ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
& ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΙΤΛΟΣ

Αναγνώριση αντικειμένων/προτύπων με χρήση βαθιάς μάθησης και εφαρμογές

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΣΙΡΤΣΑΚΗΣ ΠΑΣΧΑΛΗΣ

Επιβλέπων: Γεώργιος Φ. Φραγκούλης
Καθηγητής

16 Φεβρουαρίου 2024

HELLENIC DEMOCRACY
UNIVERSITY OF WESTERN MACEDONIA
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL
& COMPUTER ENGINEERING

ece.uowm.gr

TITLE

Deep Learning with Applications to Object/Pattern Recognition

THESIS

TSIRTSAKIS PASCHALIS

SUPERVISOR: George F. Fragulis
Professor

16 February 2024

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
& ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα Διπλωματική Εργασία με τίτλο "Αναγνώριση αντικειμένων/προτύπων με χρήση βαθιάς μάθησης και εφαρμογές" καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας και αναφέρονται ρητώς μέσα στο κείμενο που συνοδεύουν, και η οποία έχει εκπονηθεί στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας, υπό την επίβλεψη του μέλους του Τμήματος κ. Γεώργιου Φ. Φραγκούλη αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή / και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και μόνο.

Περίληψη

Τα τελευταία χρόνια η υπολογιστική όραση έχει γνωρίσει ραγδαίες εξελίξεις αφού η πρόοδος της τεχνητής νοημοσύνης (AI) και της μηχανικής μάθησης (ML) μας έχουν οδηγήσει σε μια νέα εποχή. Αυτό έχει ως αποτέλεσμα την ανάπτυξη νέων τεχνολογιών, κυρίως στον τομέα της αναγνώρισης αντικειμένων, όπου τα εμπόδια και οι περιορισμοί αυξάνονται συνεχώς . Στην αντιμετώπιση αυτών των προκλήσεων έχουν πρωταγωνιστήσει πολλά μοντέλα βαθιάς μάθησης (DL) με ιδιαίτερες δυνατότητες εκμάθησης χαρακτηριστικών και εξαιρετική απόδοση σε προβλήματα ανίχνευσης αντικειμένων. Η παρούσα διπλωματική διατριβή εστιάζει στην ανάλυση των βασικών αρχιτεκτονικών και αλγορίθμων βαθιάς μάθησης καθώς και στην αξιολόγηση της απόδοσής τους σε συγκεκριμένα σύνολα δεδομένων.

Λέξεις-Κλειδιά: Αναγνώριση Αντικειμένων, Υπολογιστική Όραση, Βαθιά Μάθηση, Συνελικτικά Νευρωνικά Δίκτυα

Abstract

Computer vision has advanced significantly in recent years leading us into a new era, due to the progress of Artificial Intelligence (AI) and Machine Learning (ML). This has resulted in the development of new technologies with many real-word applications, especially in the field of object recognition, where lots of obstacles and limitations exist. To tackle such challenges, deep learning (DL) models have effective feature learning capabilities and excellent performance on object detection tasks. This thesis focuses on the analysis of basic deep learning architectures and cutting-edge algorithms as well as on the evaluation of their performance on specific datasets.

Keywords: Object Recognition, Computer Vision, Deep Learning, Convolutional Neural Networks

Ευχαριστίες

Πρώτα από όλα, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στον καθηγητή κ. Γιώργο Φραγκούλη και στον υποψήφιο διδάκτορα κ. Γιώργο Μαρασλίδη, καθώς η καθοδήγηση και η αμέριστη υποστήριξή τους ήταν καθοριστική στην διάρκεια της ακαδημαϊκής μου πορείας. Ταυτόχρονα θα ήθελα να τους ευχαριστήσω για την εμπιστοσύνη και την βαθιά κατανόηση που έδειξαν στην εκπόνηση της συγκεκριμένης εργασίας.

Φυσικά, οφείλω να ευχαριστήσω την οικογένειά μου για την αγάπη, την ενθάρρυνση και τις θυσίες που έκανε σε όλη την διάρκεια των σπουδών μου. Είμαι επίσης ευγνώμων στον διοικητή μου κ. Φώτη Αλεξίου, ο οποίος με βοήθησε να εξισορροπήσω τις απαιτήσεις των σπουδών μου με την στρατιωτική μου θητεία. Τέλος δεν θα μπορούσα να ξεχάσω τους φίλους και τις αναμνήσεις που απέκτησα όλα αυτά τα χρόνια.

16 Φεβρουαρίου 2024

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Προσδιορισμός του θέματος

Η αναγνώριση αντικειμένων (Object Recognition) είναι ένα πεδίο της υπολογιστικής όρασης (Computer Vision) που εστιάζει στην ικανότητα εντοπισμού αντικειμένων συγκεκριμένου ενδιαφέροντος σε εικόνες και βίντεο. Ο στόχος της αναγνώρισης αντικειμένων είναι η ανάπτυξη ευέλικτων αλγορίθμων που συνδυάζουν μεγάλα ποσοστά ακρίβειας και υψηλή απόδοση. Το συγκεκριμένο επιστημονικό πεδίο ερευνάται από καταξιωμένους επιστήμονες για τουλάχιστον τρεις δεκαετίες και συναντάται σε πληθώρα καθημερινών εφαρμογών. Μερικές από αυτές είναι η αυτόνομη οδήγηση, η έξυπνη παρακολούθηση σε συστήματα ασφαλείας, αναγνώριση ασθενειών σε ιατρικές απεικονίσεις, ανάκτηση εικόνων κλπ. Το πρόβλημα της αναγνώρισης αντικειμένων είναι εξαιρετικά περίπλοκο καθώς οι προκλήσεις προέρχονται από τον αριθμό των διαφορετικών κλάσεων ανά αντικείμενο και τις άπειρες συνθήκες απεικόνισης. Μέχρι σήμερα έχει αποδειχτεί ότι η βέλτιστη επιλογή για την αντιμετώπιση του προβλήματος είναι χρήση Συνελικτικών Νευρωνικών Δικτύων.

1.2 Σκοπός και στόχοι της εργασίας

Σκοπός της εργασίας είναι αρχικά η κατανόηση της βασικής θεωρίας στην επιστήμη της βαθιάς μάθησης. Στην συνέχεια η εργασία στοχεύει στην άντληση της απαραίτητης γνώσης από την κατάλληλη βιβλιογραφία στον τομέα των αρχιτεκτονικών και αλγορίθμων βαθιάς μάθησης, που χρησιμοποιούνται για την ταξινόμηση και τον εντοπισμό αντικειμένων σε εικόνες. Κατά την μελέτη της εργασίας, ο αναγνώστης θα κατανοήσει πλήρως τον τρόπο λειτουργείας και τις βασικές διαφορές των μοντέλων και ανιχνευτών που έχουν ξεχωρίσει μέχρι σήμερα.

1.3 Παρουσίαση και δομή της εργασίας

Τα βασικά μέρη της εργασίας είναι τα τέσσερα κεφάλαια που ακολουθούν. Το δεύτερο κεφάλαιο αποτελεί το θεωρητικό υπόβαθρο της εργασίας όπου γίνεται μια απλή εισαγωγή στις πρωταρχικές έννοιες της μηχανικής και βαθιάς μάθησης και έπειτα αναπτύσσεται λεπτομερώς η θεωρία των νευρωνικών και συνελικτικών νευρωνικών δικτυών, με σκοπό την καλύτερη κατανόηση των επόμενων κεφαλαίων. Το τρίτο κεφάλαιο είναι αφιερωμένο αποκλειστικά στην αναγνώριση αντικειμένων, ξεκινώντας με την ιστορική αναδρομή, τον ορισμό βασικών εννοιών και κριτήριων αξιολόγησης και έπειτα παρουσιάζονται οι ιδιαιτερότητες των διαθέσιμων συνόλων δεδομένων. Ακολουθούν τα τελευταία κεφάλαια που εμβαθύνουν στο αντικείμενο όπου γίνεται πλήρης επεξήγηση των αρχιτεκτονικών και αλγορίθμων βαθιάς μάθησης, μαζί με την καταγραφή και σύγκριση της απόδοσής τους.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Τεχνητή Νοημοσύνη, Μηχανική Μάθηση και Βαθιά Μάθηση

2.1.1 Εισαγωγή στην Τεχνητή Νοημοσύνη

Η τεχνητή νοημοσύνη (Artificial Intelligence, AI) είναι η νοημοσύνη που επιδεικνύεται από μηχανές ικανές να ανταγωνιστούν τον άνθρωπο (Ongsulee, 2017). Στην επιστήμη των υπολογιστών, ο τομέας της τεχνητής νοημοσύνης ορίζεται ως η μελέτη των «ευφυών πρακτόρων», δηλαδή των συσκευών που αντιλαμβάνονται το περιβάλλον γύρω τους και πραγματοποιούν ενέργειες που μεγιστοποιούν την επιτυχία τους σε κάποιο πιθανό στόχο (Russell & Norvig, 2010). Γενικότερα, χρησιμοποιούμε τον όρο τεχνητή νοημοσύνη όταν αναφερόμαστε σε μια μηχανή που μιμείται τις γνωστικές λειτουργίες όπως η μάθηση και η επίλυση προβλημάτων (Shinde & Shah, 2018). Η αφετηρία και η ονομασίας της επιστήμης δόθηκε το 1956 από τον επιστήμονα John McCarthy σε συνέδριο της εποχής όπου παρουσιάστηκε το πρώτο πρόγραμμα τεχνητής νοημοσύνης (McCorduck & Cfe, 2004). Τις επόμενες δεκαετίες έγιναν τα πρώτα βήματα στην εξέλιξής της και μέχρι το 2012 η χρήση της περιοριζόταν μόνο σε εταιρίες κολοσσούς, κυβερνήσεις και ερευνητικούς φορείς. Έκτοτε, η τεχνητή νοημοσύνη έχει εισχωρήσει σε κάθε κοινωνία προσφέροντας λύσεις στα καθημερινά προβλήματα (Ongsulee, 2017).

2.1.2 Εισαγωγή στην Μηχανική Μάθηση

Η μάθηση στον άνθρωπο θεωρείται σημαντικό μέρος της νοημοσύνης του. Ο ορισμός της μάθησης είναι η αλλαγή της συμπεριφοράς, των γνώσεων, των δεξιοτήτων και των αντιλήψεων ενός ατόμου. Η αλλαγή αυτή προκαλείται από την εμπειρία και την εκπαίδευση (Ertmer & Newby, 1993). Ταυτόχρονα, η μάθηση συνδέεται με την απόκτηση καινούργιων γνώσεων και δεξιοτήτων καθώς και με την ανάπτυξη νέων πεποιθήσεων.

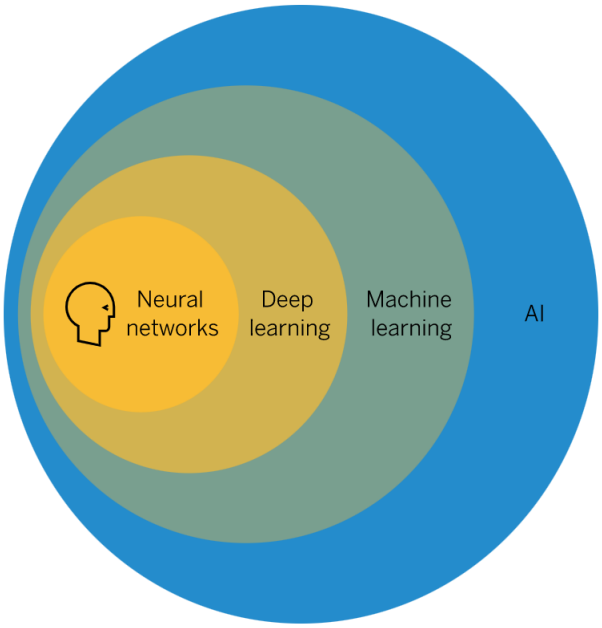
Η μηχανική μάθηση (machine learning) είναι ένας κλάδος τη επιστήμης των υπολογιστών που πραγματεύεται την ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν σε υπολογιστικά συστήματα να μαθαίνουν από δεδομένα χωρίς να χρειάζεται να προγραμματιστούν ρητά από τον άνθρωπο (Samuel, 1959). Ο θεωρητικός ορισμός για την περιγραφή της μηχανικής μάθησης προέρχεται από τον Tom Mitchell στο βιβλίο του “Machine Learning”, το οποίο εκδόθηκε από το McGraw το 1997: «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία E σε σχέση με κάποια εργασία T και κάποιο μετρό απόδοσης P, εάν η απόδοσή του στο T, όπως μετριέται με το P, βελτιώνεται με την εμπειρία E» (Mitchell, 1997).

Η μηχανική μάθηση αποτελεί μια υποκατηγορία της τεχνητής νοημοσύνης έχοντας εφαρμογή σε πολλούς επιστημονικούς κλάδους. Στόχος της μηχανικής μάθησης είναι η δυνατότητα παραγωγής σωστών εκτιμήσεων σχετικά με δεδομένα τα οποία αντιμετωπίζονται για πρώτη φορά στο σύστημα (ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019). Ο σκοπός της μηχανικής μάθησης

είναι η κατασκευή και εκπαίδευση υπολογιστικών μοντέλων για την εύρεση μοτίβων ή συσχετίσεων σε μεγάλα σύνολα δεδομένων, και την πραγματοποίηση προβλέψεων με βάση την ήδη υπάρχων πληροφορία. Η μάθηση της πληροφορίας βασίζεται σε παραδείγματα τα οποία σχετίζονται με δεδομένα και παρατηρήσεις. Η μηχανική μάθηση συνδέεται στενά με την επιστήμη της στατιστικής και την επιστήμη των μαθηματικών (Ongsulee, 2017). Ακόμα, η μηχανική μάθηση συσχετίζεται πολλές φορές και με την εξόρυξη δεδομένων, παρόλα αυτά υπάρχουν βασικές διαφορές που διαχωρίζουν τα συγκεκριμένα πεδία.

2.1.3 Εισαγωγή στην Βαθιά Μάθηση

Η βαθιά μάθηση (deep learning) είναι ο τομέας της μηχανικής μάθησης που πλαισιώνει τα δίκτυα πολλών επιπέδων. Σύμφωνα με τον LeCun, η βαθιά μάθηση επιτρέπει σε υπολογιστικά μοντέλα που αποτελούνται από πολλαπλά επίπεδα στρωμάτων να μαθαίνουν αναπαραστάσεις δεδομένων με συνθέτη επεξεργασία (LeCun et al., 2015). Οι περισσότερες μέθοδοι βαθιάς μάθησης χρησιμοποιούν αρχιτεκτονικές νευρωνικών δικτυών, δηλαδή βαθιά νευρωνικά δίκτυα. Συνοπτικά, οι αρχιτεκτονικές σχηματίζουν μια ιεραρχική δομή από επίπεδα που συλλέγουν πολύπλοκα χαρακτηριστικά και από επίπεδα που συλλέγουν απλά χαρακτηριστικά. Αυτό καθιστά την βαθιά μάθηση κατάλληλη για ανάλυση και εξαγωγή χρήσιμης γνώσης από μεγάλες ποσότητες δεδομένων αλλά και από δεδομένα που προέρχονται από διαφορετικές πήγες (Shinde & Shah, 2018).

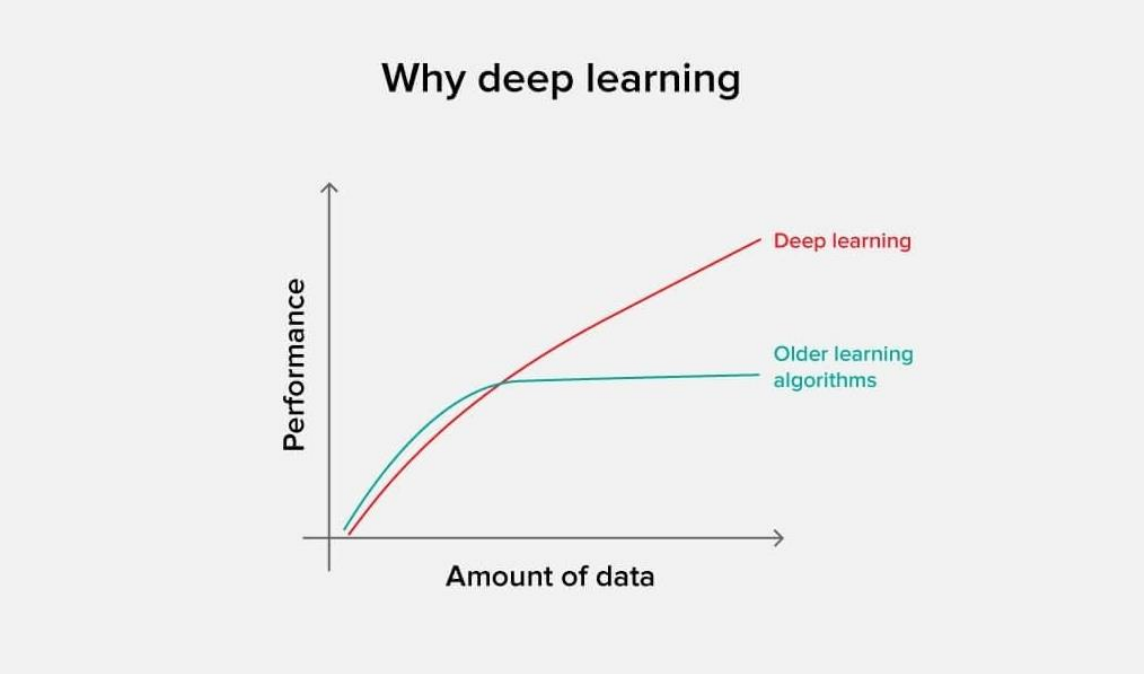


Εικόνα 1: Διάγραμμα σχέσης μεταξύ AI, Machine Learning και Deep Learning. Πηγή sap.com

Ο τρόπος με τον οποίο οι επιστημονικοί τομείς συσχετίζονται μεταξύ τους, απεικονίζεται στην εικόνα 1. Κάθε επιστήμη θεωρείται ως ένα σύνολο. Η τεχνητή νοημοσύνη περιλαμβάνει όλα τα υπόλοιπα υποσύνολα αφού η μηχανική μάθηση και η βαθιά μάθηση αποτελούν πιο εξειδικευμένα επιστημονικά πεδία. Αντίστοιχα, όλα τα προβλήματα της μηχανικής μάθησης ή βαθιάς μάθησης είναι και προβλήματα τεχνητής νοημοσύνης.

2.1.4 Βαθιά Μάθηση αντί για Μηχανική Μάθηση

Τις τελευταίες δεκαετίες η άνθηση της βαθιάς μάθησης είναι αδιαμφισβήτητη στην επιστήμη των υπολογιστών. Μάλιστα, σε αρκετές περιπτώσεις η βαθιά μάθηση είναι προτιμότερη από την μηχανική μάθηση. Το μεγαλύτερο πλεονέκτημα της βαθιάς μάθησης είναι η ικανότητά της να τροφοδοτείται με τεράστιο όγκο δεδομένων. Βέβαια, το γεγονός αυτό απαιτεί υπολογιστικά μηχανήματα προηγμένης τεχνολογίας αφού η εκπαίδευση βαθιών μοντέλων είναι υπολογιστικά δαπανηρή και χρονοβόρα. Σε αντίθεση, η μηχανική μάθηση είναι αποτελεσματική για μικρότερα σύνολα δεδομένων όπου οι παραδοσιακοί αλγόριθμοι θεωρούνται πιο αποδοτικοί. Ακόμα, στην μηχανική μάθηση συνήθως απαιτείται προεπεξεργασία δεδομένων, η οποία περιλαμβάνει ανθρώπινη παρέμβαση. Μια επιπλέον σημαντική διαφορά μεταξύ των δύο, είναι η προσέγγιση στην επίλυση προβλημάτων. Οι τεχνικές μηχανικής μάθησης τείνουν να αντιμετωπίζουν τα προβλήματα ως ξεχωριστά κομμάτια που λύνονται με διαφορετικούς τρόπους και στην συνέχεια συνδυάζονται για το τελικό αποτέλεσμα, κάτι το οποίο δεν συμβαίνει με τις τεχνικές βαθιάς μάθησης (Janiesch et al., 2021).



Εικόνα 2: Διάγραμμα αποδοτικότητας βαθιάς μάθησης σε σχέση με άλλες τεχνικές μηχανικής μάθησης. Πηγή v7labs.com

2.1.5 Εφαρμογές Βαθιάς Μάθησης και Νευρωνικών Δικτύων

Τα νευρωνικά δίκτυα είναι ιδιαίτερα δημοφιλή σε προβλήματα που δεν είναι πλήρως κατανοητά και έχουν αβέβαια συμπεριφορά. Τέτοια προβλήματα συναντώνται σε πολλές καθημερινές ανθρώπινες δραστηριότητες που σχετίζονται με την κατηγοριοποίηση, την αναγνώριση, την αποτίμηση και την πρόβλεψη. Παρακάτω περιγράφονται μερικές εφαρμογές της βαθιάς μάθησης σε συγκεκριμένους τομείς (Shinde & Shah, 2018; Vlahavas et al., 2020):

- Ιατρικός τομέας:
  - Κατηγοριοποίηση ιατρικών εξετάσεων.
  - Πρόβλεψη και διάγνωση ασθενειών.
- Αμυντικός τομέας:
  - Κατηγοριοποίηση εικόνων αμυντικών συστημάτων.
  - Παρακολούθηση στόχων.
  - Εντοπισμός κίνησης στόχων.
- Γεωργικός τομέας:
  - Έλεγχος καλλιεργειών.
  - Πρόβλεψη καλλιεργειών.
  - Εντοπισμός καλλιεργειών.
  - Πρόβλεψη καιρού.
- Χρηματοοικονομικός τομέας:
  - Κατηγοριοποίηση πελατών.
  - Αναγνώριση γνησιότητας υπογραφής.
  - Πρόβλεψη ισοτιμίας νομισμάτων.
  - Πρόβλεψη μετοχών και πωλήσεων.
  - Αποτίμηση ακίνητης περιουσίας.
  - Αποτίμηση δανείων.
- Τομέας της Τέχνης:
  - Δημιουργία πρωτότυπων εικόνων.
  - Αυτόματος υποτιτλισμός ταινιών και βίντεο.
  - Εντοπισμός αντικειμένων σε εικόνα και δημιουργία λεζάντας.
- Τομέας της Τεχνολογίας:
  - Ανάπτυξη αυτόνομων οχημάτων.
  - Ανάπτυξη έξυπνων πόλεων.
  - Διάφορες εφαρμογές ρομποτικής.
  - Εφαρμογές Αεροναυπηγικής και Αεροδιαστημικής.

2.2 Νευρωνικά Δίκτυα

2.2.1 Εισαγωγή

Τα νευρωνικά δίκτυα είναι υπολογιστικά μοντέλα εμπνευσμένα από τον εγκέφαλο. Ένα νευρωνικό δίκτυο είναι μια συλλογή κόμβων ή μονάδων που συνδέονται μεταξύ τους. Η τοπολογία και οι ιδιαιτερότητες των νευρώνων καθορίζουν τις ιδιότητες του δικτύου (Russell & Norvig, 2010). Η θεωρία των νευρωνικών δικτύων έχει επηρεαστεί από διάφορες επιστήμες και έχει βασιστεί σε έννοιες από την νευρολογία, τα μαθηματικά και την πληροφορική. Τα μοντέλα νευρωνικών δικτύων είναι μια προσπάθεια μοντελοποίησης της διαδικασίας κατά την οποία ο ανθρώπινος εγκέφαλος επεξεργάζεται την πληροφορία.

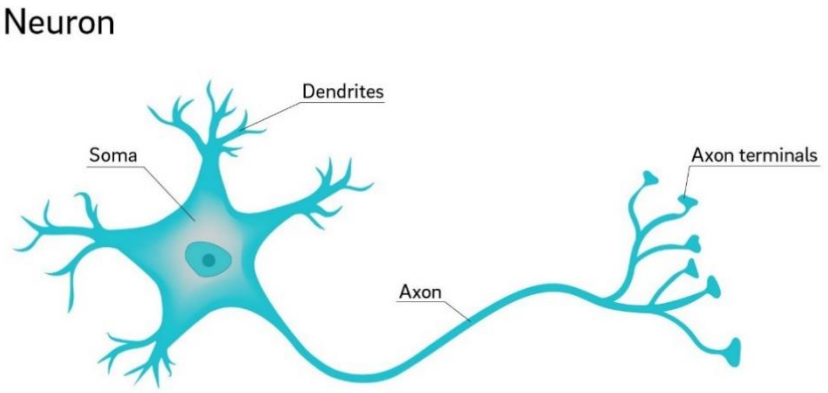
2.2.2 Βιολογικά Νευρωνικά Δίκτυα

Το νευρικό σύστημα στον άνθρωπο, ρυθμίζει και ελέγχει την λειτουργία όλων των οργάνων που βρίσκεται στο σώμα του. Το σύστημα αυτό, αποτελείται κυρίως από εξειδικευμένα κύτταρα που ονομάζονται νευρώνες. Κύρια λειτουργία των νευρώνων είναι η μετάδοση ερεθισμάτων σε ολόκληρο το νευρικό σύστημα. Τα μέρη ενός τυπικού βιολογικού νευρώνα είναι ο πυρήνας του, οι δενδρίτες που αποτελούν την είσοδο του συστήματος και οι άξονες που αποτελούν την έξοδο του συστήματος. Οι δενδρίτες λαμβάνουν σήματα από νευρώνες και μέσω του άξονα τα σήματα αυτά στέλνονται σε διαφορετικούς νευρώνες. Οι νευρώνες συνδέονται μεταξύ τους με συνάψεις, οι οποίες έχουν ως αφετηρία τους άξονες ενός νευρώνα και καταλήγουν στους δενδρίτες του επομένου νευρώνα (Basegmez, 2014).

Μεταξύ των συνάψεων πραγματοποιούνται χημικές διαδικασίες οι οποίες είτε επιταχύνουν είτε επιβραδύνουν την ροή των ηλεκτρικών φορτίων. Η δράση αυτή έχει ως αποτέλεσμα τόσο διεγερτικών όσο και ανασταλτικών εισροών προς τον νευρώνα ο οποίος με αυτόν τον τρόπο ενεργοποιείται (Gluck & Myers, 2001). Αξίζει να σημειωθεί πως η διαδικασία που περιγράφηκε είναι όμοια για τους ανθρώπους και για τα περισσότερα ζώα της φύσης οπού αυτή η δικτυακή δομή του εγκέφαλου φαίνεται να είναι η βασική προϋπόθεση για την εμφάνιση συνείδησης και συνθέτης συμπεριφοράς (Rojas, 2013).

Οι νευρώνες μπορούν να βρίσκονται σε δυο καταστάσεις, την ενεργή και την μη ενεργή κατάσταση. Ένας νευρώνας είναι ενεργός όταν παράγει ένα ηλεκτρικό σήμα το οποίο μεταφέρει δεδομένα σε έναν γειτονικό νευρώνα. Αντίστοιχα, όταν ένας νευρώνας δεν προκαλεί ηλεκτρικά σήματα βρίσκεται σε κατάσταση αδράνειας και θεωρείται μη ενεργός. Ο ηλεκτρικός παλμός παράγεται μόνο όταν το συνολικό άθροισμα φορτίου που βρίσκεται στον νευρώνα, είναι μεγαλύτερο από κατώφλι του (Yuste, 2015).

Σε αυτό το σημείο είναι κατανοητό πως η θεωρητική ανάλυση των βιολογικών νευρωνικών δικτύων είναι απαραίτητη καθώς έχουν καθοριστική σημασία στην ανάπτυξη των τεχνητών νευρωνικών δικτύων και στην δημιουργία μαθηματικών μοντέλων.



Εικόνα 3: Διάγραμμα αποδοτικότητας βαθιάς μάθησης σε σχέση με άλλες τεχνικές μηχανικής μάθησης. Πηγή v7labs.com

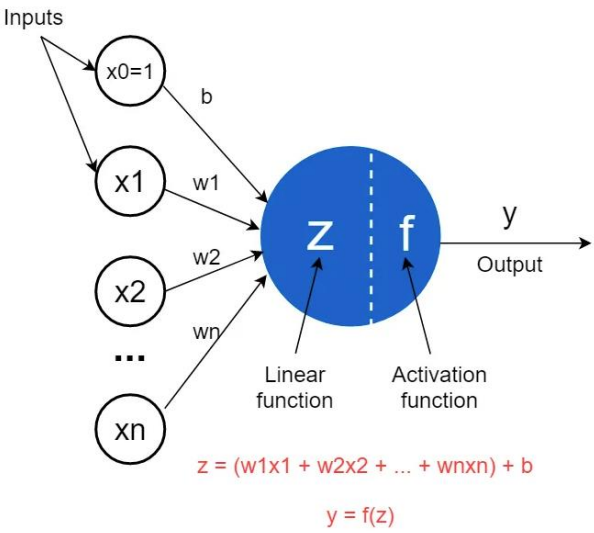
2.2.3 Ιστορική Ανασκόπηση των νευρωνικών δικτύων

Η ιστορία των νευρωνικών δικτύων είναι αρκετά μεγάλη και έχει αφετηρία την δεκαετία του 1940. Παρακάτω γίνεται ανασκόπηση των σημαντικότερων γεγονότων:

- 1943: Ο Walter Pitts και ο Warren McCulloch δημοσίευσαν το άρθρο με τίτλο «A logical calculus of the ideas immanent in nervous activity». Η έρευνα αυτή επικεντρώνεται στην κατανόηση και ανάπτυξη πολύπλοκων υπολογιστικών μοτίβων όπως αυτών που βρίσκονται στον ανθρώπινο εγκέφαλο. Η κυριότερη ιδέα που προέκυψε από αυτήν την εργασία ήταν η χρήση Boolean συναρτήσεων, δηλαδή 0/1 ή αληθές/ψευδές, στις μονάδες κατωφλιού ή εισόδου με κατάλληλα βάρη (McCulloch & Pitts, 1943).
- 1958: Ο F. Rosenblatt στο άρθρο του «The perceptron: A probabilistic model for information storage and organization in the brain» αναπτύσσει το μοντέλο perceptron εισάγοντας βάρη στις εξισώσεις των Pitts και McCulloch. Ο Rosenblatt κατέστησε ένα υπολογιστικό σύστημα της εποχής ικανό να ξεχωρίζει κάρτες που είναι σημαδεμένες αριστερά από κάρτες σημαδεμένες στα δεξιά αντίστοιχα (Rosenblatt, 1958).
- 1974: Η μέθοδος της οπισθοδρόμησης (Backpropagation) έχει εμφανιστεί με πολλούς ερευνητές να έχουν συμβάλλει στην ανάπτυξής της. Ο Paul Werbos στο άρθρο του «Beyond regression: new tools for prediction and analysis in the behavioral sciences» εισάγει την εφαρμογή της οπισθοδρόμησης στα νευρωνικά δίκτυα στο πλαίσιο της διδακτορικής του διατριβής (Werbos, 1974).
- 1989: Ο Yann LeCun δημοσίευσε το άρθρο «Backpropagation Applied to Handwritten Zip Code Recognition» στο οποίο εξηγεί πως η μέθοδος της οπισθοδρόμησης και η χρήση περιορισμών σε αυτήν μπορεί να χρησιμοποιηθεί για εκπαίδευση αλγορίθμων. Στην έρευνά του υλοποιεί ένα νευρωνικό δίκτυο το οποίο είναι ικανό να αναγνωρίζει χειρόγραφα ψηφία ταχυδρομικού κώδικα (LeCun et al., 1989).

2.2.4 Τεχνητός νευρώνας

Το μοντέλο του απλού βιολογικού νευρώνα πλαισιώνει την βάση για την δημιουργία του τεχνητού νευρώνα. Η πρώτη διαφορά εντοπίζεται στην είσοδο, για τους βιολογικούς νευρώνες η είσοδος είναι ένας ηλεκτρικός παλμός ενώ η είσοδος που λαμβάνει ένας τεχνητός νευρώνας είναι συνεχής μεταβλητές. Στους βιολογικούς νευρώνες οι συνάψεις αντιστοιχίζονται με τις συνδέσεις των τεχνητών νευρώνων. Κάθε τέτοιος σύνδεσμος συσχετίζεται με ένα αριθμητικό βάρος το οποίο καθορίζει το πρόσημο και την ισχύ της σύνδεσης. Το σώμα του τεχνητού νευρώνα χωρίζεται σε δυο μέρη. Στο πρώτο μέρος βρίσκεται η μονάδα αθροίσματος (sum), η οποία προσθέτει την είσοδο από τα καινούργια βάρη των σημάτων προκειμένου να παραχθεί μια ποσότητα. Στο δεύτερο μέρος βρίσκεται η συνάρτηση ενεργοποίησης (activation function), η οποία λειτουργεί ως φίλτρο και παράγει την τελική τιμή του σήματος εξόδου (Russell & Norvig, 2010). Ένα σπουδαίο χαρακτηριστικό των νευρώνων είναι η μοναδικότητα της τελικής τιμής εξόδου η οποία εξαρτάται αποκλειστικά και μόνο από τον ίδιο νευρώνα στον οποίο υπολογίστηκε. Στην συνέχεια, η κατάσταση του νευρώνα καθορίζεται από το πρόσημο της τελικής τιμής.



Εικόνα 4: Απεικόνιση Τεχνητού Νευρώνα. Πηγή towardsdatascience.com

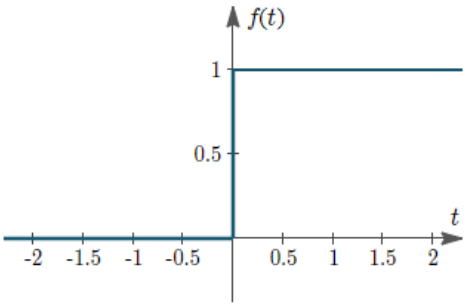
2.2.5 Συναρτήσεις Ενεργοποίησης

Η τελική τιμή εξόδου ενός νευρώνα εξαρτάται σε τεράστιο βαθμό από τις συναρτήσεις ενεργοποίησης. Η απουσία τέτοιων συναρτήσεων θα είχε ως αποτέλεσμα το σήμα εξόδου να ήταν απλώς μια πολυωνυμική συνάρτηση (Rojas, 2013). Ταυτόχρονα οι συναρτήσεις ενεργοποιήσεις έχουν καθοριστικό ρολό την αποτελεσματικότητα και την ακρίβεια των νευρωνικών δικτύων (Sharma et al., 2020). Υπάρχουν αρκετές συναντήσεις ενεργοποιήσεις εκ των οποίων οι σημαντικότερες περιγράφονται λεπτομερώς παρακάτω:

- Βηματική Συνάρτηση:

Η βηματική αποτελεί την απλούστερη συνάρτηση ενεργοποίησης και χρησιμοποιείται σε μοντέλα δομημένα από έναν νευρώνα. Ορίζεται ως μια τμηματική και ασυνεχής συνάρτηση με σταθερές τιμές y για ένα δεδομένο διάστημα τιμών x. Το βασικό μειονέκτημα της βηματικής συνάρτησης είναι ότι η παράγωγος της απειρίζεται (Sharma et al., 2020; Sibi et al., 2005). Η συνάρτηση περιγράφεται από τον μαθηματικό τύπο:

$$f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (1)$$

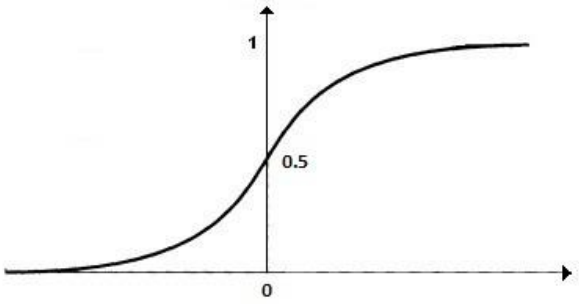


Εικόνα 5: Απεικόνιση Βηματικής Συνάρτησης.

- Σιγμοειδής Συνάρτηση:

Η σιγμοειδής θεωρείται η πιο διαδεδομένη συνάρτηση ενεργοποίησης επειδή αποτελεί προσέγγιση της βηματικής. Είναι συνεχής, παραγωγίσιμη και γνησίως αύξουσα στο πεδίο ορισμού της. Η σιγμοειδής παίρνει τιμές από 0 έως 1 γεγονός που επιτρέπει την ενεργή ή ανενεργή κατάσταση του νευρώνα. Παρόλα αυτά, η σιγμοειδής δεν είναι συμμετρική ως προς την αρχή των αξόνων με αποτέλεσμα τα πρόσημα όλων των τιμών εξόδου να είναι όμοια (Sharma et al., 2020; Sibi et al., 2005). Η συνάρτηση περιγράφεται από τον μαθηματικό τύπο:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2)$$



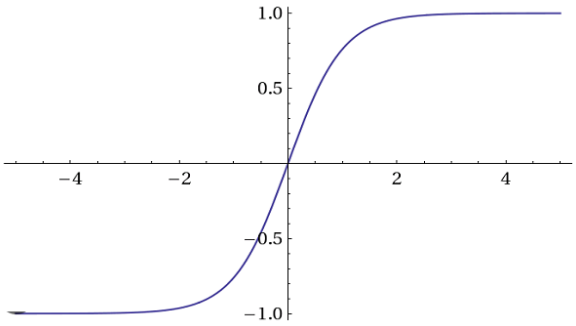
Εικόνα 6: Απεικόνιση Σιγμοειδούς Συνάρτησης.

- Υπερβολική Εφαπτομένη Συνάρτηση:

Η υπερβολική εφαπτομένη παρουσιάζει πολλές ομοιότητες με την σιγμοειδή συνάρτηση, δηλαδή είναι συνεχής και παραγωγίσιμη στο πεδίο ορισμού της. Παίρνει τιμές από -1 έως 1, είναι συμμετρική ως προς την αρχή των αξόνων και η χρήση της είναι προτιμότερη από την

σιγμοειδή συνάρτηση (Sharma et al., 2020; Sibi et al., 2005). Η συνάρτηση περιγράφεται από τον μαθηματικό τύπο:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3)$$

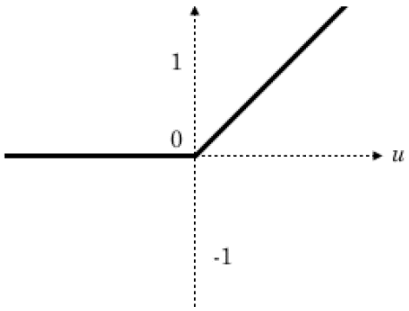


Εικόνα 7: Απεικόνιση Υπερβολικής Εφαπτομένης Συνάρτησης.

- Ανορθωμένη Γραμμική Συνάρτηση Ράμπας (ReLu):

Η συνάρτηση ράμπας θεωρείται πιο αποδοτική από τις προηγούμενες συναρτήσεις αφού έχει αποδειχτεί ότι μπορεί να εκπαιδεύσει ένα δίκτυο γρηγορότερα και αποτελεσματικότερα. Αυτό οφείλεται στο γεγονός ότι οι νευρώνες δεν ενεργοποιούνται την ίδια χρονική στιγμή. Ωστόσο, το πρόβλημα που εμφανίζεται στην συνάρτηση ράμπας είναι η αστοχία εκπαίδευσης νευρώνων όταν το άθροισμα εξόδου είναι μηδενικό ή αρνητικό (Sharma et al., 2020; Sibi et al., 2005). Η συνάρτηση περιγράφεται από τον μαθηματικό τύπο:

$$f(x) = \max(0, x) \quad (4)$$



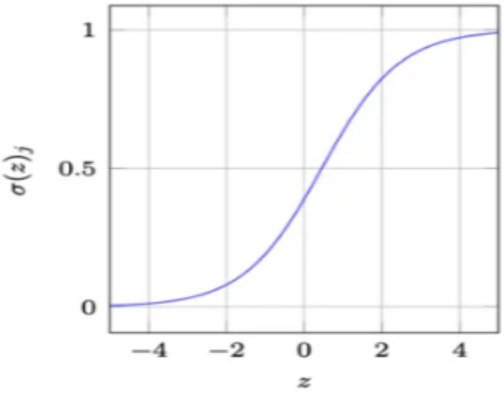
Εικόνα 8: Απεικόνιση Συνάρτησης Ράμπας (ReLu).

- Συνάρτηση SoftMax

Η συνάρτηση SoftMax αποτελεί έναν συνδυασμό πολλών σιγμοειδών συναρτήσεων. Όπως αναφέρθηκε πριν, η σιγμοειδής συνάρτηση επιστρέφει τιμές από 0 έως 1. Πρακτικά η σιγμοειδής μπορεί να χειριστεί μέχρι δυο κλάσεις, δηλαδή να χρησιμοποιηθεί για δυαδική ταξινόμηση. Έτσι, η συνάρτηση SoftMax δίνει την λύση στην αντιμετώπιση προβλημάτων κατηγοριοποίησης με περισσότερες από δυο κλάσεις. Συνήθως χρησιμοποιείται στα τελευταία στρωματά ενός δικτύου υπολογίζοντας τιμές που αντιστοιχούν σε πιθανότητες

κατηγοριοποίησης των επιμέρους κλάσεων (Sharma et al., 2020; Sibi et al., 2005). Η συνάρτηση περιγράφεται από τον μαθηματικό τύπο:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}} \quad \gamma\iota\alpha\ j = 1, \dots, k \qquad (5)$$

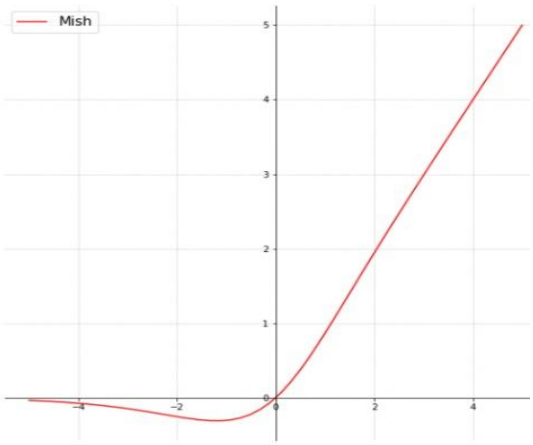


Εικόνα 9: Απεικόνιση Συνάρτησης SoftMax.

- Συνάρτηση Mish

Η συνάρτηση Mish είναι μια ομαλή, συνεχής, μη μονοτονική συνάρτηση ενεργοποίησης. Σχεδιάστηκε αποκλειστικά για προβλήματα βαθιάς μάθησης και μέχρι στιγμής αποτελεί την ιδανικότερη συνάρτηση ενεργοποίησης παρέχοντας καλύτερα αποτελέσματα κανονικοποίησης, εξαιρετικά ομαλές διαβαθμίσεις και σταθερή απόδοση. Υπερτερεί όλων των υπολοίπων συναρτήσεων και έχει αποδειχτεί ότι αυξάνει σημαντικά την απόδοση των δικτύων (Misra, 2020). Το πεδίο ορισμού της είναι [-0.31, ∞) και περιγράφεται με τον μαθηματικό τύπο:

$$f(x) = x \cdot tanh \cdot \big(softplus(x)\big) \quad \text{όπου} \quad softplus(x) = \ln(1 + e^x) \qquad (6)$$



Εικόνα 10: Απεικόνιση Συνάρτησης Mish.

Πίνακας 1: Συνοπτικός Πίνακας Συναρτήσεων Ενεργοποίησης

Συναρτήσεις Ενεργοποίησης	Μαθηματική Συνάρτηση
Βηματική Συνάρτηση	$f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$
Σιγμοειδής Συνάρτηση	$f(x) = \frac{1}{1 + e^{-x}}$
Υπερβολική Εφαπτομένη Συνάρτηση	$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$
Ανορθωμένη Γραμμική Συνάρτηση Ράμπας (ReLu)	$f(x) = \max(0, x)$
Συνάρτηση SoftMax	$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}}$
Συνάρτηση Mish	$(x) = x \cdot tanh \cdot \big(softplus(x)\big)$

2.2.6 Δομή και Αρχιτεκτονική των Νευρωνικών Δικτύων

Ο συνδυασμός ενός συνόλου νευρώνων και συνάψεων, δομούν ένα νευρωνικό δίκτυο. Η απόδοσή του σχετίζεται άμεσα με την τοπολογία του δικτύου και τον αλγόριθμο μάθησης που χρησιμοποιείται για τον υπολογισμό των βαρών. Η αρχιτεκτονική ενός νευρωνικού δικτύου καθορίζει τον τρόπο σύνδεσης και τοποθέτησης των νευρώνων, τον αριθμό των νευρώνων και το επίπεδο στο οποίο ανήκουν. Πιο συγκεκριμένα, τα περισσότερα νευρωνικά δίκτυα κατακερματίζονται σε τρεις ομάδες νευρώνων οι οποίες ονομάζονται στρωματά ή επίπεδα: το στρώμα εισόδου (input layer), το κρυφό στρώμα (hidden layer) και το στρώμα εξόδου (output layer) (da Silva et al., 2017):

- Στρώμα Εισόδου.

Το επίπεδο αυτό αποτελεί την είσοδο αρχικών τιμών στο δίκτυο. Τα δεδομένα μεταφέρονται στο επόμενο επίπεδο χωρίς να πραγματοποιηθεί κάποια διαδικασία επεξεργασίας.

- Κρυφό Στρώμα.

Μεταξύ του στρώματος εισόδου και του στρώματος εξόδου, βρίσκεται το κρυφό ή αόρατο στρώμα. Το πλήθος των επίπεδων αυτού του στρώματος δεν είναι σταθερό. Το σύνολο των νευρώνων που συγκροτούν το κρυφό στρώμα υλοποιεί την διαδικασία των υπολογισμών.

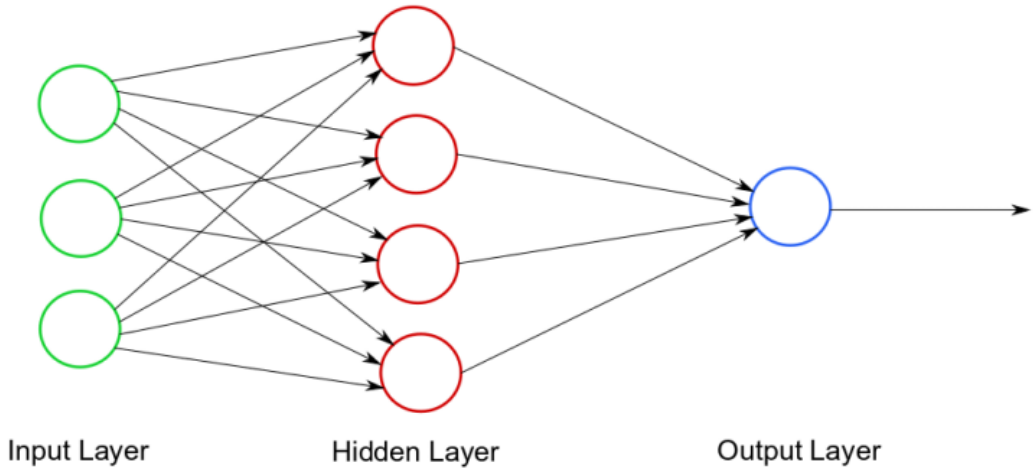
- Στρώμα Εξόδου.

Αυτό το επίπεδο είναι υπεύθυνο για τον έξοδο των τελικών αποτελεσμάτων από την επεξεργασία που πραγματοποιήθηκε στα προηγούμενα στρώματα (da Silva et al., 2017).

Σύμφωνα με τις εκάστοτε αρχιτεκτονικές, τα νευρωνικά δίκτυα διαχωρίζονται σε δυο βασικές κατηγορίες: α) Δίκτυα Εμπρός Τροφοδότησης και β) Αναδρομικά Δίκτυα.

2.2.7 Δίκτυα Εμπρός Τροφοδότησης (Feed – Forward Networks)

Τα δίκτυα εμπρός τροφοδότησης διαθέτουν ένα στρώμα εισόδου, ένα ή πολλά κρυφά στρώματα και ένα στρώμα εξόδου. Η δομική διαμόρφωση αυτών των δικτύων είναι πολύ συγκεκριμένη καθώς οι έξοδοι των κόμβων σε ένα επίπεδο συνδέονται αποκλειστικά με τις εισόδους των επόμενων κόμβων. Ταυτόχρονα, η σύνδεση μεταξύ των επιπέδων εισόδου και εξόδου είναι μονής κατεύθυνσης ενώ δεν υπάρχουν συνδέσεις μεταξύ των νευρώνων σε ένα στρώμα (Ojha et al., 2017). Με αυτόν τον τρόπο το σήμα διαδίδεται από την είσοδο προς την έξοδο του δικτύου. Η εσωτερική κατάσταση ενός δικτύου εμπρός τροφοδότησης καθορίζεται από τα βάρη του αφού στην πραγματικότητα ένα τέτοιο δίκτυο αντιστοιχίζει την είσοδό του σε συνάρτηση με την έξοδό του (Russell & Norvig, 2010). Τα πιο δημοφιλή δίκτυα που βασίζονται στην συγκεκριμένη αρχιτεκτονική είναι το πολυεπίπεδο Perceptron (Multilayer Perceptron, MLP) και τα δίκτυα ακτινικής βάσης (Radial Basis Functions, RBF) (da Silva et al., 2017).

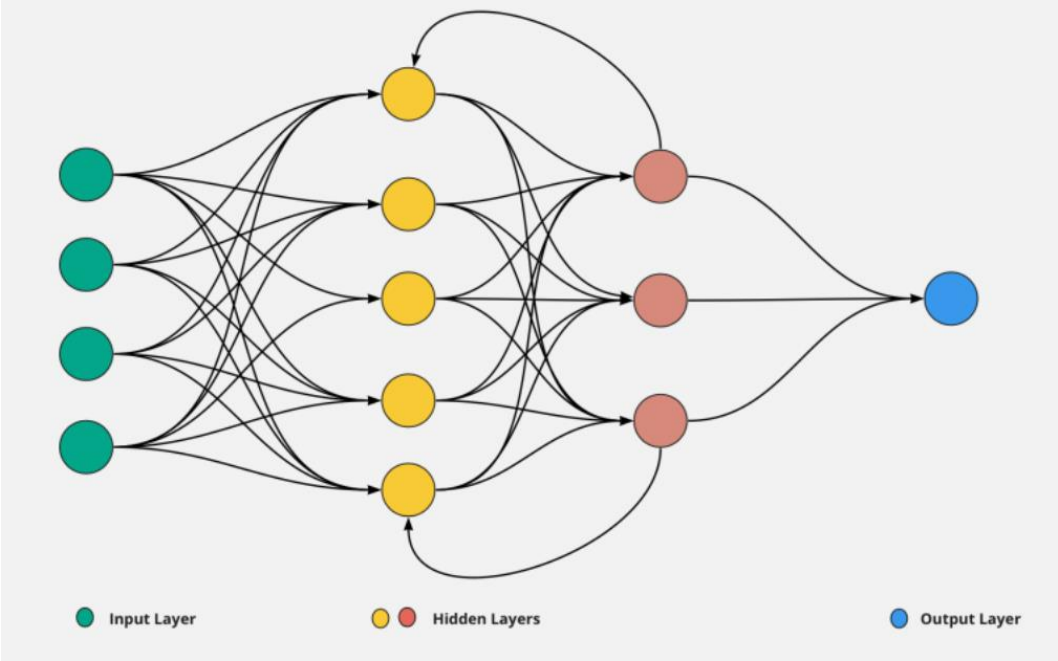


Εικόνα 11: Απεικόνιση Απλού Δικτύου Πρόσθιας Τροφοδότησης. Πηγή kdnuggets.com

2.2.8 Αναδρομικά Δίκτυα (Recurrent Networks)

Τα αναδρομικά δίκτυα είναι ισοδύναμα με τα δίκτυα εμπρός τροφοδότησης αλλά ταυτόχρονα παρουσιάζουν σημαντικές διάφορες. Σε ένα αναδρομικό δίκτυο μπορεί να υπάρχουν συνδέσεις μεταξύ των νευρώνων σε ένα στρώμα (Rojas, 2013), γεγονός που επιτρέπει οποιαδήποτε κρυφή μονάδα να αλληλοεπιδράσει με μια άλλη. Η σχηματική απεικόνιση των αναδρομικών δικτύων συνήθως μοιάζει με κυκλικούς γράφους, επομένως η ροή της πληροφορίας γίνεται με επαναλαμβανόμενο τρόπο. Πιο αναλυτικά, τα συγκεκριμένα δίκτυα έχουν την δυνατότητα της ανάδρασης, δηλαδή μπορούν να τροφοδοτήσουν τις εισόδους τους με τα σήματα των εξόδων τους. Αυτό σημαίνει ότι η κατάσταση του δικτύου είναι μεταβλητή, είτε σταθερή είτε ασταθής. Ένα σημαντικό χαρακτηριστικό των αναδρομικών δικτύων είναι η ικανότητα ανάπτυξης βραχυπρόθεσμης μνήμης, εφόσον η απόκριση του δικτύου για μια δεδομένη είσοδο εξαρτάται από την αρχική κατάστασή του (Russell & Norvig, 2010). Οι παραπάνω ιδιότητες καθιστούν τα αναδρομικά δίκτυα κατάλληλα για την επεξεργασία σειριακών και ακολουθιακών δεδομένων

αναδεικνύοντας τα ως ισχυρά εργαλεία για εξαγωγή και επιλογή χαρακτηριστικών, πρόβλεψη και ταξινόμηση δεδομένων (Basegmez, 2014).



Εικόνα 12: Απεικόνιση Απλού Αναδρομικού Δικτύου. Πηγή dataaspirant.com

2.2.9 Μάθηση & Διαδικασία Εκπαίδευσης

Ένα από τα πιο σημαντικά χαρακτηριστικά των νευρωνικών δικτύων είναι ικανότητά τους να μαθαίνουν από τα δεδομένα που δέχονται. Αφού το δίκτυο μάθει την σχέση μεταξύ της εισόδου και της εξόδου, μπορεί να γενικεύσει την λύση, δηλαδή να παράγει μια είσοδο που βρίσκεται κοντά στην επιθυμητή έξοδο. Επομένως, η διαδικασία εκπαίδευσης μεταφράζεται ως η προσαρμογή της τιμής των συναπτικών βαρών μεταξύ των συνδέσεων, γεγονός που καθορίζει την συμπεριφορά των νευρώνων και του δικτύου (Rojas, 2013). Το άθροισμα των βημάτων που απαιτείται για την εκπαίδευση του δικτύου, ονομάζεται αλγόριθμος μάθησης. Η επιλογή ενός τέτοιου αλγορίθμου εξαρτάται από την αρχιτεκτονική του δικτύου. Βασικός στόχος είναι η μείωση του σφάλματος μεταξύ της πραγματικής τιμής και της επιθυμητής τιμής εξόδου, ανεξάρτητα από το είδος της μάθησης.

Η διαδικασία της εκπαίδευσης χωρίζεται στις παρακάτω βασικές κατηγορίες:

1. Μάθηση με επίβλεψη (Supervised Learning)

   Σύμφωνα με την συγκεκριμένο μέθοδο μάθησης, το μοντέλο τροφοδοτείται με ένα σύνολο δεδομένων εισόδου και εξόδου. Πιο συγκεκριμένα, κάθε δείγμα εκπαίδευσης αποτελείται από πρότυπα εισόδου και τα αντίστοιχα επιθυμητά αποτελέσματα (Sah, 2020). Τα πρότυπα είναι συνήθως διανύσματα τα οποία αποτελούν χαρακτηριστικά του δείγματος. Κατά την διαδικασία της εκπαίδευσης, τα συναπτικά βάρη αρχικοποιούνται σε

τυχαίες τιμές και στην πορεία ενημερώνονται συνεχώς σύμφωνα με την απόκλιση των παραγόμενων τιμών από τις επιθυμητές εξόδους. Το 1948 ο Donald Hebb, πρότεινε την πρώτη στρατηγική εποπτευόμενης μάθησης, εμπνευσμένη από τις νευρολογικές παρατηρήσεις του (da Silva et al., 2017). Τα προβλήματα μάθησης με επίβλεψη χωρίζονται σε δυο μεγάλες κατηγορίες. Στα προβλήματα ταξινόμησης, οπού οι στόχοι είναι διακριτές τιμές και συσχετίζονται με κλάσεις αντικειμένων, και στα προβλήματα παλινδρόμησης οπού οι στόχοι είναι συνεχείς τιμές και συσχετίζονται με ποσότητες (Géron, 2022).

2. Μάθηση χωρίς επίβλεψη (Unsupervised Learning)

Αυτός ο τύπος μάθησης αποτελεί την αντίθετη περίπτωση της μάθησης με επίβλεψη. Αναλυτικότερα, το μοντέλο χρησιμοποιεί για την εκπαίδευσή του ένα σύνολο από δεδομένα που απαρτίζονται μόνο από πρότυπα εισόδου (Sah, 2020). Έτσι, η συγκεκριμένη διαδικασία μάθησης δεν απαιτεί την γνώση των επιθυμητών εξόδων. Ο στόχος της εκπαίδευσης του μοντέλου είναι εύρεση συσχετίσεων και μοτίβων μεταξύ των δεδομένων, που βασίζονται στις ιδιότητές τους (da Silva et al., 2017). Οι σημαντικότεροι αλγόριθμοι που υπάγονται σε αυτήν την κατηγορία, είναι οι αλγόριθμοι ομαδοποίησης, οι αλγόριθμοι μείωσης διαστάσεων και αλγόριθμοι εύρεσης κανόνων συσχέτισης (Géron, 2022).

3. Μάθηση με ενίσχυση (Reinforcement Learning)

Αυτή η διαδικασία μάθησης συναντάται σε μοντέλα που επιχειρούν να μάθουν μέσα από την άμεση αλληλεπίδρασή τους με το περιβάλλον. Θεωρείται ότι αποτελεί παραλλαγή της μάθησης με επίβλεψη, αφού αναλύει συνεχώς την διαφορά μεταξύ της παραγόμενη εξόδου και της επιθυμητής εξόδου (da Silva et al., 2017; Sah, 2020). Η ενισχυτική μάθηση εφαρμόζεται σε κατηγορίες προβλημάτων όπου δεν υπάρχει γνώση για τις ενέργειες που πρέπει να πραγματοποιηθούν προκειμένου να εκτελεστεί μια εργασία. Τα μοντέλα ονομάζονται πράκτορες και η διαδικασία εκμάθησής τους γίνεται με δοκιμή και αξιολόγηση σφαλμάτων. Εάν η απόκριση του συστήματος αξιολογηθεί ως ικανοποιητική τότε τα συναπτικά βάρη αυξάνονται με σκοπό να βελτιωθεί η συνολική συμπεριφορά του (Géron, 2022).

2.3 Αλγόριθμοι Βελτιστοποίησης

2.3.1 Αλγόριθμος Κατάβασης Δυναμικού (Gradient Descent Rule)

Όπως αναφέρθηκε παραπάνω, ένα δίκτυο χαρακτηρίζεται ως επαρκώς εκπαιδευμένο εφόσον το αποτέλεσμα σύγκρισης της παραγόμενης εξόδου και της επιθυμητής εξόδου είναι αποδεκτό. Αρχικά, η αξιολόγηση της απόδοσης του δικτύου υπολογίζεται από την συνάρτηση κόστους (cost function), η οποία ορίζεται μαθηματικά ως εξής:

$$C(w,b) = \frac{1}{2n}\sum_{x}\|y(x) - a\|^2 \quad (7)$$

Όπου το w συνοψίζει το σύνολο των βαρών, b το σύνολο των πολώσεων, n το σύνολο των εισόδων εκπαίδευσης, y είναι η επιθυμητή έξοδος και α είναι το διάνυσμα τιμών εξόδου για τις x τιμές εισόδου. Η παραπάνω συνάρτηση ονομάζεται τετραγωνική συνάρτηση κόστους (quadratic) γνωστή και ως συνάρτηση ελαχίστων τετραγώνων (Nielsen, 2015).

Η μέθοδος κατάβασης δυναμικού, είναι ένας τρόπος ελαχιστοποίησης της συνάρτησης κόστους C(w,b) ενός μοντέλου, ενημερώνοντας τις παραμέτρους στην αντίθετη κατεύθυνση της κλίσης της συνάρτησης (Ruder, 2017). Η κλασσική μέθοδος κατάβασης δυναμικού σχεδιάστηκε και παρουσιάστηκε επίσημα από τον Cauchy το 1847 (Goh et al., 2012).

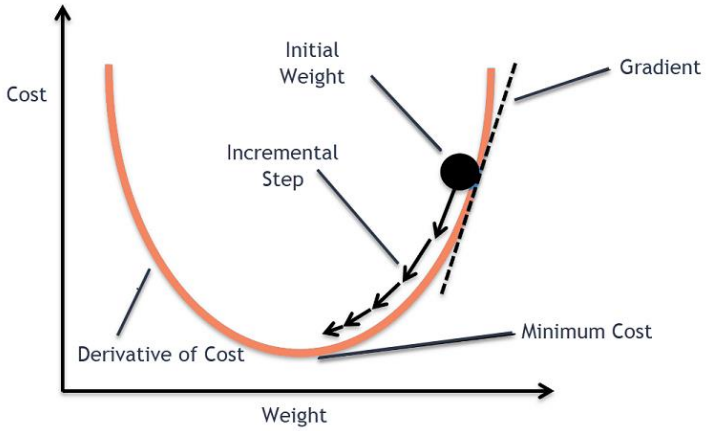
Ουσιαστικά, για να επιτευχθεί αυτή η ελαχιστοποίηση πρέπει να βρεθεί το σημείο στο οποίο η συνάρτηση κόστους παρουσιάζει ολικό ελάχιστο. Ο τρόπος για να λυθεί αυτό το πρόβλημα είναι να υπολογιστεί η κλίση και έπειτα να βρεθεί το ακρότατο της συνάρτησης. Η μέθοδος κατάβασης δυναμικού αρχικοποιείται σε ένα τυχαίο σημείο και στην συνέχεια εγγυάται ότι κινείται στην κατεύθυνση κατά την οποία η συνάρτηση κόστους μειώνεται με κάθε επανάληψη. Σε κάθε επανάληψη του αλγορίθμου, υπολογίζεται η κλίση, δηλαδή η μερική παράγωγος της συνάρτησης για κάθε παράμετρό της. Το διάνυσμα της κλίσης δείχνει προς την κατεύθυνση της ανόδου, επομένως αντίθετα βρίσκεται η κατεύθυνση της απότομης κατάβασης (Nielsen, 2015).

Ο ρυθμός εκπαίδευσης *γ* καθορίζει την ταχύτητα με την οποία ο αλγόριθμος κινείται προς το ελάχιστο και εξαρτάται από το μέγεθος του βήματος. Η επιλογή του *γ* θα πρέπει από την μια πλευρά να έχει μικρή τιμή ώστε η προσέγγιση να είναι σωστή, και από την άλλη πλευρά η ίδια τιμή να προσδίδει έναν σχετικά γρήγορο ρυθμό εκπαίδευσης. Ο υπολογισμός των συναπτικών βαρών και των πολώσεων πραγματοποιείται σύμφωνα με τις παρακάτω εξισώσεις (Basegmez, 2014):

$$w_{new} = w_{old} - \gamma \frac{\partial C}{\partial w} \quad (8)$$

$$b_{new} = b_{old} - \gamma \frac{\partial C}{\partial b} \quad (9)$$

Η διαδικασία επαναλαμβάνεται έως ότου η συνάρτηση κόστους συγκλίνει στο ελάχιστο, υποδεικνύοντας ότι το μοντέλο έχει φτάσει στην βέλτιστη κατάσταση.



Εικόνα 13: Απεικόνιση Αλγορίθμου Κατάβασης Δυναμικού. Πηγή analyticsvidhya.com

Υπάρχουν τρεις παραλλαγές της μεθόδου κατάβασης δυναμικού, οπού η βασική τους διαφορά εντοπίζεται στον όγκο των δεδομένων που χρησιμοποιούν για τον υπολογισμό της συνάρτησης κόστους.

1. Κατάβαση Δυναμικού Δέσμης (Batch Gradient Descent)

Η κατάβαση δυναμικού δέσμης υπολογίζει την διαβάθμιση της συνάρτησης κόστους για όλο το σύνολο δεδομένων εκπαίδευσης και στην συνέχεια εκτελεί μόνο μια ενημέρωση. Το πρόβλημα που δημιουργείται είναι ο μεγάλος χρόνος εκπαίδευσης του μοντέλου και η υπερβολική κατανάλωση πόρων σε περίπτωση που οι τιμές εισόδου είναι πολλές (Ruder, 2017).

2. Στοχαστική Κατάβαση Δυναμικού (Stochastic Gradient Descent)

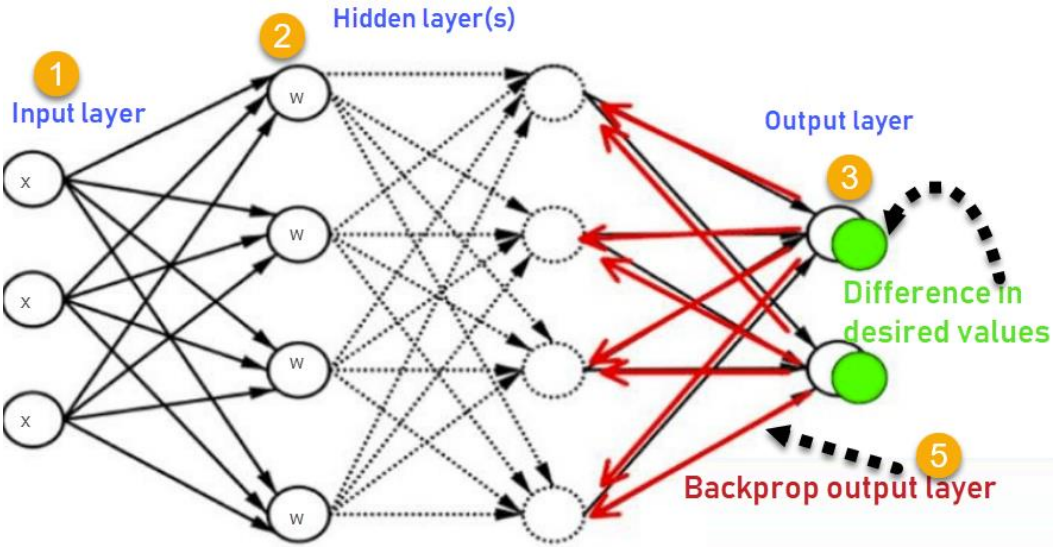
Η στοχαστική κατάβαση δυναμικού ενημερώνει τις παραμέτρους του μοντέλου υπολογίζοντας την διαβάθμιση ενός παραδείγματος εκπαίδευσης κάθε φορά. Πρώτα πραγματοποιεί μια στοχαστική εκτίμηση της πραγματικής λύσης και μετά εκτελεί μια ενημέρωση χωρίς πλεονάζοντες υπολογισμούς (Basegmez, 2014). Ακόμα η ιδιότητα της στοχαστικής επιλογής μειώνει την πιθανότητα παγίδευσης της διαδικασίας εκπαίδευσης σε τοπικά ελάχιστα (Ruder, 2017).

3. Κατάβαση Δυναμικού Μικρής Δέσμης (Mini-Batch Gradient Descent)

Η κατάβαση δυναμικού μικρής δέσμης ενημερώνει τις παραμέτρους του μοντέλου χρησιμοποιώντας ένα υποσύνολο των τιμών εισόδου κάθε φορά. Ο αριθμός των δειγμάτων εκπαίδευσης για μια επανάληψη ονομάζεται μέγεθος παρτίδας. Η συγκεκριμένη μέθοδος συνδυάζει τα πλεονεκτήματα τόσο της κατάβασης δέσμης όσο και της στοχαστικής κατάβασης, επομένως είναι υπολογιστικά αποδοτικότερη και οδηγεί σε σταθερή σύγκλιση (Ruder, 2017).

2.3.2 Αλγόριθμος Ανάστροφης Μετάδοσης Λάθους (Back Propagation)

Η ανάστροφη μετάδοση λάθους αποτελεί την πιο γνωστή μέθοδο εκπαίδευσης νευρωνικών δικτύων πολλών επιπέδων. Αυτή η τεχνική βασίζεται στην μέθοδο της κατάβασης δυναμικού και ελαχιστοποιεί την συνάρτηση κόστους με αναδρομικό τρόπο (Leung & Haykin, 1991). Η υλοποίηση του αλγορίθμου πραγματοποιείται σε δυο στάδια. Στο πρώτο στάδιο εισάγονται τα δεδομένα εκπαίδευσης και στην συνέχεια το επίπεδο εισόδου παράγει αποτελέσματα τα οποία με την σειρά τους αποτελούν είσοδο για το επόμενο κρυφό επίπεδο. Η διαδικασία αυτή επαναλαμβάνεται μέχρι το επίπεδο εξόδου και ονομάζεται πρόσθιο πέρασμα (forward pass). Στο δεύτερο στάδιο, γίνεται ο υπολογισμός του συνολικού σφάλματος το οποίο διαδίδεται προς τα πίσω. Έτσι η διαδικασία προσαρμογής των συναπτικών βαρών γίνεται από το επίπεδο εξόδου προς το επίπεδο εισόδου και ονομάζεται ανάστροφο πέρασμα (backward pass) (Nielsen, 2015; Russell & Norvig, 2010; Vlahavas et al., 2020).



Εικόνα 14: Απεικόνιση Αλγορίθμου Ανάστροφης Μετάδοσης Λάθους. Πηγή niser.ac.in

Πρώτα από όλα, πρέπει να οριστούν οι σχέσεις εισόδου (4) και εξόδου (5) ενός νευρώνα μεμονωμένα (Vlahavas et al., 2020):

$$input_j = \sum_{i=1}^{n} v_{ij}\, x_i \quad (10)$$

$$z_j = f(input_j) = f(\sum_{i=1}^{n} v_{ij}\, x_i) \quad (11)$$

Όπου το v_{ij} είναι το βάρος της σύνδεσης δυο νευρώνων i και j, το x_i είναι το σήμα εισόδου στον νευρώνα I και η συνάρτηση f είναι η συνάρτηση ενεργοποίησης του νευρώνα. Επομένως, συνολικά για όλους τους νευρώνες στο επίπεδο εξόδου προκύπτουν αντίστοιχα οι σχέσεις (6) για την είσοδο και (7) για την έξοδο (Vlahavas et al., 2020):

$$input_k = \sum_{j=1}^{q} w_{jk}\, z_i \quad (12)$$

$$y_k = f(input_k) = f\left(\sum_{i=1}^{q} w_{jk}\, z_i\right) \quad (13)$$

Την ίδια στιγμή το σφάλμα μπορεί να υπολογιστεί από την σχέση (1) ενώ είναι γνωστό και το επιθυμητό αποτέλεσμα. Επομένως μπορούν να αναπροσαρμοστούν οι τιμές των βαρών μεταξύ του επιπέδου εξόδου και του προηγούμενου κρυφού επιπέδου. Αποδεικνύεται ότι ισχύουν οι σχέσεις (Vlahavas et al., 2020):

$$\Delta w_{jk} = d \cdot \delta_k \cdot z_j \quad \gamma\iota\alpha \quad \delta_k = (t_k - y_k) \cdot f'(input_k) \quad (14)$$

Όπου δ_k είναι ο ρυθμός μεταβολής σφάλματος, δ ο ρυθμός μάθησης, t_k είναι η επιθυμητή είσοδος στον νευρώνα k και f' είναι η πρώτη παράγωγος της συνάρτησης ενεργοποίησης. Με τον ίδιο τρόπο αποδεικνύεται ότι μπορούν να αναπροσαρμοστούν τα βάρη όλων των κρυφών επιπέδων (Vlahavas et al., 2020):

$$\Delta w_{ij} = d \cdot \delta_j \cdot x_i \quad \gamma\iota\alpha \quad \delta_j = f'(input_j) \sum_{k=1}^{m} \delta_k w_{jk} \quad (15)$$

Εν τέλη, γίνεται κατανοητό ότι μπορούν στην αρχή να υπολογιστούν οι καινούργιες τιμές των βαρών στο επίπεδο εξόδου σε σχέση με το προηγούμενο κρυφό επίπεδο και στην συνέχεια μπορούν να υπολογιστούν όλες οι τιμές που συνδέουν κάθε κρυφό επίπεδο με το προηγούμενό του φτάνοντας έτσι στο επίπεδο εισόδου.

2.4 Συνελικτικά Νευρωνικά δίκτυα

2.4.1 Εισαγωγή

Τα συνελικτικά νευρωνικά δίκτυα (ΣΝΔ) ή Convolutional Neural Networks (CNN) είναι βαθιά δίκτυα πολλών στρωμάτων σχεδιασμένα για επεξεργασία δεδομένων μοτίβων (patterns) καθιστώντας τα κατάλληλα για προβλήματα αναγνώρισης εικόνων (Yamashita et al., 2018). Η αρχιτεκτονική των δικτύων σχεδιάστηκε θεωρητικά από τους Hubel & Wiesel, εμπνευσμένοι από τα τμήματα του εγκέφαλου που σχετίζονται με την όραση της γάτας (Hubel & Wiesel, 1959). Αργότερα, στο τέλος της δεκαετίας του ’90 ο LeCun κατάφερε να εκπαιδεύσει τα πρώτα υπολογιστικά μοντέλα για διαφορά προβλήματα αναγνώρισης με εξαιρετικές επιδόσεις για την εποχή (LeCun et al., 1989, 1995).

Η βασική πρόκληση ήταν η εκπαίδευση εικόνων σχετικά μικρού μεγέθους που περιλαμβάνουν εκατοντάδες εικονοστοιχεία. Ένα τυπικό νευρωνικό δίκτυο με τυχαία αρχικοποίηση βαρών δεν θα μπορούσε να εκπαιδευτεί σωστά αφού κάθε εικονοστοιχείο ξεχωριστά (pixel) θα αποτελούσε δεδομένο εισόδου το οποίο για την εκπαίδευσή του απαιτεί έναν νευρώνα στο πρώτο κρυφό επίπεδο. Επομένως η εκπαίδευση χιλιάδων δεδομένων (pixel) συνεπάγεται ένα δίκτυο αποτελούμενο από εκατομμύρια συνδέσεις και συναπτικά βάρη. Η αρχιτεκτονική των συνελικτικών δικτύων επιλύει το παραπάνω πρόβλημα μετατρέποντας τα δεδομένα εικόνων σε διανύσματα δυο διαστάσεων, δηλαδή σε πίνακες (Vlahavas et al., 2020).

Ένα συνελικτικό νευρωνικό δίκτυο περιλαμβάνει τρία βασικά στάδια επεξεργασίας, τα οποία θα αναλυθούν λεπτομερώς παρακάτω (Ajit et al., 2020; Vlahavas et al., 2020; Yamashita et al., 2018; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019):

- Στάδιο Συνέλιξης (convolution):

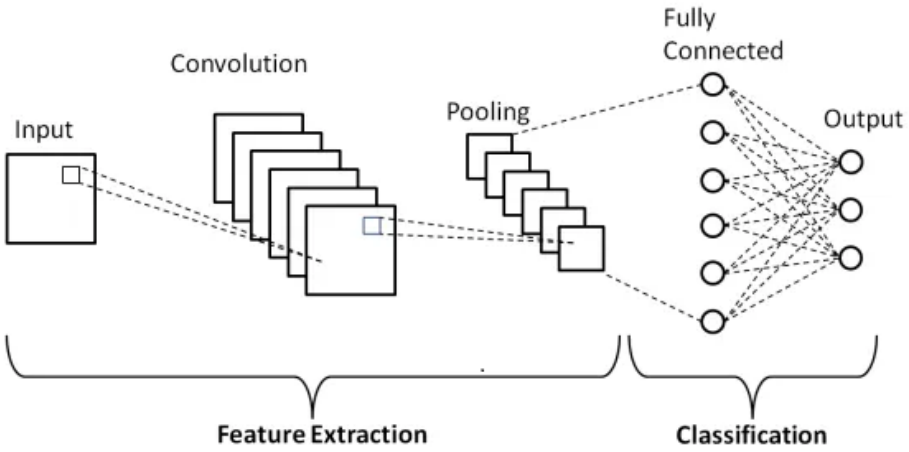
Το στάδιο αυτό περιλαμβάνει επίπεδα συνέλιξης (convolution layers) και ειδικεύεται στον εντοπισμό χαρακτηριστικών λεπτής ή αδρής υφής που βρίσκονται στην εικόνα.

- Στάδιο Υποδειγματοληψίας (sub-sampling/pooling):

Σε αυτό το στάδιο μειώνονται οι διαστάσεις των επιπέδων με σκοπό την εύρεση των περισσότερο σημαντικών χαρακτηριστικών εισόδου.

- Πλήρως Διασυνδεμένα Επίπεδα (fully connected layers):

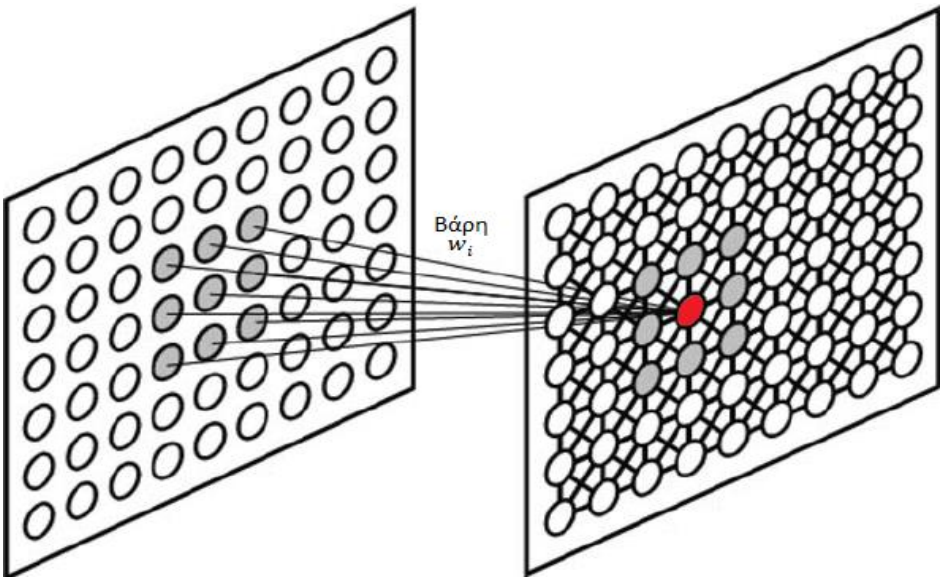
Τα επίπεδα που βρίσκονται στην έξοδο του δικτύου και είναι υπεύθυνα για την ταξινόμηση.



Εικόνα 15: Αρχιτεκτονική ενός συνηθισμένου συνελικτικού νευρωνικού δικτύου. Πηγή medium.com

2.4.2 Στάδιο Συνέλιξης

Στο στάδιο συνέλιξης πραγματοποιείται η βασική διαδικασία εντοπισμού χαρακτηριστικών. Αυτό επιτυγχάνεται με την χρήση φίλτρων των οποίων η μορφή εξαρτάται από το χαρακτηριστικό που αναζητούν. Συνήθως σε κάθε στάδιο συνέλιξης εφαρμόζονται πολλά φίλτρα με αποτέλεσμα να παράγονται αντίστοιχα επίπεδα από νευρώνες, τα οποία έχουν τροφοδοτηθεί με δεδομένα κάποιας συγκεκριμένης περιοχής (Ajit et al., 2020; Vlahavas et al., 2020; Yamashita et al., 2018; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019).



Εικόνα 16: Συνέλιξη μεταξύ δυο διαδοχικών επιπέδων. Πηγή researchgate.net

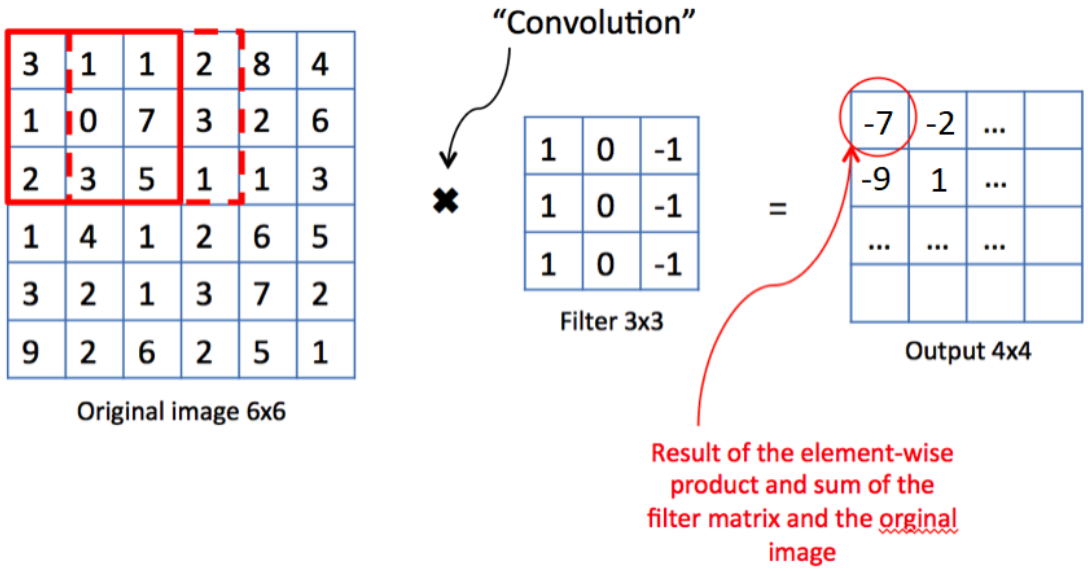
Στην εικόνα 16 διακρίνεται ένα φίλτρο 3x3 στο επίπεδο αριστερά που τροφοδοτεί έναν νευρώνα με χρώμα κόκκινο στο επίπεδο δεξιά. Οι γραμμές που συνδέουν τα δυο επίπεδα αντιστοιχούν στις τιμές των βαρών. Η έξοδος του νευρώνα με χρώμα κόκκινο περιγράφεται από την συνάρτηση (Ajit et al., 2020; Vlahavas et al., 2020):

$$y = f\left(\sum_{i=1}^{9} w_i\, x_i \;+\; b\right) \quad (16)$$

Όπου το w_i είναι το βάρος της σύνδεσης, το x_i είναι το σήμα εισόδου, f είναι η συνάρτηση ενεργοποίησης του νευρώνα και b είναι η τιμή πόλωσης του νευρώνα.

2.4.3 Η πράξη της συνέλιξης

Κάθε επίπεδο έχει διαστάσεις nxn ενώ κάθε φίλτρο έχει διαστάσεις fxf. Το φίλτρο τροφοδοτεί νευρώνες διατρέχοντας την εικόνα με μετατόπιση ένα εικονοστοιχείο (pixel) κάθε φορά.



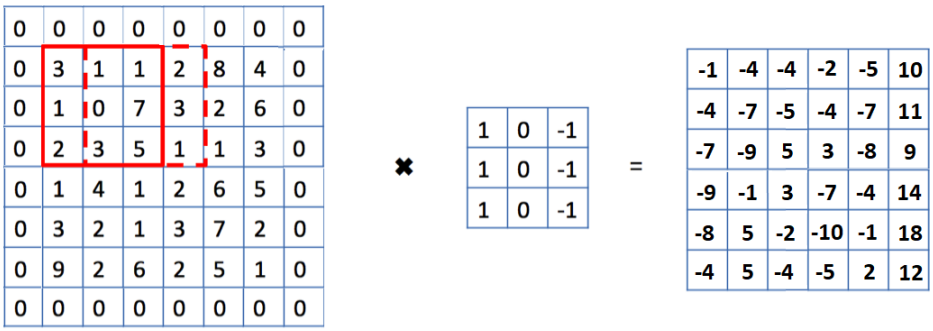
Εικόνα 17: Παράδειγμα υπολογισμού τιμών συνέλιξης. Πηγή kaggle.com

Στην εικόνα 17 υπάρχει ένας πίνακας διαστάσεων 6x6 που αντιστοιχεί σε μια εικόνα, δίπλα βρίσκεται ο πίνακας φίλτρου 3x3 και δεξιά από αυτόν υπάρχει ένας πίνακας 4x4 ο οποίος περιέχει τις τιμές συνέλιξης. Επίσης στον πρώτο πίνακα απεικονίζονται δυο επικαλυπτόμενοι πίνακες 3x3 με χρώμα κόκκινο. Ο υπολογισμός των τιμών γίνεται μεταξύ αυτών των κόκκινων υποπεριοχών και του φίλτρου. Πρώτα πολλαπλασιάζονται τα στοιχεία των αντίστοιχων θέσεων και έπειτα αθροίζονται τα γινόμενα που προκύπτουν. Πιο αναλυτικά τα αποτελέσματα του πίνακα συνέλιξης προκύπτουν ως εξής:

- 3·1 + 1·0 + 1·(-1) + 1·1 + 0·0 + 7·(-1) + 2·1 + 3·0 + 5·(-1) = **-7**
- 1·1 + 1·0 + 2·(-1) + 0·1 + 7·0 + 3·(-1) + 3·1 + 5·0 + 1·(-1) = **-2**

- 1·1 + 0·0 + 7·(-1) + 2·1 + 3·0 + 5·(-1) + 1·1 + 4·0 + 1·(-1) = **-9**
- 0·1 + 7·0 + 3·(-1) + 3·1 + 5·0 + 1·(-1) + 4·1 + 1·0 + 2·(-1) = **1**

Παρόλα αυτά ο συγκεκριμένος τρόπος δημιουργεί μερικά πρόβλημα. Το πρώτο πρόβλημα εντοπίζεται στο γεγονός ότι ο παραγόμενος πίνακας έχει αρκετά μικρότερες διαστάσεις από τον αρχικό. Το δεύτερο και σημαντικότερο πρόβλημα, είναι ότι οι τιμές που βρίσκονται περιφερικά του αρχικού πίνακα συμμετέχουν σε λιγότερους υπολογισμούς με αποτέλεσμα να χάνεται πληροφορία. Τα προβλήματα αυτά αντιμετωπίζονται με μια τεχνική που ονομάζεται padding, κατά την οποία προστίθενται στήλες και γραμμές περιμετρικά από τον αρχικό πίνακα με μηδενικές τιμές (zero – padding) (Vlahavas et al., 2020; Yamashita et al., 2018).



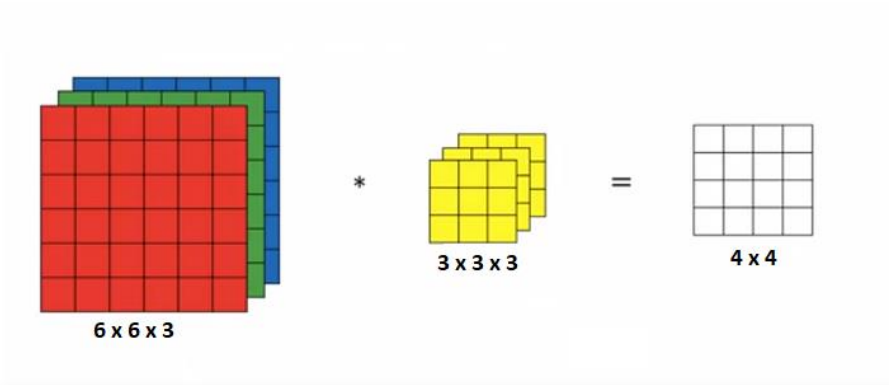
Εικόνα 18: Παράδειγμα υπολογισμού τιμών συνέλιξης με τεχνική padding. Πηγή kaggle.com

Η εικόνα 18 αποτελεί συνέχεια του προηγούμενου παραδείγματος, με την διαφορά ότι τώρα ο αρχικός πίνακας έχει αποκτήσει διαστάσεις 8x8 και έτσι ο τελικός πίνακας συνέλιξης θα κληρονομήσει τις αρχικές διαστάσεις 6x6.

Γενικότερα, έχει αποδειχθεί ότι η τιμή του padding πρέπει να είναι (f-1)/2 ώστε να μην αλλάξει η διάσταση του πίνακα συνέλιξης. Επιπλέον η μετατόπιση του φίλτρου ονομάζεται διασκελισμός (stride) και μπορεί να πάρει τιμές μεγαλύτερες από 1.

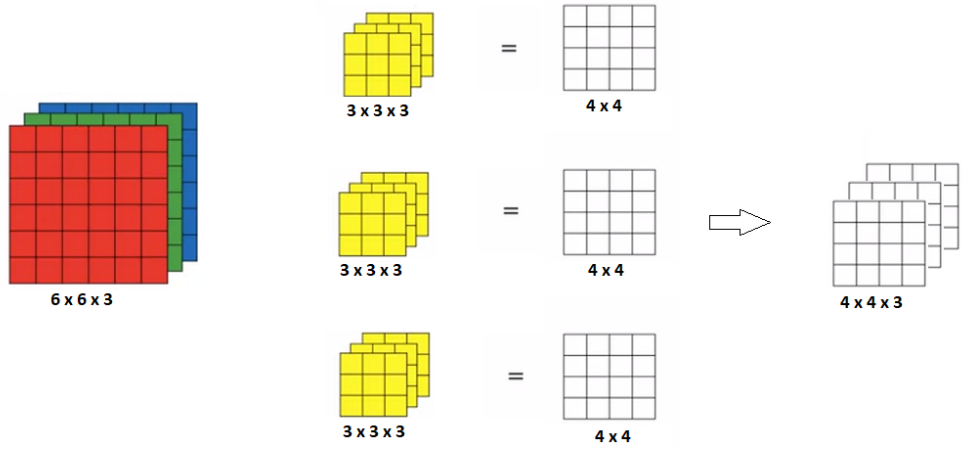
2.4.4 Συνέλιξη σε έγχρωμες εικόνες

Η διαδικασία συνέλιξης σε έγχρωμες εικόνες απαιτεί περισσοτέρους υπολογισμούς. Ο λόγος που συμβαίνει αυτό είναι επειδή η διαδικασία της συνέλιξης πραγματοποιείται ξεχωριστά για καθένα από τα τρία χρωματικά κανάλια της εικόνας (R, G, B). Αυτό σημαίνει ότι το φίλτρο χρησιμοποιείται τρεις φορές παράγοντας αντίστοιχα τρεις πίνακες συνέλιξης όπως φαίνεται στην εικόνα 19. Στην συνέχεια οι πίνακες που προέκυψαν προστίθενται μεταξύ τους δημιουργώντας τον τελικό πίνακα συνέλιξης (Vlahavas et al., 2020).



Εικόνα 19: Παράδειγμα συνέλιξης σε έγχρωμη εικόνα. Πηγή medium.com

Όπως αναφέρθηκε στην αρχή, στον στάδιο συνέλιξης εφαρμόζονται συνήθως περισσότερα από ένα φίλτρα, με σκοπό τον εντοπισμό πολλών χαρακτηριστικών. Στην εικόνα 20 απεικονίζεται η εφαρμογή τριών διαφορετικών φίλτρων τα οποία θα εφαρμοστούν σε κάθε χρωματικό κανάλι παράγοντας τρεις πίνακες 4x4 όπως στο προηγούμενο παράδειγμα. Στην συνέχεια οι παραγόμενοι πίνακες 4x4 προστίθενται και συνθέτουν τον συνολικό πίνακα (Vlahavas et al., 2020).

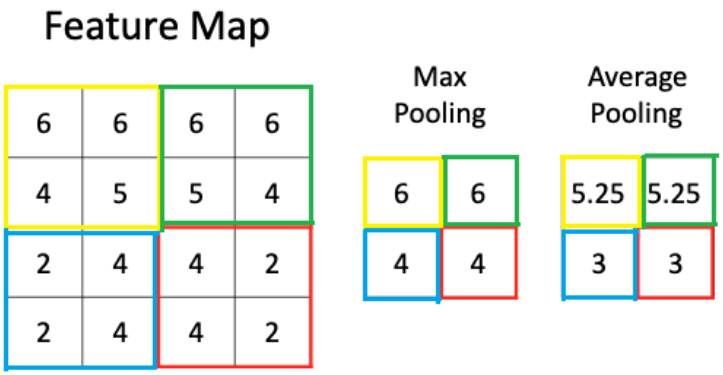


Εικόνα 20: Παράδειγμα συνέλιξης σε έγχρωμη εικόνα με χρήση τριών διαφορετικών φίλτρων. Πηγή medium.com

Οι τιμές των αρχικών επιπέδων 6x6 για τα προηγούμενα παραδείγματα αντιστοιχούν σε κάποια ιδιότητα της εικόνας, όπως για παράδειγμα στην φωτεινότητα των εικονοστοιχείων (pixel). Από την άλλη πλευρά οι τιμές που βρίσκονται στα φίλτρα αλλά και οι τιμές πόλωσης δεν είναι προκαθορισμένες. Οι τιμές αυτές προκύπτουν κατά την εκπαίδευση του δικτύου, η οποία πραγματοποιείται με τον αλγόριθμο ανάστροφης μετάδοσης λάθους (Back Propagation) (Leung & Haykin, 1991; Vlahavas et al., 2020; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019).

2.4.5 Στάδιο Υποδειγματοληψίας

Το στάδιο υποδειγματοληψίας (sub-sampling) βρίσκεται μετά από ένα ή περισσότερα επίπεδα συνέλιξης. Σκοπός αυτού του σταδίου είναι η μείωση ή αλλιώς συμπίεση διαστάσεων των επιπέδων οδηγώντας σε πράξεις με λιγότερους υπολογισμούς (ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019). Ταυτόχρονα, εξαλείφονται οι λεπτομέρειες και διατηρούνται μόνο τα έντονα χαρακτηριστικά των επιπέδων. Πρέπει να τονιστεί ότι η διαδικασία της υποδειγματοληψίας θα εφαρμοστεί σε όλα τα παραγόμενα επίπεδα που δημιουργήθηκαν από το στάδιο συνέλιξης. Τα είδη των υπολογισμών που πραγματοποιούνται σε αυτό το στάδιο είναι η υποδειγματοληψία μέσης τιμής (average pooling) και η υποδειγματοληψία μέγιστης τιμής (max pooling) (Vlahavas et al., 2020; Yamashita et al., 2018; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019). Έχει αποδειχτεί ότι η υποδειγματοληψίας μέγιστης τιμής προσφέρει καλύτερη αναπαράσταση των χαρακτηριστικών και είναι προτιμότερη από αυτήν της μέσης τιμής. Στην εικόνα 21, απεικονίζονται υπολογισμοί για έναν τυχαίο πίνακα 4x4. Για παράδειγμα, η υποπεριοχή με κίτρινο χρώμα υπολογίζεται στον πίνακα Max Pooling σύμφωνα με την συνάρτηση MAX(6, 6, 4, 5) = 6, ενώ στον πίνακα Average Pooling υπολογίζεται σύμφωνα με την συνάρτηση AVG = (6 + 6 + 4 + 5)/4 = 5,25.



Εικόνα 21: Παράδειγμα υπολογισμού υποδειγματοληψίας max pooling και average pooling σε πίνακα 4x4 με διασκελισμό 2. Πηγή kaggle.com

2.4.6 Πλήρως Διασυνδεμένα Επίπεδα

Τα πλήρως διασυνδεμένα επίπεδα βρίσκονται πάντοτε στο τελικό στάδιο επεξεργασίας ενός συνελικτικού δικτύου και είναι υπεύθυνα για την ταξινόμηση. Αφού πραγματοποιηθεί η εξαγωγή χαρακτηριστικών από τα στάδια συνέλιξης και υποδειγματοληψίας, ο ταξινομητής υλοποιεί την αντιστοίχιση του αντικειμένου με την κλάση του (Vlahavas et al., 2020). Κάθε πλήρως διασυνδεμένο στρώμα ακολουθείται από μια συνάρτηση ενεργοποίηση, όπως η Relu (Yamashita et al., 2018). Παρόλα αυτά η συνάρτηση ενεργοποίησης του τελευταίου στρώματος διαφέρει από τα προηγούμενα. Το τελικό στρώμα συνήθως έχει τον ίδιο αριθμό κόμβων εξόδου με τον αριθμό κλάσεων και ακόμα οι τιμές των κόμβων πρέπει να κυμαίνονται μεταξύ 0 και 1 προσδίδοντας την έννοια της πιθανότητας. Η πιο κατάλληλη συνάρτηση ενεργοποίησης για ταξινόμηση πολλών κλάσεων θεωρείται η συνάρτηση SoftMax (Yamashita et al., 2018).

2.5 Μεταφορά Μάθησης (Transfer Learning)

Τα συνελικτικά νευρωνικά δίκτυα αποτελούν την βάση για τα δημοφιλέστερα μοντέλα αναγνώρισης αντικειμένων, τα οποία θα αναλυθούν στο επόμενο κεφάλαιο. Η εκπαίδευση αυτών των μοντέλων αποτελεί μεγάλη πρόκληση. Αφενός ο όγκος των δεδομένων που απαιτείται είναι τεράστιος ενώ αρκετές φορές μη διαθέσιμος και αφετέρου ο χρόνος εκπαίδευσης από αρχικό στάδιο κυμαίνεται από μέρες έως εβδομάδες. Ο τρόπος για την αντιμετώπιση τέτοιων δυσκολιών γίνεται με την μεταφορά μάθησης.

Η γενική ιδέα είναι η επαναχρησιμοποίηση της γνώσης ενός εκπαιδευμένου μοντέλου σε μια νέα εργασία ώστε η διαδικασία της μάθησης να μην ξεκινήσει από το μηδέν. Η μεταφορά μάθησης δεν αποτελεί μια πραγματική τεχνική μάθησης αλλά θεωρείται μια μεθοδολογία για αποτελεσματικότερη μάθηση. Ουσιαστικά, χρησιμοποιείται ένα προεκπαιδευμένο μοντέλο σε ένα λίγο διαφορετικό πρόβλημα (Janiesch et al., 2021). Αυτό γίνεται με την αντικατάσταση του τελικού σταδίου που συνθέτει τον ταξινομητή από καινούργια επίπεδα που έχουν επανεκπαιδευτεί στα δεδομένα του νέου προβλήματος, εφόσον αυτά παρουσιάζουν κάποιες ομοιότητες με τα δεδομένα του αρχικού προβλήματος. Τα συναπτικά βάρη στα αρχικά στάδια επεξεργασίας και εξαγωγής χαρακτηριστικών δεν αλλοιώνονται. Τα κυρία πλεονεκτήματα που εξασφαλίζει η μεταφορά μάθησης είναι ο μικρός χρόνος εκπαίδευσης, η καλύτερη απόδοση του δικτύου και η μείωση της ανάγκης για πολλά δεδομένα (Vlahavas et al., 2020; Zhuang et al., 2021).

Κεφάλαιο 3

Αναγνώριση Αντικειμένων

3.1 Εισαγωγή

Η ανάπτυξη της τεχνολογίας έχει επιτρέψει την ανάδειξη διάφορων μεθόδων που χρησιμοποιούνται στην αναγνώριση αντικειμένων, συμπεριλαμβανομένων παραδοσιακών μεθόδων μηχανικής μάθησης και μεθόδων βαθιάς μάθησης. Αρκετοί επιστήμονες έχουν επισημάνει τις διαφορές ανάμεσα στους δυο προηγούμενους τρόπους προσέγγισης. Και στις δυο περιπτώσεις, ο τρόπος λειτουργίας επικεντρώνεται στην εξαγωγή χαρακτηριστικών από περιοχές έντονου ενδιαφέροντος πάνω στην εικόνα. Το πρόβλημα που προκύπτει σε αυτήν την διαδικασία είναι η ανάγκη επιλογής των πιο σημαντικών χαρακτηριστικών. Σε αυτό το σημείο οι αλγόριθμοι μηχανικής μάθησης έχουν μέτρια απόδοση ενώ στην περίπτωση που ο αριθμός των κλάσεων αυξάνεται, η επιλογή των χαρακτηριστικών γίνεται ακόμα πιο περιπλοκή. Σε αυτό το σημείο οι τεχνικές βαθιάς μάθησης και τα συνελικτικά νευρωνικά δίκτυα παρουσιάζουν πολύ καλύτερη απόδοση αφού η διαδικασία εκπαίδευσης είναι πολύ διαφορετική (O'Mahony et al., 2020). Για αυτό τον λόγο, παρακάτω αναλύονται αποκλειστικά μοντέλα και ανιχνευτές βαθιάς μάθησης.

3.2 Ιστορική Ανασκόπηση στην Αναγνώριση Αντικειμένων

Η αφετηρία της υπολογιστικής όρασης και της αναγνώρισης αντικειμένων βρίσκεται στην αρχή της δεκαετίας του '70. Εκείνη την εποχή οι ερευνητές εργάζονταν στην ανίχνευση απλών γεωμετρικών σχημάτων, όπως οβάλ και στρόγγυλων, με σκοπό την ανίχνευση προσώπων (Yakimovsky, 1976). Με αυτόν τον τρόπο αποδείχθηκε ότι η ανίχνευση ακμών είναι εξαιρετικά σημαντικό στοιχείο στην διαδικασία ανίχνευσης. Στην πορεία, η γρήγορη ανάπτυξη διαφορετικών μεθόδων εξαγωγής χαρακτηριστικών, οδήγησε στην ανάπτυξη διάσημων αλγορίθμων της εποχής, όπως ο αλγόριθμος SIFT (Sorting Intolerant From Tolerant) (Ng & Henikoff, 2003) και ο αλγόριθμος SURF (Speeded Up Robust Features) (Bay et al., 2006). Με την κυκλοφορία αυτών των αλγορίθμων οι εικόνες σταμάτησαν να αντιμετωπίζονται καθολικά, αντίθετα διαιρέθηκαν σε μικρότερα μέρη και για πρώτη φορά εφαρμόστηκαν διάφορα φίλτρα. Αυτό επιτεύχθηκε με την χρήση πυραμοειδών αναπαραστάσεων (Dollár et al., 2014). Κατά την δεκαετία του '80, υπήρξε σημαντική πρόοδος στην αναγνώριση γεωμετρικών σχημάτων, καθώς προτάθηκε μια διαδικασία αντιστοίχισης που χρησιμοποίει την θεωρία των γράφων (Bunke & Allermann, 1983). Ταυτόχρονα είχε παρατηρηθεί η επίδραση εξωτερικών παραγόντων στις εικόνες, όπως ο θόρυβος ο οποίος είχε αποτέλεσμα την εσφαλμένη αναγνώριση.

Η δεκαετία του '90 ήταν αυτή που σηματοδότησε την αρχή της χρήσης των νευρωνικών δικτυών ως ταξινομητές για όλα τα είδη αντικειμένων. Οι σημαντικότερες προκλήσεις που δημιουργηθήκαν για αυτή την προσέγγιση ήταν από την μια η ποιότητα και ο όγκος των δεδομένων και από την άλλη ο μεγάλος χρόνος εκπαίδευσης. Εκείνη την εποχή παρουσιάστηκαν μερικοί πετυχημένοι ταξινομητές στον τομέας της ανίχνευσης προσώπων στις εργασίες (Er et al., 2002; Rowley et al., 1998). Ωστόσο, μετά το 2000 η εξέλιξη των νευρωνικών δικτύων οδήγησε στην δημιουργία των συνελικτικών νευρωνικών δικτύων τα οποία πέτυχαν ρεκόρ βελτιώσεων

στην ανίχνευση αντικειμένων γενικών κατηγοριών. Η επιτυχία τους οφείλεται στα τεράστια σύνολα δεδομένων και στην εξέλιξη της τεχνολογίας που προσέφερε την δυνατότητα μεγάλης υπολογιστικής ικανότητας.

3.3 Βασικές Έννοιες

Η βιβλιογραφία και τα διαθέσιμα επιστημονικά άρθρα για την αναγνώριση αντικειμένων είναι εντυπωσιακά πολλά. Παρά την μεγάλη ενασχόληση των ερευνητών, έχει παρατηρηθεί ότι δεν υπάρχει καθολική συμφωνία σχετικά με τους ορισμούς των διάφορων διεργασιών της υπολογιστικής όρασης. Οι περισσότεροι όροι αναφέρονται συχνά και συγχέονται μεταξύ τους. Στον τομέα της υπολογιστής όρασης διακρίνονται τρεις βασικές έννοιες: ταξινόμηση, εντοπισμός και ανίχνευση (Andreopoulos & Tsotsos, 2013).

3.3.1 Ταξινόμηση

Ως ταξινόμηση (Classification) ορίζεται η διαδικασία κατά την οποία ένα ή περισσότερα αντικείμενα που βρέθηκαν στην ζητούμενη εικόνα, κατηγοριοποιούνται σε ένα συγκεκριμένο σύνολο κλάσεων. Η κατηγοριοποίηση γίνεται σύμφωνα με μια συνάρτηση που υπολογίζει την πιθανότητα του αντικειμένου να ανήκει σε συγκεκριμένη κλάση, προσδιορίζοντας έτσι την παρουσία αλλά όχι την τοποθεσία του (Andreopoulos & Tsotsos, 2013; L. Liu et al., 2020).

3.3.2 Εντοπισμός

Ως εντοπισμός (Localization) ορίζεται η διαδικασία κατά την οποία καθορίζεται η χωρική θέση και η έκταση των αντικειμένων που εντοπίστηκαν. Αυτό πραγματοποιείται με ένα πλαίσιο οριοθέτησης (Bounding Box), δηλαδή ένα ορθογώνιο πλαίσιο που υποδεικνύει την ακριβή θέση του κάθε αντικειμένου (Andreopoulos & Tsotsos, 2013; L. Liu et al., 2020).

3.3.3 Ανίχνευση

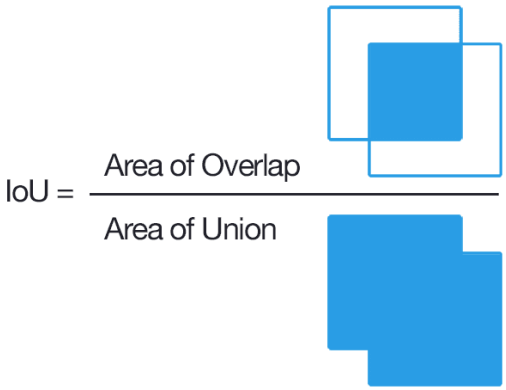
Ως ανίχνευση (Detection) ορίζεται η διαδικασία κατά την οποία αρχικά συνδυάζεται το αποτέλεσμα της ταξινόμησης και του εντοπισμού και έπειτα πραγματοποιείται η αντιστοίχιση του αποτελέσματος με μια ετικέτα. Συνήθως η περιγραφή που εμφανίζεται στην ετικέτα περιέχει την κλάση στην οποία κατηγοριοποιήθηκε το αντικείμενο μαζί με το πλαίσιο οριοθέτησης (Andreopoulos & Tsotsos, 2013; L. Liu et al., 2020).

3.4 Δείκτες Αξιολόγησης

3.4.1 Λόγος Επικάλυψης, Αληθώς/Ψευδώς Θετικό και Αρνητικό

Πρώτα πρέπει να οριστεί ο λόγος επικάλυψης IoU (Intersection over Union) ο οποίος αναφέρεται στα πλαίσια οριοθέτησης. Πιο συγκεκριμένα ο λόγος IoU εκφράζει το πηλίκο της ιδανικής επιφάνειας οριοθέτησης *A* ενός αντικειμένου και της προτεινόμενης επιφάνειας οριοθέτησης *B* του ιδίου αντικειμένου που προέκυψε από υπολογισμό ενός αλγορίθμου. Μαθηματικά ορίζεται από την σχέση:

$$\mathrm{IOU} = \frac{A \frown B}{A \smile B} \quad (17)$$



Εικόνα 22: Αναπαράσταση υπολογισμού του λόγου επικάλυψης IoU. Πηγή pyimagesearch.com

Η τυπική έξοδος ενός ανιχνευτή για μια δοκιμαστική εικόνα j περιλαμβάνει ένα σύνολο τριών στοιχείων για κάθε αντικείμενο που βρέθηκε. Το σύνολο αυτό περιγράφεται ως $\{(b_j, c_j, p_j)\}$ όπου το b_j αποτελεί το πλαίσιο οριοθέτησης (Bounding Box) του αντικειμένου, το c_j αναφέρεται στην προβλεπόμενη κλάση του και το p_j στον βαθμό εμπιστοσύνης του (L. Liu et al., 2020; Padilla et al., 2020, 2021).

Επομένως μια τέτοια πρόβλεψη θεωρείται αληθώς θετική (True Positive) αν πρώτα η πρόβλεψη της κλάσης c_j αντιστοιχεί στην πραγματικότητα και έπειτα ο λόγος επικάλυψης IoU είναι μεγαλύτερος από την τιμή του κατωφλιού ε (L. Liu et al., 2020; Padilla et al., 2020, 2021). Η τιμή ε είναι εσκεμμένα χαμηλή και συνήθως ορίζεται στο 0,5 ή 50% ώστε να ληφθούν υπόψη οι ανακρίβειες στα οριοθετημένα πλαίσια λόγω της ιδιομορφίας των αντικειμένων (Everingham et al., 2010). Σε αντίθετη περίπτωση η πρόβλεψη θεωρείται ψευδώς θετική (False Positive). Από την άλλη πλευρά μια πρόβλεψη μπορεί να θεωρηθεί ψευδώς αρνητική (False Negative) όταν κάποιο αντικείμενο βρίσκεται σε μια εικόνα αλλά δεν μπορεί να ανιχνευτεί ενώ αντίστοιχα μια πρόβλεψη μπορεί να θεωρηθεί αληθώς αρνητική (True Negative) για ένα αντικείμενο που εντοπίστηκε αλλά δεν ταξινομήθηκε στην κλάση που ανήκει (Everingham et al., 2010;).

3.4.2 Ακρίβεια και Ανάκλαση

Το πρώτο βασικό κριτήριο που χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός μοντέλου νευρωνικού δικτύου ή αλγορίθμου είναι η ακρίβεια (precision). Η ακρίβεια ορίζεται ως το πηλίκο των αληθώς θετικών προβλέψεων προς το άθροισμα των αληθώς και ψευδώς θετικών προβλέψεων που επιστρέφονται από τον αλγόριθμο. Το μέτρο της ακρίβειας εκφράζει το ποσοστό των σωστών ταξινομήσεων μεταξύ των κλάσεων (Padilla et al., 2020, 2021; Russakovsky et al., 2015).

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (18)$$

Το δεύτερο βασικό κριτήριο είναι η ανάκλαση (Recall). Η ανάκλαση ορίζεται ως το πηλίκο των αληθώς θετικών προβλέψεων προς το άθροισμα των αληθώς θετικών και ψευδώς αρνητικών προβλέψεων. Το μέτρο της ανάκλασης εκφράζει τον αριθμό των αντικειμένων που ανιχνευτήκαν με επιτυχία (Everingham et al., 2015; Padilla et al., 2020; Russakovsky et al., 2015).

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (19)$$

3.4.3 Μέση Ακρίβεια και Συνολική Μέση Ακρίβεια

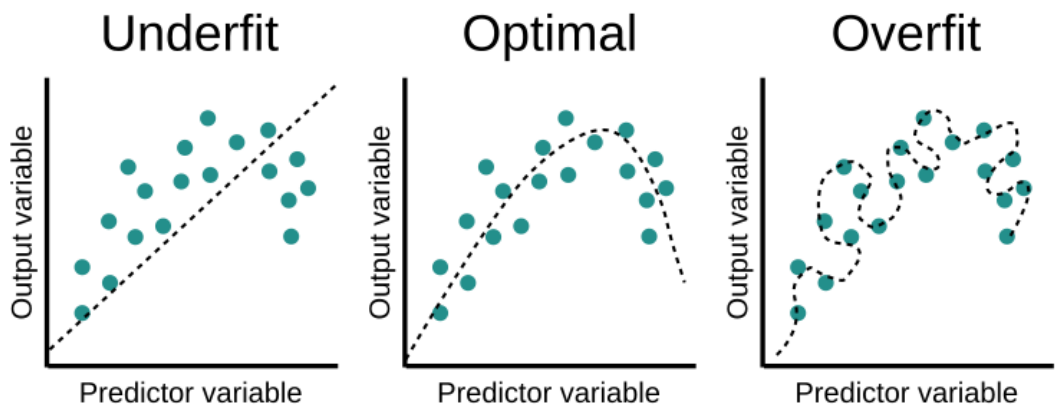
Για την μέτρηση της απόδοσης ενός μοντέλου νευρωνικού δικτύου ή αλγορίθμου χρησιμοποιείται η μέση ακρίβεια (Average Precision) η οποία υπολογίζεται για κάθε μια κατηγορία αντικειμένων ξεχωριστά. Ορίζεται ως το πηλίκο της ακρίβειας προς την ανάκλαση. Παρόλα αυτά, ο πιο διαδεδομένος δείκτης απόδοσης στην υπολογιστική όραση είναι η συνολική μέση ακρίβεια (mean Average Precision). Προέκυψε από την σύγκριση της απόδοσης για όλες τις κατηγορίες αντικειμένων και υιοθετήθηκε ως το τελικό μέτρο απόδοσης αφού υπολογίζει τον μέσο όρο μέσης ακρίβειας κάθε κατηγορίας (L. Liu et al., 2020; Padilla et al., 2020).

3.4.4 Σφάλμα Top-5

Εκτός από τα παραπάνω, ένα εξαιρετικά σημαντικό κριτήριο αξιολόγησης των αρχιτεκτονικών είναι το λεγόμενο σφάλμα ταξινόμησης Top-5 (Top-5 Classification Error). Αυτό το σφάλμα αναπαριστά το ποσοστό των περιπτώσεων όπου διάφορα αντικείμενα που περιέχονται στην εικόνα δεν περιέχονται στις πέντε κορυφαίες προβλέψεις του μοντέλου. Η συγκεκριμένη μέθοδος αξιολόγησης θεσπίστηκε αποκλειστικά για το σύνολο δεδομένων ImageNet, το οποίο αναφέρεται στην επόμενη ενότητα (Russakovsky et al., 2015; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019).

3.4.5 Υπερπροσαρμογή και Υποπροσαρμογή

Αφού ένα μοντέλο ολοκληρώσει την εκπαίδευση σε ένα σύνολο δεδομένων, αναμένεται να αποδίδει ικανοποιητικά σε δεδομένα που δεν συμμετείχαν στην διαδικασία εκπαίδευσης. Η συγκεκριμένη ικανότητα ονομάζεται γενίκευση. Ωστόσο, έχει παρατηρηθεί ότι τα νευρωνικά δίκτυα που έχουν εκπαιδευτεί με σκοπό την μείωση του τετραγωνικού σφάλματος, καταφέρνουν να μειώσουν το σφάλμα αλλά παρουσιάζουν κακή απόδοση σε δεδομένα του ιδίου προβλήματος. Δηλαδή το μοντέλο έχει πετύχει χαμηλό σφάλμα εκπαίδευσης αλλά οι προβλέψεις σε νέα δεδομένα παρουσιάζουν μεγάλες αποκλίσεις. Η κατάσταση αυτή ονομάζεται υπερπροσαρμογή (Overfitting), όπου το μοντέλο έχει χάσει την ικανότητα γενίκευσης παρουσιάζοντας μια υπερβολική εξειδίκευση στα δεδομένα εκπαίδευσης. Αντίθετα, η περίπτωση στην οποία το μοντέλο έχει χαμηλή απόδοση στο αρχικό σύνολο δεδομένων, ονομάζεται υποπροσαρμογή (Underfitting). Μερικές τεχνικές που χρησιμοποιούνται για την αντιμετώπιση και τον περιορισμό των συγκεκριμένων προβλημάτων είναι ο καλύτερος διαχωρισμός των δεδομένων εκπαίδευσης, η πρώιμη εγκατάλειψη, η ομαλοποίηση βαρών και η προσωρινή απόρριψη νευρώνων ή μείωση μεγέθους του δικτύου (Vlahavas et al., 2020).



Εικόνα 23: Διάγραμμα με τρία διαφορετικά επίπεδα προσαρμογής με βάση τα δεδομένα εκπαίδευσης. Πηγή fastaireference.com

Στην εικόνα 23 απεικονίζονται τρεις διαφορετικές περιπτώσεις μοντελοποίησης. Σε κάθε γράφημα το σφάλμα χαρακτηρίζεται ως οι αποστάσεις των πράσινων σημείων από την διακεκομμένη ευθεία. Στο αριστερό γράφημα παρουσιάζεται το φαινόμενο της υποπροσαρμογής καθώς η μάθηση είναι ατελής και το σφάλμα σχετικά μεγάλο. Στο μεσαίο γράφημα η προσαρμογή θεωρείται ιδανική αφού το σφάλμα παραμένει σταθερά χαμηλό και η μοντελοποίηση της σχέσης εισόδου - εξόδου παραπέμπει σε απλή συνάρτηση. Το γράφημα στα δεξιά παρουσιάζει μια πολύπλοκη εκπαίδευση κατά την οποία το σφάλμα σχεδόν μηδενίστηκε και η μοντελοποίηση παραπέμπει σε πολυωνυμική συνάρτηση μεγάλου βαθμού, γεγονός που συμβαίνει στην υπερπροσαρμογή.

Πίνακας 2: Συνοπτικός Πίνακας Μετρήσεων Αξιολόγησης στην Ανίχνευση Αντικειμένων

Συντομογραφία	Δείκτης Αξιολόγησης	Σύντομη Περιγραφή
IOU	Λόγος Επικάλυψης	$\mathrm{IOU} = \frac{A \frown B}{A \smile B}$
TP	True Positive	Αληθώς θετική πρόβλεψη
FP	False Positive	Ψευδώς θετική πρόβλεψη
TN	True Negative	Αληθώς αρνητική πρόβλεψη
FN	False Negative	Ψευδώς αρνητική πρόβλεψη
P	Precision - Ακρίβεια	$\mathrm{P} = \frac{True\ Positives}{True\ Positives + False\ Positives}$
R	Recall - Ανάκλαση	$\mathrm{R} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$
AP	Average Precision - Μέση Ακρίβεια	$AP = \frac{\mathrm{Precision}}{\mathrm{Recall}}$
mAP	mean Average Precision - Συνολική ΜΑ	Μέσος όρος ακρίβειας κάθε κλάσης
Top-5	Top-5 Classification Error	Σφάλμα 5 κορυφαίων προβλέψεων

3.5 Σύνολα Δεδομένων

Τα σύνολα δεδομένων (Datasets) έχουν διαδραματίσει καθοριστικό ρόλο στον τομέα της υπολογιστικής όρασης και της αναγνώρισης αντικειμένων, κυρίως στην επίλυση περίπλοκων προβλημάτων αλλά και ως αντικείμενο σύγκρισης της απόδοσης μεταξύ των διάφορων μοντέλων και αλγορίθμων. Τα σύνολα αυτά ενσωματώνουν ένα τεράστιο αριθμό εικόνων με συγκεκριμένες ιδιότητες. Αναλυτικότερα, κάθε σύνολο περιλαμβάνει εικόνες που περιέχουν αντικείμενα συγκεκριμένου ενδιαφέροντος και συνήθως κάθε εικόνα συνοδεύονται από ένα αρχείο σχολιασμού που αναγράφει την κλάση και την χωρική θέση των αντικειμένων. Την ίδια στιγμή, ένα σύνολο δεδομένων οφείλει να έχει σωστή δομή. Τα σύνολα δεδομένων χωρίζονται σε τρία μέρη με προκαθορισμένη αναλογία: το πρώτο μέρος περιέχει τις εικόνες εκπαίδευσης (train), το δεύτερο μέρος περιέχει τις εικόνες επικύρωσης (validation) και το τρίτο μέρος περιέχει δοκιμαστικές εικόνες (test). Για την ανίχνευση γενικών αντικειμένων υπάρχουν τέσσερα διάσημα σύνολα δεδομένων: α) PASCAL VOC, β) ImageNet, γ) MC COCO, δ) Open Images (L. Liu et al., 2020).

Το PASCAL VOC προτάθηκε στις εργασίες (Everingham et al., 2010, 2015) ξεκινώντας το 2005 με μόνο τέσσερεις κατηγορίες αντικειμένων και με την πάροδο του χρόνου εμπλουτίστηκε φτάνοντας τις 20 κατηγορίες. Περιέχει εικόνες σχετικά κοντά με αυτές του πραγματικού κόσμου ενώ η ποιότητα των δειγμάτων θεωρείται μέτρια. Έπειτα παρουσιάστηκε το ImageNet το 2009 στο (Russakovsky et al., 2015) έχοντας πολύ μεγαλύτερο αριθμό εικόνων ανά κατηγορία σε σύγκριση με το PASCAL VOC και καλύτερη ποιότητα με αντικείμενα που βρίσκονται στο κέντρο των εικόνων. Το MC COCO εμφανίστηκε το 2014 (Lin et al., 2014) κατά τη προσπάθεια των ερευνητών να δημιουργήσουν ένα σύνολο δεδομένων βασισμένο στην πραγματική καθημερινή ζωή. Το MC COCO περιέχει αρκετά πολύπλοκες εικόνες με πολλά αντικείμενα στην καθεμία και προσφέρει περισσότερες πληροφορίες για αυτά. Ακόμα, περιέχει δεδομένα που αφορούν τις συντεταγμένες των αντικειμένων μέσα στις εικόνες, κάτι το οποίο δεν είναι διαθέσιμο στο ImageNet. Το MC COCO θεωρείται το πιο καθιερωμένο σύνολο δεδομένων μέχρι σήμερα. Τέλος το Open Images προέρχεται από την εργασία (Kuznetsova et al., 2020) και επί του παρόντος είναι το μεγαλύτερο σύνολο δεδομένων ανίχνευσης αντικειμένων σύμφωνα με την τελευταία έκδοση (V5). Η σημαντικότερη διαφορά εντοπίζεται στο ότι ο σχολιασμός κάθε εικόνας πραγματοποιήθηκε μόνο για τα αντικείμενα που έχουν υψηλή βαθμολογία εντοπισμού. Στον παρακάτω πίνακα συνοψίζονται περισσότερες λεπτομέρειες για κάθε σύνολο (L. Liu et al., 2020):

Πίνακας 3: Συνοπτικός Πίνακας Συνόλων Δεδομένων

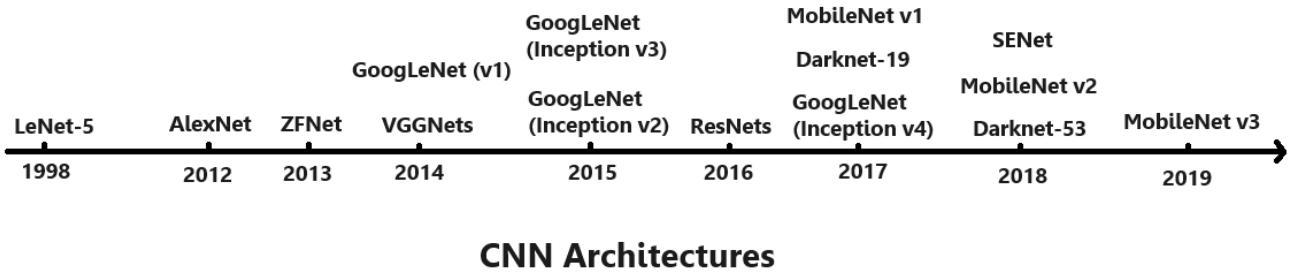
Dataset Name	Year	Classes	Number of Images	Number of Annotated Objects
PASCAL VOC	2007	20	9.963	12.608
	2008	20	8.465	10.363
	2009	20	13.704	17.218
	2010	20	19.740	23.374
	2011	20	22.534	27.450
	2012	20	22.531	27.450
ImageNet (ILSVRC)	2013	200	456.182	401.356
	2014	200	516.840	543.309
	2015	200	527.982	543.309
	2016	200	536.688	543.309
	2017	200	542.188	543.309
MS COCO	2015	80	204.721	896.782
	2016	80	204.721	896.782
	2017	80	163.957	896.782
	2018	80	163.957	896.782
Open Images	2018	500	1.910.098	12.195.144

Κεφάλαιο 4

Αρχιτεκτονικές Βαθιάς Μάθησης

4.1 Εισαγωγή

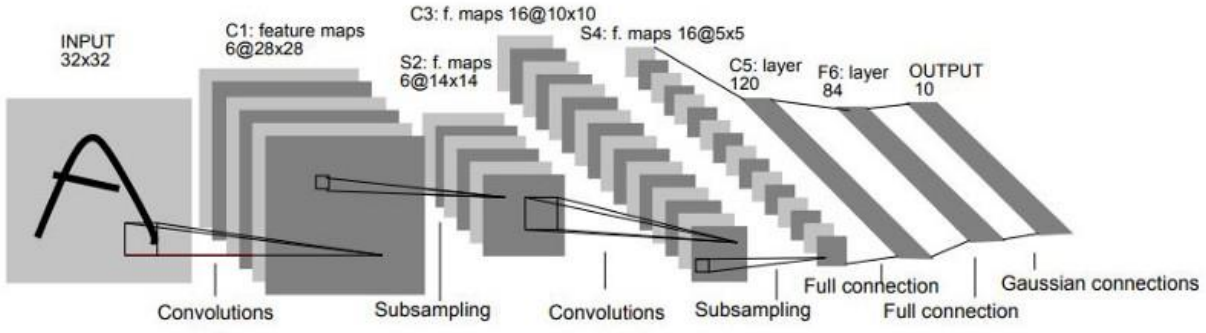
Η αρχιτεκτονική ενός μοντέλου νευρωνικού δικτύου εστιάζει στα ιδιαίτερα χαρακτηριστικά του όπως την τοπολογία, την σύνδεση και τον αριθμό είτε των υπολογιστικών στρωμάτων είτε των νευρώνων (Russell & Norvig, 2010). Ουσιαστικά η αρχιτεκτονική καθορίζει την διαδικασία εκπαίδευσης ενός μοντέλου και την απόδοσή του. Παρακάτω θα γίνει λεπτομερής περιγραφή σε μοντέλα που αποτελούν ορόσημα στην αναγνώριση αντικειμένων. Είναι σημαντικό να αναφερθεί ότι από το 2010 διενεργείται ένας ετήσιος διαγωνισμός οπτικής αναγνώρισης γνωστός ως ILSVRC (ImageNet Large Scale Visual Recognition Challenge) και διοργανώνεται από την ομάδα του ImageNet. Οι συμμετέχοντες καλούνται να υποβάλλουν μοντέλα τα οποία θα διαγωνιστούν σε δυο εργασίες. Η πρώτη είναι ο εντοπισμός αντικειμένων σε μία εικόνα που περιέχει αντικείμενα από 200 κλάσεις και η δεύτερη είναι η ταξινόμηση των αντικειμένων που εντοπίστηκαν σε μια από τις 1000 πιθανές κατηγορίες (ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019). Αξίζει να σημειωθεί ότι αρκετά μοντέλα εξελίχθηκαν με το πέρασμα του χρόνου αλλάζοντας έτσι τον βασικό πυρήνα του δικτύου τους.



Εικόνα 24: Χρονολογικό διάγραμμα επισκόπησης των αρχιτεκτονικών που έχουν προταθεί για την αναγνώριση αντικειμένων .

4.2 Δίκτυο LeNet

Το πρώτο κλασσικό συνελικτικό δίκτυο LeNet-5 (Lecun et al., 1998) προτάθηκε από τον Yann LeCun. Το δίκτυο αυτό είναι εκπαιδευμένο με τον αλγόριθμο ανάστροφης μετάδοσης λάθους (Back Propagation) (Leung & Haykin, 1991) για την αναγνώριση χειρόγραφων χαρακτήρων. Όπως φαίνεται στην εικόνα 25, υπάρχουν δυο συνελικτικά στρώματα, δυο στρώματα υποδειγματοληψίας και τρία πλήρως διασυνδεμένα στρώματα. Το LeNet πλαισιώνει την βάση των σύγχρονων ΣΝΔ παρόλο που εκείνη την εποχή δεν αναγνωρίστηκε η αξία του (Ajit et al., 2020; Z. Li et al., 2022).

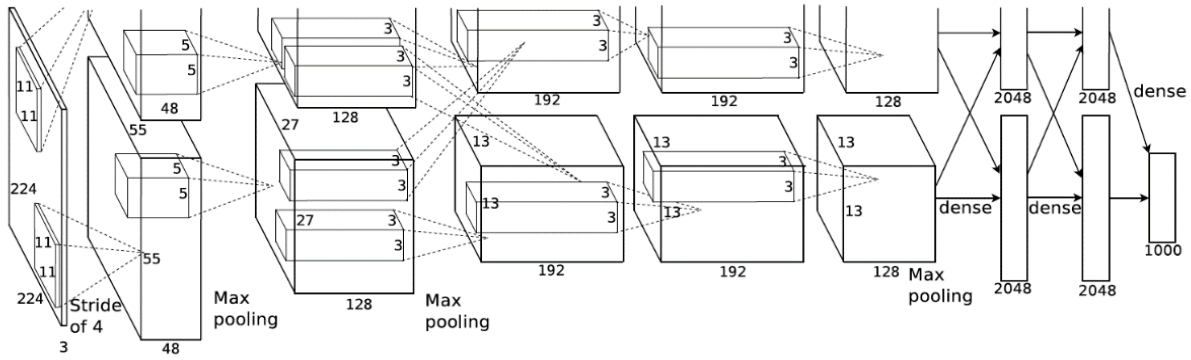


Εικόνα 25: Αναπαράσταση αρχιτεκτονικής του δικτύου LeNet. Πηγή datasciencecentral.com

4.3 Δίκτυο AlexNet

Ο διαγωνισμός ILSVRC του 2012 νικήθηκε από τον Alex Krizhevsky και τους συνεργάτες του, οι οποίοι πρότειναν το πρώτο βαθύ συνελικτικό μοντέλο με σφάλμα 15,3% (Top-5), το AlexNet (Krizhevsky et al., 2012). Το δίκτυο έχει συνολικά πέντε συνελικτικά στρώματα, τρία στρώματα υποδειγματοληψίας μέγιστης τιμής (max pooling), δύο στρώματα με πλήρως διασυνδεμένους νευρώνες και ένα τελικό στρώμα Softmax για την ταξινόμηση. Το πλήθος νευρώνων φτάνει τους 650.000. Το στρώμα εισόδου δέχεται μια τριάδα εικόνων RGB (μια για κάθε χρωματικό κανάλι) μεγέθους 224x224 pixel. Η αρχιτεκτονική απεικονίζεται στην εικόνα 26.

Η εκπαίδευση του δικτύου παρουσιάζει ιδιαίτερα χαρακτηριστικά. Οι έξοδοι των νευρώνων κάθε συνελικτικού επιπέδου κανονικοποιούνται. Έπειτα υπάρχει η πιθανότητα να χρησιμοποιηθεί η μέθοδος της προσωρινής απόσυρσης νευρώνων από την εκπαίδευση. Αυτό συμβαίνει διότι το AlexNet περιέχει μεγάλο αριθμό ελεύθερων παραμέτρων, κάτι το οποίο μπορεί να οδηγήσει σε υπερπροσαρμογή. Τέλος, κατά την διαδικασία της υποδειγματοληψίας, υπάρχει επικάλυψη στα πεδία ενδιαφέροντος γειτονικών νευρώνων, γεγονός που μειώνει ελαφρά το σφάλμα (Ajit et al., 2020; Z. Li et al., 2022; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019).

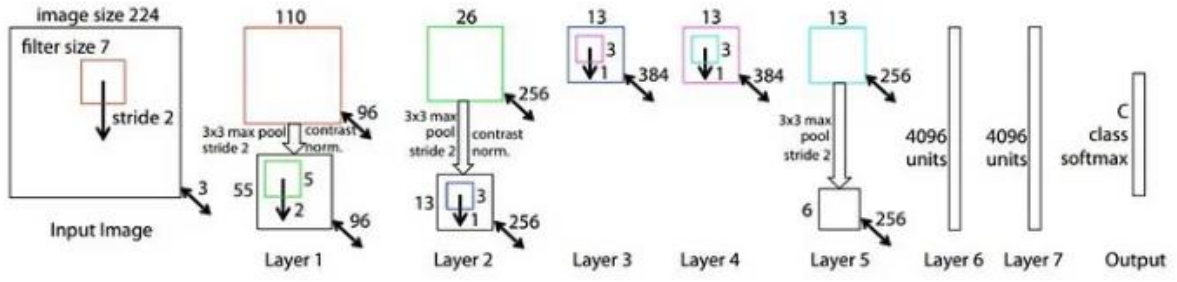


Εικόνα 26: Αναπαράσταση αρχιτεκτονικής του δικτύου AlexNet. Πηγή (Krizhevsky et al., 2012).

4.4 Δίκτυο Zeiler-Fergus

Στον διαγωνισμό ILSVRC του 2013 παρουσιάστηκε το μοντέλο ZFNet, πετυχαίνοντας σφάλμα 14,8% (Top-5) (Zeiler & Fergus, 2014). Πρόκειται για ένα μοντέλο που βασίζεται

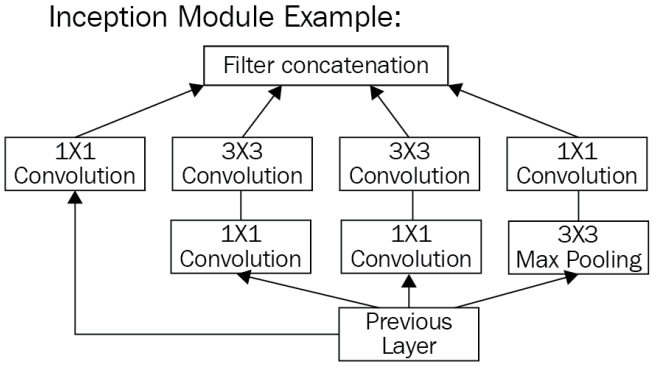
ολοκληρωτικά στο AlexNet και αποτελεί μια σχετικά απλή μετατροπή του. Το ZFNet περιέχει τον ίδιο αριθμό στρωμάτων και νευρώνων με το AlexNet και διαφοροποιείται κυρίως στα πρώτα στρώματα συνέλιξης όπου οι τιμές των φίλτρων και του διασκελισμού (stride) είναι ελαφρώς μειωμένες. Το γεγονός αυτό διακρίνεται στην εικόνα Χ όπου απεικονίζεται η αρχιτεκτονική του.



Εικόνα 27: Αναπαράσταση αρχιτεκτονικής του δικτύου ZFNet. Πηγή (Zeiler & Fergus, 2014).

4.5 Δίκτυο GoogLeNet

Ο διαγωνισμός ILSVRC του 2014 νικήθηκε από την Google με το μοντέλο GoogLeNet πετυχαίνοντας το εντυπωσιακό σφάλμα 6,67% (Top-5) (Szegedy et al., 2015). Η καινοτομία του μοντέλου εντοπίζεται στην εισαγωγή ενός μπλοκ που απαρτίζεται από στρώματα υποδειγματοληψίας μέγιστης τιμής και από συνελικτικά στρώματα. Αυτό το στοιχείο ονομάζεται Inception Module και έχουν κυκλοφορήσει τέσσερεις εκδόσεις. Αναλυτικότερα, ένα Inception μπλοκ αποτελείται από δυο στρώματα στα οποία πραγματοποιούνται συνελίξεις 1x1, 3x3, 5x5 και μια ομαδοποίηση μέγιστης τιμής 3x3, όπως φαίνεται στην εικόνα 28 .



Εικόνα 28: Αναπαράσταση αρχιτεκτονικής ενός Inception Module. Πηγή oreilly.com

Ένα επιπλέον διαφορετικό γνώρισμα του GoogLeNet είναι οι πλευρικοί ταξινομητές. Συνήθως τοποθετούνται διπλά από τα Inception Modules και απαρτίζουν ένα μικρό ολοκληρωμένο συνελικτικό δίκτυο. Σκοπός των πλευρικών ταξινομητών είναι η διόρθωση του σφάλματος στο εσωτερικό του δικτύου.

Το δίκτυο GoogLeNet στην πλήρη αρχιτεκτονική φαίνεται στη εικόνα 29 και περιέχει τρία συνελικτικά στρώματα, εννιά στρώματα Inception, τέσσερα στρώματα υποδειγματοληψίας μέγιστης τιμής, ένα στρώμα υποδειγματοληψίας μέσης τιμής, ένα πλήρως διασυνδεμένο στρώμα και τέλος ένα στρώμα εξόδου Softmax. Το στρώμα εισόδου δέχεται μια τριάδα εικόνων RGB (μια

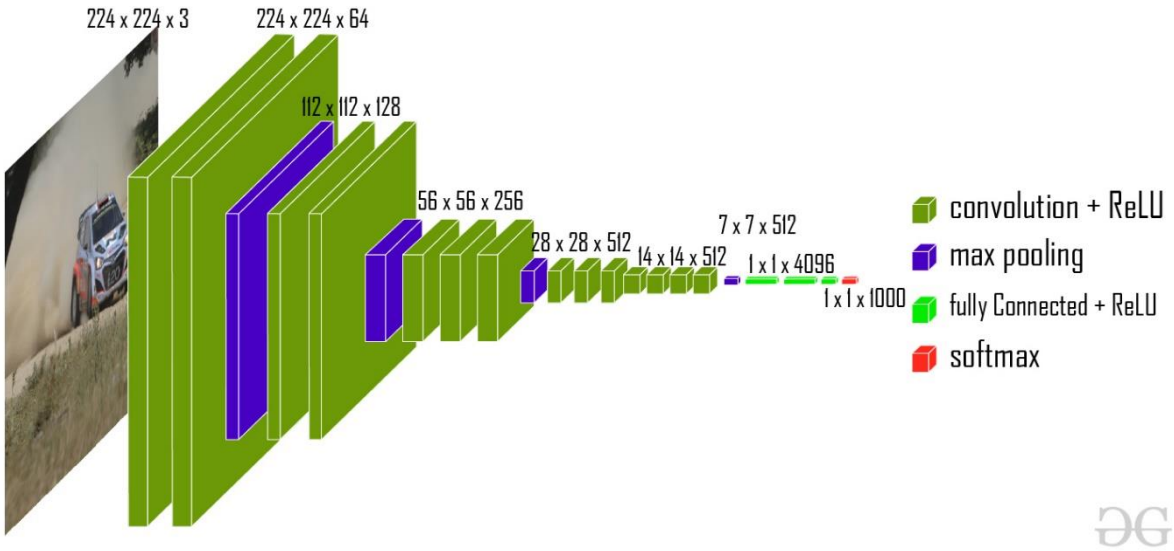
για κάθε χρωματικό κανάλι) μεγέθους 224x224 pixel (Ajit et al., 2020; Z. Li et al., 2022; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019). Στα υπόλοιπα στρώματα αντιστοιχεί μια τριάδα αριθμών M x N x C. Τα νούμερα M x N παραπέμπουν στις διαστάσεις των χαρτών χαρακτηριστικών ενώ ο αριθμός C αναφέρεται στο πλήθος των χαρτών του στρώματος. Για παράδειγμα, στο πρώτο συνελικτικό στρώμα αναγράφεται η τριάδα 112 x 112 x 64, δηλαδή το στρώμα αυτό θα παράξει 64 χάρτες χαρακτηριστικών με διαστάσεις 112 x 112.

type	patch size/ stride	output size	depth
convolution	7×7/2	112×112×64	1
max pool	3×3/2	56×56×64	0
convolution	3×3/1	56×56×192	2
max pool	3×3/2	28×28×192	0
inception (3a)		28×28×256	2
inception (3b)		28×28×480	2
max pool	3×3/2	14×14×480	0
inception (4a)		14×14×512	2
inception (4b)		14×14×512	2
inception (4c)		14×14×512	2
inception (4d)		14×14×528	2
inception (4e)		14×14×832	2
max pool	3×3/2	7×7×832	0
inception (5a)		7×7×832	2
inception (5b)		7×7×1024	2
avg pool	7×7/1	1×1×1024	0
dropout (40%)		1×1×1024	0
linear		1×1×1000	1
softmax		1×1×1000	0

Εικόνα 29: Αναπαράσταση πλήρης αρχιτεκτονικής του δικτύου GoogLeNet. Πηγή (Szegedy et al., 2015).

4.6 Δίκτυο VGGNet

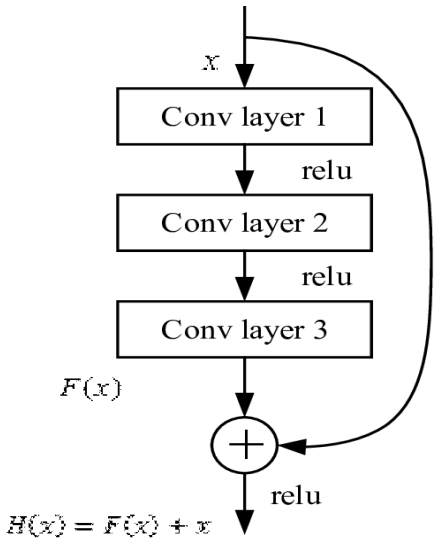
Εκτός από το GoogLeNet, στον διαγωνισμό ILSVRC του 2014 προτάθηκαν και τα δίκτυα VGGNets (Visual Geometry Group Net) (Simonyan & Zisserman, 2015) από τους Karen Simonyan και Andrew Zisserman αποσπώντας την πρώτη θέση στην εργασία του εντοπισμού αντικειμένων και την δεύτερη θέση στον πρόβλημα της ταξινόμησης με σφάλμα 6,8% (Top-5), μετά το GoogLeNet. Υπάρχουν δυο υποπαραλλαγές, οι VGG-16 και VGG-19. Οι αριθμοί 16 και 19 αναφέρονται στο πλήθος των στρωμάτων. Το δίκτυο περιέχει 13 ή 16 συνελικτικά στρώματα (για VGG-16 και VGG-19 αντίστοιχα), πέντε στρώματα υποδειγματοληψίας μέγιστης τιμής, δυο πλήρως διασυνδεμένα στρώματα και τέλος ένα στρώμα Softmax (Ajit et al., 2020; Z. Li et al., 2022; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019).



Εικόνα 30: Αναπαράσταση αρχιτεκτονικής του δικτύου VGG-16. Πηγή geeksforgeeks.org

4.7 Δίκτυο ResNet

Το δίκτυο Residual Network (ResNet) αναδείχθηκε στην πρώτη θέση του διαγωνισμού ILSVRC 2016, και υπάρχει σε πέντε υποπαραλλαγές των 18, 34, 50, 101 και 152 στρωμάτων (He et al., 2016). Διαθέτει ένα χαρακτηριστικό αρκετά όμοιο με τα Inception Modules του GoogLeNet, που ονομάζεται Residual Network Block. Η φιλοσοφία ενός μπλοκ τέτοιου τύπου φαίνεται στην εικόνα 31. Κάθε μπλοκ ενσωματώνει μια συνάρτηση *F* και αποτελείται από συνελικτικά δίκτυα τριών στρωμάτων. Μετά από αρκετά πειράματα, αποδείχτηκε ότι το ResNet μπορεί να πετύχει εξαιρετική απόδοση στην μείωση του σφάλματος, καταγράφοντας ποσοστό 3,57% (Top-5). Γενικότερα έχει παρατηρηθεί ότι η αύξηση των επίπεδων ενός δικτύου επηρεάζει αντίστροφα το σφάλμα, δηλαδή συνεισφέρει στην αύξησή του.



Εικόνα 31: Αναπαράσταση αρχιτεκτονικής ενός ResBlock. Πηγή researchgate.net

Στην εικόνα 32 παρουσιάζεται ο πίνακας αρχιτεκτονικής του δικτύου για όλες τις εκδόσεις του. Αποτελείται από ένα συνελικτικό στρώμα, τέσσερα στρώματα υποδειγματοληψίας μέγιστης

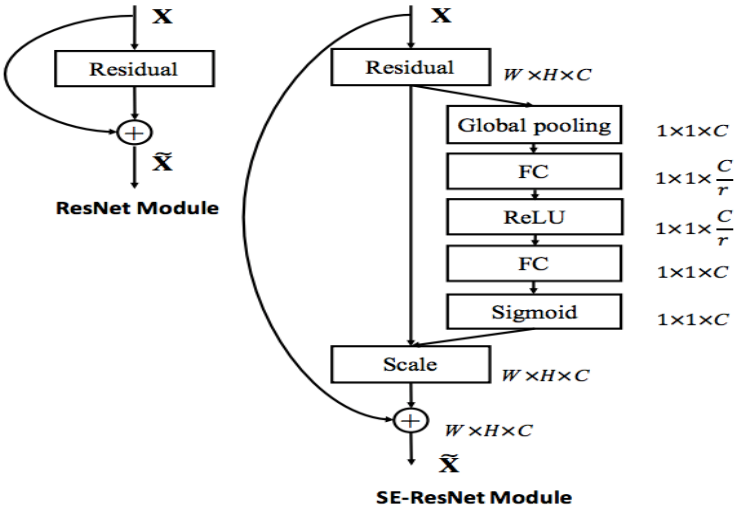
τιμής, ένα στρώμα υποδειγματοληψίας μέσης τιμής και ένα στρώμα εξόδου Softmax. Ο αριθμός των στρωμάτων ResBlock ποικίλει ανάλογα με την έκδοση του μοντέλου (He et al., 2016; Z. Li et al., 2022; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019).

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Εικόνα 32: Αναπαράσταση αρχιτεκτονικής του δικτύου ResNet. Πηγή (He et al., 2016).

4.8 Δίκτυο SENet

Τον αμέσως επόμενο χρόνο στον διαγωνισμό ILSVRC 2017 διακρίθηκε το δίκτυο Squeeze & Excitation Net (SENet) πετυχαίνοντας σφάλμα 2,25% (Top-5) (Hu et al., 2018). Το συγκεκριμένο δίκτυο βασίζεται ολοκληρωτικά στο ResNet έχοντας μια πολύ βασική διαφορά, η οποία εντοπίζεται στον τύπου του βασικού μπλοκ. Το συγκεκριμένο μπλοκ ονομάζεται SEBlock και αποτελεί μια βελτίωση του ResBlock. Στην εικόνα 33 παρουσιάζεται στα αριστερά η δομή ενός ResBlock σε σύγκριση με ένα SEBlock στα δεξιά. Αναλυτικότερα το SEBlock περιέχει ένα επιπλέον παράλληλο μικρό δίκτυο αποτελούμενο από τρία στρώματα που λειτουργεί ως μια επιπλέον συμπίεση πληροφοριών.



Εικόνα 33: Αναπαράσταση αρχιτεκτονικής ενός SEBlock (δεξιά) και ενός ResBlock (αριστερά). Πηγή (Hu et al., 2018).

Η αρχιτεκτονική του δικτύου SENet είναι παρόμοια με αυτήν του ResNet-50 που περιεγράφηκε πριν εκτός από το γεγονός όπου τα ResBlock έχουν αντικατασταθεί από τα SEBlock.

Το δίκτυο αποτελείται από ένα απλό συνελικτικό στρώμα, τέσσερεις ομάδες SEBlock όπου η καθεμιά χρησιμοποιεί διαφορετικές συναρτήσεις και παραμέτρους, ένα στρώμα υποδειγματοληψίας μέσης τιμής και ένα στρώμα εξόδου Softmax (Hu et al., 2018; Z. Li et al., 2022; ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019).

Output size	SE-ResNet-50
112×112	conv, 7×7, 64, stride 2
56×56	max pool, 3×3, stride 2
	$\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \\ fc, [16, 256] \end{bmatrix} \times 3$
28×28	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \\ fc, [32, 512] \end{bmatrix} \times 4$
14×14	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 1024 \\ fc, [64, 1024] \end{bmatrix} \times 6$
7×7	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \\ fc, [128, 2048] \end{bmatrix} \times 3$
1×1	global average pool, 1000-d fc, softmax

Εικόνα 34: Αναπαράσταση αρχιτεκτονικής του δικτύου SENet. Πηγή (Hu et al., 2018).

4.9 Δίκτυο Darknet

Επίσης, το 2017 παρουσιάστηκε το δίκτυο Darknet από τους Redmon και Farhadi το οποίο χρησιμοποιείται ως η ραχοκοκαλιά του αλγορίθμου YOLO. Βασίζεται στην αρχιτεκτονική του δικτύου VGGNet καθώς χρησιμοποιεί συνελίξεις 3x3 και ταυτόχρονα διπλασιάζει τον αριθμό καναλιών μετά από κάθε βήμα υποδειγματοληψίας. Επιπλέον, για την μείωση της διάστασης των παραμέτρων χρησιμοποιούνται αποκλειστικά συνελίξεις 1x1. Η αρχιτεκτονική του δικτύου Darknet-19 περιγράφεται στην εικόνα 35 και συνολικά περιέχει 19 συνελικτικά στρώματα, πέντε στρώματα υποδειγματοληψίας μέγιστης τιμής και ένα στρώμα εξόδου Softmax. Το Darknet-19 είναι ταχύτερο από το δίκτυο VGG αφού πραγματοποίει σημαντικά λιγότερους υπολογισμούς λόγω την απλοϊκότητάς του. Τα συνελικτικά στρώματα του VGG-16 απαιτούν περίπου 30 δισεκατομμύρια παραμέτρους για ένα πέρασμα σε μια εικόνα με ανάλυση 224x224, ενώ το Darknet-19 απαιτεί περίπου 8 δισεκατομμύρια παραμέτρους για την ίδια διαδικασία (Redmon & Farhadi, 2017). Ένα χρόνο αργότερα, οι ίδιοι ερευνητές παρουσίασαν μια αποδοτικότερη έκδοση η οποία έχει μεγαλύτερο πλήθος στρωμάτων και χρησιμοποιεί διαδοχικές συνελίξεις 1x1 και 3x3. Πρόκειται για το δίκτυο Darknet-53 το οποίο είναι μιάμιση φορές ταχύτερο από το ResNet-101 και δυο φορές ταχύτερο από το ResNet-152 πετυχαίνοντας παρόμοια απόδοση (Redmon & Farhadi, 2018). Η αρχιτεκτονική του δικτύου Darknet-53 περιγράφεται στην ίδια εικόνα (35) και συνολικά περιέχει 53 συνελικτικά στρώματα.

Type	Filters	Size/Stride	Output
Convolutional	32	3×3	224×224
Maxpool		$2 \times 2/2$	112×112
Convolutional	64	3×3	112×112
Maxpool		$2 \times 2/2$	56×56
Convolutional	128	3×3	56×56
Convolutional	64	1×1	56×56
Convolutional	128	3×3	56×56
Maxpool		$2 \times 2/2$	28×28
Convolutional	256	3×3	28×28
Convolutional	128	1×1	28×28
Convolutional	256	3×3	28×28
Maxpool		$2 \times 2/2$	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Maxpool		$2 \times 2/2$	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	1000	1×1	7×7
Avgpool		Global	1000
Softmax			

Table 6: Darknet-19.

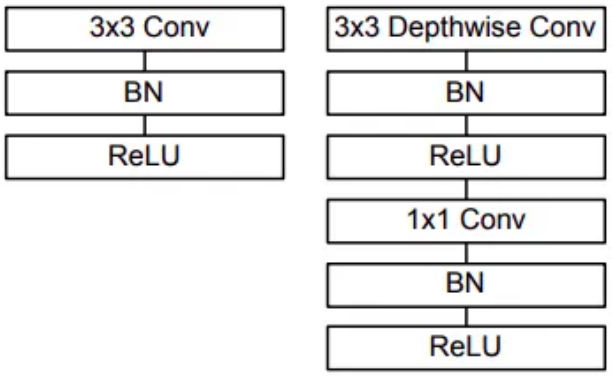
	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Table 1. **Darknet-53.**

Εικόνα 35: Αναπαράσταση αρχιτεκτονικής του δικτύου Darknet-19 και του Darknet-53. Πηγή (Redmon & Farhadi, 2017) και (Redmon & Farhadi, 2018).

4.10 Δίκτυα MobileNet

Τα δίκτυα MobileNet είναι μια σειρά ελαφρών βαθιών νευρωνικών δικτύων που προτάθηκαν από την ομάδα της Google και χρησιμοποιούνται σε συσκευές με περιορισμένους υπολογιστικούς πόρους, όπως τα κινητά τηλέφωνα (smartphones). Μέχρι σήμερα υπάρχουν τρεις εκδόσεις, τα MobileNet V1 (A. G. Howard et al., 2017), MobileNet V2 (Sandler et al., 2018) και MobileNet V3 (A. Howard et al., 2019). Στα μοντέλα αυτά χρησιμοποιούνται πολύ απλοποιημένες αρχιτεκτονικές με σκοπό την μείωση των υπολογισμών και των παραμέτρων. Αυτό επιτυγχάνεται με την τεχνική της διαχωρίσιμης συνέλιξης κατά βάθος (Depth Wise Convolution) (Chollet, 2017) και της σημειακής συνέλιξης (Point Wise Convolution) (Hua et al., 2018). Ουσιαστικά η συνέλιξη χωρίζεται σε δυο στρώματα όπου το πρώτο χρησιμοποιείται για να φιλτράρει τα κανάλια εισόδου ενώ το δεύτερο στρώμα συνδυάζει το παραγόμενο αποτέλεσμα και δημιουργεί ένα νέο χαρακτηριστικό. Στη συνέχεια, χρησιμοποιείται ένα ακόμα πρόσθετο στρώμα που υπολογίζει τον γραμμικό συνδυασμό της εξόδου από την κατά βάθος συνέλιξη (A. G. Howard et al., 2017). Στην εικόνα 36 παρουσιάζονται οι διαφορές μεταξύ μιας κλασσικής συνέλιξης (αριστερά) και μιας διαχωρίσιμης συνέλιξης (δεξιά).



Εικόνα 36: Αναπαράσταση αρχιτεκτονικής κλασσικής συνέλιξης (αριστερά) και διαχωρίσιμης συνέλιξης (δεξιά). Πηγή (Howard et al., 2017).

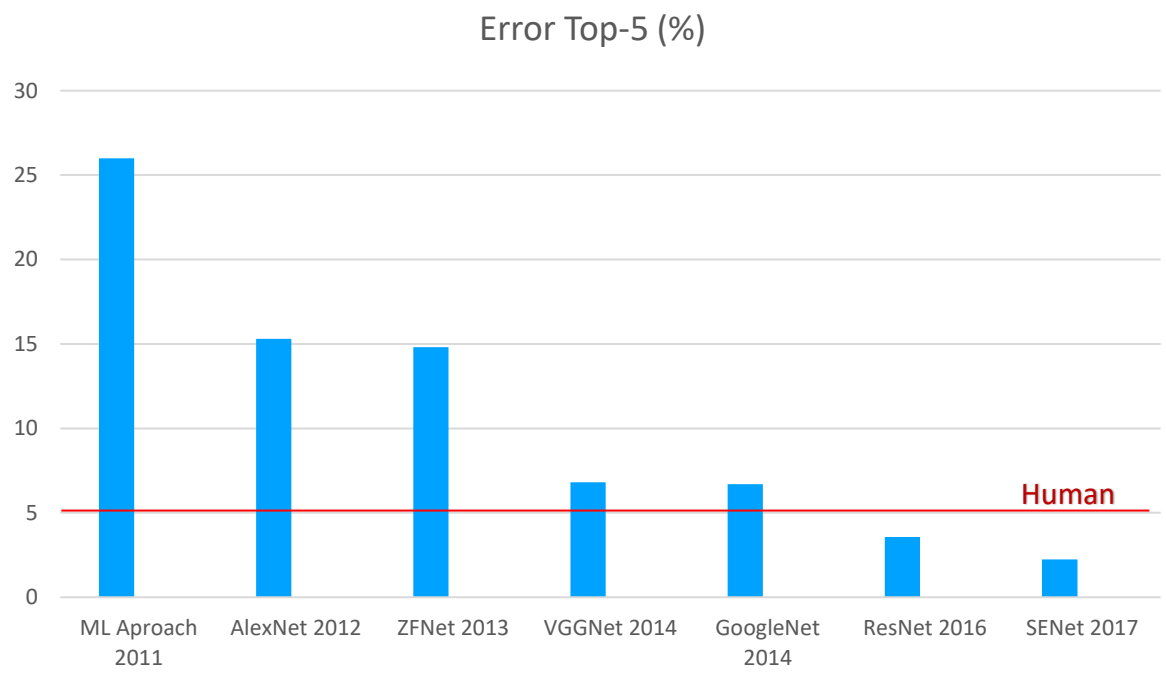
Η βάση της αρχιτεκτονικής των MobileNets αναλύεται στην εικόνα 37. Τα δίκτυα αποτελούνται από συνελικτικά στρώματα είτε κλασσικά είτε διαχωρίσιμα και η έξοδός τους περιλαμβάνει ένα στρώμα υποδειγματοληψίας μέσης τιμής, ένα πλήρως διασυνδεμένο στρώμα και τέλος ένα στρώμα εξόδου Softmax. Όλα τα στρώματα συνέλιξης ενσωματώνουν την συνάρτηση ενεργοποίησης Relu ενώ η έξοδος κάθε στρώματος κανονικοποιείται (A. G. Howard et al., 2017).

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

Εικόνα 37: Αναπαράσταση αρχιτεκτονικής του δικτύου MobileNet. Πηγή (Howard et al., 2017).

4.11 Σύνοψη Κεφαλαίου

Στα πρώτα χρόνια του διαγωνισμού ILSVRC ο νικητής είχε πέτυχει σφάλμα 26% χρησιμοποιώντας συμβατικά μοντέλα μηχανικής μάθησης που περιέχουν χαρακτηριστικά από τεχνολογίες όπως ο αλγόριθμος SIFT και SVM (Sanchez & Perronnin, 2011). Έκτοτε, το AlexNet αποτελεί την πρώτη σημαντική βελτίωση στην επίδοση του σφάλματος, σημειώνοντας 15,3% Top-5 αλλά οι πιο πρόσφατες εκδόσεις μοντέλων όπως το ResNet και το SENet κατέγραψαν επιδόσεις ρεκόρ με 3,57% και 2,25% αντίστοιχα. Αξιοσημείωτη είναι επίσης η απόδοση του GoogLeNet και των μετέπειτα εκδόσεών του, ξεκινώντας με 6,7% στην πρώτη έκδοση και 3,1% στην τελική αντίστοιχα. Στην εικόνα 38 παρουσιάζεται η διαχρονική εξέλιξη των σημαντικότερων επιδόσεων που καταγράφηκαν στον διαγωνισμό ILSVRC από το 2011 μέχρι και το 2017. Είναι προφανές ότι τα βαθιά συνελικτικά μοντέλα έχουν κυριαρχήσει έναντι των μοντέλων μηχανικής μάθησης στην αναγνώριση εικόνων, ενώ ταυτόχρονα ξεπεράσαν την επίδοση της ανθρώπινης ικανότητας, η οποία σύμφωνα με επίσημα πειράματα βρίσκεται στο 5,1% (Russakovsky et al., 2015). Στην επόμενη σελίδα βρίσκεται ο συγκεντρωτικός πίνακας αρχιτεκτονικών όπου συμπεριλαμβάνονται λεπτομερείς πληροφορίες και χαρακτηριστικά για όλες τις εκδόσεις των μοντέλων που αναλύθηκαν σε αυτό το κεφάλαιο.



Εικόνα 38: Γράφημα απόδοσης σφάλματος Top-5 κατά τους διαγωνισμούς ILSVRC 2011-2017.

Πίνακας 4: Αναλυτικός Πίνακας Αρχιτεκτονικών Βαθιάς Μάθησης

Architecture Name	Year	Layers Convolution + Fully Connected	Parameters	Top-5 Error	Trained Task	Paper
LeNet-5	1998	2+1	$6x10^3$	-	Ταξινόμηση Εικόνων	(Lecun et al., 1998)
AlexNet	2012	5+2	$57x10^6$	15,3%	Ταξινόμηση Εικόνων	(Krizhevsky et al., 2012)
ZFNet	2013	5+2	$58x10^6$	14,8%	Ταξινόμηση Εικόνων	(Zeiler & Fergus, 2014)
VGG-16	2014	13+2	$138x10^6$	6,8%	Ταξινόμηση Εικόνων	(Simonyan & Zisserman, 2015)
VGG-19	2014	16+2	$144x10^6$			
GoogLeNet	2014	22	$6x10^6$	6,7%	Ταξινόμηση Εικόνων	(Szegedy et al., 2015)
GoogLeNet Inception v2	2015	31	$31x10^6$	4,8%	Ταξινόμηση Εικόνων	(Ioffe & Szegedy, 2015)
GoogLeNet Inception v3	2015	47	$47x10^6$	3,6%	Ταξινόμηση Εικόνων	(Szegedy et al., 2016)
ResNet18	2016	17	$12x10^6$	3,57%	Ταξινόμηση Εικόνων	(He et al., 2016)
ResNet34	2016	33	$22x10^6$			
ResNet50	2016	49	$25x10^6$			
ResNet101	2016	100	$44x10^6$			
ResNet152	2016	151	$60x10^6$			
GoogLeNet Inception v4	2017	75	$43x10^6$	3,1%	Ταξινόμηση Εικόνων	(Szegedy et al., 2017)
Darknet-19	2017	19	$20x10^6$	-	Εντοπισμός Αντικειμένων	(Redmon & Farhadi, 2017)
MobileNet v1	2017	27+1	$4x10^6$	-	Ταξινόμηση Εικόνων + Εντοπισμός Αντικειμένων	(A. G. Howard et al., 2017)
MobileNet v2	2018	53	$3,5x10^6$	-	Ταξινόμηση Εικόνων + Εντοπισμός Αντικειμένων	(Sandler et al., 2018)
SENet	2018	50	$26x10^6$	2,25%	Ταξινόμηση Εικόνων	(Hu et al., 2018)
Darknet-53	2018	53	$28x10^6$	-	Εντοπισμός Αντικειμένων	(Redmon & Farhadi, 2018)
MobileNet v3 Large	2019	28	$5x10^6$	-	Ταξινόμηση Εικόνων + Εντοπισμός Αντικειμένων	(A. Howard et al., 2019)
MobileNet v3 Small	2019	14	$2,5x10^6$	-	Ταξινόμηση Εικόνων + Εντοπισμός Αντικειμένων	

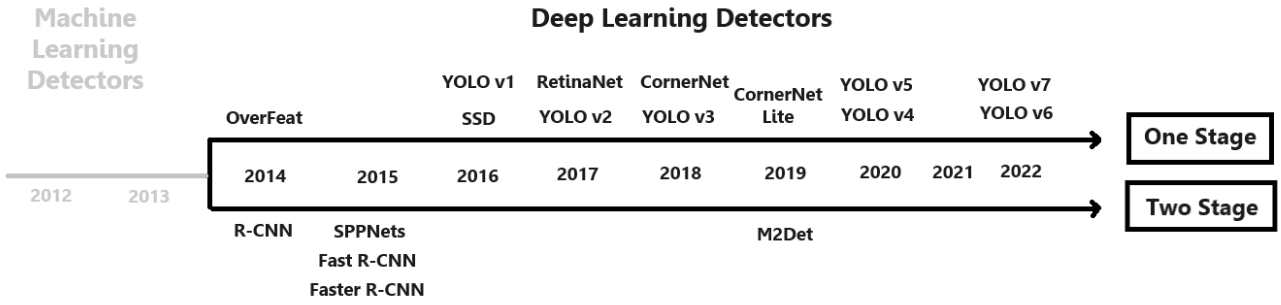
Κεφάλαιο 5

Αλγόριθμοι Βαθιάς Μάθησης

5.1 Εισαγωγή

Η ανίχνευση των αντικειμένων εκτελείται από διάφορους αλγορίθμους που χρησιμοποιούνται σε συνδυασμό με τα μοντέλα βαθιάς μάθησης. Υπάρχουν δυο κατηγορίες στους αλγορίθμους ανίχνευσης, οι ανιχνευτές ενός σταδίου και οι ανιχνευτές δυο σταδίων.

Η διαδικασία ανίχνευσης που ακολουθούν οι ανιχνευτές δυο σταδίων χωρίζεται σε δυο μέρη. Στο πρώτο μέρος, οι ανιχνευτές αναζητούν τις περιοχές ενδιαφέροντος και έπειτα εφαρμόζουν την διαδικασία ταξινόμησης στις συγκεκριμένες περιοχές. Αντίθετα, οι ανιχνευτές ενός σταδίου παραλείπουν την διαδικασία αναζήτησης περιοχής ενδιαφέροντος και προχωρούν στην ανίχνευση επιστρέφοντας απευθείας την θέση και την κατηγορία των αντικειμένων. Παρακάτω, παρουσιάζεται η χρονολογική σειρά των ανιχνευτών που έχουν ξεχωρίσει .



Εικόνα 39: Χρονολογικό διάγραμμα επισκόπησης των αλγορίθμων ανίχνευσης που έχουν προταθεί στην αναγνώριση αντικειμένων .

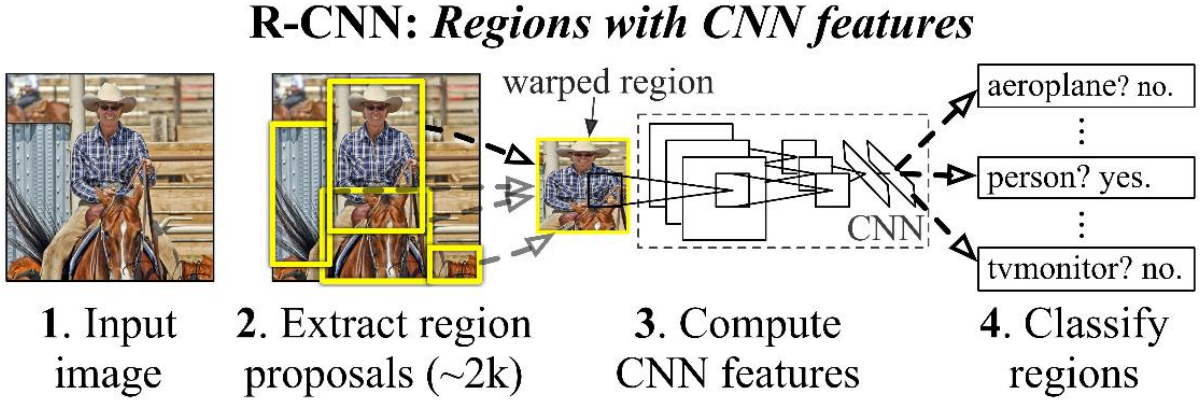
5.2 Αλγόριθμοι Δύο Σταδίων

5.2.1 R-CNN

Η πρώτη επιτυχημένη υλοποίηση δημιουργήθηκε το 2014 από τον Ross Girshick και την ομάδα του, οι οποίοι πρότειναν το αλγόριθμο R-CNN (Girshick et al., 2014). Η απόδοσή του μετρήθηκε στο σύνολο δεδομένων PASCAL VOC 2010-12 και ανέρχεται στο ποσοστό 53.3% mAP. Πρόκειται για έναν αλγόριθμο που χωρίζεται σε τρεις ενότητες. Στην πρώτη ενότητα ο αλγόριθμος αναζητά υποψήφιες περιοχές αναγνώρισης ορίζοντας πλαίσια οριοθέτησης. Στην επόμενη ενότητα, ο αλγόριθμος χρησιμοποιεί ένα βαθύ νευρωνικό δίκτυο για να πραγματοποιηθεί η εξαγωγή χαρακτηριστικών για κάθε περιοχή που κατοχυρώθηκε στο

προηγούμενο βήμα. Στην τελευταία ενότητα, βρίσκεται ο ταξινομητής όπου πραγματοποιούνται οι προβλέψεις για τα αντικείμενα που βρέθηκαν σύμφωνα με τα χαρακτηριστικά τους.

Για την δημιουργία των πλαισίων οριοθέτησης, ο R-CNN χρησιμοποιεί την μέθοδο της Επιλεκτικής Αναζήτησης (Selective Search). Σύμφωνα με αυτήν, η ομαδοποίηση γίνεται σε γειτονικά εικονοστοιχεία (pixel) που παρουσιάζουν ομοιότητες ως προς το χρώμα και την ένταση ενώ για κάθε εικόνα παράγονται 2000 προτεινόμενες περιοχές με τυχαίες διαστάσεις. Όσον αφορά την εξαγωγή χαρακτηριστικών, οι ερευνητές στην πρώτη υλοποίηση του R-CNN χρησιμοποίησαν το δίκτυο AlexNet κάνοντας μερικές παραμετροποιήσεις στην αρχιτεκτονική του. Επομένως, η έξοδος του ΣΝΔ αποτελείται από ένα διάνυσμα χαρακτηριστικών 4.096 διαστάσεων, το οποίο προκύπτει για κάθε περιοχή αναγνώρισης. Τέλος, τα διανύσματα οδηγούνται σε ένα σύνολο γραμμικών μοντέλων ταξινόμησης SVM (Support Vector Machine). Για κάθε πιθανή κλάση υπάρχει ένας εκπαιδευμένος ταξινομητής SVM.

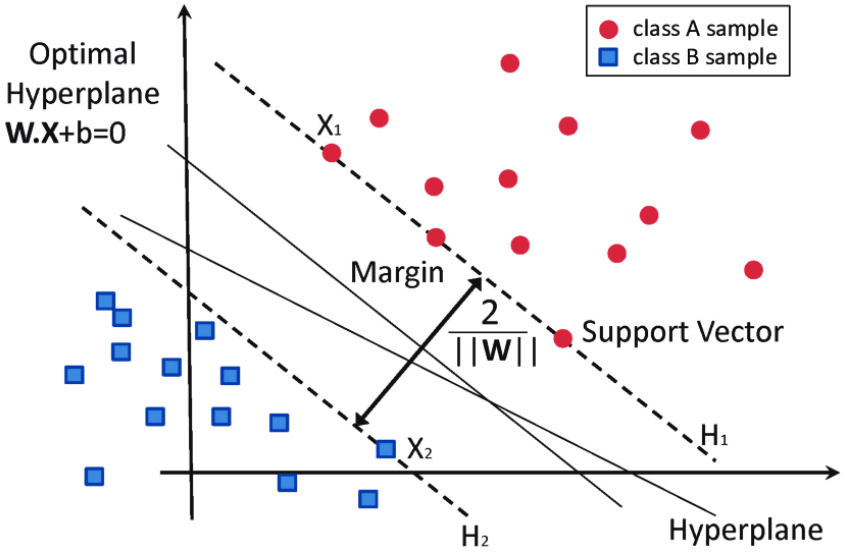


Εικόνα 40: Αναπαράσταση λειτουργίας του R-CNN. Πηγή (Girshick et al., 2014).

SVM Algorithm – Μηχανές Υποστήριξης Διανυσμάτων

Οι μηχανές υποστήριξης διανυσμάτων (Support Vector Machine - SVM) είναι μέθοδοι μηχανικής μάθησης που χρησιμοποιούνται σε προβλήματα γραμμικής ή μη γραμμικής ταξινόμησης και σε προβλήματα παλινδρόμησης. Πρωτοεμφανίστηκαν το 1963 από τους Vapnik και Chapelle, οι οποίοι πέτυχαν σημαντικές βελτιώσεις κατά την δεκαετία του 1990, καθιστώντας τα αποδοτικά μέχρι την εμφάνιση των βαθιών νευρωνικών δικτυών (Vapnik & Chapelle, 2000).

Ο κύριος στόχος της μεθόδου SVM είναι η εύρεση μιας διαχωριστικής επιφάνειας που έχει την δυνατότητα να απέχει όσο το δυνατόν περισσότερο από τα δείγματα των κλάσεων που χωρίζει. Η συγκεκριμένη επιφάνεια ονομάζεται υπερεπιφάνεια ή υπερεπίπεδο και ορίζεται από διανύσματα. Τα διανύσματα είναι υπεύθυνα για τον διαχωρισμό και ονομάζονται διανύσματα υποστήριξης (support vectors). Η διάσταση του υπερεπιπέδου εξαρτάται από τον αριθμό των χαρακτηριστικών. Για παράδειγμα εάν ο αριθμός χαρακτηριστικών εισόδου είναι ίσος με δυο, τότε το υπερεπίπεδο είναι απλώς μια ευθεία γραμμή. Σε περίπτωση που ο αριθμός χαρακτηριστικών εισόδου είναι ίσος με τρία, τότε το υπερεπίπεδο συγκροτεί ένα δισδιάστατο επίπεδο (ΔΙΑΜΑΝΤΑΡΑΣ, & ΜΠΟΤΣΗΣ, 2019).

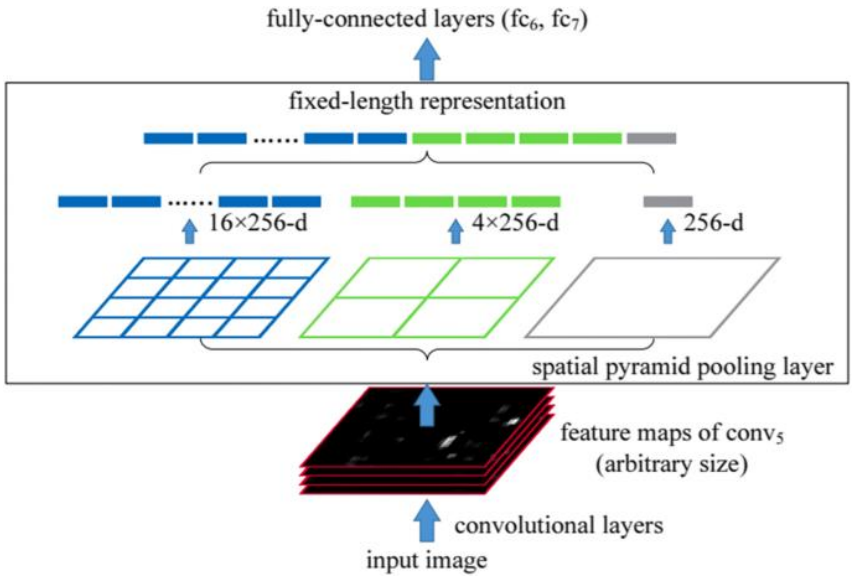


Εικόνα 41: Αναπαράσταση ταξινόμησης δεδομένων με μηχανή υποστήριξης διανυσμάτων. Πηγή researchgate.net.

5.2.2 SPPNet

Ο Kaiming He μαζί με την ερευνητική του ομάδα παρατήρησαν ένα τεχνικό ζήτημα που προκύπτει στα δεδομένα εκπαίδευσης ενός ΣΝΔ και στην εργασία τους παρουσίασαν μια πρωτοποριακή λύση. Το πρόβλημα εντοπίζεται στην απαίτηση των ΣΝΔ για το σταθερό μέγεθος της εικόνας εισόδου (συνήθως 224x224). Όπως είδαμε παραπάνω όλα τα ΣΝΔ απαρτίζονται από συνελικτικά στρώματα, από επίπεδα υποδειγματοληψίας και από πλήρως διασυνδεμένα επίπεδα. Ο περιορισμός για το σταθερό μέγεθος της εικόνας προέρχεται μόνο από τα πλήρως διασυνδεμένα επίπεδα αφού η αναλογία μεγέθους είναι σταθερή εξ ορισμού τους. Γίνεται εύκολα αντιληπτό πως στην περίπτωση που ένα ΣΝΔ τροφοδοτηθεί με μια εικόνα οποιουδήποτε μεγέθους, τότε θα πραγματοποιηθούν διαδικασίες είτε περικοπής είτε παραμόρφωσης. Το αποτέλεσμα αυτής της διαδικασίας θα είναι μια εικόνα που πιθανότατα δεν θα περιέχει ολόκληρο το αντικείμενο ή το αντικείμενο θα έχει παραμορφωθεί σε μεγάλο βαθμό, γεγονός που επηρεάζει την απόδοση του δικτύου (He et al., 2015).

Η τεχνική που χρησιμοποιήθηκε για την εξάλειψη του παραπάνω προβλήματος ονομάζεται Spatial Pyramid Pooling (SPP) και επιτρέπει από την μια πλευρά την είσοδο δεδομένων εκπαίδευσης και από την άλλη την πρόβλεψη αντικειμένων για οποιοδήποτε μέγεθος εικόνας. Πρέπει να τονιστεί ότι η συγκεκριμένη τεχνική δεν αποτελεί έναν ολοκληρωμένο αλγόριθμο ανίχνευσης αντικειμένων ούτε θεωρείται ένα ολοκληρωμένο μοντέλο, αλλά η εφαρμογή της πέτυχε αξιοσημείωτες βελτιώσεις. Αναλυτικότερα, πρόκειται για την προσθήκη ενός στρώματος πυραμίδας ακριβώς μετά το τελευταίο συνελικτικό στρώμα ενός ΣΝΔ, όπως φαίνεται στην εικόνα 39. Σε αυτό το στρώμα ο χάρτης χαρακτηριστικών, που προέκυψε στα συνελικτικά επίπεδα, χωρίζεται σε έναν προκαθορισμένο αριθμό κελιών στα οποία εκτελείται για το καθένα υποδειγματοληψία μέγιστης τιμής (max pooling). Έπειτα το αποτέλεσμα ομαδοποιείται και σχηματίζει ένα τελικό διάνυσμα χαρακτηριστικών σταθερών διαστάσεων, το οποίο με την σειρά του τροφοδοτείται στα πλήρως διασυνδεμένα επίπεδα.



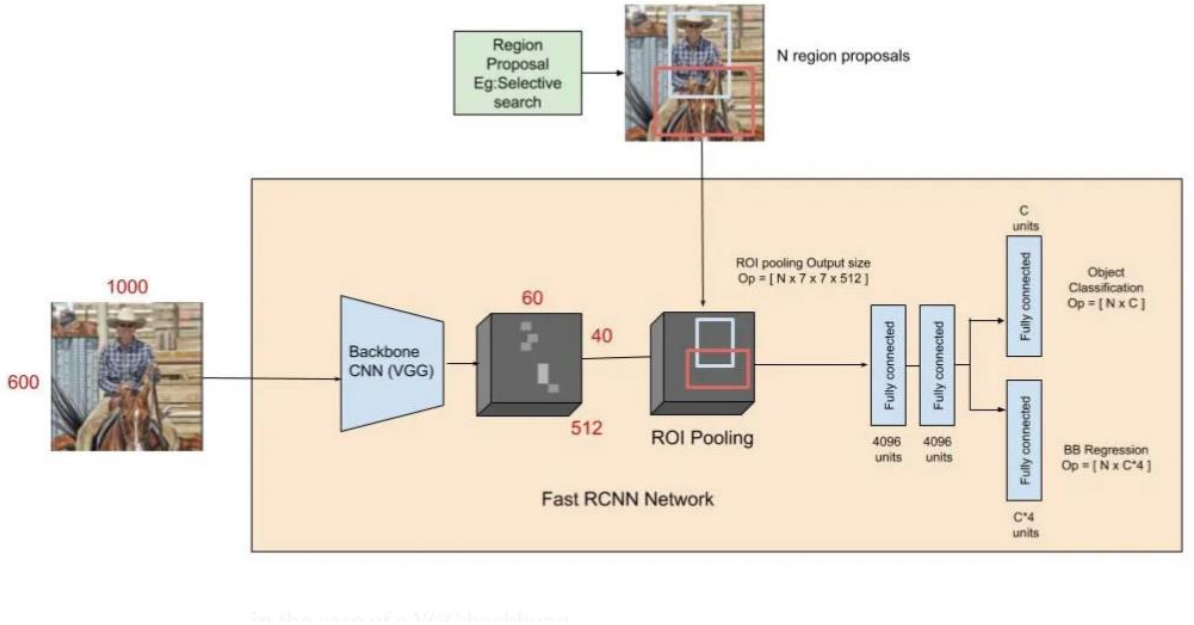
Εικόνα 42: Αναπαράσταση ενός συνελικτικού μοντέλου με την προσθήκη ενός στρώματος πυραμίδας SPP (SPPNet). Πηγή (He et al., 2015).

Οι ερευνητές εφάρμοσαν την συγκεκριμένη μέθοδο σε μοντέλα βαθιάς μάθησης αλλά και σε αλγορίθμους ανίχνευσης, συμμετέχοντας στον διαγωνισμό ILSVRC 2014 διακρίθηκαν στην δεύτερη θέση για τον εντοπισμό αντικειμένων και στην τρίτη θέση για την ταξινόμηση εικόνων. Σε μια από αυτές τις περιπτώσεις η μέθοδος χρησιμοποιήθηκε στον αλγόριθμο R-CNN καθιστώντας τον εκατό φορές γρηγορότερο από την αρχική του έκδοση (He et al., 2015).

5.2.3 Fast R-CNN

Το μεγαλύτερο μειονέκτημα του R-CNN είναι ο μεγάλος χρόνος εκπαίδευσης καθώς πρέπει να εξεταστούν 2000 προτεινόμενες περιοχές. Επιπρόσθετα η περαιτέρω βελτίωση της τεχνικής SPP στον R-CNN είναι αδύνατη, διότι η εκπαίδευση του R-CNN πραγματοποιείται με τον αλγόριθμο ανάστροφης μετάδοσης λάθους (Back Propagation) (Leung & Haykin, 1991) με αποτέλεσμα η ενημέρωση των βαρών πριν από το στρώμα πυραμίδας να μην μπορεί να επιτευχθεί. Έτσι το 2015 ο Girshick παρουσίασε έναν καινούργιο αλγόριθμο που ονομάζεται Fast R-CNN (Girshick, 2015) ο οποίος συνδυάζει τα πλεονεκτήματα των προηγουμένων R-CNN και SPP.

Το νέο μοντέλο αποτελείται από ένα προεκπαιδευμένο δίκτυο VGG-16 όπου το τελευταίο επίπεδο υποδειγματοληψίας του, αντικαθίσταται από ένα στρώμα ROI (Region Of Interest) ενώ προστέθηκαν δυο ξεχωριστά στρώματα εξόδου μετά τα πλήρως διασυνδεμένα επίπεδα. Ο αλγόριθμος Fast R-CNN λαμβάνει ως είσοδο μια ολόκληρη εικόνα και ένα σύνολο προτάσεων αντικειμένων. Πρώτα το ΣΝΔ (VGG) επεξεργάζεται την εικόνα για να δημιουργήσει τον χάρτη χαρακτηριστικών. Έπειτα ένα στρώμα περιοχής ενδιαφέροντος ROI εξάγει ένα διάνυσμα χαρακτηριστικών σταθερού μεγέθους για κάθε πρόταση αντικειμένου. Το στρώμα περιοχής ενδιαφέροντος ROI αποτελεί μια ειδική περίπτωση και λειτουργεί παρόμοια με ένα στρώμα πυραμίδας SPP. Τελικά, τα παραγόμενα διανύσματα οδηγούνται στα ξεχωριστά στρώματα εξόδου. Στην πρώτη έξοδο πραγματοποιείται η πρόβλεψη μέσω ενός στρώματος Softmax ενώ στο δεύτερο στρώμα υπολογίζονται οι ακριβείς τιμές για τα πλαίσια οριοθέτησης.

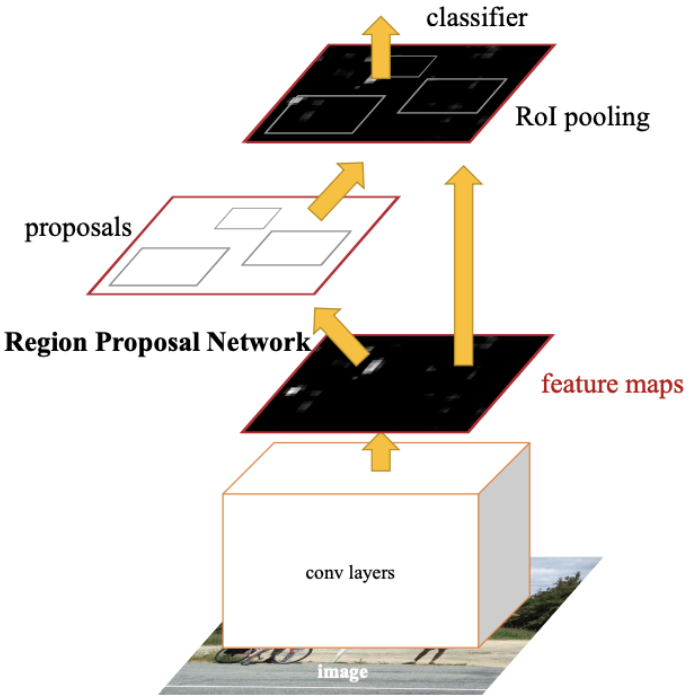


Εικόνα 43: Αναπαράσταση λειτουργίας του Fast R-CNN. Πηγή towardsdatascience.com.

5.2.4 Faster R-CNN

Σε όλες τις προηγούμενες περιπτώσεις αλγορίθμων που αναλύθηκαν, ο τρόπος εύρεσης των υποψήφιων περιοχών προς αναγνώριση είναι κοινός. Πρόκειται για την μέθοδο της Επιλεκτικής Αναζήτησης (Selective Search), η οποία είναι αρκετά χρονοβόρα. Για τον λόγο αυτό το 2016 ο Ren Shaoqing μαζί με την επιστημονική του ομάδα πρότειναν μια σημαντική βελτίωση στον Fast R-CNN που εγκαταλείπει την ιδέα της επιλεκτικής αναζήτησης αλλά όχι την αρχιτεκτονική του. Ο νέος αλγόριθμος ονομάστηκε Faster R-CNN (Ren et al., 2016).

Η βελτίωση που πέτυχαν εστιάζεται στην λειτουργία ενός δικτύου RPN (Region Proposal Network) που είναι υπεύθυνο για την δημιουργία προτεινόμενων περιοχών. Αυτό το δίκτυο είναι συνδεμένο με την έξοδο των συνελικτικών στρωμάτων του ΣΝΔ, ώστε να επεξεργαστεί τον χάρτη χαρακτηριστικών. Στο εσωτερικό του δικτύου πραγματοποιούνται συνελίξεις ενώ η έξοδός του υπολογίζει περίπου 300 υποψήφια πλαίσια οριοθέτησης και περιέχει πληροφορίες σχετικά με την πιθανότητα ύπαρξης κάποιου αντικειμένου. Η διαδικασία αναγνώρισης είναι αρκετά όμοια με αυτήν του Fast R-CNN και περιγράφεται στην εικόνα 44. Ο αλγόριθμος λαμβάνει ως είσοδο μια εικόνα ανεξάρτητου μεγέθους, το ΣΝΔ (VGG) υπολογίζει τον χάρτη χαρακτηριστικών και μετά το RPN προβλέπει τις προτεινόμενες περιοχές. Στην συνέχεια, η πληροφορία διοχετεύεται στο στρώμα περιοχής ενδιαφέροντος ROI και η διαδικασία ολοκληρώνεται όπως στον Fast R-CNN. Ο αλγόριθμος Faster R-CNN αποτελεί την ταχύτερη υλοποίηση με ποσοστό 73.2% mAP και μπορεί να χρησιμοποιηθεί σε πραγματικό χρόνο αφού χρειάζεται κάτι λιγότερο από 0.2 δευτερόλεπτα για να αναγνωρίσει διάφορα αντικείμενα.



Εικόνα 44: Αναπαράσταση λειτουργίας του Faster R-CNN. Πηγή paperswithcode.com

5.3 Αλγόριθμοι Ενός Σταδίου

5.3.1 OverFeat

Η ιδέα χρήσης ενός και μόνο ΣΝΔ που θα έχει την δυνατότητα να υλοποιεί όλες τις εργασίες μηχανικής όρασης, δηλαδή εντοπισμός και ταξινόμηση, εμφανίστηκε το 2013 με την κυκλοφορία του αλγορίθμου Overfeat (Sermanet et al., 2014). Το κύριο σημείο της εργασίας είναι ότι η εκπαίδευση ενός ΣΝΔ σε όλες τις εργασίες, μπορεί να αυξήσει την συνολική απόδοσή του χρησιμοποιώντας την τεχνική της πολλαπλής ταξινόμησης, κατά την οποία ο αλγόριθμος εξετάζει την ίδια εικόνα σε διαφορετικές διαστάσεις για τουλάχιστον πέντε φορές.

Η διαδικασία της ταξινόμησης των εικόνων βασίζεται στο AlexNet. Οι ερευνητές παρουσίασαν δυο εκδόσεις εκ των οποίων η πρώτη παρουσιάζεται στον πάνω πίνακα της εικόνας 45 και αποτελεί την γρήγορη έκδοση, ενώ η έκδοση με την μεγαλύτερη ακρίβεια βρίσκεται στον κάτω πίνακα της ίδιας εικόνας. Η βασική διαφορά ανάμεσα στα δυο μοντέλα είναι το μέγεθος του φίλτρου και η τιμή του διασκελισμού (stride) που χρησιμοποιούνται στην συνέλιξη. Η γρήγορη έκδοση του αλγορίθμου χρησιμοποιεί μεγάλες τιμές για τις προηγούμενες παραμέτρους. Το αποτέλεσμα είναι η δημιουργία ενός μικρού χάρτη χαρακτηριστικών σε σύγκριση με τον χάρτη που παράγει η έκδοση με την μεγαλύτερη ακρίβεια. Επιπλέον οι διαστάσεις των επιπέδων υποδειγματοληψίας έχουν μειωθεί.

Layer	1	2	3	4	5	6	7	Output 8
Stage	conv + max	conv + max	conv	conv	conv + max	full	full	full
# channels	96	256	512	1024	1024	3072	4096	1000
Filter size	11x11	5x5	3x3	3x3	3x3	-	-	-
Conv. stride	4x4	1x1	1x1	1x1	1x1	-	-	-
Pooling size	2x2	2x2	-	-	2x2	-	-	-
Pooling stride	2x2	2x2	-	-	2x2	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	231x231	24x24	12x12	12x12	12x12	6x6	1x1	1x1

Layer	1	2	3	4	5	6	7	8	Output 9
Stage	conv + max	conv + max	conv	conv	conv	conv + max	full	full	full
# channels	96	256	512	512	1024	1024	4096	4096	1000
Filter size	7x7	7x7	3x3	3x3	3x3	3x3	-	-	-
Conv. stride	2x2	1x1	1x1	1x1	1x1	1x1	-	-	-
Pooling size	3x3	2x2	-	-	-	3x3	-	-	-
Pooling stride	3x3	2x2	-	-	-	3x3	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	221x221	36x36	15x15	15x15	15x15	15x15	5x5	1x1	1x1

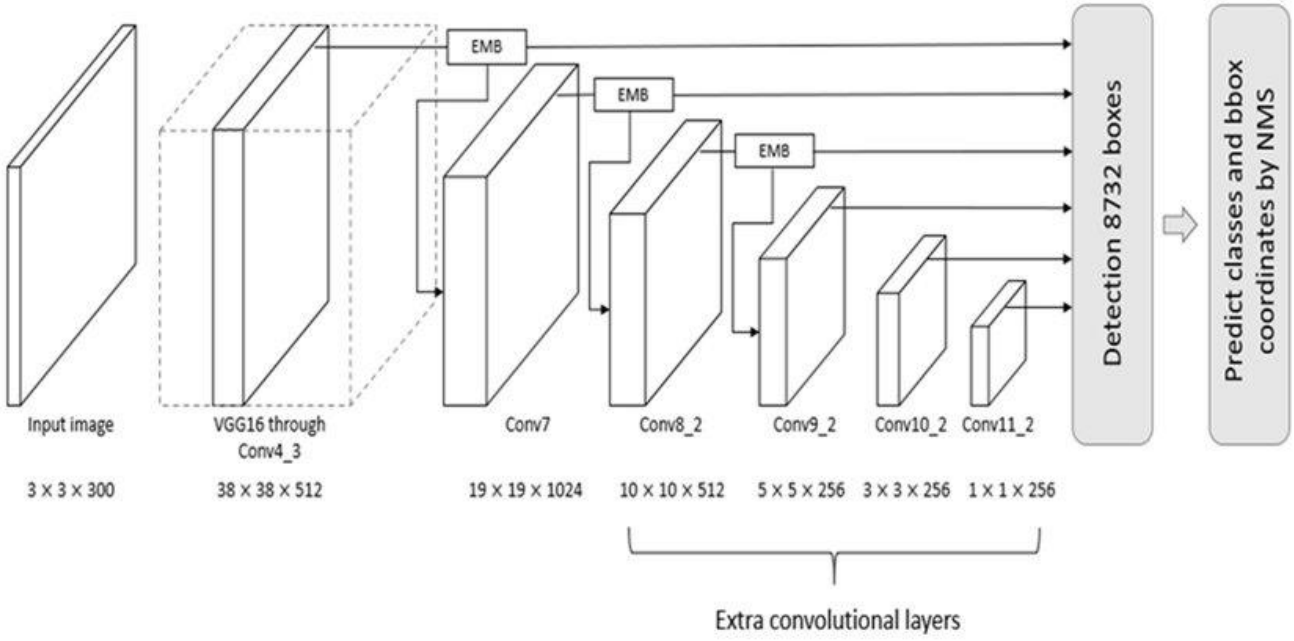
Εικόνα 45: Αναπαράσταση αρχιτεκτονικής του αλγορίθμου OverFeat. Ο πρώτος πίνακας ανήκει στην γρήγορη έκδοση και ο δεύτερος στην έκδοση με μεγαλύτερη ακρίβεια. Πηγή (Sermanet et al., 2014).

Για την διαδικασία του εντοπισμού και της ανίχνευσης οι ερευνητές τροποποίησαν το δίκτυο AlexNet αφαιρώντας το επίπεδο ταξινόμησης, δηλαδή το στρώμα εξόδου Softmax, και στην θέση του τοποθέτησαν δυο πλήρως συνδεδεμένα επίπεδα τα οποία παράγουν τις συντεταγμένες των πλαισίων οριοθέτησης και το τελικό αποτέλεσμα. Ο Overfeat έχει μεγάλο πλεονέκτημα στην ταχύτητα αλλά είναι λιγότερο ακριβής σε σχέση με τον αντίπαλό του εκείνη την εποχή, τον R-CNN. Τέλος αναδείχτηκε στην πρώτη θέση για το πρόβλημα εντοπισμού αντικειμένων του διαγωνισμού ILSVRC 2013.

5.3.2 SSD: Single Shot MultiBox Detector

Ο αλγόριθμος SSD παρουσιάστηκε το 2016 πετυχαίνοντας πολύ υψηλές επιδόσεις φτάνοντας το ποσοστό 76.8% mAP στο PASCAL VOC 2007 (W. Liu et al., 2016). Μολονότι η επίδοσή του είναι ελάχιστα καλύτερη από αυτήν του Faster R-CNN, ο SSD δεν παρουσιάζει ομοιότητες με τον τελευταίο αφού εκτελεί προβλέψεις με ένα απλό ΣΝΔ.

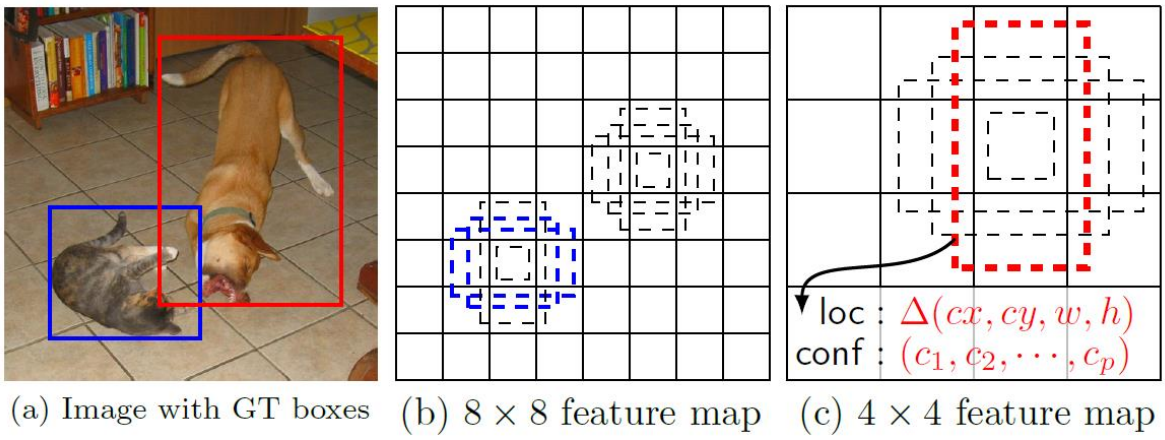
Το βασικό ΣΝΔ, που χρησιμοποιεί ο αλγόριθμος SSD, είναι πλήρως συνελικτικό όπου τα πρώτα επίπεδά του βασίζονται στο μοντέλο VGG-16 και στην συνέχεια ακολουθούν πολλά βοηθητικά επίπεδα συνέλιξης των οποίων οι διαστάσεις μειώνονται προοδευτικά σε μέγεθος. Με αυτόν τον τρόπο επιτυγχάνεται η εξαγωγή χαρακτηριστικών σε διαφορετικές κλίμακες της εικόνας και οι χάρτες χαρακτηριστικών συγκεντρώνουν μεγάλο όγκο πληροφορίας. Το γεγονός αυτό βοηθάει στον εντοπισμό αντικειμένων οποιουδήποτε μεγέθους.



Εικόνα 46: Αναπαράσταση αρχιτεκτονικής του SSD. Πηγή (W. Liu et al., 2016).

Η βασική λεπτομέρεια του SSD είναι ότι παράγει μια συλλογή από τέσσερεις προβλέψεις μαζί με κάθε χάρτη χαρακτηριστικών που δημιουργείται από τα βοηθητικά στρώματα. Κάθε πρόβλεψη περιλαμβάνει ένα πλαίσιο οριοθέτησης με συγκεκριμένες συντεταγμένες και την κλάση για το αντικείμενο.

Για παράδειγμα στην εικόνα 47 ο χάρτης χαρακτηριστικών 8x8 έχει εξάγει τέσσερα προκαθορισμένα πλαίσια οριοθέτησης για τον εντοπισμό της γάτας ενώ ο χάρτης χαρακτηριστικών 4x4 ακολουθεί την ίδια διαδικασία για τον εντοπισμό του σκύλου. Το τελικό αποτέλεσμα υπολογίζεται από ένα μικρό δίκτυο συνέλιξης 3x3 το οποίο επιλέγει τις καλύτερες προβλέψεις.



Εικόνα 47: Παράδειγμα υπολογισμού πλαισίων οριοθέτησης με τον αλγόριθμο SSD. Πηγή (W. Liu et al., 2016).

5.3.3 YOLO: You Only Look Once

Ο αλγόριθμος YOLO και οι μετέπειτα βελτιώσεις του αποτελούν την πιο δημοφιλή οικογένεια αλγορίθμων στα πλαίσια της αναγνώρισης αντικειμένων (Redmon et al., 2016). Αναπτύχθηκε από τον Joseph Redmon μαζί με την ερευνητική του ομάδα και συνεχίζει να εξελίσσεται με γοργούς ρυθμούς μέχρι σήμερα. Η προσέγγιση του αλγορίθμου περιλαμβάνει ένα ΣΝΔ το οποίο αντιμετωπίζει την αναγνώριση αντικειμένων ως πρόβλημα παλινδρόμησης. Η ιδιαιτερότητα που τον ξεχωρίζει από τα υπόλοιπα μοντέλα είναι η ταχύτητά του σε συνδυασμό με την δυνατότητα υποστήριξης υψηλού αριθμού FPS (Frames Per Second) και υψηλής ακρίβειας, καθιστώντας τον κατάλληλο για αναγνώριση σε πραγματικό χρόνο.

Στην πρώτη έκδοση YOLO v1, ο αλγόριθμος δέχεται μια εικόνα και την διαιρεί με ένα πλέγμα ορθογωνίων κελιών του οποίου οι διαστάσεις είναι *SxS*. Για κάθε ορθογώνιο κελί ο αλγόριθμος προβλέπει πλαίσια οριοθέτησης *B* και τον βαθμό εμπιστοσύνης, είτε υπάρχει είτε όχι ένα αντικείμενο μέσα σε αυτό. Ο βαθμός εμπιστοσύνης είναι ενδεικτικός της απουσίας ή της παρουσίας ενός αντικειμένου μέσα στο κελί. Επιπρόσθετα, κάθε πρόβλεψη περιέχει τιμές που σχετίζονται με τις συντεταγμένες του κέντρου σε σχέση με το κελί, τιμές που σχετίζονται με τις διαστάσεις του πλαισίου οριοθέτησης και τέλος τιμές για την κλάση *C* που ανήκει το αντικείμενο. Στην περίπτωση που το κέντρο ενός αντικειμένου ενδιαφέροντος βρίσκεται σε ένα συγκεκριμένο κελί, τότε μόνο αυτό είναι υπεύθυνο για την ανίχνευση του αντικειμένου.

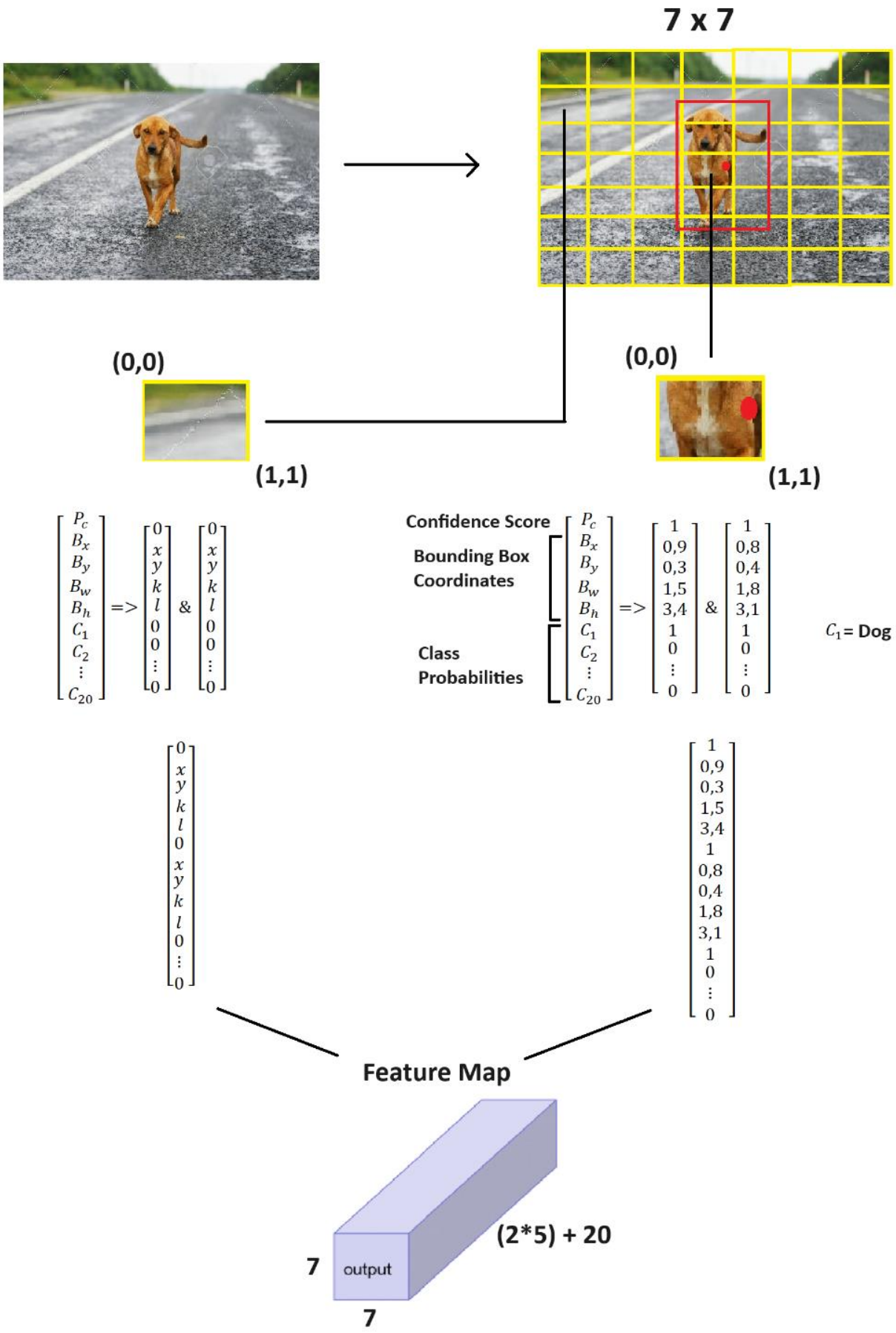
Η αρχιτεκτονική που χρησιμοποιεί ο YOLO v1 βασίζεται στο GoogLeNet και η απόδοση του μετρήθηκε στο PASCAL VOC 2007–12 με ποσοστό 63.4% mAP στα 45 FPS και 52.7% mAP στα 155 FPS. Σύμφωνα με τις παραπάνω μετρήσεις, οι συγγράφεις προτείνουν συγκεκριμένα νούμερα για τις προηγούμενες παραμέτρους: S = 7, B = 2 και C = 20. Για την καλύτερη κατανόηση του αλγορίθμου, έχει δημιουργηθεί η εικόνα 48.

Στο παράδειγμα της επόμενης σελίδας, η εικόνα χωρίζεται με ένα πλέγμα από κελιά διαστάσεων 7x7, περιέχοντας συνολικά 49 κελιά. Για κάθε κελί πραγματοποιούνται δυο προβλέψεις για πλαίσια οριοθέτησης. Την ίδια στιγμή, κάθε πρόβλεψη περιέχει ένα διάνυσμα όπου η τιμή P_c αντιστοιχεί στον βαθμό εμπιστοσύνης, τα B_x και B_y είναι η απόσταση του κέντρου (αν υπάρχει) από το σημείο (0,0), τα B_w και B_h είναι η διαστάσεις του πλαισίου οριοθέτησης και οι μεταβλητές C_1 έως C_{20} αντιστοιχούν στις 20 κλάσεις που βρίσκονται στο PASCAL VOC. Οι συγκεκριμένες μεταβλητές κλάσεων παίρνουν μόνο δυαδικές τιμές. Ο βαθμός εμπιστοσύνης σχετίζεται με τον λόγο επικάλυψης του αντικειμένου και υπολογίζεται από την σχέση:

$$\mathrm{Pc} = p\,(object) * IoU_{pred}^{truth} \quad (20)$$

Το παράδειγμα εξετάζει δυο κελιά. Στην πρώτη περίπτωση το κελί δεν περιέχει κάποιο αντικείμενο επομένως και τα δυο διανύσματα που δημιουργούνται έχουν μηδενικό βαθμό εμπιστοσύνης άρα οι υπόλοιπες τιμές δεν έχουν ιδιαίτερη σημασία. Αντίθετα, στην δεύτερη περίπτωση μέσα στο κελί περιέχεται το κέντρο του σκύλου, επομένως υπολογίζονται δυο διανύσματα για όλες τις μεταβλητές που αναφέρθηκαν. Το επόμενο βήμα είναι η συνένωση των δυο διανυσμάτων που υπολογίστηκαν για όλα τα κελιά της φωτογραφίας. Με αυτόν τον τρόπο ο χάρτης χαρακτηριστικών αποκτά διαστάσεις 7x7x30 και στο σύνολό του περιλαμβάνει 1470 προβλέψεις. Ο μαθηματικός τύπος για τον υπολογισμό αυτό είναι:

$$S\ x\ S\ x\ [B\ x\ (5 + C)] \quad (21)$$



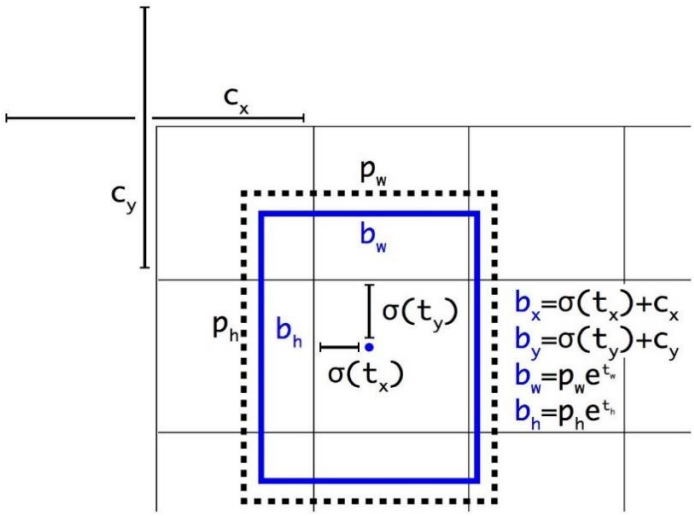
Εικόνα 48: Παράδειγμα υπολογισμού πλαισίων οριοθέτησης του αλγορίθμου YOLO v1 σε τυχαία εικόνα.

5.3.4 YOLO v2

Έναν χρόνο μετά κυκλοφόρησε η πρώτη αναβάθμιση του αλγορίθμου, γνωστή ως YOLO v2, από τους ίδιους ερευνητές (Redmon & Farhadi, 2017). Τα προβλήματα του YOLO v1 είναι η πλήρης αδυναμία ανίχνευσης μικρών αντικειμένων και η ανακριβή τοποθέτηση των πλαισίων οριοθέτησης. Αναλυτικότερα, η πρώτη αλλαγή είναι η χρήση του Batch Normalization (Liao & Carneiro, 2016), δηλαδή η κανονικοποίηση των τιμών εξόδου σε όλα τα στρωματά επεξεργασίας. Με αυτήν την εισαγωγή ο YOLO v2 βελτίωσε το mAP σε σύγκριση με την αρχική έκδοση και εξάλειψε την ανάγκη χρήσης τεχνικών για την μείωση της υπερπροσαρμογής (Overfitting) (G. Li et al., 2022), όπως η πρόωρη εγκατάλειψη (Garbin et al., 2020). Η αρχική έκδοση του YOLO δέχεται ως είσοδο εικόνες με ανάλυση 224x224 pixel κατά το στάδιο της εκπαίδευσης, ενώ στην φάση της ανίχνευσης ο αλγόριθμος δέχεται εικόνες έως και 448x448 pixel. Αυτό υποχρεώνει τον αλγόριθμο να προσαρμοστεί σε μεταβαλλόμενη ανάλυση εικόνας και τελικά μειώνει την απόδοσή του. Για την λύση αυτού του προβλήματος, οι ερευνητές εκπαίδευσαν τον νέο YOLO v2 σε εικόνες με ανάλυση 448x448 pixel για μόνο 10 εποχές στο ImageNet.

Πολλές αλλαγές έγιναν και στον τρόπο ανίχνευσης των αντικειμένων. Στο νέο μοντέλο αντικαταστάθηκαν τα πλήρως συνδεμένα επίπεδα και πλέον το δίκτυο αποτελείται από συνελικτικά επίπεδα και επίπεδα υποδειγματοληψίας. Για την πρόβλεψη των πλαισίων οριοθέτησης προστέθηκαν τα κουτιά αγκύρωσης, δηλαδή μια λίστα από ορθογώνια κουτιά με προκαθορισμένες διαστάσεις, παρόμοιας φιλοσοφίας με αυτά του SSD. Το ενδιαφέρον εντοπίζεται στον τρόπο με τον οποίο υπολογίζονται οι διαστάσεις των συγκεκριμένων κουτιών. Οι συγγραφείς χρησιμοποίησαν στα δεδομένα εκπαίδευσης τον αλγόριθμο ομαδοποίησης k-means (Ahmed et al., 2020) σε συνδιαμσό με τον λόγο επικάλυψης IoU για τον υπολογισμό τους.

Σχετικά με την διαδικασία πρόβλεψης των συντεταγμένων του κέντρου, ο YOLO v1 δεν έχει περιορισμούς και προχωράει απευθείας στον υπολογισμό. Αντίθετα στην δεύτερη έκδοση, η πρόβλεψη του κέντρου ενός αντικειμένου πραγματοποιείται με την χρήση μιας σιγμοειδούς συνάρτησης. Έτσι οι νέες συντεταγμένες παίρνουν τιμές από 0 έως 1 και υπολογίζονται με τους μαθηματικούς τύπους της εικόνας 49. Οι μεταβλητές t_x, t_y, p_w και p_h εξάγονται από το δίκτυο. Οι προβλέψεις γίνονται για όλα τα κελιά, τα οποία έχουν αυξηθεί σε 13x13, δηλαδή στο σύνολό τους 169.



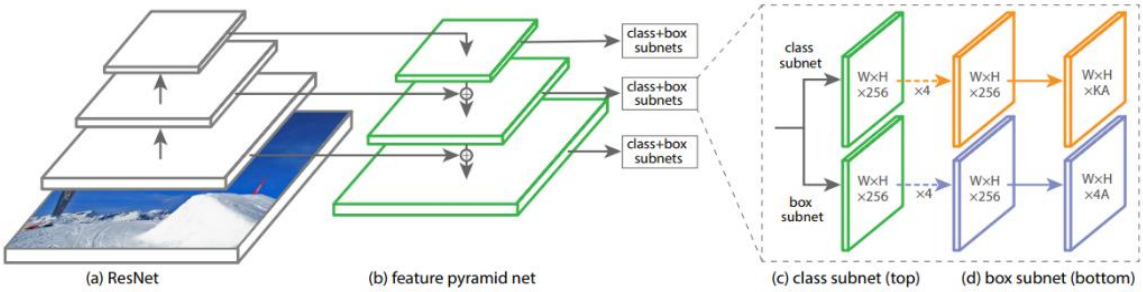
Εικόνα 49: Διαδικασία υπολογισμού των πλαισίων οριοθέτησης του αλγορίθμου YOLO v2. Πηγή (Redmon & Farhadi, 2017).

Ο YOLO v2 εκπαιδεύτηκε σε τρεις αρχιτεκτονικές, συγκεκριμένα στα VGG-16 και GoogLeNet, αλλά και στην καινούργια αρχιτεκτονική Darknet-19 που προτάθηκε στην συγκεκριμένη εργασία. Όσον αφορά την απόδοση, τα ποσοστά έχουν αυξηθεί σε μεγάλο βαθμό με 76.8% mAP στα 67 FPS και 78.6% mAP στα 40 FPS, ξεπερνώντας όλους τους προηγούμενους αλγορίθμους. Τέλος, στο ίδιο επιστημονικό άρθρο παρουσιάστηκε το μοντέλο YOLO9000 που βασίζεται στο YOLO v2 και έχει στόχο την ανίχνευση 9000 διαφορετικών αντικειμένων σε πραγματικό χρόνο. Ωστόσο δεν έλαβε την αναγνώριση που του αναλογεί εξαιτίας της μειωμένης απόδοσής του.

5.3.5 RetinaNet

Ο αλγόριθμος RetinaNet παρουσιάστηκε μετά την κυκλοφορία των YOLO v2 και SSD αποτελώντας την πρώτη ουσιαστική λύση στο πρόβλημα εντοπισμού μικρών και πυκνών αντικειμένων. Ο αλγόριθμος έρχεται σε τρεις εκδόσεις οι οποίες διαφέρουν στον αριθμό των στρωμάτων, ενώ η απόδοση της καλύτερης έκδοσης μετρήθηκε στο σύνολο δεδομένων MS COCO με ποσοστό 39.1% mAP (Lin, Goyal, et al., 2017). Εκείνη την χρονική περίοδο ο RetinaNet ξεπέρασε όλους τους υπολοίπους ανιχνευτές ανεξαρτήτως κατηγορίας, δηλαδή είτε ανιχνευτές ενός είτε δυο σταδίων. Οι βασικές καινοτομίες του αλγορίθμου εμφανίζονται στον πυρήνα της αρχιτεκτονικής του, όπου εκεί υιοθετείται το Feature Pyramid Network (FPN) (Lin, Dollar, et al., 2017), σε συνδυασμό με τρεις εκδόσεις του μοντέλου ResNet (ResNet-50, ResNet-101, ResNet-152).

Το FPN είναι ένα υποδίκτυο εξαγωγής χαρακτηριστικών, πλήρως ανεξάρτητο από το βασικό δίκτυο (ResNet), το οποίο λαμβάνει εικόνες ανεξάρτητου μεγέθους και δημιουργεί χάρτες χαρακτηριστικών πολλών επιπέδων σε μορφή πυραμίδας. Χρησιμοποιείται κυρίως για την εύρεση αντικειμένων με μικρές διαστάσεις και χαμηλή ανάλυση. Η διαδικασία δημιουργίας ενός χάρτη χαρακτηριστικών από τον RetinaNet υλοποιείται με δυο διαδρομές: α) διαδρομή από τα κάτω προς τα πάνω στρώματα (bottom-up) και β) διαδρομή από τα πάνω προς τα κάτω στρώματα (top-down). Στην πρώτη περίπτωση εκτελούνται οι συνηθισμένοι υπολογισμοί από τα στρώματα συνέλιξης του δικτύου ResNet και έπειτα πραγματοποιούνται οι υπολογισμοί στο υποδίκτυο FPN. Αφού ολοκληρωθεί το πρώτο πέρασμα, ξεκινάει η διαδικασία της δεύτερης περίπτωσης, όπου η πληροφορία από τα τελευταία στρώματα της πυραμίδας μεταφέρεται και συγχωνεύεται στα χαμηλότερα επιθυμητά επίπεδα μέσω πλευρικών συνδέσεων. Εκτός από τους χάρτες χαρακτηριστικών, υπολογίζονται και τα πλαίσια αγκύρωσης για όλα τα επίπεδα του FPN μαζί με την τιμή της πιθανότητας κατανομής για τις κλάσεις των αντικειμένων. Ο αλγόριθμος επιλέγει μέχρι 1000 πλαίσια αγκύρωσης από κάθε επίπεδο με την μεγαλύτερη βαθμολογία εμπιστοσύνης (>0.5) και στη συνέχεια χρησιμοποιεί το υποδίκτυο παλινδρόμησης για την διαδικασία της πρόβλεψης.



Εικόνα 50: Αναπαράσταση αρχιτεκτονικής του αλγορίθμου RetinaNet. Πηγή (Lin, Goyal, et al., 2017).

Πέρα από την χρήση του FPN για τον εντοπισμό δύσκολων αντικειμένων, εισάγεται και η έννοια της εστιακής απόκλισης (Focal Loss) για την εκπαίδευση του αλγορίθμου σε δύσκολα παραδείγματα. Η εστιακή απόκλιση σχεδιάστηκε για την αντιμετώπιση ενός προβλήματος που εντοπίζεται σε όλους τους ανιχνευτές ενός σταδίου, κατά το οποίο εμφανίζεται μια ανισορροπία στην διασταυρούμενη εντροπία (Cross Entropy) των αντικείμενα που βρίσκονται στο προσκήνιο σε σχέση με αυτά που βρίσκονται στο βάθος μιας εικόνας. Δηλαδή, τα εύκολα προς ανίχνευση αντικείμενα παρότι έχουν μικρές τιμές απώλειας, μπορούν να κατακλύσουν συνολικά τους χάρτες χαρακτηριστικών καλύπτοντας τα αντικείμενα με ιδιαίτερη δυσκολία ανίχνευσης. Έτσι, οι ερευνητές αναδιαμόρφωσαν την συνάρτηση απώλειας με την χρήση της εστιακής απόκλισης, εστιάζοντας σε αντικείμενα με χαμηλή τιμή πιθανότητας κατανομής.

5.3.6 YOLO v3

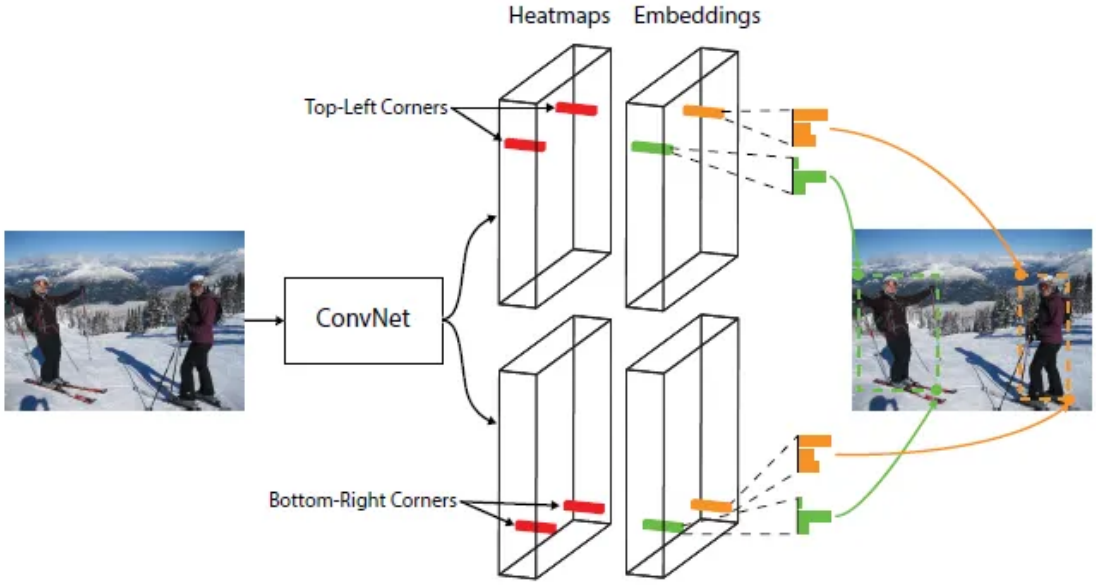
Στην τρίτη έκδοση του αλγορίθμου YOLO παρουσιάστηκαν μικρές αλλά αξιόλογες αλλαγές πάνω στον YOLO v2, με σκοπό την βελτίωση της ακρίβειας (Redmon & Farhadi, 2018). Οι ερευνητές υλοποίησαν ξανά αλλαγές στην αρχιτεκτονική του δικτύου, χρησιμοποιώντας την νέα αρχιτεκτονική Darknet-53. Πλέον, το δίκτυο είναι πλήρως συνελικτικό και το νέο μοντέλο περιλαμβάνει συνολικά 106 στρώματα συνέλιξης. Τα πρώτα 53 προέρχονται από το Darknet-53 και χρησιμοποιούνται για την εξαγωγή χαρακτηριστικών ενώ τα υπόλοιπα 53 προστέθηκαν για την ανίχνευση. Ο YOLO v3 πραγματοποιεί προβλέψεις σε τρεις χάρτες χαρακτηριστικών με διαφορετικές κλίμακες. Οι συγκεκριμένοι χάρτες δημιουργούνται από επίπεδα που χρησιμοποιούν διαφορετικό βήμα (stride) στην μετατόπιση των φίλτρων. Για παράδειγμα, το δίκτυο θα διαπεράσει μια εικόνα με διαστάσεις 416x416 pixel τρεις φορές, με βήμα 32,16 και 8 δημιουργώντας χάρτες χαρακτηριστικών με διαστάσεις 13x13, 26x26 και 52x52 αντίστοιχα. Αυτή η μέθοδος χρησιμοποιείται για τον εντοπισμό μεγάλων, μεσαίων και μικρών αντικειμένων.

Κάθε κελί παράγει τρία πλαίσια οριοθέτησης και τελικά οι προβλέψεις γίνονται από συνελίξεις 1x1. Η απόδοση του YOLO v3 μετρήθηκε στο MS COCO dataset για 80 κλάσεις αντικειμένων, πετυχαίνοντας ποσοστό 28.2% mAP με ταχύτητα 22 ms. Χρησιμοποιώντας τον τύπο 15, οι διστάσεις των χαρτών που προκύπτουν για το παράδειγμα της εικόνας 416x416 είναι: (13x13x255), (26x26x255), (52x52x255). Επειδή το τελικό νούμερο των προβλέψεων είναι μεγάλο, ο YOLO v3 χρησιμοποιεί δυο τεχνικές για την μείωση των υπολογισμών. Ο αλγόριθμος θέτει ένα ελάχιστο κατώφλι στον βαθμό εμπιστοσύνης, δηλαδή ο υπολογισμός μιας πρόβλεψης ολοκληρώνεται μόνο όταν η βαθμολογία της είναι μεγαλύτερη από το κατώφλι. Ακόμα, στην περίπτωση που αλγόριθμος προβλέψει πολλά διαφορετικά πλαίσια για ένα αντικείμενο, τότε χρησιμοποιεί την τεχνική της μη μέγιστης καταστολής (Non-maximum Suppression) και επιλεγεί αυτό με την καλύτερη βαθμολογία.

5.3.7 CornerNet

Ο πρωτοποριακός αλγόριθμος CornerNet παρουσιάστηκε το 2018 και επιχείρησε να αμφισβητήσει τον κυρίαρχο ρόλο των κουτιών αγκύρωσης που χρησιμοποιούν όλοι οι προηγούμενοι ανιχνευτές (Law & Deng, 2018). Σύμφωνα με τους συγγραφείς, η χρήση των κουτιών αγκύρωσης, ειδικά σε ανιχνευτές ενός σταδίου, έχει μειονεκτήματα όπως το μεγάλο ποσοστό αναλογίας μεταξύ θετικών και αρνητικών παραδειγμάτων και η εισαγωγή πολλών υπερπαραμέτρων, επιβραδύνοντας την εκπαίδευση του μοντέλου.

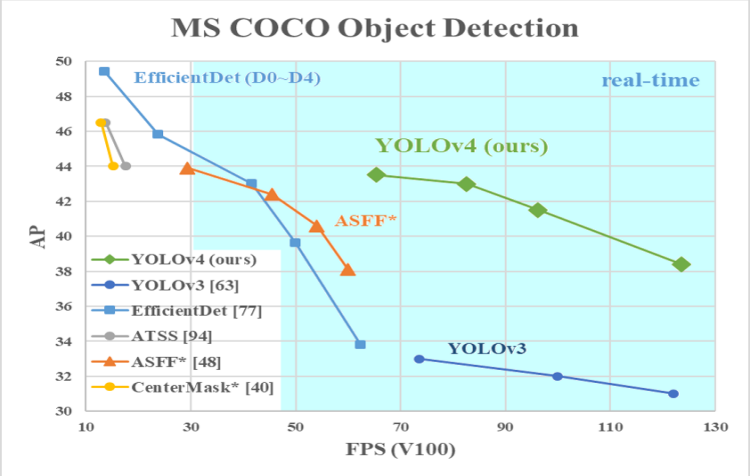
Ο αλγόριθμος ανιχνεύει ένα πλαίσιο οριοθέτησης ως ένα ζεύγος σημείων – κλειδιών, αποτελούμενο από την πάνω αριστερή και κάτω δεξιά γωνιά, εξαλείφοντας την ανάγκη σχεδιασμού ενός συνόλου προκαθορισμένων πλαισίων αγκύρωσης. Οι συγγράφεις εμπνευσμένοι από προβλήματα που σχετίζονται με την εκτίμηση της ανθρώπινης θέσης και στάσης (human pose estimation), χρησιμοποιούν ένα διάνυσμα ενσωμάτωσης για να ομαδοποιήσουν τα σημεία. Για την εύρεση αυτών των σημείων χρησιμοποιούνται δυο ενωμένα δίκτυα Hourglass (Newell et al., 2016), τα οποία αντιπροσωπεύουν τον κορμό του CornerNet. Τα Hourglass Networks αποτελούν έναν νέο τύπο συνελικτικού δικτύου που υιοθετούν μια απλή μέθοδο υποδειγματοληψίας, γνωστή ως Corner Pooling. Ο CornerNet πέτυχε το εντυπωσιακό ποσοστό 42,1% στο σύνολο MS COCO, αλλά παρόλα αυτά ήταν σημαντικά πιο αργός από τους SSD και YOLO. Μάλιστα σε αρκετές περιπτώσεις παρατηρήθηκε ότι ο αλγόριθμος δημιουργεί λανθασμένα πλαίσια οριοθέτησης. Έναν χρόνο αργότερα οι ίδιοι ερευνητές παρουσίασαν την βελτιωμένη έκδοση του αλγορίθμου CornerNet-Lite, πετυχαίνοντας ποσοστό 47% mAP στο MS COCO (Law et al., 2020).



Εικόνα 51: Διαδικασία υπολογισμού πλαισίων οριοθέτησης του αλγορίθμου CornerNet. Πηγή (Law & Deng, 2018).

5.3.8 YOLO v4

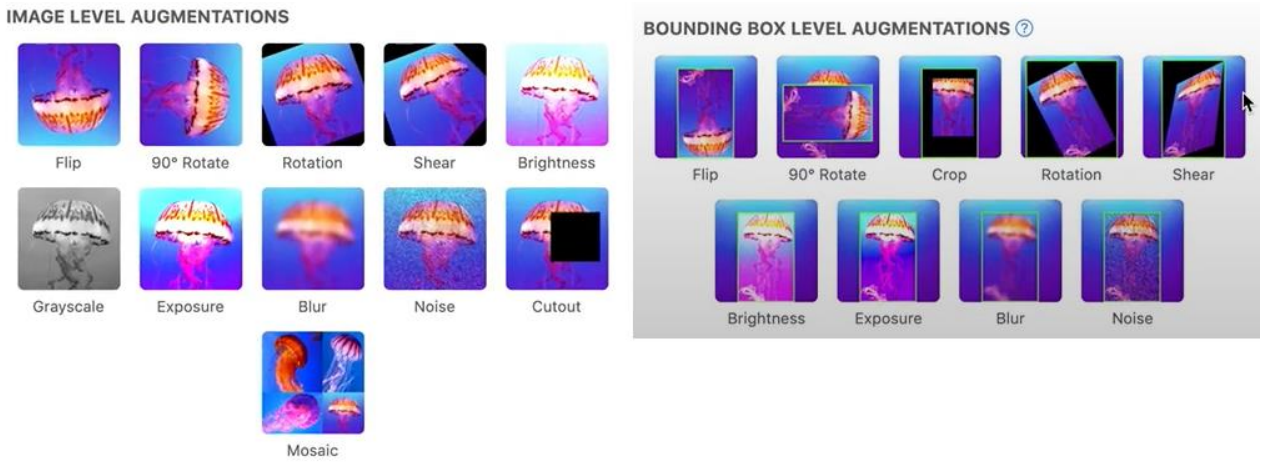
Τον Απρίλιο του 2020, προτάθηκε η τέταρτη έκδοση, YOLO v4, από τον Alexey Bochkovsky. Ήταν η πρώτη επίσημη βελτίωση στην οποία δεν συμμετείχαν οι ερευνητές που δημιούργησαν τον αλγόριθμο. Στο συγκεκριμένο άρθρο διεξάχθηκαν πολλά πειράματα σε διαφορετικές GPU, δοκιμάστηκαν πολλές καινούργιες τεχνικές και τα νέα αποτελέσματα ήταν εξαιρετικά. Η εικόνα 52 είναι ενδεικτική και αποδεικνύει την υπεροχή του YOLO έναντι των υπόλοιπων ανιχνευτών (Bochkovskiy et al., 2020). Μάλιστα, η απόδοσή του έχει αυξηθεί κατά 10% σε σχέση με τον προκάτοχό του. Τα σημαντικά σημεία που θα εξεταστούν παρακάτω είναι η προσθήκη των α) Bag of freebies β) Bag of specials και γ) η αρχιτεκτονική του.



Εικόνα 52: Διάγραμμα απόδοσης mAP σε συνάρτηση με τα FPS του YOLO v4 μαζί με άλλους ανιχνευτές στο MC COCO Dataset. Πηγή (Bochkovskiy et al., 2020).

Bag of freebies

Με τον όρο bag of freebies οι ερευνητές αναφέρονται σε μεθόδους που αλλάζουν πτυχές στην διαδικασία εκπαίδευσης, επηρεάζοντας μόνο το κόστος της. Η πιο συχνά χρησιμοποιούμενη μέθοδος είναι η αύξηση των δεδομένων εκπαίδευσης (data augmentation), χωρίς την προσθήκη νέων εικόνων. Αυτό πραγματοποιείται με φωτομετρικές ή γεωμετρικές παραμορφώσεις στις ήδη υπάρχοντες εικόνες. Μερικά τέτοια παραδείγματα απεικονίζονται στην εικόνα 53 και έχουν να κάνουν με την προσαρμογή της φωτεινότητας, της απόχρωσης, του θορύβου αλλά και με την περικοπή, την αναστροφή και περιστροφή της εικόνας. Ταυτόχρονα υπάρχουν περισσότερο σύνθετες τεχνικές που μπορούν να εφαρμοστούν, όπως οι CutOut, DropOut, DropConnect, DropBlock και MixUp. Ιδιαίτερη έμφαση δόθηκε στην τεχνική Mosaic για την οποία αποδείχτηκε ότι σε συνδυασμό με το Batch Normalization (Liao & Carneiro, 2016) βελτιώνει σημαντικά την απόδοση. Με την χρήση αυτής, συνενώνονται τέσσερεις εικόνες του συνόλου εκπαίδευσης σε μία.



Εικόνα 53: Παραδείγματα τεχνικών επαύξησης των εικόνων εκπαίδευσης. Πηγή roboflow.com.

Bag of specials

Με τον όρο bag of specials οι ερευνητές αναφέρονται σε μεθόδους που βελτιώνουν την συνολική μέση ακρίβεια του μοντέλου, αυξάνοντας το κόστος υπολογισμού της ανίχνευσης. Αυτό μπορεί να επιτευχθεί με την ενσωμάτωση τεχνικών όπως οι SPP (αναλύθηκε στην προηγουμένη ενότητα), SAM, RFB, BiFPN. Ένας άλλος τρόπος είναι χρήση αποδοτικότερων συναρτήσεων ενεργοποίησης. Συγκεκριμένα, ο YOLO v4 χρησιμοποιεί την συνάρτηση ενεργοποίησης Mish (Misra, 2020) στην ραχοκοκαλιά του δικτύου διότι δημιουργεί πλουσιότερους σε πληροφορία χάρτες χαρακτηριστικών. Επιπλέον, μια εξίσου σημαντική προσθήκη είναι η Mini-Batch Normalization (Yao et al., 2021) σε όλα τα υπολογιστικά επίπεδα, γεγονός που επιτρέπει την εκπαίδευση του αλγορίθμου σε οποιαδήποτε GPU.

Αρχιτεκτονική

Οι ερευνητές χωρίζουν την αρχιτεκτονική του αλγορίθμου σε τρεις βασικούς πυλώνες: ο κορμός (Backbone), ο λαιμός (Neck) και η κεφαλή(Head). Ο κορμός καθορίζει κυρίως την ικανότητα εξαγωγής χαρακτηριστικών και ο σχεδιασμός του παίζει κρίσιμο ρόλο στην αποτελεσματικότητα όλου το δικτύου. Ο λαιμός συγκεντρώνει όλα τα χαρακτηριστικά χαμηλού και υψηλού επιπέδου και έπειτα χτίζει τον συγκεντρωτικό χάρτη χαρακτηριστικών. Η κεφαλή χρησιμοποίει τον χάρτη χαρακτηριστικών και πραγματοποιεί την τελική διαδικασία της πρόβλεψης.

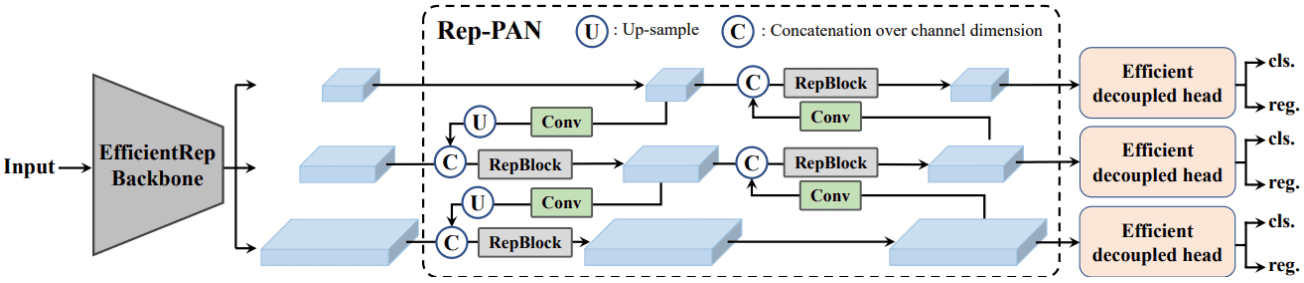
Λεπτομερέστερα, στον κορμό του YOLO v4 χρησιμοποιείται το δίκτυο CSPDarknet53 για την εξαγωγή των χαρακτηριστικών. Η νέα προσθήκη του CSPNet επιτρέπει τον διαχωρισμό της διαδικασίας υπολογισμού σε δύο μέρη και στην συνέχεια την συγχώνευσή τους με σκοπό την καλύτερη κατανόηση των χαρακτηριστικών από το δίκτυο. Στα επιπλέον επίπεδα του λαιμού, ο YOLO v4 χρησιμοποιεί τον SPP σε συνδυασμό με το PAN (Path Aggregation Network) (S. Liu et al., 2018). Το τελευταίο αποτελεί ένα δίκτυο που βελτιώνει την διαδικασία τμηματοποίησης των εικόνων και επιλέχθηκε λόγω της ικανότητάς του να διατηρεί με ακρίβεια χωρικές πληροφορίες. Τέλος, στο κομμάτι της κεφαλής χρησιμοποιούνται τα επίπεδα ανίχνευσης που αναλυθήκαν στον αλγόριθμο YOLO v3, ειδικά σχεδιασμένα για την ανίχνευση αντικειμένων διαφορετικού μεγέθους για τις 80 κλάσεις που περιέχονται στο MS COCO Dataset.

5.3.9 YOLO v5

Τον Ιούνιο του 2020, μόλις δυο μήνες μετά την κυκλοφορία του YOLO v4, ο Glenn Jocher δημιούργησε την πέμπτη έκδοση του αλγορίθμου δημοσιεύοντας τον κώδικα στο αποθετήριο GitHub χωρίς συνοδευόμενο επιστημονικό άρθρο. Ο YOLO v5 (Jocher et al., 2020) είναι παρόμοιος με την προηγουμένη έκδοση αλλά χτισμένος πάνω στην βιβλιοθήκη PyTorch (Jocher et al., 2021), εγκαταλείποντας οριστικά το πλαίσιο του Darknet. Αν και το Darknet είναι περισσότερο ευέλικτος αφού βασίζεται στη γλώσσα προγραμματισμού C, ο συγγραφέας θέλησε να καταργήσει τους περιορισμούς του. Μια επιπλέον αξιοσημείωτη βελτίωση είναι η αυτοματοποιημένη εκμάθηση των κουτιών αγκύρωσης. Ο μηχανισμός που χρησιμοποιήθηκε στους YOLO v2, v3 και v4 με την χρήση του k-means, παρουσιάζει αδυναμίες στην προσαρμογή του ως προς το MS COCO Dataset. Με τον καινούργιο τρόπο το δίκτυο μαθαίνει αυτόματα τα βέλτιστα κουτιά αγκύρωσης για το συγκεκριμένο σύνολο δεδομένων, με αποτέλεσμα την μείωση του χρόνου εκπαίδευσης. Ο YOLO v5 διατίθεται σε διάφορες παραλλαγές και η απόδοσή του εξαρτάται από το πλήθος των παραμέτρων.

5.3.10 YOLO v6

Η έκτη έκδοση του αλγορίθμου κυκλοφόρησε τον Ιούνιο του 2022 από την ερευνητική ομάδα Κινέζων Meituan, όπου οι συγγραφείς επικεντρωθήκαν στην παραγωγή ενός ανιχνευτή σχεδιασμένο για βιομηχανική χρήση (C. Li et al., 2022). Για να καλύψουν τις περισσότερες ανάγκες των βιομηχανικών εφαρμογών, δημιουργήθηκαν διάφορες παραλλαγές του αλγορίθμου, ξεκινώντας από τον YOLO v6 nano ως τον ταχύτερο και καταλήγοντας στον YOLO v6 large υψηλής ακρίβειας. Η εντυπωσιακή απόδοση των YOLO v6 οφείλεται σε ριζικές αλλαγές στην αρχιτεκτονική του αλγορίθμου. Η ερευνητική ομάδα εμπνευσμένη από τις νέες εκδόσεις του VGG, σχεδίασε μια νέα αρχιτεκτονική για τον κορμό του δικτύου, που ονομάζεται EfficientRep. Τα κύρια συστατικά της τελευταίας είναι τα RepBlock, RepConv και CSPStackRep Block, πρόκειται για μπλοκ που περιέχουν συνελίξεις 1x1 και 3x3 σε συνδυασμό με την συνάρτηση ReLu. Στα επίπεδα του λαιμού, υιοθετήθηκε το υποδίκτυο PAN από τις προηγούμενες εκδόσεις και με την προσθήκη μικρών αλλαγών η νέα τοπολογία αναφέρεται ως Rep-PAN (Ding et al., 2021). Ιδιαίτερες αλλαγές εφαρμόστηκαν και στην κεφαλή του δικτύου η οποία χαρακτηρίζεται ως αποσυνδεμένη. Αυτό συνέβη με την προσθήκη νέων στρωμάτων που διαχωρίζουν τα χαρακτηριστικά, σύμφωνα με τις ιδιότητες τους, και στην συνέχεια τα καθοδηγούν κατάλληλα είτε για την διαδικασία της ταξινόμησης είτε της παλινδρόμησης.



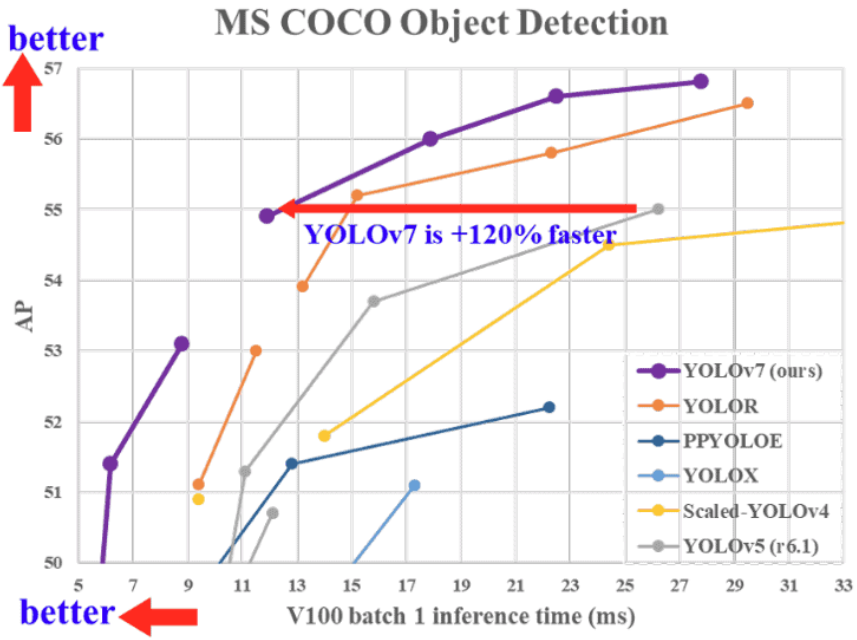
Εικόνα 54: Αναπαράσταση αρχιτεκτονικής του αλγορίθμου YOLO v6. Πηγή (C. Li et al., 2022).

Παράλληλα, οι ερευνητές καινοτόμησαν στις συναρτήσεις απώλειας. Εφόσον η αναγνώριση αντικειμένων περιέχει ταξινόμηση και εντοπισμό, ο νέος αλγόριθμος συνδυάζει τεχνικές για την απώλεια ταξινόμησης αλλά και για την απώλεια παλινδρόμησης πλαισίου. Στην περίπτωση της ταξινόμησης, οι ερευνητές χρησιμοποιούν την απώλεια VFL (VariFocal Loss) (Zhang et al., 2021), η οποία αντιμετωπίζει διάφορα θετικά και αρνητικά δείγματα με βαθμολόγηση, βοηθώντας στην εξισορρόπηση των δειγμάτων εκμάθησης. Για την παλινδρόμηση πλαισίου, χρησιμοποιούν την απώλεια DFL (Distribution Focal Loss) (X. Li et al., 2020), η οποία αντιμετωπίζει την κατανομή των θέσεων του πλαισίου οριοθέτησης ως διακριτή κατανομή πιθανότητας.

5.3.11 YOLO v7

Η πιο πρόσφατη έκδοση του αλγορίθμου κυκλοφόρησε τον Ιούλιο του 2022, ένα μήνα μετά τον YOLO v6. Βέβαια, έχουν δημοσιευτεί μερικές ακόμα εκδόσεις, οι YOLO X (Ge et al., 2021) και YOLO R (Wang et al., 2021) που επικεντρώθηκαν στην καλύτερη εκμετάλλευση των GPU. Επί του παρόντος, ο YOLO v7 θεωρείται ο ισχυρότερος ανιχνευτής πραγματικού χρόνου,

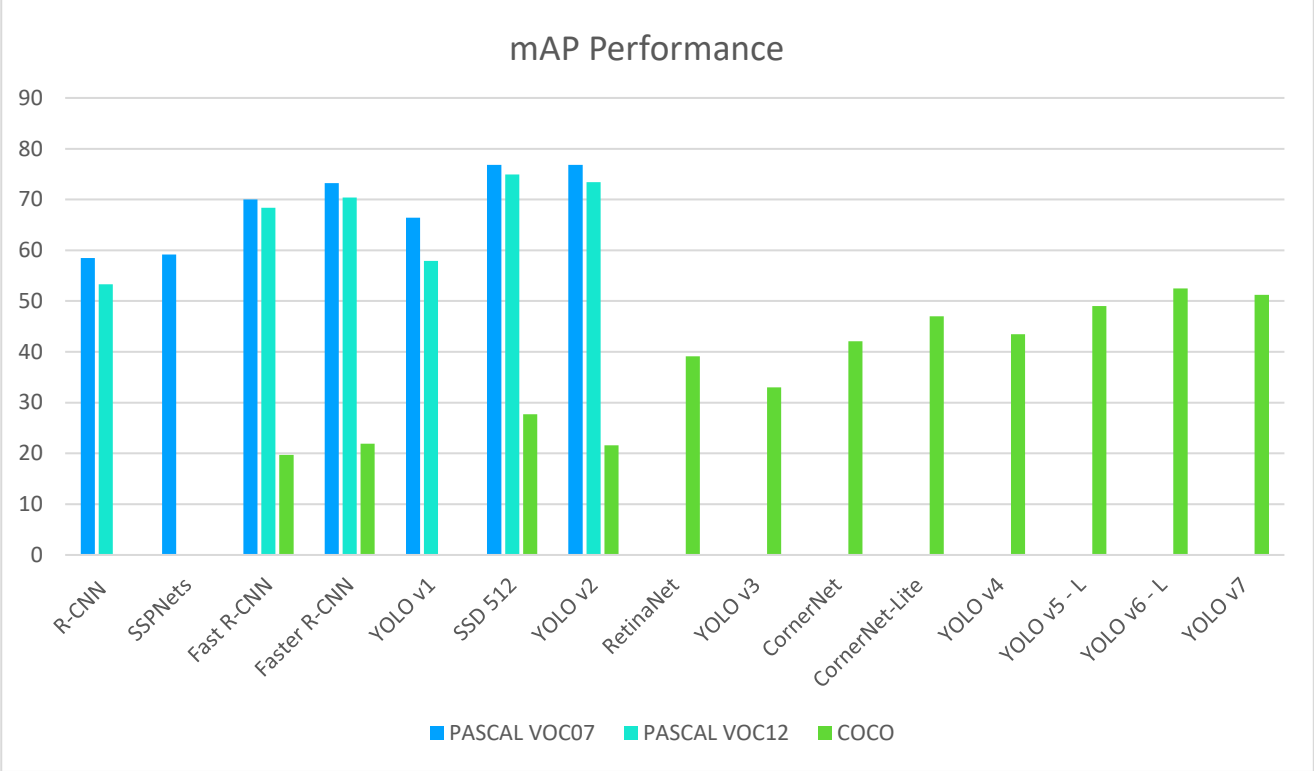
εξοπλισμένος με τις πιο προηγμένες τεχνολογίες εκπαίδευσης νευρωνικών δικτύων (Wang et al., 2023). Μια από αυτές είναι η εφαρμογή του E-ELAN (Extended Efficient Layer Aggregation Network) (Wu et al., 2021) στον κορμό του YOLO v7. Με την συγκεκριμένη τεχνολογία, οι ερευνητές κατάφεραν να ελέγξουν την ποσότητα μνήμης που χρειάζονται τα επίπεδα του δικτύου, μαζί με την απόσταση που χρειάζεται το σφάλμα για να διαδοθεί προς τα πίσω. Τέλος, το νέο μοντέλο έχει την ικανότητα να κλιμακώνει το βάθος και το πλάτος του μέσω παραμέτρων κλιμάκωσης για οποιαδήποτε απαίτηση, διατηρώντας πάντα την βέλτιστη αρχιτεκτονική του.



Εικόνα 55: Διάγραμμα απόδοσης του YOLO v7 σε σύγκριση με προηγούμενες εκδόσεις του YOLO. Πηγή (Wang et al., 2023).

5.4 Σύνοψη Κεφαλαίου

Η σύγκριση των αλγορίθμων αναγνώρισης αντικειμένων πραγματοποιείται στα σύνολα δεδομένων, υπολογίζοντας την συνολική μέση ακρίβεια (mAP) και την ταχύτητα (ms). Με το πέρασμα του χρόνου, η αξιολόγηση των ανιχνευτών στο MS COCO εδραιώθηκε, λόγω των ιδιαιτεροτήτων του. Παρόλα αυτά, οι πρώτοι ανιχνευτές που δημοσιεύτηκαν, αξιολογήθηκαν στα σύνολα δεδομένων PASCAL VOC07 και VOC12, όπου ο R-CNN απέδωσε 58,5% ενώ οι βελτιώσεις που ακολουθήσαν έφτασαν την απόδοση του στο 73,2%. Από την πλευρά των ανιχνευτών ενός σταδίου, ξεχώρισαν οι SSD και RetinaNet με ποσοστά 27,7% και 29,1% στο MS COCO αντίστοιχα, όμως η παρουσία του YOLO με την συνεχή εξέλιξή του, αποτελεί την κορυφαία επιλογή φτάνοντας το 52,5%. Στην εικόνα 56 παρουσιάζεται η διαχρονική εξέλιξη των επιδόσεων συνολικής μέσης ακρίβειας, που καταγράφηκαν από το 2014 μέχρι και σήμερα.



Εικόνα 56: Γράφημα απόδοσης mAP στα σύνολα δεδομένων PASCAL VOC και MS COCO.

Παρακάτω βρίσκεται ο συγκεντρωτικός πίνακας αλγορίθμων βαθιάς μάθησης, όπου αναγράφονται με λεπτομέρεια τα χαρακτηριστικά για όλες τις εκδόσεις των ανιχνευτών που αναλύθηκαν σε αυτό το κεφάλαιο. Για το PASCAL VOC χρησιμοποιείται $mAP^{IoU=.50}$ δηλαδή το ποσοστό μέσης ακρίβειας σε μια κλάση αντικειμένου υπολογίζεται για IoU > 50%. Ωστόσο, στο MS COCO χρησιμοποιείται η μέτρηση $mAP^{IoU=.50:0.95}$ που αντιστοιχεί στον μέσο όρο του mAP σε 10 διαφορετικές τιμές IoU για κάθε κλάση (δηλαδή: 0.50, 0.55, 0.60, …, 0.95). Επιπλέον συναντάται η έκφραση ‘07+12’ η οποία συμβολίζει ότι το σύνολο δεδομένων εκπαίδευσης περιέχει εικόνες από τα σετ εκπαίδευσης και επικύρωσης (trainval) του PASCAL VOC 2007 μαζί με τα σετ εκπαίδευσης και επικύρωσης (trainval) του PASCAL VOC 2012. Παρομοίως η έκφραση ‘07++12’ συμβολίζει ότι το σύνολο δεδομένων εκπαίδευσης περιέχει εικόνες από τα σετ εκπαίδευσης, επικύρωσης και δοκιμής (trainval+test) του PASCAL VOC 2007 μαζί με τα σετ εκπαίδευσης και επικύρωσης (trainval) του PASCAL VOC 2012.

Πίνακας 5: Αναλυτικός Πίνακας Αλγορίθμων Βαθιάς Μάθησης

Algorithm Name	Year	Backbone	mAP				Stage Category	Paper
			ILSVRC	VOC07	VOC12	COCO		
R-CNN	2014	AlexNet	-	58,5%	53,3%	-	Two stage	(Girshick et al., 2014)
OverFeat	2014	AlexNet	24,3% (ILSVRC 2012)	-	-	-	One stage	(Sermanet et al., 2014)
SSPNets	2015	AlexNet	-	58,5%	-	-	-	(He et al., 2015)
	2015	ZFNet	-	59,2%	-	-	Two stage	
Fast R-CNN	2015	VGG-16	-	70,0% (07+12)	68,4% (07++12)	19,7%	Two stage	(Girshick, 2015)
Faster R-CNN	2015	VGG-16	-	73,2% (07+12)	70,4% (07++12)	21,9%	Two stage	(Ren et al., 2016)
YOLO v1	2016	GoogLeNet	-	66,4% (07+12)	57,9% (07++12)	-	One stage	(Redmon et al., 2016)
SSD 512	2016	VGG-16	-	76,8% (07+12)	74,9% (07++12)	27,7%	One stage	(W. Liu et al., 2016)
SSD 300	2016	VGG-16	-	74,3% (07+12)	72,4% (07++12)	23,4%	One stage	
YOLO v2	2017	Darknet-19	-	76,8% (07+12)	73,4% (07++12)	21,6%	One stage	(Redmon & Farhadi, 2017)
RetinaNet	2017	ResNet-101-FPN	-	-	-	39,1%	One stage	(Lin, Goyal, et al., 2017)
YOLO v3	2018	Darknet-53	-	-	-	33,0%	One stage	(Redmon & Farhadi, 2018)
CornerNet	2018	Hourglass-104	-	-	-	42,1%	One stage	(Law & Deng, 2018)
CornerNet-Lite	2019	Hourglass-104	-	-	-	47,0%	One stage	(Law et al., 2020)
YOLO v4	2020	CSPDarknet53	-	-	-	43,5%	One stage	(Bochkovskiy et al., 2020)
YOLO v5 - S	2020	CSPDarknet53	-	-	-	37,4%	One stage	(Jocher et al., 2020)
YOLO v5 - M	2020	CSPDarknet53	-	-	-	45,4%	One stage	(Jocher et al., 2020)
YOLO v5 - L	2020	CSPDarknet53	-	-	-	49,0%	One stage	(Jocher et al., 2020)
YOLO v6 - S	2022	EfficientRep	-	-	-	43,5%	One stage	(C. Li et al., 2022)
YOLO v6 - M	2022	EfficientRep	-	-	-	49,5%	One stage	
YOLO v6 - L	2022	EfficientRep	-	-	-	52,5%	One stage	
YOLO v7	2022	E-ELAN	-	-	-	51.2%	One stage	(Wang et al., 2023)

Κεφάλαιο 6

Επίλογος

6.1 Προκλήσεις

Η υπολογιστική όραση και ειδικότερα η αναγνώριση αντικειμένων ήδη συναντώνται σε αρκετές τεχνολογίες που αναπτυχθήκαν τα τελευταία χρόνια. Η εξέλιξη τέτοιων τεχνολογιών πρέπει να συμβαδίζει με την ευημερία των ανθρώπων προσφέροντας ανωτέρα προϊόντα και υπηρεσίες. Ωστόσο, πρέπει να αντιμετωπιστούν ορισμένες προκλήσεις, προκειμένου να αξιοποιηθούν πλήρως οι δυνατότητες των συγκεκριμένων τεχνολογιών.

Περιορισμένα Σύνολα Δεδομένων: Το κυριότερο πρόβλημα που συναντάται στην αναγνώριση αντικειμένων είναι ο περιορισμένος αριθμός εικόνων, παρά τις πολλές προσπάθειες που έγιναν από τους ερευνητές στην δημιουργία μεγάλων συνόλων εκπαίδευσης. Αυτά τα σύνολα συνήθως περιέχουν εικόνες που απεικονίζονται με σταθερές διαστάσεις και σε ιδανικές συνθήκες. Την ίδια στιγμή, η διαδικασία σχολιασμού κάθε εικόνας είναι υποχρεωτική για κάθε αντικείμενο που βρίσκεται σε μια εικόνα, ενώ ότι η εκπαίδευση των ανιχνευτών απαιτεί μια προκαθορισμένη μορφή των δεδομένων σχολιασμού, για παράδειγμα ο YOLO χρησιμοποιεί απλά αρχεία κειμένου txt, ενώ ο SSD χρησιμοποιεί αρχεία σε μορφή XML. Επομένως η κατασκευή νέων συνόλων εκπαίδευσης είναι εξαιρετικά χρονοβόρα και κουραστική, εφόσον πραγματοποιείται αποκλειστικά για έναν ανιχνευτή.

Αδυναμία ανίχνευσης πολλαπλών, αλληλοκαλυπτόμενων και μικρών αντικειμένων: Οι περισσότεροι αλγόριθμοι ανιχνεύουν αποτελεσματικά αντικείμενα κανονικών διαστάσεων για τα οποία εκπαιδευτήκαν επιτυχώς. Παρόλα αυτά, αρκετοί από αυτούς που αναλύθηκαν σε αυτήν την εργασία, ανιχνεύουν είτε λανθασμένα είτε παραβλέπουν αντικείμενα μικρών διαστάσεων. Μάλιστα, σε αρκετές περιπτώσεις έχει παρατηρηθεί ότι ο εντοπισμός ενός αντικειμένου που είναι εν μέρη ορατό ή καλύπτεται από ένα διαφορετικό αντικείμενο, είναι αρκετά δύσκολος.

6.2 Μελλοντικές επεκτάσεις

Το μέλλον της αναγνώρισης αντικειμένων είναι λαμπρό και πολλά υποσχόμενο, καθώς η τεχνολογία εξελίσσεται και εισέρχεται σε περισσότερες πτυχές της καθημερινής ζωής. Ωστόσο, η επιστημονική κοινότητα οφείλει να λάβει υπόψη τα ηθικά και νομικά ζητήματα που προκύπτουν από την χρήση μια τέτοιας τεχνολογίας, όπως η ασφάλεια των προσωπικών δεδομένων και το απόρρητο. Το πρώτο βήμα για την μελλοντική ανάπτυξη της συγκεκριμένης τεχνολογίας, είναι η πετυχημένη αντιμετώπιση των πολυμερών προκλήσεων που αναλυθήκαν προηγουμένως. Η μεγαλύτερη προτεραιότητα, είναι η δημιουργία ενός καθολικού συνόλου δεδομένων ή ακόμα και η βελτίωση των ήδη υπαρχόντων. Η προσπάθεια αυτή απαιτεί την συλλογή και τον σχολιασμό εκατομμυρίων εικόνων, με μεγαλύτερη ποικιλία αντικειμένων, με διαφορετικές διαστάσεις και ανάλυση, αλλά κυρίως με μια καθολική μορφή σχολιασμού των αντικειμένων. Επιπλέον, θα ήταν ωφέλιμη η ανάπτυξη πιο ισχυρών αλγορίθμων με την ικανότητα να διαχειρίζονται ταυτόχρονα αντικείμενα διαφορετικών διαστάσεων, διαφορετικές συνθήκες φωτισμού, διαφορετικές οπτικές γωνίες και μάλιστα θα αποδίδουν καλυτέρα σε αναγνώριση με πραγματικό χρόνο. Μερικοί τρόποι για να επιτευχθεί αυτό είναι η αναβάθμιση των δικτυών εξαγωγής χαρακτηριστικών, η

χρήση νέων συναρτήσεων ενεργοποίησης και η επιλογή συνθέτων συναρτήσεων κόστους. Τέλος, η αναγνώριση αντικειμένων παραμένει απρόσιτη και ακατανόητη για το μεγαλύτερο μέρος των νέων ενδιαφερόμενων. Το γεγονός αυτό τονίζει την ανάγκη για δημιουργία είτε πιο φιλικών προς τον χρήστη εργαλείων και διεπαφών, είτε περισσότερων προεκπαιδευμένων μοντέλων με εύκολη παραμετροποίηση και χρήση.

6.3 Συμπεράσματα

Η παρούσα διπλωματική εργασία εμβάθυνε στην υπολογιστική όραση και κυρίως στην αναγνώριση αντικειμένων. Με την μελέτη της εργασίας ευκολά κάποιος μπορεί να καταλήξει σε μερικά ενδιαφέροντα συμπεράσματα σχετικά με την μέχρι τώρα πορεία και κατεύθυνση της συγκεκριμένης τεχνολογίας. Οι δυνατότητες της βαθιάς μάθησης καθώς και τα διάσημα σύνολα δεδομένων κυριαρχούν δίχως αμφισβήτηση. Στην εργασία αυτή, πρώτα αναλύσαμε βασικές ορολογίες και περιγράψαμε θεμελιώδεις θεωρητικές έννοιες, απαραίτητες για την κατανόηση του αντικειμένου, ώστε να γνωρίσουμε τον κόσμο των Συνελικτικών Νευρωνικών Δικτυών. Αμέσως μετά ακολούθησε λεπτομερής περιγραφή των σημαντικότερων μοντέλων βαθιάς μάθησης, ξεκινώντας με το πρώτο δίκτυο LeNet και καταλήγοντας στα πιο σύνθετα και αποδοτικά δίκτυα (AlexNet, GoogLeNet, VGG, ResNet, SENet) όπου το κύριο μέτρο που χρησιμοποιήθηκε για την σύγκρισή τους είναι το σφάλμα ταξινόμησης Top-5. Μάλιστα, έγινε ιδιαίτερη αναφορά στην διαχρονική εξέλιξη του διαγωνισμού ILSVRC 2011-2017 σύμφωνα με τις επιδόσεις των δικτύων. Από την πλευρά των αλγορίθμων, χωρίσαμε τους ανιχνευτές σε δυο κατηγορίες, σύμφωνα με τα βήματα που ακολουθούν για ολόκληρη την διαδικασία ανίχνευσης. Πρώτα, εξετάστηκαν οι R-CNN και οι επόμενες εκδόσεις του, που αντιπροσωπεύουν τους αλγορίθμους δυο σταδίων, οι όποιοι ενώ εμφανίζουν μεγάλα ποσοστά ακρίβειας, χρειάζονται αρκετό χρόνο επεξεργασίας. Έπειτα, αναλύθηκαν οι ανιχνευτές ενός σταδίου που σε αντίθεση με τους προηγούμενους είναι αρκετά γρηγορότεροι και αποτελεσματικοί για αναγνώριση σε πραγματικό χρόνο. Βέβαια οι πρώτοι αλγόριθμοι που παρουσιάστηκαν (OverFeat, SSD), δεν κατάφεραν να ξεπεράσουν τους τότε καλυτέρους Fast R-CNN και Faster R-CNN, αλλά οι μετέπειτα RetinaNet, CornerNet αποδείχτηκαν καλύτεροι και πρωτίστως ο YOLO κατέκτησε την πρώτη θέση εξαιτίας της συνεχής προσπάθειας για την βελτίωσή του. Η σύγκριση των αλγορίθμων έγινε με βάση το ποσοστό συνολικής μέσης ακρίβειας mAP που έχει υπολογιστεί για κάθε αλγόριθμο πάνω στα σύνολα PASCAL VOC και MS COCO. Αξίζει να σημειωθεί ότι είναι εξαιρετικά δύσκολο να απαντηθεί ποιος από όλους είναι ο καλύτερος, καθώς οι πραγματικές εφαρμογές είναι περίπλοκες και ιδιαιτέρα απαιτητικές. Τέλος, πρέπει να τονιστεί ότι είναι δύσκολο να πραγματοποιηθεί μια ίση και δίκαιη σύγκριση μεταξύ όλων των αλγορίθμων ειδικότερα στην περίπτωση που θέλουμε να μετρήσουμε με ακρίβεια την ταχύτητα κάθε ανιχνευτή, διότι η διαδικασία αυτή προϋποθέτει ισοδύναμες προδιαγραφές υλικού (hardware).

Βιβλιογραφία

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, *9*(8), Article 8. https://doi.org/10.3390/electronics9081295

Ajit, A., Acharya, K., & Samanta, A. (2020). A Review of Convolutional Neural Networks. *2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE)*, 1–5. https://doi.org/10.1109/ic-ETITE47903.2020.049

Andreopoulos, A., & Tsotsos, J. K. (2013). 50 Years of object recognition: Directions forward. *Computer Vision and Image Understanding*, *117*(8), 827–891. https://doi.org/10.1016/j.cviu.2013.04.005

Basegmez, E. (2014). *NEUROSCIENTIFIC SYSTEM THEORY Technische Universitat Munchen*.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded Up Robust Features. In A. Leonardis, H. Bischof, & A. Pinz (Eds.), *Computer Vision – ECCV 2006* (pp. 404–417). Springer. https://doi.org/10.1007/11744023_32

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). *YOLOv4: Optimal Speed and Accuracy of Object Detection* (arXiv:2004.10934). arXiv. https://doi.org/10.48550/arXiv.2004.10934

Bunke, H., & Allermann, G. (1983). Inexact graph matching for structural pattern recognition. *Pattern Recognition Letters*, *1*(4), 245–253. https://doi.org/10.1016/0167-8655(83)90033-8

Chollet, F. (2017). *Xception: Deep Learning With Depthwise Separable Convolutions*. 1251–1258. https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html

da Silva, I. N., Hernane Spatti, D., Andrade Flauzino, R., Liboni, L. H. B., & dos Reis Alves, S. F. (2017). Artificial Neural Network Architectures and Training Processes. In I. N. da Silva, D. Hernane Spatti, R. Andrade Flauzino, L. H. B. Liboni, & S. F. dos Reis Alves (Eds.), *Artificial Neural Networks: A Practical Course* (pp. 21–28). Springer International Publishing. https://doi.org/10.1007/978-3-319-43162-8_2

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). *RepVGG: Making VGG-Style ConvNets Great Again*. 13733–13742. https://openaccess.thecvf.com/content/CVPR2021/html/Ding_RepVGG_Making_VGG-Style_ConvNets_Great_Again_CVPR_2021_paper.html

Dollár, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(8), 1532–1545. https://doi.org/10.1109/TPAMI.2014.2300479

Er, M. J., Wu, S., Lu, J., & Toh, H. L. (2002). Face recognition with radial basis function (RBF) neural networks. *IEEE Transactions on Neural Networks*, *13*(3), 697–710. https://doi.org/10.1109/TNN.2002.1000134

Ertmer, P. A., & Newby, T. J. (1993). Behaviorism, Cognitivism, Constructivism: Comparing Critical Features from an Instructional Design Perspective.

Performance Improvement Quarterly, *6*(4), 50–72. https://doi.org/10.1111/j.1937-8327.1993.tb00605.x

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2015). The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, *111*(1), 98–136. https://doi.org/10.1007/s11263-014-0733-5

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, *88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Garbin, C., Zhu, X., & Marques, O. (2020). Dropout vs. batch normalization: An empirical study of their impact to deep learning. *Multimedia Tools and Applications*, *79*(19), 12777–12815. https://doi.org/10.1007/s11042-019-08453-9

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). *YOLOX: Exceeding YOLO Series in 2021* (arXiv:2107.08430). arXiv. https://doi.org/10.48550/arXiv.2107.08430

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Girshick, R. (2015). *Fast R-CNN*. 1440–1448. https://openaccess.thecvf.com/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation*. 580–587.

https://openaccess.thecvf.com/content_cvpr_2014/html/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.html

Gluck, M. A., & Myers, C. E. (2001). *Gateway to Memory: An Introduction to Neural Network Modeling of the Hippocampus and Learning*. MIT Press.

Goh, K. W., Mamat, M., Mohd, I., & Yosza, D. (2012). A novel of step size selection procedures for steepest descent method. *Applied Mathematical Sciences (Ruse)*, *6*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications* (arXiv:1704.04861). arXiv. https://doi.org/10.48550/arXiv.1704.04861

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). *Searching for MobileNetV3*. 1314–1324.

https://openaccess.thecvf.com/content_ICCV_2019/html/Howard_Searching_for_MobileNetV3_ICCV_2019_paper.html

Hu, J., Shen, L., & Sun, G. (2018). *Squeeze-and-Excitation Networks*. 7132–7141. https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html

Hua, B.-S., Tran, M.-K., & Yeung, S.-K. (2018). *Pointwise Convolutional Neural Networks*. 984–993. https://openaccess.thecvf.com/content_cvpr_2018/html/Hua_Pointwise_Convolutional_Neural_CVPR_2018_paper.html

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591.

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning*, 448–456. https://proceedings.mlr.press/v37/ioffe15.html

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, Hogan, A., Lorenzomammana, Tkianai, YxNONG, AlexWang1900, Diaconu, L., Marc, Wanghaoyang0106, Ml5ah, Doug, Hatovix, Poznanski, J., …

Yzchen. (2020). ultralytics/yolov5: V3.0. *Zenodo*. https://doi.org/10.5281/zenodo.3983579

Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, Tkianai, YxNONG, Hogan, A., Lorenzomammana, AlexWang1900, Chaurasia, A., Diaconu, L., Marc, Wanghaoyang0106, Ml5ah, Doug, Durgesh, … 于力军 L. Y. (2021). ultralytics/yolov5: V4.0 - nn.SiLU() activations, Weights & Biases logging, PyTorch Hub integration. *Zenodo*. https://doi.org/10.5281/zenodo.4418161

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., & Ferrari, V. (2020). The Open Images Dataset V4. *International Journal of Computer Vision*, *128*(7), 1956–1981. https://doi.org/10.1007/s11263-020-01316-z

Law, H., & Deng, J. (2018). *CornerNet: Detecting Objects as Paired Keypoints*. 734–750. https://openaccess.thecvf.com/content_ECCV_2018/html/Hei_Law_CornerNet_Detecting_Objects_ECCV_2018_paper.html

Law, H., Teng, Y., Russakovsky, O., & Deng, J. (2020). *CornerNet-Lite: Efficient Keypoint Based Object Detection* (arXiv:1904.08900). arXiv. https://doi.org/10.48550/arXiv.1904.08900

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), Article 7553. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U. A., Sackinger, E., Simard, P., Vapnik, V., & Laboratories, T. B. (1995). *LEARNING ALGORITHMS FOR CLASSIFICATION: A COMPARISON ON HANDWRITTEN DIGIT RECOGNITION*.

Leung, H., & Haykin, S. (1991). The complex backpropagation algorithm. *IEEE Transactions on Signal Processing*, *39*(9), 2101–2104. https://doi.org/10.1109/78.134446

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., Li, Y., Zhang, B., Liang, Y., Zhou, L., Xu, X., Chu, X., Wei, X., & Wei, X. (2022). *YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications* (arXiv:2209.02976). arXiv. https://doi.org/10.48550/arXiv.2209.02976

Li, G., Jian, X., Wen, Z., & AlSultan, J. (2022). Algorithm of overfitting avoidance in CNN based on maximum pooled and weight decay. *Applied Mathematics and Nonlinear Sciences*, *7*(2), 965–974. https://doi.org/10.2478/amns.2022.1.00011

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., & Yang, J. (2020). Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. *Advances in Neural Information Processing Systems*, *33*, 21002–21012. https://proceedings.neurips.cc/paper/2020/hash/f0bda020d2470f2e74990a07a607ebd9-Abstract.html

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

Liao, Z., & Carneiro, G. (2016). On the importance of normalisation layers in deep learning with piecewise linear activation units. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–8. https://doi.org/10.1109/WACV.2016.7477624

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). *Feature Pyramid Networks for Object Detection*. 2117–2125. https://openaccess.thecvf.com/content_cvpr_2017/html/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.html

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). *Focal Loss for Dense Object Detection*. 2980–2988. https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html

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 D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision*, *128*(2), 261–318. https://doi.org/10.1007/s11263-019-01247-4

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). *Path Aggregation Network for Instance Segmentation*. 8759–8768. https://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Path_Aggregation_Network_CVPR_2018_paper.html

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 21–37). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2

McCorduck, P., & Cfe, C. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. CRC Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Misra, D. (2020). *Mish: A Self Regularized Non-Monotonic Activation Function* (arXiv:1908.08681; Version 3). arXiv. https://doi.org/10.48550/arXiv.1908.08681

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Newell, A., Yang, K., & Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 483–499). Springer International Publishing. https://doi.org/10.1007/978-3-319-46484-8_29

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

Nielsen, M. (2015). *Neural Networks and Deep Learning*.

Ojha, V. K., Abraham, A., & Snášel, V. (2017). Metaheuristic design of feedforward neural networks: A review of two decades of research. *Engineering Applications of Artificial Intelligence*, *60*, 97–116. https://doi.org/10.1016/j.engappai.2017.01.013

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., & Walsh, J. (2020). Deep Learning vs. Traditional Computer Vision. In K. Arai & S. Kapoor (Eds.), *Advances in Computer Vision* (pp.

128–144). Springer International Publishing. https://doi.org/10.1007/978-3-030-17795-9_10

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 1–6. https://doi.org/10.1109/ICTKE.2017.8259629

Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). A Survey on Performance Metrics for Object-Detection Algorithms. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130

Padilla, R., Passos, W. L., Dias, T. L. B., Netto, S. L., & da Silva, E. A. B. (2021). A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit. *Electronics*, *10*(3), Article 3. https://doi.org/10.3390/electronics10030279

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). *You Only Look Once: Unified, Real-Time Object Detection*. 779–788. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html

Redmon, J., & Farhadi, A. (2017). *YOLO9000: Better, Faster, Stronger*. 7263–7271. https://openaccess.thecvf.com/content_cvpr_2017/html/Redmon_YOLO9000_Better_Faster_CVPR_2017_paper.html

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An Incremental Improvement* (arXiv:1804.02767). arXiv. https://doi.org/10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, *28*. https://proceedings.neurips.cc/paper_files/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

Rojas, R. (2013). *Neural Networks: A Systematic Introduction*. Springer Science & Business Media.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408. https://doi.org/10.1037/h0042519

Rowley, H. A., Baluja, S., & Kanade, T. (1998). Rotation invariant neural network-based face detection. *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, 38–44. https://doi.org/10.1109/CVPR.1998.698585

Ruder, S. (2017). *An overview of gradient descent optimization algorithms* (arXiv:1609.04747). arXiv. https://doi.org/10.48550/arXiv.1609.04747

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Russell, J. S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach. 3rd Edition*. Pearson Education.

Sah, S. (2020). *Machine Learning: A Review of Learning Types* (2020070230). Preprints. https://doi.org/10.20944/preprints202007.0230.v1

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Sanchez, J., & Perronnin, F. (2011). High-dimensional signature compression for large-scale image classification. *CVPR 2011*, 1665–1672. https://doi.org/10.1109/CVPR.2011.5995504

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 4510–4520. https://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014). *OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks* (arXiv:1312.6229). arXiv. https://doi.org/10.48550/arXiv.1312.6229

Sharma, S., Sharma, S., & Athaiya, A. (2020). ACTIVATION FUNCTIONS IN NEURAL NETWORKS. *International Journal of Engineering Applied Sciences and Technology*, *04*(12), 310–316. https://doi.org/10.33564/IJEAST.2020.v04i12.054

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857

Sibi, P., Jones, S. A., & Siddarth, P. (2005). ANALYSIS OF DIFFERENT ACTIVATION FUNCTIONS USING BACK PROPAGATION NEURAL NETWORKS. . . *Vol.*, *47*.

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition* (arXiv:1409.1556). arXiv. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1), Article 1. https://doi.org/10.1609/aaai.v31i1.11231

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). *Going Deeper With Convolutions*. 1–9. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the Inception Architecture for Computer Vision*. 2818–2826. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html

Vapnik, V., & Chapelle, O. (2000). Bounds on Error Expectation for Support Vector Machines. *Neural Computation*, *12*(9), 2013–2036. https://doi.org/10.1162/089976600300015042

Vlahavas, Kefalas, Bassiliades, Kokkoras, & Sakellariou. (2020). *Artificial Intelligence, 4th edition (in Greek – Τεχνητή Νοημοσύνη, Δ΄Έκδοση) – Intelligent Systems Lab*. https://intelligence.csd.auth.gr/publication/books/artificial-intelligence-4th-edition-in-greek-%CF%84%CE%B5%CF%87%CE%BD%CE%B7%CF%84%CE%AE-%CE%BD%CE%BF%CE%B7%CE%BC%CE%BF%CF%83%CF%8D%CE%BD%CE%B7-%CE%B4-%CE%AD%CE%BA%CE%B4%CE%BF%CF%83%CE%B7/

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). *YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors*. 7464–7475. https://openaccess.thecvf.com/content/CVPR2023/html/Wang_YOLOv7_Trainable_Bag-of-Freebies_Sets_New_State-of-the-Art_for_Real-Time_Object_Detectors_CVPR_2023_paper.html

Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2021). *You Only Learn One Representation: Unified Network for Multiple Tasks* (arXiv:2105.04206). arXiv. https://doi.org/10.48550/arXiv.2105.04206

Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. *PhD Thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA*. https://cir.nii.ac.jp/crid/1571135649638605440

Wu, W., Zhao, Y., Xu, Y., Tan, X., He, D., Zou, Z., Ye, J., Li, Y., Yao, M., Dong, Z., & Shi, Y. (2021). DSANet: Dynamic Segment Aggregation Network for Video-Level Representation Learning. *Proceedings of the 29th ACM International Conference on Multimedia*, 1903–1911. https://doi.org/10.1145/3474085.3475344

Yakimovsky, Y. (1976). Boundary and Object Detection in Real World Images. *Journal of the ACM*, *23*(4), 599–618. https://doi.org/10.1145/321978.321981

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

Yao, Z., Cao, Y., Zheng, S., Huang, G., & Lin, S. (2021). *Cross-Iteration Batch Normalization*. 12331–12340. https://openaccess.thecvf.com/content/CVPR2021/html/Yao_Cross-Iteration_Batch_Normalization_CVPR_2021_paper.html

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, *16*(8), Article 8. https://doi.org/10.1038/nrn3962

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 818–833). Springer International Publishing. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, H., Wang, Y., Dayoub, F., & Sunderhauf, N. (2021). *VarifocalNet: An IoU-Aware Dense Object Detector*. 8514–8523. https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VarifocalNet_An_IoU-Aware_Dense_Object_Detector_CVPR_2021_paper.html

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

ΔΙΑΜΑΝΤΑΡΑΣ, Δ. Κ., & ΜΠΟΤΣΗΣ, Α. Δ. (2019). *ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ*. ΚΛΕΙΔΑΡΙΘΜΟΣ. https://www.politeianet.gr/books/9789604619955-diamantaras-konstantinos-kleidarithmos-michaniki-mathisi-310786

Συντομογραφίες - Ακρωνύμια

AI Artificial Intelligent

ML Machine Learning

DL Deep Learning

MLP Multilayer Perceptron

CPU Central Processing Unit

GPU Graphics Processing Unit

TPU Tensor Processing Unit

RELU Rectified Linear Unit

API Application Programming Interface

RBF Radial basis function

ΣΝΔ Συνελικτικό Νευρωνικό Δίκτυο

CNN Convolutional Neural Networks

SIFT Sorting Intolerant From Tolerant

SURF Speeded Up Robust Features

IoU Intersection over Union

ROI Region of Interest

mAP mean Average Precession

ILSVRC ImageNet Large Scale Visual Recognition Challenge

SVM Support Vector Machine

VGG Visual Geometry Group

RPN Region Proposal Network

SENet Squeeze & Excitation Net

SPP Spatial Pyramid Pooling

SSD Single Shot MultiBox Detector

YOLO You Only Look Once

PAN Path Aggregation Network

FPN Feature Pyramid Network

VFL VariFocal Loss

DLF Distribution Focal Loss

E-ELAN Extended Efficient Layer Aggregation Network

CUDA Compute Unified Device Architecture

ms milliseconds

Απόδοση Αγγλικών Όρων

Απόδοση	Ξενόγλωσσος όρος
Τεχνητή Νοημοσύνη	Artificial Intelligent
Μηχανική Μάθηση	Machine Learning
Βαθιά Μάθηση	Deep Learning
Νευρωνικά Δίκτυα	Neural Networks
Ανάστροφη Μετάδοση Λάθους	Back Propagation
Συναρτήσεις Ενεργοποίησης	Activation Functions
Στρώμα	Layer
Δίκτυα Εμπρός Τροφοδότησης	Feed Forward Networks
Αναδρομικά Δίκτυα	Recurrent Networks
Συνάρτηση Κόστους	Cost Function
Αλγόριθμος Κατάβασης Δυναμικού	Gradient Descent Rule
Συνελικτικά Νευρωνικά Δίκτυα	Convolutional Neural Networks
Συνέλιξη	Convolution
Υποδειγματοληψία	Sub Sampling
Πλήρως Συνδεδεμένα Επίπεδα	Fully Connected Layers
Υποδειγματοληψία Μέγιστης Τιμής	Max Pooling
Υποδειγματοληψία Μέσης Τιμής	Average Pooling
Αναγνώριση Αντικειμένων	Object Recognition
Ταξινόμηση	Classification
Εντοπισμός	Localization
Ανίχνευση	Detection
Μάθηση Με Επίβλεψη	Supervised Learning

Αναγνώριση αντικειμένων/προτύπων με χρήση βαθιάς μάθησης και εφαρμογές.

Απόδοση	Ξενόγλωσσος όρος
Μάθηση Χωρίς Επίβλεψη	Unsupervised Learning
Μάθηση Με Ενίσχυση	Reinforcement Learning
Μεταφορά Μάθησης	Transfer Learning
Εικονοστοιχείο	Pixel
Λόγος Επικάλυψης IoU	Intersection over Union
Πλαίσιο Οριοθέτησης	Bounding Box
Ακρίβεια	Precision
Ανάκλαση	Recall
Συνολική Μέση Ακρίβεια	mean Average Precision
Υπερπροσαρμογή	Overfitting
Υποπροσαρμογή	Underfitting
Σύνολο Δεδομένων	Dataset
Επιλεκτική Αναζήτηση	Selective Search
Μηχανές Υποστήριξης Διανυσμάτων	Support Vector Machine
Μη Μέγιστη Καταστολή	Non-maximum Suppression
Επαύξηση Δεδομένων	Data Augmentation
Κορμός	Blackbone
Λαιμός	Neck
Κεφαλή	Head