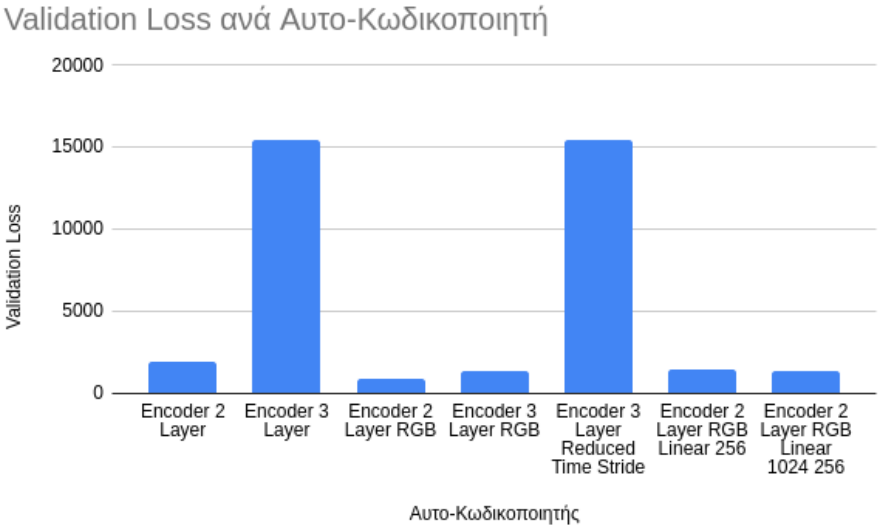
Παρατηρώντας το αποτέλεσμα, φαίνεται αντίστοιχο με αυτό της προηγούμενης

εκπαίδευσης, αρκετά ικανοποιητικό σε σχέση με το βαθμό της συμπίεσης της πληροφορίας των δειγμάτων. 4.14

Σχήμα 4.14: Εικόνα αποτελεσμάτων κωδικοποίησης του '2-Layer RGB - 1024-256 Linear'. Αριστερά ένα τυχαίο στιγμιότυπο του βίντεο και δεξιά η εικόνα μετά την κωδικοποίηση και αποκωδικοποίηση.

4.2.1 Σύγκριση των μοντέλων αυτοκωδικοποιητών

Στην ενότητα αυτή θα συγκριθούν τα αποτελέσματα των αυτοκωδικοποιητών με τη χρήση του loss που προέκυψε για το καθένα. Το loss χαρακτηρίζει την ικανότητα του κάθε μοντέλου να συμπιέσει το βίντεο, διατηρώντας τα πιο σημαντικά του χαρακτηριστικά, με την μικρότερη απώλεια πληροφορίας. Τα αποτελέσματα παρουσιάζονται στο σχήμα 4.15

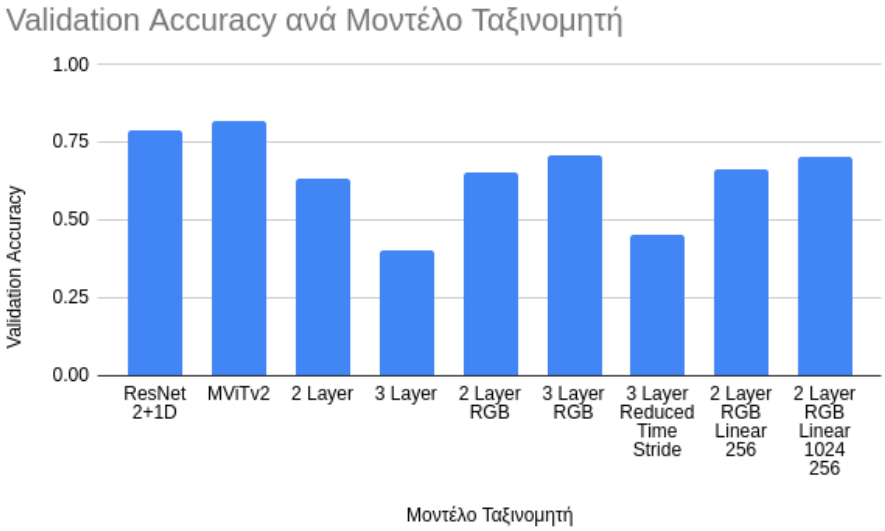


Σχήμα 4.15: Διάγραμμα του Validation loss ανά Μοντέλο Αυτο-κωδικοποιητή.

Όπως φαίνεται στο σχήμα 4.15 ο βαθμός της συμπίεσης των αυτο-κωδικοποιητών επηρεάζει άμεσα το loss του καθενός.

4.3 Σύγκριση πειραμάτων ταξινόμησης

Στην ενότητα αυτή θα συγκριθούν τα 2 καλύτερα μοντέλα που έφεραν οι εκπαιδεύσεις των δικτύων ResNet 2+1D και MViTv2, καθώς και αυτά που προέκυψαν από τις εκπαιδεύσεις των μοντέλων ταξινόμησης της κωδικοποιημένης πληροφορίας των αυτοκωδικοποιτών.



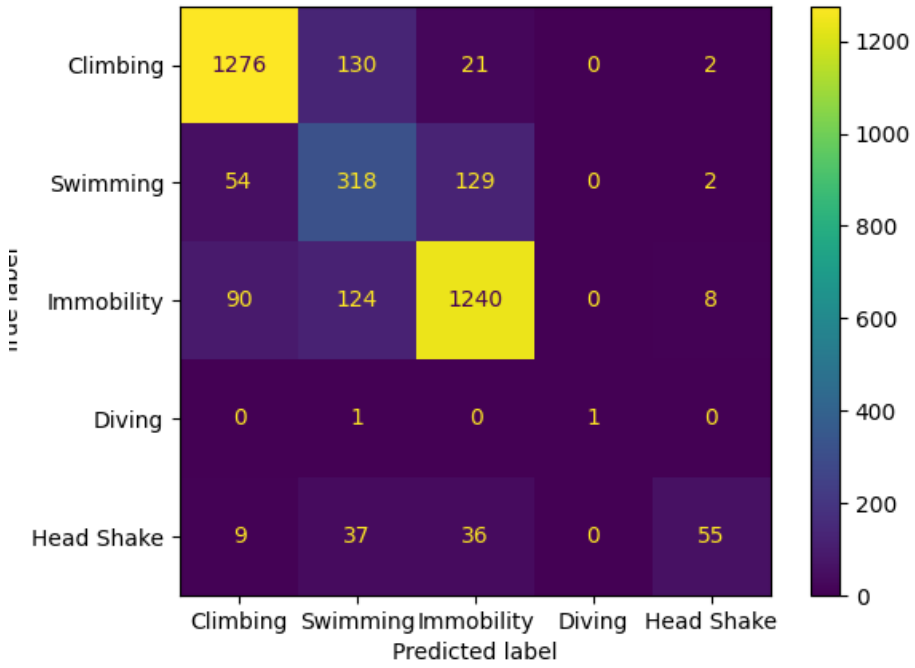
Σχήμα 4.16: Διάγραμμα του Validation Accuracy ανά Μοντέλο ταξινομητή.

Για την κατανόηση των αποτελεσμάτων δημιουργήθηκε το διάγραμμα του σχήματος 4.16. Στο σχήμα φαίνεται η ξεκάθαρη υπεροχή των μοντέλων τεχνολογιών αιχμής. Το καλύτερο μοντέλο είναι το MViTv2 με accuracy 83%.

Ως προς τα μοντέλα των αυτο-κωδικοποιητών, παρατηρείται συσχέτιση του loss των αυτοκωδικοποιητών με το accuracy των ταξινομητών τους. Εξαίρεση αποτελεί το '3 Layer RGB' το οποίο είναι το καλύτερο μοντέλο εξ' αυτών, ενώ περιλαμβάνει μεγάλη συμπίεση πληροφορίας. Φαίνεται λοιπόν ότι η αρχιτεκτονική λαμβάνει σημαντικό ρόλο και όχι μόνο ο βαθμός της συμπίεσης της πληροφορίας.

4.4 Εμβάθυνση στο μοντέλο MViT2

Όπως προέκυψε, το μοντέλο MViT2 είναι το πιο πετυχημένο με accuracy 82.5%. Για το λόγο αυτό, γίνεται περαιτέρω παρουσίαση των αποτελεσμάτων του.



Σχήμα 4.17: Πίνακας σύγχυσης του μοντέλου MViTv2.

Συγκεκριμένα, στο σχήμα 4.17 φαίνεται ο πίνακας σύγχυσης του μοντέλου.

Επιπλέον στον παρακάτω πίνακα παρουσιάζονται τα στατιστικά στοιχεία του μοντέλου MViTv2 στα δεδομένα επαλήθευσης του συνόλου δεδομένων.

	Precision	**Recall**	**F1-Score**	**Support**
Class Statistics				
Immobility	0.87	0.87	0.87	1426
Swimming	0.63	0.52	0.57	610
Climbing	0.89	0.89	0.89	1429
Head Shake	0.40	0.82	0.54	67
Diving	0.50	1	0.67	1
Overall Statistics				
Accuracy		0.82		3533
Macro Average	0.66	0.82	0.73	3533
Weighted Average	0.83	0.82	0.83	3533

Πίνακας 4.1: Στατιστικά στοιχεία των αποτελεσμάτων του μοντέλου MViTv2

Παρατηρήθηκε ότι οι κατηγορίες με τα περισσότερα δείγματα, Climbing και Immobility, έχουν τα καλύτερα F1 Score μεγαλύτερα από 87%, ενώ οι υπόλοιπες κατηγορίες κυμαίνονται από 54% έως 67%.

Κεφάλαιο 5

Συμπεράσματα και Μελλοντικές Επεκτάσεις

5.1 Συμπεράσματα

Στην διπλωματική αυτή, πραγματεύτηκε το πρόβλημα της ταξινόμησης της συμπεριφοράς των επιμυών, κατά τη Δοκιμασία Εξαναγκασμένης Κολύμβησης. Το πείραμα αυτό αποτελεί ένα σημαντικό εργαλείο για τη μελέτη και σχεδίαση αντικαταθλιπτικών φαρμάκων. Η ταξινόμηση γίνεται για περίπου κάθε δευτερόλεπτο του πειράματος, σε πέντε διαφορετικές κατηγορίες ενδιαφέροντος. Τη δεδομένη στιγμή, υπάρχουν εμπορικές εφαρμογές υψηλού κόστους, για την αυτοματοποίηση του προβλήματος αυτού, και συνήθως, έπειτα από καθορισμό παραμέτρων για την ευαισθησία των αλγορίθμων. Για το λόγο αυτό, η διαδικασία ταξινόμησης πολλές φορές συμβαίνει με χειροκίνητη εργασία από ειδικούς παρατηρητές.

Για την επίλυση του προβλήματος χρησιμοποιήθηκαν τόσο οι τεχνολογίες αιχμής στην ταξινόμηση βίντεο που αποτελούν τα νευρωνικά δίκτυα με τρισδιάστατες συνελίξεις και χρήση transformers με αξιοποίηση προεκπαιδευμένων βαρών, καθώς και νευρωνικά δίκτυα που σχεδιάστηκαν με προσπάθεια προσαρμογής στο συγκεκριμένα πρόβλημα, και συγκεκριμένα auto-encoders για την κωδικοποίηση των χαρακτηριστικών των βίντεο τα οποία τροφοδοτήθηκαν σε νέο νευρωνικό με έναν απλό ταξινομητή.

Σύμφωνα με τα αποτελέσματα, προέκυψε ότι οι τεχνολογίες αιχμής υπερέχουν κατά πολύ σε σχέση με τα προσαρμοσμένα δίκτυα που σχεδιάστηκαν, καθώς το δίκτυο MViTv2 πέτυχε accuracy 83% σε σύγκριση με το καλύτερο auto-encoder μοντέλο που πέτυχε accuracy 73%, δηλαδή κατά 10% υψηλότερο.

Οι αλγόριθμοι βίντεο classification είναι ειδικά σχεδιασμένοι για την ταξινόμηση συμπεριφορών σύμφωνα με το περιεχόμενο τους, και χρησιμοποιούν διάφορες τεχνικές όρασης υπολογισμών για την ανάλυση των οπτικών στοιχείων των βίντεο. Αντίθετα, οι autoencoders που εκπαιδεύτηκαν, σχημάτισαν τα χαρακτηριστικά τους με κριτήριο την συμπίεση των δεδομένων με την ελάχιστη απώλεια πληροφορίας, και όχι με κριτήριο την σωστή ταξινόμηση της συμπεριφοράς των επιμυών. Η πληροφορία που χάθηκε κατά τη φάση της κωδικοποίησης μπορεί να ήταν χρήσιμη για την αναγνώριση των κατηγοριών, όπως προκύπτει από τα αποτελέσματα. Επιπλέον, η χρήση των αταξινόμητων δεδομένων για την εκπαίδευση των αυτοκωδικοποιητών, πιθανότητα δεν ήταν αρκετά βοηθητική σε σύγκριση με τα προεκπαιδευμένα βάρη του dataset Kinetics400 που διατίθενται στους αλγόριθμους video classification.

Για την επαλήθευση αυτού του ισχυρισμού, δοκιμάστηκε η εκπαίδευση του μοντέλου MViTv2 χωρίς τη χρήση προεκπαιδευμένων βαρών. Σε αυτή τη περίπτωση το μοντέλο πέτυχε accuracy 74%, δηλαδή πάρα πολύ κοντά σε αυτή των προσαρμοσμένων δικτύων με autoencoders που κατασκευάστηκαν. Το γεγονός αυτό αναδεικνύει τη σημασία των προεπαιδευμένων βαρών που καθιστούν πολύ ισχυρότερες τις αρχιτεκτονικές που έχουν εκπαιδευτεί σε μεγάλα datasets όπως το Kinetics400.

Παρόλο που οι αυτοκωδικοποιητές εξήγαγαν με επιτυχία χαρακτηριστικά από τα βίντεο, το γεγονός ότι δεν είναι ειδικά σχεδιασμένα για την ταξινόμηση του ζητούμενου προβλήματος φαίνεται να μην τους έθεσε ως τον πιο αποτελεσματικό τρόπο για την ταξινόμηση του προβλήματος εξαναγκασμένης κολύμβησης.

Ως προς τους αυτοκωδικοποιητές που υλοποιήθηκαν, παρατηρήθηκε ότι το μέγε-

θος της συμπίεσης των δεδομένων, δεν ήταν αντιστρόφως ανάλογο με το accuracy των μοντέλων. Σε πολλές περιπτώσεις, όπως αυτή του '2 Layer RGB Linear 1024 256', εξυπνότερες αρχιτεκτονικές με μεγάλη συμπίεση, έφεραν πολύ καλύτερα αποτελέσματα από πολύ λιγότερο συμπιεσμένα δεδομένα. Αυτό αποδεικνύει τη σημασία του σωστού σχεδιασμού αρχιτεκτονικής. Φαίνεται ότι η διατήρηση καναλιών στα κωδικοποιημένα δεδομένα, όπως συνέβη στη περίπτωση των μοντέλων '2 Layer RGB', '3 Layer RGB', '2 Layer RGB Linear 256' και '2 Layer RGB Linear 1024 256', βοήθησε στην επίτευξη μεγαλύτερου accuracy. Ακόμη, η διατήρηση αποκλειστικά ενός μονοδιάστατου διανύσματος ως χαρακτηριστικά του βίντεο, δηλαδή τα μοντέλα '2 Layer Lineaer 256' και '2 Layer Linear 1024 256' έδωσαν τα καλύτερα αποτελέσματα με ευστοχία 73% και 71%.

Σχετικά με την κωδικοποίηση των βίντεο μέσω των αυτοκωδικοποιητών, προκύπτουν κάποια εύλογα συμπεράσματα τα οποία είναι:

- Τα δίκτυα με 3 layers, δηλαδή 3 υποδιπλασιασμούς της πληροφορίας και χωρίς τη διατήρηση 3 καναλιών, δηλαδή τα 'Encoder 3 Layer' και 'Encoder 3 Layer Reduced Time Stride' υστερούν κατά πολύ σε σχέση με τα υπόλοιπα μοντέλα. Φαίνεται ότι ο βαθμός αυτός συμπίεσης είναι υπερβολικός.

- Τα υπόλοιπα μοντέλα έχουν παρόμοιο loss, με σημαντική υπεροχή των RGB μοντέλων. Φαίνεται λοιπόν ότι η διατήρηση των 3 καναλιών στο κωδικοποιημένο βίντεο βοηθάει σημαντικά στην διατήρηση χαμηλού loss του αυτοκωδικοποιητή.

Ωστόσο η παραπάνω σύγκριση προκύπτει από το μέσο τετραγωνικό σφάλμα (MSE Loss) των μοντέλων που εκπαιδεύτηκαν, και δεν αφορά την μετέπειτα χρήση τους για την ταξινόμηση του προβλήματος της Εξαναγκασμένης κολύμβησης. Αφορούν αποκλειστικά την ικανότητα των μοντέλων να κωδικοποιήσουν την πληροφορία των βίντεο με την ελάχιστη απώλεια πληροφορίας.

Όσον αφορά τους αλγορίθμους ταξινόμησης, είναι φανερή η εξέλιξη τους στο χρόνο που φέρνει αναλογικές βελτιώσεις. Όπως αναφέρθηκε, η τεχνολογία αιχμής του 2017 'Inception 3D' πέτυχε 79% στο dataset, στη συνέχεια αυτή του 2019, η αρχιτεκτονική ResNet 2+1D πέτυχε accuracy 80%, ενώ τέλος η αρχιτεκτονική MvITv2, του 2021, πέτυχε 83%. Η σταδιακή βελτίωση τους είναι φανερή στο σύνολο δεδομένων που χρησιμοποιήθηκε και εκτιμάται αντίστοιχη βελτίωση και στο μέλλον.

Σαν αποτέλεσμα υλοποιήθηκε σύστημα το οποίο με την είσοδο βίντεο της Δοκιμασίας Εξαναγκασμένης Κολύμβησης, εξάγει αυτομάτως τη συμπεριφορά του επιμύ κάθε μισό δευτερόλεπτο, ταξινομώντας μικρά χρονικά διαστήματα στις πέντε κατηγορίες ενδιαφέροντος.

5.2 Μελλοντικές Επεκτάσεις

Κατά την εκπόνηση της διπλωματικής, υπήρξαν διάφορες ιδέες για την περαιτέρω βελτίωση του προβλήματος που ενδέχεται να βελτιώσουν τα υπάρχοντα αποτελέσματα. Για την ταξινόμηση των δειγμάτων, δυνατή είναι η παράλληλη τροφοδότηση του δικτύου με τον ήχο των αντίστοιχων δειγμάτων. Στην συγκεκριμένη εργασία, δεν θεωρήθηκε προτεραιότητα καθώς τα πειράματα του συνόλου δεδομένων υλοποιήθηκαν σε δυάδες, και έτσι σε κάθε βίντεο υπήρχαν ήχοι από δύο διαφορετικά πειράματα ταυτόχρονα, γεγονός που θα μπορούσε να είναι παραπλανητικό κατά την ταξινόμηση. Παρ' όλα αυτά, ακόμη και σε αυτή την περίπτωση, η συνεισφορά του ήχου στην ταξινόμηση θα μπορούσε να βελτιώσει σημαντικά την ευστοχία της ταξινόμησης, καθώς σύμφωνα με την κίνηση των επιμυών, προκύπτουν χαρακτηριστικοί ήχοι του νερού.

Όσον αφορά την συμπίεση των βίντεο, θα μπορούσαν να δοκιμαστούν τεχνικές που χρησιμοποιούνται στον τομέα του βίντεο για την κωδικοποίηση και συμπίεση των βίντεο. Οι κωδικοποιήσεις αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για την τροφοδότηση ειδικά διαμορφωμένων νευρωνικών δικτύων. Η απλότητα τον χαρακτηριστικών αυτών σε συνδυασμό με την ελάχιστη απώλεια δεδομένων ενδεχομένως θα βελτίωνε την ακρίβεια σε σχέση με την συμπίεση του βίντεο με τη χρήση των αυτοκωδικοποιητών.

Στη συγκεκριμένη διπλωματική, εξετάζονται τεχνικές για την ταξινόμηση αποκομμένων βίντεο. Αυτό σημαίνει ότι κάθε βίντεο περιλαμβάνει μόνο μία κατηγορία. Ωστόσο υπάρχει προσπάθεια για ανάπτυξη τεχνικών που διαχωρίζουν το βίντεο σε επιμέρους τμήματα με βάση τα διαφορετικά χαρακτηριστικά κάθε τμήματος, και ταξινομούν το κάθε ένα απ' αυτά. Το πρόβλημα αυτό ονομάζεται αναγνώριση δράσης συνεχών βίντεο (Untrimmed / Continuous Video Action Recognition). Η εφαρμογή του στη συγκεκριμένη εφαρμογή έχει ιδιαίτερο ερευνητικό ενδιαφέρον.

Τέλος, σημαντικός είναι ο εμπλουτισμός του dataset με πειράματα διαφορετικών εργαστηριών. Εφόσον τα βίντεο τα οποία τροφοδότησαν τα μοντέλα, προέρχονται από το ίδιο εργαστήριο, δεν μπορεί να θεωρηθεί ότι το μοντέλο έχει γενικευτεί στο πρόβλημα, αλλά στο συγκεκριμένο περιβάλλοντα χώρο του εργαστηρίου και στη συγκεκριμένη ράτσα επιμυών. Η ενίσχυση των δειγμάτων με νέα πειράματα διαφορετικής προέλευσης θα ενίσχυε αδιαμφισβήτητα τα αποτελέσματα της εκτίμησης.

Κατάλογος σχημάτων

Κατάλογος πινάκων

Βιβλιογραφία

[1] R. D. Porsolt, M. Le Pichon και M. Jalfre, «Depression: A new animal model sensitive to antidepressant treatments», *Nature*, τόμ. 266, αριθμ. 5604, σσ. 730–732, Απρ. 1977, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/266730a0. διεύθν.: http://www.nature.com/articles/266730a0.

[2] R. D. Porsolt, A. Bertin και M. Jalfre, «Behavioral despair in mice: A primary screening test for antidepressants», *Archives internationales de pharmacodynamie et de therapie*, τόμ. 229, αριθμ. 2, σσ. 327–336, 1 Οκτ. 1977, ISSN: 0003-9780. pmid: 596982.

[3] R. D. Porsolt, G. Anton, N. Blavet και M. Jalfre, «Behavioural despair in rats: A new model sensitive to antidepressant treatments», *European Journal of Pharmacology*, τόμ. 47, αριθμ. 4, σσ. 379–391, 15 Φεβ. 1978, ISSN: 0014-2999. DOI: 10.1016/0014-2999(78)90118-8. διεύθν.: https://www.sciencedirect.com/science/article/pii/0014299978901188.

[4] G. Hedou, C. Pryce, L. Di Iorio, C. Heidbreder και J. Feldon, «An Automated Analysis of Rat Behavior in the Forced Swim Test», *Pharmacology Biochemistry and Behavior*, τόμ. 70, σσ. 65–76, 1 Οκτ. 2001. DOI: 10.1016/S0091-3057(01)00575-5.

[5] H. Fröhlich, A. Hoenselaar, J. Eichner, H. Rosenbrock, G. Birk και A. Zell, «Automated classification of the behavior of rats in the forced swimming test with support vector machines», *Neural Networks*, τόμ. 21, αριθμ. 1, σσ. 92–101, 1 Ιαν. 2008, ISSN: 0893-6080. DOI: 10.1016/j.neunet.2007.09.019. διεύθν.: https://www.sciencedirect.com/science/article/pii/S0893608007002055.

[6] A. Nandi, G. Virmani, A. Barve και S. Marathe, «DBscorer: An Open-Source Software for Automated Accurate Analysis of Rodent Behavior in Forced Swim Test and Tail Suspension Test», *eNeuro*, τόμ. 8, αριθμ. 6, ENEURO.0305–21.2021, 2 Νοέ. 2021, ISSN: 2373-2822. DOI: 10.1523/ENEURO.0305-21.2021. pmid: 34625460. διεύθν.: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8570685/.

[7] H. Wang και C. Schmid, «Action Recognition with Improved Trajectories», σσ. 3551–3558, 2013. διεύθν.: https://openaccess.thecvf.com/content_iccv_2013/html/Wang_Action_Recognition_with_2013_ICCV_paper.html.

[8] K. Simonyan και A. Zisserman, «Very Deep Convolutional Networks for Large-Scale Image Recognition», αριθμ. arXiv:1409.1556, 10 Απρ. 2015. DOI: 10.48550/arXiv.1409.1556. arXiv: 1409.1556 [cs]. διεύθν.: http://arxiv.org/abs/1409.1556.

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama κ.ά., «Long-Term Recurrent Convolutional Networks for Visual Recognition and Description», σσ. 2625–2634, 2015. διεύθν.: https://openaccess.thecvf.com/content_cvpr_2015/html/Donahue_Long-Term_Recurrent_Convolutional_2015_CVPR_paper.html.

[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani και M. Paluri, «Learning Spatiotemporal Features with 3D Convolutional Networks», αρθμ. arXiv:1412.0767, 6 Οκτ. 2015. DOI: 10.48550/arXiv.1412.0767. arXiv: 1412.0767 [cs]. διεύθν.: http://arxiv.org/abs/1412.0767.

[11] J. Carreira και A. Zisserman, «Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset», σσ. 6299–6308, 2017. διεύθν.: https://openaccess.thecvf.com/content_cvpr_2017/html/Carreira_Quo_Vadis_Action_CVPR_2017_paper.html.

[12] C. Szegedy, W. Liu, Y. Jia κ.ά., «Going Deeper With Convolutions», σσ. 1–9, 2015. διεύθν.: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html.

[13] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun και M. Paluri, «A Closer Look at Spatiotemporal Convolutions for Action Recognition», έκδ. 3, αρθμ. arXiv:1711.11248, 11 Απρ. 2018. DOI: 10.48550/arXiv.1711.11248. arXiv: 1711.11248 [cs]. διεύθν.: http://arxiv.org/abs/1711.11248.

[14] C. Feichtenhofer, H. Fan, J. Malik και K. He, «SlowFast Networks for Video Recognition», σσ. 6202–6211, 2019. διεύθν.: https://openaccess.thecvf.com/content_ICCV_2019/html/Feichtenhofer_SlowFast_Networks_for_Video_Recognition_ICCV_2019_paper.html.

[15] H. Fan, B. Xiong, K. Mangalam κ.ά., «Multiscale Vision Transformers», αρθμ. arXiv:2104.11227, 22 Απρ. 2021. DOI: 10.48550/arXiv.2104.11227. arXiv: 2104.11227 [cs]. διεύθν.: http://arxiv.org/abs/2104.11227.

[16] D. Bahdanau, K. Cho και Y. Bengio, «Neural Machine Translation by Jointly Learning to Align and Translate», αρθμ. arXiv:1409.0473, 19 Μάι. 2016. DOI: 10.48550/arXiv.1409.0473. arXiv: 1409.0473 [cs, stat]. διεύθν.: http://arxiv.org/abs/1409.0473.

[17] Y. Li, C.-Y. Wu, H. Fan κ.ά., «MViTv2: Improved Multiscale Vision Transformers for Classification and Detection», αρθμ. arXiv:2112.01526, 30 Μαρ. 2022. DOI: 10.48550/arXiv.2112.01526. arXiv: 2112.01526 [cs]. διεύθν.: http://arxiv.org/abs/2112.01526.

[18] B. Zhang, L. Wang, Z. Wang, Y. Qiao και H. Wang, «Real-Time Action Recognition With Enhanced Motion Vector CNNs», σσ. 2718–2726, 2016. διεύθν.: https://openaccess.thecvf.com/content_cvpr_2016/html/Zhang_Real-Time_Action_Recognition_CVPR_2016_paper.html.

[19] C.-Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola και P. Krähenbühl, «Compressed Video Action Recognition», σσ. 6026–6035, 2018. διεύθν.: https://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Compressed_Video_Action_CVPR_2018_paper.html.

[20] Z. Shou, X. Lin, Y. Kalantidis κ.ά., «DMC-Net: Generating Discriminative Motion Cues for Fast Compressed Video Action Recognition», σσ. 1268–1277, 2019. διεύθν.: https://openaccess.thecvf.com/content_CVPR_2019/html/Shou_DMC-Net_Generating_Discriminative_Motion_Cues_for_Fast_Compressed_Video_Action_CVPR_2019_paper.html.

[21] B. Korbar, D. Tran και L. Torresani, «SCSampler: Sampling Salient Clips from Video for Efficient Action Recognition», αριθμ. arXiv:1904.04289, 30 Αύγ. 2019. DOI: 10.48550/arXiv.1904.04289. arXiv: 1904.04289 [cs]. διεύθν.: http://arxiv.org/abs/1904.04289.