Κεφάλαιο 1

Classification stage

Στα προηγούμενα 2 κεφάλαια παρουσιάσαμε την διαδικασία που χρησιμοποιούμε για να δημιουργήσουμε υποψήφια action tubes, τα οποία πιθανώς να περιέχουν κάποια πραγματοποιούμενη δράση ή μπορεί όχι. Τις περισσότερες φορές τα προτεινόμενα action tubes ανήκουν στο φόντο, και γι΄ αυτό, όπως αναφέρθηκε και στον προηγούμενο κεφάλαιο, είναι σημαντικό να επιλέξουμε έναν καλό αλγόριθμο που προτείνει καλές ακολουθίες από πλαίσια. Ωστόσο, είναι αρκετά σημαντικό να επιλέξουμε και τον κατάλληλο ταξινομητή ο οποίος θα είναι σε θέση με μεγάλη ακρίβεια να προβλέψει αν ένα υποψήφιο action tube ανήκει σε μια γνωστή κατηγορία από δράσεις ή ανήκει στο φόντο. Κι αυτό γιατί μπορεί να παράγουμε καλές προτάσεις για υποψήφιες δράσεις, αλλά αν ο ταξινομητής μας δεν λειτουργεί στο έπακρο, το σύστημα μας πάλι θα αποτυγχάνει να αναγνωρίσει τις δράσεις.

Η σωστή επιλογή ενός ταξινομητή είναι μια μεγάλη απόφαση που καλούμαστε να πάρουμε. Ωστόσο, αυτός ο ταξινομητής θα δεχθεί ορισμένους χάρτες ενεργοποίησης τους οποίους θα κληθεί να ταξινομήσει. Συνεπώς, εκτός από την καλή επιλογή ταξινομητή, εξίσου σημαντική είναι η καλή επιλογή χαρακτηριστικών. Τέλος, μεγάλο ρόλο παίζει και η διαδικασία εκπαίδευσης του ταξινομητή προκειμένου να είναι σε θέση να γενικεύει και καταστάσεις overfitting να αποφεύγονται.

Σε αυτό το κεφάλαιο παρουσιάζουμε διάφορες μεθόδους που χρησιμοποιήσαμε οι οποίες περιλαμβάνουν ένα Γραμμικό ταξινομητή, ένα Recursive Neural Network (RNN), ένα Support Vector Machine (SVM)και ένα Multilayer Perceptron(MLP). Επίσης, πειραματιζόμαστε χρησιμοποιώντας χάρτες χαρακτηριστικών που εξηχθησαν μέσω του 3D RoiAlign χρησιμοποιώντας παράλληλα avg ή max pooling. Τελευταίο αλλά εξίσου σημαντικό είναι το γεγονός ότι προσπαθήσαμε να βρούμε το καλύτερο ποσοστό μεταξύ actio tubes προσκηνίου και φόντο αλλά και τον συνολικό αριθμό τους που είναι απαραίτητα κατά την διάρκεια της εκπαίδευσης προκειμένου ο ταξινομητής να λειτουργεί αποδοτικά.

Η όλη διαδικασία ταξινόμησης αποτελείται από τα ακόλουθα βήματα:

1. Διαχωρίζουμε το βίντεο σε μικρά βίντεο κλιπ και τροφοδοτούμε το δίκτυο TPN με αυτά τα βίντεο κλιπ και παίρνουμε ως αποτέλεσμα k-προτεινόμενα ToIs και τα αντίστοιχα χαρακτηριστικά τους για κάθε κλιπ βίντεο.

2. Συνδέουμε τα προτεινόμενα ToIs για να πάρουμε action tubes που μπορεί να περιέχουν μια ενέργεια.

3. Για κάθε υποψήφιο action tube, η οποία είναι μια ακολουθία του ToIς, τροφοδοτούμε τους χάρτης ενεργοποίησης του στον ταξινομητή για ταξινόμηση.

Στα πρώτα βήματα του σταδίου ταξινόμησης αναφερόμαστε μόνο στο σύνολο δεδομένων JHMDB, επειδή έχει μικρότερο αριθμό βίντεο από το σύνολο δεδομένων UCF το οποίο μας βοήθησε να εξοικονομήσουμε πολύ χρόνο και πόρους. Αυτό συμβαίνει επειδή κάναμε τα περισσότερα πειράματα μόνο JHMDB και αφού βρήκαμε τη βέλτιστη υλοποίηση, την υλοποιήσαμε για το UCF, επίσης.

1.1 JHDMB dataset

1.1.1 Ταξινομητές Linear, SVM και RNN

Training Για να εκπαιδεύσουμε τον ταξινομητή μας, πρέπει να εκτελέσουμε τα προηγούμενα βήματα, για κάθε βίντεο. Ωστόσο, κάθε βίντεο έχει διαφορετικό αριθμό καρέ και καταλαμβάνει μεγάλη ποσότητα μνήμης στη ΓΠΥ. Για να αντιμετωπίσουμε αυτή την κατάσταση και έχοντας 4 διαθέσιμες GPU, δίνουμε ως είσοδο ένα βίντεο ανά GPU. Έτσι μπορούμε να χειριστούμε 4 βίντεο ταυτόχρονα. Αυτό σημαίνει ότι ένα κλασσικό training παίρνει πάρα πολύ χρόνο για μόλις 1 εποχή. Η λύση με την οποία ήρθαμε, είναι να προυπολογίσουμε τους χάρτες χαρακτηριστικών τόσο για action tubes προσκηνίου όσο και φόντου και στη συνέχεια να τροφοδοτήσουμε αυτούς τους χάρτες στον ταξινομητή μας για να τον εκπαιδεύσουμε.Αυτή η λύση περιλαμβάνει τα ακόλουθα βήματα:

1. Αρχικά, εξαγουμε τους χάρτες χαρακτηριστικών από τα πραγματικά action tubes . Ακόμα εξάγουμε τα χαρακτηριστικά από action tubes φόντου τα οποία είναι διπλάσια στον αριθμό από αυτά του φόντου. Επιλέξαμε αυτή την αναλογία μεταξύ του αριθμού των θετικών και αρνητικών action tubes εμπνευσμένοι από τους Yang et al. 2017, των οποίων η μέθοδος χρησιμοποιεί ποσοστό 25% μεταξύ των περιοχών ενδιαφέροντος προσκηνίου και των συνολικών περιοχών, και συνολικά επιλέγει 128 τέτοιες περιοχές. Αντίστοιχα, επιλέγουμε ένα λίγο μεγαλύτερο ποσοστό επειδή έχουμε μόνο ένα πραγματικό action tube σε κάθε βίντεο. Έτσι, για κάθε βίντεο λαμβάνουμε 3 action tubes συνολικά, 1 προσκηνίου και 2 φόντου. Θεωρούμε ως background action tubes εκείνα που το σκορ επικάλυψης τους με οποιοδήποτε action tube είναι μεγαλύτερο απο 0.1 αλλά μικρότερο από 0.3 . Φυσικά,προκειμενου να εξάγουμε αυτά τα action tubes, χρησιμοποιούμε ένα προεκπαιδευμένο TPN, για να μας προτείνει ToIs για κάθε τμήμα βίντεο και τον προτεινόμενο αλγόριθμο σύνδεσης για να συνδέσουμε αυτά τα ToIs. Τελικώς, για κάθε action tube λαμβάνουμε τους αντίστοιχους χάρτες ενεργοποίησης χρησιμοποιώντας 3D RoiAlign.

2. Αφού εξαγουμε αυτά τα χαρακτηριστικά, εκπαιδεύουμε τους ταξινομητές μας. Ο Γραμμικός ταξινομητής χρειάζεται ένα σταθερό μέγεθος εισόδου,

συνεπώς χρησιμοποιούμε μια συνάρτηση pooling στην διάσταση του αριθμού των βίντεο. Έτσι, αρχικά έχουμε ένα χάρτη χαρακτηριστικών μεγέθους *3,512,16* και μετά λαμβάνουμε ως έξοδο έναν χάρτη χαρακτηριστικών μεγέθους *512,16*. Πειραματιζόμαστε χρησιμοποιώντας αμφότερα max και avg pooling όπως φαίνεται στον Πίνακα χρησιμοποιώντας 1.1. Για τον ταξινομητή RNN δεν χρειαζόμαστε καμία pooling διαδικασία ενώ για τον ταξινομητή SVM πειραματιζόμαστε ξανά χρησιμοποιώντας και τις δύο αυτές συναρτήσεις τα αποτελέσματα του οποίου φαίνονται στον Πίνακα 1.2.

Validation Το στάδιο επικύρωσης περιλαμβάνει τη χρήση τόσο προεκπαιδευμένου TPN όσο και του ταξινομητή. Έτσι, για κάθε βίντεο λαμβάνουμε σκορ ταξινόμησης για τα προτεινόμενα action tubes. Οι περισσότερες προσεγγίσεις συνήθως θεωρούν ένα κατώφλι σκορ εμπιστοσύνης πάνω από το οποίο θεωρούν ένα action tube ως προσκήνιο. Ωστόσο, εμείς δεν χρησιμοποιούμε κανένα σκορ εμπιστοσύνης. Αντιθέτως, επειδή γνωρίζουμε ότι JHMDB έχει κομμένα βίντεο με μόνο 1 εκτελούμενη δράση ανά βίντεο, εμείς απλά θεωρούμε το καλύτερο ως προς το σκορ action tube ως πρόβλεψη.

Classifier	**Pooling**	**mAP**		
		0.5	0.4	0.3
Linear	mean	14.18	19.81	20.02
	max	13.67	16.46	17.02
RNN	-	11.3	14.14	14.84

Table 1.1: First classification results using Linear and RNN classifiers

Dimensions		**Pooling**	**mAP precision**		
before	after		0.5	0.4	0.3
(k,64,8,7,7)	(1,64,8,7,7)	mean	3.16	4.2	4.4
(k,64,8,7,7)	(1,64,8,7,7)	max	1.11	2.35	2.71
(k,256,8,7,7)	(1,256,8,7,7)	mean	11.41	11.73	11.73
(k,256,8,7,7)	(1,256,8,7,7)	max	**22.07**	**24.4**	**25.77**

Table 1.2: Our architecture's performance using 5 different policies and 2 different feature maps while pooling in tubes' dimension. With bold is the best scoring case

1.1.2 Temporal pooling

Μετά τη λήψη των πρώτων αποτελεσμάτων, εφαρμόζουμε μια συνάρτηση χρονικής ομαδοποίησης (temporal pooling) εμπνευσμένη από το Hou, ἤεν και Σηαη

2017. Χρειαζόμαστε ένα σταθερό μέγεθος εισόδου για το Σ΄Μ. Ωστόσο, το χρονικό stride των action tube μας ποικίλλει από 2 έως 5, αφού ένα βίντεο με 15 καρέ αποτελείεται από 2 συνεχόμενες ToIs ενώ ένα βίντεο με 40 καρέ αποτελείται απο 5. Έτσι χρησιμοποιούμε ως σταθερή χρονική διάσταση ίσον με 2. Ως λειτουργία pooling χρησιμοποιούμε 3D max poolign, για κάθε φίλτρο του χάρτη χαρακτηριστικών ξεχωριστά. Για παράδειγμα, για ένα action tube με 4 συνεχόμενες ToIs, έχουμε $(4, 256, 8, 7, 7)$ ως μέγεθος του χάρτη χαρακτηριστικών. Διαχωρίσουμε το feature map σε 2 ομάδες χρησιμοποιώντας την συνάρτηση *linspace* και αναδιαμορφώνουμε το χάρτη χαρακτηριστικών σε $(256, k, 8, 7, 7)$ όπου k είναι το μέγεθος της κάθε ομάδας. Αφού κάνουμε χρήση 3D max pooling, θα πάρουμε ένα χαρακτηριστικό χάρτη διαστάσεων $(256, 8, 7, 7)$, ακολούθως τους ενώνουμε και τελικά λαμβάνουμε χαρακτηριστικών μεγέθους $(2, 256, 8, 7, 7)$. Σε αυτή την περίπτωση δεν πειραματιζόμαστε με χάρτες χαρακτηριστικών μεγέθους (64, 8, 7, 7) επειδή με βάση τα παραπάνω αποτελέσματα, δεν θα έχουμε καλύτερη επίδοση απ΄ τα χαρακτηριστικών μεγέθους $(256, 8, 7, 7)$. Τα αποτελέσματα παρουσιάζονται στον πίνακα 1.3, όπου περιλαμβάνεται η καλύτερη προηγούμενη μέθοδος η οποία χρησιμοποιεί max pooling αντί για temporal pooling.

Dimensions		**Temp Pooling**	**mAP precision**		
before	after		0.5	0.4	0.3
k,256,8,7,7	1,256,8,7,7	-	22.07	24.4	25.77
k,256,8,7,7	2,256,8,7,7	Yes	25.07	26.91	29.11

Table 1.3: mAP results using temporal pooling for both RoiAlign approaches

1.2 Προσθήκη περισσότερων groundtruth tubes

Τα προηγούμενα αποτελέσματα προήλθαν από την εκπαίδευση των ταξινομητών χρησιμοποιώντας μόνο 1 action tube προσκηνίου και 2 φόντο. Σκεφτήκαμε ότι θα έπρεπε να πειραματιστούμε με τον αριθμό των action tubes προσκηνίου καθώς επίσης και την αναλογία μεταξύ των action tubes προσκηνίου και φόντου, επειδή στις προηγούμενες προσεγγίσεις λειτουργήσαμε λιγάκι αυθαίρετα. Έτσι, επιλέγουμε να εκπαιδεύσουμε τους προηγούμενους ταξινομητές μας χρησιμοποιώντας 2, 4 και 8 action tubes προσκηνίου και αναλογία 2:3, 1:2, 1:3 και 1:4 μεταξύ του αριθμού των tubes προσκηνίου και του συνολικού αριθμού τους.

Πρώτον, εκπαιδεύουμε το RNN ταξινομητή χρησιμοποιώντας χάρτες χαρακτηριστικών με διαστάσεις $(256, 8, 7, 7)$. Οι επιδόσεις τους με βάση την μετρική mAP παρουσιάζονται στον πίνακα 1.4 για το όριο επικάλυψης ίσο με 0,5, 0,4 και 0,3.

F. map	**FG tubes**	**Total tubes**	**mAP**		
			0.5	0.4	0.3

(k,256,8,7,7)	1	3	11.3	14.14	14.84
	2	3	1.96	5.07	7.27
		4	3	5.03	5.77
		6	1.34	3.89	4.49
		8	0.77	1.51	2.72
	4	6	13.23	21.74	25.4
		8	20.73	28.25	29.50
		12	16.55	24.35	25.22
		16	20.11	25.50	27.62
	8	12	13.82	19.93	22.80
		16	15.47	23.08	24.19
		24	15.88	23.44	24.48
		32	12.66	23.50	25.61

Table 1.4: RNN results

Σύμφωνα με τον πίνακα 1.4, πρώτον, μπορούμε να δούμε ότι η αύξηση του αριθμού των action tubes προσκηνίου από 1 έως 2 οδηγούν στη απότομη μείωση της απόδοσης του mAP. Αλλά, όταν θέτουμε τα action tubes προσκηνίου ίσα με 4 έχουμε καλύτερα αποτελέσματα. Πάνω σ′ αυτό, έχουμε την καλύτερη απόδοση όταν η αναλογία είναι ίση με 1:2 και 1:4. Τέλος, όταν ορίζουμε τον αριθμό των tubes προσκηνίου ίσο με 8, η απόδοση βελτιώνεται ελαφρώς σε σύγκριση με τις αρχικές επιλογές (1 action tube προσκηνίου και 3 συνολικά) , αλλά η κατάσταση αυτή δεν να μας φέρεις τα καλύτερα αποτελέσματα.

Στη συνέχεια, είναι καιρός να πειραματιστούμρ χρησιμοποιώντας τη γραμμική ταξινόμηση. Χρησιμοποιούμε ξανά το ίδιες υποθέσεις όπως κάναμε και για την ταξινόμηση με RNN. Όπως προαναφέρθηκε, χρειαζόμαστε μια μέθοδο ομαδοποίησης (pooling) πριν από το βήμα ταξινόμησης. Σύμφωνα με τον πίνακα 1.1, η μέθοδος του avg pooling έχει ως αποτέλεσμα καλύτερη απόδοση mAP από το max pooling , οπότε χρησιμοποιούμε avg pooling για όλες τις ακόλουθες περιπτώσεις. Τα αποτελέσματα περιλαμβάνονται στον πίνακα 1.5.

F. map	**FG tubes**	**Total tubes**	**mAP**		
			0.5	0.4	0.3
(k,256,8,7,7)	1	3	14.18	19.81	20.02
	2	3	12.68	13.38	15.14
		4	11.5	14.95	16.22
		6	10.74	13.36	15.18
		8	8.00	9.83	11.17
	4	6	15	17.55	19.39
		8	17.04	20.12	22.07
		12	17.57	19.9	21.88
		16	14.24	17.24	17.95

	8	12	17.91	22.51	24.62
		16	16.76	20.34	22.72
		24	17.61	19.12	24.48
		32	14.45	18.07	19.14

Table 1.5: Linear results

Πρώτα απ ′ όλα, μετά την εξέταση των αποτελεσμάτων που παρουσιάστηκαν στους δύο πίνακες 1.4 και 1.5, είναι σαφές ότι όταν ορίζουμε τον αριθμό των action tubes προσκηνίου ίσο με 2, και για τις δύο περιπτώσεις, έχουμε χειρότερα αποτελέσματα απ΄ το αρχικό. Αυτό μάλλον οφείλεται στο γεγονός ότι αυξάνουμε επίσης τον αριθμό των action tubes φόντου για περιπτώσεις όταν η αναλογία είναι 1:2, 1:3 και 1:4 με αποτέλεσμα να θεωρούν οι ταξινομητές τα περισσότερα προτεινόμενα action tubes ότι είναι φόντου. Από την άλλη πλευρά, όταν έχουμε ορίσει αναλογία ίση με 2:3, αντί να θεωρήσουν τα περισσότερα προτεινόμενα action tubes, ως φόντο, τα ταξινομούν ως μια συγκεκριμένη δράση τάξη, που σημαίνει ότι καταλήγομε σε κατάσταση overfitting. Έτσι, αν και πιστεύουμε ότι δεν θα πρέπει να ερευνήσουμε για περιπτώσεις με 2 action tubes που ανήκουν στο προσκήνιο, θα εκπαιδεύσουμε τον SVM ταξινομητή μας χρησιμοποιώντας 2 action tubes προσκηνίου και όλα τα προαναφερθέντα ποσοστά επειδή θέλουμε να είμαστε βέβαιοι για την υπόθεσή μας. Από την άλλη πλευρά, παρατηρούμε ότι η χρήση 4 ή 8 action tubes μας οδηγεί σε καλύτερα αποτελέσματα από το τα αρχικά αποτελέσματα. Οι καλύτερες επιδόσεις έρχονται όταν η αναλογία μεταξύ των αριθμών των action tubes προσκηνίου και συνολικών είναι 1:3 και για τις δύο περιπτώσεις. Παράλληλα, έχουμε καλά αποτελέσματα για τις αναλογίες 2:3 και 1:2, και λαμβάνουμε την χειρότερη επίδοση όταν χρησιμοποιούμε αναλογία 1:4. Αυτό προκαλείται μάλλον από το μεγάλο αριθμός action tubes φόντου σε σχέση με τον αριθμό των action tubes προσκηνίου. Όπως προαναφέρθηκε, εκπαιδεύουμε τον ταξινομητή SVM χρησιμοποιώντας τις προαναφερθείσες περιπτώσεις Οι επιδόσεις ταξινόμησης με χρήση της μέτρησης μΑΠ εμφανίζονται στον πίνακα 1.6. .

F. map	**FG tubes**	**Total tubes**	**mAP**		
			0.5	0.4	0.3
(2,256,8,7,7)	1	3	24.97	26.91	29.11
	2	3	13.87	18.74	21.29
		4	14.21	19.67	21.75
		6	12.88	18.62	21.59
		8	12.66	18.7	21.97
	4	6	25.04	26.91	27.82
		8	24.34	25.67	26.34
		12	23.47	25.31	25.9
		16	21.94	23.55	24.23
	8	12	24.83	27.13	27.46

		16	23.97	26.38	26.94
		24	24.17	26.24	26.76
		32	24.17	26.24	26.76

Table 1.6: SVM results

Τα αποτελέσματα μας δείχνουν κάποια ενδιαφέροντα γεγονότα. Πρώτον, επιβεβαιώνουν την υπόθεσή μας ότι το δίκτο είναι αδύνατον να εκπαιδευτεί με μόνο 2 action tubes προσκηνίου. Επίσης, παρατηρούμε ότι έχουμε σχεδόν τα ίδια αποτελέσματα με τα αποτελέσματα που προέκυψαν για τη χρήση της πολιτικής 1, μόνο ένα action tube προσκηνίου, 3 συνολικά και χρονικό pooling, γεγονός το οποίο είναι λίγο παράξενο. Αυτό είναι μάλλον επειδή κατά τη διάρκεια του υπολογισμού της κλίμακας, στο στάδιο εκπαίδευσης, δεν έχουμε τόσο καλό δείγμα βίντεο όπως κάναμε κατά τη διάρκεια της προαναφερθείσας περίπτωσης. Αλλά θεωρούμε ότι είναι καλύτερο να συνεχίσουμε τις δοκιμές χρησιμοποιώντας 4 ή 8 action tubes προσκηνίου. Τελευταίό αλλά όχι λιγότερο σημαντικό, είναι σαφές ότι έχουμε το καλύτερο αποτέλεσμα όταν έχουμε μια ποσοστό 2:3 μεταξύ του αριθμού των action tubes προσκηνίου και των συνολικών. Επίσης, είναι πιο προτιμότερο να έχουμε 4 action tube προσκηνίου αντί για 8. Αυτό σημαίνει ότι επειδή έχουν δοθεί πάρα πολλά το SVM μπερδεύεται, και έτσι αποτυγχάνει να λειτουργήσει αποτελεσματικά.

1.3 Ταξινομητής MultiLayer Perceptron (MLP)

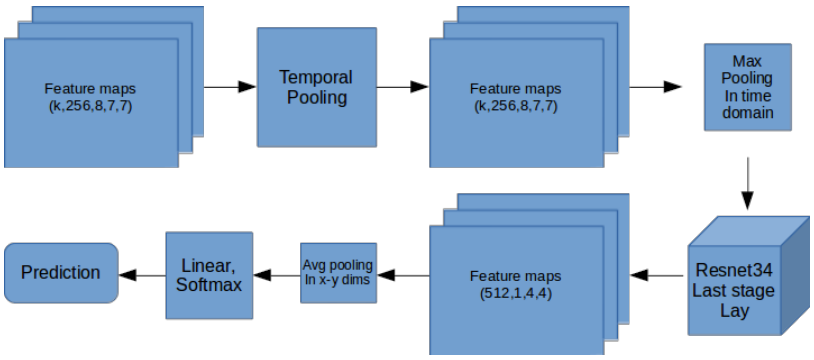

Figure 1.1: Structure of the MLP classifier

Σε προηγούμενες ενότητες χρησιμοποιήσαμε κλασσικούς ταξινομητές όπως τον Γραμμικό, ένα RNN και SVM. Τελευταίοι αλλά εξίσου σημαντικοι, μια άλλη ευρέως κατηγορία ταξινομητών είναι οι Multilayer Perceptron (MLP). Σχεδιάζουμε ένα ΜΛΠ όπως φαίνεται στο σχήμα 1.1 για διάρκεια δείγματος ίση με 8, και περιγράφεται κατωτέρω:

- Στην αρχή, μετά το 3D Roi Align και για διάρκεια του δείγματος ίση με 8 καρέ, λαμβαουμε ένα χάρτη ενεργοποίησης μεγέθους $(k, 256, 8, 7, 7)$ όπου κ είναι ο αριθμός των συνδεδεμένων ToIς. Εμπνευσμένοι από προηγούμενες ενότητες, εκτελούμε temporal pooling ακολουθούμενο από max pooling στην διάσταση της διάρκειας του δείγπματος. Έτσι, έχουμε τώρα έναν χάρτη χαρακτηριστικών με διαστάσεις ίσες με (2, 256, 7, 7), τις οποίες αναδιαμορφώνουμε σε (256, 2, 7, 7) και τροφοδοτούμε layers που εξήχθησαν από το τελευταίο στάδιο του ResNet34. Αυτά τα στάδια περιλαμβάνουν 3 Residual Layers με στριδε ίσο με 2 σε όλες τις 3 διαστάσεις και αριθμός εξόδου φίλτρων ίσου με 512.

- Μετά τα Residual Layers, κάνουμε avg pooling για τις διαστάσεις ξ-ψ. Έτσι, έχουμε ως χάρτες ενεργοποίησης εξόδου με μέγεθος διαστάσεων ίσο με (512,). Τέλος, τροφοδοτούμε αυτούς τους χαρακτηριστικούς χάρτες σε ένα γραμμικό layer προκειμένου να εξάγουμε την κλάση του υποψήφιου action tube, μετά την εφαρμογή της λειτουργίας Σοφτ-Μαξ.

1.3.1 Κλασσικό training

Όπως προαναφέρθηκε προηγουμένως, ο κώδικας εκπαίδευσης απαιτεί την εκτέλεση ενός μόνο βίντεο ανά ΓΠΥ, επειδή τα βίντεο έχουν διαφορετική διάρκεια. Για προηγούμενες προσεγγίσεις, μας ήρθδε η ιδέα του προυπολογισμού των χαρακτηριστικά των action tubes του βίντεο και στη συνέχεια εκπαιδεύουμε μόνο τον ταξινομητή. Ωστόσο, για αυτό το βήμα, εκπαιδεύσαμε τον ταξινομητή μας με τον κλασικό τρόπο για να λάβουμε αποτελέσματα ταξινόμησης. Φυσικά, χρησιμοποιήσαμε ένα προεκπαιδευμένο TPN, του οποίο παγώσαμε τα layers για να μην εκπαιδευτούν. Προσπαθήσαμε να εξερευνήσουμε διαφορετικές αναλογίες μεταξύ του αριθμού των action tubes προσκηνίου και του συνολικού αριθμού των action tubes ανά βίντεο. Οι πρώτες 3 προσομοιώσεις περιλαμβάνουν σταθερό αριθμό συνολικών action tubes και μεταβλητή αναλογία μεταξύ του αριθμού των action tubes προσκηνίου και φόντου. Αρχίσαμε χρησιμποιώντας μόνο action tubes προσκηνίου, το οποίο σημαίνει ότι 32 από 32 action tube είναι προσκηνίου, μετά τα μισά από τα προτεινόμενα action tubes, δηλαδή 16 από 32 και τέλος λιγότερο από το ήμισυ, δηλαδή 14 από τις 32. Μετά από αυτό, πειραματιζόμαστε χρησιμοποιώντας έναν σταθερό αριθμό action tubes προσκηνίου και μεταβλητού αριθμού συνολικών, ο οποίος είναι 16, 24 και 32. Τα αποτελέσματα των επιδόσεων παρουσιάζονται στον πίνακα 1.7.

FG tubes	**Total tubes**	**mAP**		
		0.5	0.4	0.3
32	32	1.28	1.73	1.87
16		3.98	4.38	4.38
14		0.40	0.40	0.40
8	16	9.41	12.59	14.61
	24	12.32	15.53	18.57

	32	7.16	10.92	13.00

Table 1.7: MLP's mAP performance for regular training procedure

Τα αποτελέσματα δείχνουν ότι όταν οι πρώτες 3 προσεγγίσεις μας δίνουν πολύ άσχημα αποτελέσματα. Συγκρίνοντας τους με τους υπόλοιπους 3, ήρθαμε με το συμπέρασμα ότι χρειαζόμαστε το πολύ 8 action tubes προσκηνίου, ακόμη και όταν ο λόγος μεταξύ του αριθμού action tubes του προσκηνίου και του φόντου είναι υπέρ του δεύτερου. Πιθανότατα, πάρα πολλοί action tubes προσκηνίου κάνουν την αρχιτεκτονική μας να έρθει σε κατάσταση overfitting και συνεπώς να είναι ανίκανη να γενικεύει.

1.3.2 Εξαγωγή χαρακτηριστικών

Όπως εκτελέστηκε προηγουμένως, εκπαιδεύσαμε τον ταξινομητή MLP χρησιμοποιώντας προ-υπολογισμένους χάρτες χαρακτηριστικών. Αυτοί οι χάρτες περιλαμβάνουν τόσο action tubes που είναι στο προσκήνιο όσο φόντου. Με βάση τα συμπεράσματα που προέκυψαν στα προηγούμενα τμήματα, θα εκπαιδεύσουμε μόνο για αριθμό action tube προσκηνίου ίσο με 4 και 8. Επιπλέον Θα εκπαιδεύσουμε τον ταξινομητή μας για 3 διαφορετικές αναλογίες, οι οποίες είναι 1:1, 1:2 και 1:3. Ο πίνακας 1.8 δείχνει αυτές τις περιπτώσεςι καθώς και τις αντίστοιχες επιδόσεις του mAP κατά τη διάρκεια του βήματος επικύρωσης.

FG tubes	**Total tubes**	**mAP**		
		0.5	0.4	0.3
4	6	4,37	8,54	10,12
	8	5.89	9.54	13.61
	12	9.51	12.8	14.6
	16	6.80	13.17	14.67
8	12	8,62	12,32	14,74
	16	8.49	13.94	15.09
	24	6.72	12.17	15.30
	32	13.27	17.64	18.97

Table 1.8: mAP results for MLP trained using extracted features

Συγκρίνοντας τα αποτελέσματα από τους πίνακες 1.8 και 1.7, είναι σαφές ότι χρειαζόμαστε 8 tubes προσκηνίου για να λειτουργεί καλά ο ταξινομητής MLP. Ωστόσο, δεν είναι πολύ σαφές ποια από τις δύο προτεινόμενες εκπαιδευτικές διαδικασίες είναι καλύτερη, αλλά αν πρέπει να αποφασίσουμε μία μέθοδο, θα επιλέξουμε τη χρήση προϋπολογισμένων χαρακτηριστικών. Η προσέγγιση αυτή κατορθώνει

να επιτύχει τα καλύτερα αποτελέσματα, και ειδικά όταν έχουμε 8 tubes προσκηνίου και 32 συνολικά. Επίσης, συγκρίνοντας τις μεθόδους με 4 ή 8 θετικά action tubes, είναι σαφές ότι θα προτιμούσαμε να χρησιμοποιούμε 8 γενικά. Ωστόσο, δεν είναι σαφές ποια αναλογία είναι καλύτερη, επειδή, έχουμε καλύτερα αποτελέσματα όταν έχουμε 8 action tubes και αναλογία 1:4 ενώ έχουμε καλύτερα αποτελέσματα όταν η αναλογία είναι 1:3 με 4 action tubes.

1.4 UCF dataset

1.4.1 Εισαγωγή

Κατά τη διάρκεια της προηγούμενης ενότητας, διερευνούμε διαφορετικές μεθόδους ταξινόμησης χρησιμοποιώντας διάφορες τάξεις. Λαμβάνοντας υπόψη την απόδοση των recall και MABO που παρουσιάζονται στο κεφάλαιο 4, είναι σαφές ότι το δίκτυό μας θα αποτύχει να αναγνωρίσει τα περισσότερα πραγματικά χωροχρονικά action tubes και να τα ταξινομήσει σωστά. Ωστόσο στις περισσότερες περιπτώσεις, η απόδοση του MABO πήρε Βαθμολογία περίπου 92-94%. Οπότε, μας ήρθε η ιδέα να μην εκτελέσουμε χωροχρονικό εντοπισμό και ταξινόμηση, για το σύνολο δεδομένων ΥΣΦ, αλλά μόνο χρονικό εντοπισμό. Αυτό σημαίνει ότι προσπαθούμε να ανιχνεύσουμε τα τμήματα βίντεο στα οποία εκτελείται μια ενέργεια, και επίσης προσπαθούμε να προσδιορίσουμε την κλάση της εκτελεσμένης ενέργειας.

1.4.2 Χρονικός εντοπισμός

Όπως παρουσιάζεται στο κεφάλαιο 4, ο αλγόριθμος σύνδεσής μας είναι σε θέση να πάρει καλή απόδοση χρονικού recall και MABO. Για τον χρονικό εντοπισμό μιας δράσης σε βίντεο, χρησιμοποιούμε μόνο τις χρονικές πληροφορίες που περιέχουν τα προτεινόμενα action tubes, που ισοδυναμούν με το πρώτο και το τελευταίο καρέ του tube. Θα ταξινομήσουμε τα προτεινόμενα action tubes χωρίς να πραγματοποιήσουμε χωροχρονικό εντοπισμό, αλλά μόνο χρονικό. Αν και δεν χρησιμοποιούμε τα προτεινόμενα πλαίσια ανά καρέ για την ταξινόμηση, εκμεταλλευόμαστε τις χωρικές πληροφορίες τους για να να εκτελέσουμε καλύτερο χρονικό εντοπισμό. Διαισθητικά, αυτό συμβαίνει επειδή, για να να εξαγάγουμε τα action tubes, λαμβάνουμε υπόψιν μας τη χωρική επικάλυψη μεταξύ των συνδεδεμένων ToIs.

Η προαναφερθείσα προσέγγιση περιλαμβάνει τα ακόλουθα βήματα:

1. η γροθιά, χρησιμοποιούμε ΤΠΝ για να προτείνουμε χωροκροταφικό ToIs, όπως ακριβώς κάναμε σε προηγούμενες προσεγγίσεις. Στη συνέχεια, συνδέουμε αυτά τα ToIs με βάση την προτεινόμενη στο κεφάλαιο 5, με τη χρήση του χωροκροταφικού αλγορίθμου όριο ίσο με 0,9, για την αφαίρεση επικαλυπτόμενων σωλήνων δράσης.

2. τα προηγούμενα βήματα είναι ακριβώς τα ίδια με τις προηγούμενες προσεγγίσεις ταξινόμησης. Ωστόσο, σε αυτή την προσέγγιση, δεν χρησιμοποιούμε

κανένα είδος στοίχισης POI για να χάρτες λειτουργιών εκχυλίσματος σωλήνων δράσης. Αντιθέτως, για όλες τις προτεινόμενες σωλήνες δράσης, βρίσκουμε τη διάρκειά τους, γνωστή και ως την πρώτη και την τελευταία τους κορνίζα. Μετά από αυτό, θα κάνουμε χρονικές ενέργειες για να αφαιρέσουμε την αλληλοεπικάλυψη Σωλήνες. Η μόνη διαφορά μεταξύ χωροχρονικών και κροταφικών το κριτήριο επικάλυψης, το οποίο χρησιμοποιείται. Για τα χωροχρονικά, χρησιμοποιούμε και αντίστοιχα, για τη χρονική χρήση των χρονικών που παρουσιάζονται στο κεφάλαιο 2.

3. Φυσικά, τα προτεινόμενα σωληνάκια δράσης διαρκούν περισσότερο από 16 που ορίζεται ως διάρκεια δείγματος. Έτσι, διαχωρίσουμε τους σωλήνες δράσης σε βίντεο κλιπ διάρκειας 16 πλαισίων (όπως η διάρκεια του δείγματός μας). Αυτά τα τμήματα βίντεο είναι και πάλι σε 3Δ ρεσΝετ34 ([;]), αλλά αυτή τη φορά, δεν το χρησιμοποιούμε μόνο για δυνατότητα εξαγωγής, αλλά και για ταξινόμηση για κάθε τμήμα βίντεο.

4. έτσι, για κάθε βίντεο κλιπ, για κάθε κλάση έχουμε ένα σκορ εμπιστοσύνης μετά την λειτουργία. Τέλος, έχουμε τη μέση βαθμολογία εμπιστοσύνης για σε κάθε κλάση, και θεωρούμε την κλάση βέλτιστης βαθμολόγησης ως ετικέτα σωλήνα δράσης. Φυσικά, μερικοί σωλήνες δράσης μπορεί να μην περιέχουν καμία ενέργεια, Έτσι έχουμε ορίσει μια βαθμολογία εμπιστοσύνης για ξεχωριστούς σωλήνες δράσης προσκηνίου με Φόντο.

Training Το μόνο εκπαιδεύον μέρος αυτής της αρχιτεκτονικής είναι το ΡεσΝετ34. Χρησιμοποιούμε ένα προεκπαιδευμένο ΤΠΝ, όπως παρουσιάζεται στο κεφάλαιο 4. Η διαδικασία κατάρτισης ΡεσΝετ34 με βάση τον κωδικό που δόθηκε από το [;]. Το τροποποιήσαμε για να μπορούμε να εκπαιδευόμαστε για το σύνολο δεδομένων ΥΎΦ-101, μόνο για τις 24 τάξεις, για τις οποίες υπάρχουν και η ΤΠΝ μας είναι εκπαιδευμένη.

Validation Με βάση τα προαναφερθέντα βήματα, είναι σαφές ότι οι παράμετροι που μπορούν να τροποποιηθούν είναι τα χρονικά όρια και το όριο να αποφασίζουν εάν μια ενέργεια περιέχεται ή όχι. Όλοι οι διαφορετικοί συνδυασμοί που χρησιμοποιούνται κατά τη διάρκεια της επικύρωσης παρουσιάζονται στον πίνακα 1,20.

ΝΜΣ τηρεση	**ὂνφ τηρεση**	**μΑΠ**		
		0.5	0.4	0.3
0.9	0.6	0.3	0.54	0.64
	0.75	0.25	0.45	0.55
	0.85	0.2	0.38	0.49
0.7	0.6	0.63	1.02	1.27
	0.75	0.5	0.84	1.05
	0.85	0.4	0.68	0.89
0.5	0.6	0.96	1.21	1.75
	0.75	0.63	0.93	1.38
	0.85	0.57	0.72	1.03

0.4	0.6	1.07	1.52	2.03
	0.75	0.79	1.18	1.63
	0.85	0.71	0.98	1.33
0.3	0.6	1.1	1.66	2.53
	0.75	0.93	1.39	2.08
	0.85	0.81	1.12	1.6
0.2	0.6	0.84	1.38	2.17
	0.75	0.73	1.13	1.78
	0.85	0.65	0.81	1.31

Πίνακας 1.9: Υ˜Φ΄ς τεμποραλ λοςαλιζατιον μΑΠ περφορμανςε

ΝΜΣ τηρεση	**Ρεςαλλ**			**MABO**
	0.9	0.8	0.7	
0.9	0.7361	0.8935	0.9422	0.9138130172
0.7	0.3194	0.6875	0.9293	0.8412186326
0.5	0.1757	0.3331	0.6281	0.7471525429
0.4	0.1483	0.2829	0.4707	0.6986400756
0.3	0.111	0.2038	0.3848	0.6429232202

Πίνακας 1.10: Υ˜Φ΄ς τεμποραλ λοςαλιζατιον ρεςαλλ ανδ MABO περφορμανςες

Σύμφωνα με τον πίνακα 1,20, οι επιδόσεις του μΑΠ για τη χρονική η ταξινόμηση είναι πολύ κακή. Η καλύτερη απόδοση είναι περίπου 2πολύ χαμηλά. Συγκρίνοντας αυτά τα αποτελέσματα με τα αποτελέσματα που εμφανίζονται στον πίνακα 1,21, συμπεραίνουμε ότι η μέθοδός μας δεν είναι καθόλου αποδοτική. Παρόλο που τα αποτελέσματα του μΑΠ αυξάνονται και η απόδοση του MABO μειώνεται ταχέως. Φυσικά, αυτό το αποτέλεσμα Δεδομένου ότι, μειώνοντας το όριο του νμμς, ο αριθμός των απορριπτθέντων σωλήνες δράσης αυξάνεται. Δοκιμάσαμε μια άλλη προσέγγιση, η οποία εφαρμόζει τον αλγόριθμο και όχι πριν από αυτό όπως κάναμε προηγουμένως. Επίσης, παρατηρήσαμε σε προηγούμενες στις περισσότερες περιπτώσεις, έχουμε αυτές τις χαμηλές επιδόσεις λόγω της με ψευδή θετικά αποτελέσματα, τα οποία δεν αφαιρούνται κατά τη διάρκεια της διαδικασίας. Να είμαι πιο συγκεκριμένα, ο πίνακας 1,22 δείχνει όλα τα πραγματικά και ψευδώς θετικά ορίζεται το όριο του ννς ίσο με 0,2, όριο επικάλυψης μΑΠ ίσο με 0,3 και όριο εμπιστοσύνης ίσο με 0,6 για τις δύο προαναφερθείσες προσεγγίσεις.

˜λασς	**Απρ 1**		**Απρ 2**		**˜λασς**	**Απρ 1**		**Απρ 2**	
	ΤΠ	ΦΠ	ΤΠ	ΦΠ		ΤΠ	ΦΠ	ΤΠ	ΦΠ

Βασκετβαλλ	5	279	6	403	ΒασκετβαλλΔυνκ	7	7	12	13
Βικινγ	0	3	0	5	̈λιφφΔινγ	11	55	1	1
̈ριςκετΒοωλινγ	0	0	10	75	Δινγ	20	189	23	272
Φενςινγ	11	222	25	336	ΦλοορΓψμναστιςς	2	86	6	131
ΓολφΣωινγ	4	51	6	78	ΗορσεΡιδινγ	0	33	4	58
ΙςεΔανςινγ	8	29	6	38	ΛονγΘυμπ	1	24	6	43
Πολεα̈υλτ	0	202	9	296	Ροπε̈λιμβινγ	1	24	4	43
ΣαλσαΣπιν	3	158	5	237	ΣκατεΒοαρδινγ	0	10	0	13
Σκιινγ	0	0	0	0	Σκιθετ	1	27	6	43
ΣοςςερΘυγγλινγ	3	94	1	153	Συρφινγ	11	102	23	159
ΤεννισΣωινγ	0	125	0	166	ΤραμπολινεΘυμπινγ	4	18	4	32
ο̈λλεψβαλλΣπικινγ	20	704	20	1044	ΩαλκινγΩιτηΔογ	0	5	0	9

Πίνακας 1.11: ο̈μπαρινγ ΤΠ ανδ ΦΠ φορ βοτη αππροαχηες

Λαμβάνοντας υπόψη αυτά τα δύο γεγονότα, ήρθαμε με την ακόλουθη λύση. Στην προσέγγιση, ταξινομήσαμε τους υποψήφιους σωλήνες δράσης χρησιμοποιώντας βαθμολογίες σύνδεσης που από τη σύνδεση του αλγορίθμου και μετά αφαιρέσαμε τους σωλήνες δράσης. Ιν η νέα προσέγγισή μας, αφαιρούμε πρώτα τους σωλήνες δράσης με τα ίδια χρονικά όρια, για να πάρετε μοναδικούς σωλήνες χρονικής δράσης. Στη συνέχεια, ταξινομούμε όλες τις προτεινόμενες σωλήνες δράσης ακριβώς όπως κάναμε στο βήμα 3 προηγουμένως. Μετά από αυτό, θα εκτελέσουμε με τις βαθμολογίες εμπιστοσύνης που εξάγονται από το τελευταίο στρώμα της 3Δ ΡεσΝετ34 και τέλος κρατάμε αυτά που το σκορ εμπιστοσύνης τους είναι πάνω από ένα προκαθορισμένο όριο.

ΝΜΣ τηρεση	**ο̈νφ τηρεση**	**μΑΠ**		
		0.5	0.4	0.3
0.9	0.6	0.31	0.54	0.65
	0.75	0.26	0.46	0.55
	0.85	0.2	0.39	0.49
0.7	0.6	0.66	0.95	1.22
	0.75	0.55	0.80	1.01
	0.85	0.41	0.67	0.87
0.5	0.6	0.98	1.43	1.63
	0.75	0.75	1.14	1.29
	0.85	0.64	0.92	1.04
0.4	0.6	1.19	1.73	2.15
	0.75	0.9	1.35	1.63
	0.85	0.79	1.16	1.38
0.3	0.6	1.12	1.85	2.23
	0.75	0.96	1.54	1.7
	0.85	0.83	1.28	1.43

0.2	0.6	2.05	2.68	3.7
	0.75	1.61	2.17	3
	0.85	1.51	1.88	2.54

Πίνακας 1.12: ΥΦ΄ς τεμποραλ λοςαλιζατιον μΑΠ περφορμανςε

Συγκρίνοντας τους πίνακες 1,23 και 1,21, έχουμε τα ίδια αποτελέσματα για την επικάλυψη κατώτατα όρια 0,9, 0,7, 0,5, 0,4 και 0,3. Αλλά για επικάλυψη όριο 0,2 έχουμε παρατηρήσει Οι επιδόσεις του μΑΠ βελτιώνονται περίπου στο 1να χρησιμοποιούν ακόμη μικρότερα κατώτατα όρια επικάλυψης, τα οποία είναι 0,15, 0,1 και 0,05 τα οποία παρουσιάζονται στον πίνακα 1,24.

ΝΜΣ τηρεση	**ὃνφ τηρεση**	**μΑΠ**		
		0.5	0.4	0.3
0.15	0.6	2.01	2.66	3.62
	0.75	1.62	2.21	2.97
	0.85	1.51	1.91	2.56
0.10	0.6	1.87	2.74	3.77
	0.75	1.62	2.28	3.08
	0.85	1.5	2	2.7
0.05	0.6	1.85	2.71	3.73
	0.75	1.61	2.28	3.1
	0.85	1.5	2	2.7

Πίνακας 1.13: ΥΦ΄ς τεμποραλ λοςαλιζατιον μΑΠ περφορμανςε φορ εεν σμαλλερ ΝΜΣ τηρεσηολδ

Η χρήση πολύ μικρού κατώτατου ορίου του ΝΣ οδηγεί σε μια μικρή βελτίωση του μΑΠ Απόδοση. Ωστόσο, αυτές οι βαθμολογίες απέχουν πολύ από τα υπερκαλλιτεχνικά αποτελέσματα. Έτσι, νομίζουμε ότι πρέπει να επανεξετάσουμε τον τρόπο που εκπαιδεύσαμε την τάξη μας και την τον τρόπο με τον οποίο ταξινομεί τους σωλήνες δράσης, προκειμένου να είναι σε θέση να ταξινομήσει αποτελεσματικά. Αυτή η έρευνα έχει απομείνει για μελλοντικές εργασίες, και δεν θα γίνει ανάληψη καθηκόντων σε αυτή τη διατριβή.