**Δράση: ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ - ΚΑΙΝΟΤΟΜΩ**

**MediaPot [ΤΑΕΔΚ-06196]: Πλατφόρμα συλλογής, ανάλυσης και σύνθεσης πολυμεσικού περιεχομένου από κοινωνικά δίκτυα στην υπηρεσία των Ψηφιακών Μέσων**

# Π4.1 Μεθοδολογία Εξόρυξης Πληροφορίας από Πολυμεσικό Περιεχόμενο

| | |
|---|---|
| **Ενότητα Εργασίας** | Μεθοδολογία Εξόρυξης Πληροφορίας από Πολυμεσικό Περιεχόμενο |
| **Ημερομηνία** | 31/05/2024 |
| **Τύπος εγγράφου** | Αναφορά 1.0 |
| **Υπεύθυνος Φορέας** | ΕΚΕΤΑ |
| **Συμμετέχοντες Φορείς** | |
| **Επιμελητές** | Βασιλική Κουτσουπιά, Μανώλης Μυλωνάς, Συμεών Παπαδόπουλος, Βασίλειος Μεζάρης |
| **Συνοπτική περιγραφή** | Το Παραδοτέο επικεντρώνεται στην τεκμηρίωση της μεθοδολογίας εξόρυξης πληροφορίας από πολυμεσικό περιεχόμενο, με σκοπό τη διευκόλυνση της ειδησεογραφικής έρευνας, παρέχοντας στους χρήστες εργαλεία ώστε να εξάγουν δομημένη πληροφορία και να διαχειρίζονται με ευκολία το πολυμεσικό περιεχόμενο που εντοπίζεται από διάφορες πηγές στο Διαδίκτυο. Στο πλαίσιο της προτεινόμενης μεθοδολογίας, τα εργαλεία που αναπτύσσονται αφορούν τις ακόλουθες λειτουργίες 1) επισημείωση πολυμεσικού περιεχομένου 2) αντίστροφη αναζήτηση πολυμεσικού περιεχομένου 3) κατάτμηση περιεχομένου βίντεο 4) περίληψη περιεχομένου βίντεο |



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# Περιεχόμενα





Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# Πίνακες





Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# Εικόνες





Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# Υπόμνημα

Η εργασία υλοποιήθηκε στο πλαίσιο της Δράσης ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ – ΚΑΙΝΟΤΟΜΩ συγχρηματοδοτήθηκε από το Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης (ΕΤΠΑ) της Ευρωπαϊκής Ένωσης και εθνικούς πόρους μέσω του Ε.Π. Ανταγωνιστικότητα, Επιχειρηματικότητα & Καινοτομία (ΕΠΑνΕΚ) (κωδικός έργου MediaPot: ΤΑΕΔΚ-06196)



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# 1 Εισαγωγή

Το ερευνητικό έργο MediaPot έχει στόχο να αναπτύξει ένα σύγχρονο περιβάλλον σύνθεσης ειδησεογραφικών ιστοριών που να αξιοποιεί καινοτόμα εργαλεία ανάλυσης περιεχομένου και επαλήθευσης των ειδήσεων της επικαιρότητας. Συγκεκριμένα, στο πλαίσιο της Ενότητας Εργασίας 4 (ΕΕ4), αναπτύσσονται μέθοδοι και εργαλεία επεξεργασίας και ανάλυσης πολυμεσικού περιεχομένου (εικόνες, βίντεο) από διαδικτυακές πηγές. Με σκοπό οι χρήστες/-ριες[1] της προτεινόμενης πλατφόρμας να διαχειρίζονται αποδοτικά τις ειδήσεις που εντοπίζονται στα μέσα κοινωνικής δικτύωσης και να τις αξιοποιούν- όταν κρίνονται αξιόπιστες- για την δική τους ειδησεογραφική έρευνα.

Το έργο έχει σκοπό να γεφυρώσει τον ρόλο των παραδοσιακών Μέσων Μαζικής Ενημέρωσης (ΜΜΕ) με τον νέο, ταχύτατα αναπτυσσόμενο, ρόλο των κοινωνικών δικτύων αναφορικά με την καθημερινή ενημέρωση των πολιτών. Η προτεινόμενη πλατφόρμα θα μπορεί να αναλύει πολυμεσικό περιεχόμενο (κείμενο, εικόνες, βίντεο) από τα μέσα κοινωνικής δικτύωσης, να συμβάλει στην επαλήθευση του και να παράγει με ημιαυτόματο τρόπο συναφείς ειδήσεις, σχετικές με αυτές που παράγονται από τα ΜΜΕ, μέσω του συνδυασμού υπηρεσιών τεχνητής νοημοσύνης και εξόρυξης δεδομένων.

Ο τομέας και τα εργαλεία τεχνητής νοημοσύνης είναι δυνατόν να διευκολύνουν την αυτόματη ανάλυση μεγάλου όγκου πολυμέσων που εντοπίζονται στο διαδίκτυο και μπορούν να δώσουν τη δυνατότητα στα παραδοσιακά ΜΜΕ, να διαχειριστούν με καλύτερο τρόπο τα δεδομένα που εντοπίζουν, καθώς και να ανακτήσουν πολύτιμες πληροφορίες από αυτά.

Σε αυτό το πλαίσιο, αναπτύσσονται μέθοδοι και εργαλεία προ-επεξεργασίας και ανάλυσης πολυμεσικού περιεχομένου (εικόνες, βίντεο). Επιπλέον, εξετάζεται πρότερη εμπειρία αναφορικά με την προ-επεξεργασία και την ανάλυση εικόνων και βίντεο. Συνεπώς, αξιοποιούνται επιπλέον γνώσεις και μέθοδοι που έχουν ήδη αποκτηθεί στους συγκεκριμένους τομείς ανάλυσης πολυμέσων από την ομάδα του ΕΚΕΤΑ– η οποία έχει σημαντική εμπειρία στα εργαλεία που παρουσιάζονται– σε προηγούμενα προγράμματα Έρευνας και Καινοτομίας. Η ανάπτυξη, η αξιοποίηση και η σύνδεση των εργαλείων και των αποτελεσμάτων προηγούμενων και παρόντων μελετών, έχει ως τελικό σκοπό την αποδοτική διαχείριση των πολυμέσων από τους χρήστες της πλατφόρμας, για την διευκόλυνση χρήσης και έγκαιρης ενσωμάτωσης τους στην ειδησεογραφική έρευνα.

Ένα σημαντικό στοιχείο της πλατφόρμας MediaPot αφορά την ανάλυση μεμονωμένων στοιχείων περιεχομένου από διάφορες πηγές του διαδικτύου, με σκοπό την σημασιολογική απεικόνιση τους και

[1] *Με σκοπό την όσο το δυνατόν πιο ολοκληρωμένη συμπεριληπτική γλώσσα επιλέγεται στο κείμενο η χρήση των καταλήξεων σε όλα τα γένη. Ωστόσο, για λόγους συντομίας και μόνο, στο υπόλοιπο κείμενο δεν αναγράφονται πάντα όλες οι καταλήξεις αναλυτικά, αλλά εννοούνται.*



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

την άντληση πληροφορίας από κάθε στοιχείο χωριστά. Προς εκπλήρωση αυτής της ανάγκης, η προτεινόμενη μεθοδολογία εξόρυξης πληροφορίας από πολυμεσικό περιεχόμενο περιλαμβάνει εργαλεία για την επισημείωση και την αντίστροφη αναζήτηση πολυμεσικού περιεχομένου (εικόνες, βίντεο) καθώς και την κατάτμηση και τη δημιουργία αυτόματων περιλήψεων βίντεο περιεχομένου. Κάθε ένα από τα επόμενα τέσσερα κεφάλαια αυτού του παραδοτέου περιγράφει λεπτομερώς τις επιστημονικές μεθόδους καθώς και τις τεχνικές λεπτομέρειες υλοποίησης των τεσσάρων αυτών λειτουργιών, που αποτελούν τη βάση της μεθοδολογίας, της έρευνας και της ανάπτυξης που επιτελείται στο πλαίσιο της Ενότητας 4 του έργου (ΕΕ4).



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# 2 Επισημείωση Πολυμεσικού Περιεχομένου

## 2.1 Περιγραφή

Η επισημείωση (annotation) πολυμέσων (εικόνες, βίντεο) αποτελεί σημαντική λειτουργία για την ανάλυση ψηφιακού περιεχομένου σε μεγάλη κλίμακα. Η διαδικασία της επισημείωσης πολυμεσικού περιεχομένου έχει στόχο να προσδιορίσει, να οργανώσει και να καταγράψει πληροφορίες και σημαντικά χαρακτηριστικά για μεγάλες συλλογές πολυμέσων. Ένα από τα οφέλη της είναι η καλύτερη αναζήτηση και εξαγωγή πληροφοριών από τα δεδομένα, επιτρέποντας έτσι στους χρήστες να εντοπίζουν και να οργανώνουν ευκολότερα πολυμεσικό περιεχόμενο- που μπορεί να αφορά σε αντικείμενα, πρόσωπα ή συμβάντα- με βάση συγκεκριμένα κριτήρια ή κατηγορίες. Επιπλέον, η επισημείωση είναι σημαντική για την ανάπτυξη και την εκπαίδευση των μοντέλων μηχανικής και βαθιάς μάθησης, καθώς παρέχονται δεδομένα που είναι δυνατόν να χρησιμοποιηθούν για την βελτίωση της ακρίβειας και της απόδοσης των μοντέλων.

Το σύστημα επισημείωσης εικόνων και βίντεο της πλατφόρμας MediaPot σχεδιάζεται με στόχο να παρέχει αυτόματα μεταδεδομένα για τα αρχεία πολυμέσων. Τα μεταδεδομένα καλύπτουν διαφορετικές πτυχές του σημασιολογικού περιεχομένου των πολυμέσων και παράγονται έπειτα από επεξεργασία μέσω μιας σειράς αλγορίθμων βαθιάς μάθησης που βρίσκονται στην αιχμή των τεχνολογικών εξελίξεων. Δίνεται η δυνατότητα στους χρήστες να μπορούν εύκολα να οργανώνουν και να αναζητούν περιεχόμενο με βάση τις ανάγκες τους σε συλλογές μεγάλης κλίμακας. Τα πολυμέσα που υποστηρίζονται στην τρέχουσα έκδοση αφορούν εικόνες, βίντεο καθώς και αντικείμενα 3D.

Πιο συγκεκριμένα, μέσω του συστήματος επισημείωσης υποστηρίζεται η αυτόματη παραγωγή κειμένου φυσικής γλώσσας για την περιγραφή του περιεχομένου, όπως και η εξαγωγή οντοτήτων και κατηγοριών όπως δραστηριοτήτων, προσώπων, αντικειμένων, ειδικών κατηγοριών και άλλες επισημειώσεις. Συγκεκριμένα, υποστηρίζεται η αναγνώριση 16000 διεθνών διασημοτήτων (π.χ. αθλητές, καλλιτέχνες, πολιτικοί), 400 ειδών δραστηριότητας (π.χ. γυμναστική, περπάτημα, μπάσκετ), 6500 αντικειμένων (π.χ. αυτοκίνητο, τραπέζι, άνθρωπος, τρένο) και ειδικών κατηγοριών (π.χ. επαγγέλματα, χρώματα, είδος σκηνής). Επιπλέον, υποστηρίζεται η αναγνώριση περιεχομένου που θεωρείται ακατάλληλο ή σκληρό, καθώς και η ειδική επισημείωση εικόνων τύπου meme. Τέλος, παρέχεται επιπλέον μία διανυσματική αναπαράσταση του πολυμέσου η οποία είναι δυνατό να χρησιμοποιηθεί για την ανάκτηση σημασιολογικά όμοιων στοιχείων περιεχομένου.

Το σύστημα της αυτόματης επισημείωσης (annotation) πολυμεσικού περιεχομένου πρόκειται να δώσει τη δυνατότητα στους χρήστες της πλατφόρμας να κατανοούν και να αξιοποιούν τον μεγάλο όγκο περιεχομένου- ειδησεογραφικού ενδιαφέροντος- που δημοσιεύεται συνεχώς στα μέσα κοινωνικής δικτύωσης. Συνεπώς, περιλαμβάνει αρκετές επιμέρους μεθόδους βαθιάς μάθησης, οι οποίες βρίσκονται στην αιχμή των τεχνολογικών εξελίξεων. Κάθε μία από αυτές έχει μελετηθεί



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

εκτενώς στη σχετική βιβλιογραφία και έχουν παρουσιαστεί ενδιαφέρουσες και καινοτόμες προτάσεις και εφαρμογές, που μας οδήγησαν στην εφαρμογή των συγκεκριμένων μεθόδων και εργαλείων. Στη συνέχεια, παρουσιάζουμε μια πολύ σύντομη σύνοψη των σχετικών εργασιών ανά περιοχή. Έπειτα παρουσιάζονται οι μέθοδοι και τα μοντέλα βαθιάς μάθησης που χρησιμοποιούνται στο πλαίσιο της προτεινόμενης υλοποίησης της επισημείωσης περιεχομένου στην πλατφόρμα MediaPot.

Επιπλέον, παρουσιάζονται συνολικά οι πειραματικές μέθοδοι που ακολουθήθηκαν, η αξιολόγηση των εργαλείων, καθώς και τα αποτελέσματα τους στις δοκιμές που διενεργήθηκαν. Σκοπός της ανάπτυξης και της ενσωμάτωσης των εργαλείων σε ένα περιβάλλον είναι να διευκολύνει τον χρήστη να επεξεργαστεί και να οργανώσει το πολυμεσικό υλικό που αναπαράγεται στο διαδίκτυο. Για τον σκοπό αυτό, η αλληλεπίδραση του χρήστη με τα εργαλεία προτείνεται να πραγματοποιείται σε υψηλό επίπεδο χωρίς να απαιτείται καμία προηγούμενη γνώση του τομέα της βαθιάς μάθησης.

Τέλος, όπως αναφέρθηκε το σύστημα της αυτόματης επισημείωσης πολυμεσικού περιεχομένου περιλαμβάνει διάφορα εργαλεία βαθιάς μάθησης με σκοπό να δίνει τη δυνατότητα στα παραδοσιακά ΜΜΕ να επεξεργάζονται και να αξιοποιούν καλύτερα τις ειδήσεις που αναρτώνται στα μέσα κοινωνικής δικτύωσης. Συνεπώς, παρουσιάζεται συνοπτικά πως αυτά τα μοντέλα- για κάθε μία μέθοδο- ενσωματώνονται και υλοποιούνται με σκοπό να δημιουργηθεί ένα ολοκληρωμένο σύστημα για την επισημείωση πολυμεσικού περιεχομένου (εικόνες, βίντεο).

Για την εκτέλεση των μοντέλων χρησιμοποιείται η τεχνολογία του Nvidia Triton Inference Server[2] που περιλαμβάνει εκτέλεση μοντέλων βαθιάς μάθησης - που υλοποιούνται σε γλώσσα προγραμματισμού Python σε περιβάλλον docker container.

## 2.2 Ανίχνευση Αντικειμένων

### 2.2.1 Υπόβαθρο – Σχετικές Δουλειές

Στον τομέα της **ανίχνευσης αντικειμένων** σε εικόνες οι πλέον δημοφιλείς μέθοδοι σήμερα αφορούν το μοντέλο Faster R-CNN (Ren et al, 2016) το οποίο προτείνει τα Region Proposal Networks (RPN), καθιστώντας τις διαδικασίες εκπαίδευσης και εκτίμησης ακόμα πιο γρήγορες ενοποιώντας τη συνολική αρχιτεκτονική. Το μοντέλο You Only Look Once (YOLO) (Redmon et al, 2015; Redmon & Farhadi, 2016; Wang et al, 2020) είναι λιγότερο πολύπλοκο αλλά πολύ πιο γρήγορο από το R-CNN, επιτρέποντας την ανίχνευση αντικειμένων σε πραγματικό χρόνο. Εκείνο όμως που επιτυγχάνει σήμερα την καλύτερη επίδοση είναι το EfficientDet που χρησιμοποιεί δίκτυα bi-directional feature

[2] https://developer.nvidia.com/nvidia-triton-inference-server



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

pyramid και σύνθετη κλιμάκωση (Tan et al, 2020). Σχετικά με την ανίχνευση αντικειμένων σε βίντεο, που αξιοποιεί τις χρονικές πληροφορίες και τα χαρακτηριστικά των frames των βίντεο, οι ανιχνευτές αντικειμένων διακρίνονται σε flow-based (Zhu et al., 2017; Zhu et al., 2017a; Zhu et al., 2018), LSTM-based (Liu et al., 2019; Liu & Zhu, 2018; Zhang & Kim, 2019), attention-based (Chen et al., 2020; Guo et al., 2017; Wu et al., 2019) tracking-based (Yang et al., 2019; Feichtenhofer et al., 2017) και υβριδικές μεθόδους (Bertasius et al., 2018; Xiao et al., 2017; Wang et al., 2019a). Η πλειονότητα των προσεγγίσεων χρησιμοποιεί το σύνολο δεδομένων ImageNet VID[3] για αξιολόγηση της απόδοσης.

Ακόμη, αναφορικά με την δημιουργία ετικετών για εικόνες το μοντέλο SAM - Segment Anything Model (Kirillov et al., 2023) έχει πολύ καλή απόδοση σε μεγάλα σύνολα δεδομένων χωρίς εκπαίδευση, αντιμετωπίζοντας μερικές προκλήσεις σχετικά με τις διεργασίες ανάγνωσης ετικετών. Αντίθετα, ένα πραγματικά καινοτόμο μοντέλο είναι το RAM - Recognize Anything Model (Zhang et al., 2023), το οποίο καθιερώνει ένα καθολικό κι ενιαίο σύστημα παραγωγής ετικετών σε μεγάλα σύνολα δεδομένων εικόνας-κειμένου. Ο σχεδιασμός του μοντέλου συνδέσει τις ετικέτες εικόνας με την περιγραφή της κι έτσι επιτρέπει την γενίκευση σε κατηγορίες που δεν έχει δει προηγουμένως.

Η επιλογή του κατάλληλου μοντέλου εξαρτάται από το είδος του προβλήματος, τους περιορισμούς χρόνου και πόρων, καθώς και τις απαιτήσεις ακρίβειας και απόδοσης της εφαρμογής. Για παράδειγμα, ένα μοντέλο EfficientDet παρέχει καλή ισορροπία μεταξύ απόδοσης και αποτελεσματικότητας υπολογισμού, καθιστώντας το κατάλληλο για εφαρμογές που απαιτούν ταχύτητα και απόδοση. Από την άλλη πλευρά, ένα μοντέλο Faster R-CNN παρέχει ακρίβεια και αξιοπιστία στην ανίχνευση αντικειμένων, παρόλο που μπορεί να απαιτεί περισσότερους υπολογιστικούς πόρους. Ενώ, τα μοντέλα SAM – Segment Anything Model και RAM – Recognize Anything Model φαίνεται πως αποδίδουν πολύ καλά σε κατηγορίες για τις οποίες δεν έχουν εκπαιδευτεί και χαρακτηρίζονται ως τα πιο καινοτόμα μοντέλα ανίχνευσης αντικειμένων.

### 2.2.2 Μέθοδος

Τα μοντέλα που επιλέχθηκαν να χρησιμοποιηθούν για την ανίχνευση και τον εντοπισμό αντικειμένων σε εικόνες και καρέ βίντεο, είναι τα προ-εκπαιδευμένα μοντέλα Faster R-CNN και RAM - Recognize Anything Model.

Παρόλο που το μοντέλο EfficientDet φαίνεται να επιτυγχάνει την καλύτερη επίδοση σε διάφορες εφαρμογές, έχει ένα σημαντικό μειονέκτημα αναφορικά με την ακρίβεια εντοπισμού των αντικειμένων. Έπειτα από δοκιμές με ενδεικτικό πολυμεσικό περιεχόμενο, διαπιστώσαμε πως στα πλαίσια του συστήματος που μελετάμε το μοντέλο Faster R-CNN (Ren et al., 2015) εντοπίζει με

[3] https://image-net.org/



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

περισσότερη ακρίβεια και σε λιγότερο χρόνο τα αντικείμενα στις εικόνες στις οποίες δοκιμάστηκε και για αυτό το λόγο επιλέχθηκε. Επιπλέον, συνδυάζεται με το μοντέλο InceptionV2 (Ioffe & Szegedy, 2015) που χρησιμοποιείται για την εξαγωγή χαρακτηριστικών και την κατηγοριοποίηση των εικόνων. Η δομή που χρησιμοποιείται αφορά δύο δίκτυα, ένα για την εξαγωγή χαρακτηριστικών (backbone network) και ένα για την ανίχνευση των περιοχών ενδιαφέροντος (region proposal network). Το Faster R-CNN είναι εκπαιδευμένο σε 80 κατηγορίες αντικειμένων από το σύνολο δεδομένων MS COCO.

Επιπλέον, επιλέχθηκε και το μοντέλο RAM με σκοπό να αυξηθούν οι δυνατότητες του συστήματος σχετικά με την ανίχνευση αντικειμένων. Το RAM είναι ένα μοντέλο για την ανίχνευση αντικειμένων που αντιμετωπίζει τις προκλήσεις προηγούμενων μοντέλων αναφορικά με την συλλογή και τη διαχείριση μεγάλων συνόλων δεδομένων, προσφέροντας ευελιξία, γενίκευση σε νέο περιεχόμενο και ισχυρές δυνατότητες επισημείωσης (Zhang et al., 2023). Η ανάπτυξη του RAM περιλαμβάνει τέσσερα βασικά στάδια. Αρχικά, αξιολογούνται ετικέτες εικόνων χωρίς αναφορές μεγάλης κλίμακας μέσω της αυτόματης ανάλυσης του κειμένου. Έπειτα, εκπαιδεύεται ένα προκαταρκτικό μοντέλο για αυτόματη αναφορά, συνδυάζοντας τις δραστηριότητες λεζάντας και δημιουργίας ετικετών, με επίβλεψη από τα αρχικά κείμενα και τις parsed ετικέτες αντίστοιχα. Στη συνέχεια, χρησιμοποιείται ένας μηχανισμός δεδομένων για τη δημιουργία πρόσθετων αναφορών και τον καθαρισμό των ανεπιθύμητων. Τέλος, το μοντέλο επανεκπαιδεύεται με τα επεξεργασμένα δεδομένα και προσαρμόζεται με μια μικρότερη, αλλά ποιοτικά ανώτερη συλλογή δεδομένων.

Τελικά, με την χρήση των δύο μοντέλων το σύστημα καταφέρνει να εντοπίσει συνολικά 6500 αντικείμενα και να δημιουργήσει ένα αποδοτικό σύστημα για την ανίχνευση αντικειμένων που δεν έχει ξανά συναντήσει. Αναγνωρίζεται οποιαδήποτε κατηγορία με υψηλή ακρίβεια, ξεπερνώντας την απόδοση τόσο των πλήρως επιβλεπόμενων μοντέλων όσο και των υπαρχόντων γενικών προσεγγίσεων όπως το CLIP και το BLIP (Zhang et al., 2023).

### 2.2.3 Πειραματική Αξιολόγηση

Για την ανίχνευση αντικειμένων σε εικόνες χρησιμοποιήθηκαν τα προ-εκπαιδευμενα μοντέλα Faster R-CNN InceptionV2 και RAM-14M στο σύνολο δεδομένων MSCOCO. Σημειώσαμε το μέτρο αξιολόγησης mAP που αφορά το σύνολο δεδομένων MSCOCO. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 2-1. Ακόμη, το μοντέλο R-CNN InceptionV2 έχει χρόνο εκτέλεσης/ταχύτητα 58ms.


Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Πίνακας 2-1: Μέση Μέγιστη Ακρίβεια mAP (COCO) για δύο μοντέλα ανίχνευσης αντικειμένων**

| Μοντέλο | mAP(COCO) |
|---|---|
| Faster RCNN Inception v2 | 0.28 |
| RAM-14M | 0.80 |

### 2.2.4 Υλοποίηση και Ενσωμάτωση

Το μοντέλο για την ανίχνευση αντικειμένων σε εικόνες και καρέ βίντεο όπως είδαμε είναι το Faster R-CNN (Ren et al., 2015) σε συνδυασμό με το InceptionV2 (Ioffe & Szegedy, 2015) και το μοντέλο RAM-14M. Για τις εικόνες, ανιχνεύεται και αποθηκεύεται το "πλαίσιο περιορισμού" (bounding box) που περιέχει το αντίστοιχο αντικείμενο, ενώ στα βίντεο παρέχονται και πληροφορίες αναφορικά με τον χρόνο. Ενώ και στις δύο περιπτώσεις περιλαμβάνεται ένα σκορ βεβαιότητας. Είναι εφικτό να βρεθούν εικόνες που περιέχουν ένα συγκεκριμένο αντικείμενο, αν ο χρήστης το επιθυμεί, χρησιμοποιώντας ένα φίλτρο αντικειμένων. Έτσι θα ανιχνευτούν όλες οι εικόνες που έχουν επισημανθεί με μια συγκεκριμένη ετικέτα. Συνολικά, εντοπίζονται 6500 αντικείμενα (π.χ. αυτοκίνητο, τραπέζι, άνθρωπος, τρένο) και ειδικές ετικέτες (π.χ. χρώματα, είδος σκηνής).

**Εικόνα 2-1: Παράδειγμα επισημείωσης και υποτιτλισμού σε μια τυχαία εικόνα**




Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Πίνακας 2-2: Αποτελέσματα επισημείωσης και υποτιτλισμού για την Εικόνα 2-1**

| Εργαλεία | Εικόνα 2-1 | |
|---|---|---|
| Υποτιτλισμός (captioning) | A large crowd of protesters matching down the street with the capitol building in the background. | |
| Επισημείωση (annotation) | Περιεχόμενο | not meme, not disturbing, SFW, not artificial |
| | **Αντικείμενα (objects) & Ετικέτες (tagging)** | building, crowd, fill, gather, hold, large, march, people, protest, protester, sign, street, Cyrus the Great, Danny Carey |

## 2.3 Αυτόματος Υποτιτλισμός Περιεχομένου

### 2.3.1 Υπόβαθρο – Σχετικές Δουλειές

Την τελευταία δεκαετία, πολλές μελέτες έχουν πραγματοποιηθεί αναφορικά με τον αυτόματο **υποτιτλισμό (captioning) πολυμεσικού περιεχομένου** (Bernardi et al., 2017). Αυτές οι μελέτες εστιάζουν στο πρώτο στάδιο κωδικοποίησης εικόνων, όπου χρησιμοποιούνται προεκπαιδευμένα CNN για την εξαγωγή χαρακτηριστικών από την εικόνα, και στο δεύτερο στάδιο αποκωδικοποίησης, όπου πραγματοποιείται η δημιουργία της λεζάντας-περιγραφής (Vinyals et al., 2015; Xu et al., 2015). Πιο σύγχρονες προσεγγίσεις αφορούν διαφορετικούς μηχανισμούς προσοχής (attention mechanism).

Σχετικά με τον υποτιτλισμό βίντεο διακρίνονται δύο κατηγορίες: οι CNN-RNN και RNN-RNN. Οι μέθοδοι CNN-RNN χρησιμοποιούν δίκτυα συνέλιξης 2D ή 3D για την εξαγωγή χαρακτηριστικών και τα διανύσματα χαρακτηριστικών που εξάγονται τροφοδοτούνται στη συνέχεια στην αρχιτεκτονική RNN μέσω ενός γραμμικού Fully-Connected επιπέδου για την δημιουργία ακολουθίας λέξεων (Shen et al., 2017; Hou et al., 2019; Zhang and Peng, 2019; Zhang et al., 2020). Στις μεθόδους RNN-RNN, τα μοντέλα βασίζονται σε δίκτυα RNN για την εξαγωγή χαρακτηριστικών και για την παραγωγή κειμένου (Pasunuru and Bansal, 2017; Wang et al., 2018; Yang et al., 2018). Ο κωδικοποιητής χρησιμοποιεί έναν συνδυασμό CNN με μια παραλλαγή RNN γνωστή ως LSTM ενώ ο αποκωδικοποιητής χρησιμοποιεί LSTM για τη παραγωγή κειμένου.

### 2.3.2 Μέθοδος

Ο υποτιτλισμός είναι δυνατόν να ενισχύσει σημαντικά την ανάκτηση πολυμέσων δημιουργώντας μια περιγραφή κειμένου για κάθε πολυμέσο, περιγράφοντας περαιτέρω το περιεχόμενο. Το μοντέλο που χρησιμοποιείται είναι το OFA (Wang et al., 2022), ένα προηγμένο μοντέλο παραγωγής περιγραφικού



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

κειμένου το οποίο έχει εκπαιδευτεί σε 20 εκατομμύρια ζεύγη εικόνας-κειμένου που είναι δημόσια διαθέσιμα και έχει αναδειχτεί πειραματικά ως ένα από τα πιο αποτελεσματικά συστήματα υποτιτλισμού.

### 2.3.3 Πειραματική Αξιολόγηση

Για τον αυτόματο υποτιτλισμό πολυμεσικού περιεχομένου, δηλαδή για την παραγωγή μιας περιγραφικής λεζάντας, το μοντέλο OFA αξιολογήθηκε στο σύνολο δεδομένων refCOCO και στο ImageNet επιτυγχάνοντας ακρίβεια της τάξης του 94% και 85.6% αντίστοιχα.

**Πίνακας 2-3: Η ακρίβεια του μοντέλου OFA σε σύνολα δεδομένων**

| Σύνολο Δεδομένων | Ακρίβεια |
|---|---|
| refCOCO | 94.03% |
| ImageNet-1k | 85.6% |

### 2.3.4 Υλοποίηση και Ενσωμάτωση

Για τον υποτιτλισμό των πολυμέσων χρησιμοποιείται το μοντέλο OFA και στη συνέχεια τα αποτελέσματα ευρετηριάζονται με την χρήση του Elasticseach. Στην Εικόνα 2-2 φαίνονται δύο παραδείγματα εικόνων και οι αντίστοιχες περιγραφές τους. Είναι εφικτό να παραχθούν οι αντίστοιχες περιγραφές χωρίς ο χρήστης να μας παρέχει σχετικές πληροφορίες που αφορούν κείμενο ή μεταδεδομένα.

**Εικόνα 2-2: Παραδείγματα υποτιτλισμού και επισημείωσης εικόνων (α) – (β)**




Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Πίνακας 2-4: Αποτελέσματα υποτιτλισμού και επισημείωσης για την Εικόνα 2-2 (α) – (β)**

| Εικόνες | Εργαλεία | | |
|---|---|---|---|
| Εικόνα 2-2 (α) | **Υποτιτλισμός (captioning)** | A large crowd of people walking down a street with police officers | |
| | Επισημείωση (annotation) | Περιεχόμενο | not meme, not disturbing, SFW, not AI generated |
| | | Δράσεις (actions) | applauding |
| | | Αντικείμενα (objects) | person, 2 traffic lights, umbrella |
| | | Ετικέτες (tags) | city, crowd, demonstration Large, march, people, police, police officer, protest, protester, street |
| Εικόνα 2-2 (β) | **Υποτιτλισμός (captioning)** | A crowd of people holding protest signs and wearing masks. | |
| | Επισημείωση (annotation) | Περιεχόμενο | not meme, not disturbing, SFW, not AI generated |
| | | Αντικείμενα (objects) | person, sign |
| | | Ετικέτες (tags) | crowd, demonstration, gather, hold, large, march, people, protest, protester, rally, street, wear |

## 2.4 Ανίχνευση Εικόνων Meme

### 2.4.1 Υπόβαθρο – Σχετικές Δουλειές

Το **διαδικτυακό meme** εικόνας είναι ένα ειδικό είδος εικόνας που εμπλουτίζεται με κείμενο και χρησιμοποιείται για να εκφράσει συγκεκριμένες έννοιες ή συναισθήματα όπως χιούμορ, ειρωνεία, σαρκασμό, απογοήτευση ακόμα και μίσος. Το θέμα της ανίχνευσης meme εικόνων δεν έχει λάβει ακόμη σημαντική προσοχή. Το προτεινόμενο εργαλείο βασίζεται στη μεθοδολογία Visual Part Utilization (Koutlis et al., 2022) η οποία βασίζεται στην αρχιτεκτονική Vision Transformer (ViT) για την εκπαίδευση ενός μοντέλου ταξινόμησης εικόνων σε meme/non-meme εκπαιδεύοντας το με


Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

κατάλληλα θετικά και αρνητικά παραδείγματα που δημιουργούνται στη βάση μιας αντιπαραθετικής (contrastive) λογικής.

### 2.4.2 Μέθοδος

Για τον αυτόματο εντοπισμό ή τον χαρακτηρισμό μιας εικόνας με τον όρο meme επιλέχθηκε να χρησιμοποιηθεί το μοντέλο MemeTector που βασίζεται στη μέθοδο Visual Part Utilization. Το μοντέλο εστιάζει στα σημαντικά τμήματα της εικόνας και εκπαιδεύεται ώστε να αναγνωρίζει αποτελεσματικά αν μια εικόνα χαρακτηρίζεται ως meme (Koutlis et al., 2022). Χρησιμοποιεί έναν εκπαιδευμένο μηχανισμό προσοχής σε μια τυπική αρχιτεκτονική ViT ώστε να εστιάζει σε συγκεκριμένα τμήματα της εικόνας που είναι σημαντικά. Το μοντέλο εκπαιδεύεται να κατατάσσει σωστά τις εικόνες meme και τις κανονικές εικόνες, χρησιμοποιώντας ένα σύνολο δεδομένων που περιλαμβάνει τόσο εικόνες meme (Hateful Memes του Facebook), όσο και ένα σύνολο δεδομένων που περιλαμβάνει κανονικές εικόνες (Google's Conceptual Captions). Το MemeTector ξεπερνά τις προηγούμενες προσεγγίσεις στην ανίχνευση εικόνων meme, προσφέροντας την καλύτερη απόδοση αναφορικά με τις προβλέψεις του.

### 2.4.3 Πειραματική Αξιολόγηση

Για την ανίχνευση εικόνων τύπου meme συγκρίνεται το μοντέλο MemeTector που επιλέξαμε με άλλα δημοφιλή μοντέλα. Στον Πίνακα 2-5 αναγράφεται η απόδοση των μοντέλων με όρους μέσης ακρίβειας και παρατηρείται πως το μοντέλο MemeTector επιτυγχάνει ακρίβεια 94.98%

**Πίνακας 2-5: Η μέση ακρίβεια στα διάφορα μοντέλα για την ανίχνευση εικόνων meme**

| Μοντέλο | Ακρίβεια |
|---|---|
| VGG16 | 91.36% |
| ResNet50 | 92.31% |
| EfficientNetB5 | 90.05% |
| ViT | 94.17% |
| **MemeTector** | **94.98%** |

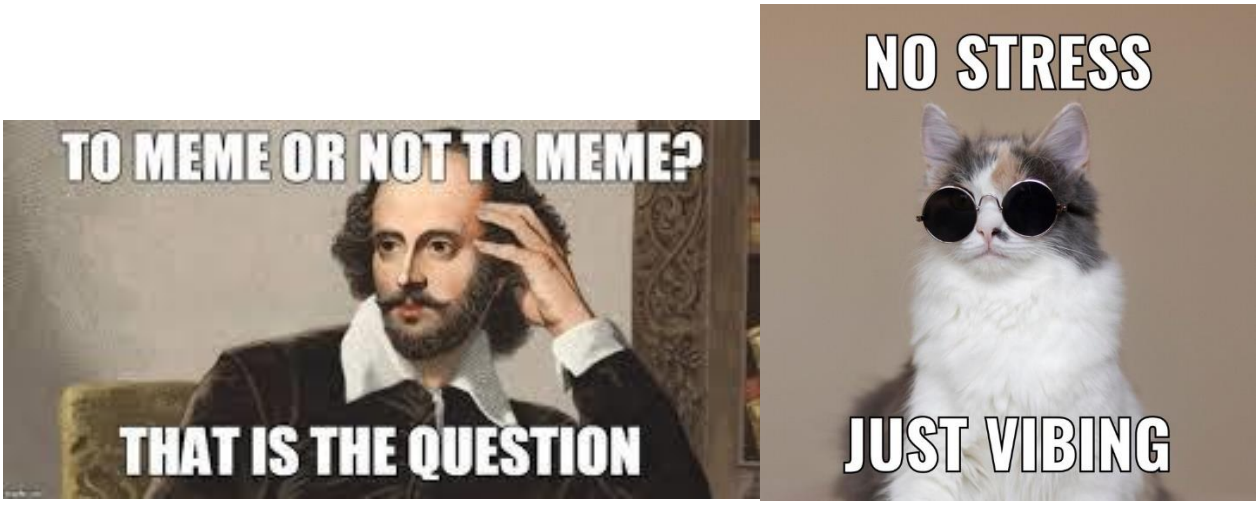
### 2.4.4 Υλοποίηση και Ενσωμάτωση

Εφαρμόζοντας το μοντέλο ανίχνευσης meme εικόνων MemeTector για δύο τυχαία παραδείγματα εικόνων της Εικόνας 2-3, τα αποτελέσματα που προκύπτουν παρουσιάζονται στον Πίνακα 2-6 αναλυτικά.



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Εικόνα 2-3: Παραδείγματα εικόνων meme (α) – (β)**




**Πίνακας 2-6: Αποτελέσματα περιεχομένου για την Εικόνα 2-3 (α) – (β)**

| Εικόνες | Εργαλεία | | |
|---|---|---|---|
| Εικόνα 2-3 (α) | Υποτιτλισμός (captioning) | A man with a beard and a question to meme or not to meme that. | |
| | Επισημείωση (annotation) | **Περιεχόμενο** | **meme**, not disturbing, SFW |
| | | Αντικείμενα (objects) | person |
| | | Ετικέτες (tags) | beard, man, picture, text |
| Εικόνα 2-3 (β) | Υποτιτλισμός (captioning) | No stress just voting cat with sunglasses. | |
| | Επισημείωση (annotation) | **Περιεχόμενο** | **meme**, not disturbing, SFW |
| | | Αντικείμενα (objects) | cat |
| | | Ετικέτες (tags) | animal, cat, glass, sunglass, wear |

## 2.5 Αναγνώριση Προσώπων

### 2.5.1 Υπόβαθρο – Σχετικές Δουλειές

Η **αναγνώριση προσώπων** με εργαλεία βαθιάς μάθησης έχει εξελιχθεί σημαντικά τα τελευταία χρόνια. Οι πιο σύγχρονες τεχνικές εκμεταλλεύονται μεγάλα σύνολα δεδομένων σε συνδυασμό με



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

συνελικτικά νευρωνικά δίκτυα. Μια από τις πιο σημαντικές μεθόδους μάθησης χαρακτηριστικών στην αναγνώριση προσώπων παρουσιάστηκε από τους Taigmain et al. (2014) και βασίστηκε στην εργασία των Quoc V. Le et al. (2012), εκπαιδεύοντας έναν αραιό αυτό-κωδικοποιητή (sparse autoencoder). Επίσης, έχει προταθεί η χρήση τροποποιημένων αρχιτεκτονικών Transformer για την αναγνώριση προσώπων. Εξαιτίας της έλλειψης διάθεσης μεγάλων συνόλων δεδομένων για την αναγνώριση προσώπων επιδιώκεται η ανάπτυξη κατάλληλων εργαλείων και αρχιτεκτονικών για την όσο το δυνατόν καλύτερη αξιοποίηση μικρότερων συνόλων δεδομένων.

Όσον αφορά την αναγνώριση προσώπων σε βίντεο, στόχος είναι να εντοπίζονται πολλά και διαφορετικά πρόσωπα με ακρίβεια κατά τη διάρκεια ενός βίντεο. Η πιο σύγχρονη προσέγγιση αφορά το μοντέλο SENResNet (Zheng et al., 2021)- που είναι παρόμοιο με την υλοποίηση του VGGFace2, διότι χρησιμοποιεί Squeeze και excitation blocks κι ένα μοντέλο Regression Network-based Face Tracking (RNFT) για την εξαγωγή χαρακτηριστικών προσώπου από γειτονικά πλαίσια και πρόβλεψη της θέσης του προσώπου στο επόμενο frame.

### 2.5.2 Μέθοδος

Για την αναγνώριση προσώπου επιλέχθηκε το μοντέλο InceptionResNetV1 το οποίο εκπαιδεύεται να αναγνωρίζει πρόσωπα σε διαφορετικές συνθήκες προσανατολισμού, φωτισμού και εκφράσεων. Η συλλογή δεδομένων (VGGFace2) με βάση την οποία προ-εκπαιδεύεται περιλαμβάνει 3.31 εκατομμύρια εικόνες και σχεδόν 9 χιλιάδες πρόσωπα. Το συγκεκριμένο μοντέλο μπορεί να εφαρμοστεί σε εικόνες και βίντεο. Έπειτα, πραγματοποιείται μια συλλογή σελίδων της Wikipedia- μέσω της διεπαφής BDpedia Lookup API- με αποτέλεσμα τη συλλογή 104,861 εικόνων, με σκοπό να αναγνωριστούν όσο το δυνατόν περισσότερα δημοφιλή πρόσωπα. Συλλέγονται χαρακτηριστικά και χρησιμοποιείται ο αλγόριθμος ομαδοποίησης K-mean ώστε να χωριστούν τα δεδομένα σε ομάδες και έπειτα να ταξινομηθούν. Για τη ταξινόμηση χρησιμοποιήθηκαν μοντέλα SVMs και CNNs. Τελικά, το μοντέλο καταφέρνει να εντοπίζει με ακρίβεια περίπου 16 χιλιάδες πρόσωπα διασημοτήτων.

### 2.5.3 Πειραματική Αξιολόγηση

Για την αναγνώριση προσώπων χρησιμοποιείται το μοντέλο InceptionResNetV1 που προ-εκπαιδεύεται στο σύνολο δεδομένων VGGFace2 που περιέχει 3.31 εκατομμύρια εικόνες με 9 χιλιάδες διαφορετικές αντικείμενα/πρόσωπα, και σε ένα σύνολο σελίδων/εικόνων από την Wikipedia που περιέχει 104,861 εικόνες και συνολικά περίπου 16 χιλιάδες πρόσωπα.

Θεωρούμε κάθε μία από τις 104,861 εικόνες ως query και τις υπόλοιπες ως υποψήφιες (candidates). Για κάθε ζεύγος προσώπου query-candidate υπολογίζεται η ομοιότητα συνημίτονου μεταξύ των διανυσματικών χαρακτηριστικών καταλήγοντας σε $10^{10}$ υπολογισμούς ομοιότητας. Η

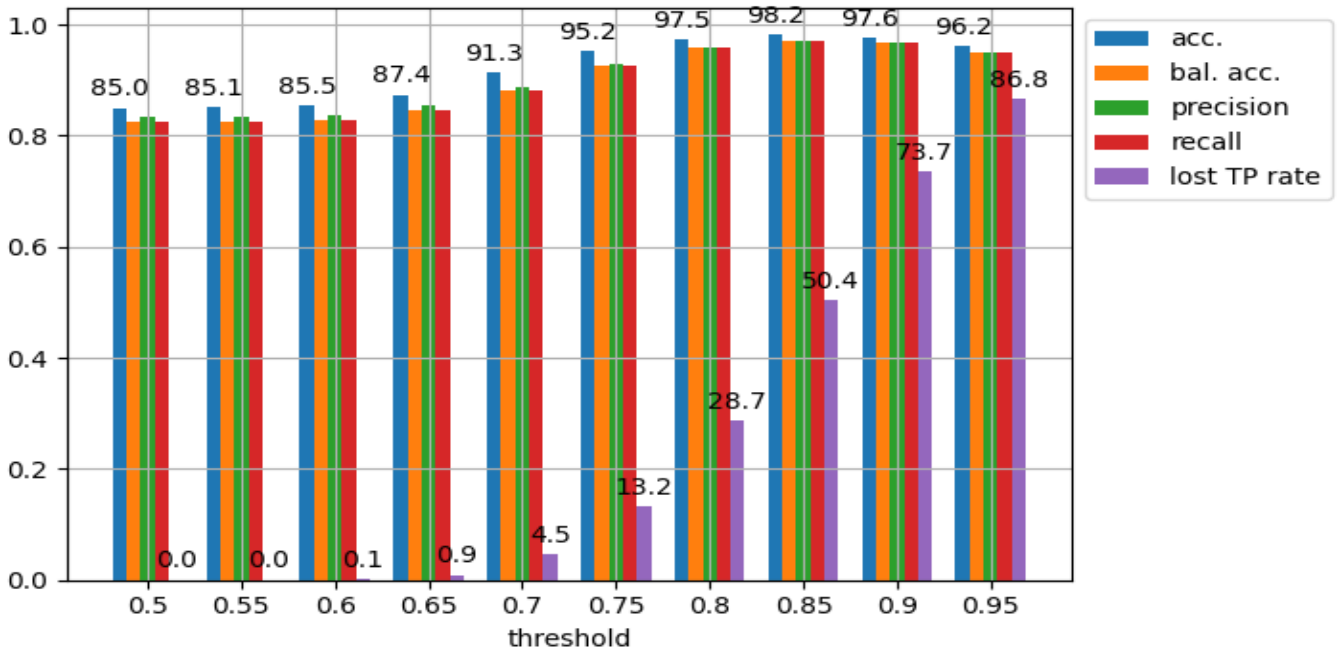


Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

προσωπικότητα, στην οποία ανήκει η εικόνα υποψηφίου (candidate) με τη μέγιστη ομοιότητα, επιστρέφεται εάν το επίπεδο ομοιότητας υπερβεί ένα κατώφλι.

Έγιναν δοκιμές με διάφορες τιμές κατωφλίου (threshold) και υπολογίστηκαν οι μετρικές: accuracy, balanced accuracy, precision (macro), recall (macro). Ακόμη, υπολογίζεται το ποσοστό των true positives (TP) που δεν επιστράφηκαν λόγω του κατωφλιού και ονομάζεται ρυθμός απώλειας TP. Ιδανικά αναζητείται ένα κατώφλι που να παρέχει υψηλή ακρίβεια για ένα ελεγχόμενο επίπεδο απωλειών TP. Αυτές οι μετρικές παρείχαν ενδείξεις για την αποτελεσματικότητα των μοντέλων στην ακριβή αναγνώριση και κατηγοριοποίηση προσώπων από σελίδες της Wikipedia.

**Εικόνα 2-4: Τα μέτρα bal. accuracy , precision, recall, lost TP rate σε σχέση με τις διάφορες τιμές κατωφλιού**




Τέλος υπολογίζεται το μέσο επίπεδο ομοιότητας (average similarity level) μεταξύ εικόνων που αφορούν το ίδιο άτομο και εικόνων που αφορούν διαφορετικό άτομο. Σε κάθε πρόσωπο είναι δυνατόν να αντιστοιχεί παραπάνω από μία εικόνα. Η μέση ομοιότητα για το ίδιο πρόσωπο είναι 0.67 (0.19) και η μέση ομοιότητα για διαφορετικό πρόσωπο είναι 0.02 (0.15).

Παρατηρείται ότι αν οριστεί το κατώφλι ομοιότητας πάνω από 0.7 τότε επιτυγχάνεται ακρίβεια πάνω από 90%. Η καλύτερη απόδοση προκύπτει όταν το κατώφλι έχει τιμή 0.85, τότε η ακρίβεια φτάνει το 98.2% αλλά τα μισά TP χάνονται. Τελικά, η ισορροπία προκύπτει αν το κατώφλι έχει την τιμή 0.75 όπου η ακρίβεια διατηρείται σε υψηλά επίπεδα και τα χαμένα TP σε χαμηλά επίπεδα.



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Τα κύρια ευρήματα των πειραμάτων υπογράμμισαν την αποτελεσματικότητα της προτεινόμενης μεθόδου στην ακριβή αναγνώριση και κατηγοριοποίηση προσώπων από σελίδες της Wikipedia. Τα μοντέλα πέτυχαν υψηλά επίπεδα ακρίβειας, αποδεικνύοντας την εφικτή ενσωμάτωση της μεθόδου σε πραγματικές εφαρμογές.

### 2.5.4 Υλοποίηση και Ενσωμάτωση

Το μοντέλο InceptionResNetV1 εφαρμόζεται εξίσου σε εικόνες και σε βίντεο στο σύστημα, εντοπίζοντας με καλή ακρίβεια, λόγω της πολυπλοκότητας των προσώπων που καλείται να εντοπίσει, τις λεπτομέρειες και τα χαρακτηριστικά των προσώπων. Στο σύστημα εντοπίζονται 16000 διεθνείς διασημότητες (π.χ. αθλητές, καλλιτέχνες, πολιτικοί). Στην Εικόνα 2-5 φαίνονται δύο παραδείγματα αναγνώρισης προσώπου καθώς και τα αποτελέσματα που παράγονται για την κάθε εικόνα παρουσιάζονται στον Πίνακα 2-7.

**Εικόνα 2-5: Παράδειγμα ανίχνευσης προσώπου και δημιουργίας ετικετών (α) – (β)**



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Πίνακας 2-7: Αποτελέσματα ανίχνευσης προσώπου και δημιουργίας ετικετών για την Εικόνα 2-5 (α) – (β)**

| Εικόνες | Εργαλεία | | |
|---|---|---|---|
| Εικόνα 2-5 (α) | Υποτιτλισμός (captioning) | A man sitting at a desk in front of two flags. | |
| | Επισημείωση (annotation) | Περιεχόμενο | not meme, not disturbing, SFW, not AI generated |
| | | **Πρόσωπο (face)** | Kyriakos Mitsotakis |
| | | Αντικείμενα (objects) & Ετικέτες (tags) | chair, person, tie, desk, flag, man, office, politician, president, sit, table |
| Εικόνα 2-5 (β) | Υποτιτλισμός (captioning) | A footballer holding a yellow and a red flag with his arm outstretched. | |
| | Επισημείωση (annotation) | Περιεχόμενο | not meme, not disturbing, SFW |
| | | **Πρόσωπο (face)** | Lionel Messi |
| | | Αντικείμενα (objects) & Ετικέτες (tags) | person, ball, football player, player, soccer player, stadium |

## 2.6 Αναγνώριση Ενεργειών-Δράσεων

### 2.6.1 Υπόβαθρο – Σχετικές Δουλειές

Για την **αναγνώριση ενεργειών-δράσεων** σε βίντεο έχουν χρησιμοποιηθεί δίκτυα two-stream είτε χρησιμοποιώντας μεμονωμένα frames ως είσοδο στο πρώτο stream και optical flow στο δεύτερο stream (Simonyan & Zisserman 2014) είτε χρησιμοποιώντας ένα αργό stream με χαμηλό frame rate και ένα γρήγορο stream με υψηλό frame rate, αλλά χαμηλότερη χωρητικότητα καναλιών, που είναι γνωστή και ως αρχιτεκτονική SlowFast (Feichtenhofer et al., 2019). Η νεότερη αρχιτεκτονική X3D (Feichtenhofer, 2020) πετυχαίνει εξαιρετικές επιδόσεις με το να επεκτείνει μια μικρή αρχιτεκτονική ταξινόμησης εικόνας 2D σε πολλαπλούς άξονες, όπως του χώρου, του χρόνου, του πλάτους και του βάθους. Παρόλο που όλες οι προηγούμενες αρχιτεκτονικές- έως το 2020- βασίζονταν σε συνελίξεις, το 2021 τα μοντέλα που βασίζονται σε μετασχηματιστές, που επεκτείνουν το Vision Transformer (ViT), αρχίζουν να επικρατούν (Dosovitskiy et al., 2020).



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

### 2.6.2 Μέθοδος

Για την αναγνώριση ενεργειών/δράσεων σε βίντεο χρησιμοποιείται το μοντέλο SlowFast R50 (Feichtenhofer et al., 2019) εκπαιδευμένο στο σύνολο δεδομένων Kinetics400 (Smaira, Carreita et al., 2020) που περιέχει 400 διαφορετικές κατηγορίες δράσεων. Το SlowFast R50 είναι μια βελτιωμένη έκδοση του SlowFast και απαιτεί λιγότερους υπολογιστικούς πόρους. Για εικόνες χρησιμοποιείται ένα μοντέλο ResNet152 (He et al., 2016) με βάση το TSN (Wang et al., 2016) για εκπαίδευση σε frame level. Το TSN βασίζεται στην ιδέα της μοντελοποίησης του long-range temporal structure ενώ το ResNet152 είναι ένα βαθύ νευρωνικό δίκτυο βασισμένο στη σειρά των μοντέλων ResNet.

### 2.6.3 Πειραματική Αξιολόγηση

Για την αναγνώριση στατικών ενεργειών/δράσεων, δηλαδή δράσεων σε εικόνες, χρησιμοποιείται το μοντέλο ResNet152 με βάση το TSN για εκπαίδευση σε frame level.

Για βίντεο χρησιμοποιείται ένα προεκπαιδευμένο μοντέλο SlowFast R50 στο σύνολο δεδομένων Kinetics400. Η πειραματική αξιολόγηση βασίστηκε στις δόκιμές μεταξύ του μοντέλου SlowFast R50 και του TimeSforrmer, όπου στον Πίνακα 2-8 παρουσιάζεται η ακρίβεια για κάθε μοντέλο στο σύνολο δεδομένων. Έπειτα, τα δύο μοντέλα δοκιμάζονται σε τρία παραδείγματα βίντεο τριών διαφορετικών δράσεων που κάποια ενδεικτικά frames παρουσιάζονται στις Εικόνες 2-6,7,8 και τα αποτελέσματα αναγράφονται στους Πίνακες 2-9,10,11 αναλυτικά.

Η αξιολόγηση γίνεται αναφορικά με την ταχύτητα φόρτωσης, την ταχύτητα επισημείωσης και την ακρίβεια επισημείωσης. Το μοντέλο SlowFastR50 επιλέχθηκε σε σύγκριση με το TimeSformer. Συγκεκριμένα, το SlowFast παρέχει ισοδύναμη ποιότητα επισήμανσης με το TimeSformer, ενώ φορτώνει τα δεδομένα 4-5 φορές ταχύτερα. Ωστόσο, παρουσιάζει μια ελαφρώς μεγαλύτερη καθυστέρηση στην επεξεργασία των βίντεο σε σύγκριση με το TimeSformer. Επιπλέον, η επεξεργασία μπορεί να είναι πιο αργή για μεγάλα βίντεο λόγω της απαιτούμενης προεπεξεργασίας με τρίτες βιβλιοθήκες. Συνολικά, το SlowFast είναι μια αξιόπιστη επιλογή με καλή απόδοση και λίγο γρηγορότερη φόρτωση σε σύγκριση με το TimeSformer.

**Πίνακας 2-8: Η ακρίβεια για τα δύο μοντέλα SlowFast και TimeSformer**

| | Μοντέλα | |
|---|---|---|
| | SlowFast R50 | TimeSformer |
| Ακρίβεια Kinetics-400 top-1 | 75.3% | 77.9% |



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Εικόνα 2-6: Τυχαία Frames του 1ου τεστ βίντεο «News Video Template»**


**Εικόνα 2-7: Τυχαία Frames του 2ου τεστ βίντεο «Aeroplane Crash»**

**Εικόνα 2-8: Τυχαία Frames του 3ου τεστ βίντεο «Tea Pouring»**



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Πίνακας 2-9: Αποτελέσματα 1ου τεστ βίντεο «News Video Template»**

| 1ο τεστ βίντεο | SlowFast | TimeSformer |
|---|---|---|
| Παραγόμενες Ετικέτες | extinguishing fire 100%<br>news anchoring 75%<br>marching 53%<br>sailing 39%<br>dancing ballet 4%<br>stretching leg 2% | extinguishing fire 99%<br>marching 54%<br>answering questions 38%<br>sailing 26%<br>exercising arm 7% |
| Φόρτωση βίντεο | 1.08 sec | 1.03 sec |
| Φόρτωση μοντέλου | 0.48 sec | 2.21 sec |
| Επισημείωση | 3.87 sec | 1.98 sec |
| Σύνολο | 5.43 sec | 5.22 sec |

**Πίνακας 2-10: Αποτελέσματα 2ου τεστ βίντεο «Aeroplane Crash»**

| 2ο τεστ βίντεο | SlowFast | TimeSformer |
|---|---|---|
| Παραγόμενες Ετικέτες | hurdling 97%<br>extinguishing fire 68%<br>abseiling 49%<br>water skiing 21%<br>driving car 17%<br>flying kite 8% | extinguishing fire 97%<br>using remote controller (not gaming) 74%<br>flying kite 42%<br>skydiving 29% |
| Φόρτωση βίντεο | 1.93 sec | 1.95 sec |
| Φόρτωση μοντέλου | 0.58 sec | 2.45 sec |
| Επισημείωση | 9.80 sec | 4.02 sec |
| Σύνολο | 12.31 sec | 8.42 sec |



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Πίνακας 2-11: Αποτελέσματα 3ου τεστ βίντεο «Tea Pouring»**

| 3ο τεστ βίντεο | SlowFast | TimeSformer |
|---|---|---|
| Παραγόμενες Ετικέτες | making_tea 93% | making tea 99% |
| Φόρτωση βίντεο | 3.11 sec | 3.07 sec |
| Φόρτωση μοντέλου | 0.59 sec | 2.35 sec |
| Επισημείωση | 1.01 sec | 0.41 sec |
| Σύνολο | 4.72 sec | 5.84 sec |

### 2.6.4 Υλοποίηση και Ενσωμάτωση

Το μοντέλο SlowFastR50 επιλέχθηκε σε σύγκριση με το TimeSformer με βάση την πειραματική αξιολόγηση της ενότητας 2.6.3. Στην Εικόνα 2-9 παρουσιάζονται δύο παραδείγματα εικόνων αναγνώρισης δράσεων και δημιουργίας ετικετών και στον Πίνακα 2-12 παρουσιάζονται τα αποτελέσματα των Εικόνων 2-9 (α) – (β).

**Εικόνα 2-9: Παραδείγματα αναγνώρισης δράσεων και δημιουργίας ετικετών στις εικόνες (α) – (β)**



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Πίνακας 2-12: Αποτελέσματα αναγνώρισης δράσεων και δημιουργίας ετικετών για την Εικόνα 2-4 (α) – (β)**

| Εικόνες | Εργαλεία | | |
|---|---|---|---|
| Εικόνα 2-9 (α) | Υποτιτλισμός (captioning) | Silhouettes of people running on the beach at sunset. | |
| | Επισημείωση (annotation) | Περιεχόμενο | not meme, not disturbing, SFW |
| | | **Δράσεις (actions)** | running |
| | | Ετικέτες (tags) | beach, people, run, runner, silhouette, sun, sunset, water, woman |
| Εικόνα 2-9 (β) | Υποτιτλισμός (captioning) | A person playing a guitar with their hands. | |
| | Επισημείωση (annotation) | Περιεχόμενο | not meme, not disturbing, SFW, not AI generated |
| | | **Δράσεις (actions)** | playing bass guitar |
| | | Ετικέτες (tags) | guitar, hand, man, person, play |

## 2.7 Επισημείωση Ακατάλληλου Περιεχομένου

### 2.7.1 Υπόβαθρο – Σχετικές Δουλειές

Τέλος, η ανάγκη **επισημείωσης του περιεχομένου** ως ακατάλληλου με βάση τις κατηγορίες ενοχλητικό (disturbing) και Not-Safe-For-Work (NSFW), οδήγησε στην ανάπτυξη του μοντέλου CM-Refinery, το οποίο χρησιμοποιεί μεγάλα σύνολα δεδομένων για να επεκτείνει αυτόματα με δύσκολα παραδείγματα τα αρχικά σύνολα εκπαίδευσης, βελτιώνοντας τα μοντέλα και μειώνοντας την ανάγκη παρέμβασης του ανθρώπινου παράγοντα. Η μέθοδος εφαρμόζεται σε δύο κατηγορίες περιεχομένου ώστε να αντιμετωπιστούν οι προκλήσεις συλλογής δεδομένων και εισάγει ένα κριτήριο ποικιλομορφίας με σκοπό να βελτιωθεί η απόδοση του. Αξιολογείται στην ανίχνευση δύο κατηγοριών (disturbing, NSFW) επιτυγχάνοντας σημαντικές βελτιώσεις στην ακρίβεια συγκριτικά με άλλα υπάρχοντα μοντέλα και μειώνοντας σημαντικά την ανάγκη για ανθρώπινη εμπλοκή (Sarridis, et al., 2022).



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

### 2.7.2 Μέθοδος

Εξαιτίας του όγκου του διαδικτυακού περιεχομένου που προέρχεται από ανεξέλεγκτες πηγές κρίνεται αναγκαίο να ενσωματωθούν μοντέλα διαχείρισης περιεχομένου με σκοπό την προστασία του συνόλου των χρηστών από τα δεδομένα που δέχονται από τις ψηφιακές πλατφόρμες, τα οποία είναι δυνατόν να προκαλέσουν πολύ έντονη δυσφορία και ανησυχία. Συνεπώς, χρησιμοποιούνται δύο μοντέλα διαχείρισης αναφορικά με το φιλτράρισμα του περιεχομένου των πολυμέσων ώστε να κρίνεται αυτόματα η καταλληλότητά του. Αυτά τα δύο μοντέλα εκπαιδεύονται χρησιμοποιώντας μια επαναληπτική προσέγγιση ή οποία αξιοποιεί μεγάλα σύνολα εικόνων με ένα ημι-αυτόματο σχήμα επισημείωσης (Sarridis et al., 2022). Έτσι τα μοντέλα που προέκυψαν είναι δυνατόν να ανιχνεύουν αν το περιεχόμενο είναι ενοχλητικό (disturbing) ή μη-ασφαλές (Not Safe For Work-NSFW).

### 2.7.3 Πειραματική Αξιολόγηση

Το μοντέλο που χρησιμοποιείται για την ανίχνευση ακατάλληλου περιεχομένου (disturbing/NSFW) είναι το CM-Refinery το οποίο σε δύο σύνολα δεδομένων επιτυγχάνει ακρίβεια πάνω από 95%. Συγκεκριμένα στο σύνολο δεδομένων Pornography-2K για frames επιτυγχάνει ακρίβεια 95.7% ενώ για βίντεο 97.7%. Επίσης στο σύνολο δεδομένων DID επιτυγχάνει ακρίβεια 95%.

Εκπαιδεύονται τα δύο αυτά moderation μοντέλα χρησιμοποιώντας μια επαναληπτική προσέγγιση ή οποία αξιοποιεί μεγάλα σύνολα εικόνων με ένα ημι-αυτόματο σχήμα επισημείωσης. Έτσι τα μοντέλα που προέκυψαν είναι δυνατόν να ανιχνεύουν αν το περιεχόμενο είναι ανατριχιαστικό (disturbing) ή μη-ασφαλές (Not Safe For Work-NSFW). Για τις εικόνες τα αποτελέσματα, συμπεριλαμβανομένου και ενός confident score, αποθηκεύονται και ευρετηριάζονται . Για τα βίντεο, τα μοντέλα εφαρμόζονται σε keyframes και ένα βίντεο κατηγοριοποιείται ως NSFW ή disturbing αν έχει τουλάχιστον μία σκηνή που λαμβάνει αυτή την ετικέτα. Κατά τη διάρκεια της ανάκτησης, οι χρήστες έχουν την δυνατότητα να συμπεριλάβουν ή να αποκλείσουν αυτό το περιεχόμενο ενώ ή διεπαφή χρήστη θολώνει εξαρχής αυτού του είδος το περιεχόμενο. Στη συνέχεια, έχουν την επιλογή να αποκαλύψουν το disturbing περιεχόμενο αν και εφόσον το επιθυμούν.

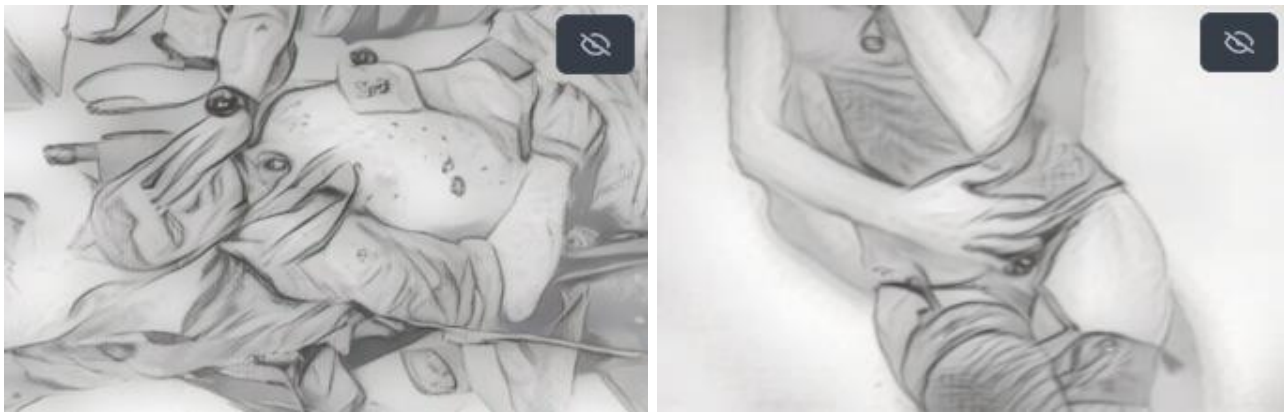
### 2.7.4 Υλοποίηση και Ενσωμάτωση

Το μοντέλο CM-Refinery εφαρμόζεται με σκοπό εντοπίζει με ακρίβεια πολυμεσικό περιεχόμενο το οποίο χαρακτηρίζεται ως ενοχλητικό/ακατάλληλο ή Not-Safe-For-Wok (NSFW). Στην Εικόνα 2-10 (α) – (β) παρουσιάζονται δύο παραδείγματα εικόνων που αφορούν ενοχλητικό/ακατάλληλο περιεχόμενο και περιεχόμενο που χαρακτηρίζεται ως Not-Safe-For-Work. Οι εικόνες αυτές επειδή χαρακτηρίζονται ως «ακατάλληλες» εμφανίζονται φιλτραρισμένες. Επιπλέον ο χρήστης έχει την δυνατότητα να δει την αρχική εκδοχή της εικόνας, αν το επιθυμεί. Τα αποτελέσματα της


Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

επισημείωσης ακατάλληλου/ενοχλητικού περιεχομένου και περιεχομένου NSFW παρουσιάζονται στον Πίνακα 2-13.

**Εικόνα 2-10: Παραδείγματα επισημείωσης ακατάλληλου περιεχομένου (α) - (β)**



**Πίνακας 2-13: Αποτελέσματα επισημείωσης ακατάλληλου περιεχομένου, δημιουργίας ετικετών και υποτιτλισμού περιεχομένου**

| Εικόνες | Εργαλεία | | |
|---|---|---|---|
| Εικόνα 2-10 (α) | Υποτιτλισμός (captioning) | A soldier laying on the ground with his face covered in mud. | |
| | **Επισημείωση (annotation)** | **Περιεχόμενο** | **disturbing**, not meme, SFW |
| | | Ετικέτες (tags) | blood, dead, head, lay, lie, man, person |
| Εικόνα 2-10 (β) | Υποτιτλισμός (captioning) | A woman sitting on a wall with her hands on her thighs. | |
| | **Επισημείωση (annotation)** | **Περιεχόμενο** | **NSFW**, not meme, not disturbing |
| | | Ετικέτες (tags) | person, black, dress, girl, phone, sit, sit on, stocking, wall, woman, young |

## 2.8 Υλοποίηση και Ενσωμάτωση

Τα πολυμέσα με την χρήση της βιβλιοθήκης **Elasticsearch** ταξινομούνται και είναι εφικτό οι χρήστες να τα αναζητήσουν με την χρήση οποιουδήποτε query επιλέξουν που θα περιγράφει καλύτερα το θέμα για το οποίο ενδιαφέρονται. Η βιβλιοθήκη Elasticsearch αποτελεί μια πολύ δημοφιλή υλοποίηση ανοιχτού κώδικα για αναζήτηση, ευέλικτη αποθήκευση και ανάλυση πολυμεσικών δεδομένων σε



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

πραγματικό χρόνο. Κάθε ετικέτα που παράγεται ανήκει και σε μια κατηγορία που προκύπτει με βάση την χρήση των μοντέλων που αναγράφηκαν προηγουμένως. Το γεγονός αυτό καθιστά την ταξινόμηση και την αναζήτηση πιο εύκολη. Για παράδειγμα, στον Πίνακα 2-14 παρουσιάζονται κάποιες από τις ετικέτες που παράχθηκαν σε μία εικόνα. Οι ετικέτες truck και person είναι τύπου object ενώ η ετικέτα person είναι τύπου tag. Κάτι τέτοιο, μας βοηθά να τις διαχωρίσουμε και να ταξινομούμε τα πολυμέσα αντίστοιχα.

**Πίνακας 2-14: Περιγραφή τυχαίων ετικετών κατηγορίας object & tag**

| ετικέτα (tag) | τύπος (type) | τιμή (value) | confidence |
|---|---|---|---|
| truck | object | truck | 78.32% |
| man | tag | man | 85.57% |
| person | object | person | 90.48% |

Ακόμη, χρησιμοποιείται η μέθοδος συμπίεσης **InDistill** (Sarridis et al., 2022) εξαιτίας του μεγάλου αριθμού μοντέλων που χρησιμοποιούνται στο σύστημα επισημείωσης πολυμεσικού περιεχομένου, για την μείωση του μεγέθους και του χρόνου πρόβλεψης των μοντέλων. Το InDistill ενσωματώνει τη διαδικασία εκμάθησης και την τεχνική κατάτμησης καναλιών σε ένα ενιαίο πλαίσιο, ώστε να μεταφέρει τις σημαντικές διαδρομές ροής πληροφοριών από μοντέλα μεγάλου μεγέθους σε ελαφρύτερα μοντέλα.

## Τεχνική υλοποίηση & API

Για την εκτέλεση των μοντέλων χρησιμοποιείται ένας Nvidia Triton Inference Server[4], ο οποίος παρέχει μια ευέλικτη και κλιμακούμενη λύση για την εκτέλεση μοντέλων τεχνητής νοημοσύνης. Το Triton

μπορεί να φιλοξενήσει μοντέλα τεχνητής νοημοσύνης που έχουν υλοποιηθεί στις περισσότερες δημοφιλείς βιβλιοθήκες βαθιάς μάθησης, όπως η TensorFlow, PyTorch, ONNX, κ.λπ., καθιστώντας το σύστημα εύκολα επεκτάσιμο με νέα μοντέλα που βασίζονται στην τεχνολογική εξέλιξη. Επιτρέπει την παράλληλη επεξεργασία μέσα από κάρτες γραφικών με σκοπό τη μεγιστοποίηση της απόδοσης. Υποστηρίζει την επεξεργασία βάσει της μονάδας επεξεργαστή, προσφέρει προηγμένες δυνατότητες όπως σύνολα μοντέλων και συνεχής επεξεργασία, καθώς επίσης παρέχει μία σειρά από προηγμένα χαρακτηριστικά ροή εκτέλεσης (streaming inferencing) και όπως σύνολα μοντέλων (model ensemble).

[4] https://developer.nvidia.com/nvidia-triton-inference-server



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Τα μέρη του συστήματος εκτελούνται σε περιβάλλον docker container[5]. Συγκεκριμένα, για τη φόρτωση των μοντέλων, πρέπει να δημιουργηθεί ένα αρχείο διαμόρφωσης. Το αρχείο διαμόρφωσης θα περιέχει την πλατφόρμα που χρησιμοποιείται (δηλαδή TensorFlow, PyTorch κλπ.), το μέγιστο μέγεθος παρτίδας και το σχήμα εισόδου/εξόδου για τα δεδομένα, μαζί με τα ονόματα των πρώτων και τελευταίων στρωμάτων. Στη συνέχεια, το αρχείο διαμόρφωσης μαζί με το saved_model θα φορτωθούν από τον διακομιστή Triton χρησιμοποιώντας το Docker.

Αξιοποιώντας το Triton, το σύστημα επισημείωσης πολυμέσων παρέχει μια αποδοτική λύση για την εκτέλεση μοντέλων τεχνητής νοημοσύνης, επιτρέποντας την -σχεδόν πραγματικού χρόνου- επισήμανση των πολυμέσων. Πρόκειται για έναν gRPC[6] server που εκτελείται σε γλώσσα Python και ο οποίος είναι δυνατόν να δέχεται αιτήματα μέσω ενός gRPC AsyncIO API. Για να επικοινωνήσουμε τον διακομιστή, η Nvidia υλοποιεί τα πρωτόκολλα HTTP (REST) και gRPC. Επιλέξαμε το πρωτόκολλο gRPC λόγω της ευκολίας υποστήριξης για πολλές γλώσσες, της δυνατότητας ροής δεδομένων και της υψηλής απόδοσης σε σύγκριση με το REST. Η κύρια ιδέα του είναι να καλούμε μια μέθοδο σε έναν server σαν να ήταν τοπικό αντικείμενο. Ο κώδικας πρωτοκόλλου γράφεται σε ένα αρχείο Protocol buffer (.proto), και έπειτα γίνεται η δημιουργία πρόσβασης δεδομένων σε πολλές γλώσσες. Το κάθε μοντέλο πρέπει να υλοποιήσει έναν hundler που δέχεται τρία ορίσματα (arguments), τα οποία χρησιμοποιούνται για την εκτέλεση του μοντέλου. Η γενική ροή του συστήματος είναι απλή και εύκολα επαναχρησιμοποιήσιμη, ενώ η ενημέρωση ενός μοντέλου απαιτεί μόνο τον ορισμό ενός νέου αρχείου config και του αντίστοιχου μοντέλου.

[5] *https://www.docker.com/*

[6] *https://grpc.io/*



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# 3 Αντίστροφη Αναζήτηση πολυμεσικού περιεχομένου

## 3.1 Περιγραφή Προβλήματος

Το σύστημα αντίστροφης αναζήτησης εικόνας και βίντεο παρέχει τη δυνατότητα αποδοτικής ανάκτησης πολυμέσων (εικόνων και βίντεο) με βάση το οπτικό και ακουστικό περιεχομένου τους σε συλλογές μεγάλης κλίμακας. Συγκεκριμένα, παρέχει τη δυνατότητα αναζήτησης σχεδόν όμοιων (near-duplicate) εικόνων και βίντεο, χρησιμοποιώντας άλλες εικόνες ή βίντεο ως ερωτήματα (queries). Η ομοιότητα μπορεί να εκτείνεται από τα αντίγραφα ενός πολυμέσου, έως οπτικοακουστικό περιεχόμενο που αναφέρεται στο ίδιο γεγονός, παρέχοντας τη δυνατότητα ρύθμισής της με βάση τις ανάγκες της εκάστοτε εφαρμογής. Ταυτόχρονα, υποστηρίζει τη δυνατότητα αναζήτησης σκηνής προς σκηνή (shot-to-shot), ενώ δίνει τη δυνατότητα επιλογής πραγματοποίησης της αναζήτησης είτε βάση της εικόνας, είτε του ήχου. Το πολυμεσικό περιεχόμενο οργανώνεται σε συλλογές, το μέγεθος των οποίων μπορεί να εκτείνεται ακόμη και σε εκατομμύρια στοιχεία, δίχως να επηρεάζεται δραστικά η απόδοση της αναζήτησης. Ακόμη, παρέχεται ευρεία υποστήριξη διαδικτυακών πηγών για τη λήψη περιεχομένου.

Η τεχνολογία αναζήτησης βασίζεται σε μία αρθρωτή (modular) αρχιτεκτονική βαθιάς μηχανικής μάθησης δύο σταδίων, σχεδιασμένη για την αποδοτική ανάκτηση πολυμέσων σε συλλογές μεγάλης κλίμακας. Επιμέρους μοντέλα βαθιάς μηχανικής μάθησης εστιάζουν στις ιδιαιτερότητες του κάθε τύπου πολυμεσικού περιεχομένου (εικόνα, ήχος), μετατρέποντάς το σε ιεραρχικές διανυσματικές απεικονίσεις που χρησιμοποιούνται στα επιμέρους στάδια της αναζήτησης, καταλαμβάνοντας μικρό αποθηκευτικό χώρο.

Στο σύστημα που αναπτύσσεται εμπεριέχει δύο ειδών ομοιότητες. Συγκεκριμένα, χρησιμοποιείται η **οπτική ομοιότητα** η οποία λαμβάνει ως είσοδο μία εικόνα που επιλέγεται από τον χρήστη, με σκοπό να εντοπίσει παρόμοιο περιεχόμενο με κριτήριο το θέμα και τα στοιχεία της εικόνας. Έπειτα, αναπτύσσεται η ομοιότητα για την ανίχνευση (κοντινών) διπλότυπων (**Near-Duplicate Detection**), που αφορά το πρόβλημα εντοπισμού διπλότυπων σε πολυμεσικό περιεχόμενο το οποίο αποσκοπεί στον εντοπισμό όμοιων πολυμέσων και ίσως αποτελούν κομμάτι ενός άλλου πολυμέσου.

Το σύστημα αντίστροφης αναζήτησης μπορεί να είναι εξαιρετικά χρήσιμο για χρήστες που δυσκολεύονται να διαμορφώσουν την αναζήτηση του περιεχομένου που επιθυμούν μέσω κειμένου ή δεν είναι εφικτό να προσδιορίσουν με σαφήνεια αυτό που αναζητούν. Ακόμη, η αντίστροφη αναζήτηση παρόμοιου περιεχομένου είναι χρήσιμη όταν η αναζήτηση αφορά πολύ μεγάλο αριθμό δεδομένων, όπου η αναζήτηση μέσω κειμένου θα μπορούσε να είναι ανεπαρκής ή χρονοβόρα.


Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

## 3.2 Υπόβαθρο και Σχετικές Δουλειές

Τα τελευταία χρόνια, έχουν αναπτυχθεί διάφορες μέθοδοι για την ενσωμάτωση της φυσικής γλώσσας στην υπολογιστική όραση. Μέθοδοι όπως η σημασιολογική ανάλυση, χρησιμοποιήθηκαν για την μετατροπή της φυσικής γλώσσας σε χαρακτηριστικά ή σε επιπλέον ετικέτες εκπαίδευσης (Srivastava et al., 2017; Hancock et al., 2018), βελτιώνοντας έτσι την απόδοσης της οπτικής ταξινόμησης.

Το CLIP (Radford et al., 2021) είναι μια σημαντική προσπάθεια σε αυτόν τον τομέα, αξιοποιώντας την φυσική γλώσσα για εργασίες πέραν αυτής και βελτιστοποιώντας την ανάκτηση κειμένου-εικόνας. Πρόσφατες εργασίες έχουν επεκτείνει την έννοια της φυσικής γλωσσικής επίβλεψης σε πρόσθετες μορφές όπως τα βίντεο (Miech et al., 2019; 2020).

Στο πλαίσιο της ανάκτησης εικόνας-κειμένου, η δημιουργία μεγάλης κλίμακας συνόλων δεδομένων είναι κρίσιμη. Ενώ υπάρχουσες συλλογές δεδομένων όπως το Pascal1K και το Flickr8K είναι σημαντικές προσπάθειες για την αυτόματη δημιουργία μεγαλύτερων συνόλων δεδομένων (Ordonez et al., 2011; Mithun et al., 2018). Η διαδικασία δημιουργίας συνόλου δεδομένων του CLIP εκμεταλλεύεται τις ακολουθίες κειμένου που σχετίζονται με εικόνες. Τέλος, το CLIP σχετίζεται ευρύτερα με τον τομέα της μάθησης κοινών μοντέλων όρασης-γλώσσας. Ενώ συναφείς προσεγγίσεις επικεντρώνονται στη σύνδεση της όρασης και της γλώσσας για την επίλυση πολύπλοκων εργασιών όπως η απάντηση σε οπτικές ερωτήσεις κ.α.. Το μοντέλο CLIP έχει μια σημαντική ιδιότητα που αφορά την δυνατότητα του να μαθαίνει μοντέλα όρασης από το μηδέν μέσω φυσικής γλωσσικής επίβλεψης, χωρίς να συνδέει τους δύο τομείς με ένα κοινό μοντέλο προσοχής (Lu et al., 2019; Tan & Bansal, 2019; Chen et al., 2019; Li et al., 2020b).

Σχετικά με τον εντοπισμό και την ανάκτηση κοντινών διπλότυπων πολυμέσων κάποιες από τις πιο πρόσφατες και σημαντικές μέθοδοι αναφέρονται παρακάτω. Οι Yuan et al. (2020) σχετικά με την προσέγγιση coarse-grained, προτείνουν την δημιουργία hash codes για ολόκληρα βίντεο σε συνδυασμό με την απόσταση Hamming για τον υπολογισμό της ομοιότητας. Ο κατακερματισμός πραγματοποιείται μέσω ενός δικτύου που εκπαιδεύεται να διατηρεί τις σχέσεις μεταξύ των βίντεο. Αυτή η μέθοδος προσφέρει αποδοτική ανάκτηση κατάλληλη για εφαρμογές μεγάλης κλίμακας. Ωστόσο, η απόδοση ανάκτησης της μπορεί να είναι περιορισμένη σε σύγκριση με τις προσεγγίσεις fine-grained.

Οι προσεγγίσεις fine-grained είναι μια σημαντική μέθοδος αφορά ένα multi-attentional δίκτυο (Wang et al., 2021), που εξάγει πολλαπλές αναπαραστάσεις βίντεο με σκοπό την σύγκριση και το aggregation. Αυτή η προσέγγιση χρησιμοποιεί σχήματα attention-based που επιτυγχάνουν υψηλή απόδοση ανάκτησης. Ωστόσο οι fine-grained μέθοδοι ενδέχεται να μην προσαρμόζονται επαρκώς σε μεγάλα σύνολα δεδομένων λόγω των απαιτήσεών τους σε υπολογιστικούς πόρους και αποθηκευτικό χώρο.



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σχετικά με το re-ranking των βίντεο, παρουσιάζεται μια μέθοδος που συνδυάζει τις παραπάνω δύο προσεγγίσεις με σκοπό την αποτελεσματική και ακριβή ανάκτηση (Liang & Wang, 2020). Χρησιμοποιείται μια μέθοδος coarse-grained με σκοπό την άμεση κατάταξη και το φιλτράρισμα του βίντεο και έπειτα μια fine-grained μέθοδος για την βελτίωση του υπολογισμού της ομοιότητας για τα επιλεγμένα βίντεο. Αυτή η μέθοδος ενισχύει την απόδοση της ανάκτησης διατηρώντας την αποτελεσματικότητα.

Τέλος, μια από τις πιο σημαντικές μεθόδους αφορά το μοντέλο DnS (Kordopatis-Zilos et al., 2022) εξετάζει την αποτελεσματική ανάκτηση περιεχομένου σε πολύ μεγάλα σύνολα δεδομένων. Οι υφιστάμενες μέθοδοι χωρίζονται σε δύο κατηγορίες: οι προσεγγίσεις που είναι αρκετά λεπτομερείς προσφέρουν πολύ καλή απόδοση αλλά απαιτούν υψηλό υπολογιστικό κόστος, ενώ οι γενικές προσεγγίσεις παρέχουν χαμηλό κόστος αλλά και χαμηλή απόδοση. Το DnS ανήκει στην κατηγορία μοντέλων που χρησιμοποιούν Teacher, Student και Selector δίκτυα για την καλύτερη εκπαίδευση του μοντέλου σε πολλά και νέα δεδομένα με στόχο να επιτευχθεί όσο το δυνατόν καλύτερη απόδοση.

## 3.3 Μέθοδος

Το σύστημα αντίστροφης αναζήτησης για πολυμεσικό περιεχόμενο (εικόνες, βίντεο) επιτρέπει στους χρήστες να αναζητούν παρόμοιο- ή σχεδόν παρόμοιο- περιεχόμενο μέσω εικόνων και βίντεο που παρέχουν οι ίδιοι ως είσοδο για το σύστημα. Προσφέρει προηγμένες δυνατότητες ομοιότητας σε οπτικό επίπεδο χρησιμοποιώντας εργαλεία οπτικής ομοιότητας, επιτρέποντας στους χρήστες να εντοπίζουν με ακρίβεια πολυμέσα που μεταφέρουν μια παρόμοια ιδέα ή έννοια ακόμη κι αν δεν είναι οπτικά πανομοιότυπα. Υποστηρίζεται η αναζήτηση με βάση διάφορες σημασιολογικές έννοιες αλλά και με βάση αυστηρά οπτική αναζήτηση. Είναι σημαντικό να τονιστεί ότι το σύστημα αντίστροφης αναζήτησης δεν αναζητά παρόμοιο περιεχόμενο μέσω κειμένου αλλά μέσω εικόνων ή βίντεο και εμπεριέχει δύο μεθόδους αναζήτησης παρόμοιου/διπλότυπου πολυμεσικού περιεχομένου.

Η **οπτική ομοιότητα** με το προ-εκπαιδευμένο μοντέλο CLIP αναφέρεται στη δυνατότητα του μοντέλου CLIP να κατανοεί το περιεχόμενο των εικόνων μέσω του συνδυασμού της οπτικής πληροφορίας με τη γλωσσική. Το CLIP εκπαιδεύτηκε σε ένα μεγάλο σύνολο δεδομένων εικόνων-κειμένου και μπορεί να εξάγει αναπαραστάσεις για εικόνες και κείμενο με βάση τη σημασία τους. Η οπτική ομοιότητα με το CLIP επιτρέπει την αναζήτηση εικόνων βάσει του περιεχομένου τους, όπως η αναζήτηση εικόνων που περιέχουν συγκεκριμένα αντικείμενα ή έννοιες.

Για να εντοπίζεται η ομοιότητα εικόνων χρησιμοποιείται το προ-εκπαιδευμένο μοντέλο CLIP (Radford et al. 2021), το οποίο εκπαιδεύτηκε σε ένα μεγάλο σύνολο δεδομένων εικόνας-κειμένου, που περιλαμβάνει περίπου 400 εκατομμύρια ζεύγη και κωδικοποιεί το σημασιολογικό περιεχόμενο των πολυμέσων σε ένα πυκνό διάνυσμα σχετικά λίγων διαστάσεων. Χρησιμοποιείται η έκδοση ViT-B/32



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ένας κωδικοποιητής (encoder) εικόνας. Επίσης, χρησιμοποιούμε **Elasticsearch** ως ευρετήριο ώστε να εξάγουμε παρόμοιο πολυμεσικό περιεχόμενο με την χρήση της ταξινόμησης και αναζήτησης **k-Nearest Neighbors** (kNN).

Η αναζήτηση οπτικής ομοιότητας χρησιμοποιεί το μοντέλο CLIP και την αναζήτηση k-NN με σκοπό την πυκνή αναπαράσταση του περιεχομένου μέσω ενός διανύσματος, το οποίο κωδικοποιεί τις σημασιολογικές πληροφορίες του οπτικού και ακουστικού περιεχομένου και την αναζήτηση του αντίστοιχου περιεχομένου. Από την άλλη, η αντίστροφη αναζήτηση εικόνας και βίντεο μπορεί να εντοπίζει περιεχόμενο που έχει οπτικές παραλλαγές. Έτσι, είναι δυνατόν να εντοπίζονται στοιχεία που είναι παρόμοια μεταξύ τους, ακόμα κι αν δε είναι ακριβή αντίγραφα/διπλότυπα. Κάτι τέτοιο αποτελεί χρήσιμο εργαλείο σε εφαρμογές που αφορούν τον εντοπισμό διπλότυπου περιεχομένου, παραβιάσεων πνευματικών δικαιωμάτων, εντοπισμό ψευδών και παραποιημένων ειδήσεων κλπ.

Η **ανίχνευση κοντινών διπλοτύπων** (Near-Duplicate Detection) αφορά τον εντοπισμό βίντεο που είναι σχεδόν πανομοιότυπα παρόμοια μεταξύ τους. Για αυτό το σκοπό, χρησιμοποιείται το μοντέλο Distill-and-Select (DnS) (Kordopatis-Zilos et al., 2022) που παρέχει δύο βασικές λειτουργίες: ευρετηριοποίηση και αναζήτηση εικόνων και βίντεο. Η πρώτη λειτουργία αναλύει το οπτικό περιεχόμενο του πολυμεσικού περιεχομένου που παρέχεται και το εντάσσει στο κατάλληλο ευρετήριο. Στη συνέχεια η δεύτερη λειτουργία αναζητά το ευρετήριο που δημιουργήθηκε για παρόμοιο υλικό ενός βίντεο ή μιας εικόνας και κατατάσσει τα αποτελέσματα που ανακτήθηκαν βάσει της ομοιότητάς τους με την εικόνα ή το βίντεο εισόδου. Το μοντέλο ξεκινά με ένα εξαιρετικά αποδοτικό Teacher δίκτυο και έπειτα εκπαιδεύει τα δίκτυα Student σε διαφορετικά επίπεδα απόδοσης και αποδοτικότητας υπολογισμού. Παράλληλα εκπαιδεύει και ένα δίκτυο Selector το οποίο κατά την δοκιμή ταξινομεί γρήγορα τα διάφορα δείγματα στο κατάλληλο δίκτυο Student ώστε να διατηρηθεί η υψηλή απόδοση. Μια σημαντική παράμετρος του μοντέλου είναι το κατώφλι με βάση το οποίο ορίζεται το εύρος που θα γίνει η αναζήτηση. Για παράδειγμα, αν η τιμή του κατωφλιού είναι πολύ κοντά στη μονάδα τα βίντεο που θα επιστραφούν θα αφορούν πιστά αντίγραφα του αρχικού query βίντεο.

## 3.4 Πειραματική Αξιολόγηση

Αναγράφονται μερικά παραδείγματα των εργαλείων σε σχέση με τα απαιτούμενα καθήκοντα και τα αποτελέσματα των δοκιμών. Το κύριο μέλημα της ανάπτυξης και ενσωμάτωσης των εργαλείων σε ένα περιβάλλον είναι να διευκολύνει τον χρήστη να διαχειρίζεται και οργανώνει το τεράστιο σε ποσότητα πολυμεσικό περιεχόμενο που εντοπίζεται στο διαδίκτυο. Προτείνεται η διεξαγωγή της χρήσης των εργαλείων σε υψηλό επίπεδο αλληλεπίδρασης, χωρίς να απαιτείται προηγούμενη εξοικείωση με παρόμοια ή τα ίδια εργαλεία βαθιάς μάθησης.



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Για να υποστηρίζεται η ομοιότητα χρησιμοποιείται το προ-εκπαιδευμένο μοντέλο CLIP και ένας κωδικοποιητής εικόνας ViT/B-32, που αποκτήθηκε από το GitHub. Για κάθε εικόνα δημιουργείται ένα διάνυσμα που κωδικοποιεί την πληροφορία της, στη συνέχεια οργανώνεται κατάλληλα και τελικά ανακτάται όταν αναζητείται σημασιολογικά παρόμοιο περιεχόμενο μέσω της λειτουργίας προσέγγισης του k-nearest neighbor (kNN) του Elasticsearch.

Στην Εικόνα 3-1 ανακτώνται παρόμοιες εικόνες με βάση την εικόνα που φαίνεται στην πρώτη σειρά η οποία απεικονίζει την Ακρόπολη/Παρθενώνα. Αντίστοιχα στην Εικόνα 3-2 επιστρέφονται παρόμοιο περιεχόμενο σχετικό με την εικόνα που επιλέχθηκε και οι εικόνες αφορούν ένα ελληνικό νησιώτικο τοπίο, όπως το νησί της Σαντορίνης. Τέλος, στην Εικόνα 3-3 φαίνεται πως με βάση την μέθοδο της οπτικής ομοιότητας επιστρέφονται εικόνες για το κίνημα Black Lives Matter στις ΗΠΑ.

**Εικόνα 3-1: Ανάκτηση εικόνων με παρόμοιο περιεχόμενο με βάση την 1η εικόνα (Ακρόπολη/Παρθενώνας)**



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Εικόνα 3-2: Ανάκτηση εικόνων με παρόμοιο περιεχόμενο με βάση την 1η εικόνα (Σαντορίνη/Νησιά)**

**Εικόνα 3-3: Ανάκτηση εικόνων με παρόμοιο περιεχόμενο με βάση την 1η εικόνα (Black Lives Matter)**




Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Το μοντέλο DnS (Distill-and-Select) παρέχει αντίστοιχα δύο βασικές λειτουργίες την ευρετηριοποίηση και αναζήτηση εικόνων και βίντεο. Στην Εικόνα 3-4 παρουσιάζεται ένα παράδειγμα ανάκτησης παρόμοιου περιεχομένου σε βίντεο με κοινό θέμα όλων των βίντεο που ανακτήθηκαν το θέμα του αρχικού βίντεο, που αφορά μια τρομοκρατική ενέργεια στο Παρίσι.

**Εικόνα 3-4: Παράδειγμα ανάκτησης διπλότυπων βίντεο (DnS)**


## 3.5 Υλοποίηση και Ενσωμάτωση

Η τεχνολογία αναζήτησης βασίζεται σε μία αρθρωτή (modular) αρχιτεκτονική βαθιάς μηχανικής μάθησης δύο σταδίων, η οποία έχει σχεδιαστεί για την αποδοτική ανάκτηση πολυμέσων σε συλλογές μεγάλης κλίμακας. Επιμέρους μοντέλα βαθιάς μηχανικής μάθησης εστιάζουν στις ιδιαιτερότητες του κάθε τύπου πολυμεσικού περιεχομένου (εικόνα, ήχος), μετατρέποντάς το σε ιεραρχικές διανυσματικές απεικονίσεις που χρησιμοποιούνται στα επιμέρους στάδια της αναζήτησης, καταλαμβάνοντας μικρό αποθηκευτικό χώρο.

Η αρχιτεκτονική του συστήματος στηρίζεται στο κατανεμημένο (distributed) μοντέλο των microservices, παρέχοντας τη δυνατότητα για κλιμάκωση (scaling) δίχως περαιτέρω αλλαγές στη δομή του συστήματος. Υλοποιείται με χρήση Docker[7] containers, RabbitMQ[8] και Protocol Buffers[9] για

[7] https://www.docker.com/
[8] https://rabbitmq.com/
[9] https://protobuf.dev/



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

την απομόνωση και επικοινωνία των microservices, MongoDB[10] και FAISS[11] για την αποθήκευση των δεδομένων, PyTorch[12] για την εκτέλεση των μοντέλων βαθιάς μηχανικής μάθησης και FastAPI[13] για την παροχή του REST API, ενώ το documentation του συστήματος ακολουθεί το πρότυπο OpenAPI[14]. Οι δυνατότητες αναζήτησης δύναται να ενσωματωθούν σε τρίτες εφαρμογές μέσω του παρεχόμενου REST API όπως στην Ενότητα 2.8.

Το σύστημα αντίστροφης αναζήτησης ακολουθεί μια αρχιτεκτονική που βασίζεται στο DnS και υλοποιεί υπηρεσίες για την εξαγωγή χαρακτηριστικών, την ευρετηριοποίηση και την αναζήτηση διπλότυπων σε πολυμεσικό περιεχόμενο (εικόνες, βίντεο) από τη μία, και από την άλλη στον εντοπισμό παρόμοιου περιεχομένου χρησιμοποιώντας έναν κωδικοποιητή εικόνας. Ο υπολογισμός της οπτικής ομοιότητας παρέχεται στον χρήστη με σκοπό τον εντοπισμό και την ανάκτηση παρόμοιου ή διπλότυπου πολυμεσικού περιεχομένου με βάσει τις απαιτήσεις του.

Με τον ίδιο τρόπο, χρησιμοποιώντας ήδη αναγνωρισμένο περιεχόμενο, η λειτουργία ομοιότητας είναι εφικτό να χρησιμοποιηθεί ώστε να επεκτείνει το περιεχόμενο αυτό με στοιχεία που είναι εννοιολογικά παρόμοια. Όπως είδαμε για την διαδικασία επισημείωσης του περιεχομένου έχουν αξιοποιηθεί διάφορα μοντέλα ώστε να παρέχουν χρήσιμες πληροφορίες και χαρακτηριστικά των πολυμέσων τα οποία είναι φιλικά προς τον χρήστη.

Ο εντοπισμός παρόμοιου περιεχομένου γίνεται όπως αναφέρθηκε ήδη, με την χρήση του CLIP**,** η έκδοση Vit-B/32 του κωδικοποιητή (encoder) εικόνας, το Elasticsearch και η αναζήτηση k-Nearest Neighbors (kNN). Για κάθε εικόνα το σύστημα επισημείωσης λαμβάνει μια πυκνή αναπαράσταση ενός διανύσματος 512 διαστάσεων που κωδικοποιεί τις σημασιολογικές πληροφορίες του οπτικού περιεχομένου της εικόνας. Αυτό στη συνέχεια ευρετηριάζεται στο Elasticsearch και χρησιμοποιείται για την ανάκτηση εικονικά παρόμοιου περιεχομένου μέσω της λειτουργίας προσεγγιστικής αναζήτησης των k πλησιέστερων γειτόνων (kNN) του Elasticsearch. Η δυνατότητα αυτή μπορεί να συνδυαστεί με άλλες λειτουργίες του Elasticsearch όπως οι αναζητήσεις κειμένου, φίλτρων κλπ. Αυτή η λειτουργία μπορεί να χρησιμοποιηθεί σε συνδυασμό με όλα τα άλλα χαρακτηριστικά του Elasticsearch, επιτρέποντας τη συνδυασμένη χρήση ερωτημάτων kNN με αναζητήσεις κειμένου, φίλτρα και συγκεντρώσεις.

[10] https://www.mongodb.com/
[11] https://github.com/facebookresearch/faiss
[12] https://pytorch.org/
[13] https://fastapi.tiangolo.com/
[14] https://swagger.io/specification/



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Τέλος, σχετικά με το μοντέλο DnS και τον εντοπισμό διπλότυπου πολυμεσικού υλικού (βίντεο) κατά την οποία ο χρήστης παρέχει ως είσοδο ένα URL ενός βίντεο περιεχομένου και με βάση την ομοιότητα επιστρέφονται τα αντίστοιχα διπλότυπα βίντεο. Είναι κρίσιμης σημασίας πως ο χρήστης παρέχει πολυμεσικό υλικό με σκοπό να αναζητήσει εξίσου πολυμέσα. Η ανάκτηση του πολυμεσικού υλικού υλοποιείται με βάση μια τιμή κατωφλιού που ορίζει πόσο ίδια είναι τα βίντεο και παίρνει τιμές από 0 ως 1 (0: καμία ομοιότητα, 1: όμοια/διπλότυπα).



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# 4 Κατάτμηση Βίντεο

## 4.1 Περιγραφή Προβλήματος

Ένα βίντεο συνήθως αποτελείται από διαφορετικά πλάνα, τα οποία έχουν συρραφεί σε αργότερο στάδιο επεξεργασίας σε μια σειρά, συχνά με τη χρήση κάποιου οπτικού εφέ, για να δημιουργήσουν το τελικό αποτέλεσμα. Κάθε πλάνο αποτελείται από μια ακολουθία καρέ (frames) που έχουν καταγραφεί χωρίς διακοπή από μια βίντεο κάμερα. Τα πλάνα μπορεί να διαφέρουν μεταξύ τους είτε λόγω νέας γωνίας της κάμερας, είτε λόγω τελείως διαφορετικού σκηνικού.

Ο διαχωρισμός των πλάνων είναι ιδιαίτερα χρήσιμος για να παραχθούν αρκετά μεταδεδομένα που μπορούν να χρησιμεύσουν στην κατηγοριοποίηση και την αποθήκευση ενός βίντεο, ενώ όπως θα δούμε και στην Ενότητα 5, είναι απαραίτητος για τη δημιουργία περίληψης βίντεο. Ωστόσο, τα υπάρχοντα format αποθήκευσης βίντεο, δεν παρέχουν πληροφορίες για το που αλλάζουν τα πλάνα. Έτσι, έχουν δημιουργηθεί αρκετοί αλγόριθμοι, οι οποίοι είναι σε θέση να αναγνωρίζουν την αλλαγή πλάνων.

Παράλληλα, πολλά βίντεο, τα οποία ανεβαίνουν κυρίως σε Μέσα Κοινωνικής Δικτύωσης (ΜΚΔ) από χρήστες, είναι τραβηγμένα από κινητά, ή από μηχανές τύπου GoPro. Αυτά τα βίντεο είναι κατά κανόνα ενός πλάνου και μπορούν να διαρκούν από λίγα δευτερόλεπτα έως αρκετά λεπτά. Είναι εξίσου χρήσιμο να μπορέσουν και τα συγκεκριμένα βίντεο να διαχωριστούν σε υποπλάνα, ανάλογα με τις αλλαγές που συμβαίνουν κατά τη διάρκεια τους.

## 4.2 Υπόβαθρο

Ο στόχος της κατάτμησης βίντεο είναι να τμηματοποιήσει το βίντεο στα πλάνα από τα οποία αποτελείται. Αυτό το πρόβλημα έχει μελετηθεί σε μεγάλο βαθμό, με τους τελευταίους αλγόριθμους να είναι αρκετά αποτελεσματικοί.

Οι πρώτες προσπάθειες κατάτμησης βίντεο επικεντρώθηκαν στο να συγκρίνουν τα pixel διαδοχικών καρέ (Zhang et al., 1993), ή αργότερα να εξάγουν χαρακτηριστικά των καρέ από τα κυρίαρχα χρώματα, με τεχνικές όπως ιστογράμματα χρωμάτων, διανύσματα συνοχής χρωμάτων (Pass et al., 1996) και SURF περιγραφείς (Apostolidis et al., 2014).



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Πιο πρόσφατα, υπήρξε στροφή προς την Τεχνητή Νοημοσύνη (ΤΝ) και συγκεκριμένα τη βαθιά μάθηση (deep learning). Στην εργασία (Gygli et al., 2018) χρησιμοποιούνται 3D συνελικτικά δίκτυα (CNN), με την τρίτη διάσταση να είναι ο χρόνος. Αν και είναι υπολογιστικά σημαντικά πιο πολύπλοκα, καταφέρνουν να έχουν πολύ βελτιωμένα αποτελέσματα σε ακρίβεια και ταχύτητα. Στο πλαίσιο του MediaPot, για την κατάτμηση βίντεο σε πλάνα, αξιοποιούμε το μοντέλο TransNet V2 (Souček et al., 2020), το οποίο πάλι στηρίζεται σε 3D συνελικτικά δίκτυα, αλλά μειώνει τις ανάγκες υπολογιστικής ισχύς χρησιμοποιώντας διεσταλμένες συνελίξεις.

Παράλληλα, για την κατάτμηση σε υποπλάνα, υπάρχουν δύο προσεγγίσεις. Η πρώτη προσέγγιση θεωρεί ένα υποπλάνο, μια σειρά καρέ τα οποία δεν διαφέρουν σημαντικά μεταξύ τους. Έτσι, για να ορίσουν τα υποπλάνα ενός βίντεο, συγκρίνουν την ομοιότητα διαδοχικών καρέ. Οι μέθοδοι σύγκρισης διαδοχικών καρέ είναι αντίστοιχες με μεθόδους για την γενική κατάτμηση σε πλάνα, και χρησιμοποιούν χρωματικά ιστογράμματα (Dumont et al., 2001, Pan et al., 2007) ή συγκρίνουν πόσο σημαντικά αλλάζουν τα καρέ σε ένα χρονικό παράθυρο. Στην (Ojutkangas et al., 2012), υπολογίζουν την φωτεινότητα, καθώς και την κίνηση της κάμερας και των αντικειμένων κάθε καρέ, χρησιμοποιώντας YUV ιστογράμματα και διανύσματα οπτικής ροής, και ορίζουν την αλλαγή υποπλάνου εκεί που τα παραπάνω χαρακτηριστικά εμφανίζουν κάποια σημαντική διαφορά.

Η δεύτερη προσέγγιση προσπαθεί να διαχωρίσει υποπλάνα με βάση τις διαφορές της δραστηριότητας της κάμερας σε διαδοχικά καρέ. Διάφορες υλοποιήσεις υπολογίζουν τις κινήσεις της κάμερας υπολογίζοντας διανύσματα θέσης με βάση αντικείμενα που υπάρχουν σε διαδοχικά καρέ (Kim et al, 2000). Η υλοποίηση της (Cooray et al., 2010) πέρα από τα διανύσματα θέσης, μελετάει διάφορους υπολογισμούς πεδίου οπτικής ροής, χρησιμοποιώντας τοπικούς περιγραφείς, όπως οι SIFT και SURF (Lowe, 1999; Bay et al., 2006), και τη χρήση του PLK αλγορίθμου (Bouguet, 2001). Στο MediaPot, χρησιμοποιούμε μία υλοποίηση, που ανήκει στην πρώτη προσέγγιση, και υπολογίζει με μαθηματικά εργαλεία όπως η ομοιότητα συνημιτόνων τα σκορ ομοιότητας καρέ, για να ομαδοποιήσει διαδοχικά καρέ σε υποπλάνα.

## 4.3 Μέθοδος

### 4.3.1 Κατάτμηση σε πλάνα

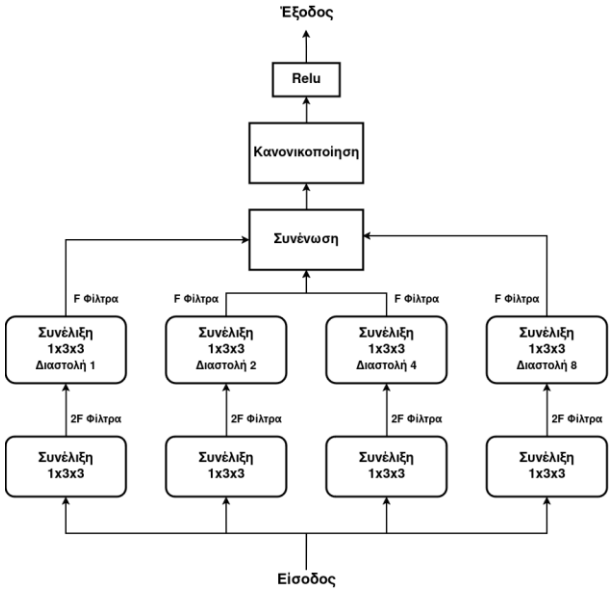
Το μοντέλο TransNet V2 βασίζεται σε 3D συνελικτικά δίκτυα, τα οποία είναι διεσταλμένα, δηλαδή παρουσιάζουν κενά ανάμεσα στα στοιχεία τους. Η διαστολή αυτή παίρνει διαφορετικές τιμές (1, 2, 4, 8), κατά τη χρονική διάσταση. Έτσι, οι συνολικοί παράμετροι που χρησιμοποιεί το δίκτυο μειώνονται σημαντικά. Ως είσοδο, το μοντέλο παίρνει ένα βίντεο σε ακολουθία καρέ, και επιστρέφει ένα σκορ για



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

κάθε καρέ. Το σκορ αυτό υποδηλώνει πόσο πιθανό είναι το καρέ να αποτελεί όριο ενός πλάνου. Έτσι στο τέλος, ξεχωρίζοντας τα πιο σημαντικά καρέ με βάση το σκορ τους, μπορούμε να κατατμήσουμε ένα βίντεο. Κάθε 3D συνελικτικό δίκτυο υλοποιείται στην πραγματικότητα από δύο διαδοχικά συνελικτικά δίκτυα, το πρώτο 2D και το δεύτερο 1D. Ένα τέτοιο άνοιγμα έχει αποδειχθεί (Oord et al., 2016) πως βοηθάει στη διαδικασία της εκπαίδευσης. Στο παρακάτω διάγραμμα φαίνεται ακριβώς το πώς υλοποιείται ένα διεσταλμένο 3D συνελικτικό δίκτυο, με **F** φίλτρα:

**Εικόνα 4-1: Διεσταλμένο 3D συνελικτικό δίκτυο**




Το TransNet V2 αποτελείται από πολλαπλά μπλοκ διεσταλμένων συνελικτικών δικτύων, τα οποία εναλλασάνονται από στρώματα που κανονικοποίησης, και στρώματα επιλογής μέγιστης τιμής (max pooling). Ταυτόχρονα, κάθε δεύτερο συνελικτικό περιέχει μια σύνδεση παράλειψης, ακολουθούμενη από ένα στρώμα επιλογής χωρικού μέσου όρου (spatial average pooling) για να μειώσει τη χωρική διάσταση.

Εκτός από τα χαρακτηριστικά που προκύπτουν από τα διεσταλμένα συνελικτικά δίκτυα, στο μοντέλο προσθέτονται επίσης χαρακτηριστικά για κάθε καρέ από ιστογράμματα χρωμάτων, τα οποία εξάγουν ομοιότητες με βάση τα RGB χρώματα του καρέ, και χαρακτηριστικά ομοιότητας καρέ. Τα χαρακτηριστικά ομοιότητας κάθε καρέ συγκρίνονται με τα προηγούμενα και τα επόμενα 50 καρέ με ομοιότητα συνημίτονου (cosine similarity) και μεταφέρονται με ένα πυκνό στρώμα στο δίκτυο.

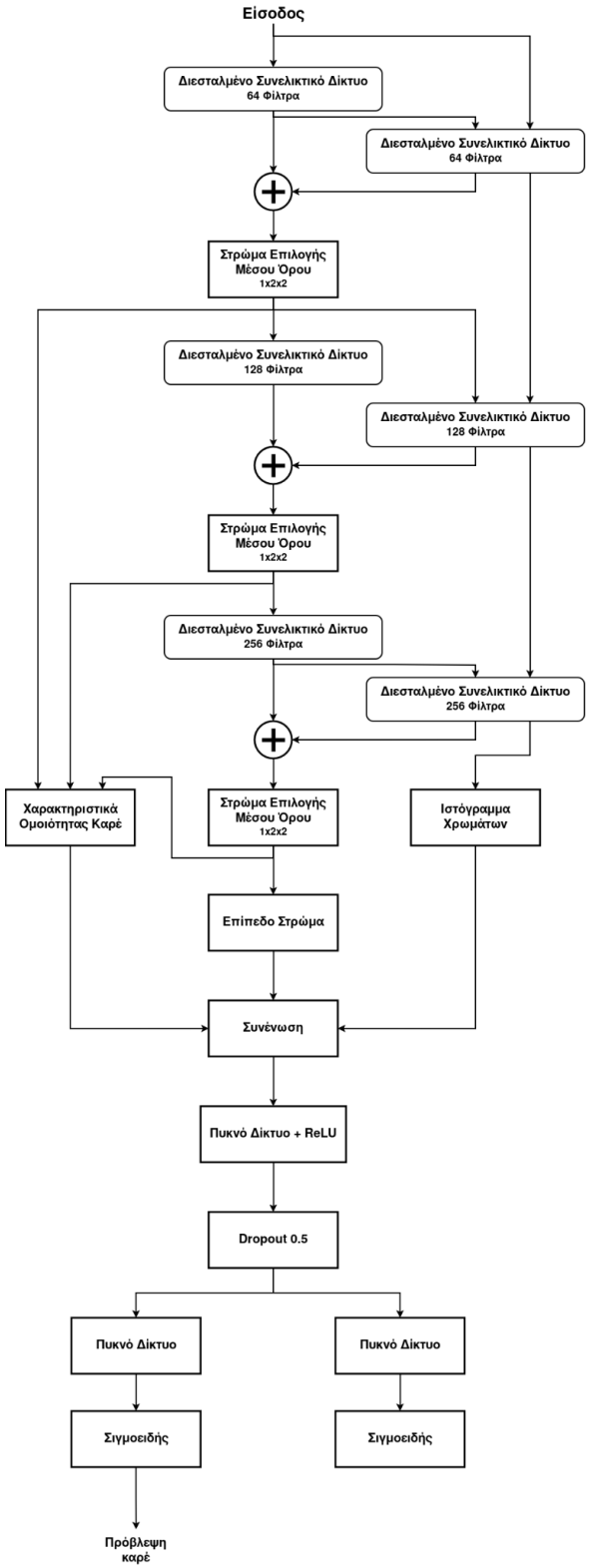
Για την τελική πρόβλεψη, χρησιμοποιούνται δύο κεφάλες για την ανίχνευση των καρέ που αποτελούν όρια ενός πλάνου. Η πρώτη ανιχνεύει ένα όριο τη φορά, ενώ η δεύτερη ανιχνεύει όλα τα καρέ που είναι όρια μιας ακολουθίας. Ωστόσο, η δεύτερη κεφαλή χρησιμοποιείται μόνο κατά τη διάρκεια της εκπαίδευσης.

Η συνολική αρχιτεκτονική του TransNet V2 παρουσιάζεται στο παρακάτω διάγραμμα:



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Εικόνα 4-2: Το δίκτυο TransNet V2 που χρησιμοποιείται στο MediaPot**






Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

### 4.3.2 Κατάτμηση σε υποπλάνα

Όταν ένα βίντεο αποτελείται από ένα μόνο πλάνο, ή όταν είναι σημαντικό να διαχωριστεί ένα πλάνο σε περισσότερα κομμάτια, γίνεται διαχωρισμός σε υποπλάνα σύμφωνα με το μοντέλο της εργασίας (Apostolidis et al., 2018).

Το προτεινόμενο μοντέλο τμηματοποιεί πλάνα (ή βίντεο που αποτελούνται από ένα μόνο πλάνο) σε υποπλάνα, ανιχνεύοντας οπτικά συνεκτικά κομμάτια, δηλαδή συνεχόμενα καρέ τα οποία διαφέρουν ελάχιστα και παρουσιάζουν οπτική συνέχεια. Ο συνηθισμένος ρυθμός καταγραφής καρέ είναι 30 καρέ το δευτερόλεπτο, ενώ οι κάμερες στα smartphones και τις GoPro, όπου συνήθως τα βίντεο ενός πλάνου καταγράφονται, έχουν ακόμα υψηλότερο, γύρω στα 240 καρέ το δευτερόλεπτο. Συνεπώς, διαδοχικά καρέ δεν παρουσιάζουν ιδιαίτερη διαφορά μεταξύ τους. Αυτό σημαίνει πως χρειάζεται να συγκριθούν καρέ δειγματοληπτικά, κρατώντας λίγα καρέ ανά δευτερόλεπτο, δηλαδή περίπου το 10% των καρέ του αρχικού βίντεο.

Η σύγκριση των καρέ γίνεται με βάση τις κινήσεις της κάμερας που καταγράφει το βίντεο είτε στον τρισδιάστατο χώρο, είτε στην αλλαγή της εστίασης (zoom in / out). Ο αλγόριθμος εξάγει χαρακτηριστικά χωροχρονικής (spatio-temporal) ανάλυσης κίνησης, εξετάζοντας ζευγάρια γειτονικών καρέ. Συγκεκριμένα, κάθε καρέ χωρίζεται σε τέσσερα τεταρτημόρια (πάνω/κάτω αριστερά/δεξιά), και οι πιο κυρίαρχες γωνίες κάθε τεταρτημόριου συμμετέχουν στον υπολογισμό του διανύσματος τοπικής οπτικής ροής του ζευγαριού γειτονικών καρέ, με τη χρήση του PLK αλγορίθμου (Bouguet, 2001). Παίρνοντας τον μέσο όρο των τεσσάρων διανυσμάτων, υπολογίζουμε το διάνυσμα της οπτικής ροής του ζευγαριού καρέ, το οποίο συμβολίζει την κίνηση που επικρατεί στο καρέ. (Εικόνες 4-3α και 4-3β). Σε περίπτωση διαγώνιας κίνησης, το διάνυσμα αναλύεται στον κάθετο και τον οριζόντιο άξονα του Ευκλείδειου χώρου (Εικόνα 4-3γ).

Η παραπάνω διαδικασία ανιχνεύει κινήσεις της κάμερας στις δύο διαστάσεις (κάθετη και οριζόντια κίνηση) για το ζευγάρι καρέ που γίνεται η ανάλυση. Για την ανίχνευση κίνησης στην τρίτη διάσταση (βάθος), επαναλαμβάνεται ο υπολογισμός των διανυσμάτων κάθε τεταρτημόριου. Ωστόσο, πριν τον υπολογισμό του μέσου διανύσματος οπτικής ροής, τα διανύσματα του πάνω και του κάτω αριστερά τεταρτημόριου αντιστρέφονται, ώστε να δείχνουν προς την αντίθετη κατεύθυνση. Έτσι, στην περίπτωση που υπάρχει κίνηση μόνο στους δύο πρώτους άξονες (κάθετος και οριζόντιος), το τελικό διάνυσμα θα είναι μηδενικό (Εικόνα 4-4α). Αντίθετα, όταν υπάρχει κίνηση στο βάθος, το τελικό διάνυσμα θα δείχνει δεξιά για κίνηση προς τα μέσα (zoom in, Εικόνα 4-4β) και αριστερά για κίνηση προς τα έξω (zoom out, Εικόνα 4-4γ).

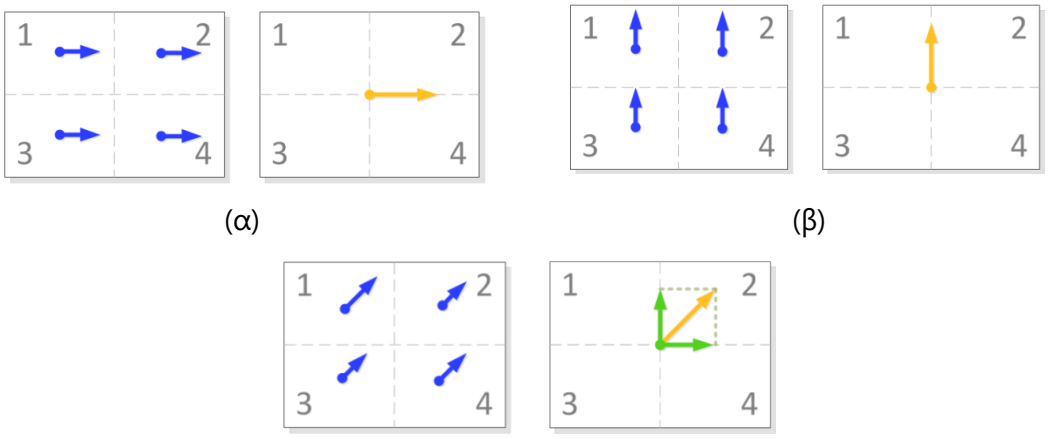


Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Εικόνα 4-3: Παράδειγμα διανυσμάτων οπτικής ροής στους δύο άξονες**

Υπολογισμός διανυσμάτων κίνησης στους δύο άξονες (οριζόντιος και κάθετος) για (α) δεξιά κίνηση (β) πάνω κίνηση (γ) διαγώνια κίνηση

Πηγή (Apostolidis et al., 2018)



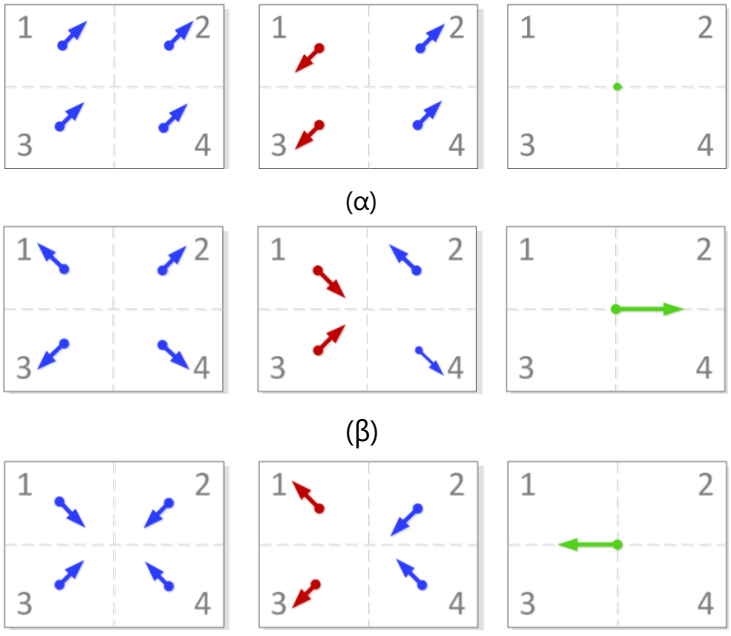

(α) (β)

(γ)

**Εικόνα 4-4: Παράδειγμα διανυσμάτων οπτικής ροής στον τρίτο άξονα**

Υπολογισμός διανυσμάτων κίνησης στον τρίτο άξονα (βάθος) για (α) κίνηση μόνο στους δύο άξονες (κάθετος και οριζόντιος) (β) Κίνηση προς τα μέσα (zoom in) (γ) Κίνηση προς τα έξω (zoom out)

Πηγή (Apostolidis et al., 2018)




(α)

(β)

(γ)

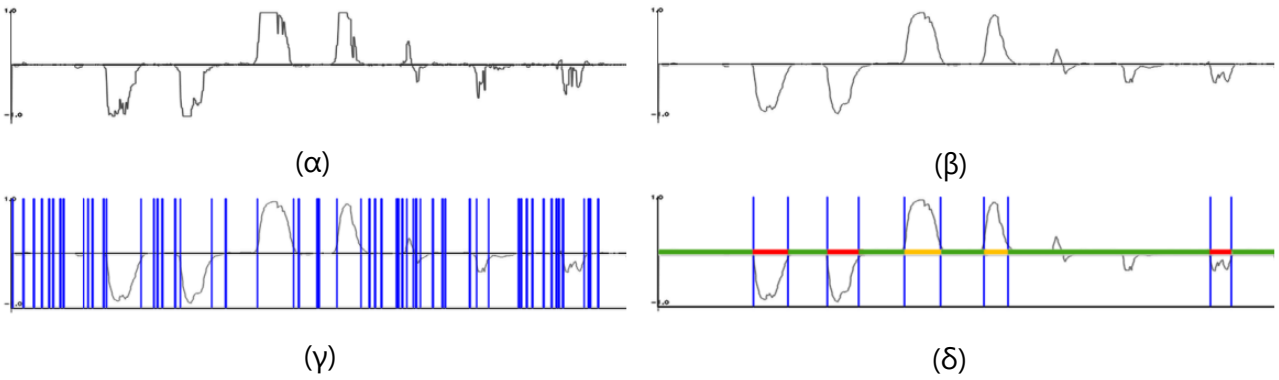


Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Στο τέλος, η μέθοδος καταλήγει με ένα τρισδιάστατο διάνυσμα οπτικής ροής, *V*, το οποίο αναλύεται σε τρεις συνιστώσες, *V = (Vx, Vy, Vz)*, όπου η κάθε συνιστώσα δηλώνει την κίνηση σε κάθε άξονα (οριζόντιο, κάθετο και βάθος) στο ζευγάρι καρέ που μελετάται. Για όλη τη διάρκεια του πλάνου, συλλέγονται τα διανύσματα οπτικής ροής, και δημιουργούνται τρεις χρονοσειρές, που έχουν τις τιμές των *Vx*, *Vy* και *Vz* για κάθε ζευγάρι γειτονικών καρέ. Η κάθε χρονοσειρά περνάει από ένα χαμηλοπερατό φίλτρο, κόβοντας μικρές αλλαγές που οφείλονται σε θόρυβο. Οι τελικές χρονοσειρές συμβολίζονται με *V'x*, *V'y* και *V'z*.

Για τον υπολογισμό των υποπλάνων, η μέθοδος υπολογίζει πόσες φορές υπάρχει αλλαγή στο πρόσημο κάθε χρονοσειράς, καθώς αλλαγή στο πρόσημο σημαίνει πως υπάρχει και αλλαγή στην κίνηση της κάμερας. Τα τελικά υποπλάνα ορίζονται ως τα κομμάτια τα οποία διαρκούν περισσότερο από 1 δευτερόλεπτο και στα οποία έστω μία χρονοσειρά παρουσιάζει περισσότερες από *t* αλλαγές στο πρόσημο της. Οι συγγραφείς της μεθόδου προτείνουν την επιλογή *t = 12*.
Η ανάλυση της χρονοσειράς του οριζόντιου άξονα *Vx* φαίνεται στην παρακάτω εικόνα:

**Εικόνα 4-5: Παράδειγμα διανυσμάτων οπτικής ροής στον τρίτο άξονα**
Υπολογισμός υποπλάνων χρησιμοποιώντας τα διανύσματα οπτικής ροής (α) Η αρχική χρονοσειρά με τις τιμές του *Vx* (β) Η χρονοσειρά V'x μετά το χαμηλοπερατό φίλτρο (γ) Τα σημεία στα οποία η χρονοσειρά μηδενίζεται και αλλάζει το πρόσημό της (δ) Τα όρια των προτεινόμενων υποπλάνων. Τα πράσινα κομμάτια δείχνουν ελάχιστη ή καθόλου κίνηση, τα κόκκινα κίνηση προς τα αριστερά και τα πορτοκαλί κίνηση προς τα δεξιά.
Πηγή (Apostolidis et al., 2018)





Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

## 4.4 Πειραματική Αξιολόγηση

### 4.4.1 Κατάτμηση σε πλάνα

Το μοντέλο TransNet V2 αξιολογείται με βάση τη μετρική F-score, και συγκρίνεται με άλλα ανταγωνιστικά μοντέλα (Souček et al., 2020). Τα σύνολα δεδομένων που χρησιμοποιούνται είναι τα ClipShots (Shitao Tang et al 2018), BBC (Baraldi et al., 2015), και RAI (Baraldi et al., 2015). Παρατηρούμε ότι στα πρώτα δύο σύνολα δεδομένων, ClipShots και BBC, το TransNet V2 παρουσιάζει σημαντική βελτίωση από τα υπόλοιπα ανταγωνιστικά μοντέλα της βιβλιογραφίας. Στο τρίτο σύνολο δεδομένων, RAI, το TransNet V2 παρουσιάζει μικρότερο F-score σε σχέση με το πρώτο μοντέλο, TransNet. Ωστόσο, η διαφορά θεωρείται οριακή, και σε συνδυασμό με τη μεγάλη βελτίωση της απόδοσης στα άλλα δύο σύνολα δεδομένων, η επιλογή του TransNet V2 είναι δικαιολογημένη.

**Πίνακας 4-1: Σύγκριση F-score του μοντέλου TransNet V2 που χρησιμοποιείται στο MediaPot με άλλα ανταγωνιστικά μοντέλα**

| Μοντέλο | ClipShots | BBC | RAI |
|---|---|---|---|
| TransNet | 73.5 | 92.9 | **94.3** |
| Hassanien et al | 75.9 | 92.6 | 93.9 |
| Tang et al | 76.1 | 89.3 | 92.8 |
| **TransNet V2** | **77.9** | **96.2** | 93.9 |

### 4.4.2 Κατάτμηση σε υποπλάνα

Οι συγγραφείς της εργασίας (Apostolidis et al., 2018) δημιούργησαν ένα νέο σύνολο δεδομένων για να αξιολογήσουν μεθόδους κατάτμησης σε υποπλάνα.

Το σύνολο δεδομένων αυτό αποτελείται συνολικά από 33 βίντεο, όλα ενός πλάνου. Τα 15 βίντεο είναι τραβηγμένα από τους ίδιους και έχουν συνολική διάρκεια 6 λεπτά, τα 5 βίντεο προέρχονται από το YouTube, με συνολική διάρκεια 17 λεπτών, και τα 13 τελευταία βίντεο προέρχονται από πλάνα ταινιών, με συνολική διάρκεια 46 λεπτών.
Αξιολογούν το προτεινόμενο μοντέλο συγκρίνοντας τις μετρικές Precision, Recall και F-Score με άλλα ανταγωνιστικά μοντέλα της βιβλιογραφίας. Παρατηρούμε ότι η προτεινόμενη μέθοδος παρουσιάζει με μεγάλη διαφορά το καλύτερο Precision και το καλύτερο F-score, ενώ στο Recall βρίσκεται στις πρώτες θέσεις, με μικρή σχετικά απόκλιση από το πρώτο μοντέλο. Συνεπώς, θεωρούμε την επιλογή του προτεινόμενου μοντέλου δικαιολογημένη.


Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Πίνακας 4-2: Σύγκριση Precision, Recall και F-score του μοντέλου που χρησιμοποιείται στο MediaPot με άλλα ανταγωνιστικά μοντέλα**

| Μέθοδος | Precision | Recall | F-score |
|---|---|---|---|
| S_HSV | 0.28 | 0.36 | 0.32 |
| S_DCT | 0.22 | **0.84** | 0.36 |
| B_HSV | 0.44 | 0.11 | 0.18 |
| B_DCT | 0.41 | 0.27 | 0.32 |
| A_OF | 0.27 | 0.78 | 0.40 |
| A_SIFT | 0.33 | 0.17 | 0.23 |
| A_SURF | 0.36 | 0.29 | 0.33 |
| A_ORB | 0.38 | 0.05 | 0.08 |
| H_OF | 0.37 | 0.60 | 0.45 |
| H_SIFT | 0.34 | 0.66 | 0.45 |
| H_SURF | 0.36 | 0.66 | 0.46 |
| H_ORB | 0.28 | 0.72 | 0.40 |
| **Επιλεγμένη μέθοδος MediaPot** | **0.52** | 0.70 | **0.59** |



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# 5 Περίληψη Βίντεο

## 5.1 Περιγραφή Προβλήματος

Ο όγκος της πληροφορίας που ανεβαίνει σε μορφή βίντεο στα Μέσα Κοινωνικής Δικτύωσης (ΜΚΔ) είναι ιδιαίτερα μεγάλος. Αυτό σημαίνει πως οι υπεύθυνοι οργανισμοί ΜΜΕ έχουν πολλές, νέες ευκαιρίες για να εμπλουτίσουν τις πηγές τους, αλλά και πως παράλληλα οποιαδήποτε προσπάθεια εύρεσης και επεξεργασίας της διαθέσιμης πληροφορίας απαιτεί τη διάθεση εξίσου μεγαλύτερου χρόνου. Μια συνοπτική περίληψη ενός βίντεο βοηθάει το θεατή να καταλάβει τα το κεντρικό μήνυμα του βίντεο και να αποφασίσει εάν είναι σχετικό με αυτό που ψάχνει.

Η περίληψη βίντεο περιεχομένου στοχεύει στο να εξάγει αυτόματα τα σημαντικά σημεία του αρχικού βίντεο, ώστε να κρατάει τα κεντρικά σημεία της ιστορίας του αρχικού βίντεο. Περιλήψεις βίντεο μπορούν να χρησιμοποιηθούν για πιο γρήγορη μελέτη και κατανόηση του διαθέσιμου περιεχομένου, αλλά και για την πιο αποτελεσματική αποθήκευση, αρχειοθέτηση και αναζήτηση από οργανισμούς ΜΜΕ.

Παράλληλα, νέες τεχνικές τροποποίησης και παραμετροποίησης των προτεινόμενων αλγορίθμων περίληψης επιτρέπουν στους χρήστες να επιλέξουν ακριβώς τα στοιχεία στα οποία επιθυμούν να επικεντρωθούν και να μελετήσουν, διευκολύνοντας τους σε μεγάλο βαθμό.

## 5.2 Υπόβαθρο

Η αυτόματη περίληψη βίντεο βασίζεται στην Τεχνητή Νοημοσύνη (ΤΝ) και άρχισε να γίνεται δημοφιλής τα τελευταία δέκα χρόνια, παράλληλα με τη ραγδαία εξέλιξη των εργαλείων βαθιάς μάθησης (deep learning).

Οι σχετικοί αλγόριθμοι χωρίζονται σε δύο μεγάλες κατηγορίες (Truong et al., 2007): στατική περίληψη και δυναμική περίληψη. Η στατική περίληψη ξεχωρίζει και επιστρέψει τα πιο σημαντικά καρέ ενός βίντεο, ενώ η δυναμική περίληψη διαλέγει σκηνές του αρχικού βίντεο, τις ενώνει, δημιουργώντας ένα νέο βίντεο περίληψης. Η δυναμική περίληψη δίνει τη δυνατότητα προσθήκης κίνησης και ήχου, κάνοντας την περίληψη πιο διαδραστική, σημαντικό στοιχείο στη χρήση βίντεο για δημοσιογραφικούς σκοπούς. Για αυτό, η παρούσα υλοποίηση ακολουθεί τη λογική δυναμικής περίληψης.


Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Οι πιο παλιές προσεγγίσεις περίληψης χρησιμοποιούν επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks - RNN) και πιο συγκεκριμένα τις παραλλαγές των RNN όπως τα LSTMs και τα GRUs δίκτυα (Cho et al., 2014; Zhang et al., 2016). Οι αρχιτεκτονικές αυτά καταφέρνουν να μοντελοποιήσουν τις χρονικές εξαρτήσεις που παρουσιάζουν διαφορετικά καρέ, μαθαίνοντας έτσι να ξεχωρίσουν τα σημαντικότερα καρέ. Παρουσιάζουν ωστόσο αδυναμίες σύνδεσης καρέ που απέχουν αρκετά μεταξύ τους, καθιστώντας τες ακατάλληλες για περίληψη βίντεο μεγάλης διάρκειας.

Πολλές εργασίες βασίζονται στη χρήση μηχανισμών προσοχής (attention mechanism), οι οποίοι εκπαιδεύονται ώστε να εστιάζουν στα σημαντικά σημεία από κάθε καρέ. Πιο παλιές προσεγγίσεις (Feng et al., 2018; Fu et al. 2019) συνδυάζουν τα RNN με τους μηχανισμούς προσοχής, κι αν και δείχνουν να έχουν θετικά αποτελέσματα, οι μηχανισμοί προσοχής εστιάζουν σε ολόκληρο το βίντεο, δημιουργώντας προβλήματα σε βίντεο μεγάλης διάρκειας όπου μακρινά καρέ διαφέρουν σημαντικά. Έτσι, πιο πρόσφατες υλοποιήσεις που βασίζονται σε μηχανισμούς προσοχής (Apostolidis et al., 2022; Apostolidis et al., 2021) χρησιμοποιούν κάποιο ενδιάμεσο μηχανισμό που επιτρέπει στους μηχανισμούς προσοχής να εστιάζουν σε μικρότερα κομμάτια του βίντεο.

Αξιόλογα αποτελέσματα παρουσιάζουν και προσεγγίσεις που χρησιμοποιούν γράφους, είτε με νευρωνικά δίκτυα γράφων (Chaves et al., 2024) και συνελικτικά νευρωνικά δίκτυα γράφων (Wu et al., 2021), είτε με κάποιον γράφο μάθησης (knowledge graph), υπεύθυνο για να καταγράψει τις σχέσεις και τα χαρακτηριστικά των καρέ (Zhu et al., 2022).

Τελευταίες δουλείες προσπαθούν να εισάγουν περισσότερη πληροφορία πέρα από αυτήν που βρίσκεται στα καρέ για να οδηγηθούν σε μια καλύτερη περίληψη. Στην εργασία (Ghauri et al., 2021) συνδυάζουν τα χαρακτηριστικά των καρέ, με χαρακτηριστικά που αποτυπώνουν την κίνηση ανάμεσα σε γειτονικά καρέ. Στις εργασίες (Wang et al., 2023), (Narasimhan et al., 2021), (Zhong et al., 2022) προσθέτουν κείμενο στην είσοδο των μοντέλων. Το κείμενο αυτό μπορεί να είναι κάποια σύντομη περιγραφή του βίντεο, αναλυτική περιγραφή του βίντεο, δημιουργημένη από κάποιο Large Language Model (LLM), ή κάποιο άλλο στοχευμένο σχόλιο στο οποίο θα εστιάσει μετέπειτα η περίληψη.

## 5.3 Επιλογή Μοντέλου

Σε αντίθεση με το διαχωρισμό ενός πλάνου σε υποπλάνα (Ενότητα 4), η περίληψη βίντεο αποτελεί ένα πιο περίπλοκο πρόβλημα.



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Υπάρχουν πολλές εργασίες, που εκτός από διαφορετικές υλοποιήσεις, διαφέρουν και στον τρόπο που ορίζουν τον στόχο. Αυτό οδηγεί σε μεγάλες διαφορές μεταξύ υλοποιήσεων, που επηρεάζουν την πολυπλοκότητα των υλοποιήσεων και την καταλληλόλητα τους για το παρόν έργο.

Η υλοποίηση της αυτόματης περίληψης βίντεο στα πλαίσια του παρόντος έργου έχει ως στόχο να αποτελέσει ένα εύχρηστο και κατανοητό εργαλείο, το οποίο θα συμβάλλει στην πιο στοχευμένη και εύκολη ανάλυση βίντεο από τους συνεργάτες. Με αυτό ως βασικό γνώμονα, κάναμε διεξοδική μελέτη στη βιβλιογραφία και συγκεκριμένα στα μοντέλα που αναφέρθηκαν στην προηγούμενη ενότητα, με σκοπό να επιλέξουμε το καταλληλότερο. Τα κριτήρια αξιολόγησης των μοντέλων είναι:

1. **<u>Στόχος μοντέλου</u>:**
   Ο βασικός στόχος του επιλεγμένου μοντέλου πρέπει να ταιριάζει με τον στόχο του έργου Mediapot. Το μοντέλο πρέπει να είναι σε θέση να δημιουργεί δυναμικές περιλήψεις με τα σημαντικότερα πλάνα του αρχικού βίντεο και όχι απλά να ξεχωρίζει σημαντικά καρέ, ώστε οι συνεργάτες να μπορούν να εκμεταλλευτούν άμεσα το αποτέλεσμα.
2. **<u>Απόδοση μοντέλου</u>:**
   Το επιλεγμένο μοντέλο πρέπει να αποδίδει σε ικανοποιητικό βαθμό στα κλασσικά τεστ αξιολόγησης.
3. **<u>Πολυπλοκότητα μοντέλου</u>:**
   Περίπλοκες αρχιτεκτονικές οδηγούν σε ανάγκη για μεγαλύτερους υπολογιστικούς πόρους, περισσότερο χρόνο για την λήψη αποτελεσμάτων, ενώ ανεβάζουν τη δυσκολία μιας ακριβής υλοποίησης. Για να είναι το παραδιδόμενο εργαλείο εύχρηστο, πρέπει να μπορεί να δίνει αποτελέσματα σχετικά άμεσα και γρήγορα.
4. **<u>Εξηγισιμότητα (Explainability) μοντέλου</u>:**
   Ένα πρόβλημα με τα μοντέλα ΤΝ είναι πως συχνά είναι δύσκολο να ερμηνεύσουμε τις αποφάσεις τους, και τα αντιμετωπίζουμε ως «μαύρα κουτιά». Στο παρόν έργο, είναι χρήσιμο να επιλέξουμε ένα μοντέλο, το οποίο θα μπορούμε να ακολουθήσουμε λίγο πιο εύκολα την πορεία επιλογής του και να δώσουμε καλύτερες και πιο ακριβείς απαντήσεις στους συνεργάτες.
5. **<u>Δυνατότητα επέκτασης</u>:**
   Καθώς η ΤΝ τρέχει με γρήγορους ρυθμούς, νέες ιδέες δημοσιεύονται συνεχώς. Η επιλογή ενός μοντέλου που μπορεί εύκολα να επεκταθεί ή αναβαθμιστεί, ώστε να ενσωματώσει κάποια καινοτομία που αυξάνει την απόδοση, είναι καταλληλότερη, σε σύγκριση με κάποιο άλλο ειδικό και πιο στατικό μοντέλο.

Με βάση τα παραπάνω κριτήρια, καταλήξαμε στην επιλογή του μοντέλου **PGL-SUM** (Apostolidis et al., 2021). Το PGL-SUM παίρνει ως είσοδο ένα βίντεο και δημιουργεί μια δυναμική περίληψη, ικανοποιώντας το πρώτο κριτήριο. Η απόδοση του θεωρείται κορυφαία σε σύγκριση με άλλα καινοτόμα μοντέλα, όπως θα αναλύσουμε στην υποενότητα 5.5.

Η υλοποίηση βασίζεται σε έναν μηχανισμό, ο οποίος συνδυάζει έναν γενικό μηχανισμό προσοχής (global attention mechanism) που κοιτάει σε όλο το βίντεο, με τοπικούς μηχανισμούς προσοχής (local



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

attention mechanisms) οι οποίοι εστιάζουν σε διαφορετικά κομμάτια του βίντεο. Παράλληλα, εισάγεται και η κωδικοποίηση χρονικής θέσης των καρέ ως επιπλέον πληροφορία που τροφοδοτείται στο μοντέλο. Αποφεύγει να χρησιμοποιήσει πιο πολύπλοκες αρχιτεκτονικές, όπως Transformers και νευρωνικά δίκτυα με γράφους, και χρησιμοποιεί ένα απλό δίκτυο παλινδρόμησης για τον υπολογισμό των σκορ των καρέ. Αυτό σημαίνει πως η υλοποίηση του παραμένει απλή, ενώ μελετώντας τον γενικό και τους τοπικούς μηχανισμούς προσοχής, μπορούμε εύκολα και ξεκάθαρα να εξάγουμε συμπεράσματα για την απόφαση του μοντέλου.

Η σχετικά απλή δομή του PGL-SUM, μας επιτρέπει να το τροποποιήσουμε χωρίς ιδιαίτερο κόπο στην προσπάθεια να βελτιώσουμε την απόδοσή του. Στην παρούσα φάση του έργου, εφαρμόζουμε την αρχική του υλοποίηση, αλλά παράλληλα εξερευνούμε τροποποιήσεις του μοντέλου που μπορεί να οδηγήσουν σε καλύτερα αποτελέσματα.

## 5.4 Μέθοδος

Στην παρούσα υποενότητα θα αναλύσουμε λεπτομερώς το μοντέλο που υιοθετήθηκε. Το **PGL-SUM** δέχεται ένα βίντεο ως αλληλουχία καρέ στην είσοδο και επιστρέφει το σκορ σημαντικότητας του κάθε καρέ. Με βάση τα σημαντικότερα καρέ, επιλέγονται τα κομμάτια του βίντεο τα οποία περιέχουν κυρίως σημαντικά καρέ, και δημιουργούν τη δυναμική περίληψη του βίντεο. Το μοντέλο στηρίζεται σε μηχανισμούς προσοχής πολλών κεφαλών (multi-head attention mechanisms).

Όπως αναφέρθηκε και στην υποενότητα 5.3, το μοντέλο συνδυάζει γενικούς μηχανισμούς προσοχής, οι οποίοι μελετάνε το σύνολο των καρέ για γενικές διαφορές, με τοπικούς μηχανισμούς προσοχής που εστιάζουν στις διαφορές μεταξύ διαδοχικών καρέ. Παράλληλα, οι μηχανισμοί προσοχής ενισχύονται με την πληροφορία της κωδικοποιημένης χρονικής θέσης που τροφοδοτείται.

Η λειτουργία του μοντέλου περιγράφεται στη συνέχεια της υποενότητας.

Έστω ένα βίντεο με **T** καρέ. Το κάθε καρέ περνάει από ένα προ-εκπαιδευμένο συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network - CNN) ώστε να επιλεγούν τα σημαντικά χαρακτηριστικά του καρέ, τα οποία συμβολίζονται με το διάνυσμα xt, διάστασης **D** και τα συνολικά χαρακτηριστικά για όλα τα καρέ στο βίντεο συμβολίζονται με **X** ={xt}t=1T..

Τα συνολικά χαρακτηριστικά ακολουθούν δύο διαδρομές. Η μία σχετίζεται με τον γενικό μηχανισμό προσοχής πολλών κεφαλών, ο οποίος στοχεύει να ανακαλύψει τις συσχετίσεις των καρέ από όλο το βίντεο. Ο γενικός μηχανισμός προσοχής βασίζεται στο Δίκτυο των Transformers (Vaswani et al., 2017). Στην κάθε κεφαλή δημιουργούνται τρεις πίνακες κατά την εκπαίδευση του μοντέλου, ο Query (**Q**



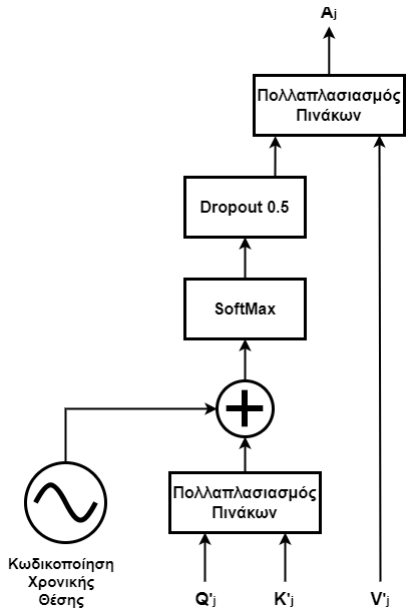
Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

={qt}t=1T), ο Key (**K** ={kt}t=1T) και ο Value (**V** ={vt}t=1T). Οι τρεις αυτοί πίνακες τροφοδοτούνται ο καθένας σε ένα γραμμικό στρώμα και μετά συνδυάζονται με dot-product, δημιουργώντας τις τιμές προσοχής για κάθε καρέ
(**A** ={at}t=1T). Στη dot-product διαδικασία, προστίθεται και πληροφορία σχετικά με την χρονική θέση των καρέ, αυτό που ονομάζουμε κωδικοποίηση χρονικής θέσης. Η κωδικοποίηση δημιουργείται χρησιμοποιώντας την απόλυτη θέση του καρέ, και ημιτονοειδής συναρτήσεις σε διαφορετικές συχνότητες:

$$PE(pos,2i) = \sin(pos/100002i/D)$$
$$PE(pos,2i+1) = \cos(pos/100002i/D)$$

όπου **pos** είναι η απόλυτη θέση και **i** ο δείκτης του κάθε καρέ. Η κωδικοποίηση εφαρμόζεται σε κάθε καρέ του βίντεο, δημιουργώντας έναν **TxT** πίνακα. Τελικά η dot-product διαδικασία εφαρμόζεται σύμφωνα με το παρακάτω διάγραμμα:

**Εικόνα 5-1: Dot-product διαδικασία**

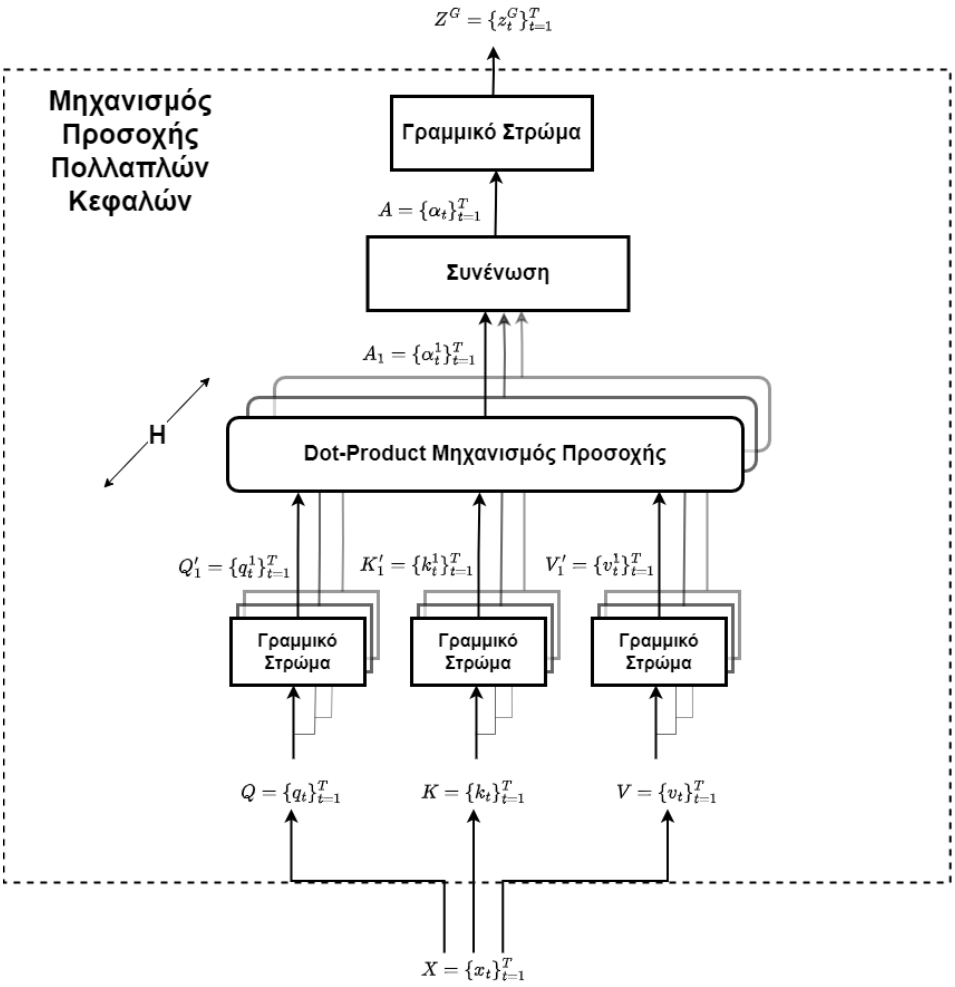



Η διαδικασία του μηχανισμού προσοχής επαναλαμβάνεται σε κάθε κεφαλή, και τελικά οι τιμές προσοχής από κάθε κεφαλή συνδυάζονται μεταξύ τους για να παράγουν την τελική, συνολική, τιμή προσοχής. Η τιμή αυτή περνάει από ένα γραμμικό στρώμα, παράγοντας την έξοδο του μηχανισμού ZG. Ο αριθμός των κεφαλών **H** σε κάθε μηχανισμό αποτελεί υπερπαράμετρο, ρυθμίζεται δηλαδή έτσι ώστε να μεγιστοποιεί την ακρίβεια του μοντέλου. Στην παρούσα υλοποίηση είναι ίσος με 8. Η παραπάνω διαδικασία φαίνεται στο διάγραμμα που ακολουθεί:



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

**Εικόνα 5-2: Γενικός μηχανισμός προσοχής**




Η δεύτερη διαδρομή των συνολικών χαρακτηριστικών σχετίζεται με τους τοπικούς μηχανισμούς προσοχής πολλαπλών κεφαλών. Τα χαρακτηριστικά χωρίζονται σε M συνεχόμενα και μη επικαλυπτόμενα κομμάτια, και το καθένα τροφοδοτείται σε έναν τοπικό μηχανισμό προσοχής. Η πορεία των χαρακτηριστικών είναι αντίστοιχη με αυτή που περιγράφηκε στο γενικό μηχανισμό προσοχής. Οι τοπικοί μηχανισμοί προσοχής εστιάζουν αυτήν τη φορά στο να βρουν συσχετίσεις μεταξύ των καρέ όχι όλου του βίντεο όπως στο γενικό μηχανισμό, αλλά μόνο των καρέ που ανήκουν στο ίδιο κομμάτι. Στην έξοδο του κάθε τοπικού μηχανισμού παράγεται το ZL={zLi}i=1M. Το **M** αποτελεί υπερπαράμετρο που ρυθμίζεται ώστε να βελτιστοποιεί τα αποτελέσματα. Στην παρούσα υλοποίηση είναι ίσο με xxx.

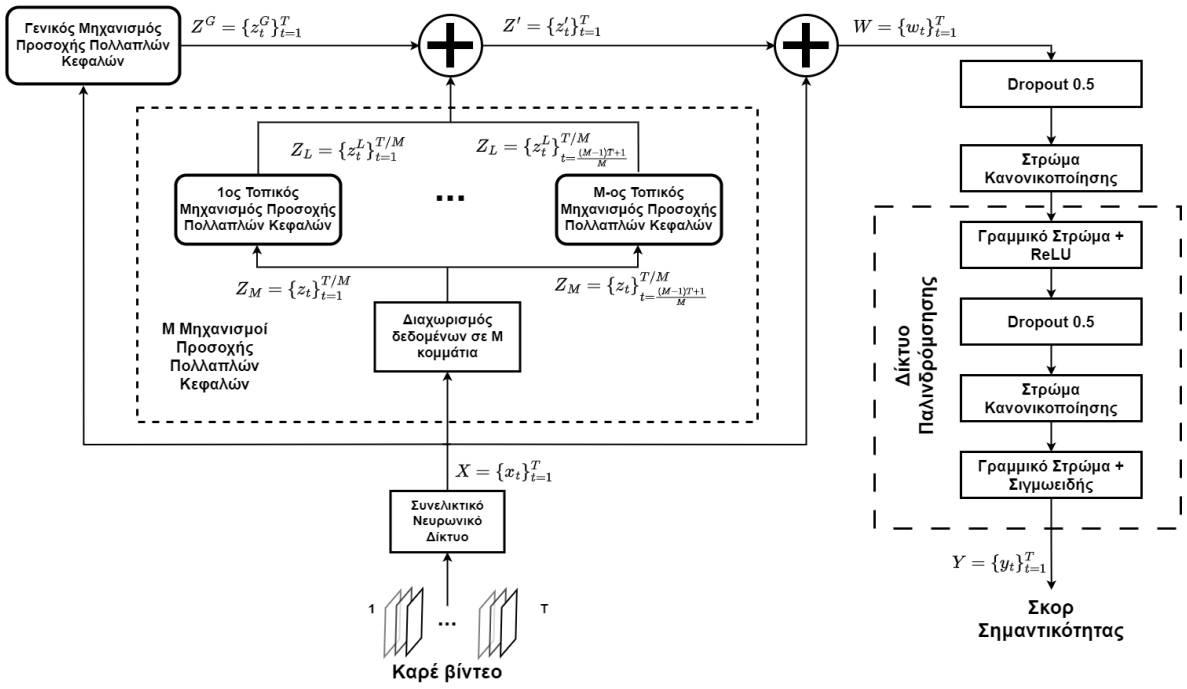
Τα τελικά χαρακτηριστικά προσοχής (**Z'** ={z't}t=1T) προκύπτουν προσθέτοντας τα ZG και ZL. Στη συνέχεια, τα χαρακτηριστικά προσοχής **Z'** προστίθενται στα αρχικά χαρακτηριστικά **X**. Το



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

αποτέλεσμα της πράξης, **W**, περνάει από ένα Dropout στρώμα και μετά κανονικοποιείται, για να τροφοδοτήσει τέλος ένα δίκτυο Παλινδρόμησης, που υπολογίζει το σκορ σημαντικότητας του κάθε καρέ, επιστρέφοντας μια τιμή στο διάστημα [0, 1], για κάθε καρέ. Ο συνολικός μηχανισμός παρουσιάζεται στο παρακάτω διάγραμμα.

**Εικόνα 5-3: Το μοντέλο PGL-SUM που χρησιμοποιείται στο MediaPot**




Το μοντέλο εκπαιδεύεται ελαχιστοποιώντας το σφάλμα μέσων ελαχίστων τετραγώνων ανάμεσα στη πρόβλεψη των σημαντικών καρέ και των πραγματικών σκορ του συνόλου δεδομένων.

## 5.5 Πειραματική Αξιολόγηση

Η αξιολόγηση του **PGL-SUM** μοντέλου γίνεται στα δύο βασικά σύνολα δεδομένων που χρησιμοποιούνται στα πλαίσια της περίληψης βίντεο. Το πρώτο είναι το **TVSum**, το οποίο αποτελείται από 50 βίντεο, διάρκειας 1-11 λεπτών. 20 χρήστες έχουν προσθέσει σκορ σημαντικότητας για κάθε καρέ σε κάθε βίντεο. Συνδυάζοντας τα σκορ κάθε χρήστη, βγαίνει μία περίληψη για κάθε βίντεο. Το δεύτερο σύνολο είναι το **SumMe**, το οποίο αποτελείται από 25 βίντεο, διάρκειας 1-6 λεπτών. Για κάθε βίντεο, αρκετοί χρήστες (15 με 18) έχουν διαλέξει τα σημαντικότερα κομμάτια τα οποία δημιουργούν την περίληψη. Όπως και οι περισσότερες καινοτόμες δουλειές, το μοντέλο αξιολογείται με το πρωτόκολλο σημαντικών κομματιών του βίντεο. Αυτό σημαίνει πως με



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

βάση τα σκορ σημαντικότητας καρέ, επιλέγονται τα σημαντικά κομμάτια του βίντεο που δημιουργούν την περίληψη. Το SumMe έχει έτοιμα τα σημαντικά κομμάτια, ενώ το TVSum απαιτεί την επιλογή των σημαντικών κομματιών με βάση τα σκορ των καρέ. Στο τέλος, συγκρίνονται τα σημαντικά κομμάτια του βίντεο που επιλέγονται από το προτεινόμενο μοντέλο, με τα κομμάτια που έχει επιλέξει ο κάθε χρήστης στο εκάστοτε σύνολο δεδομένων και υπολογίζεται το **F-Score**. Στο τέλος, υπολογίζεται ο μέσος όρος από όλα τα **F-Score** που έχουν υπολογιστεί με βάση τη σύγκριση πρόβλεψης και κάθε χρήστη. Το 80% των δεδομένων σε κάθε σύνολο χρησιμοποιείται για την διαδικασία της εκπαίδευσης, ενώ το άλλο 20% για τη διαδικασία της αξιολόγησης.

Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα, όπου το επιλεγμένο μοντέλο συγκρίνεται με άλλα καινοτόμα μοντέλα. Παρατηρούμε ότι το PGL-SUM παρουσιάζει με διαφορά το υψηλότερο score στο σύνολο δεδομένων SumMe σε σύγκριση με τα υπόλοιπα καινοτόμα μοντέλα, ενώ στο σύνολο δεδομένων TVSum, έρχεται δεύτερο, με οριακή διαφορά από το πρώτο μοντέλο, διατηρώντας ωστόσο σημαντική διαφορά από τα υπόλοιπα καινοτόμα μοντέλα. Συνεπώς, η επιλογή του PGL-SUM μοντέλου για το κομμάτι της περίληψης βίντεο, με βάση τη μελέτη της βιβλιογραφίας και τη σύγκριση καινοτόμων μοντέλων, είναι κατάλληλη.

**Πίνακας 5-1: Σύγκριση F-score του μοντέλου PGL-SUM που χρησιμοποιείται στο MediaPot**

| Μοντέλο | SumMe | TVSum |
|---|---|---|
| VASNet | 49.7 | 61.42 |
| M-AVS | 44.4 | 61.0 |
| DSNet | 53.0 | 62.1 |
| RR-STG | 54.5 | **63.0** |
| **PGL-SUM** | **57.1** | 62.7 |

## 5.6 Υλοποίηση και Ενσωμάτωση

Η υλοποίηση, τόσο της περίληψης βίντεο, όσο και της κατάτμησης βίντεο σε πλάνα και υποπλάνα που περιγράφηκε στην ενότητα 4, έχει σχεδιαστεί ως υπηρεσία αρχιτεκτονικής REST, χρησιμοποιώντας microservises, αντίστοιχα με αυτά που περιγράφθηκαν στην υποενότητα 3.5. Ο χρήστης συνδέεται με την εφαρμογή μέσω API.



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Η εφαρμογή θα ανακτήσει το βίντεο και θα το αποθηκεύσει τοπικά. Στη συνέχεια κατατμεί το βίντεο σε πλάνα και υποπλάνα, σύμφωνα με τη μέθοδο της υποενότητας 4.3, ενώ επιστρέφει και τα τρία σημαντικότερα καρέ για κάθε υποπλάνο.

Μόλις ολοκληρωθεί η διαδικασία της κατάτμησης, η εφαρμογή συνεχίζει με την περίληψη. Για να εξάγει τα χαρακτηριστικά των καρέ, περνάει κάθε δεύτερο καρέ από ένα προεκπαιδευμένο δίκτυο GoogleNet, και τα τροφοδοτεί στο μοντέλο της υποενότητας 5.3, για να εξάγει τα σκορ σημαντικότητας κάθε καρέ. Με βάση αυτά τα σκορ, δημιουργεί την περίληψη του βίντεο.

Στο τέλος, με κατάλληλη εντολή από τον χρήστη, η εφαρμογή επιστρέφει σε αρχεία json τις χρονικές στιγμές και τα keyframes της αρχής κάθε πλάνου και υποπλάνου, καθώς και τα frames που ανήκουν στην τελική περίληψη. Επιπλέον, υπάρχει δυνατότητα δημιουργίας αρχείου mp4 με την περίληψη του βίντεο.



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# 6 Συμπεράσματα και Επόμενα Βήματα

Η συλλογή και η διαχείριση πολυμεσικού περιεχομένου απαιτεί πολυεπίπεδη προσέγγιση, η οποία περιλαμβάνει τη χρήση τεχνητής νοημοσύνης και προηγμένων τεχνολογιών. Με βάση εργαλεία που αναπτύχθηκαν, δοκιμάστηκαν και αξιολογήθηκαν προκύπτουν συγκεκριμένα συμπεράσματα.

Αρχικά, η χρήση μοντέλων μηχανικής μάθησης, όπως τα CNNs και οι ViTs, είναι καίρια για την ανίχνευση και την ταξινόμηση του πολυμεσικού περιεχομένου. Η εκπαίδευση αυτών των μοντέλων μεγάλων διαστάσεων με χρήση επαναληπτικών προσεγγίσεων και η ανάπτυξη ημι-αυτόματων συστημάτων επισήμανσης είναι ουσιώδεις για τη δημιουργία αποτελεσματικών λύσεων. Ακόμη, η επέκταση της λειτουργικότητας για την αναγνώριση διαφορετικών μορφών πολυμεσικού περιεχομένου, όπως εικόνες και βίντεο, απαιτεί προηγμένες τεχνικές επεξεργασίας και ανάλυσης.

Η χρήση της βιβλιοθήκης Elasticsearch παρέχει έναν αποδοτικό μηχανισμό δεικτοδότησης και αναζήτησης πολυμεσικού περιεχομένου, βελτιώνοντας σημαντικά τη διαχείριση των δεδομένων και την ταχύτητα απόκρισης του συστήματος.

Επιπλέον, η τεχνική υλοποίηση μέσω του Nvidia Triton Inference Server παρέχει ένα ευέλικτο και κλιμακούμενο περιβάλλον για την εκτέλεση μοντέλων τεχνητής νοημοσύνης, επιτρέποντας την παράλληλη επεξεργασία και τη σταθερή απόδοση του συστήματος, ενώ η υιοθέτηση του πρωτοκόλλου gRPC για την επικοινωνία με τον Triton Inference Server εξασφαλίζει υψηλή απόδοση και ευελιξία στη διασύνδεση των συστημάτων.

Συνολικά, ο συνδυασμός αυτών των τεχνολογιών και τεχνικών παρέχει μια ισχυρή ευέλικτη μεθοδολογία διαχείρισης πολυμεσικού περιεχομένου. Επόμενα βήματα δύναται να περιλαμβάνουν τη συνεχή βελτίωση των αλγορίθμων μάθησης, την επέκταση της λειτουργικότητας και την προσαρμογή σε περιεχόμενο, έννοιες και κατηγορίες που είναι χρήσιμα για τα Ελληνικά ΜΜΕ με σκοπό να ανταποκρίνονται στις σημερινές ανάγκες και απαιτήσεις που προκύπτουν από το συνεχές μεταβαλλόμενο πεδίο των κοινωνικών δικτύων και της ενημέρωσης.

Τα εργαλεία που παρουσιάζονται στις παραπάνω ενότητες μπορούν να ενσωματωθούν εύκολα με τη χρήση API στη συνολική πλατφόρμα MediaPot, ώστε να προσφερθούν στους τελικούς χρήστες. Στη σημερινή εποχή, όπου μεγάλο κομμάτι της διαθέσιμης πληροφορίας στο διαδίκτυο βρίσκεται σε μορφή βίντεο, τα εργαλεία αυτά μπορούν να βοηθήσουν σε μεγάλο βαθμό στην εξαγωγή χρήσιμης πληροφορίας. Με τη χρήση αυτών των συστημάτων, οι δημοσιογράφοι μπορούν να αναλύουν και να επεξεργάζονται μεγάλες συλλογές εικόνων και βίντεο πιο αποδοτικά, εξοικονομώντας χρόνο και


Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

βελτιώνοντας την ακρίβεια των ρεπορτάζ τους, με την καλύτερη και ταχύτερη ενημέρωση του κοινού.



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# 7 Βιβλιογραφικές Αναφορές

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. (2016). WaveNet: A Generative Model for Raw Audio. CoRR abs/1609.03499 https://doi.org/10.48550/arXiv.1609.03499

Apostolidis E., G. Balaouras, V. Mezaris and I. Patras, (2021). "Combining Global and Local Attention with Positional Encoding for Video Summarization," 2021 IEEE International Symposium on Multimedia (ISM), Naple, Italy, 2021, pp. 226-234, https://doi.org/10.1109/ISM52913.2021.00045.

Apostolidis E. and Mezaris V. (2014). Fast shot segmentation combining global and local visual descriptors. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2014), 6583–6587. DOI:10.1109/ICASSP.2014.6854873

Apostolidis, K., Apostolidis, E., Mezaris, V. (2018). A Motion-Driven Approach for Fine-Grained Temporal Segmentation of User-Generated Videos. In: Schoeffmann, K., *et al.* MultiMedia Modeling. MMM 2018. Lecture Notes in Computer Science(), vol 10704. Springer, Cham. https://doi.org/10.1007/978-3-319-73603-7_3

Apostolidis E., Balaouras G., Mezaris V., and Patras I. 2022. Summarizing Videos using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames. In Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22). Association for Computing Machinery, New York, NY, USA, 407–415. https://doi.org/10.1145/3512527.3531404

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349

Bay, H., et al. (2006). Surf: Speeded up robust features. Int. Jour. of Comp. Vis. pp. 404–417 https://doi.org/10.1007/11744023_32

Baraldi L.,Grana C., and Cucchiara R. (2015). A Deep Siamese. Network for Scene Detection in Broadcast Videos. In Proceedings of the 23rd ACM International Conference on Multimedia (Brisbane, Australia) (MM âĂŹ15). Association for Computing Machinery, New York, NY, USA, 1199âĂŞ1202., https://doi.org/10.1145/2733373.2806316

Baraldi L., Grana C., and Cucchiara R. (2015). Shot and Scene. Detection via Hierarchical Clustering for Re-using Broadcast Video. In Computer Analysis of Images and Patterns, George Azzopardi and Nicolai Petkov (Eds.). Springer International Publishing, Cham, 801–811, https://doi.org/10.1007/978-3-319-23192-1_67



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A. & Plank, B. (2017). Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures, IJCAI. https://doi.org/10.48550/arXiv.1601.03896

Bertasius, G., Torresani, L., Shi, J. (2018). Object Detection in Video with Spatiotemporal Sampling Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany. https://doi.org/10.48550/arXiv.1803.05549

Bouguet, J.Y., (2001). Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. Intel Corporation 5(1-10), 4

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A Dataset for Recognising Faces across Pose and Age. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Published. https://doi.org/10.1109/fg.2018.00020

Chaves, J. M., & Tripathi, S. (2024). VideoSAGE: Video Summarization with Graph Representation Learning. ArXiv. /abs/2404.10539. https://doi.org/10.48550/arXiv.2404.10539

Chen, H., Li, J., Hu, X. (2020). Delving deeper into the decoder for video captioning. arXiv preprint arXiv:200105614. https://doi.org/10.3233/FAIA200204

Chen, Y., Cao, Y., Hu, H., Wang, L. (2020). Memory Enhanced Global-Local Aggregation for Video Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition https://doi.org/10.48550/arXiv.2003.12063

Chen, J., Zhang, L., Bai, C. & Kpalma, K. (2020a). Review of Recent Deep Learning Based Methods for Image-Text Retrieval. IEEE 3rd International Conference on Multimedia Information Processing and Retrieval. DOI:10.1109/MIPR49039.2020.00042

Cho K., B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

Cooray, S.H., et al., (2010). Identifying an efficient and robust sub-shot segmentation method for home movie summarisation. In: 10th Int. Conf. on Intell. Syst. Design and Appl. pp. 1287–1292 https://doi.org/10.1109/ISDA.2010.5687086

Deep Feature Flow for Video Recognition. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; (pp. 4141–4150). https://doi.org/10.48550/arXiv.1611.07715



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. IEEE conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Dumont, E., et al., (2008). Rushes video summarization using a collaborative approach. In: Proc. of the 2nd ACM TRECVID Vid. Summar. Workshop. pp. 90–94 https://doi.org/10.1145/1463563.1463579

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6202-6211). doi: 10.1109/ICCV.2019.00630.

Feichtenhofer, C. (2020). X3D: Expanding Architectures for Efficient Video Recognition. CVPR. https://doi.org/10.48550/arXiv.2004.04730

Feichtenhofer, C., Pinz, A., Zisserman, A. (2017). Detect to Track and Track to Detect. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, (pp. 3057–3065). https://doi.org/10.48550/arXiv.1710.03958

Feng L., Z. Li, Z. Kuang, and W. Zhang. (2018). Extractive video summarizer with memory augmented neural networks. In Proceedings of the 26th ACM International Conference on Multimedia, MM '18, pages 976–983, New York, NY, USA, 2018. ACM. https://doi.org/10.1145/3240508.3240651

Fu T., S. Tai, and H. Chen. (2019). Attentive and adversarial learning for video summarization. In IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, January 7-11, 2019, pages 1579–1587 https://doi.org/10.1109/WACV.2019.00173

Ghauri J. A., Hakimov S. and Ewerth R., "Supervised Video Summarization Via Multiple Feature Sets with Parallel Attention," 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 2021, pp. 1-6s, https://doi.org/10.1109/ICME51207.2021.9428318

Greg Pass, Ramin Zabih, and Justin Miller. (1996). Comparing Images Using Color Coherence Vectors. In Proceedings of the Fourth ACM International Conference on Multimedia (MULTIMEDIA '96). ACM, New York, NY, USA, 65–73. https://doi.org/10.1145/244130.244148

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

He, B., Wang, J., Qiu, J., Bui, T., Shrivastava, A., & Wang, Z. (2023). Align and Attend: Multimodal Summarization with Dual Contrastive Losses. *ArXiv*. /abs/2303.07284. https://doi.org/10.48550/arXiv.2303.07284

Hong Jiang Zhang, Atreyi Kankanhalli, and Stephen W. Smoliar. (1993). Automatic partitioning of full-motion video. Multimedia Systems 1, 1 (01 Jan 1993), 10–28. https://doi.org/10.1007/BF01210504

Hou, J., Wu, X., Zhao, W., Luo, J., Jia, Y. (2019). Joint syntax representation learning and visual cue translation for video captioning. In: Proceedings of the IEEE International Conference on Computer Vision (pp. 8918–8927). DOI:10.1109/ICCV.2019.00901

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). pmlr. https://doi.org/10.48550/arXiv.1502.03167

Kim, J.G., et al., (2000). Efficient camera motion characterization for mpeg video indexing. In: Proc. of the IEEE Int. Conf. on Mult. and Expo. vol. 2, pp. 1171–1174 https://doi.org/10.1109/ICME.2000.871569

Kordopatis-Zilos, G., Tzelepis, C., Papadopoulos, S., Kompatsiaris, I., & Patras, I. (2022). DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval. International Journal of Computer Vision, 130(10), 2385- 2407. https://doi.org/10.48550/arXiv.2106.13266

Koutlis, C., Schinas, M., & Papadopoulos, S. (2022). MemeTector: Enforcing deep focus for meme detection. https://doi.org/10.48550/arXiv.2205.13268

Lin T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham https://doi.org/10.48550/arXiv.1405.0312

Liu, M., Zhu, M., White, M., Li, Y., Kalenichenko, D. (2019). Looking Fast and Slow: Memory-Guided Mobile Video Object Detection. arXiv, arXiv:1903.10172. https://doi.org/10.48550/arXiv.1903.10172

Liu, M., Zhu, M. (2018). Mobile Video Object Detection with Temporally-Aware Feature Maps. In Proceedings of the 2018 IEEE/Cvf Conference on Computer Vision and Pattern Recognition (pp. 5686–5695) https://doi.org/10.48550/arXiv.1711.06368

Lowe, D.G. (1999). Object recognition from local scale-invariant features. In: Proc. of the 7th IEEE Int. Conf. on Comp. Vis. vol. 2, pp. 1150–1157 https://doi.org/10.1109/ICCV.1999.790410

Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, Andrew Zisserman. (2020). A Short Note on the Kinetics-700-2020 Human Action Dataset. arXiv:2010.10864. https://doi.org/10.48550/arXiv.2010.10864



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Michael Gygli. (2018). Ridiculously Fast Shot Boundary Detection with Fully Convolutional Neural Networks. In 2018 International Conference on ContentBased Multimedia Indexing, CBMI 2018, La Rochelle, France, September 4-6, 2018. 1–4. https://doi.org/10.1109/CBMI.2018.8516556

Narasimhan, Medhini, Anna Rohrbach and Trevor Darrell. "CLIP-It! Language-Guided Video Summarization." Neural Information Processing Systems (2021). https://doi.org/10.48550/arXiv.2107.00650

Ojutkangas, O., et al., (2012). Location Based Abstraction of User Generated Mobile Videos, pp. 295–306. Springer Berlin Heidelberg https://doi.org/10.1016/j.image.2012.01.017

Pan, C.M., et al., (2007). NTU TRECVID-2007 fast rushes summarization system. In: Proc. of the 1st ACM TRECVID Vid. Summar. Workshop. pp. 74–78 https://doi.org/10.1145/1290031.1290045

Pasunuru, R., Bansal, M. (2017) Multi-task video captioning with video and entailment generation. arXiv preprint arXiv:170407489. https://doi.org/10.48550/arXiv.1704.07489

Quoc V. Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., and Ng, A. (2012). Building high-level features using large scale unsupervised learning, International Conference in Machine Learning. https://doi.org/10.48550/arXiv.1112.6209

Qiu, S., Anwar, S., & Barnes, N. (2021). Semantic Segmentation for Real Point Cloud Scenes via Bilateral Augmentation and Adaptive Fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1757-1767). https://doi.org/10.48550/arXiv.2103.07074

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020. https://doi.org/10.48550/arXiv.2103.00020

Ren, S., He, K., Girshick, R. & Sun, J. (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497. https://doi.org/10.48550/arXiv.1506.01497

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640. https://doi.org/10.48550/arXiv.1506.02640

Redmon, J. & Farhadi, A. (2016). YOLO9000: Better, Faster, Stronger. arXiv:1612.08242. https://doi.org/10.48550/arXiv.1612.08242

Redondo, R., & Gibert, J. (2020, June 24). Extended Labeled Faces in-the-Wild (ELFW): Augmenting Classes for Face Segmentation. arXiv:2006.13980 https://doi.org/10.48550/arXiv.2006.13980



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Sarridis, I., Koutlis, C., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, I. (2022). InDistill: Information flowpreserving knowledge distillation for model compression. arXiv preprint arXiv:2205.10003v2. https://doi.org/10.48550/arXiv.2205.10003

Schinas M., Galopoulos P., & Symeon Papadopoulos S. (2023). MAAM: Media Asset Annotation and Management. In Proceedings of the 2023 ACM International Conference on Multimedia Retrieval (ICMR '23). Association for Computing Machinery, New York, NY, USA, 659–663. https://doi.org/10.1145/3591106.3592232

Shen, Z., Li, J., Su, Z., Li, M., Chen, Y., Jiang, YG., Xue, X. (2017) Weakly supervised dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1916–1924. https://doi.org/10.48550/arXiv.1704.01502

Sheng-Hua Zhong, Jingxu Lin, Jianglin Lu, Ahmed Fares, and Tongwei Ren. (2022). Deep Semantic and Attentive Network for Unsupervised Video Summarization. ACM Trans. Multimedia Comput. Commun. Appl. 18, 2, Article 55, 21 pages. https://doi.org/10.1145/3477538

Shitao Tang, Litong Feng, Zhanghui Kuang, Yimin Chen, and Wei Zhang. (2018) Fast Video Shot Transition Localization with Deep Structured Models. CoRR abs/1808.04234 (2018). arXiv:1808.04234 http://arxiv.org/abs/1808.04234

Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. NeurIPS Proceedings. https://doi.org/10.48550/arXiv.1406.2199

Souček, T., & Lokoč, J. (2020). TransNet V2: An effective deep network architecture for fast shot transition detection. ArXiv. /abs/2008.04838. https://doi.org/10.48550/arXiv.2008.04838

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification, Conference on Computer Vision and Pattern Recognition. CVPR. doi: 10.1109/CVPR.2014.220.

Tan, M., Pang, R. & Le, Q.V. (2020). EfficientDet: Scalable and Efficient Object Detection. CVPR. https://doi.org/10.1109/CVPR42600.2020.01079.

Truong B. T. and S. Venkatesh. (2007). Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl., 3(1), Feb. 2007. https://doi.org/10.1145/1198302.1198305

Vinyals, O., Toshev, A., Bengio, A. & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. CVPR https://doi.org/10.48550/arXiv.1411.4555



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision (pp. 20-36). Springer, Cham. https://doi.org/10.48550/arXiv.1608.00859

Wang, T., Xiong, J., Xu, X., Shi, Y. (2019a) SCNN: A General Distribution Based Statistical Convolutional Neural Network with Application to Video Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA (pp. 5321–5328).

Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, Hongxia Yang. (2022). OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. arXiv:2202.03052v2. https://doi.org/10.48550/arXiv.2202.03052

Wu, H., Chen, Y., Wang, N.. Zhang, Z. (2019). Sequence Level Semantics Aggregation for Video Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV) https://doi.org/10.48550/arXiv.1907.06390

Wu, G., Lin, J., & Silva, C. T. (2021). IntentVizor: Towards Generic Query Guided Interactive Video Summarization. ArXiv. /abs/2109.14834 https://doi.org/10.48550/arXiv.2109.14834

Xiao, F., Lee, J. (2017). Y. Video Object Detection with an Aligned Spatial-Temporal Memory. arXiv 2017, arXiv:1712.06317. https://doi.org/10.48550/arXiv.1712.06317

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2048-2057. https://doi.org/10.48550/arXiv.1502.03044

Yang, Y., Zhou, J., Ai, J., Bin, Y., Hanjalic, A., Shen, H.T., Ji, Y. (2018) Video captioning by adversarial lstm. IEEE Transactions on Image Processing (pp. 5600–11). doi: 10.1109/TIP.2018.2855422.

Yang, W., Liu, B., Li, W., Yu, N. (2019). Tracking Assisted Faster Video Object Detection. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo, Shanghai, China (pp. 1750–1755). doi: 10.1109/ICME.2019.00301.

Zhang K., W.-L. Chao, F. Sha, and K. Grauman. (2016). Video summarization with long shortterm memory. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, Computer Vision – ECCV 2016, pages 766–782, Cham, 2016. Springer International Publishing. https://doi.org/10.1007/978-3-319-46478-7_47

Zhang, C., Kim, J. (2019). Modeling Long—And Short-Term Temporal Context for Video Object Detection. In Proceedings of the 2019 IEEE International Conference on Image Processing, Taipei, Taiwan (pp. 71–75). doi: 10.1109/ICIP.2019.8802920



Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Zhang, J., Peng, Y. (2019). Hierarchical vision-language alignment for video captioning. In: International Conference on Multimedia Modeling, Springer (pp. 42–54). DOI:10.1007/978-3-030-05710-7_4

Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J.. (2020). Object relational graph with teacherrecommended learning for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (pp. 13278–13288). https://doi.org/10.48550/arXiv.2002.11566

Zheng, G., & Xu, Y. (2021). Efficient face detection and tracking in video sequences based on deep learning. Information Sciences, 568, 265–285. https://doi.org/10.1016/j.ins.2021.03.027

Zhu, X., Dai, J., Yuan, L., Wei, Y. (2018). Towards High Performance Video Object Detection. In Proceedings of the 2018 IEEE/Cvf Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA (pp. 7210–7218). https://doi.org/10.48550/arXiv.1711.11577

Zhu X., Wang Y., Dai J., Yuan L., Wei Y. (2017). Flow-Guided Feature Aggregation for Video Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, (pp. 408–417). Zhu, X., Xiong Y., Dai J., Yuan L., Wei Y. (2017a) https://doi.org/10.48550/arXiv.1703.10025

Zhu W., Lu J., Li J. and Zhou J. (2021). "DSNet: A Flexible Detect-to-Summarize Network for Video Summarization," in IEEE Transactions on Image Processing, vol. 30, pp. 948-962. https://doi.org/10.1109/TIP.2020.3039886

Zhu W., Lu J., Li J. and Zhou J. (2022). "Relational Reasoning Over Spatial-Temporal Graphs for Video Summarization," in IEEE Transactions on Image Processing, vol. 31, pp. 3017-3031. https://doi.org/10.1109/TIP.2022.3163855



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης