ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Θεματα Ορασης Υπολογιστων

Συνελικτικα Νευρωνικα Δικτυα

Ανιχνευση & Εντοπισμος Αντικειμενων

Διδάσκων: Αναπλ. Καθηγητής Εμμανουήλ Ζ. Ψαράκης
Επικουρικό έργο: Παναγιώτης Κάτσος, Παναγιώτης Γεωργαντόπουλος

Πάτρα Δεκέμβριος 2021

ΣΤΟΙΧΕΙΩΔΗΣ ΘΕΩΡΙΑ

Ο ανίχνευση (detection) και ο εντοπισμός (localization) αντικειμένων αποτελεί ένα από τα βασικά προβλήματα του πεδίου της υπολογιστικής όρασης με αναρίθμητες εφαρμογές. Οι απαιτήσεις αυτών των εφαρμογών σε πεδία όπως:

- της αυτόνομης οδήγησης
- των συστημάτων ασφαλείας
- του κλάδου της ιατρικής κ.ά.

απαιτούν συστήματα τα οποία θα αναγνωρίζουν τα αντικείμενα ενδιαφέροντος με μεγάλη ακρίβεια και μικρό ποσοστό εσφαλμένων ανιχνεύσεων.

Το πρόβλημα της ανίχνευσης ενός αντικειμένου απαιτεί την επίλυση των ακόλουθων δύο (2) επιμέρους υποπροβλημάτων:

- της **κατηγοριοποίησης** (classification) ενός αντικειμένου : διαδικασία κατά την οποία αναθέτουμε ένα αντικείμενο ενδιαφέροντος σε μια κλάση
- ο **προσδιορισμός της θέσης** (localization) του αντικειμένου : διαδικασία κατά την οποία το αντικείμενο ενδιαφέροντος οριοθετείται μέσα σε ένα πλαίσιο (bounding box) και διαχωρίζεται από άλλα αντικείμενα σε μια εικόνα.

Τα τελευταία χρόνια με την ανάπτυξη των νευρωνικών δικτύων βαθέων αρχιτεκτονικών, πολλοί από τους πιο διαδεδομένους ανιχνευτές αιχμής βασίζονται στην χρήση συνελικτικών νευρωνικών δικτύων καθώς έχει παρατηρηθεί ότι τα δίκτυα αυτά έχουν πολύ καλές επιδόσεις σε εφαρμογές που σχετίζονται με την οπτική πληροφορία.

Στη συνέχεια θα υποθέσουμε ότι έχουμε στην διάθεσή μας το παρακάτω σύνολο δεδομένων:

$$\mathcal{D} = \left\{ I_m \right\}_{m=1}^{M}$$

που θα χρησιμοποιήσουμε για την εκπαίδευση των νευρωνικών δικτύων. Για το σύνολο αυτό θεωρούμε επίσης ότι:

- υπάρχουν L κλάσεις αντικειμένων $\mathcal{C}_l,\ l = 1, 2, \cdots, L$
- για την εικόνα I_m του παραπάνω συνόλου έχουμε στην διάθεσή μας το σύνολο:

  $$\mathcal{G}_m = \{\mathbf{g}^j\}_{j=1}^{J_m}$$

  το οποίο περιέχει τα ground truth πλαίσια που περιέχονται στην εικόνα αυτή, όπου J_m το πλήθος των.

Πρέπει να τονίσουμε εδώ ότι αφού η εκπαίδευση του δικτύου είναι επιβλεπόμενη, κάθε ground truth πλαίσιο απεικονίζεται και σε μία και μόνο μία κλάση αντικειμένων. Αυτή η παρατήρηση, όπως θα εξηγήσουμε και θα δούμε παρακάτω είναι καθοριστικής σημασίας.

Τέλος, στα παρακάτω θα χρησιμοποιήσουμε τον ακόλουθο συμβολισμό για τις συναρτήσεις είτε αυτές αναφέρονται στις συναρτήσεις κόστους χρησιμοποιούνται για την περιγραφή ενός νευρωνικού δικτύου, ή την συνάρτηση πυκνότητας πιθανότητας σππ :

- $f(., \theta)$ αν οι παράμετρες θ **μετέχουν** κατά την εκπαίδευση ενός δικτύου (είναι δηλαδή learnable)
- $f_\theta(.)$ αν οι παράμετρες θ **δεν μετέχουν** κατά την εκπαίδευση ενός δικτύου.

Συνελικτικά δίκτυα βασιζόμενα σε περιόχες

R-CNN

Το R-CNN (Region-Based Convolutional Neural Network) [1] αποτελεί την πρώτη γενιά της συγκεκριμένης οικογενείας ανιχνευτών. Το συγκεκριμένο σύστημα ανίχνευσης βασίζεται στα ακόλουθα τρία (3) βασικά υποσυστήματα:

- Υποσύστημα δημιουργίας **υποψήφιων περιοχών** ανίχνευσης
- Υποσύστημα **Εξαγωγής Χαρακτηριστικών** και
- Υποσύστημα **Κατηγοριοποίησης Κλάσεων**.

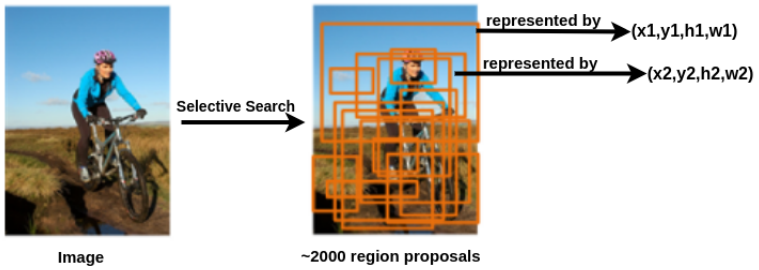
Για κάθε εικόνα από το σύνολο εκπαίδευσης που μας δίνεται, γίνεται η παρακάτω διαδικασία, αρχής γενομένης από το υποσύστημα δημιουργίας προτεινόμενων περιοχών.

1. **Υποσύστημα δημιουργίας υποψήφιων περιοχών ανίχνευσης**: Το υποσύστημα αυτό, για κάθε εικόνα της δέσμης των δεδομένων (batch size) με την χρήση του αλγορίθμου **επιλεκτικής αναζήτησης** δημιουργεί N_0 (περίπου 2000) υποψήφιες περιοχές ενδιαφέροντος (ROIs) οι οποίες πιθανόν να περιέχουν αντικείμενα. Όπως μπορούμε να δούμε στο Σχήμα 1 οι περιοχές αυτές, οι οποίες συμβολίζονται με:

   $$\mathbf{p}^n = [p_{x_n}\ p_{y_n}\ p_{w_n}, p_{h_n}]^t\ n = 1, 2, \cdots, N_0,$$

   οριοθετούνται με την βοήθεια τεσσάρων προτεινόμενων (από τον αλγόριθμο επιλεκτικής αναζήτησης) τιμών και συγκεκριμένα από:

   - τις συντεταγμένες $[p_{x_n}\ p_{y_n}]^t$ της πάνω αριστερής κορυφής της n-οστής προτεινόμενης περιοχής
   - το προτεινόμενο πλάτος (width) p_{w_n} και
   - το προτεινόμενο ύψος (height) p_{h_n} της.



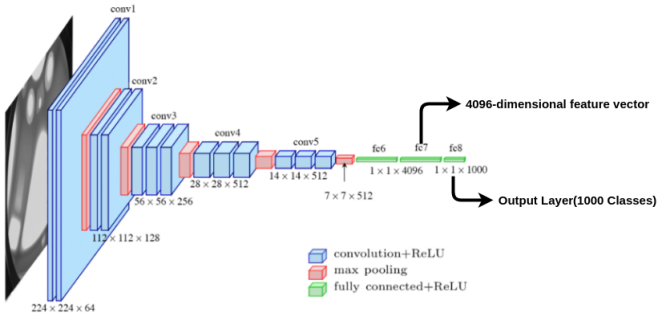
Σχήμα 1: Οι υποψήφιες περιοχές που έχουν προκύψει από τον αλγόριθμο επιλεκτικής αναζήτησης [2].

2. **Εξαγωγή Χαρακτηριστικών**: Οι υποψήφιες περιοχές που έχουν παραχθεί από το προηγούμενο υποσύστημα, μετασχηματίζονται ώστε να έχουν όλες την ίδια διάσταση (227×227) και κάθε μία από αυτές οδηγείται στην είσοδο ενός συνελικτικού νευρωνικού δικτύου (CNN) το οποίο αποτελείται από:

(α΄) πέντε συνελικτικά επίπεδα και

(β΄) δύο πλήρως συνδεδεμένα επίπεδα.

Το δίκτυο, για κάθε μιά υποψήφια περιοχή $\mathbf{p}^n$, υπολογίζει ένα διάνυσμα χαρακτηριστικών της περιοχής μεγέθους 1×4096, όπως μπορούμε να δούμε στο Σχήμα 2,



Σχήμα 2: Αρχιτεκτονική ενός συνελικτικού δικτύου. Στην υλοποίηση του R-CNN το τελευταίο επίπεδο fc8 αποκόπτεται [2].

και η υποψήφια περιοχή θα πρέπει να αντιστοιχιθεί σε ένα μοναδικό ground truth πλαίσιο του συνόλου $\mathcal{G}_m$ που όπως ήδη αναφέραμε περιέχει όλα τα ground truth πλαίσια που υπάρχουν στην εικόνα I_m την οποία επεξεργάζεται το δίκτυο.

Πρέπει να ξανατονίσουμε εδώ ότι αφού η εκπαίδευση του δικτύου είναι επιβλεπόμενη, κάθε ground truth πλαίσιο απεικονίζεται σε μί μό μί κλάση αντικειμένων. Επομένως, αν κάνουμε την αντιστοίχιση του $\mathbf{p}^n$ σε ένα $\mathbf{g}^{j^\star}$ εμμέσως έχει προσδιορισθεί και η κλάση στην οποία ανήκει το αντικείμενο που υπάρχει σε αυτή την περιοχή. Άρα είναι καθοριστικής σημασίας η αντιστοίχιση της προτεινόμενης περιοχής σε ένα ground truth πλαίσιο. Για το σκοπό αυτό εκτελούμε τον Αλγόριθμο 1 και αντιστοιχίζουμε το υποψήφιο πλαίσιο της περιοχής με εκείνο το ground truth πλαίσιο με το οποίο εμφανίζει την μεγαλύτερη επικάλυψη IoU, εφόσον βέβαια αυτή ξεπερνά ένα προκαθορισμένο κατώφλι T_0. Παρατηρήστε ότι με το πέρας της παραπάνω διαδικασίας έχουμε πετύχει τα ακόλουθα:

Αλγόριθμος

1:Αντιστοίχιση προτεινόμενης περιοχής σε ground truth πλαίσιο**;;**

Input $\mathbf{p}^n$
for $j = 1 : J_m$
    Compute $s_m(j)$ =IoU($\mathbf{g}^j,\ \mathbf{p}^n$)
endfor
Compute $M^\star = \max_{j=1,2\cdots,J_m} \{s_m(j)\}$
Compute $j^\star = \arg\max_{j=1,2\cdots,J_m} \{s_m(j)\}$
If $M^\star \geq\ T_o$
  $n = j^\star$
else
  $n = 0$
endIf

(α′) Αν $n = 0$: η προτεινόμενη περιοχή, αφού δεν αντιστοιχήθηκε σε κανένα ground truth πλαίσιο, θεωρείται ως μία περιοχή υποβάθρου.

(β′) Αν $n \neq 0$: η προτεινόμενη περιοχή έχει αντιστοιχηθεί στο $j^\star$ ground truth πλαίσιο, δηλαδή:

$$\mathbf{p}^n \rightarrow \mathbf{g}^{j^\star}$$

και το αντικείμενο που υπάρχει στην προτεινόμενη περιοχή (ή ισοδύναμα στο $\mathbf{g}^{j^\star}$ ground truth πλαίσιο) ανήκει στην κλάση $\mathcal{C}_l$. Κατά συνέπεια, διαθέτουμε, εκτός του διανύσματος χαρακτηριστικών το οποίο οδηγείται στην είσοδο όλων των SVMs:

- ένα θετικό παράδειγμα για το SVM της κλάσης $\mathcal{C}_l$ και
- ένα αρνητικό παράδειγμα εκπαίδευσης για όλα τα υπόλοιπα $L-1$ SVM που χρησιμοποιούνται για τις άλλες κλάσεις αντικειμένων που υπάρχουν στα δεδομένα μας.

Επαναλαμβάνοντας την παραπάνω διαδικασία για κάθε εικόνα του batch, δημιουργούμε τα ακόλουθα L σύνολα:

$$\mathcal{S}_l = \{\mathbf{p}_l^n, \mathbf{g}_l^n\}_{n=1}^{N_l},\ l = 1, 2, \cdots, L \tag{1}$$

με N_l ζευγάρια κάθε ένα από αυτά, όπου:

$$\sum_{l=1}^{L} N_l = |\mathcal{D}|$$

με $|\mathcal{D}|$ ο πληθικός αριθμός του συνόλου $\mathcal{D}$ και:

$$\mathbf{p}_l^n = [p_{x_l}^n, p_{y_l}^n, p_{w_l}^n, p_{h_l}^n]^t$$

ένα διάνυσμα το οποίο,όπως ήδη έχουμε αναφέρει, περιέχει τις προβλεπόμενες συντεταγμένες της πάνω αριστερής κορυφής του υποψήφιου πλαισίου καθώς και το ύψος και το μήκος του. Αντίστοιχα, το διάνυσμα $\mathbf{g}_l^n$ περιέχει τις αντίστοιχες πληροφορίες για το ground truth πλαίσιο που έχει επιλεχθεί μέσω της διαδικασίας αντιστοίχισης που προαναφέρθηκε.

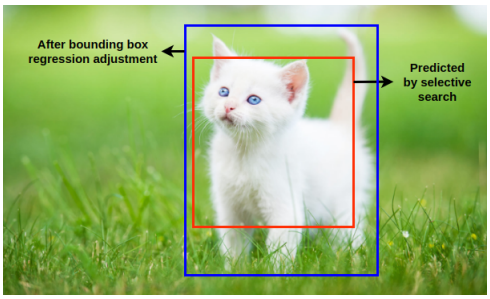
Εχοντας δημιουργήσει τα προαναφερθέντα σύνολα και τα αντίστοιχα χαρακτηριστικά διανύσματα μπορούμε να προχωρήσουμε στην εκπαίδευση, με την ανανέωση των βαρών του δικτύου, και συγκεκριμένα την:

(α΄) **κατηγοριοποίηση των Κλάσεων** με την εκπαίδευση ενός SVM για κάθε κλάση $\mathcal{C}_l$. Δηλαδή, την εκπαίδευση συνολικά L SVM. Πρέπει να αναφέρουμε εδώ ότι κατά την εκπαίδευση των SVM μπορούμε να εκπαιδεύουμε συγχρόνως το CNN ή να το έχουμε προεκπαιδεύσει, και

(β΄) **παλινδρόμηση πλαισίων**: κατά την οποία για κάθε υποψήφια περιοχή που έχουμε επεξεργαστεί στο προηγούμενο βήμα υπολογίζεται το νέο πλαίσιο της περιοχής, ώστε να βελτιωθεί ο προσδιορισμός της θέσης της και του μεγέθους της. Για παράδειγμα, μπορεί τα πλαίσια που έχουν προκύψει στο πρώτο βήμα της διαδικασίας να μην περικλείουν ολόκληρο το αντικείμενο, αλλά μόνο μέρος του, όπως φαίνεται στο Σχήμα 3 που ακολουθεί.

Αναλυτικότερα, σκοπός του συγκεκριμένου βήματος είναι:

να υπολογιστούν κατάλληλοι μετασχηματισμοί κάθε πλαισίου $\mathbf{p}_l^n,\ n = 1, 2, \cdots, N_l$ **που έχει παραχθεί από την επι-**



Σχήμα 3: Σύγκριση πλαισίων αναγνώρισης πριν και μετά την παλινδρόμηση [2].

λεκτική αναζήτηση σε σχέση με τα ground truth για κάθε κλάση $\mathcal{C}_l,\ l = 1, 2, \cdots, L$.[1]

Ο παραπάνω μετασχηματισμός, για κάθε προτεινόμενη περιοχή, παραμετροποιείται βάσει τεσσάρων διαφορετικών συναρτήσεων και συγκεκριμένα των

(α') $d_x(\mathbf{p}^n, \mathbf{w}_x)$

(β') $d_y(\mathbf{p}^n, \mathbf{w}_y)$

(γ') $d_w(\mathbf{p}^n, \mathbf{w}_w)$ και

(δ') $d_h(\mathbf{p}^n, \mathbf{w}_h)$.

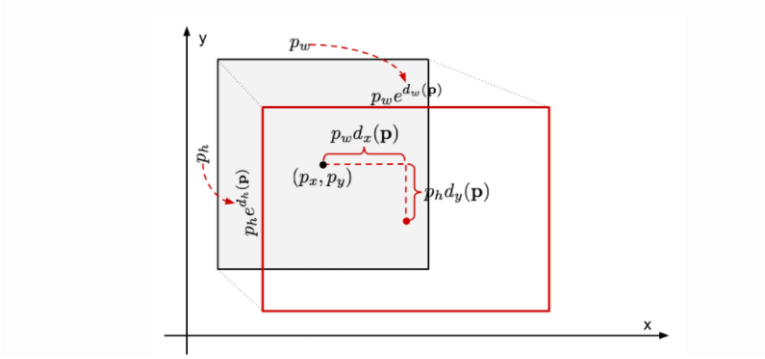
Γνωρίζοντας αυτές τις τέσσερεις συναρτήσεις μπορούμε να ορίσουμε τους παρακάτω μετασχηματισμούς:

$$\begin{aligned} \hat{g}_x^n(\mathbf{w}_x) &= p_w^n d_x(\mathbf{p}^n, \mathbf{w}_x) + p_x^n \\ \hat{g}_y^n(\mathbf{w}_y) &= p_h^n d_y(\mathbf{p}^n, \mathbf{w}_y) + p_y^n \\ \hat{g}_w^n(\mathbf{w}_w) &= p_w^n \exp(d_w(\mathbf{p}^n, \mathbf{w}_w)) \\ \hat{g}_h^n(\mathbf{w}_h) &= p_h^n \exp(d_h(\mathbf{p}^n, \mathbf{w}_h)) \end{aligned} \tag{2}$$

με την βοήθεια των οποίων μπορούν να προκύψουν εκτιμήσεις των ground truth συντεταγμένων και του μεγέθους της περιοχής. Ένα

[1]Από το σημείο αυτό και πέρα για απλοποίηση του συμβολισμού δεν ξαναχρησιμοποιούμε τον υποδείκτη l. Ουσιαστικά, η διαδικασία που ακολουθεί από το σημείο αυτό και μέχρι την ολοκλήρωση της εκπαίδευσης του δικτύου, μπορεί να θεωρηθεί ότι αναφέρεται στην εκπαίδευση μίας συγκεκριμένης κλάσης αντικειμένων.

παράδειγμα του αποτελέσματος των παραπάνω μετασχηματισμών για μία προβλεπόμενη περιοχή, φαίνεται στο Σχήμα 4.



Σχήμα 4: Γεωμετρική αναπαράσταση του μετασχηματισμού μεταξύ των πραγματικών και των προβλεπόμενων πλαισίων

Κάθε συνάρτηση $d_*(\mathbf{p}^n, \mathbf{w}_*)$, όπου $* \in \{x,\ y,\ h,\ w\}$, ορίζεται από το εσωτερικό γινόμενο των χαρακτηριστικών που προκύπτουν από το τελευταίο επίπεδο υποδειγματοληψίας του συνελικτικού δικτύου, δηλαδή της συνολικής εξόδου του δικτύου $\mathbf{y}_{5\theta}(.)$ για την περιοχή $\mathbf{p}^n$ και του διανύσματος $\mathbf{w}_*$, δηλαδή:

$$d_*(\mathbf{p}^n, \mathbf{w}_*) =< \mathbf{w}_*,\ \mathbf{y}_{5\theta}(\mathbf{p}^n) > .$$

όπου θ οι παράμετροι του CNN[2].

Οι τιμές των στοιχείων του διανύσματος $\mathbf{w}_*$ προκύπτουν από την ελαχιστοποίηση του ακόλουθου προβλήματος κανονικοποιημένων ελαχίστων τετραγώνων (ridge regression):

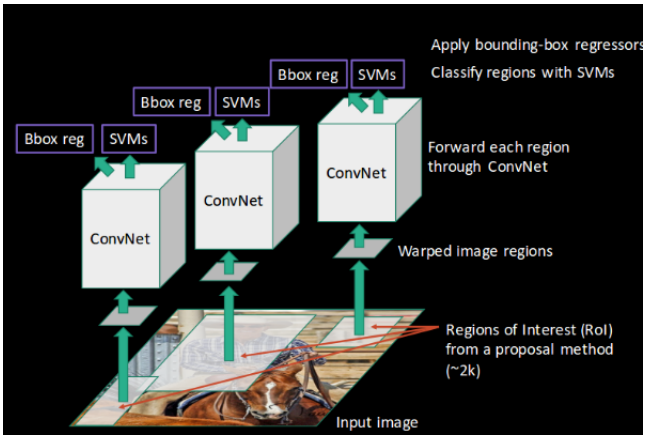
$$\mathbf{w}_*^\star = \underset{w_*}{\operatorname{argmin}} \sum_{n=1}^{N_l} \left(t_*^n - \mathbf{w}_*^t \mathbf{y}_{5\theta}(\mathbf{p}^n) \right)^2 + \lambda ||\mathbf{w}_*||_2^2 \qquad (3)$$

όπου $||\mathbf{x}||_2$ η l_2 στάθμη του διανύσματος $\mathbf{x}$ και για τις μεταβλητές

[2]Στην παραπάνω συνάρτηση κόστους οι παράμετροι του CNN θεωρούνται γνωστοί.

t_*^n ισχύουν οι ακόλουθες σχέσεις:

$$\begin{aligned} t_x^n &= (g_x^n - p_x^n)/p_w^n \\ t_y^n &= (g_y^n - p_y^n)/p_h^n \\ t_w^n &= \log(g_w^n/p_w^n) \\ t_h^n &= \log(g_h^n/p_h^n). \end{aligned} \tag{4}$$



Σχήμα 5: Συνολική απεικόνιση του ανιχνευτή R-CNN [1].

Η λύση του προβλήματος (1) είναι κλειστής μορφής και δίνεται από την ακόλουθη σχέση:

$$\mathbf{w}_*^\star = \Big(\lambda I + \Phi\Big)^{-1}\mathbf{b},$$

όπου

$$\begin{aligned} \Phi &= \sum_{n=1}^{N_l} \mathbf{y}_{5\theta}(\mathbf{p}^n)\mathbf{y}_{5\theta}(\mathbf{p}^n)^t \\ \mathbf{b} &= \sum_{n=1}^{N_l} t_*^n \mathbf{y}_{5\theta}(\mathbf{p}^n). \end{aligned}$$

Η τιμή της παραμέτρου κανονικοποίησης προτάθηκε να τεθεί $\lambda = 1000$. Μια απεικόνιση του ανιχνευτή R-CNN φαίνεται στο Σχήμα 5.

Το R-CNN, αν και είχε πολύ καλές επιδόσεις για την εποχή που προτάθηκε, έχει τα ακόλουθα βασικά μειονεκτήματα:

- (α΄) Εξαιτίας του γεγονότος ότι ο ανιχνευτής R-CNN αποτελείται από τρία διαφορετικά υποσυστήματα (επιλεκτική αναζήτηση, CNN, SVMs) απαιτείται η εγγραφή και η αποθήκευση των αποτελεσμάτων.

  (β΄) Κάθε εικόνα υποδιαιρείται σε πάρα πολλές υποψήφιες περιοχές και επομένως ο χρόνος εκπαίδευσης είναι μεγάλος. Ο χρόνος εκπαίδευσης επιβαρύνεται ακόμα περισσότερο, αν σκεφτούμε ότι χρησιμοποιούμε τόσα διαφορετικά SVM, όσες οι διαφορετικές κλάσεις.

  Επομένως η εκπαίδευση είναι ακριβή τόσο από άποψη χώρου όσο και από άποψη χρόνου.

- Η χρήση του συγκεκριμένου ανιχνευτή είναι απαγορευτική για εφαρμογές πραγματικού χρόνου καθώς κατά την διαδικασία δοκιμής, η πρόβλεψη απαιτεί 47 δευτερόλεπτα/εικόνα.
- Η διαδικασία εκπαίδευσης εξαρτάται σε μεγάλο βαθμό από τον αλγόριθμο **επιλεκτικής αναζήτησης** (δείτε στο Παράρτημα), ο οποίος δεν είναι αλγόριθμος μάθησης.

Οι παραπάνω λόγοι οδήγησαν στην ανάπτυξη ταχύτερων και πιο ρωμαλέων ανιχνευτών, όπως αυτός που περιγράφεται στην συνέχεια.

Fast R-CNN

Το δίκτυο αυτό [3] αποτελεί την μετεξέλιξη του R-CNN. Η ειδοποιός διαφορά, σε σχέση με τον R-CNN, είναι πως **αντί** να επεξεργάζεται την κάθε υποψήφια περιοχή αντικειμένου ως εικόνα, λαμβάνει υπόψην μόνο:

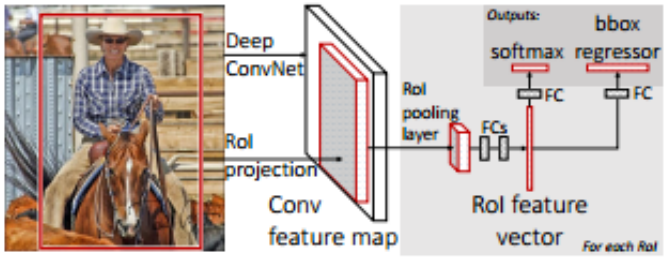
- τις συντεταγμένες $[p_x^n \; p_y^n]^t$ της πάνω αριστερής γωνίας της, και
- τις διαστάσεις $p_w^n, \; p_h^n$ της υποψήφιας περιοχής.

Αναλυτικότερα, με την χρήση ενός συνελικτικού δικτύου παράγονται οι χάρτες χαρακτηριστικών ολόκληρης της εικόνας όπως φαίνεται στο Σχήμα 6. Βάσει των συντεταγμένων της κάθε περιοχής ενδιαφέροντος:

1. εντοπίζεται η προβολή της πάνω στους χάρτες χαρακτηριστικών της περιοχής και

2. οδηγείται στην είσοδο ενός επιπέδου υποδειγματοληψίας (ROI max pooling layer). Η χρήση αυτού του επιπέδου βασίζεται σε μία παράμοια ιδέα με αυτή του SPPnet, και συγκεκριμένα στην χρήση πυραμίδας (στον Fast R-CNN έχουμε πυραμίδα ενός επιπέδου) που αναλύθηκε σε προηγούμενη άσκηση και έχει ως σκοπό μια σταθερή αναπαράσταση των χαρτών των περιοχών ενδιαφέροντος ώστε να εισαχθούν στα πλήρη συνδεδεμένα επίπεδα.

Στη συνέχεια με την χρήση δύο πλήρως συνδεδεμένων επιπέδων παράγεται το μήκους 4096 διάνυσμα χαρακτηριστικών (ROI feature vector) το οποίο τροφοδοτεί τις εισόδους των υποσυστημάτων ί και της όμ ί, όπως φαίνεται και στο Σχήμα 6.



Σχήμα 6: Απεικόνιση του Fast R-CNN [3].

Όπως μπορούμε να δούμε στο παραπάνω σχήμα, υπάρχει μια σημαντική διαφορά στην αρχιτεκτονική του Fast R-CNN σε σχέση με αυτή του Fast R-CNN και συγκεκριμένα:

η αντικατάσταση των SVMs με ένα επίπεδο softmax για την κατηγοριοποίηση των αντικειμένων.

Αυτός είναι ο λόγος για τον οποίο ο ανιχνευτής Fast R-CNN είναι ταχύτερος εξαιτίας του χρόνου που γλιτώνει με τον τρόπο που εντοπίζει τις υποψήφιες περιοχές. Ωστόσο, η εκπαίδευσή του βασίζεται και πάλι στον αλγόριθμο επιλεκτικής αναζήτησης με αποτέλεσμα η διάρκεια εκπαίδευσής του να είναι και πάλι μεγάλη.

Η εκπαίδευση του Fast R-CNN βασίζεται σε μια διαφορετική συνάρτηση κόστους από αυτή του R-CNN και συγκεκριμένα, στη συνάρτηση κόστους που χρησιμοποιεί το δίκτυο αυτό ενσωματώνονται:

- η διαδικασία της κατηγοριοποίησης όπως και
- αυτή της παλινδρόμησης των πλαισίων

με ένα βέβαια διαφορετικό τρόπο από αυτή του R-CNN.

Συγκεκριμένα, στην έξοδο του CNN παράγεται για κάθε προτεινόμενη περιοχή $\mathbf{p}^n$ ένα διάνυσμα μήκους L κάθε στοιχείο του οποίου περιέχει την πιθανότητα το προτεινόμενο πλαίσιο να ανήκει στην αντίστοιχη κλάση, δηλαδή:

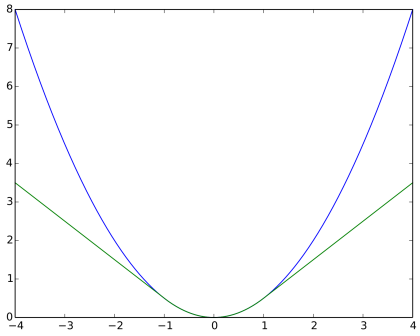
$$\mathbf{c}(\theta) = [c_0(\theta)\ c_1(\theta)\ \cdots\ c_L(\theta)]^t,$$

όπου, όπως ήδη έχουμε αναφέρει, L είναι ο αριθμός των **διαφορετικών** κλάσεων αντικειμένων. Στην περίπτωση του Fast R-CNN, θεωρούμε εξ αρχής και μια επιπρόσθετη κλάση η οποία αναφέρεται σε οτιδήποτε δεν ανήκει στις κλάσεις $\mathcal{C}_l,\ l = 1, 2, \cdots, L$ και την συμβολίζουμε ως $\mathcal{C}_0$.

Με βάση τα όσα αναφέραμε στον ανιχνευτή R-CNN ας υποθέσουμε ότι έχουν παραχθεί τα σύνολα $\mathcal{S}_l$ της Σχέσης (1). Τότε, για κάθε ένα από αυτά τα σύνολα, χρησιμοποιώντας τη Σχέση (4) μπορούμε να παράξουμε τα ακόλουθα διανύσματα (offsets):

$$\mathcal{T}_l = \{\mathbf{t}_l^n\}_{n=1}^{N_l}$$

όπου $\mathbf{t}_l^n = [t_{lx}^n\ t_{ly}^n\ t_{lw}^n\ t_{lh}^n]^t$ και να ορίσουμε την ακόλουθη συνάρτηση



Σχήμα 7: $S_{L1}(x)$ (πράσινο) και $||x||_2^2$ (μπλε).

κόστους:

$$\mathcal{L}_{FAST}(W, \theta) = \mathcal{L}_{CLASS}(\theta) + \lambda \frac{1}{L} \sum_{l=1}^{L} \sum_{*\in\{x,y,w,h\}} \mathcal{L}_{LOC}(\mathbf{w}_{l*}, \theta)$$

όπου:

$$\mathcal{L}_{CLASS}(\theta) = -\frac{1}{1+L}\sum_{l=0}^{L}\frac{1}{N_l}\sum_{n=1}^{N_l}\log(c_l^n(\theta))$$

$$\mathcal{L}_{LOC}(\mathbf{w}_{l*},\theta) = \frac{1}{N_l}\sum_{n=1}^{N_l} S_{L_1}(t_{l*}^n - < \mathbf{w}_{l*},\ y_5(\mathbf{p}^n,\theta) >)$$

και

$$S_{L_1}(x) = \begin{cases} x^2, & \text{για } |x| < 1 \\ |x| - \frac{1}{2}, & \text{διαφορετικά,} \end{cases}$$

η οποία, όπως μπορούμε να δούμε και στο Σχήμα 7 είναι πολύ λιγότερο ευαίσθητη σε ακραίες τιμές (outliers) σε σχέση με την στάθμη l_2 που χρησιμοποιείται, όπως είδαμε, στο R-CNN. Η τιμή του λ τίθεται μονάδα, θ οι παράμετροι του δικτύου και:

$$W = \begin{bmatrix} \mathbf{w}_{1x}^t & \mathbf{w}_{1y}^t & \mathbf{w}_{1w}^t & \mathbf{w}_{1h}^t \\ \mathbf{w}_{2x}^t & \mathbf{w}_{2y}^t & \mathbf{w}_{2w}^t & \mathbf{w}_{2h}^t \\ \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{w}_{Lx}^t & \mathbf{w}_{Ly}^t & \mathbf{w}_{Lw}^t & \mathbf{w}_{Lh}^t \end{bmatrix}$$

ένα μητρώο που περιέχει τις παραμέτρους των παλίνδρομων.

Faster R-CNN

Προκειμένου να αρθούν οι περιορισμοί που επιβάλει ο αλγόριθμος επιλεκτικής αναζήτησης στους προηγούμενους ανιχνευτές, προτάθηκε μια νέα μορφή του Fast R-CNN.

Ο Faster R-CNN [4] βασίζει την λειτουργία του εντοπισμού στον Fast R-CNN συνδυάζοντας ένα συνελικτικό δικτύο το οποίο εντοπίζει υποψήφιες ROIs, όπως φαίνεται στο Σχήμα 8. Το τελευταίο δίκτυο ονομάζεται Region Proposal Network (RPN) και αντικαθιστά τον αλγόριθμο επιλεκτικής αναζήτησης. Βασικά το δίκτυο RPN δέχεται ως είσοδο τον χάρτη χαρακτηριστικών του Fast R-CNN και εντοπίζει σε αυτούς τις υποψήφιες περιοχές. Αυτό επιτυγχάνεται με την χρήση ενός κυλιόμενου πυρήνα συνέλιξης μεγέθους 3×3. Συγκεκριμένα, κάθε σημείο του χάρτη

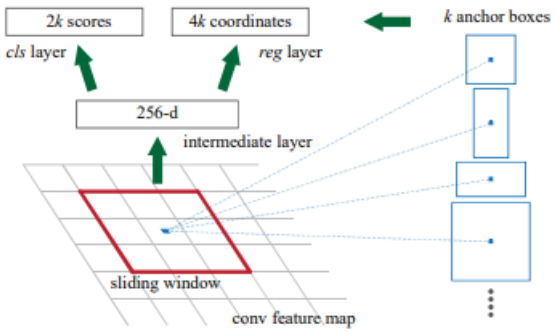
χαρακτηριστικών ορίζεται ως ένα anchor point και για κάθε τέτοιο σημείο παράγονται μέσω του κυλιόμενου πυρήνα τα anchor boxes με κέντρο το anchor point. Τα παραγόμενα anchor boxes (Σχήμα 8) έχουν:

- τρεις (3) προκαθορισμένες κλίμακες γύρω απο το σημείο και
- τρεις (3) προκαθορισμένους λόγους προοπτικής, δηλαδή:

$$\alpha_r = \left\{\frac{1}{2},\ 1, 2\right\}.$$

Στη συνέχεια έχοντας παράξει αυτά τα πλαίσια, το δίκτυο απορρίπτει ένα υποσύνολό τους βάσει των ακόλουθων απλών κανόνων:

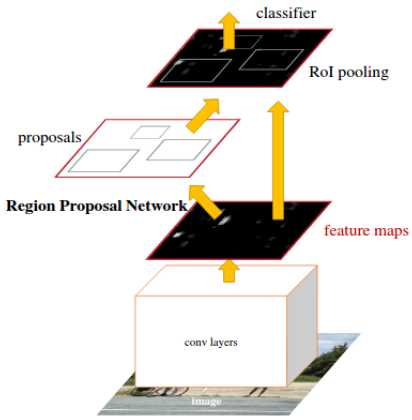
- απορρίπτονται όσα ξεπερνούν τις αρχικές διαστάσεις του χάρτη χαρακτηριστικών και
- στα υπόλοιπα ακολουθεί η διαδικασία του Non-Maximum Suppresion (NMS) (βλέπε Παράρτημα).



Σχήμα 8: Οπτική απεικόνιση των anchor boxes [4].

Στη συνέχεια, τα πλαίσια τα οποία έχουν επικρατήσει οδηγούνται στα επόμενα επίπεδα του δικτύου όπου πραγματοποιείται μια διακασία δυαδικής κατηγοριοποίησης (object, not object) και παλινδρόμησης τον πλαισίων για βελτίωση του τοπικού προσδιορισμού. Τέλος μέσω αυτής της διαδκασίας το δίκτυο τροφοδοτεί τον Fast R-CNN με τις υποψήφιες περιοχές.

Είναι φανερό ότι ο ανιχνευτής Faster R-CNN συντίθεται από δύο νευρωνικά δίκτυα:



Σχήμα 9: Απεικόνιση του Faster R-CNN [4].

1. το Fast R-CNN (τις παραμέτρους του οποίου θα συμβολίζουμε με το θ) και

2. το RPN (τις παραμέτρους του οποίου θα συμβολίζουμε με το ϕ).

Αυτό έχει ως αποτέλεσμα η εκπαίδευση του Faster R-CNN να μπορεί να γίνει με παραπάνω από έναν τρόπους. Συγκεκριμένα:

- εναλλάξ εκπαίδευση του RPN και του Fast R-CNN, αρχίζοντας με την εκπαίδευση του RPN

- από κοινού εκπαίδευση των δύο δικτύων.

Η συνάρτηση κόστους του RPN είναι παρόμοιας λογικής με την συνάρτηση κόστους του Fast R-CNN με την διαφορά ότι ο κατηγοριοποιητής είναι δυαδικός (υπάρχει ή δεν υπάρχει αντικείμενο).

Τα anchors τα οποία έχουν:

- 1. το μεγαλύτερο IoU[3] με κάποιο πραγματικό πλαίσιο,
  2. ή το IoU των οποίων υπερβαίνει την τιμή 0.7 με κάποιο πραγματικό πλαίσιο

  κατηγοριοποιούνται με πιθανότητα $\mathbf{c}^{\star}_{\theta i} = [1\ 0]^t$ ως **θετικά** παραδείγματα,

[3]Είναι προφανές ότι το IoU εξαρτάται από τις παράμετρες του RPN.

- τα anchors για τα οποία το IoU είναι μικρότερο του 0.3 κατηγοριοποιούνται με πιθανότητα $\mathbf{c}^{\star}_{\theta j} = [0\ 1]^t$ ως **αρνητικά**, και τέλος

- όσα anchors δεν πληρούν τις παραπάνω προυποθέσεις, δηλαδή $0.3 \leq \text{IoU} \leq 0.7$, δεν λαμβάνουν μέρος στην εκπαίδευση.

Η συνάρτηση κόστους του RPN δικτύου, ορίζεται ως ακολούθως:

$$\mathcal{L}_{RPN_\theta}(\mathbf{w}, \phi) = \mathcal{L}_{CLASS_\theta}(\phi) + \lambda \sum_{*\in x,y,w,h} \mathcal{L}_{LOC_\theta}(\mathbf{w}_*, \phi)$$

όπου $\mathbf{w} = [\mathbf{w}_x^t\ \mathbf{w}_y^t\ \mathbf{w}_w^t\ \mathbf{w}_h^t]^t$ και:

$$\begin{aligned}
\mathcal{L}_{CLASS_\theta}(\phi) &= \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \log(< \mathbf{c}_{\theta i}(\phi), \mathbf{c}^*_{\theta i} >) \\
&\quad + \frac{1}{N_{neg}} \sum_{j=1}^{N_{neg}} \log(< \mathbf{c}_{\theta j}(\phi), \mathbf{c}^*_{\theta j} >) \\
\mathcal{L}_{LOC_\theta}(\mathbf{w}_*, \phi) &= \sum_{i=1}^{N_{pos}} S_{L_1}(t^i_{\theta *}(\mathbf{w}_*, \phi) - \hat{t}^i_{\theta *}(\phi))
\end{aligned}$$

όπου $N_{pos},\ N_{neg}$ το συνολικό πλήθος των θετικών και των αρνητικών παραδειγμάτων αντίστοιχα, $\mathbf{c}^*_{\theta i}$ η πραγματική ετικέτα της κλάσης ($[1\ 0]^t$ ή $[0\ 1]^t$), $< ., . >$ ο τελεστής εσωτερικού γινομένου[4] και τα στοιχεία $\hat{t}_{\theta *}(\mathbf{w}_*, \phi),\ t_*(\phi)$ ορίζονται απο τις ακόλουθες σχέσεις:

$$\begin{aligned}
\hat{t}_{\theta x}(\mathbf{w}_x, \phi) &= \frac{\hat{g}_{\theta x}(\mathbf{w}_x) - x_{anchor}(\phi)}{w_{anchor}(\phi)} \\
\hat{t}_{\theta y}(\mathbf{w}_y, \phi) &= \frac{\hat{g}_{\theta y}(\mathbf{w}_y) - y_{anchor}(\phi)}{h_{anchor}(\phi)} \\
\hat{t}_{\theta w}(\mathbf{w}_w, \phi) &= log\Big(\frac{\hat{g}_{\theta w}(\mathbf{w}_w)}{w_{anchor}(\phi)}\Big) \\
\hat{t}_{\theta h}(\mathbf{w}_h, \phi) &= log\Big(\frac{\hat{g}_{\theta h}(\mathbf{w}_h)}{h_{anchor}}\Big)
\end{aligned}$$

[4]Παρατηρήστε ότι το όρισμα του $\log(.)$ είναι πάντα θετικό.

$$
\begin{aligned}
t_x(\phi) &= \frac{g_x - x_{anchor}(\phi)}{w_{anchor}(\phi)} \\
t_y(\phi) &= \frac{g_y - y_{anchor}(\phi)}{h_{anchor}(\phi)} \\
t_w(\phi) &= log\Big(\frac{g_w}{w_{anchor}(\phi)}\Big) \\
t_h(\phi) &= log\Big(\frac{g_h}{h_{anchor}(\phi)}\Big)
\end{aligned}
$$

Στην περίπτωση που η εκπαίδευση των δύο παραπάνω δικτύων γίνεται από κοινού, όπως έχει γίνει στο δίκτυο που θα χρησιμοποιήσετε στο πλαίσιο της άσκησης, η συνάρτηση κόστους είναι ένας γραμμικός συνδιασμός των συναρτήσεων κόστους $\mathcal{L}_{FAST}(W, \theta, \phi)$ και $\mathcal{L}_{RPN}(\mathbf{w}, \theta, \phi)$, δηλαδή:

$$\mathcal{L}(W, \mathbf{w}, \theta, \phi) = \mathcal{L}_{FAST}(W, \theta, \phi) + \lambda \mathcal{L}_{RPN}(\mathbf{w}, \theta, \phi)$$

με $\lambda \in \mathbb{R}$ και τις συναρτήσεις κόστους $\mathcal{L}_{FAST}(W, \theta, \phi)$ και $\mathcal{L}_{RPN}(\mathbf{w}, \theta, \phi)$ ορισμένες κατάλληλα.

ΔΙΑΔΙΚΑΣΙΑ

Στο πλαίσιο της παρούσας άσκησης επιλέξτε έναν ανιχνευτή της αρεσκείας σας από τα προεκπαιδευμένα μοντέλα ($FasterR-CNN, SSD$) που διαθέτει η βιβλιοθήκη $torchvision$ της Pytorch. Επιπρόσθετα σας δίνετε ένα σύνολο εικόνων δοκιμής καθώς και τα αντίστοιχα annotations τα οποία περιέχουν πληροφορίες όπως: τα χαρακτηριστικά κλειδιά κάθε εικόνας (ids),τις ground truth ετικέτες των κατηγοριών καθώς και τις τέσσερις συντεταγμένες των ground truth πλαισίων ($bbox$). Τα δεδομένα αυτά έχουν προέλθει από το σύνολο δεδομένων ($COCO2017$), στο οποίο έχουν προεκπαιδευτεί τα μοντέλα που θα χρησιμοποιήσετε. Μαζί με το σύνολο δεδομένων που θα χρησιμοποιήσετε σας δίνετε και ένα αρχείο .ipnyb για δικιά σας διευκόλυνση, το οποίο μπορείτε να αξιοποιήσετε στο περιβάλλον https://colab.research.google.com καθώς και δύο αρχεία .pth που περιέχουν τις αποθηκεύμενες παραμέτρους από την προεκπαίδευση των δικτύων.

Τα παραπάνω αρχεία μπορείτε να τα βρείτε εδώ.

Στο πλαίσιο της άσκησης καλείστε να :

1. Υπολογίστε την ακρίβεια πρόβλεψης $(Precision)$ του ανιχνευτή για το σύνολο εικόνων δοκιμής.

2. Υπολογίστε τον λόγο ανάκλησης $(Recall)$.

3. Υπολογίστε την μετρική F_1 ανάμεσα στην ακρίβεια πρόβλεψης και στον λόγο ανάκλησης.

4. Υπολογίστε την μέση ακρίβεια πρόβλεψης ($MeanAveragePrecision$).

5. Να σχεδιάσετε την χαρακτηριστική καμπύλη ανάμεσα στην ακρίβεια πρόβλεψης και στον λόγο ανάκλησης που υπολογίσατε στα προηγούμενα ερωτήματα.

6. Επαναλάβετε την διαδικασία 1-5 για τον αλλό ανιχνευτή από το ζευγάρι που αναφέρεται παραπάνω.

7. Καταγράψτε τον μέσο χρόνο πρόβλεψης που απαιτείται ανα εικόνα για κάθε ανιχνευτή και συγκρίνετε τα αποτελέσματα τους.

Παράρτημα

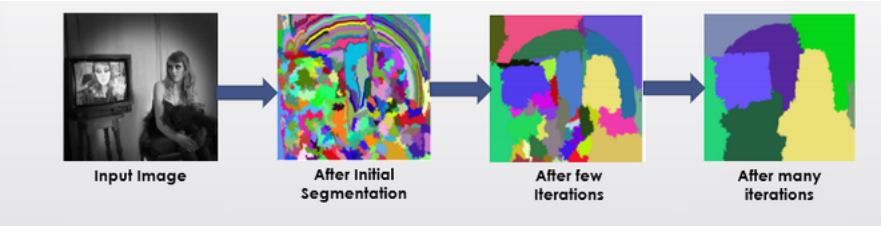
1. Αλγόριθμος Επιλεκτικής Αναζήτησης: Ο συγκεκριμένος αλγόριθμος χρησιμοποιείται ευρέως στο πρόβλημα του εντοπισμού αντικειμένων παράγωντας υποψήφιες περιοχές που ενδέχεται να περιέχουν αντικείμενα [5]. Ο αλγόριθμος συνοψίζεται στα ακόλουθα βήματα:

    - **Υπερκατάτμηση εικόνας**: Ο αλγόριθμος δέχεται μια εικόνα ως είσοδο και εφαρμόζει μια τεχνική κατάτμησης που βασίζεται στην θεωρία γράφων [6].

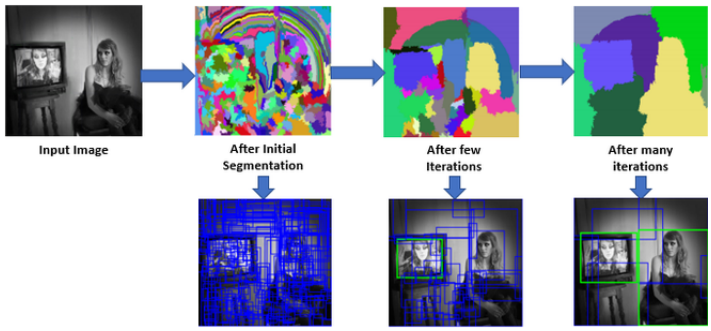
Σχήμα 10: Δεξία διακρίνεται η αρχική εικόνα μετά την κατάτμησης της [5].

    - **Συνένωση περιοχών**: Σε αυτό το βήμα εφαρμόζεται ένας αναδρομικός αλγόριθμος ο οποίος σε κάθε επανάληψη του υπολογίζει κάποιες μετρικές ομοιότητας μεταξύ των περιοχών. Όσες περιοχές είναι όμοιες ενώνονται και σχηματίζουν μεγαλύτερες. Οι μετρικές ομοιότητας βασίζονται στο χρώμα, την υφή, το μεγέθος και το fill της εικόνας.
    - **Εξαγωγή υποψήφιων περιοχών**: Πραγματοποιείται παράλληλα με το προηγούμενο βήμα. Για κάθε περιοχή που έχει προκύψει είτε από την αρχική κατάτμηση είτε από συνένωση περιοχών ορίζεται και ένα πλαίσιο γύρω από αυτήν και κατατάσσεται ως υποψήφια περιοχή.

2. Non-Maximum Supression: Αποτελεί αναδρομικό αλγόριθμο. Χρησιμοποιείται για να μειώνει τον αριθμό των προτεινόμενων πλαισίων



Σχήμα 11: Αποτέλεσμα κατάτμησης για διαφορετικό πλήθος επαναλήψεων [5].

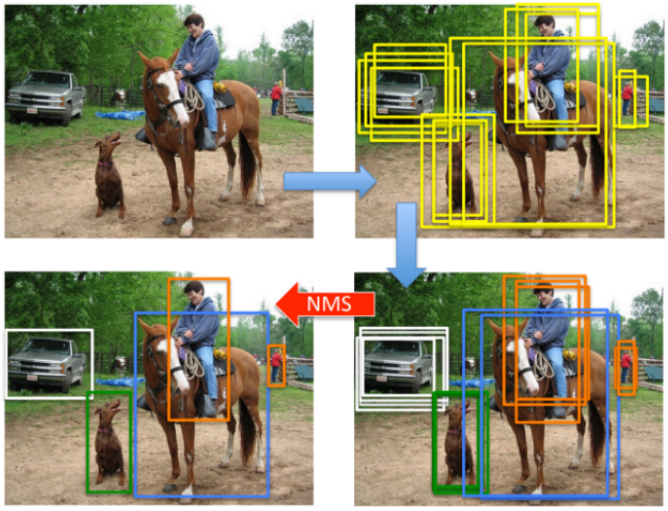


Σχήμα 12: Εξαγωγή υποψήφιων περιοχών για διαφορετικό πλήθος επαναλήψεων [5].

αναγνώρισης.Για κάθε κλάση ξεχωριστά επιλέγει το πλαίσιο που έχει την μεγαλύτερη πιθανότητα κατηγοριοποίησης και εξετάζει την μετρική IoU με τα υπόλοιπα πλαίσια της κλάσης. Η IoU μετρική ορίζεται ως εξής:

$$IoU = \frac{A \cap B}{A \cup B},$$

όπου A, B τα πλαίσια. Όταν το IoU μεταξύ δύο πλαισίων ξεπερνά κάποιο ορισμένο κατώφλι αυτό σημαίνει ότι επικαλύπτονται σε μεγάλο πόσοστο και ο αλγόριθμος απορρίπτει αυτό με την μικρότερη πιθανότητα. Στη συνέχεια συνεχίζει με το πλαίσιο που έχει την δεύτερη μεγαλύτερη πιθανότητα κοκ.



Σχήμα 13: Ο αλγόριθμος NMS για διαφορετικό πλήθος επαναλήψεων [2].

Βιβλιογραφία

[1] Ross B Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. corr abs/1311.2524 (2013). *arXiv preprint arXiv:1311.2524*, 2013.

[2] Gaurav Sinha. Deep learning method for object detection: R-cnn explained, 2020.

[3] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[5] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[6] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.