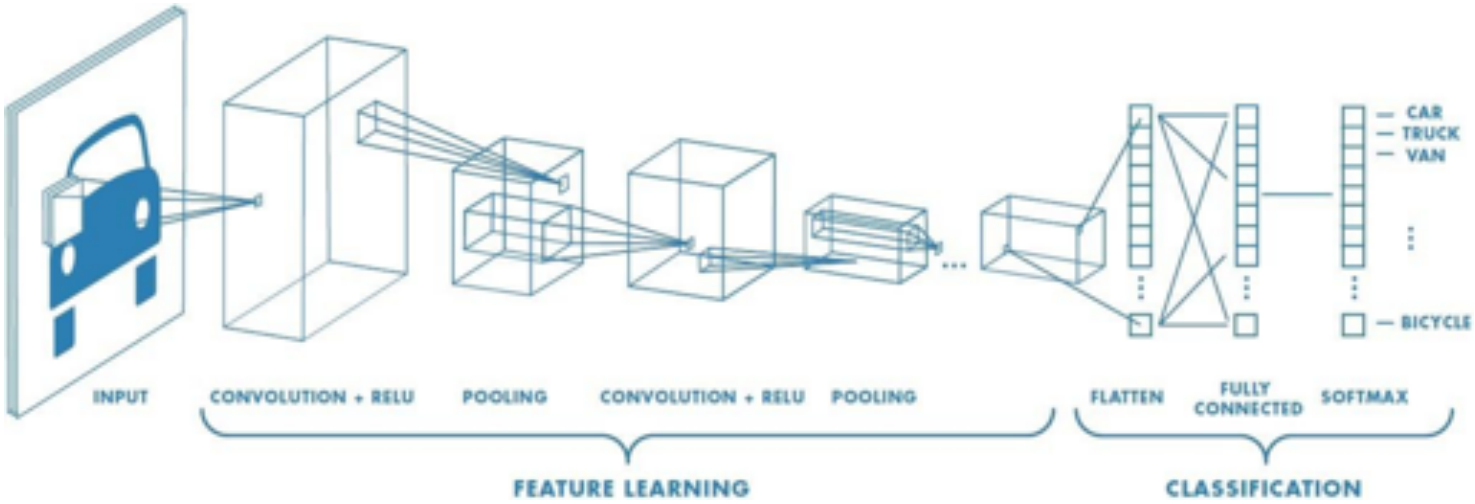
AI|LS
ARTIFICIAL INTELLIGENCE
& LEARNING SYSTEMS LABORATORY

Convolutional Neural Networks

Νευρωνικά Δίκτυα και Βαθιά Μάθηση

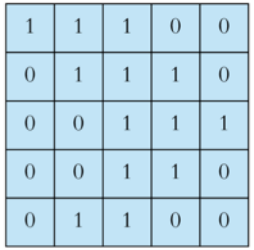
Convolutional Neural Network



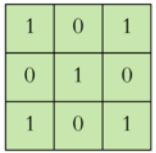
Understanding of Convolutional Neural Network (CNN) — Deep Learning, Prabhu in Towards Data Science

Convolutional Neural Network

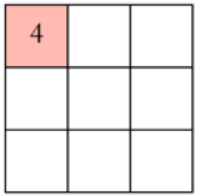
Convolution Layer



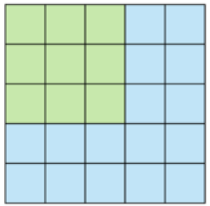
Input



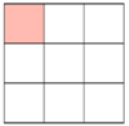
Filter / Kernel



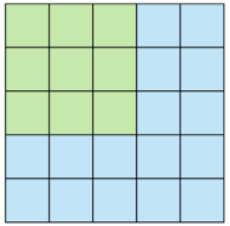
Stride



Stride 1



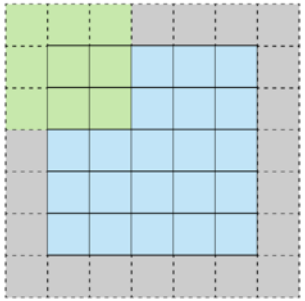
Feature Map



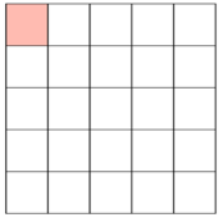
Stride 2

Feature Map

Padding



Stride 1 with Padding



Feature Map

Convolutional Neural Network

Οπότε, αν στην είσοδο του Convolution Layer έχω τένσορες διάστασης $W_1 x H_1 x D_1$ απαιτείται ο ορισμός 4 υπερπαραμέτρων:

1. το πλήθος των φίλτρων (K)
2. το μέγεθος των φίλτρων (F)
3. το stride (S)
4. το padding (P)

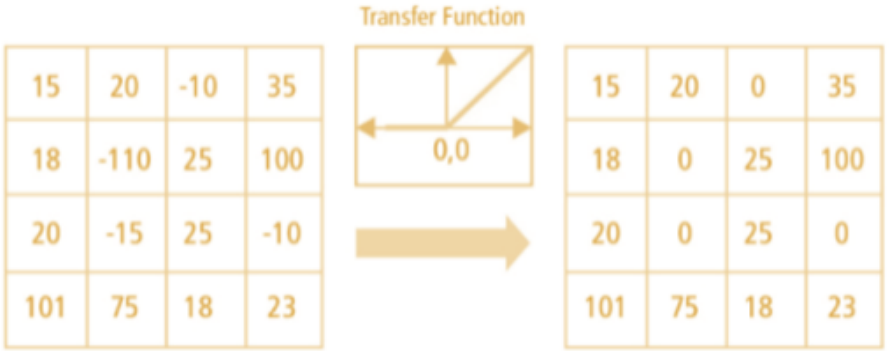
Στην έξοδο του Convolution Layer θα πάρω τένσορες μεγέθους $W_2 x H_2 x D_2$ όπου:

1. $W_2 = \frac{(W_1 - F + 2*P)}{S} + 1$
2. $H_2 = \frac{(H_1 - F + 2*P)}{S} + 1$
3. $D_2 = K$

Convolutional Neural Network

ReLU: Rectified Linear Unit for a non-linear operation

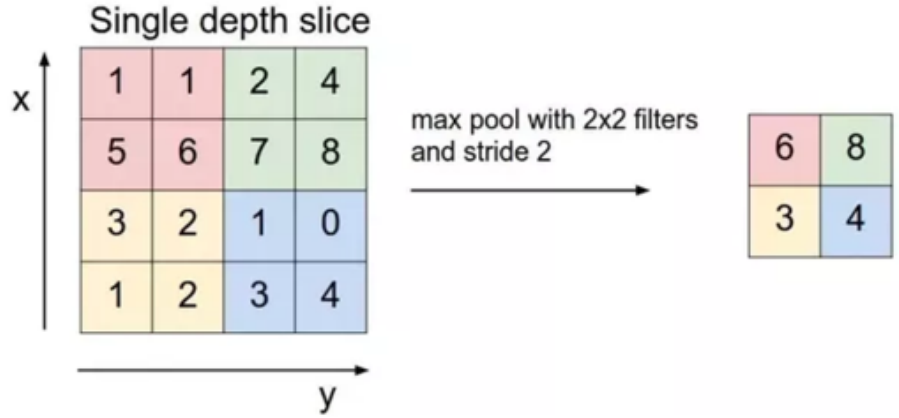
ƒ(x) = max(0,x).



Convolutional Neural Network

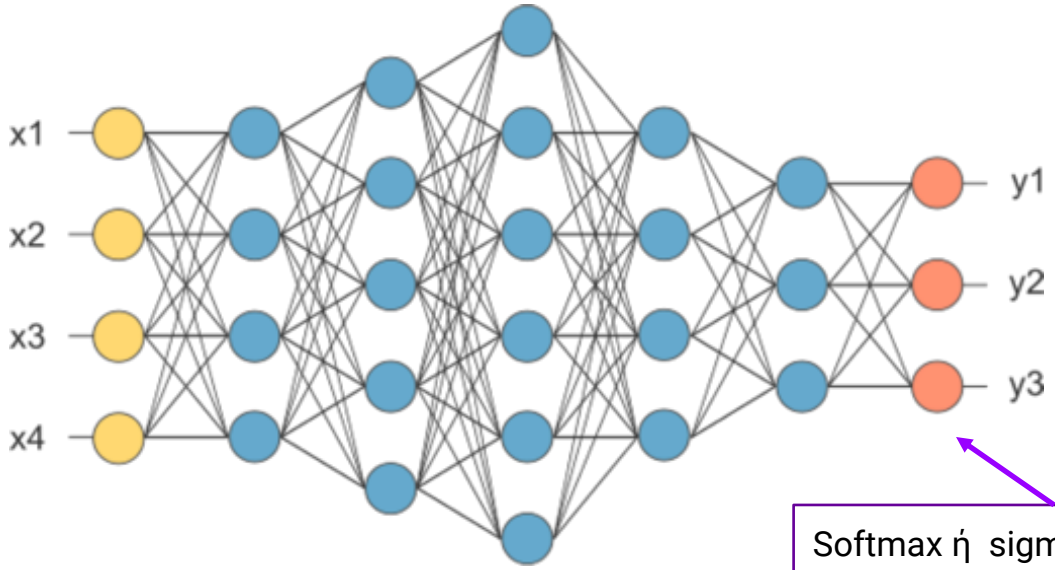
Pooling Layer

- Max Pooling
- Average Pooling
- Sum Pooling



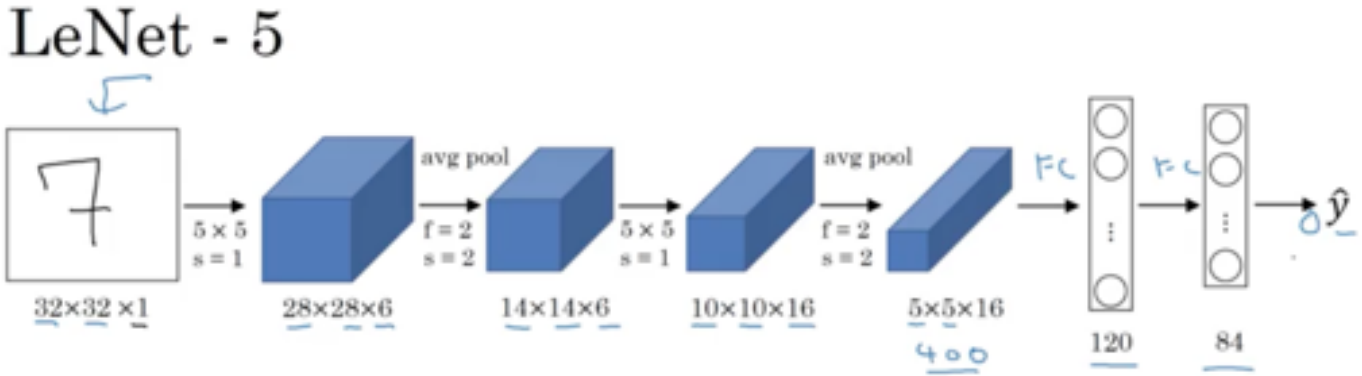
Convolutional Neural Network

Fully Connected Layer



Softmax ή sigmoid για κατηγοριοποίηση της εξόδου (π.χ. cat, dog, car, truck etc.)

CNN Architectures: LeNet



W x H → 32 x 32 (Width x Height)

F(w x h) → 5 x 5 (Filter)

S → 1 (Stride)

P → 0 (Pooling)

$$\left(\frac{W - Fw + 2P}{Sw}\right) + 1 => \left(\frac{32 - 5 + 0}{1}\right) + 1 => 27 + 1 => 28$$

$$\left(\frac{H - Fh + 2P}{Sh}\right) + 1 => \left(\frac{32 - 5 + 0}{1}\right) + 1 => 27 + 1 => 28$$

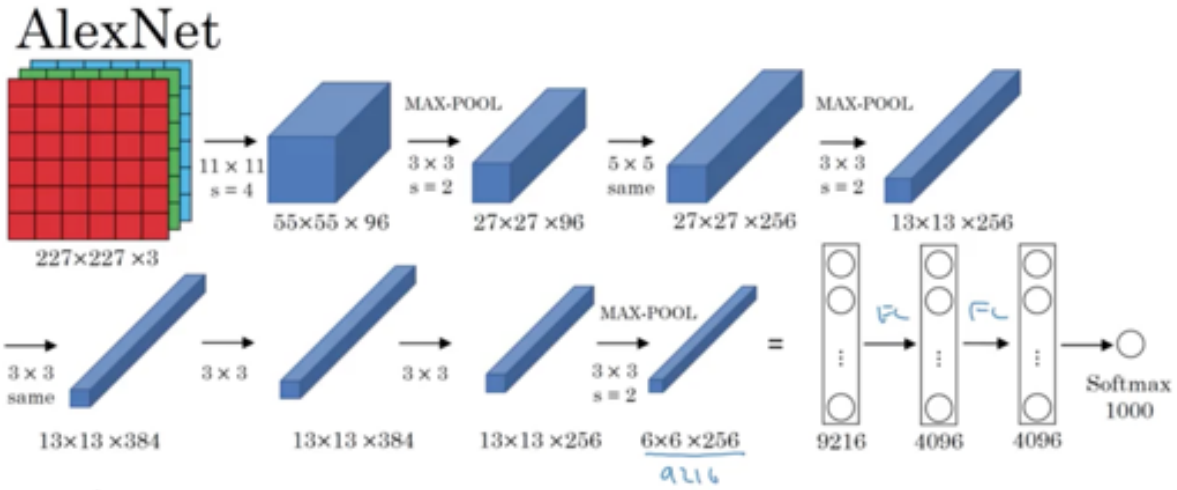
$$\textit{Output Volume} = 28 \text{ x } 28$$

LeCun et al., 1998. Gradient-based learning applied to document recognition]

Andrew Ng

CNN Architectures: AlexNet

- Activation function
  - ReLU (όχι Sigmoid ή Tanh)
    - 5 x ταχύτητα,
    - ίδια ακρίβεια
- OverFitting
  - Dropout
    - Διπλασιασμός χρόνου εκπαίδευσης
- Περισσότερα δεδομένα και μεγαλύτερο μοντέλο
  - 7 hidden layers, 650K units και 60M parameters.



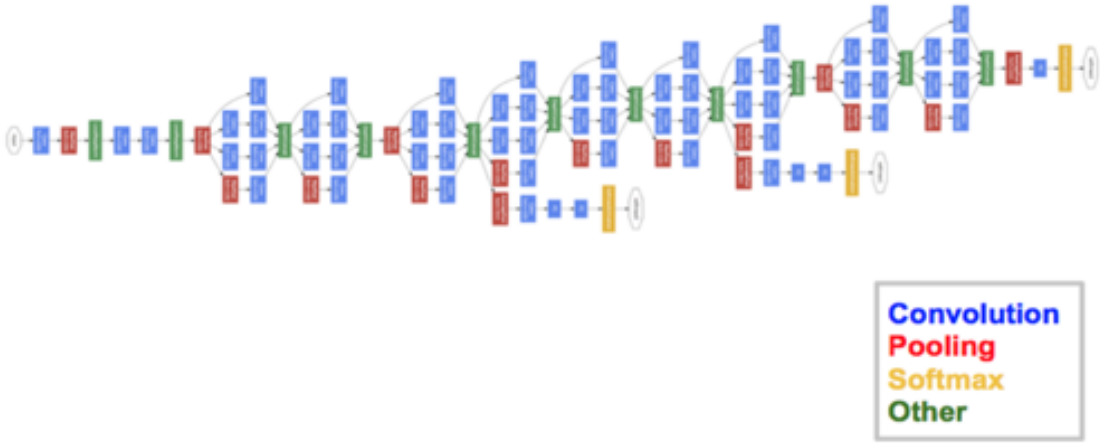
[Krizhevsky et al., 2012. ImageNet classification with deep convolutional neural networks]

Andrew Ng

CNN Architectures: Inception (GoogLeNet)

C. Szegedy *et al.*, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

- Διαδοχικά αλλά και παράλληλα CNN (error rate 6.7%)



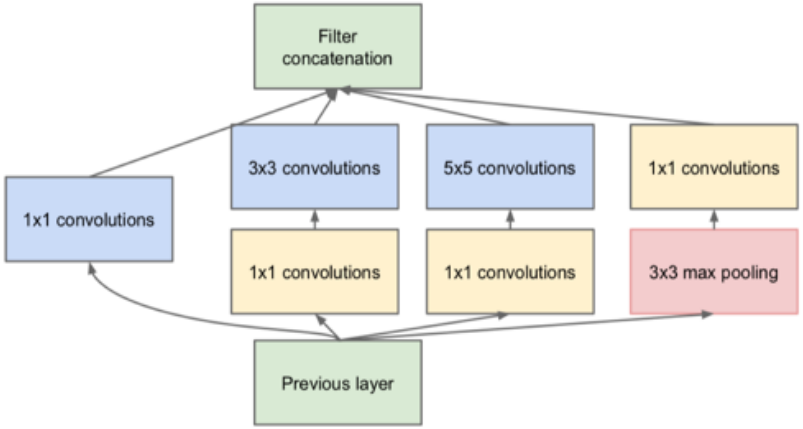
- Πολλαπλοί πυρήνες διαφορετικών μεγεθών εφαρμόζονται στο ίδιο επίπεδο με σκοπό τη ανίχνευση συγκεκριμένων χαρακτηριστικών περιοχής
  - Μεγάλοι πυρήνες → καθολικά χαρακτηριστικά που κατανέμονται σε μεγάλη περιοχή της εικόνας,
  - Μικροί πυρήνες → ανίχνευση συγκεκριμένων χαρακτηριστικών περιοχής που κατανέμονται σε ολόκληρο το πλαίσιο εικόνας.

CNN Architectures: Inception (GoogLeNet)

Μονάδα Inception :

Καταγράφει προεξέχοντα χαρακτηριστικά (salient features) σε διαφορετικά επίπεδα.

- ➔ 4 παράλληλες λειτουργίες
  - 1x1 conv layer, μείωση βάθους
  - 3x3 conv layer, Κατανεμημένα χαρακτηριστικά (distributed features)
  - 5x5 conv layer, Γενικά χαρακτηριστικά (global features)
  - max pooling, Χαμηλού επιπέδου χαρακτηριστικά (low level features)



Inception Module (source: original paper)

- ➔ Φίλτρο συνένωσης
  π.χ. εάν οι εικόνες στο σύνολο δεδομένων έχουν πολλά καθολικά χαρακτηριστικά και ελάχιστα χαρακτηριστικά χαμηλού επιπέδου, τότε το εκπαιδευμένο δίκτυο Inception θα έχει πολύ μικρά βάρη που αντιστοιχούν στον πυρήνα 3x3 σε σύγκριση με τον πυρήνα 5x5.

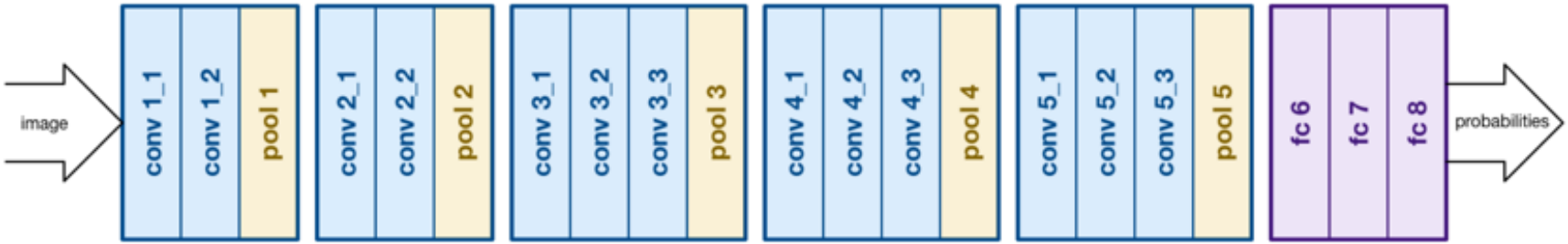
CNN Architectures : VGG

Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations (ICLR2015).

13 συνελικτικά και 3 πλήρως συνδεδεμένα επίπεδα, ReLU,
φίλτρα μικρότερου μεγέθους (2 × 2 και 3 × 3) από το AlexNet ,
138M παραμέτρους, 500MB.

Ανέδειξαν τη σημασία του βάθους του δικτύου ως σημαντικής παραμέτρου για την αποτελεσματικότητά του.

- ➔ Μείωση του αριθμού των παραμέτρων στα επίπεδα CONV
- ➔ Βελτίωση του χρόνου εκπαίδευσης
- ➔ Σχεδίασαν επίσης βαθύτερες παραλλαγές, VGG-16, VGG-19.



CNN Architectures : VGG-16

Η ιδέα πίσω από την ύπαρξη πυρήνων σταθερού μεγέθους είναι ότι όλοι οι conv πυρήνες μεταβλητού μεγέθους που χρησιμοποιούνται στο Alexnet (11x11, 5x5, 3x3) μπορούν να αναπαραχθούν χρησιμοποιώντας πολλαπλούς πυρήνες 3x3 ως δομικά στοιχεία.

π.χ. Έστω επίπεδο εισόδου μεγέθους 5x5x1

Περίπτωση 1: 1ο conv επίπεδο: ένας πυρήνας 5x5 και βήμα 1 → Έξοδος: χάρτης χαρακτηριστικών 1x1

Πλήθος μεταβλητών 5x5x1=25 ((m*n+1)*k, k: πλήθος πυρήνων)

Περίπτωση 2: 1ο conv επίπεδο: δύο πυρήνες 3x3 και βήμα 1 → Έξοδος: χάρτης χαρακτηριστικών 1x1.

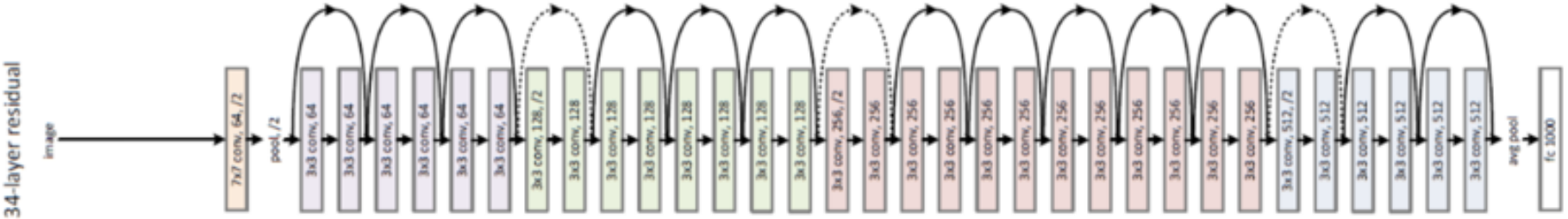
Πλήθος μεταβλητών 3x3x2=18 → Μείωση 28%

Αντίστοιχα αν αντί για χρήση πυρήνων 7x7 (11x11) εφαρμόσουμε 3 (5) 3x3 πυρήνες → μείωση αριθμού εκπαιδευόμενων μεταβλητών κατά 44,9% (62,8%)

★ Ταχύτερη εκμάθηση ★ Αποφυγή overfitting

CNN Architectures: ResNet (MSRA)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun Deep Residual Learning for Image Recognition, CVPR 2015



Resnet18

- 152 επίπεδα, 11M παράμετροι, πυρήνες, 3x3 (όπως το VGGNet), 2 pooling επίπεδα

Σύνδεση ταυτότητας (Identity connection) ανά δύο επιπέδων CONV, διάσταση εισόδου ίδια με της εξόδου

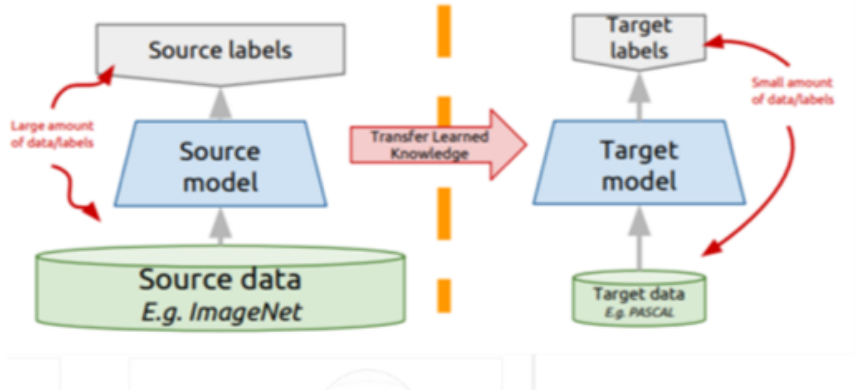
Σύνδεση προβολής (Projection connection) όπου οι διαστάσεις εισόδου διαφέρουν με της εξόδου.

Υπάρχουν πολλές εκδόσεις αρχιτεκτονικών ResNetXX όπου το «XX» υποδηλώνει τον αριθμό των επιπέδων (ResNet50, ResNet101)

Transfer Learning για Deep Learning

Ορισμός

Με δεδομένη μια εργασία Transfer Learning που ορίζεται από $\langle D_s, T_s , D_t, T_t, f_T(\cdot)\rangle$, η "μεταφορά μάθησης» στοχεύει στη μάθηση της μη γραμμικής συνάρτησης f_T που αντικατοπτρίζει ένα βαθύ νευρωνικό δίκτυο.



Tan C., Sun F., Kong T., Zhang W., Yang C., Liu C. (2018) A Survey on Deep Transfer Learning. Artificial Neural Networks and Machine Learning – ICANN 2018

Στρατηγικές Deep Transfer Learning

➔ Προεκπαιδευμένα μοντέλα για εξαγωγή χαρακτηριστικών

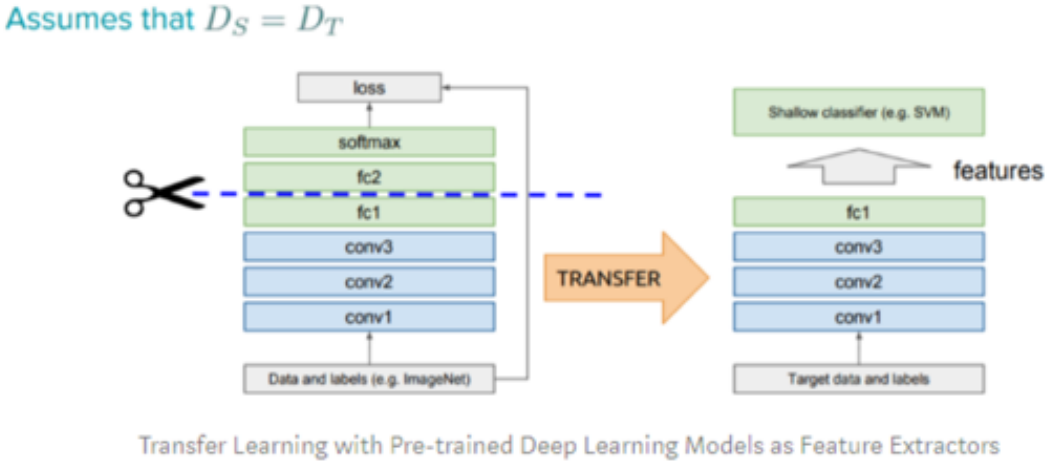
Off-the-shelf Pre-trained Models as fixed Feature Extractors

➔ Ακριβής προσαρμογή προεκπαιδευμένων μοντέλων

Fine Tuning Off-the-shelf Pre-trained Models

Προεκπαιδευμένα μοντέλα για εξαγωγή χαρακτηριστικών

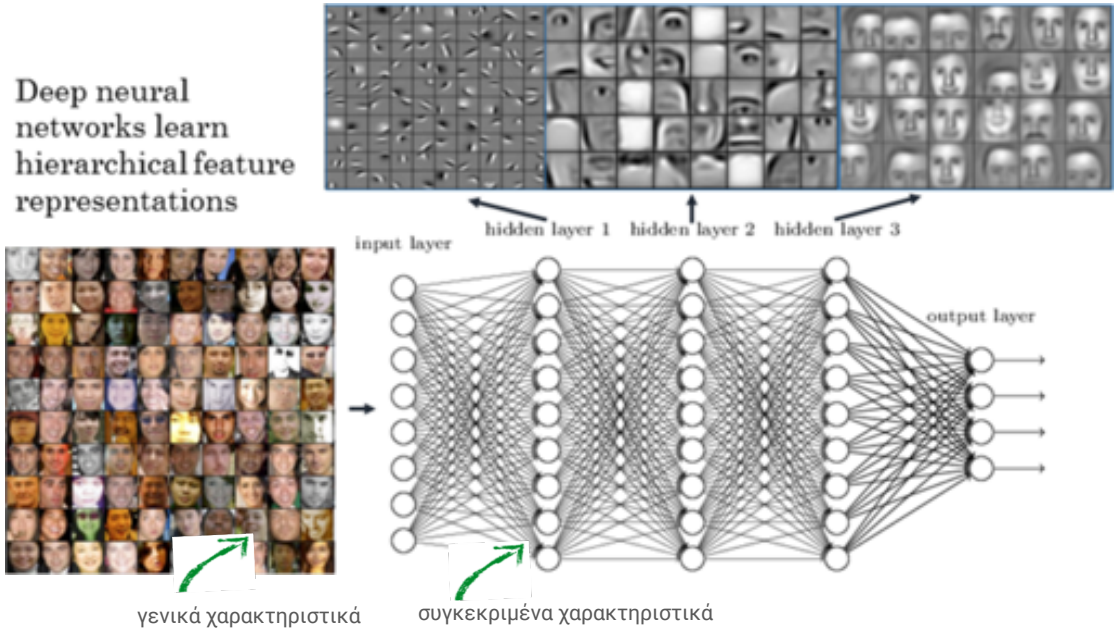
- Η έξοδος μετά από κάποιο επίπεδο ενός δικτύου βαθιάς μάθησης, που εκπαιδεύτηκε σε διαφορετική εργασία ($T_s \neq T_t$), χρησιμοποιείται ως γενικευμένος ανιχνευτής χαρακτηριστικών.
- Εκπαίδευση νέου μοντέλου (π.χ. SVM) με μεταφορά αυτών των χαρακτηριστικών.



Transfer Learning with Pre-trained Deep Learning Models as Feature Extractors

Ακριβής προσαρμογή προεκπαιδευμένων μοντέλων

Δεν αντικαθιστούμε απλώς το τελικό επίπεδο (για ταξινόμηση / παλινδρόμηση), αλλά επανεκπαιδεύουμε επιλεκτικά ορισμένα από τα προηγούμενα επίπεδα.



Πρακτικές συμβουλές

➔ Περιορισμοί από προκατασκευασμένα μοντέλα.

- Η χρήση ενός προκαθορισμένου δίκτυου, ενδέχεται να είναι δεσμευτική ως προς την αρχιτεκτονική που μπορείτε να χρησιμοποιήσετε για το νέο σύνολο δεδομένων.
    - π.χ. δεν μπορείτε να αφαιρέσετε αυθαίρετα Conv επίπεδα από το προκαθορισμένο δίκτυο.
- Συνήθως χρησιμοποιούμε μικρότερο learning rate για τα ρυθμισμένα βάρη ConvNet, σε σύγκριση με τα (τυχαία αρχικοποιημένα) βάρη που θα χρησιμοποιούσαμε για το νέο γραμμικό ταξινομητή που υπολογίζει τα βάρη ταξινόμησης του νέου συνόλου δεδομένων μας.
    - Αυτό συμβαίνει επειδή περιμένουμε ότι τα ρυθμισμένα βάρη ConvNet είναι σχετικά καλά, επομένως δεν θέλουμε να τα παραμορφώσουμε πολύ γρήγορα και πάρα πολύ.

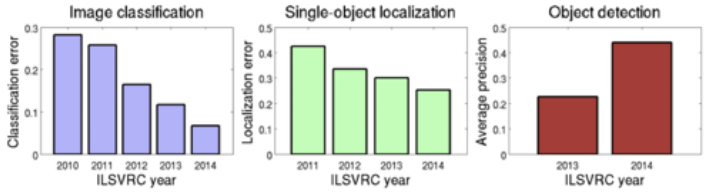
ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

- ILSVRC: ετήσιος διαγωνισμός που χρησιμοποιεί υποσύνολα από το σύνολο δεδομένων ImageNet για ανάπτυξη και συγκριτική αξιολόγηση αλγορίθμων τελευταίας τεχνολογίας.

- ImageNet: πολύ μεγάλη συλλογή χαρακτηρισμένων (Amazon Mechanical Turk Worker) φωτογραφιών για την ανάπτυξη αλγορίθμων όρασης υπολογιστή.

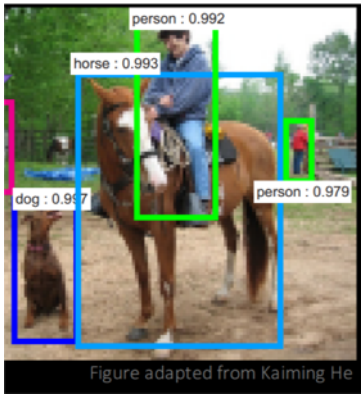
- Οι εργασίες του ILSVRC οδήγησαν σε σημαντικές αρχιτεκτονικές μοντέλων και τεχνικές σύνδεσης της όρασης υπολογιστή και της βαθιάς μάθησης

IMAGENET



Κατηγορίες

Image classification
Πρόβλεψη των κατηγοριών των αντικειμένων που υπάρχουν στην εικόνα

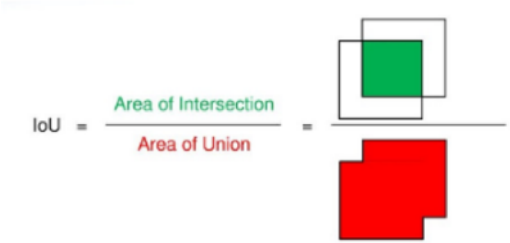


Single-object localization
Image classification + σχεδιασμός bounding box

Object detection
Image classification + σχεδιασμός bounding box γύρω από κάθε αντικείμενο.

Object Segmentation
Ανίχνευση όλων των αντικειμένων της εικόνας σε pixel level

Εισαγωγικές Έννοιες Ανίχνευσης Αντικειμένων



1. Μετρική της ομοιότητας μεταξύ δύο αντικειμένων

Σκοπός : σύγκριση και αξιολόγηση ομοιότητας αντικειμένων
(π.χ. Ανιχνευμένο αντικείμενο με το αληθινό αντικείμενο (ground truth))

Ευρέως χρησιμοποιούμενο μέγεθος: Intersection over Union - IoU

Το IoU ορίζεται ως το εμβαδόν της τομής των δύο πλαισίων προς το εμβαδόν της ένωσής τους.

2. Πλαίσιο Οριοθέτησης (Bounding Box)

- Ως πλαίσιο οριοθέτησης ενός αντικειμένου σε μία εικόνα ορίζεται το μικρότερο δυνατό ορθογώνιο τμήμα της εικόνας στο εσωτερικό του οποίου βρίσκεται ολόκληρο το αντικείμενο.
- Για την περιγραφή ενός πλαισίου οριοθέτησης είναι απαραίτητες 4 τιμές. π.χ.
  - οι συντεταγμένες της κάτω αριστερής και της πάνω δεξιάς γωνίας του
  - οι συντεταγμένες της πάνω αριστερής γωνίας, το πλάτος w και το ύψος h του πλαισίου
  - οι συντεταγμένες του κέντρου του πλαισίου, το πλάτος του w και το ύψος του h

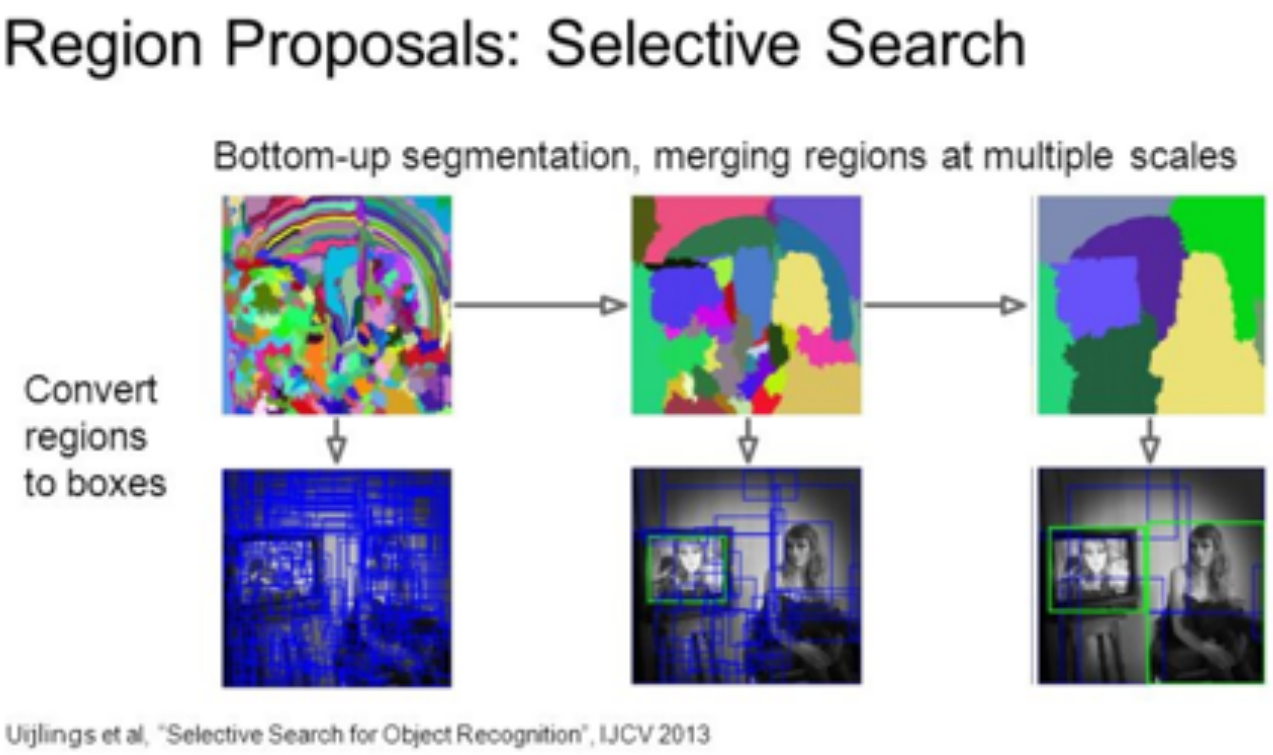
3. Περιοχή Ενδιαφέροντος (Region of Interest - ROI)

Ορίζεται μία ορθογώνια περιοχή της εικόνας εισόδου η οποία θεωρητικά είναι πιθανό να περιέχει ένα αντικείμενο.

Οι περιοχές αυτές μπορούν να υπολογιστούν:

- με χρήση κάποιου εξωτερικού αλγορίθμου όπως το Selective Search ή το Edge Box detection,
- με χρήση ενός Δικτύου Προτάσεων Περιοχών (Region Proposal Network - RPN).

Region-Based CNN: περιοχή ενδιαφέροντος (RoI)



4. Καταστολή μη μεγίστων (Non-Maximum Suppression)

Πρόβλημα: ύπαρξη πολλών προβλέψεων με μικρές διαφορές οι οποίες αντιστοιχούν στο ίδιο αντικείμενο.

Λύση: Καταστολή μη μεγίστων (Non-Maximum Supression - NMS)

- Άπληστος (greedy) αλγόριθμος που συγχωνεύει αυτά τα αλληλοεπικαλυπτόμενα πλαίσια οριοθέτησης:
    - Ταξινομεί όλα τα πλαίσια οριοθέτησης σε αύξουσα σειρά ως προς την πιθανότητά τους να αντιστοιχούν σε κάποιο αντικείμενο.
    - Επιλέγει το πλαίσιο οριοθέτησης με τη μεγαλύτερη πιθανότητα και, συγκρίνοντάς το με κάθε ένα από τα Bounding Box με μικρότερη πιθανότητα, απορρίπτει όσα έχουν επικάλυψη IoU μικρότερη από μία προκαθορισμένη τιμή.

      (Η τιμή αυτή αποτελεί μία από τις υπερπαραμέτρους του συστήματος) και επαναλαμβάνει τα βήματα όσες φορές είναι απαραίτητο.

Object Detection: απλή προσέγγιση με CNN

1. Χωρίζουμε την εικόνα σε περιοχές και τροφοδοτούμε την κάθε περιοχή ως ξεχωριστή εικόνα στο CNN το οποίο τις ταξινομεί σε διάφορες τάξεις.

2. Αφού χωρίσουμε κάθε περιοχή στην αντίστοιχη κλάση, μπορούμε να συνδυάσουμε όλες αυτές τις περιοχές για να πάρουμε την αρχική εικόνα με τα αντικείμενα που εντοπίστηκαν.

(-) : Τα αντικείμενα μπορεί να έχουν διαφορετικά aspect ratios, χωρικές θέσεις και να έχουν υποστεί διάφορους μετασχηματισμούς

(-) : Χρειάζεται πολύ μεγάλος αριθμός περιοχών, μεγάλη υπολογιστική ισχύς

Λύση: Region-based CNN

Πρόβλημα: Ανίχνευσης Αντικειμένων

Σύνθεση δύο διαφορετικών προβλημάτων:

- ένα πρόβλημα ταξινόμησης και
- ένα πρόβλημα παλινδρόμησης, γνωστό και ως bounding box regression.

Με δεδομένη μία εικόνα εισόδου, πρέπει να προβλεφθεί η τοποθεσία και η έκταση των αντικειμένων της εικόνας που ανήκουν σε ένα σύνολο προκαθορισμένων κλάσεων, και να αποδοθεί η σωστή κλάση στο κάθε αντικείμενων.

- Η τοποθεσία των αντικειμένων συνήθως εκφράζεται ως το ελάχιστο πλαίσιο οριοθέτησης που περικλείει εξ ολοκλήρου το αντικείμενο.

Κατηγορίες μοντέλων ανίχνευσης αντικειμένων

Χωρίζονται σε δύο κατηγορίες ως προς τη δομή τους:

- Τα μοντέλα ενός σταδίου (one-step models) χρησιμοποιούν :
  - ένα feed forward CNN για να προσδιορίσουν την τοποθεσία των αντικειμένων ενδιαφέροντος.
    - απλούστερα και ταχύτερα, αφού δεν παρέχουν region proposals
    - η απόδοσή τους είναι μειωμένη, κυρίως όταν απαιτείται και κατάτμηση της εικόνας

  π.χ. YOLO, Multibox, AttentionNet, G-CNN

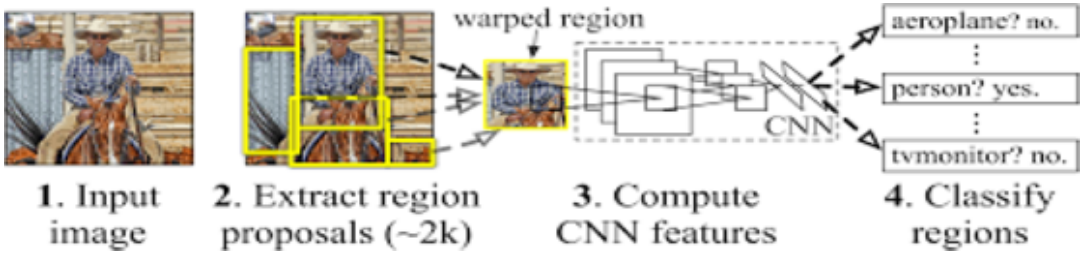
Κατηγορίες μοντέλων ανίχνευση αντικειμένων

Τα μοντέλα δύο σταδίων (two-step models ή region-based models), χρησιμοποιούν

1. Έναν αλγόριθμο (π.χ .Selective Search) ή ένα μοντέλο (π.χ. region-based CNN) που δέχεται ως είσοδο την εικόνα και προτείνει διαφορετικές πιθανές περιοχές ενδιαφέροντος
2. Έναν feature extractor π.χ. CNN ώστε να υπολογιστεί ο χάρτης χαρακτηριστικών κάθε περιοχής ενδιαφέροντος ο οποίος δίνεται σε ένα πλήρως συνδεδεμένο υπεύθυνο για την ταξινόμηση.

π.χ. R-CNN, Fast R-CNN, FPN, Faster R-CNN
- έχουν αρκετές διαφορές αλλά περίπου κοινή δομή

Region-Based Convolutional Neural Network: R-CNN



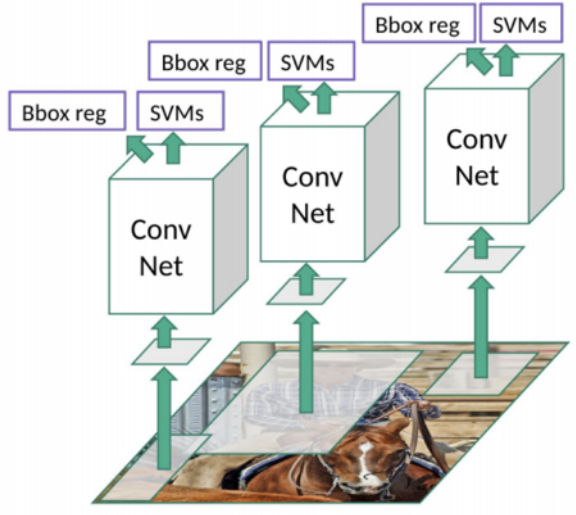
R-CNN Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5) Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik UC Berkeley

1. Χρησιμοποιεί τον αλγόριθμο Selective Search και παράγει 2000 προτάσεις περιοχών ανά εικόνα.
2. **Η κάθε περιοχή**, μετά από προσαρμογή του μεγέθους της, δίνεται ως είσοδος σε ένα προεκπαιδευμένο CNN
3. Η έξοδος από το ConvNet είναι ένα διάνυσμα 4096 χαρακτηριστικών
4. Εκπαιδεύουμε το τελευταίο επίπεδο του δικτύου **της κάθε περιοχής** ένα ταξινομητή με βάση τον αριθμό των κατηγοριών που πρέπει να εντοπιστούν

Region-Based Convolutional Neural Network: R-CNN

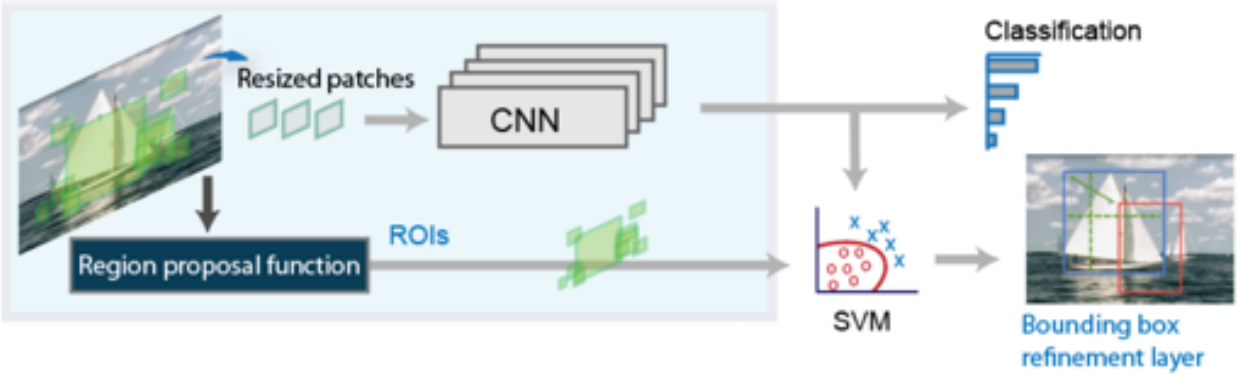
5. Αφού αποκτήσουμε τις περιοχές, εκπαιδεύουμε ένα δυαδικό SVM ανά περιοχή για να ταξινομήσουμε αντικείμενα και φόντο.

6. Τέλος, εκπαιδεύουμε ένα μοντέλο γραμμικής παλινδρόμησης για τη δημιουργία αυστηρότερων bounding boxes για κάθε αναγνωρισμένο αντικείμενο στην εικόνα.



R-CNN Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5)
Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik UC Berkeley

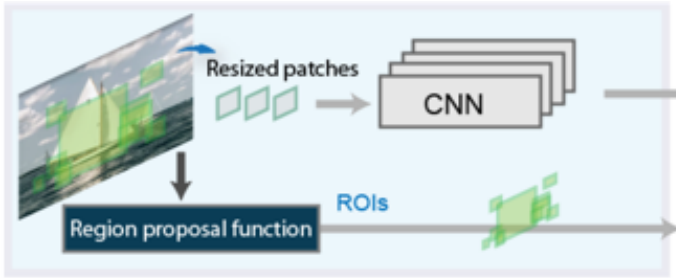
Region-Based Convolutional Neural Network: R-CNN

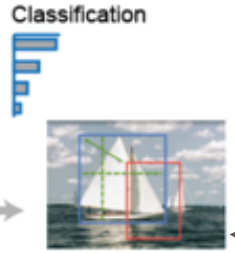


Μειονεκτήματα

- Οι 2000 προτάσεις περιοχών ανά εικόνα → πολύ μεγάλο χρόνο εκπαίδευσης
- Ο χρόνος του testing είναι απαγορευτικά μεγάλος → μη χρήση του μοντέλου για εφαρμογές πραγματικού χρόνου.
- Προκαθορισμένη συμπεριφορά του αλγορίθμου Selective Search → η αναγνώριση δε βελτιώνεται μέσω εκπαίδευσης.

Region-Based Convolutional Neural Network: R-CNN



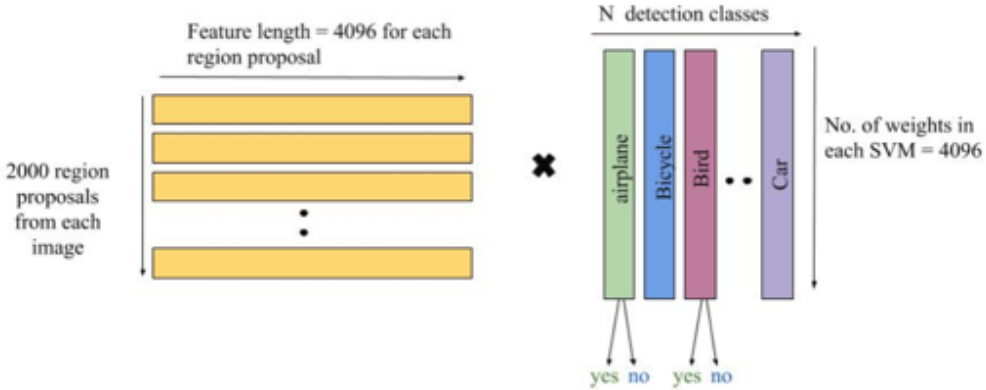


$$P^i = (P^i_x, P^i_y, P^i_w, P^i_h)$$
$$G = (G_x, G_y, G_w, G_h)$$

(1)

$$t_x = (G_x - P_x)/P_w$$
$$t_y = (G_y - P_y)/P_h$$
$$t_w = \log(G_w/P_w)$$
$$t_h = \log(G_h/P_h).$$

(2)



$$\hat{G}_x = P_w d_x(P) + P_x$$
$$\hat{G}_y = P_h d_y(P) + P_y$$
$$\hat{G}_w = P_w \exp(d_w(P))$$
$$\hat{G}_h = P_h \exp(d_h(P)).$$

(3)

$$d_\star(P) = \mathbf{w}_\star^{\mathrm{T}} \phi_5(P)$$

$$\mathbf{w}_\star = \underset{\hat{\mathbf{w}}_\star}{\operatorname{argmin}} \sum_i^N (t^i_\star - \hat{\mathbf{w}}_\star^{\mathrm{T}} \phi_5(P^i))^2 + \lambda \|\hat{\mathbf{w}}_\star\|^2$$

(4)

Bounding Box (x,y,w,h)

Learn a target transformation

P: Predicted
G: Target

Learn ground-truth transformations

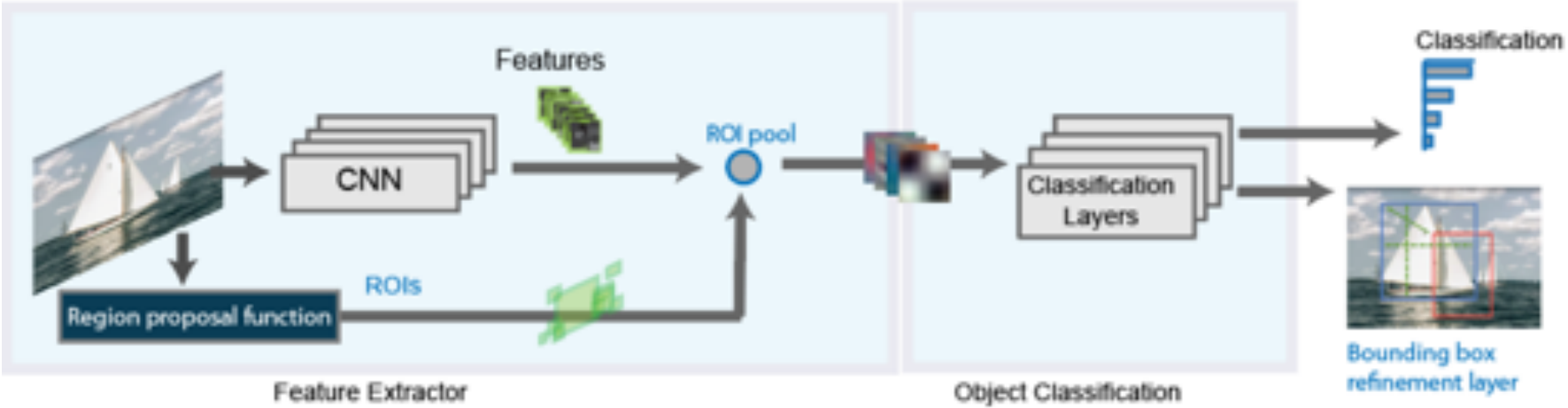
- Scale-invariant translation of the center of P(x,y)
- Specify log space transformations of the width w and height h.

Learn predicted transformation $d_k(P)$

- $\hat{G}$: corrected predicted box calculated
- $d_k(P) = w_k^T \Phi_5(P)$
  - w_k: learnable model parameters. is learnt by optimizing the regularized least-squares objective function
  - Φ_5: is dependent on the actual image features.

Ο αλγόριθμος μαθαίνει από ένα προβλεπόμενο πλαίσιο P μόνο αν βρίσκεται κοντά σε τουλάχιστον ένα πλαίσιο groundtruth.
Κάθε προβλεπόμενο πλαίσιο P αντιστοιχίζεται στο groundtruth του επιλέγοντας το πλαίσιο groundtruth με το οποίο έχει μέγιστη επικάλυψη (υπό την προϋπόθεση ότι έχει επικάλυψη IoU> 0,5).

Region-Based CNN: Fast R-CNN



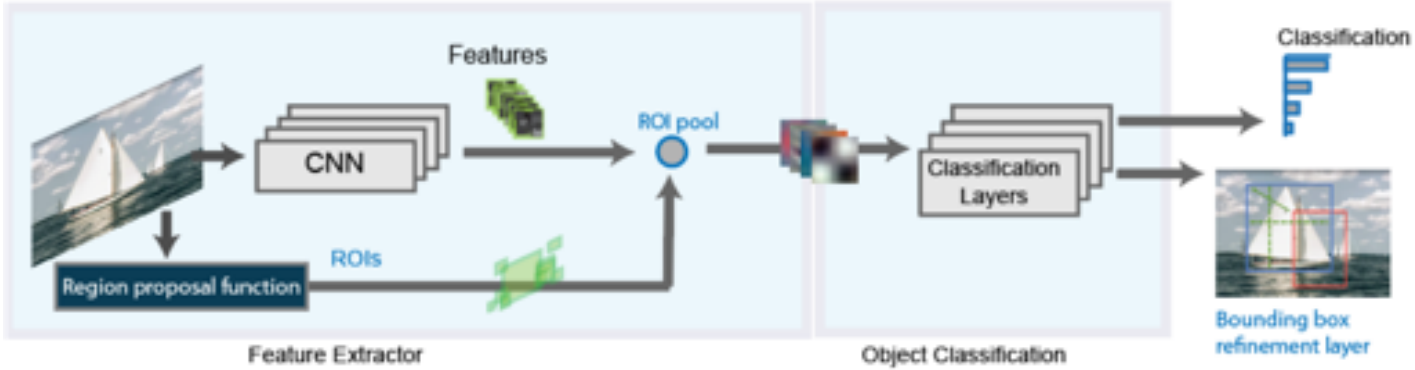
- ➔ Βελτιώνει σημαντικά την ταχύτητα του μοντέλου αλλάζοντας απλώς τη σειρά των επιπέδων του.
- ➔ Αντί να δίνεται ως είσοδος στο CNN κάθε μία από τις περιοχές ενδιαφέροντος, το CNN εξάγει τα χαρακτηριστικά ολόκληρης της εικόνας σε μορφή χάρτη χαρακτηριστικών, και στη συνέχεια για κάθε περιοχή απομονώνεται το αντίστοιχο τμήμα.

Με αυτό τον τρόπο, η εξαγωγή των χαρακτηριστικών γίνεται μόνο μία φορά αντί για 2000, γεγονός που είναι προφανές ότι βελτιώνει κατά πολύ την ταχύτητα της εκπαίδευσης και της πρόβλεψης.

Fast R-CNN Ross Girshick, ICCV 2015

Region-Based CNN: Fast R-CNN

1. Επεξεργάζεται ολόκληρη την εικόνα,
2. Ενώ ο R-CNN detector κατηγοριοποιεί κάθε περιοχή, ο Fast R-CNN συγκεντρώνει τα features maps από το CNN που αντιστοιχούν σε κάθε προτεινόμενη περιοχή (region proposal),
3. Κάθε περιοχή περνά από ένα fully connected network και ένα softmax layer δίνει τις κατηγορίες εξόδου.
4. Μαζί με το στρώμα softmax, χρησιμοποιείται επίσης ένα linear regression layer, παράλληλα για την παραγωγή των συντεταγμένων του bounding box για προβλεπόμενες κατηγορίες.



Region-Based CNN: Faster R-CNN

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun

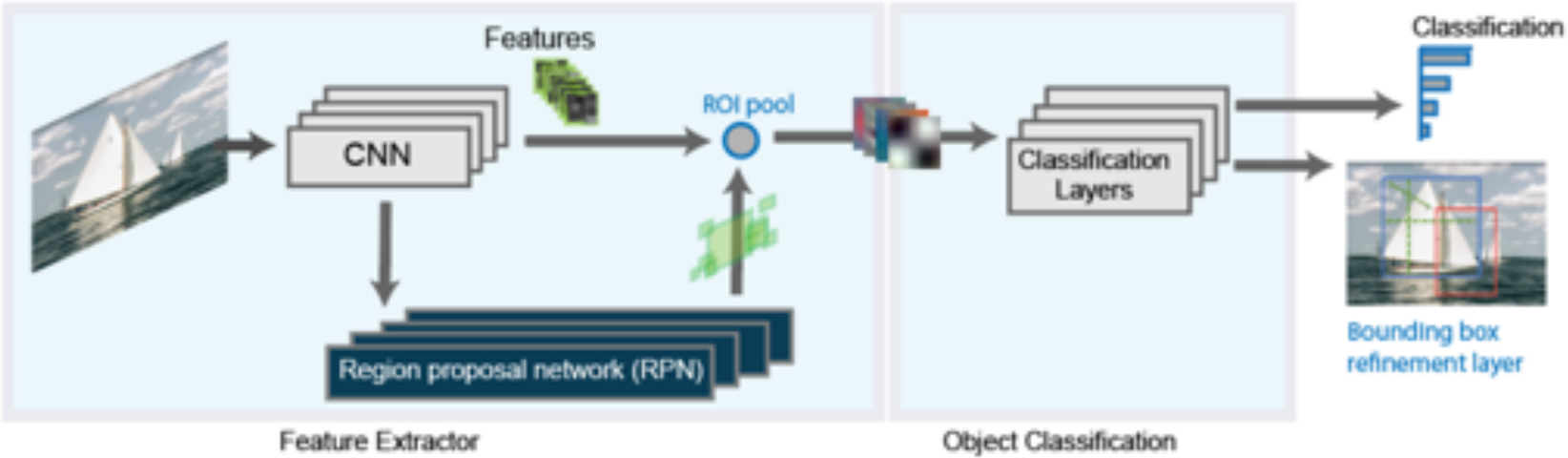
- Πρότειναν την αντικατάστασή του Selective Search, από τον δικό τους αλγόριθμο:
  Δίκτυο Πρότασης Περιοχών

    - Συγκεκριμένα, αντιλήφθηκαν ότι ο χάρτης χαρακτηριστικών που παράγεται από το συνελικτικό τμήμα του Fast R-CNN μπορεί να χρησιμοποιηθεί αποτελεσματικά και για το πρόβλημα της πρότασης περιοχών, αντικαθιστώντας τις αργές μεθόδους όπως η Selective Search με ένα εκπαιδεύσιμο νευρωνικό δίκτυο.

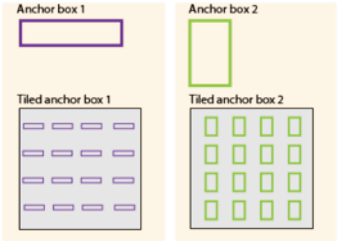
Region-Based CNN: Faster R-CNN

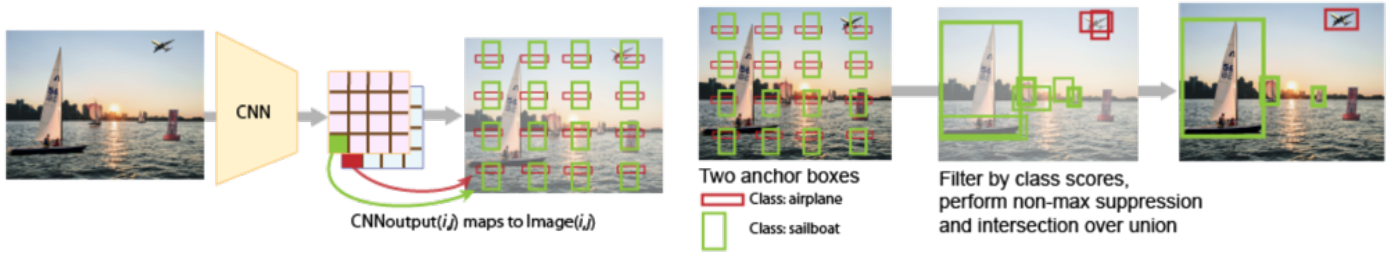
- Το Faster R-CNN προσθέτει ένα region proposal network (RPN) για να δημιουργήσει region proposals απευθείας μέσω δικτύου

- Το RPN χρησιμοποιεί Anchor Boxes για το Object Detection



Region-Based CNN: Faster R-CNN → Anchor Boxes





Τα anchor boxes είναι ένα σύνολο από προκαθορισμένα bounding boxes με συγκεκριμένο πλάτος και ύψος:

- ορίζονται για να καταγράφουν την κλίμακα και το λόγο διαστάσεων συγκεκριμένων κατηγοριών αντικειμένων που θέλουμε να εντοπίσουμε, (μπορούμε να έχουμε anchor boxes διαφορετικών μεγεθών)
- επιλέγονται συνήθως με βάση τα μεγέθη αντικειμένων στα training datasets,

Κατά τη διάρκεια της ανίχνευσης:

- τα predefined anchor boxes διατρέχουν την εικόνα,
- το δίκτυο προβλέπει την πιθανότητα και άλλα χαρακτηριστικά (background, IoU, offsets) για κάθε anchor box καθώς διατρέχει την εικόνα,
- επιστρέφεται ένα μοναδικό σύνολο προβλέψεων για κάθε καθορισμένο bounding box

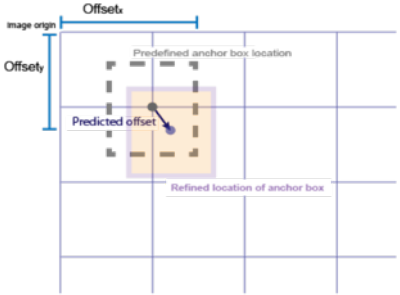
Το τελικό feature map αντιπροσωπεύει ανιχνεύσεις αντικειμένων για κάθε κατηγορία

Η χρήση anchor boxes επιτρέπει σε ένα δίκτυο να ανιχνεύει πολλά αντικείμενα, αντικείμενα διαφορετικών κλιμάκων και αλληλεπικαλυπτόμενα αντικείμενα.

Anchor Boxes for Object Detection - MATLAB & Simulink

Σφάλματα εντοπισμού και βελτίωση

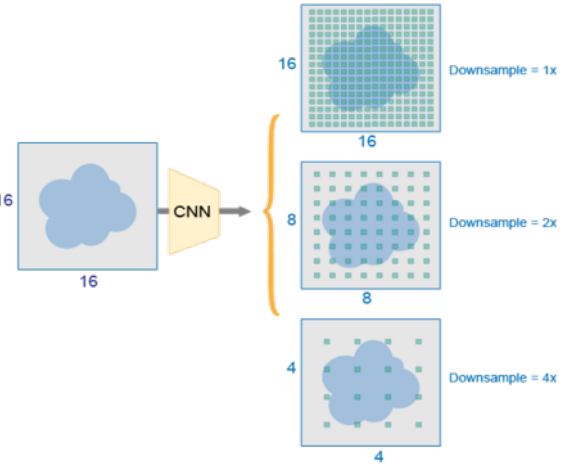
Σφάλματα εντοπισμού



- Οι ανιχνευτές αντικειμένων βαθιάς μάθησης μαθαίνουν αντισταθμίσεις για να εφαρμόζονται σε κάθε Tiled Anchor Box που βελτιώνει τη θέση και το μέγεθος του Anchor Box.

Βελτίωση

- Η απόσταση μεταξύ των Anchor Boxes είναι συνάρτηση του ποσού της δειγματοληψίας που υπάρχει στο CNN (max Pooling 2d Layer και του stride του Conv Layer)
- Τα feature maps που παράγουν τα αρχικά επίπεδα του CNN έχουν υψηλότερη χωρική ανάλυση, αλλά μπορεί να εξαγάγουν λιγότερες σημασιολογικές πληροφορίες σε σύγκριση με τα επίπεδα που βρίσκονται πιο κάτω στο δίκτυο



Δημιουργία ανιχνευτών αντικειμένων

- Αφαιρούνται τα Tiled Anchor Boxes που ανήκουν στην κατηγορία φόντου και τα υπόλοιπα φιλτράρονται από τη βαθμολογία εμπιστοσύνης τους.
- Τα Anchor Boxes με τη μεγαλύτερη βαθμολογία εμπιστοσύνης επιλέγονται χρησιμοποιώντας non-max suppression (NMS).

Algorithm 1 Non-Max Suppression

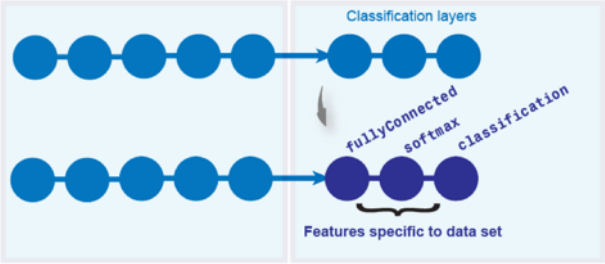
```
procedure NMS(B,c)
    B_nms ← ∅                                   Initialize empty set
    for b_i ∈ B do                              Iterate over all the boxes
        discard ← False                         Take boolean variable and set it as false. This variable indicates whether b(i) should be kept or discarded
        for b_j ∈ B do                          Start another loop to compare with b(i)
            if same(b_i, b_j) > λ_nms then      If both boxes having same IOU
                if score(c, b_j) > score(c, b_i) then
                    discard ← True              Compare the scores. If score of b(i) is less than that of b(j), b(i) should be discarded, so set the flag to True.
        if not discard then                     Once b(i) is compared with all other boxes and still the discarded flag is False, then b(i) should be considered. So add it to the final list.
            B_nms ← B_nms ∪ b_i
    return B_nms                                Do the same procedure for remaining boxes and return the final list
```

Region-Based CNN

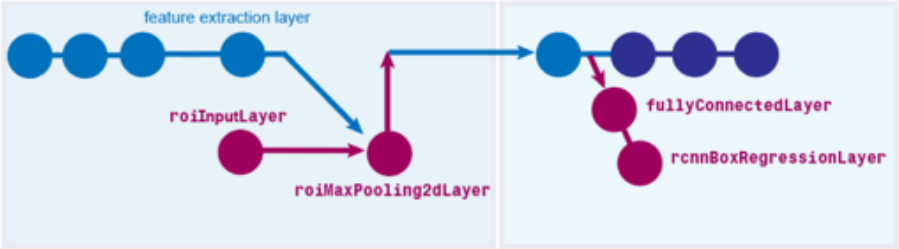
Transfer Learning model

'alexnet' ,'vgg16', 'vgg19', 'resnet50', 'resnet101', 'inceptionv3', 'googlenet', 'inceptionresnetv2', 'squeezenet'

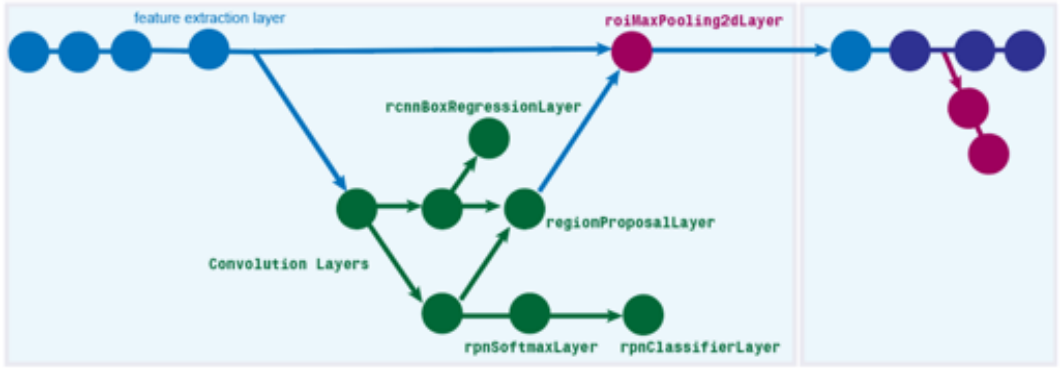
R-CNN



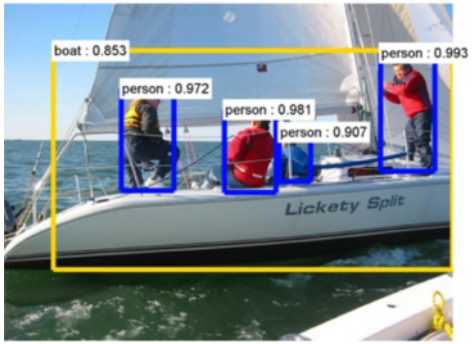
Fast R-CNN



Faster R-CNN



YOLO- You Only Look Once

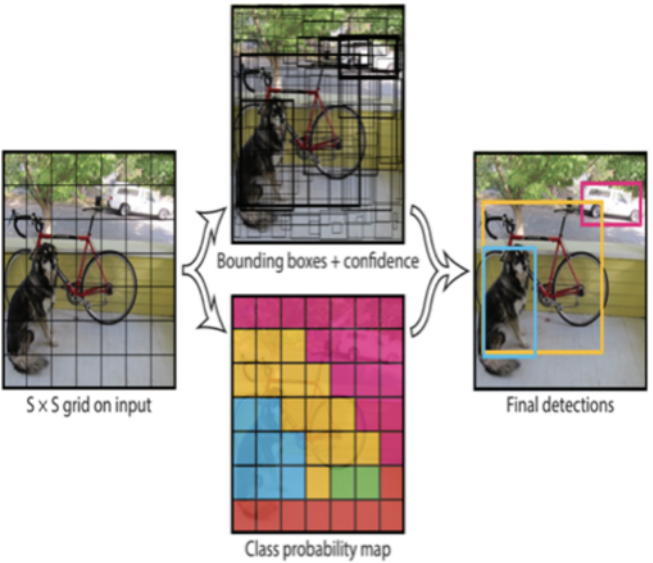


You only look once (YOLO) at an image to predict
what objects are present and where they are present.

- Χρησιμοποιεί ένα απλό CNN και βλέπει ολόκληρη την εικόνα κατά τη διάρκεια του training και του validation, οπότε κωδικοποιεί σιωπηρά πληροφορίες για τις κλάσεις καθώς και τις εμφανίσεις τους, σε αντίθεση με τις τεχνικές sliding window ή region-based (κάνοντας έτσι λιγότερο από το ήμισυ του αριθμού των σφαλμάτων σε σύγκριση με το Fast R-CNN).
- Το YOLO χρησιμοποιεί features από ολόκληρη την εικόνα για να προβλέψει κάθε bounding box
- Προβλέπει επίσης όλα τα bounding box σε όλες τις κλάσεις για μια εικόνα ταυτόχρονα με τις αντίστοιχες πιθανότητες
- Αντιμετωπίζει την ανίχνευση ως πρόβλημα παλινδρόμησης
- Εξαιρετικά γρήγορος και ακριβής αλγόριθμος

You Only Look Once: Unified, Real-Time Object Detection Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi

Λειτουργία YOLO- You Only Look Once



y =	
	pc
	bx
	by
	bh
	bw
	c1
	c2
	c3

y =	
	0
	?
	?
	?
	?
	?
	?

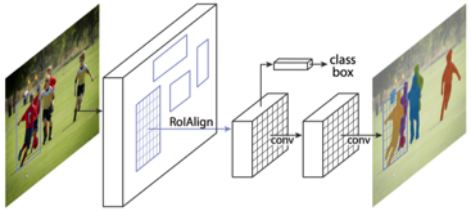
- Παίρνει μια εικόνα και τη χωρίζει σε πλέγμα SxS.
- Κάθε κελί πλέγματος προβλέπει μόνο ένα αντικείμενο.
- Η ταξινόμηση εικόνας και ο εντοπισμός εφαρμόζονται σε κάθε κελί του πλέγματος.
- Εάν το κέντρο ενός αντικειμένου πέσει σε ένα κελί πλέγματος, αυτό το κελί πλέγματος είναι υπεύθυνο για την ανίχνευση αυτού του αντικειμένου.
- Κάθε ένα από τα κελιά πλέγματος προβλέπει bounding boxes B με βαθμολογίες εμπιστοσύνης για αυτά τα bounding boxes
    - Οι βαθμολογίες εμπιστοσύνης αντικατοπτρίζουν το πόσο σίγουρο είναι το μοντέλο ότι το πλέγμα περιέχει ένα αντικείμενο και πόσο ακριβές πιστεύει ότι το πλαίσιο είναι αυτό που προβλέπει.Εάν δεν υπάρχουν αντικείμενα, τότε οι βαθμολογίες εμπιστοσύνης θα είναι μηδέν.
- Bounding box όταν ένα αντικείμενο υπάρχει στο κελί πλέγματος
- Πιθανότητα για class C

[Region-Based CNN]

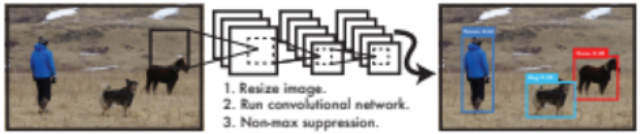
[Object Detection and Tracking in 2020 | by Borijan Georgievski]

[TensorFlow Hub Object Detection Colab]

[Mask R-CNN]



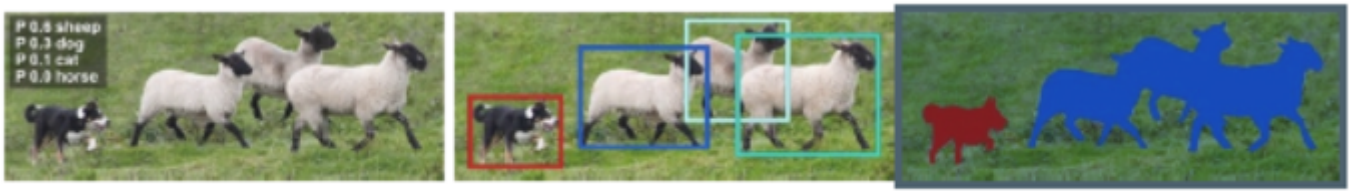
[YOLO original paper]



[YOLOv2 YOLO9000]

[DarkNet implementation]

Πρόβλημα Σημασιολογικής Κατάτμησης



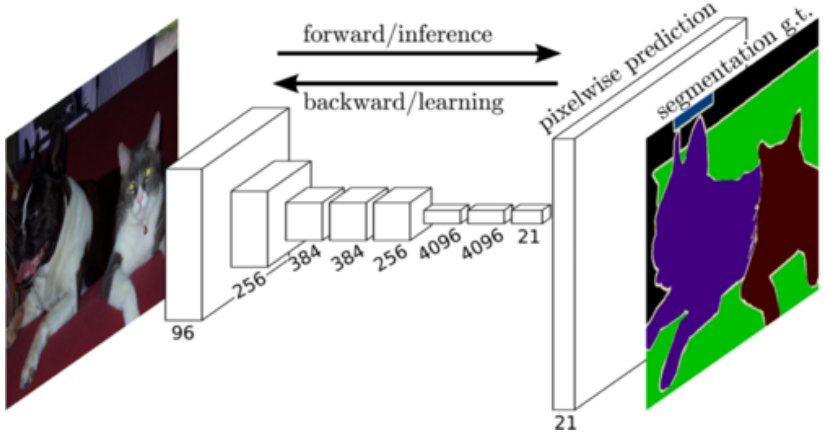
(a) Ταξινόμηση Εικόνας (b) Ανίχνευση Αντικειμένου (c) Σημασιολογική Κατάτμηση

➔ Ταξινόμηση σε επίπεδο εικονοστοιχείου
- Ανάθεση κλάσης σε κάθε εικονοστοιχείο της εικόνας
- Μέσω χωρικής ανάλυσης ενός εικονοστοιχείου.

➔ Αφελής πρώτη προσέγγιση: Υλοποίηση ενός μοντέλου με διαδοχικά συνελικτικά επίπεδα, του οποίου η έξοδος θα είχε την ίδια διάσταση με την είσοδο.
- το κάθε εικονοστοιχείο της εξόδου θα αποτελούσε την πρόβλεψη για την κλάση του αντίστοιχου εικονοστοιχείου της αρχικής εικόνας.
- απαγορευτική υπολογιστική πολυπλοκότητα.

Πρόβλημα Σημασιολογικής Κατάτμησης

Συνηθισμένες προσεγγίσεις: encoder-decoder.



1. Η διάσταση της εικόνας μειώνεται αρχικά (encoder), παράγοντας χαμηλότερης ανάλυσης χάρτες χαρακτηριστικών οι οποίοι έχουν πολύ καλά αποτελέσματα για την ταξινόμηση μεταξύ των κλάσεων,
2. Στη συνέχεια αυξάνεται και πάλι (decoder), μέχρι να προκύψει ο τελικός χάρτης κατάτμησης.

Πρόβλημα Σημασιολογικής Κατάτμησης

Πλήρως Συνελικτικό Δίκτυο (Fully Convolutional Network - FCN)

- CNN : Fully Connected Layers → Fully Convolutional Network
  - Κωδικοποιητής: Εξαγάγει τα χαρακτηριστικά και είναι εκπαιδευμένος με βάση το πρόβλημα της ταξινόμησης.
  - Αποκωδικοποιητής: Προβάλλει το χάρτη χαρακτηριστικών χαμηλής ανάλυσης που προέκυψε από τον κωδικοποιητή στην αρχική εικόνα.
    - Αποτελείται από μία σειρά συνελίξεις (backwards convolutions ή deconvolutions)
      - πραγματοποιούν αύξηση της χωρικής ανάλυσης με χρήση διγραμμικής παρεμβολής (bilinear interpolation).
    - Κάνει χρήση παρακαμπτήριων συνδέσεων (skip connections), που εκμεταλλεύονται τις παρόμοιες διαστάσεις των εκατέρωθεν επιπέδων του FCN και συνδέουν σειριακά τους χάρτες ενεργοποίησης του κωδικοποιητή με την αντίστοιχη δομή που προκύπτει μετά από κάθε αποσυνέλιξη.

Πρόβλημα Κατάτμησης Στιγμιοτύπων (Instance Segmentation)

(a) Ταξινόμηση Εικόνας (b) Ανίχνευση Αντικειμένου (c) Σημασιολογική Κατάτμηση (d) Κατάτμηση Στιγμιοτύπων

Κατάτμηση Στιγμιοτύπων = Ανίχνευσης Αντικειμένων + Σημασιολογική Κατάτμηση

Στοχεύει στον εντοπισμό των διαφορετικών αντικειμένων σε μία εικόνα όχι με χρήση πλαισίων οριοθέτησης αλλά με ακρίβεια εικονοστοιχείου.

➔ Κάθε εικονοστοιχείο ταξινομείται σε μία κλάση, όπως στη Σημασιολογική Κατάτμηση, αλλά τα διαφορετικά αντικείμενα θα έχουν άλλη μάσκα, ακόμα κι αν ανήκουν στην ίδια κλάση.

B. Hariharan, P. Arbelaez, R. B. Girshick, and J. Malik. Simultaneous Detection and Segmentation. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, Computer Vision – ECCV 2014, volume 8695 of Lecture Notes in Computer Science. Springer, Cham, 2014.

Πρόβλημα Κατάτμησης Στιγμιοτύπων

Mask R-CNN

- Faster R-CNN με δίκτυο RPN, για την πρόταση των υποψηφίων περιοχών,
- Τμήμα για τον υπολογισμό των μασκών, αντίστοιχο με ένα Πλήρως Συνελικτικό Δίκτυο (FCN).