ΕΚΕΦΕ Δημόκριτος – Πανεπιστήμιο Πειραιώς

ΔΠΜΣ στην «Τεχνητή Νοημοσύνη»

Εργασία στο μάθημα:

«Μηχανική Μάθηση σε Πολυμεσικά Δεδομένα»

Θέμα:

Μέτρηση οχημάτων που διέρχονται από στατικό σημείο με χρήση κάμερας.

1. *Πρέπει να χρησιμοποιηθεί ακουστική και οπτική πληροφορία σε πραγματικό χρόνο.*
2. *Πρέπει εκτός από την υλοποίηση της μεθοδολογίας, ο κώδικας να εκτελείται σε online mode (realtime).*
3. *Το evaluation πρεπει να γίνει σε 10 videos με συγκεκριμένα πλήθη οχημάτων και το αναφερόμενο σφάλμα να είναι το mean absolute error.*

Εισηγητής:
Θεόδωρος Γιαννακόπουλος

Παύλος Πατσώνης
ΑΜ: 2213

Ημερομηνία Παράδοσης:
3 Ιουλίου 2023

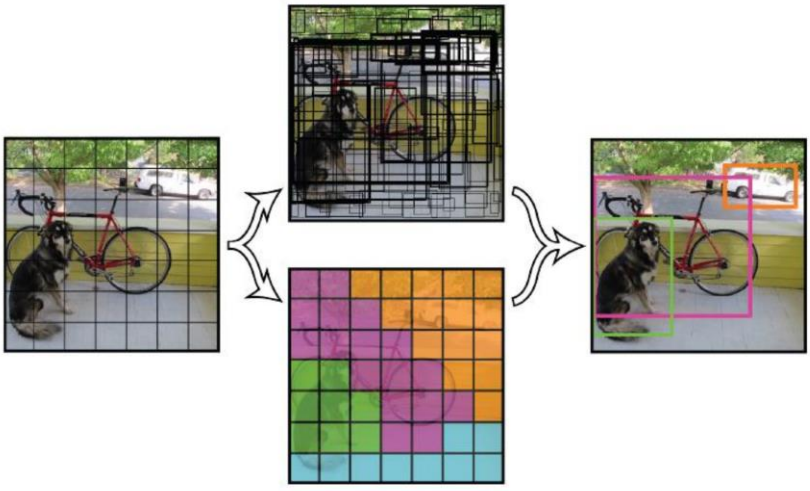
1. Εισαγωγή

Στην παρακάτω αναφορά επιχειρείται η καταμέτρηση οχημάτων από στατικό σημείο με χρήση κάμερας. Για την επίτευξη αυτού, αξιοποιείται εικονική και ηχητική πληροφορία με συνδιαστικό τρόπο, ώστε να παρθεί η απόφαση για καταμέτρηση ή μη κάποιου διερχόμενου αντικειμένου. Κομβικό στοιχείο της υλοποίησης είναι ότι οφείλει να πραγματοποιείται σε πραγματικό χρόνο, δηλαδή το εκάστοτε video προς αξιολόγηση του μοντέλου αναλύεται “frame by frame” εικόνας αλλά και ήχου. Η μετρική αξιολόγησης είναι το mean absolute error, δηλαδη ο μέσος όρος της διαφοράς καταμετρήσεων του μοντέλου από τις πραγματικές. Αρχικά, γίνεται συνοπτική θεωρητική αναφορά στα εργαλεία που χρησιμοποιούνται, στη συνέχεια αναλύεται η στρατηγική για την υλοποίηση και τέλος περιγράφονται τα αποτελέσματα αλλά και κάποιες ιδέες για μελλοντική βελτιστοποίηση του μοντέλου.

2. Στοιχεία από τη θεωρία

2.1. Ο αλγόριθμος YOLO (You Look Only Once)

Το δίκτυο YOLO (You Only Look Once), είναι το πρώτο μοντέλο νευρωνικού δικτύου συνέλιξης που επιλύει το πρόβλημα της ταυτόχρονης αναγνώρισης και εντοπισμού αντικειμένων σε εικόνες, με ένα forward pass. Η ιδιαιτερότητά του είναι ότι αντιμετωπίζει το πρόβλημα ως ένα πρόβλημα regression και όχι classification. Μία ακόμη ιδιαιτερότητα του συγκεκριμένου δικτύου είναι ότι στοχεύει κυρίως στην ταχύτητα της αναγνώρισης καθώς θέτει σαν απαίτηση την εφαρμογή του σε προβλήματα σχεδόν πραγματικού χρόνου με αντάλλαγμα τη μείωση της ακρίβειας αναγνώρισης, η οποία είναι χαμηλότερη σε σχέση με άλλα μοντέλα, όπως για παράδειγμα τα δίκτυα Fast-RCNN. Η έξοδος του δικτύου αντιστοιχείται τόσο στις κλάσεις των αντικειμένων που αναγνωρίστηκαν, καθώς και στις συντεταγμένες των οριοθετημένα πλαίσια (bounding boxes) γύρω από τα αντικείμενα όπου αυτά εντοπίστηκαν. Το πρώτο βήμα είναι να χωρίσει την εικόνα εισόδου σε ένα πλέγμα διαστάσεων S x S. Κάθε κελί του πλέγματος προβλέπει οριοθετημένα πλαίσια (bounding boxes) μαζί με ένα σκορ ”εμπιστοσύνης” για το κάθε πλαίσιο. Το σκορ εμπιστοσύνης ερμηνεύεται ως η βεβαιότητα να ανήκει ένα αντικείμενο στο συγκεκριμένο πλαίσιο μαζί με την ακρίβεια ότι το συγκεκριμένο αντικείμενο ανήκει σε αυτό το πλαίσιο.



Σχήμα 1: Παράδειγμα τρόπου λειτουργίας του αλγορίθμου YOLO

2.1.1. Non-max suppression

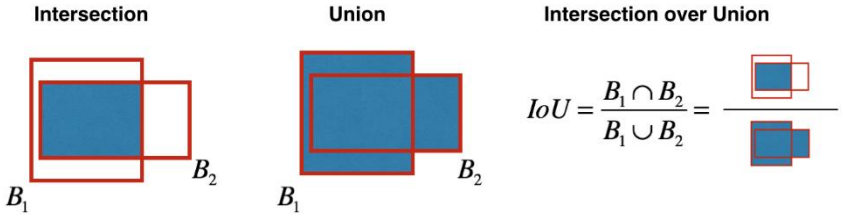
Η non-max suppression (NMS) είναι μια τεχνική που χρησιμοποιείται σε αλγορίθμους ανίχνευσης αντικειμένων όπως ο YOLO, για τη βελτίωση της εξόδου με την αφαίρεση των περιττών προβλέψεων του bounding box, βοηθώντας στην εξάλειψη των διπλών ή επικαλυπτόμενων ανιχνεύσεων του ίδιου αντικειμένου.

Κατά τη διάρκεια της NMS, τα προβλεπόμενα bounding boxes ταξινομούνται βάσει των σκορ εμπιστοσύνης τους, τα οποία αντιπροσωπεύουν την πιθανότητα να περιέχουν ένα αντικείμενο. Ξεκινώντας από το bounding box με το υψηλότερο σκορ εμπιστοσύνης, το NMS επαναλαμβάνει τον

ταξινομημένο κατάλογο και καταστέλλει τυχόν οριοθετημένα πλαίσια που έχουν μεγάλη επικάλυψη με ένα πλαίσιο με υψηλότερο σκορ.

Σχήμα 2: Φιλτράρισμα των bounding boxes μέσω της non-max suppression.

Η NMS χρησιμοποιεί μια θεμελιώδη συνάρτηση με όνομα *Intersection over Union* (IoU). Αποτελεί μια μέτρηση που χρησιμοποιείται για την ποσοτικοποίηση της επικάλυψης μεταξύ δύο bounding boxes. Υπολογίζεται διαιρώντας το εμβαδόν της τομής μεταξύ του προβλεπόμενου πλαισίου (box) και του πλαισίου ground truth με το εμβαδόν της ένωσής τους. Η IoU κυμαίνεται από 0 έως 1, με το 0 να υποδηλώνει μηδενική επικάλυψη και το 1 να υποδηλώνει τέλεια ταύτιση.



Σχήμα 3: Σχηματική απεικόνιση υπολογισμού της IoU.

Στο YOLO, αφού ληφθεί το αρχικό σύνολο προβλέψεων οριοθετημένων πλαισίων, εφαρμόζεται το NMS για την αφαίρεση περιττών ανιχνεύσεων. Ορίζεται ένα κατώφλι για το IoU για να καθοριστεί αν μια ανίχνευση θεωρείται έγκυρη ή false positive. Τα bounding boxes με IoU πάνω από το κατώφλι θεωρούνται έγκυρες ανιχνεύσεις, ενώ όσες είναι κάτω από το κατώφλι απορρίπτονται. Με τη χρήση των NMS και IoU, το YOLO μπορεί να βελτιώσει την έξοδο ανίχνευσης αντικειμένων επιλέγοντας τα πιο αξιόπιστα και μη επικαλυπτόμενα bounding boxes, βελτιώνοντας την ακρίβεια και εξαλείφοντας τον πλεονασμό στις τελικές ανιχνεύσεις.

2.1.2. YOLOv8

Το YOLOv8 είναι η πιο πρόσφατη επανάληψη αυτών των μοντέλων YOLO (από τις αρχές του 2023). Είναι προ-εκπαιδευμένο σε τεράστια σύνολα δεδομένων όπως το COCO και το ImageNet. Παρέχουν πολύ ακριβείς προβλέψεις για τις κλάσεις στις οποίες είναι προεκπαιδευμένοι (master ability) και μπορούν επίσης να μάθουν νέα μαθήματα σχετικά εύκολα (ιδιότητα μαθητή) με κάποια διαδικασία transfer learning. Έχουν επέλθει μερικές σημαντικές αλλαγές από τους προγόνους του, όπως η

ανίχνευση χωρίς άγκυρα (anchor-free boxing), ωστόσο οι βασικές αρχές λειτουργίας παραμένουν οι ίδιες.

Ο αλγόριθμος YOLOv8n που χρησιμοποιείται στην υλοποίηση είναι μια πιο ελαφριά, συμπιεσμένη έκδοση του YOLO. Η έκδοση αυτή βασίζεται πάνω στην ολοκληρωμένη έκδοση και απλοποιώντας την δομή του δικτύου και μειώνοντας κάποιες παραμέτρους γίνεται εφικτό να υλοποιηθεί σε συσκευές με μικρότερη υπολογιστική ισχύ. Το βασικό του πλεονέκτημα είναι η ταχύτητα του και λόγω αυτού επιλέχθηκε αφού σημαντικό στοιχείο για την εργασία είναι η πρόβλεψη σε πραγματικό χρόνο. Το YOLOv8n είναι μικρότερο σε μέγεθος από την κανονική έκδοση του YOLOv8 κάνοντάς το ταχύτερο στην εκτέλεση σε σημαντικό βαθμό. Είναι λογικό να σκεφτεί κανείς πως αυτή η αύξηση στην ταχύτητα που προσφέρει θα έχει κάποιο αντίτιμο. Πράγματι, το YOLOv8n προσφέρει μικρότερη ακρίβεια στις προβλέψεις του για την αναγνώριση των αντικειμένων σε σχέση με το YOLOv8. Το YOLOv8n αποτελείται από πολύ λιγότερα επίπεδα συνέλιξης και επίπεδα συγκέντρωσης, γεγονός που το κάνει να απαιτεί συνολικά πολύ λιγότερα layers από το YOLOv8.

2.2. Ενέργεια RMS

Στην επεξεργασία σήματος ήχου, το Root Mean Square (RMS) είναι ένα χαρακτηριστικό που χρησιμοποιείται συνήθως για τη μέτρηση του συνολικού πλάτους ή ενέργειας ενός σήματος. Υπολογίζεται λαμβάνοντας την τετραγωνική ρίζα του μέσου όρου των τετραγωνικών τιμών των δειγμάτων σήματος ήχου.

Αποτελεί ένα από τα βασικότερα χαρακτηριστικά για εργασίες ανάλυσης ήχου, όπως η αναγνώριση ομιλίας, η ταξινόμηση ειδών μουσικής και η ανίχνευση συμβάντων ήχου, καθώς παρέχει πληροφορίες σχετικά με τη συνολική ενέργεια του σήματος.

Στη βιβλιοθήκη Librosa, το RMS υπολογίζεται χρησιμοποιώντας τη συνάρτηση **`librosa.feature.rms`**. Η συνάρτηση υπολογίζει τις τιμές RMS σε σύντομα, επικαλυπτόμενα πλαίσια του σήματος ήχου. Υπολογίζοντας το RMS σε μικρά καρέ, η Librosa επιτρέπει μια χρονικά μεταβαλλόμενη ανάλυση της ενέργειας του ηχητικού σήματος. Αυτό μπορεί να είναι χρήσιμο για διάφορες εφαρμογές ήχου, όπως τμηματοποίηση ήχου, ανίχνευση έναρξης ή εξαγωγή χαρακτηριστικών για εργασίες ταξινόμησης.

Η εξίσωση για τον υπολογισμό της τιμής Root Mean Square (RMS) ενός σήματος είναι η εξής:

$$RMS = \sqrt{\frac{1}{N}\sum_{i}^{N} {x_i}^2}$$

όπου: N ο συνολικός αριθμός δειγμάτων στο σήμα,

x_i τα μεμονωμένα δείγματα του σήματος.

3. Υλοποίηση

3.1. Συγκέντρωση δεδομένων

Με κάμερα κινητού και χρήση τρίποδου για την επίτευξη στατικότητας, καταγράφονται ~10 min κίνησης οχημάτων σε υπερυψωμένο σημείο και κατακόρυφα, πάνω σε γέφυρα παραδρόμου Εθνικής Οδού. Η τοποθεσία επιλέχθηκε με βασικό σκεπτικό το να είναι ευδιάκριτα και τα δύο ρεύματα κίνησης αλλά και να καταγράφεται ομοιόμορφα η ηχητική πληροφορία. Στη συνέχεια, το βίντεο χωρίζεται σε 10 μικρότερα διάρκειας 15-25 sec έκαστο, διαλέγοντας τα σημεία με πιο αυξημένη ροή οχημάτων. Σε αυτά τα βίντεο που δημιουργήθηκαν, καταμετρώ τον αριθμό οχημάτων σε λίστα με όνομα **`actual_counts`**, που θα χρησιμοποιηθούν στο τέλος για τον υπολογισμό του απόλυτου σφάλματος του κάθε video και εν τέλει το mean absolute error (MAE). Τα video αποθηκεύονται σε φάκελο με όνομα *evaluation_vids* και τοποθετούνται σε Google Drive, σύνδεσμος του οποίου παρατίθεται στο repo της εργασίας.

3.2. Εξαγωγή χρήσιμων πληροφοριών

Για τον σχεδιασμό της οριζόντιας γραμμής για την απόφαση καταμέτρησης, αρχικά βρίσκω τις διαστάσεις του κάθε frame (1280×720). Από αυτές, «κατατοπίζομαι» για την γεωμετρία του κάθε frame ώστε να επιλέξω αντίστοιχα τις συντεταγμένες για τα δύο άκρα της γραμμής διέλευσης. Στη συνέχεια, δημιουργείται μια λίστα που εμπεριέχει τους δείκτες των επιθυμητών κλάσεων του YOLO (COCO dataset) που αφορούν την εργασία και αποθηκεύετια με όνομα vehicle_classes. Αυτές είναι:

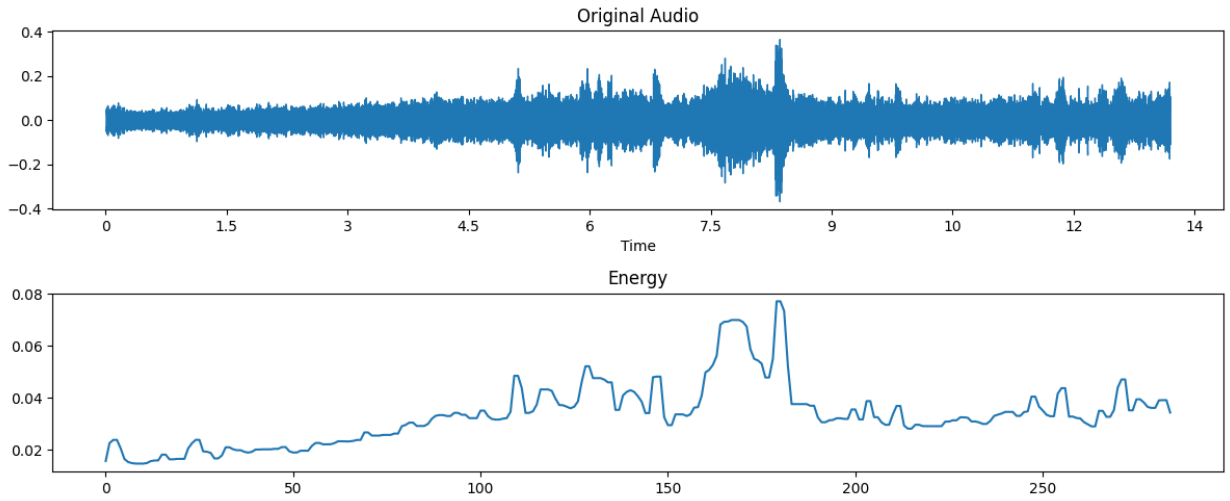
COCO Vehicle Classes	
index	**class**
1	bicycle
2	car
3	motorcycle
5	bus
6	train
7	truck

Πίνακας 1: Δείκτες και αντίστοιχες κλάσεις που επιλέγονται για εντοπισμό αντικειμένων στην υλοποίηση.

Ακόμη, βρίσκω τα FPS (frames per second = 30) των video, για να υπολογιστεί εν συνεχεία η χρονική διάρκεια που αντιστοιχεί στο κάθε frame και να συγχρονιστούν. Πιο συγκεκριμένα:

$$duration = \frac{1}{30} = 0.0333\ sec$$

Κατασκευάζεται ακόμη γράφημα χρονικής εξέλιξης της ενέργειας για την διαλογή τιμής κατωφλίου που θεωρείται θόρυβος, παράδειγμα του οποίου ακολουθεί παρακάτω:



Σχήμα 2: Γραφήματα χρονικής εξέλιξης του **`audio_data`** (πάνω) και **`energy`** (κάτω) για το 4ο video αξιολόγησης του μοντέλου.

Η τιμή ενέργειας κάτω από την οποία θεωρώ πως η ηχητική πληροφορία αποτελεί θόρυβο είναι 0.02.

3.3. Κατασκευή καταμετρητή

Η βασική ιδέα για την υλοποίηση στηρίζεται στα εξής βήματα:

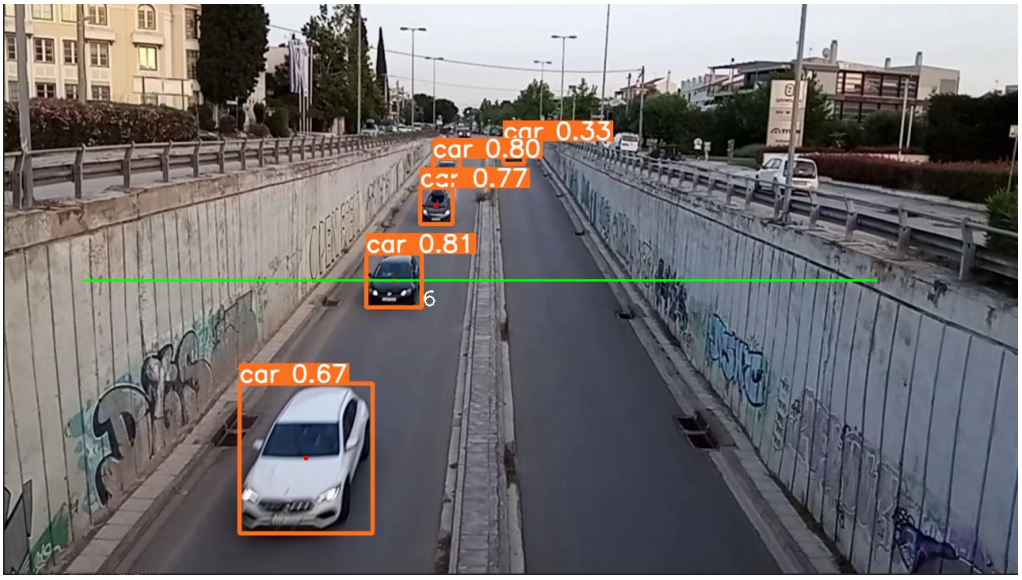
1. Εντοπισμός και παρακολούθηση των αντικειμένων που θεωρούνται οχήματα κατά το YOLO και η τοποθέτηση σημείου στο κέντρο (**`center_point`**) του κάθε bounding box.

2. Σχεδιασμός και κατάλληλη τοποθέτηση οριζόντιας γραμμής επί της εικόνας, η οποία χρησιμοποιείται ως «σκανδάλη» και ενεργοποιείται κάθε φορά που περνάει το **`center_point`** από αυτήν.

3. Συνδιασμός της προαναφερθείσας πληροφορίας με την ενέργεια του ηχητικού σήματος, η οποία οφείλει να ξεπερνάει την τιμή κατωφλίου που τίθεται ως θόρυβος, για την πάρση απόφασης για καταμέτρηση του εκάστοτε αντικειμένου.

4. Υπολογισμός και αποθήκευση αποκλίσεων των καταμετρήσεων του μοντέλου (**`model_counts`**) από το ground truth (**`actual_counts`**).

5. Η αυτοματοποιημένη επανάληψη των παραπάνω βημάτων για 10 video αξιολόγησης και ο υπολογισμός του mean absolute error (MAE).



Σχήμα 3: Εντοπισμός οχημάτων και καταμέτρηση μέσω ενεργοποίησης της γραμμής διέλευσης.

4. Αποτελέσματα

Στη συνέχεια, καταγράφονται τα αποτελέσματα που προέκυψαν από τον καταμετρητή της υλοποίησης και συγκρίνονται με τις πραγματικές τιμές:

video index	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
`model_counts`	6	3	10	7	2	5	8	3	5	4
`actual_counts`	5	4	10	7	4	6	7	5	5	4

Πίνακας 2: Σύγκριση πειραματικών και πραγματικών τιμών των οχήματα για τα 10 video αξιολόγησης.

Παρατηρώ ότι δεν υπάρχει μεγάλη απόκλιση μεταξύ παρατηρούμενων και πραγματικών τιμών. Υπήρξαν 2 περιπτώσεις που το μοντέλο μέτρησε περισσότερα οχήματα ενώ 4 στις οποίες μέτρησε λιγότερα. Θα μπορούσαμε να πούμε ότι ο καταμετρητής είναι «συντηρητικός», με την έννοια του ότι υπερτερούν τα false negatives. Στις υπόλοιπες 6 περιπτώσεις, ο καταμετρητής προσδιόρισε σωστά τον αριθμό διέλευσης οχημάτων.

Τέλος, μένει να υπολογιστεί το mean absolute error, το οποίο βρίσκεται από την σχέση:

$$MAE = \frac{\sum_{i=1}^{10} |model\_counts_i - actual\_counts_i|}{10} = 0.8 \text{ οχήματα}$$

5. Μελλοντικές βελτιστοποιήσεις

Στο τελικό κεφάλαιο της αναφοράς, παραθέτω κάποια γενικά σχόλια καθώς και ιδέες για την μελλοντική βελτιστοποίηση του μοντέλου.

Μελέτη False Positive – False Negative καταμετρήσεων

Η συγκεκριμένη τιμή $MAE = 0.8$ δεν είναι αντιπροσωπευτική του πραγματικού απολύτου σφάλματος του μοντέλου. Λόγω των FPS του video κάποια οχήματα ενεργοποιούν την γραμμή διέλευσης διπλά επομένως μετρώνται ως δύο, ενώ σε άλλες περιπτώσεις δεν καταμετρώνται καθόλου. Ωστόσο τα δύο αυτά διαφορετικά σφάλματα κατά μία έννοια «αλληλοαναιρούνται» με αποτέλεσμα να επιτυγχάνεται η προαναφερθείσα τιμή. Για την επίλυση αυτού, θα είχε ενδιαφέρον μια μελέτη και καταγραφή των false positive και false negative καταμετρήσεων. Μια λύση για την εξάλειψη των false negative καταμετρήσεων είναι η χρήση κάμερας με περισσότερα FPS.

Επανασχεδιασμός πειράματος

Η ηχητική πληροφορία για καταμέτρηση οχημάτων δεν είναι εύκολα αξιοποιήσιμη. Λόγω της μορφολογίας του σημείου που γίνεται η λήψη, η ένταση του σήματος από το κάθε αυτοκίνητο που περέρχεται δεν είναι αρκετή ώστε να ξεχωρίζει με ευκολία. Για να επιτευχθεί κάτι τέτοιο, χρειάζεται επανασχεδιασμός του πειράματος με χρήση κατευθυντικών μικροφώνων κοντά στην ροή των αυτοκινήτων, τα οποία θα «βλέπουν» την γραμμή διεύλευσης από το YOLO. Ωστόσο, κάτι τέτοιο αυξάνει την πολυπλοκότητα χωρίς ωστόσο αναγκαστικά να επιφέρει καλύτερα αποτελέσματα από τη καταμέτρηση του YOLO.

Καλύτερη αξιοποίηση ηχητικής πληροφορίας

ένταση θα μπορούσε εύκολα να αυξηθεί και από άλλες αιτίες πχ ομιλία, μουσική κλπ, χωρίς επομένως να προέρχεται από όχημα. Το παραπάνω θα μπορούσε να βελτιοποιηθεί, εξετάζοντας την εν δυνάμει συνεισφορά φασματικής ανάλυσης. Ενδεχομένως θα είχε νόημα και μια διαφορετική προσέγγιση από την “rule-based” που επέλεξα, αξιοποιώντας και άλλα χαρακτηριστικά του ήχου, όπως φασματογράμματα κλπ ή η κατασκευή και εκπαίδευση “from scratch” κάποιου μοντέλου για αναγνώριση οχημάτων μέσω ηχητικής πληροφορίας.