ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Τομέας Μαθηματικών

Panoptic Segmentation with Deep Neural Networks

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Νικόλα Ιωάννου

Επιβλέπων: Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Γεώργιος Σιόλας
Ε.ΔΙ.Π. Ε.Μ.Π.

ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΛΟΓΟΥ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ

Αθήνα, Ιούλιος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Panoptic Segmentation with Deep Neural Networks

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του
Νικόλα Ιωάννου

Επιβλέπων: Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Γεώργιος Σιόλας
Ε.ΔΙ.Π. Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 1η Ιουλίου, 2025.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
.....	.....	.....
Παναγιώτης Τσανάκας	Αντώνιος Συμβώνης	Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2025

.....

ΙΩΑΝΝΟΥ ΝΙΚΟΛΑΣ

Διπλωματούχος σχολής Εφαρμοσμένων

Μαθηματικών και Φυσικών Επιστημών Ε.Μ.Π.

Περίληψη

Στην

Abstract

in

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ.Γεώργιο Σιόλα για την επίβλεψη της παρούσας διπλωματικής εργασίας. Θα ήθελα να ευχαριστήσω επίσης την οικογένεια μου γιατί χωρίς αυτούς δεν θα μπορούσα να βρίσκομαι στην θέση την οποία βρίσκομαι τώρα.

Contents

List of Figures

List of Tables

1 Μαθηματικό υπόβαθρο

2 Εισαγωγή

Η όραση υπολογιστών αποτελεί προσομοίωση της βιολογικής όρασης, κάνοντας χρήση υπολογιστών και συναφούς εξοπλισμού.Αποσκοπεί στην κατανόηση της τρισδιάστατης δομής του περιβάλλοντος, μέσω της επεξεργασίας εικόνων και βίντεο και έχει ως απότερο σκοπό την κατανόηση του οπτικού περιεχομένου. Το πεδίο αυτό περιλαμβάνει μεταξύ άλλων την

- Επεξεργασία εικόνας (Image Processing)
- Αναγνώριση προτύπων (Pattern Recognition)
- Γεωμετρική μοντελοποίηση (Geometric modeling)
- Αναγνώριση αντικειμένων (Recognition Processes)

[1]. Η όραση υπολογιστών μπορεί να εφαρμοσθεί σε μια ευρύα γκάμα αντικειμένων όπως η ιατρική στον εντοπισμό κακοήθη όγκου [2], στην αυτόνομη οδήγηση για την κατανόηση του περιβάλλοντος γύρω του οχήματος [3] και στην ρομποτική [4].

Η ιστορική της πορεία ξεκινά στις αρχές της δεκαετίας του 1960 όταν η έρευνα επικεντρώθηκε σε βασικές τεχνικές επεξεργασίας εικόνας, όπως το φιλτράρισμα (Filtering), η οριοθέτηση και η ανίχνευση ακμών. Αρχικά ο στόχος ήταν η ανάλυση των τιμών των εικονοστοιχείων, όπως και η αναγνώριση απλών σχημάτων. Κατά την διάρκεια της δεκαετίας του 1980 αναπτύχθηκαν πιο σύνθετες τεχνικές που επέτρεπαν την αναγνώριση πιο σύνθετων σχημάτων και την εξαγωγή χαρακτηριστικών. Τη δεκαετία του 1990 η όραση υπολογιστών πέρασε στην φάση της μηχανικής μάθησης, όπου υιοθετήθηκαν στατιστικές μεθόδοι όπως για παράδειγμα μεθόδοι που έκαναν χρήση μηχανών διανυσμάτων υποστήριξης (Support Vector Machines) και τυχαίων δασών (Random Forest). Σημείο καμπής αποτέλεσε η δεκαετία του 2010 όπου με την τεράστια πρόοδο της υπολογιστικής ισχύς και την δημιουργία συνόλων δεδομένων μεγάλου μεγέθους με τις κατάλληλες επισημειώσεις, άνοιξε την πόρτα σε αλγορίθμους βασισμένους στην βαθιά μάθηση και συγκεκριμένα στα νευρωνικά δίκτυα και τα συνελικτικά νευρωνικά δίκτυα, οι οποίοι σημείωσαν τεράστια πρόοδο. Σημαντικό παράδειγμα σε αυτό αποτέλεσε το AlexNet [5], ένα βαθύ συνελικτικό νευρωνικό δίκτυο που κατασκεύασε ο Geoffrey E. Hinton με την ομάδα του, το οποίο πέτυχε 15.3% top-5 error rate στο ILSVRC (ImageNet Large Scale Visual Recognition Challenge) [6]. Σήμερα, η όραση υπολογιστών επικεντρώνεται κυρίως στην ανάπτυξη μοντέλων που μπορούν να εξάγουν αποτελέσματα σε πραγματικό χρόνο, χωρίς αισθητή χρονική καθυστέρηση, σε ηθικά ζητήματα όπως η αμεροληψία των αποτελεσμάτων όπως και σε ζητήματα ιδιωτικότητας [7].

Η όραση υπολογιστών περιλαμβάνει μια πληθώρα εργασιών, κάθε μια με διαφορετικό σκοπό και επίπεδο ανάλυσης της οπτικής πληροφορίας [8], [9]. Μεταξύ αυτών έχουμε την

- Κατάτμηση εικόνας (Image Segmentation)
- Ταξινόμηση εικόνας (Image Classification)
- Ανίχνευση αντικειμένων (Object Detection)
- Ανακατασκευή τρισδιάστασης εικόνας (3 Dimensional Image Reconstruction)

2.1 Κατάτμηση εικόνας

Η κατάτμηση εικόνας αποτελεί τεχνική στην όραση υπολογιστών που περιλαμβάνει την διαίρεση εικόνων ή βίντεο σε πλήθος αντικειμένων ή περιοχών. Κατά την διάρκεια των ετών έχουν αναπτυχθεί πολυάριθμοι αλγόριθμοι που προσπαθούν να αντιμετωπίσουν το συγκεκριμένο πρόβλημα με κάποιους απο τους πιο αρχικούς να βασίζονται σε μεθόδους κάνοντας χρήση κατωφλίου (Threshhold), ομαδοποίησης βάση ιστογράμματος, ομαδοποίησης μέσω του K-means όπως και σε πιο προχωρημένους όπως αλγορίθμους όπως οι ενεργές καμπύλες (active contours) , οι τομές γράφων, στα υπο συνθήκη και τυχαία πεδία Markov όπως και σε αλγορίθμους βασισμένους στην αραιότητα.Κατατάλλα, τα τελευταία χρόνια τα μοντέλα βαθιάς μάθησης(Deep learning models) έχουν οδηγήσει σε μια νεα γενιά μοντέλων κατάτμησης εικόνας με εντυπωσιακές αποδόσεις, συχνά επιτυγχάνοντας τις υψηλότερες αποδόσεις σε διάσημα σύνολα αναφοράς (Benchmarks) [10]. Χωρίζεται σε κατηγορίες όπως

- Σημασιολογική κατάτμηση εικόνας (Semantic Image Segmentation)
- Κατάτμηση αντικειμένων (Instance Segmentation)
- Πανοπτική κατάτμηση εικόνας (Panoptic Image Segmentation)

2.1.1 Σημασιολογική κατάτμηση εικόνας

Ζητούμενο της σημασιολογικής κατάτμησης εικόνας είναι ο προσδιορισμός της σημασιολογικής κατηγορίας κάθε εικονοστοιχείου μιας εικόνας. Κατά κανόνα, το πρόβλημα αποτελεί πρόβλημα επιπτεβόμενης μάθησης (Supervised learning), αξιοποιώντας ένα σύνολο εικόνων όπου κάθε εικόνα είναι επισυμασμένη σε επίπεδο εικονοστοιχείου, με σκοπό την εκπαίδευση ενός μοντέλου για την εκτέλεση του έργου. Οι σημασιολογικές ετικέτες χωρίζονται σε δύο κατηγορίες, στα ”αντικείμενα” (Things), όπως για παράδειγμα σκύλος, αυτοκίνητο, πεζός κ.ο.κ. και στα ”σκηνικά” στοιχεία (Stuff), όπως για παράδειγμα ουρανός, βλάστηση, δρόμος κ.ο.κ. Οι δύο αυτοί όροι χρησιμοποιούνται εκτενώς στην κατάτμησης εικόνας . Η πρώτη κατηγορία αναφέρετε σε μετρήσιμα αντικείμενα, ενώ η δεύτερη κατηγορία σχετίζεται με στοιχεία του σκηνικού [11]. Παρακάτω παρουσιάζεται μια αναπαράσταση της συγκεκριμένης τεχνικής.

Όπως μπορούμε να δούμε στην εικόνα κάθε εικονοστοιχείο της αντιστοιχείζεται σε κάποια σημασιολογική κατηγορία. Μπορούμε να παρατηρήσουμε πως τα εικονοστοιχεία που αντιστοιχείζονται στην ίδια σημασιολογική κατηγορία είναι τα εικονοστοιχεία τα οποία αντιστοιχούν στο ίδιο

Figure 1: Αυθεντική εικόνα

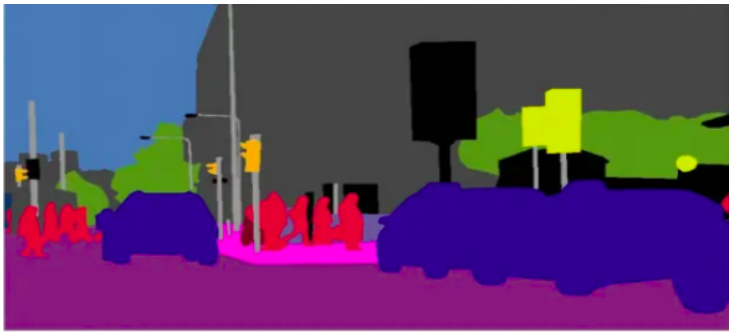

Figure 2: Σημασιολογική κατάτμηση εικόνας

”αντικείμενο” ή ”σκηνικό” στοιχείο.

Μερικές σημαντικές μετρικές απόδοσης της σημασιολογικής κατάτμησης εικόνας είναι η Intersection over Union (IoU), όπως επίσης και η Pixel Accuracy (PA) [10]. Μερικά παραδείγματα σημείων αναφοράς (benchmarks) που χρησιμοποιούνται για την ποσοτικοποίηση της απόδοσης συγκεκριμένων αλγορίθμων που χρησιμοποιούν την τεχνική αυτή είναι τα Cityscapes [12], PASCAL VOC [13] και ADE20K [14].

2.1.2 Κατάτμηση αντικειμένων

Η ανίχνευση αντικειμένων αποτελεί διαδικασία κατά την οποία ο αλγόριθμος εντοπίζει και ταξινομεί τα ”αντικείμενα” μιας εικόνας προσδιορίζοντας την θέση τους μέσω ορθογώνιων πλαισίων. Η σημασιολογική κατάτμηση εικόνας, όπως αναφέραμε παραπάνω προσδιορίζει την σημασιολογική κατηγορία κάθε εικονοστοιχείου μιας εικόνας, χωρίς όμως να διαχωρίζει μεταξύ διαφορετικών ”αντικειμένων” της ίδιας κατηγορίας. Προχωρώντας ένα βήμα παραπέρα, η κατάτμηση αντικειμένων συνδυάζει αυτές τις 2 τεχνικές και παρέχει διαφορετικές ετικέτες για ξεχωριστές εμφανίσεις ”αντικειμένων” που ανήκουν στην ίδια κατηγορία, αγνοώντας εντελώς τα ”σκηνικά” στοιχεία [15].

Όπως μπορούμε να δούμε στην εικόνα παραπάνω το μοντέλο κατηγοριοποιεί μόνο τα εικονοστοιχεία

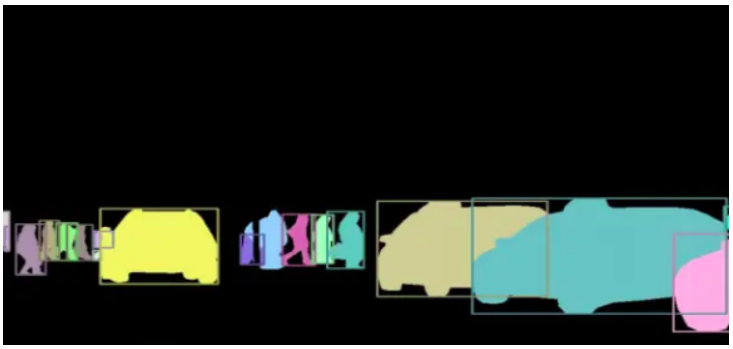

Figure 3: Κατάτμηση παραδειγμάτων

τα οποία αντιστοιχούν σε ”αντικείμενα”, αδιαφορώντας για τα υπόλοιπα. Εύκολα μπορούμε να παρατηρήσουμε επίσης πως γίνετε διάκριση μεταξύ των ”αντικειμένων” που ανήκουν στην ίδια κατηγορία.

Κάποιες σημαντικές μετρικές απόδοσης της κατάτμησης παραδείγματος είναι οι Average Precision (AP) και Mask Average Precision (Mask AP) [15]. Μερικά παραδείγματα συνόλων αναφοράς που χρησιμοποιούνται για την ποσοστικοποίηση της απόδοσης συγκεκριμένων αλγορίθμων που χρησιμοποιούν την τεχνική αυτή είναι τα COCO [16], Cityscapes [12] και ADE20K [14].

2.1.3 Πανοπτική κατάτμηση εικόνας

Η πανοπτική κατάτμηση εικόνας αποτελεί διαδικασία κατά την οποία γίνετε συνδυασμός της ανίχνευσης αντικειμένων και της σημασιολογικής κατάτμησης εικόνας. Συγκεκριμένα, στην πανοπτική κατάτμηση εικόνας πραγματοποιήτε κατηγοριοποίηση όλων των εικονοστοιχείων της εικόνας, ανεξάρτητα εάν τα εικονοστοιχεία αντιστοιχούν σε ”αντικείμενα” ή ”σκηνικά” στοιχεία και παράλληλα γίνετε διαχωρισμός μεταξύ των ”αντικειμένων” που αντιστοιχούν στην ίδια σημασιολογική κατηγορία [17].

Figure 4: Πανοπτική κατάτμηση εικόνας

Όπως μπορούμε να δούμε στην εικόνα παραπάνω το μοντέλο κατηγοριοποιεί όλα τα εικονοστοιχεία της εικόνας και ταυτόχρονα διαχωρίζει τα ”αντικείμενα”, τα οποία ανήκουν στην ίδια σημασιολογική κατηγορία.

Μερικές σημαντικές μετρικές της απόδοσης της πανοπτικής κατάτμησης εικόνας είναι η Panoptic Quality (PQ), η Segmentation Quality (SQ) και η Recognition Quality (RQ). Ωστόσο, γνωρίζουμε πως η πανοπτική κατάτμηση εικόνας αποτελεί συνδυασμό της σημασιολογικής κατάτμησης εικόνας και της κατάτμησης αντικειμένων υπάρχουν μετρικές που έχουν ώς σκοπό την ποσοτικοποίηση της απόδοσης των μοντέλων πανοπτικής κατάτμησης εικόνας στις 2 προηγούμενες εργασίες. Συγκεκριμένα για τα προαναφερθέντα υπάρχουν οι μετρικές απόδοσης Panoptic Quality Things (PQ_{th}) και Panoptic Quallity Stuff (PQ_{st}) [17]. Μερικά παραδείγματα σημείων αναφοράς που χρησιμοποιούνται για την ποσοτικοποίηση της απόδοσης συγκεκριμένων αλγορίθμων που κάνουν χρήση της τεχνικής αυτής είναι τα COCO [16], Cityscapes [12] και ADE20K [14].

2.2 Σύνολα δεδομένων για πανοπτική κατάτμηση εικόνας

Στην σύγχρονη εποχή η όραση υπολογιστών βασίζεται σχεδόν αποκλειστικά σε μεθόδους βαθιάς μάθησης. Γνωρίζουμε πως η λειτουργία τέτοιων μεθόδων απαιτεί ένα πολύ μεγάλο όγκο δεδομένων συνδυασμένο με τις κατάλληλες επισημειώσεις, ανάλογα πάντα με την εργασία που θέλουμε να πραγματοποιήσουμε. Για τον σκοπό αυτό έχουν δημιουργηθεί κατάλληλα σύνολα δεδομένων [7] που έχουν ώς σκοπό την εκπαίδευση των μοντέλων σε συγκεκριμένες εργασίες όπως και για την ποσοτική αξιολόγιση της απόδοσης τους στις εργασίες αυτές. Μερικά απο αυτά παρουσιάζονται στην συνέχεια.

2.2.1 Σύνολο δεδομένων COCO

Το COCO (Common Objects in Context), αποτελεί σύνολο δεδομένων μεγάλου μεγέθους. Περιέχει συνολικά περισσότερες απο 330.000 εικόνες και είναι διαθέσιμο σε 2 κύριες εκδόσεις, την COCO 2014 και COCO 2017, οι οποίες περιλαμβάνουν σε μεγάλο βαθμό κοινές εικόνες αλλά με διαφορετικό διαχωρισμό σε σύνολα εκπαίδευσης και επικύρωσης. Κάθε μια απο αυτές τις εικόνες είναι επισημειωμένη με 80 κατηγορίες αντικειμένων και 5 ετικέτες που περιγράφουν

την σκηνή. Το σύνολο δεδομένων χωρίζεται σε 2 κατηγορίες. Απο την μια έχουμε τις εικόνες, ενώ απο την άλλη τις αντίστοιχες επισημειώσεις. Οι εικόνες είναι οργανωμένες ιεραρχικά σε φακέλους, με τον φάκελο που βρίσκεται στο υψηλότερο επίπεδο να περιέχει φακέλους για το σύνολο δεδομένων εκπαίδευσης (Train set), το σύνολο δεδομένων επικύρωσης (Validation set) και το σύνολο δεδομένων δοκιμής (Test set) [16], [18].

Οι επισημειώσεις δίνονται σε JSON αρχεία, όπου κάθε αρχείο αντιστοιχεί σε μια εικόνα. Κάθε τέτοιο αρχείο περιέχει:

- Το όνομα του αρχείου
- Το μέγεθος της εικόνας
- 5 ετικέτες που περιγράφουν την σκηνή
- Λίστα με τα αντικείμενα που υπάρχουν μέσα στην εικόνα (Για κάθε αντικείμενο περιέχεται η κατηγορία του, οι συντεταγμένες του ορθοφωνίου που το περιβάλλει, τα εικονοστοιχεία που αντιστοιχούν σε αυτό το αντικείμενο και τα σημεία κλειδιά του)

Το σύνολο δεδομένων COCO περιέχει επίσης την άδεια χρήσης, επισημειώσεις για τα ”σκηνικά” στοιχεία, σε επίπεδο εικονοστοιχείου και υπερκατηγορίες αντικειμένων (Αποτελούν ευρύτερες κατηγορίες που περιλαμβάνουν πιο συγκεκριμένες υποκατηγορίες π.χ. dog $\subset$ animal). Το σύνολο δεδομένων COCO μπορεί να χρησιμοποιηθεί σε εργασίες όπως η ανίχνευση αντικειμένων, η σημασιολογική κατάτμηση εικόνας, η πανοπτική κατάτμηση εικόνας κ.ο.κ. [18].

Παρακάτω δίνονται όλες οι κατηγορίες αντικειμένων που περιέχονται στο σύνολο δεδομένων COCO.

person	fire hydrant	elephant	skis	wine glass	broccoli	dining table	toaster
bicycle	stop sign	bear	snowboard	cup	carrot	toilet	sink
car	parking meter	zebra	sports ball	fork	hot dog	tv	refrigerator
motorcycle	bench	giraffe	kite	knife	pizza	laptop	book
airplane	bird	backpack	baseball bat	spoon	donut	mouse	clock
bus	cat	umbrella	baseball glove	bowl	cake	remote	vase
train	dog	handbag	skateboard	banana	chair	keyboard	scissors
truck	horse	tie	surfboard	apple	couch	cell phone	teddy bear
boat	sheep	suitcase	tennis racket	sandwich	potted plant	microwave	hair drier
traffic light	cow	frisbee	bottle	orange	bed	oven	toothbrush

Figure 5: Κατηγορίες αντικειμένων συνόλου δεδομένων COCO

Είναι σημαντικό να αναφέρουμε πως το σύνολο δεδομένων COCO δεν περιέχει ισορροπημένο αριθμό αντικειμένων στις εικόνες κάτι που οδηγεί σε μεροληψία. Όπως αναφέραμε παραπάνω

το σύνολο δεδομένων COCO έρχεται σε 2 κύριες εκδόσεις. Η COCO 2014 περιέχει 82.783 εικόνες στο σύνολο δεδομένων εκπαίδευσης, 40.504 εικόνες στο σύνολο δεδομένων επικύρωσης και 40.775 εικόνες στο σύνολο δεδομένων δοκιμής. Στα αντίστοιχα σύνολα το COCO 2017 έχει 118.287, 5.000 και 40.670 εικόνες [19].

Για την εργασία της πανοπτικής κατάτμησης εικόνας το σύνολο δεδομένων COCO περιέχει επίσης κατάλληλες επισημειώσεις για την συγκεκριμένη εργασία, όπου οι επισημειώσεις αυτές περιέχουν πληροφορία για 91 διαφορετικές κατηγορίες "σκηνικών" στοιχείων [16], [18]. Η δομή του συνόλου δεδομένων δίνετε παρακάτω:

```
coco/
  annotations/
    instances_{train,val}2017.json
    panoptic_{train,val}2017.json
  {train,val}2017/
    # image files that are mentioned in the corresponding json
  panoptic_{train,val}2017/          # png annotations
  panoptic_semseg_{train,val}2017/
```

Figure 6: Δομή συνόλου δεδομένων COCO

2.2.2 Σύνολο δεδομένων Cityscapes

Η κατανόηση περίπλοκων αστικών σκηνών αποτελεί καθοριστικό παράγοντα για ένα ευρύ φάσμα εφαρμογών. Ωστόσο, δεν υπάρχουν πολλά σύνολα δεδομένων που να αποτυπόνουν επαρκώς την πολυπλοκότητα των σκηνών που παρουσιάζονται στον πραγματικό κόσμο. Λύση σε αυτό το πρόβλημα ήρθε να φέρει το σύνολο δεδομένων Cityscapes [12].

Το σύνολο δεδομένων Cityscapes δημιουργήθηκε μέσω επιλεγμένων καρέ (Frame), τα οποία εξήχθησαν απο στερεοσκοπικές ακολουθίες βίντεο που καταγράφηκαν απο κινούμενο όχημα στους δρόμους 50 διαφορετικών πόλεων, κυρίως της Γερμανίας. Απο τις εικόνες αυτές, οι 5.000 διαθέτουν υψηλής ποιότητας επισημειώσεις σε επίπεδο εικονοστοιχείου, ενώ οι υπόλοιπες 20.000 διαθέοτουν χαμηλότερης ποιότητας επισημειώσεις, έτσι ώστε να μπορούν να χρησιμοποιηθούν για την εκπαίδευση μοντέλων σε όχι πλήρως επισημειωμένα δεδομένα (Ασθενώς επιβλεπόμενη μάθηση). Απο τις 5.000 πλήρως επισημειωμένες εικόνες, οι 2.975 απο αυτές αποτελούν εικόνες του συνόλου δεδομένων εκπαίδευσης, οι 500 του συνόλου δεδομένων επικύρωσης και οι υπόλοιπες 1.525 του συνόλου δεδομένων δοκιμής. Οι πλήρως επισημειωμένες εικόνες εξήχθησαν χεροκίνητα απο 27 απο τις 50 πόλεις του συνόλου δεδομένων και συκγκεριμένα απο το 20ό καρέ αποσπασμάτων βίντεο διαρκείας 30 καρέ. Αντίθετα, οι υπόλοιπες ασθενώς επισημειωμένες εικόνες του συνόλου δεδομένων έχουν προέλευση απο τις υπόλοιπες 23 πόλεις και εξήχθησαν απο τα αντίστοιχα

βίντεο εξάγωντας μια εικόνα ανά 20 δευτερόλεπτα λήψης βίντεο ή 20 μέτρα οδήγησης, ανάλογα με το πιο απο τα 2 συναίβενε πρώτο. Όπως μπορούμε να δούμε παρακάτω, το σύνολο δεδομένων Cityscapes περιλαμβάνει ετικέτες για συνολικά 30 ”αντικείμενα”, απο τα οποία μόνο τα 19 χρησιμοποιούνται στην επικύρωση των αποτελεσμάτων [12].

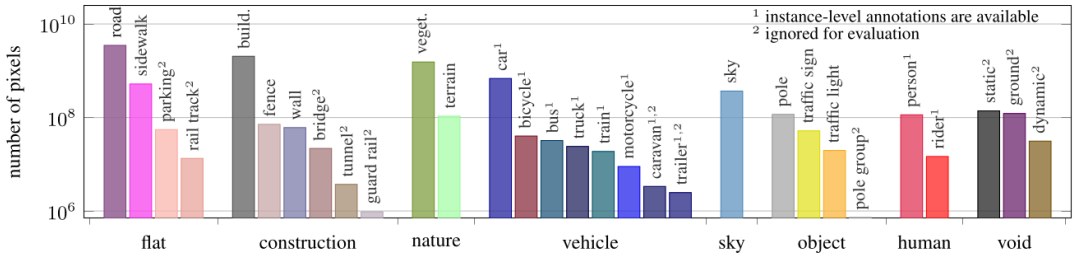

Figure 7: Αντικείμενα συνόλου δεδομένων Cityscapes

Το σύνολο δεδομένων περιλαμβάνει εικόνες καταγεγραμμένες κατά την διάρκεια της Άνοιξης, του Καλοκαιριού και του Φθινωπόρου και καλύπτει αρκετούς μήνες του έτους. Δεν περιλαμβάνει εικόνες με δυσμενείς καιρικές συνθήκες, όπως έντονη βροχόπτωση ή χιόνι επειδή οι συνθήκες αυτές απαιτούν την χρήση εξειδικευμένων τεχνικών και συνόλων δεδομένων. Όλες οι εικόνες είναι διαθέσιμες σε 8-bit (Low Dynamic Range) και 16-bit (High Dynamic Range), ώς πρός το βάθος χρώματος. Εκτός απο τις 8-bit, 16-bit εικόνες και τις αντίστοιχες επισημειώσεις, το σύνολο δεδομένων Cityscapes περιλαμβάνει πληροφορίες ώς πρός την εξωτερική θερμοκρασία, την διαδρομή που ακολούθησε το όχημα (GPS) και την οδομετρία του [12].

Το σύνολο δεδομένων Cityscapes μπορεί να χρησιμοποιηθεί για εργασίες όπως η σημασιολογική κατάτμηση εικόνας, η πανοπτική κατάτμηση εικόνας, η εκτίμηση βάθους κ.ο.κ. [12], [20], [21].

2.2.3 Σύνολο δεδομένων ADE20K

Σε αντίθεση με την απλή κατανόηση του περιεχομένου μιας εικόνας (Image-level recognition), η κατανόηση σε επίπεδο εικονοστοιχείου (Pixel level scene understanding) απαιτεί σύνολα δεδομένων με πολύ πιο πυκνές επισημειώσεις και ένα ευρύ σύνολο ”αντικειμένων”. Ωστόσο, τα περισσότερα σύνολα δεδομένων παρουσιάζουν ένα περιορισμένο αριθμό ”αντικειμένων” (π.χ. COCO [16], Pascal VOC [13]) και συχνά περιλαμβάνουν κατηγορίες που δεν είναι συνηφασμένες με τα πιο κοινά αντικείμενα που μπορούμε να συναντήσουμε στον πραγματικό κόσμο ή καλύπτουν μόνο ένα περιορισμένο φάσμα σκηνών (π.χ. Cityscapes [12]). Εξαιρεση σε αυτό αποτελεί το σύνολο δεδομένων Pascal-Context [22], όπως και η βάση δεδομένων Sun [23], με το πρώτο να επικεντρώνεται κυρίως σε μόνο 20 κατηγορίες ”αντικειμένων” ενώ το δεύτερο περιλαμβάνει επισημειώσεις ”αντικειμένων” με υψηλό επίπεδο θορύβου. Η ανάγκη δημιουργίας ενός συνόλου δεδομένων, το οποίο να ανταποκρίνετε στα παραπάνω προβλήματα οδήγησε στη δημιουργία

του συνόλου δεδομένων ADE20K [14], [24].

Το ADE20K αποτελεί ένα εκτενώς επισημειωμένο σύνολο δεδομένων, με την έννοια πως για κάθε εικόνα παρέχονται λεπτομερείς ετικέτες που καλύπτουν ”αντικείμενα”, όπως και μέρη ”αντικειμένων” (”υπο-αντικείμενα”). Αποτελείτε απο συνολικά 25.210 εικόνες ,όπου οι 20.210 απο αυτές αποτελούν εικόνες του συνόλου δεδομένων εκπαίδευσης, οι 2.000 απο αυτές αποτελούν εικόνες του συνόλου δεδομένων επικύρωσης και οι υπόλοιπες 3.000 του συνόλου δεδομένων δοκιμής. Στις εικόνες υπάρχουν συνολικά 3169 επισημειώσεις απο τις οποίες οι 2693 απο αυτές αποτελούν τα ”αντικείμενα” και ”σκηνικά” στοιχεία και οι υπόλοιπες 476 αποτελούν μέρη μεγαλύτερων ”αντικειμένων”¨. Στο σύνολο δεδομένων ADE20K υπάρχουν ”υπο-αντικείμενα” μέχρι επιπέδου το πολύ 3. Παρακάτω δίνετε μια αναπαράσταση κάποιων απο τα ”αντικείμενα” του συνόλου δεδομένων μαζί με τα αντίστοιχα ”υπο-αντικείμενα” τους [14], [24].

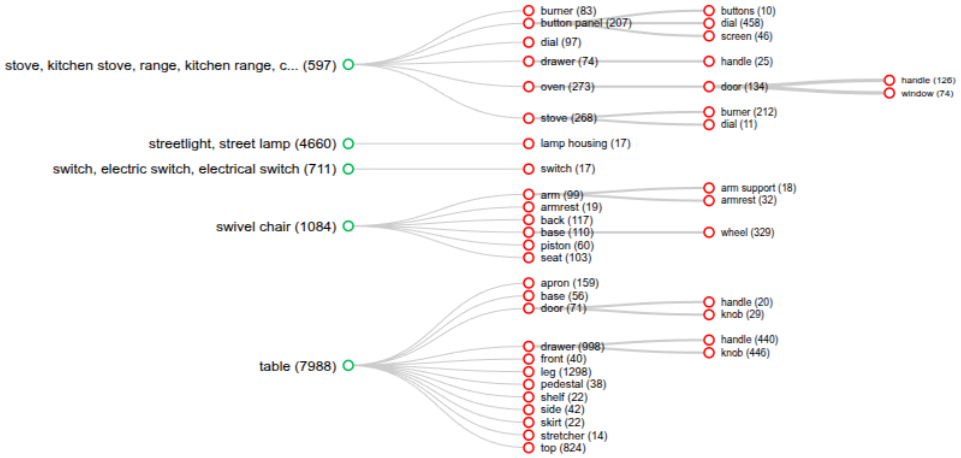

Figure 8: Αντικείμενα συνόλου δεδομένων ADE20K

Στο σύνολο δεδομένων ADE20K, το 76% των ”αντικειμένων” περιλαμβάνει ”υπο-αντικείμενα” με μέσο όρο ”υπο-αντικειμένων” που περιλαμβάνουν τα ”αντικείμενα” αυτά ίσο με 3. Κατά μέσο όρο υπάρχουν 19.5 εμφανίσεις ”αντικείμενων” και 10.5 διαφορετικές κατηγορίες ”αντικείμενων” σε κάθε εικόνα. Ο ελάχιστος αριθμός ”αντικειμένων” που υπάρχουν σε κάποια απο τις εικόνες του συνόλου δεδομένων είναι ίσος με 5 με κάποιες εικόνες να έχουν μέχρι και 273, χωρίς την συμπερίληψη των ”υπο-αντικείμενων” σε αυτά. Συμπεριλαμβανομένων των ”υπο-αντικειμένων” φτάνουμε μέχρι και 419 [14], [24].

Το σύνολο δεδομένων ADE20K μπορεί να χρησιμοποιηθεί για εργασίες όπως η σημασιολογική κατάτμηση εικόνας, η κατάτμηση αντικειμένων, η πανοπτική κατάτμηση εικόνας κ.ο.κ. [14], [20].

2.3 Μετρικές απόδοσης πανοπτικής κατάτμησης εικόνας

Οι μετρικές απόδοσης αποτελούν βασικό εργαλείο και διαδραματίζουν σημαντικό ρόλο για τη σύγκριση της αποτελεσματικότητας διαφορετικών μεθόδων σε διάφορες εργασίες. Η πανοπτική κατάτμηση εικόνας, όπως αναφέραμε και πρίν αποτελεί εργασία κατά την οποία συνδυάζεται η σημασιολογική κατάτμηση εικόνας και η κατάτμηση αντικειμένων. Αν και οι υπάρχουσες μετρικές απόδοσης της σημασιολογικής κατάτμησης εικόνας, όπως και της κατάτμησης αντικειμένων μπορούν σε κάποιο βαθμό να εφαρμοσθούν στην πανοπτική κατάτμηση εικόνας, δεν αρκούν απο μόνες τους. Συνήθως, για την συγκεκριμένη εργασία χρησιμοποιούνται μετρικές όπως η πανοπτική ποιότητα (Panoptic Quality - PQ), η ποιότητα κατάτμησης (Segmentation Quality - SQ) και η ποιότητα αναγνώρισης (Recognition Quality - RQ). Ωστόσο, μπορούν να χρησιμοποιηθούν και άλλες μετρικές για την σύγκριση της απόδοσης των μεθόδων αυτών όσον αφορά την σημασιολογική κατάτμηση και την κατάτμηση αντικειμένων, όπως ο περιορισμός των πιο πάνω μετρικών μόνο σε ”αντικείμενα” (Things) ή μόνο σε ”σκηνικά” στοιχεία (Stuff). Θα μπορούσαν επίσης να χρησιμοποιήσουμε μετρικές απόδοσης, όπως ο μέσος όρος ακρίβειας (Average Precision - AP) και η Intersection over Union (IoU) [17].

2.3.1 Πίνακας σύγχυσης (Confusion matrix)

Για να μπορέσουμε να ορίσουμε μερικές απο τις μετρικές απόδοσης που αναφέρθηκαν παραπάνω, χρειάζεται πρώτα να ορίσουμε τον πίνακα σύγχυσης. Ο πίνακας σύγχυσης αναπαριστά την ακρίβεια ενός μοντέλου ταξινόμησης. Παρουσιάζει τις τιμές των True positives (TP), των True negatives (TN), των False positives (FP) και των False negatives (FN). Ο πίνακας αυτός έχει μέγεθος NxN, όπου N είναι το πλήθος των κλάσεων ταξινόμησης και κάθε κελί του πίνακα αντιστοιχεί στο πλήθος των δειγμάτων που έχουν πραγματική τιμή ίδια με την πραγματική τιμή που αντιστοιχεί στο κελί και προβλεπόμενη τιμή ίδια με την προβλεπόμενη τιμή που αντιστοιχεί στο κελί. Στην ουσία συγκρίνεται η πραγματική ετικέτα με την ετικέτα η οποία προβλέφθηκε απο το μοντέλο. Για ένα δυαδικό πρόβλημα ταξινόμησης θα είχαμε τον παρακάτω πίνακα.

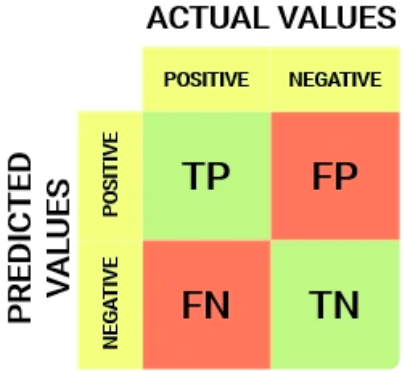

Figure 9: Μορφή πίνακα σύγχυσης για δυαδική ταξινόμηση

Στην περίπτωση αυτή το πρόβλημα είναι δυαδικό και ταξινομείτε σε 2 κλάσεις, την θετική (Positive) και την αρνητική (Negative). Όπως μπορούμε να δούμε και παραπάνω οι στήλες αναπαριστούν τις πραγματικές και οι γραμμές τις προβλεπόμενες τιμές. Παρακάτω δίνετε η ερμηνεία των TP, FP, FN και TN συκγκεκριμένα για δυαδικό πρόβλημα ταξινόμησης.

TP : Η προβλεπόμενη τιμή ανιστοιχεί με την πραγματική τιμή και η πραγματική τιμή ανήκει στην θετική κλάση (Positive).

TN : Η προβλεπόμενη τιμή ανιστοιχεί με την πραγματική τιμή και η πραγματική τιμή ανήκει στην αρνητική κλάση (Negative).

FP : Η προβλεπόμενη τιμή δεν ανιστοιχεί με την πραγματική τιμή. Η πραγματική τιμή ανήκει στην αρνητική κλάση αλλά το μοντέλο προέβλεψε πως ανήκει στην θετική. Το τιμή αυτή καλείτε και σφάλμα τύπου 1.

FN : Η προβλεπόμενη τιμή δεν ανιστοιχεί με την πραγματική τιμή. Η πραγματική τιμή ανήκει στην θετική κλάση αλλά το μοντέλο προέβλεψε πως ανήκει στην αρνητική. Η τιμή αυτή καλείτε και σφάλμα τύπου 2.

Στην περίπτωση που το πρόβλημα ταξινόμησης δεν είναι δυαδικό υπάρχουν διαφορές. Συγκεκριμένα οι τιμές των TP, FP, FN και TN, υπολογίζονται για κάθε κατηγορία. Επομένως θα έχουμε TP_i, FP_i, FN_i, TN_i, i = 1,...,N. Για κλάση i θα έχουμε:

TP$_i$: Ισούται με την τιμή του κελιού (i,i) στον πίνακα σύγχυσης.

TN$_i$: Ισούται με το άθροισμα $\sum_{p\neq i}^{N}\sum_{k\neq i}^{N} value(p,k)$, όπου value(i,j) η τιμή του πίνακα σύγχυσης στο κελί (i,j).

FP$_i$: Ισούται με το άθροισμα $\sum_{p\neq i}^{N} value(i,p)$, όπου value(i,j) η τιμή του πίνακα σύγχυσης στο κελί (i,j).

FN$_i$: Ισούται με το άθροισμα $\sum_{p\neq i}^{N} value(p,i)$, όπου value(i,j) η τιμή του πίνακα σύγχυσης στο κελί (i,j).

Αποφασίζουμε μέσω αυτής της μετρικής πως το μοντέλο ταξινόμησης είναι αποδοτικό εάν το πλήθος των True Positive και True Negative είναι μεγάλος σε σχέση με το συνολικό πλήθος των παρατηρήσεων [25].

2.3.2 Intersection over Union (Intersection over Union - IoU)

Αναφέρεται επίσης και ως δείκτης Jaccard. Πρόκειτε ουσιαστικά για ένα τρόπο ποσοτικοποίησης του ποσοστού επικάλυψης μεταξύ της μάσκας στόχου και της μάσκας πρόβλεψης. Συγκεκριμένα, η μετρική IoU μετρά το πλήθος των εικονοστοιχείων που είναι κοινά μεταξύ της μάσκας στόχου και της μάσκας πρόβλεψης, διαιρούμενο με τον συνολικό πλήθος των εικονοστοιχείων που υπάρχουν και στις δύο μάσκες μαζί [17].

$$IoU = \frac{target \cap prediction}{target \cup prediction} \tag{2.1}$$

Εύκολα μπορούμε να συμπεράνουμε πως $0 \leq IoU \leq 1$.

2.3.3 Μέσος Όρος Ακρίβειας (Average Precision - AP)

Ο μέσος όρος ακριβείας αποτελεί την πιο ευρέως χρησιμοποιημένη μετρική απόδοσης στην κατάτμηση αντικειμένων [26]. Για την διατύπωση της μετρικής αυτής χρειάζεται πρώτα να ορίσουμε κάποιες άλλες μετρικές. Αρχικά, η ακρίβεια (Precision) για μια συγκεκριμένη κλάση μετρά το ποσοστό των προβλέψεων που ανήκουν στην κλάση αυτή και συμφωνούν με την πραγματική τιμή [27]. Η ακρίβεια για μια κλάση k δίνεται απο την σχέση [28].

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \tag{2.2}$$

Στη συνέχεια θα ορίσουμε μια άλλη μετρική, την ανάκληση (Recall). Η ανάκληση για μια συγκεκριμένη κλάση μετρά το ποσοστό των ”αντικειμένων” που ανήκουν στην κλάση αυτή και έχουν προβλεφθεί σωστά [27]. Η ανάκληση για μια κλάση k δίνεται απο την σχέση [28].

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \tag{2.3}$$

Έπειτα ορίζουμε την καμπύλη ακριβείας-ανάκλησης (Precision-Recall curve). Η καμπύλη αυτή προκύπτει απεικονίζοντας τις τιμές ακριβείας και ανάκλησης του μοντέλου για μια συγκεκριμένη κλάση για όλες τις δυνατές τιμές του βαθμού εμπιστοσύνης. Ο μέσος όρος ακριβείας για κλάση k προκύπτει υπολογίζοντας το εμβαδόν κάτω απο την καμπύλη αυτή και δίνετε απο την σχέση.

$$AP_k = \int_0^1 Precision(Recall)\, d(Recall) \tag{2.4}$$

Ο μέσος όρος ακριβείας είναι άμεσα συνδεδεμένος με την τιμή του κατωφλίου της IoU που θα θέσουμε. Η τιμή του κατωφλίου επηρεάζει το πλήθος των προβλεπόμενων τμημάτων ίδιας κατηγορίας που αντιστοιχούν σε κάποιο πραγματικό τμήμα της εικόνας (Αντιστοίχιση τμημάτων) [29], επιρεάζει δηλαδή τη σύνθεση των συνόλων TP_k, FP_k και FN_k. Κατά συνέπεια, αυτό σημαίνει πως για διαφορετικές τιμές κατωφλίου λαμβάνουμε διαφορετική καμπύλη ακριβείας-ανάκλησης.

Για τον υπολογισμό της μετρικής mAP υπολογίζουμε την μέση τιμή των μέσων όρων ακριβείας (AP) για όλες τις κατηγορίες για κάποια τιμή του κατωφλίου της IoU. Αυτό δίνετε απο την σχέση.

$$mAP = \frac{1}{n} \sum_{k=1}^{n} AP_k \tag{2.5}$$

Για τον υπολογισμόυ του τελικού mAP υπολογίζουμε την μέση τιμή των mAP για διαφορετικές τιμές του κατωφλίου της IoU. Ο τρόπος υπολογισμού της τελικής mAP διαφέρει ανάλογα με τον εκάστοτε διαγωνισμό ανίχνευσης αντικειμένων και συγκεκριμένα ώς πρός τις τιμές του κατωφλίου της IoU και των τιμών του βαθμού εμπιστοσύνης που επιλέγονται. Ενδεικτικά στον διαγωνισμό ανίχνευσης αντικειμένων COCO 2017 [16] γίνετε χρήση συνολικά 10 διαφορετικών τιμών κατωφλίου, συγεκριμενα απο 0.5 εώς 0.95 με βήμα 0.05, καθώς και 101 τιμές βαθμών εμπιστοσύνης απο 0 μέχρι 1 με βημα 0.01 [30].

2.3.4 Πανοπτική ποιότητα (Panoptic Quality - PQ)

Η μετρική αυτή, σε αντίθεση με τις μετρικές που παρουσιάστηκαν προηγουμένως, αξιολογεί συνολικά την πανοπτική κατάτμηση εικόνας, λαμβάνοντας υπόψη τόσο τα ”αντικείμενα” όσο και τα ”σκηνικά” στοιχεία. Για την διατύπωση της μετρικής αυτής, απαιτείτε πρώτα ο ορισμός της έννοιας της αντιστοίχησης τμημάτων (Segment matching). Η αντιστοίχιση τμημάτων είναι η διαδικασία κατά την οποία αποφασίζεται ποιά απο τα προβλεπόμενα τμήματα που εξήγαγε το μοντέλο αντιστοιχούν σε ποιά πραγματικά τμήματα της εικόνας (Ground Truth). Για κάθε τμήμα τόσο για τα πραγματικά όσο και για τα προβλεπόμενα, κατασκευάζεται μια δυαδική μάσκα μεγέθους όσο και οι χωρικές διαστάσεις της εικόνας, όπου τα εικονοστοιχεία που αντιστοιχούν

στο τμήμα παίρνουν την τιμή 1 ενώ τα υπόλοιπα την τιμή 0. Ένα προβλεπόμενο και ένα πραγματικό τμήμα, ίδιας κατηγορίας αντιστοιχούν μεταξύ τους εάν η μετρική IoU είναι μεγαλύτερη απο 0.5 (Κατώφλι). Η συνθήκη αυτή σε συνδυασμό με την ιδιότητα της μη επικάλυψης των τμημάτων που ισχύει στην πανοπτική κατάτμηση εικόνας εξασφαλίζει μοναδική αντιστοίχηση, δηλαδή μπορούμε να αντιστοιχίσουμε το πολύ ένα προβλεπόμενο για κάθε πραγματικό τμήμα της εικόνας. Η απόδειξη βρίσκεται στο [29].

Για χαμηλότερες τιμές του κατωφλίου της IoU απαιτούνται διαφορετικές τεχνικές αντιστοίχησης. Ωστόσο, στην πράξη ταιριάσματα με $IoU \leq 0.5$ είναι σπάνια, επομένως χαμηλότερα κατώφλια είναι περιττά.

Για τον υπολογισμό της πανοπτικής ποιότητας, υπλογίζουμε πρώτα την πανοπτική ποιότητα για κάθε κατηγορία ξεχωριστά και στη συνέχεια υπολογίζουμε τον μέσο όρο. Αυτό καθιστά την πανοπτική ποιότητα ανεπηρέαστη απο ανισσοροπία μεταξύ των κατηγοριών. Για κάθε κατηγορία, η μοναδική αντιστοίχιση που προέκυψε χωρίζει τα πραγματικά και προβλεπόμενα τμήματα σε 3 σύνολα, το True Positives (TP), το False Positives (FP) και το False Negatives (FN), όπου το πρώτο σύνολο συμβολίζει τα ζευγάρια πραγματικών και προβλεπόμενων τμημάτων που αντιστοιχίστηκαν μεταξύ τους, το δεύτερο σύνολο συμβολίζει τα προβλεπόμενα τμήματα που δεν αντιστοιχίστηκαν με κανένα πραγματικό τμήμα και το τρίτο σύνολο τα πραγματικά τμήματα που δεν αντιστοιχήστηκαν με κανένα προβλεπόμενο τμήμα. Αυτό αποτυπώνεται και στο παρακάτω παράδειγμα, όπου φαίνετε πώς τα τμήματα της κατηγορίας ”person” διαχωρίζονται στα σύνολα True Positive, False Negative και False Positive [29].

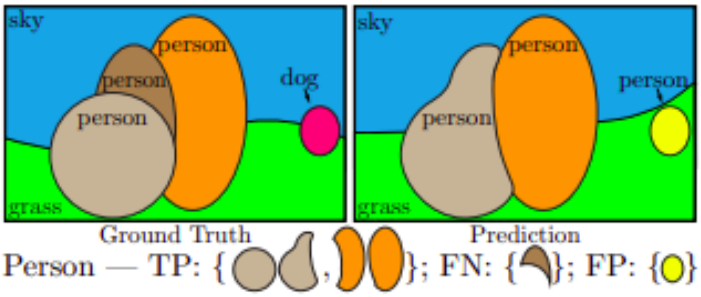

Figure 10: Παράδειγμα υπολογισμού συνόλων TP, FP και FN για την κατηγορία ”person”

Όπως μπορούμε να δούμε παραπάνω, το πραγματικό τμήμα κατηγορίας ”person” χρώματος καφέ δεν αντιστοιχήστηκε με κάποιο προβλεπόμενο τμήμα, επομένως προστίθεται στο σύνολο FN. Το προβλεπόμενο τμήμα κατηγορίας ”person” χρώματος κίτρινο δεν αντιστοιχήστηκε με κάποιο πραγματικό τμήμα, επομένως προστίθεται στο σύνολο FP. Τέλος, τα προβλεπόμενα τμήματα κατηγορία ”person” χρώματων γκρί και πορτοκαλί, αντιστοιχήστηκαν με 2 διαφορετικά πραγματικά τμήματα, επομένως τα ζευγάρια αυτά προστίθενται στο σύνολο TP.

Η πανοπτική ποιότητα μιας συγκεκριμένης κατηγορίας υπολογίζεται μέσω της παρακάτω σχέσης.

$$PQ = \frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{2.6}$$

Παρατηρώντας προσεκτικά, βλέπουμε πως η τιμή $\frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP|}$, είναι απλώς ο μέσος όρος της IoU για τα ταιριασμένα τμήματα, ενώ οι όροι $\frac{1}{2}|FP| + \frac{1}{2}|FN|$ προσθίτενται στον παρονομαστή για να επιβάλλουν ποινή στα τμήματα που δεν έχουν ταίριασμα. Παρατηρούμε πως όλα τα τμήματα μετράνε το ίδιο στον υπολογισμό της πανοπτικής ποιότητας ανεξάρτητα απο των αριθμό των εικονοστοιχείων που καταλαμβάνουν στην εικόνα. Επίσης, μπορούμε να παρατηρήσουμε πως πολλαπλασιάζοντας και διαιρώντας την μετρική με |TP|, η πανοπτική ποιότητα μπορεί να γραφεί ώς ο πολλαπλασιαμός των μετρικών της ποιότητας κατάτμησης (Segmentation Quality - SQ) και ποιότητας αναγνώρισης (Recognition Quality - RQ) [29].

$$PQ = \underbrace{\frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}} \tag{2.7}$$

Επομένως, ισχύει πως PQ = SQ x RQ.

Μπορούμε να παρατηρήσουμε πως η ποιότητα αναγνώρισης αντιστοιχεί στο μισό της τιμής του γνωστού F1-score, το οποίο χρησιμοποιείτε ως μετρική ποιότητας της ανίχνευσης αντικειμένων [17]. Η ποιότητα κατάτμησης απο την άλλη αποτελεί την μέση τιμή των IoU των ταιριασμένων τμημάτων μιας συγκεκριμένης κατηγορίας. Προκειμένου να υπολογίσουμε την συνολική πανοπτική ποιότητα υπολογίζουμε τον μέσο όρο της πανοπτικής ποιότητας κατά μήκος όλων των κατηγοριών [29].

$$PQ_{\text{overall}} = \frac{1}{C}\sum_{c=1}^{C} PQ_c \tag{2.8}$$

Περιορίζοντας τις μετρικές PQ, SQ και RQ μόνο στα ”αντικείμενα”, λαμβάνουμε τις μετρικές απόδοσης PQ_{th}, SQ_{th} και RQ_{th}. Αντίστοιχα περιορίζοντας στα ”σκηνικά” στοιχεία λαμβάνουμε τις μετρικές PQ_{st}, SQ_{st} και RQ_{st}. Οι μετρικές αυτές χρησιμοποιούνται για να αποτυπώσουν την ποιότητα στην σημασιολογική κατάτμηση εικόνας και την κατάτμηση αντικειμένων [17].

Στη συνέχεια θα αναλύσουμε πως αντιμετωπίζονται οι κενές ετικέτες (void labels) και οι ετικέτες ομάδας (group labels). Οι κενές ετικέτες δηλώνουν, σε επίπεδο εικονοστοιχείου, περιοχές της εικόνας που δεν αντιστοιχούν σε κάποιο ”αντικείμενο” ή ”σκηνικό” στοιχείο. Στις εικόνες αναφοράς (Ground Truth), αυτό οφείλετε είτε στο ότι τα εικονοστοιχεία αυτών των τμημάτων είναι πολύ δύσκολο να κατηγοριοποιηθούν με σιγουριά, είτε τα εικονοστοιχεία αυτά δεν ανήκουν σε καμιά απο τις κατηγορίες ”αντικειμένων” ή ”σκηνικών” στοιχείων που ορίζονται στο πρόβλημα. Εικονοστοιχεία τα οποία αντιστοιχούν σε κενή ετικέτα στις εικόνες αναφοράς, δεν λαμβάνονται υπόψη στην αξιολόγηση. Συγκεκριμένα, στην διαδικασία της αντιστοίχισης μεταξύ των προβλεπόμενων και πραγματικών τμημάτων, όλα τα εικονοστοιχεία στο προβλεπόμενο τμήμα τα οποία είναι επισημειωμένα ώς κενά στην εικόνα αναφοράς αφαιρούνται απο το προβλεπόμενο τμήμα και δεν λαβμάνονται υπόψη στον υπολογισμό της IoU. Επίσης, όσα προβλεπόμενα τμήματα δεν ταίριαξαν με κανένα πραγματικό τμήμα και περιλαμβάνουν ποσοστό εικονοστοιχείων τα οποία αντιστοιχούν σε κενές ετικέτες στις εικόνες αναφοράς μεγαλύτερο απο την τιμή του κατωφλίου της IoU, αφαιρούνται και δεν προστίθενται στο σύνολο False Positive (FP). Ακόμη, η τελική πρόβλεψη είναι δυνατό να περιέχει εικονοστοιχεία με κενές ετικέτες, δηλαδή το μοντέλο να μην προέβλεψε κάποια κατηγορία ”αντικειμένου” ή ”σκηνικού” στοιχείου για αυτά. Τα εικονοστοιχεία αυτά δεν λαμβάνονται υπόψη στην αξιολόγηση. Όσον αφορά τώρα τις ετικέτες ομάδας, χρησιμοποιούνται σε περιπτώσεις όπου υπάρχει πλήθος όμοιων ”αντικειμένων” συγκεντρωμένο και είναι δύσκολος ο διαχωρισμός τους σε ξεχωριστά ”αντικείμενα”. Στον υπολογισμό της πανοπτικής ποιότητας, κατά την διαδικασία της αντιστοίχισης δεν λαμβάνονται υπόψη τα τμήματα τα οποία έχουν ετικέτα ομάδας. Επιπρόσθετα, για τα προβλεπόμενα τμήματα τα οποία δεν ταίριαξαν με κανένα πραγματικό τμήμα και περιλαμβάνουν ποσοστό εικονοστοιχείων τα οποία αντιστοιχούν σε ετικέτα ομάδας στις εικόνες αναφοράς, μεγαλυτερο απο την τιμή του κατωφλίου της IoU, αφαιρούνται και δεν προστίθενται στο σύνολο False Positive [29].

3 Θεωρητικό υπόβαθρο

3.1 Ο νευρώνας

Η έρευνα σχετικά με τα τεχνητά νευρωνικά δίκτυα είναι εμπνευσμένη απο την δομή και λειτουργία του εγκεφάλου. Βασικό δομικό του στοιχείο είναι οι νευρώνες. Κίνητρο για την μελέτη του νευρώνα είναι η ανακάλυψη ενός μοντέλου το οποίο θα προσομοιώνει την λειτουργία και τις δυνατότητες του εγκεφάλου. Ωστόσο, τα τεχνητά νευρωνικά δίκτυα χρησιμοποιούν πολύ απλοποιημένα μοντέλα νευρώνων, τέτοια ώστε να διατηρούν μόνο τα πολύ αδρά χαρακτηριστικά των λεπτομερών μοντέλων που χρησιμοποιούνται στη νευρολογία. Οι λεπτομέρειες πιστεύεται πως δεν παίζουν ιδιαίτερη σημασία στην κατανόηση της ευφυούς συμπεριφοράς των βιολογικών νευρωνικών συστημάτων. Ακόμα και αυτά τα απλά μοντέλα νευρώνων μπορούν να δημιουργήσουν ιδιαιτέρως ενδιαφέροντα δίκτυα, αρκεί να πληρούν 2 βασικά χαρακτηριστικά.

- Οι νευρώνες πρέπει να έχουν ρυθμιζόμενες παραμέτρους ώστε να διευκολύνεται η διαδικασία της μάθησης (Πλαστικότητα των νευρώνων).
- Το δίκτυο πρέπει να αποτελείτε απο μεγάλο πλήθος νευρώνων ώστε να επιτυγχάνεται παραλληλισμός της επεξεργασίας και κατανομή της πληροφορίας.

Ο νευρώνας αποτελεί ένα μεγάλο σε μέγεθος κύτταρο το οποίο αποτελείτε από τον πυρήνα, τους δενδρίτες, τον άξονα και τις συνάψεις που συνδέουν τις διακλαδώσεις του άξονα με τους δενδρίτες άλλων νευρώνων [31].

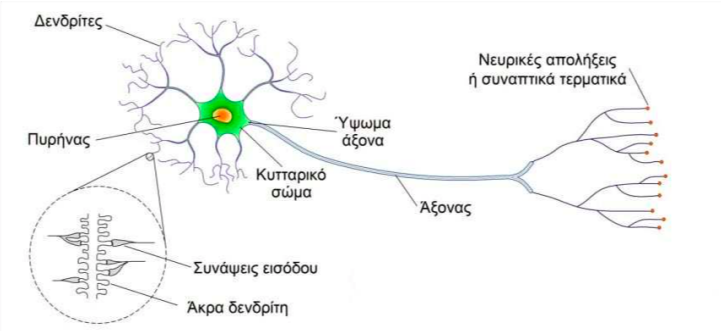

Figure 11: Δομή νευρώνα

Λειτουργικά, τα τμήματα του νευρώνα παίζουν διαφορετικούς ρόλους. Οι δενδρίτες είναι οι πύλες του νευρώνα και δέχονται ηλεκτρικά σήματα από άλλους νευρώνες. Ο άξονας είναι η πύλη εξόδου του νευρώνα και στέλνει σήματα πρός άλλους νευρώνες υπό μορφή ηλεκτρικών παλμών σταθερού πλάτους αλλά μεταβλητής συχνότητας. Τέλος, οι συνάψεις είναι τα σημεία ένωσης μεταξύ των διακλαδώσεων του άξονα ενός νευρώνα και των δενδριτών από άλλους νευρώνες. Το πλάτος της σύναψης, η απόσταση της απο τον δενδρίτη και η πυκνότητα του

ηλεκτροχημικού υλικού επηρεάζουν την ευκολία με την οποία η ηλεκτρική δραστηριότητα διαδίδεται απο τον άξονα στον δενδρίτη. Το ποσοστό της ηλεκτρικής δραστηριότητας που μεταδίδεται τελικά στον δενδρίτη ονομάζεται συναπτικό βάρος. Οι συνάψεις χωρίζονται σε ενισχυτικές (Exitatory) και ανασταλτικές (Inhibitory), ανάλογα με το αν το φορτίο που ελκύεται απο τη σύναψη διεγείρει τον νευρώνα για να παράγει παλμούς ή αντίθετα τον αναστέλλει εμποδίζοντας τον.

Στους βιολογικούς νευρώνες, οι φορείς των πληροφοριών είναι οι ηλεκτρικοί παλμοί, που ταξιδεύουν στον άξονα κάθε νευρώνα και μέσω των συνάψεων διαδίδονται στους δενδρίτες των νευρώνων. Κάθε νευρώνας συλλέγει όλο το ηλεκτρικό φορτίο που δέχεται απο κάθε σύναψη στους δενδρίτες του, σταθμόζοντας το εισερχόμενο φορτίο με το αντίστοιχο συναπτικό βάρος. Έτσι, όσο πιο ισχυρή είναι η συναπτική ζεύξη τόσο πιο πολύ συμμετέχει το συγκεκριμένο φορτίο εισόδου στο συνολικό άθροισμα. Άν το άθροισμα αυτό υπερβαίνει κάποιο κατώφλι (Threshhold), ο άξονας του νευρώνα αρχίζει να παράγει ηλεκτρικούς παλμούς με μεγάλη συχνότητα, άν όμως δεν υπερβαίνει το συγκεκριμένο όριο, τότε ο νευρώνας παράγει πολύ αραιά παλμούς σε τυχαίες χρονικές στιγμές (Αδρανής νευρώνας). Τελικά οι παλμοί που παράγονται ταξιδεύουν κατά μήκος του άξονα και τροφοδοτούν τους άλλους νευρώνες με τους οποίους συνδέεται ο νευρώνας που παρείγαγε τον παλμό [31].

3.2 Το μοντέλο McCulloch-Pitts

Το μοντέλο McCulloch-Pitts αποτελεί το πρώτο μαθηματικό μοντέλο τεχνητού νευρώνα. Προτάθηκε το 1943 από τους Warren McCulloch και Walter Pitts και σχεδιάστηκε για να προσομοιώσει τη λειτουργία ενός βιολογικού νευρώνα, χρησιμοποιώντας λογικές πράξεις. Η κατάσταση του νευρώνα περιγράφεται απο ένα δυαδικό αριθμό y.

- y = 0, ο νευρώνας είναι αδρανής
- y = 1, ο νευρώνας πυροδοτεί παλμούς (Δεν είναι αδρανής)

Οι συνάψεις περιγράφονται απο τα συναπτικά βάρη $w_i \in \mathbb{R}$, i = 1,...,n. Έστω πως x_1,x_2,...,x_n είναι οι εισόδοι του νευρώνα, ελέγχουμε εάν το άθροισμα x_1w_1+...+x_nw_n του φορτίου που δέχεται ο νευρώνας είναι μεγαλύτερο απο κάποιο κατώφλι θ. Εάν ισχύει τότε ο νευρώνας πυροδοτεί παλμούς. Διαφορετικά ο νευρώνας παραμένει αδρανής. Αυτό είναι ισοδύναμο με το εάν εάν η ποσότητα

$$u = \sum_{i=1}^{n} w_i x_i - \theta \tag{3.1}$$

είναι μεγαλύτερη ή μικρότερη από το μηδέν, οπότε η έξοδος του νευρώνα ισούται με

$$y = f(u) = \begin{cases} 1 & \text{εάν } u > 0 \\ 0 & \text{εάν } u \leq 0 \end{cases} \tag{3.2}$$

Η συνάρτηση u καλείτε διέγερση του νευρώνα και η f(·) συνάρτηση ενεργοποίησης. Στο μοντέλο McCulloch-Pitts η συνάρτηση ενεργοποίησης είναι η βηματική συνάρτηση 0/1. Η διέγερση u μπορεί να γραφεί επίσης και με την παρακάτω συνοπτική μορφή.

$$u = \bar{w}^\top \bar{x} - \theta \tag{3.3}$$

,όπου $\bar{w} = [w_1, ..., w_n]^\top$ είναι το διάνυσμα των συναπτικών βαρών και $\bar{x} = [x_1, ..., x_n]^\top$ είναι το διάνυσμα εισόδου. Το κατώφλι θ είναι ενας πραγματικός αριθμός όπως και τα $w_1, ..., w_n$. Κατ'αυτή την έννοια μπορούμε να απλοποιήσουμε την εξίσωση θέτοντας w_0 = -θ, το οποίο θα ονομάζεται πόλωση και θα είναι συνδεδεμένο με μια σταθερή είσοδο x_0 = 1. Επομένως, τώρα έχουμε

$$u = \sum_{i=0}^{n} w_i x_i \tag{3.4}$$

Παρακάτω δίνετε η σχηματική αναπαράσταση του μοντέλου McCulloch-Pitts.

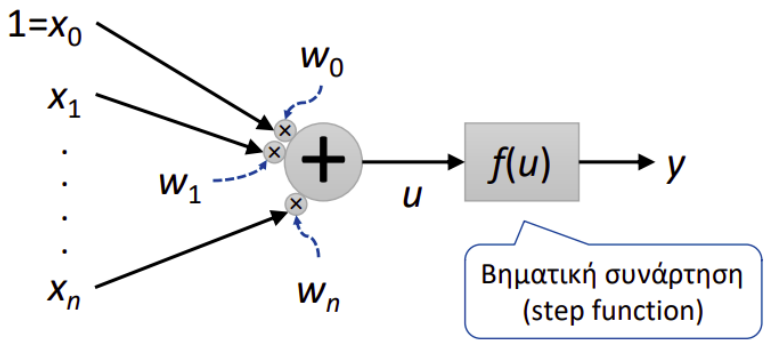

Figure 12: Μοντέλο McCulloch και Pitts του νευρώνα

3.2.1 Εναλλακτικές συναρτήσεις ενεργοποίησης

Υπάρχουν πολλές διαφορετικές μοντελοποιήσεις του νευρώνα που αποκλίνουν απο το μοντέλο McCulloch-Pitts. Η πιο σημαντική διαφορά εντοπίζεται στη μορφή της μη γραμμικής συνάρτησης f(·) που χρησιμοποιείτε στην έξοδο. Παρακάτω δίνετε η σχηματική αναπαράσταση μοντελοποίησης του νευρώνα που περιγράψαμε παραπάνω.

Η συνάρτηση ενεργοποίησης μπορεί να πάρει εναλλακτικά μεταξύ άλλων τις παρακάτω μορφές.

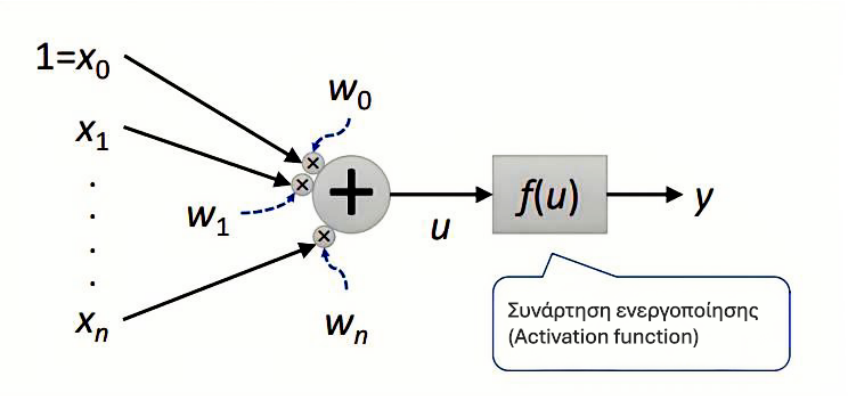

Figure 13: Σχηματική αναπαράσταση του νευρώνα

Βηματική συνάρτηση -1/1 (Step function -1/1)

Η βηματική συνάρτηση -1/1, όπως και η βηματική συνάρτηση 0/1 αποτελούν ειδικές περιπτώσεις μιας οικεγένειας συναρτήσεων που ονομάζονται βηματικές συναρτήσεις 2 επιπέδων (Binary step functions). Συγκεκριμένα, οι συναρτήσεις αυτού του τύπου, παραμένουν ανενεργές για τιμές μικρότερες ή ίσες με το κατώφλι (Στην περίπτωση μας το κατώφλι είναι ίσο με 0) και ενεργοποιούνται για τιμές μεγαλύτερες απο αυτό. Η βηματική συνάρτηση -1/1 δίνεται απο την σχέση

$$y = f(u) = \begin{cases} 1 & \text{εάν } u > 0 \\ -1 & \text{εάν } u \leq 0 \end{cases} \tag{3.5}$$

Το σημαντικότερο μειονέκτημα των συναρτήσεων αυτής της οικογένειας είναι ότι σε όλα τα σημεία τους έχουν είτε μηδενική κλίση είτε δεν είναι διαφορίσιμες (Συγκεκριμένα στο σημείο του κατωφλίου). Αυτό το μειονέκτημα σηνιστά σημαντικό εμπόδιο για την εκπαίδευση νευρωνικών δικτύων πολλών στρωμάτων. Για αυτό το λόγο, συναρτήσεις αυτής της μορφής χρησιμοποιούνται αποκλειστικά σε νευρωνικά δίκτυα με ένα μόνο στρώμα [32]. Η γραφική μορφή της βηματικής συνάρτησης -1/1 δίνετε παρακάτω.

Σιγμοειδής συνάρτηση (Sigmoid function)

Αναφέρεται επίσης και ως λογιστική συνάρτηση. Αποτελεί μη γραμμική συνάρτηση, η οποία χρησιμοποιείτε συνήθως σε νευρωνικά δίκτυα μιας κατεύθυνσης (Feedforward neural networks). Είναι φραγμένη και διαφορίσιμη με θετική παράγωγο για όλα τα σημεία του πεδίου ορισμού της, είναι δηλαδή γνησίως αύξουσα. Έχει πεδίο ορισμού το $\mathbb{R}$ και σύνολο τιμών το (0,1). Δίνετε απο την σχέση

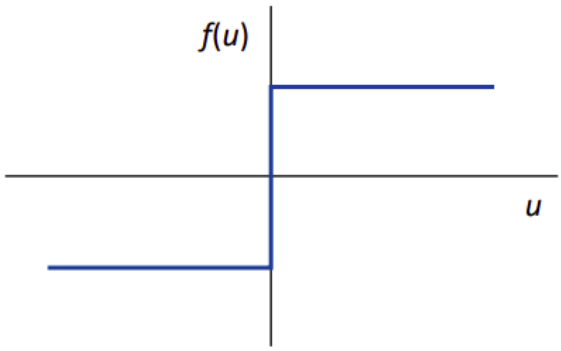

Figure 14: Γραφική μορφή βηματικής συνάρτησης -1/1

$$f(u) = \frac{1}{1 + e^{-u}} \tag{3.6}$$

Η σιγμοειδής συνάρτηση εμφανίζεται συνήθως στα στρώματα εξόδου των δικτύων βαθιάς μάθησης όταν θέλουμε η έξοδος να είναι υπό μορφή πιθανότητας. Παρά τα πλεονεκτήματα της όπως η εύκολη κατανόηση της και η καλή απόδοση της σε δίκτυα μικρού βάθους έχει κάποια σημαντικά μειονεκτήματα. Μερικά απο αυτά τα μειονεκτήματα περιλαμβάνουν την απότομη εξασθένιση των κλίσεων (Gradient) καθώς μεταφέρεται απο τα βαθύτερα πρός τα αρχικά στρώματα κατά την διαδικασία της οπισθοδρόμισης (Backpropagation), τον κορεσμό της λόγω των πολύ μικρών κλίσεων για μικρές αρνητικές και μεγάλες θετικές τιμές της εισόδου της συνάρτησης ενεργοποίησης καθώς και στο γεγονός πως η σιγμοδειδής συνάρτηση δεν είναι κεντραρισμένη στο 0, κάτι που μπορεί να οδηγήσει τις ανανεώσεις των βαρών να πραγματοποιούνται σε διαφορετικές κατευθύνσεις. Τα μειονεκτήματα αυτά μπορεί να δημιουργήσουν προβλήματα κατά την διαδικασία της εκπαίδευσης του μοντέλου όπως αργή σύγκλιση και αστάθεια κατά την διάρκεια της εκπαίδευσης [33]. Η γραφική μορφή της σιγμοειδής συνάρτησης δίνετε παρακατω.

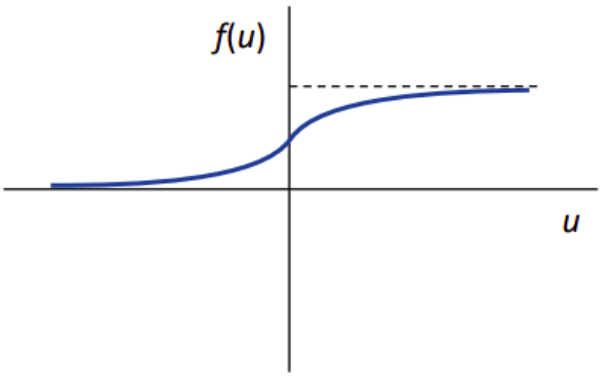

Figure 15: Γραφική μορφή σιγμοειδής συνάρτησης

Υπερβολική εφαπτομένη (Hyperbolic tangent)

Για την αντιμετώπιση ορισμένων προβλημάτων που εμφανίζονται χρησιμοποιώντας την σιγμοειδή συνάρτηση ως συνάρτησης ενεργοποίησης, προτάθηκε η χρήση της υπερβολικής εφαπτομένης. Η υπερβολική εφαπτομένη αποτελεί συνάρτηση κεντραρισμένη στο 0, με πεδίο ορισμού το $\mathbb{R}$ και σύνολο τιμών το (-1,1). Αποτελεί προτιμόμενη επιλογή σε σχέση με την σιγμοειδή συνάρτηση, καθώς διευκολύνει την εκπαίδευση νευρωνικων δικτύων πολλών στρωμάτων περιορίζοντας φαινόμενα που μπορεί να εμφανιστούν χρησιμοποιώντας ώς συνάρτηση ενεργοποίησης την σιγμοειδή συνάρτηση. Ωστόσο, η υπερβολική εφαπτομένη δεν επιλύει το πρόβλημα της εξασθένισης των κλίσεων που εμφανίζεται χρησιμοποιώντας την σιγμοειδή συνάρτηση. Το βασικό της πλεονέκτημα είναι πως παράγει έξοδο κεντραρισμένη στο 0, διευκολύνοντας έτσι την διαδικασία της εκπαίδευσης του μοντέλου. Δίνεται απο την σχέση

$$f(u) = tanh(u) = \frac{1 - e^{-u}}{1 + e^{-u}} \tag{3.7}$$

Είναι σημαντικό να αναφέρουμε πως η υπερβολική εφαπτομένη έχει κλίση ίση με 1, μόνο όταν η τιμή της εισόδου είναι ίση με 0. Αυτό έχει σαν αποτέλεσμα η συνάρτηση να δημιουργεί νεκρούς νευρώνες. Νεκρό νευρώνα ονομάζουμε μια κατάσταση κατά την οποία ο νευρώνας δεν χρησιμοποιείτε σχεδόν καθόλου κατά την διάρκεια της εκπαίδευσης ώς αποτέλεσμα των σχεδόν μηδενκών κλίσεων. Λύση στο πρόβλημα αυτό ήρθε να δώσει η συνάρτηση ενεργοποίησης ράμπας (ReLU activation function) [33]. Η γραφική μορφή της συνάρτησης υπερβολικής εφαπτομένης δίνετε παρακάτω.

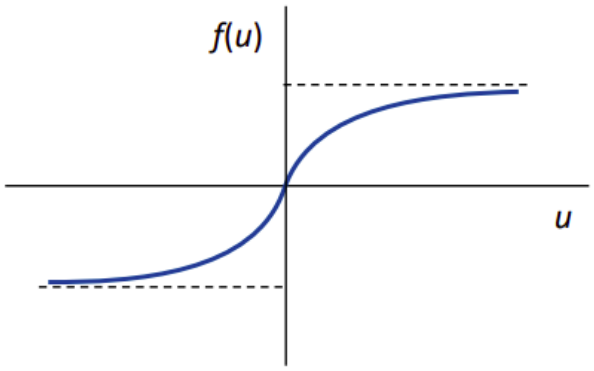

Figure 16: Γραφική μορφή υπερβολικής εφαπτομένης

Συνάρτηση ράμπας (ReLU)

Η συνάρτηση αυτή αποτελεί την πιο ευρέως χρησιμοποιημένη συνάρτηση ενεργοποίησης στα μοντέλα βαθιάς μάθησης με κορυφαία αποτελέσματα μέχρι και σήμερα. Η συνάρτηση ράμπας επιτρέπει γρήγορη εκπαίδευση του μοντέλου και αποτελεί την πιο επιτυχημένη συνάρτηση ενεργοποίησης μέχρι και σήμερα. Έχει παρουσιάσει καλύτερη απόδοση από τη σιγμοειδή και την υπερβολική εφαπτομένη σε μοντέλα βαθιάς μάθησης, ενώ επιπλέον προσφέρει καλύτερη ικανότητα γενίκευσης σε άγνωστα δεδομένα. Έχει πεδίο ορισμού το $\mathbb{R}$ και σύνολο τιμών το $[0,\infty)$. Δίνεται απο την σχέση

$$y = f(u) = \begin{cases} u & \text{if } u > 0 \\ 0 & \text{if } u \leq 0 \end{cases} \tag{3.8}$$

Η ReLU χρησιμοποιείτε κυρίως στα κρυφά στρώματα βαθιών νευρωνικών δικτύων, ενώ για τα στρώματα εξόδου χρησιμοποιούνται διαφορετικές συναρτήσεις ενεργοποίησης. Το βασικό της πλεονέκτημα είναι πως επιτρέπει πολύ γρήγορους υπολογισμούς λογω της πολύ μικρής πολυπλοκότητας υπολογισμού της, ενώ παράλληλα εισάγει αραιότητα στο μοντέλο, καθώς μηδενίζει τις εξόδους πολλών νευρώνων, απλοποιώντας έτσι τη δομή του. Ωστόσο η συνάρτηση αυτή δεν έρχετε χωρίς μειονεκτήματα. Η συνάρτηση ράμπας είναι πιο επιρρεπής απο την σιγμοειδή συνάρτηση στην υπερπροσαρμογή (overfitting). Για την αντιμετώπιση αυτού του προβλήματος χρησιμοποιείτε συχνά η τεχνική της απενεργοποίησης νευρώνων (Dropout). Σημαντικό είναι να αναφέρουμε πως κατά την διάρκεια της εκπαίδευσης είναι πιθανό, λόγω μηδενισμού της εξόδου, να δημιουργηθούν αρκετοί νεκροί νευρώνες, δηλαδή νευρώνες που τα βάρη τους έχουν σταματήσει να ενημερώνονται και δεν συμμετέχουν στην εκπαίδευση [33]. Η γραφική μορφή της συνάρτηση ράμπας δίνεται παρακάτω.

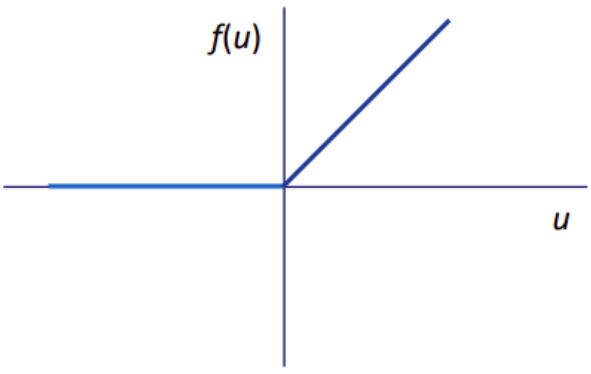

Figure 17: Γραφική μορφή συνάρτησης ReLU

Αφού παρουσιάσαμε την συνάρηση ράμπας ώς συνάρτηση ενεργοποίησης, αξίζει να δούμε και μια πιο σύγχρονη παραλλαγή της, την Gaussian Error Linear Unit (GELU). Η αυξημένη

πολυπλοκότητα των βαθιών μη γραμμικών μοντέλων τα καθιστά ικανά να προσαρμόζονται υπερβολικά καλά στα δεδομένα, οδηγώντας σε υπερπροσαρμογή. Αυτό συχνά απαιτεί απο τους σχεδιαστές των μοντέλων να επιλέξουν επιπλέον τεχνικές κανονικοποίησης, όπως η εισαγωγή θορύβου στα στρώματα και η απενεργοποίηση νευρώνων (Dropout). Αυτές οι τεχνικές δρούν συμπληρωματικά ώς πρός τις συναρτήσεις ενεργοποίησης και μπορούν να βελτιώσουν την ικανότητα γενίκευσης του μοντέλου. Η συνάρτηση GELU έχει σχεδιαστεί για να ενσωματώνει αυτήν τη στοχαστική συμπεριφορά πιο φυσικά, προσφέροντας μια πιο πιθανοκρατική ερμηνεία της εξόδου ενός νευρώνα. Στην πράξη, έχει δείξει ίση ή και καλύτερη απόδοση από την ReLU σε συγκεκριμένες εργασίες της όρασης υπολογιστών, της επεξεργασίας φυσικής γλώσσας και της αναγνώρισης ομιλίας. Η συνάρτηση GeLU συνδυάζει χαρακτηριστικά της απενεργοποίησης νευρώνων, της συνάρτησης ράμπας και του zoneout. Το zoneout αποτελεί τεχνική κανονικοποίησης που στοχαστικά κρατά κάποιες τιμές εξόδου των νευρώνων ίδιες με το προηγούμενο πέρασμα. Η όλη ιδέα υλοποιείται πολλαπλασιάζοντας την είσοδο του νευρώνα με 0 ή 1, όπου η επιλογή μεταξύ 0 και 1 γίνεται στοχαστικά και εξαρτάται από την τιμή της εισόδου. Συγκεκριμένα πολλαπλασιάζουμε την τιμή της εισόδου u με την κατανομή m $\sim$ Bernoulli$\big(\Phi(u)\big)$, όπου $\Phi(u) = P(X \leq u),\ X \sim \mathcal{N}(0,1)$. Η κατανομή αυτή επιλέχθηκε επειδή οι τιμές των εισόδων των νευρώνων τείνουν να ακολουθούν κανονική κατανομή. Στο πλαίσιο αυτό, όσο μικραίνει η τιμή της εισόδου, τόσο πιο πιθανό είναι να μηδενιστεί. Η διαδικασία είναι τυχαία, αλλά εξαρτάται από την ίδια την τιμή της εισόδου. Η τεχνική αυτή ουσιαστικά αυτό που κάνει μηδενίζει ή αφήνει την είσοδο ίδια. Η συνάρτηση GeLU δίνεται απο την σχέση

$$\mathrm{f}(u) = uP(X \leq u) = u\Phi(u) = u \cdot \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{u}{\sqrt{2}}\right)\right]. \tag{3.9}$$

, όπου $X \sim \mathcal{N}(0,1)$ και erf$(\cdot)$ η συνάρτηση σφάλματος, η οποία δίνεται απο την σχέση

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt \tag{3.10}$$

Η συνάρτηση GeLU μπορεί να προσεγγιστεί μέσω της σχέσης

$$f(u) \approx 0.5u\left(1 + \tanh\left[\sqrt{\frac{2}{\pi}}\left(u + 0.044715u^3\right)\right]\right) \tag{3.11}$$

ή μέσω της

$$f(u) \approx u\sigma(1.702u) \tag{3.12}$$

,όπου σ(·) η σιγμοειδής συνάρτηση. Η προσεγγιστικές μορφές της GeLU χρησιμοποιούνται σε περίπτωση όπου η ταχύτερος υπολογισμός της αξίζει το κόστος της ακρίβειας [34]. Η γραφική μορφή της συνάρτησης αυτής δίνετε παρακάτω.

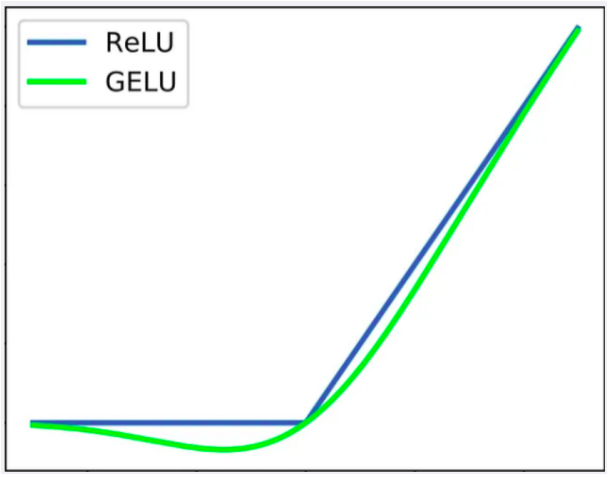

Figure 18: Γραφική μορφή συνάρτησης GeLU

Γραμμική συνάρτηση (Linear function)

Η γραμμική ή διαφορετικά ταυτοτική (Identity) συνάρτηση έχει πεδίο ορισμού και σύνολο τιμών το $\mathbb{R}$. Δίνεται απο την σχέση

$$f(u) = u \tag{3.13}$$

Παρά την απλότητά της, παρουσιάζει σημαντικά μειονεκτήματα. Η χρήση γραμμικής συνάρτησης ως συνάρτησης ενεργοποίησης περιορίζει το μοντέλο, καθώς λόγω της γραμμικής μορφής του δεν μπορεί να μάθει πολύπλοκες μη γραμμικές σχέσεις [35]. Επιπλέον, δεν είναι δυνατή η χρήση οπισθοδρόμισης για την εκπαίδευση του μοντέλου, καθώς η παράγωγος της συνάρτησης είναι σταθερή και δεν έχει σχέση με την τιμή της εισόδου u [36]. Η γραφική μορφή της γραμμικής συνάρτησης δίνετε παρακάτω

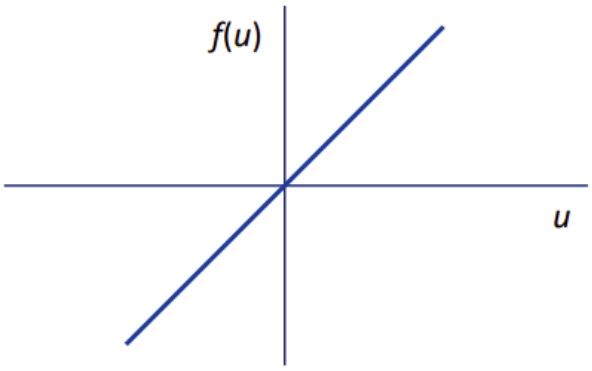

Figure 19: Γραφική μορφή γραμμικής συνάρτησης

Η ειδική περίπτωση τεχνητού νευρώνα με συνάρτηση ενεργοποίησης την βηματική συνάρτηση καλείτε δίκτυο Perceptron [37]. Το δίκτυο Perceptron είναι το πιο απλό νευρωνικό δίκτυο που μπορεί να σχεδιαστεί.

3.3 Νευρωνικά δίκτυα πολλών στρωμάτων

3.3.1 Το δίκτυο MLP

Στα γραμμικά μοντέλα όπως το δίκτυο Perceptron οι δυνατότητες αναπαράστασης διαχωριστικών επιφανειών είναι περιορισμένες, επειδή το δίκτυο μπορεί να αναπαραστήσει μόνο επίπεδες επιφάνειες. Με άλλα λόγια το υπερεπίπεδο u=0, χωρίζει τον χώρο $\mathbb{R}^n$ σε 2 μέρη, όπου στο ένα ισχύει f(u) = 1 και στο άλλο f(u) = 0 (Στην περίπτωση που χρησιμοποιούμε την 0/1 βηματική συνάρτηση αντί την -1/1). Η κατάσταση που προκύπτει μπορεί να οπτικοποιηθεί καλύτερα στις 2 διαστάσεις. Στον χώρο $\mathbb{R}^2$ η εξίσωση u=$w_1x_1 + w_2x_2 + b$ = 0 ορίζει μια ευθεία κάθετη στο διάνυσμα των συναπτικών βαρών $\bar{w} = [w_1, w_2]^T$. Η ευθεία αυτή χωρίζει το επίπεδο σε 2 τμήματα.

- Το τμήμα πρός την κατεύθυνση του $\bar{w}$ περιέχει τα $\bar{x}$ για τα οποία ισχύει u > 0 (και άρα f(u) = 1)
- Το τμήμα πρός την αντίθετη κατεύθυνση του $\bar{w}$ περιέχει τα σημεία $\bar{x}$ για τα οποία ισχύει u < 0 (και άρα f(u) = 0)

Η απόσταση της ευθείας απο την αρχή των αξόνων εξαρτάται απο την τιμή της πόλωσης w_0 = b. Η σχηματική μορφή της παραπάνω ευθείας στον χώρο $\mathbb{R}^2$ δίνεται παρακάτω.

Ο περιορισμός αυτός αίρεται με τη χρήση περισσότερων νευρώνων. Η χρήση περισσότερων κρυφών νευρώνων, θα μπορούσε να ορίσει περισσότερες διαχωριστικές ευθείες. Ο συνδυασμός των ευθειών αυτών μπορεί να μας δώσει μεγάλη ποικιλία περιοχών που θα μπορούσαμε να διαχωρίσουμε στην έξοδο. Υπάρχουν άπειρα παραδείγματα τέτοιων σχηματισμών, τα οποία μπορούν να απεικονιστούν σε διαφορετικές διαστάσεις. Στη συνέχεια παρουσιάζεται μια ενδεικτική

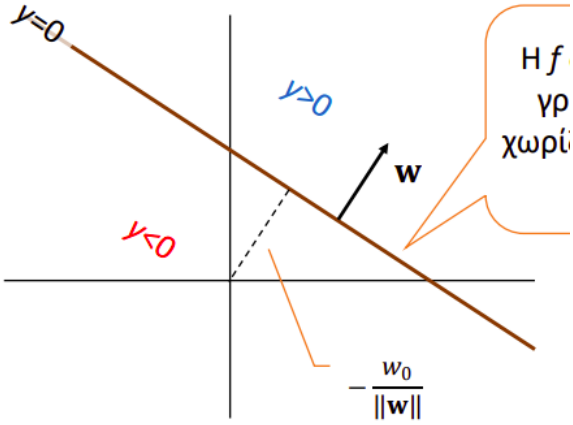

Figure 20: Ταξινόμηση στον $\mathbb{R}^2$ χρήση δικτύου Perceptron

σχηματική απεικόνιση της μορφής που μπορεί να προκύψει στον χώρο $\mathbb{R}^2$.

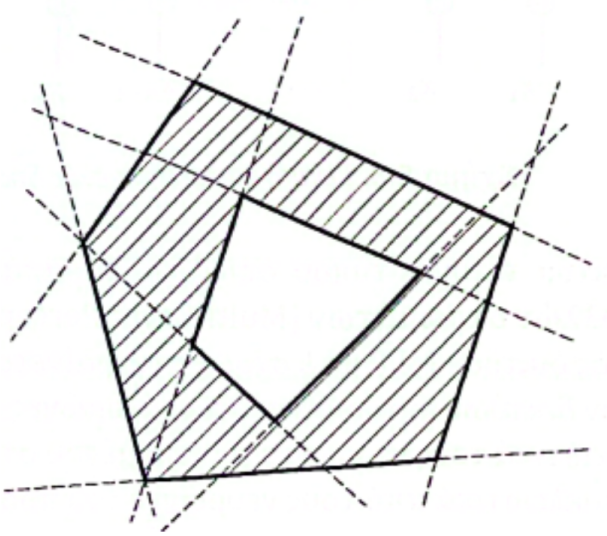

Figure 21: Ταξινόμηση στον $\mathbb{R}^2$ χρήση MLP

Δίκτυα τέτοιου τύπου καλούνται δίκτυα Perceptron πολλών στρωμάτων (Multi-Layer Perceptron - MLP). Η γενική αρχιτεκτονική ενός δικτύου MLP με L στρώματα φαίνετε παρακάτω. Το χαρακτηριστικό των δικτύων αυτών είναι πως οι νευρώνες του στρώματος l τροφοδοτούν αποκλειστικά τους νευρώνες του επόμενου στρώματος l+1 και τροφοδοτούνται αποκλειστικά από τους νευρώνες του προηφούμενου στρώματος l-1.

Τα δίκτυα Perceptron πολλών στρωμάτων στα οποία οι νευρώνες χρησιμοποιούν την βηματική συνάρτηση 0/1 ή -1/1, όπως έχουμε ήδη διαπιστώσει μπορούν να υλοποιήσουν συναρτήσεις που δεν είναι εφικτό να υλοποιηθούν με ένα απλό δίκτυο Perceptron. Ωστόσο, η χρήση της βηματικής συνάρτησης δεν προτιμάται. Ο λόγος είναι ότι οι περισσότεροι κανόνες εκπαίδευσης

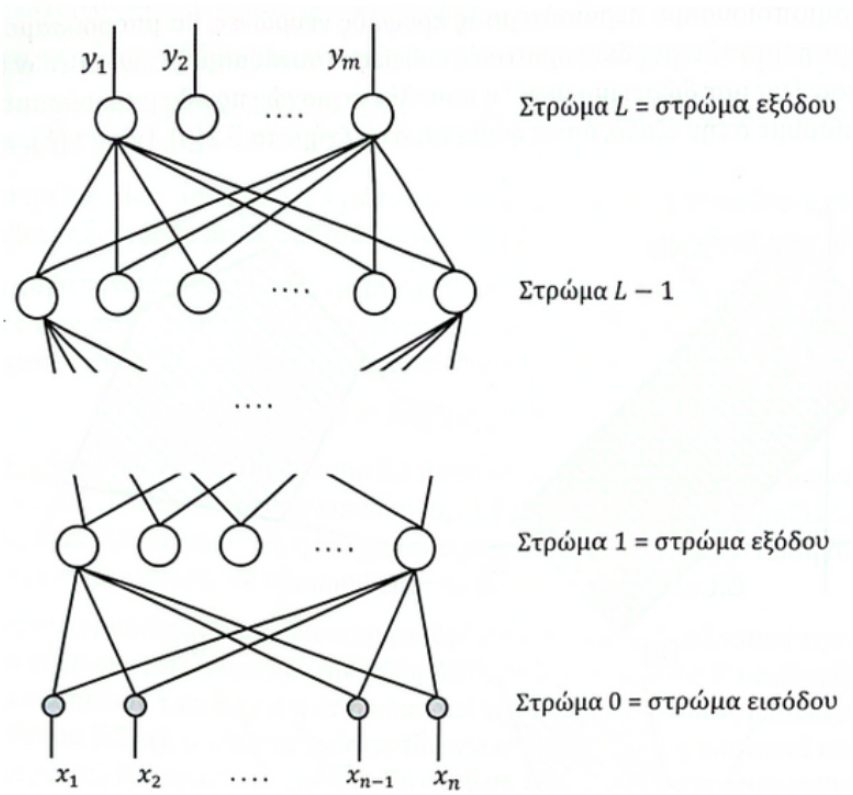

Figure 22: Γενική σχηματική μορφή δικτύου Perceptron πολλών στρωμάτων

βασίζονται σε μεθόδους βελτιστοποίησης, οι οποίες χρησιμοποιούν παραγώγους, ενώ η βηματική συνάρτηση δεν είναι παραγωγίσιμη. Αυτή είναι μια τεχνική δυσκολία η οποία παρ'όλα αυτά ξεπερνιέται με την χρήση της σιγμοειδούς συνάρτησης. Η σιγμοειδής συνάρτηση είναι παραγωγίσιμη και μοιάζει πολύ με την βηματική συνάρτηση 0/1.

Εναλλακτικά μπορούν να χρησιμοποιηθούν συναρτήσεις όπως η υπερβολική εφαπτομένη, η οποία είναι παραγωγίσιμη και μοιάζει με την βηματική -1/1, η συνάρτηση ράμπας κ.ο.κ.

Πολύ σημαντικό είναι να αναφέρουμε πως ένα δίκτυο MLP μπορεί να υλοποιήσει οποιαδήποτε διαχωριστική επιφάνεια σε n διαστάσεις, σε αντίθεση με το απλό δίκτυο Perceptron που μπορεί να υλοποιήσει μόνο ευθείες επιφάνειες. Η απόδειξη δίνεται στο [38]. Πράγματι, άν θέλουμε να προσεγγίσουμε οποιαδήποτε διαχωριστική επιφάνεια στον χώρο $\mathbb{R}^n$ χρησιμοποιώντας ένα MLP, αρκεί να βρούμε μια συνάρτηση g(x) τέτοια ώστε για κάποιο κατώφλι θ, ο χώρος $\mathbb{R}^n$ να χωρίζεται σε δύο τμήματα. Για τα μισά θα ισχύει g(x) > θ και για τα άλλα μισά g(x) < θ.

Ανάκληση είναι η διαδικασία υπολογισμού των τιμών όλων των νευρώνων του δικτύου με δεδομένες τις τιμές των εισόδων. Ορίζουμε αρχικά ως

- L το πλήθος των στρωμάτων του δικτύου εκτός του στρώματος εισόδου.

- N(l) είναι το πλήθος των νευρώνων του στρώματος l, l = 0,...,L.
- $\alpha_i(l)$ είναι οι ενεργοποιήσεις των νευρώνων του στρώματος l.
- $w_{ij}(l)$ είναι το συναπτικό βάρος που συνδέει τον νευρώνα $\alpha_j(l-1)$ του στρώματος l-1 με τον νευρώνα $\alpha_i(l)$ του στρώματος l.
- $w_{i0}(l)$ είναι η πόλωση του νευρώνα $alpha_i(l)$ του στρώματος l.
- $x_i = \alpha_i(0)$ είναι οι είσοδοι του δικτύου.
- $y_i = \alpha_i(L)$ είναι οι έξοδοι του δικτύου.

Οι ενεργοποιήσεις των νευρώνων για οποιαδήποτε στρώμα δίνονται από την σχέση

$$a_i(l) = f\left(\sum_{j=1}^{N(l-1)} w_{ij}(l)\, a_j(l-1) + w_{i0}(l)\right) \tag{3.14}$$

Αυτός είναι ο τύπος ενεργοποίησης ενός νευρώνα. Όπως δείχνει η παραπάνω σχέση, ο νευρώνας i του στρώματος l δέχεται ως εισόδους τις ενεργοποιήσεις $\alpha_i(l-1)$ των νευρώνων απο το στρώμα l-1 και ως πόλωση την τιμή $w_{i0}(l)$. Κατά την ανάκληση μας δίνονται οι τιμές x_i των εισόδων του δικτύου, οπότε με βάση της εισόδους υπολογίζουμε πρώτα τις ενεργοποιήσεις των νευρώνων του στρώματος 1, κατόπιν βάση αυτές υπολογίζουμε τις ενεργοποιήσεις του στρώματος 2 κ.ο.κ.

3.3.2 Ο αλγόριθμος εκπαίδευσης Back-Propagation

Η εκπαίδευση ενός δικτύου πολλών στρωμάτων είναι η διαδικασία καθορισμού των συναπτικών βαρών του έτσι ώστε να ικανοποιείται κάποιο κριτήριο καταλληλότητας. Αυτός είναι και ο λόγος της εκπαίδευσης σε οποιοδήποτε νευρωνικό δίκτυο. Κυριότερος εκπρόσωπος των αλγορίθμων εκπαίδευσης Peceptron πολλών στρωμάτων είναι ο αλγόριθμος Back-Propagation. Ο αλγόριθμος Back-Propagation προτάθηκε από τον Paul Werbos στη δεκαετία του 1970 στα πλαίσια της ανάλυσης μοντέλων οικονομικής και πολιτικής πρόβλεψης. Στη δεκαετία του 1980 έγινε αντιληπτό πως η μέθοδος μπορούσε να μεταφερθεί αυτούσια στην εκπαίδευση νευρωνικών δικτύων πολλών στρωμάτων και έκτοτε έγινε η πιο γνωστή και η πιο διαδεδομένη μέθοδος για τον σκοπό αυτό.

Βασικό χαρακτηριστικό της μεθόδου είναι η ύπαρξη στόχων. Συνεπώς, το μοντέλο ανήκει στην κατηγορία των δικτύων που εκπαιδεύονται με επίβλεψη. Έστω δίκτυο με L στρώματα, n εισόδους και m εξόδους. Ορίζουμε ως

- $\mathbf{x}^{(p)} = \left[x_1^{(p)}, \ldots, x_n^{(p)}\right]^T$ το p-οστό διάνυσμα εισόδου
- $\mathbf{y}^{(p)} = \left[y_1^{(p)}, \ldots, y_m^{(p)}\right]^T$ το p-οστό διάνυσμα εξόδου
- $\mathbf{t}^{(p)} = \left[t_1^{(p)}, \ldots, t_m^{(p)}\right]^T$ το p-οστό διάνυσμα στόχων

Τα δεδομένα που απαιτούνται για να εκπαιδευτεί το δίκτυο είναι τα ζεύγη διανυσμάτων $\{\mathbf{x}^{(i)}, \mathbf{t}^{(i)}\}$, i = 1,...,P. Θα ήταν ιδανικό να πετυχέναμε τάυτιση εξόδων και στόχων για κάθε πρότυπο εισόδου, ωστόσο αυτό μπορεί να μην είναι απολύτως εφικτό. Για αυτό τον λόγο επιζητούμε τη βέλτιστη προσέγγιση της επιθυμητής κατάστασης χρησιμοποιώντας ένα κριτήριο κόστους.

Βιβλιογραφική αναφορά

[1] K. Gao, G. Mei, F. Piccialli, S. Cuomo, J. Tu, and Z. Huo, "Julia language in machine learning: Algorithms, applications, and open issues," *Computer Science Review*, vol. 37, p. 100254, Aug. 2020.

[2] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *npj Digital Medicine*, vol. 4, no. 1, p. 5, 2021.

[3] X. Dong and M. Cappuccio, "Applications of computer vision in autonomous vehicles: Methods, challenges and future directions," 11 2023.

[4] M. T. Shahria, M. S. Sunny, I. Islam, J. Ghommam, S. Ahamed, and M. Rahman, "A comprehensive review of vision-based robotic applications: Current state, components, approaches, barriers, and potential solutions," *Robotics*, vol. 11, 12 2022.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2015.

[7] E. Chris, J. Kate, and A. Juliet, "The evolution of computer vision: From pixels to perception," 12 2025.

[8] G. Boesch, "Computer vision tasks (comprehensive 2025 guide)," October 2024.

[9] L. Zhou, G. Wu, Y. Zuo, X. Chen, and H. Hu, "A comprehensive review of vision-based 3d reconstruction methods," *Sensors*, vol. 24, p. 2314, April 2024.

[10] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *CoRR*, vol. abs/2001.05566, 2020.

[11] G. Csurka, R. Volpi, and B. Chidlovskii, "Semantic image segmentation: Two decades of research," 2023.

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016.

[13] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[14] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, 2017.

[15] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: state of the art," *International Journal of Multimedia Information Retrieval*, vol. 9, p. 171–189, July 2020.

[16] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015.

[17] O. Elharrouss, S. Al-Maadeed, N. Subramanian, N. Ottakath, N. Almaadeed, and Y. Himeur, "Panoptic segmentation: A review," 2021.

[18] D. Shah, "Coco dataset: All you need to know to get started," 2023.

[19] Z. Fan, Y. Wang, Y. Zhu, and J. Zhao, "Preparing state-of-the-art models for classification and object detection with nvidia tao toolkit," February 2021. Accessed: 2025-06-24.

[20] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2022.

[21] P. Taghavi, R. Langari, and G. Pandey, "Swinmtl: A shared architecture for simultaneous depth estimation and semantic segmentation from monocular camera images," 2024.

[22] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014.

[23] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

[24] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," 2018.

[25] A. Bhandari, "Confusion matrix in machine learning," 2025.

[26] V. Kookna, "Semantic vs. instance vs. panoptic segmentation," 2022.

[27] B. T., "Comprehensive guide to multiclass classification metrics," 2021.

[28] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," 2020.

[29] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[30] D. Shah, "Mean average precision (map) explained: Everything you need to know," 2022.

[31] J. Zhang, "Basic neural units of the brain: Neurons, synapses and action potential," 2019.

[32] A. Navlani, "Activation functions," 2022.

[33] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018.

[34] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2023.

[35] S. M. K, "Linear activation function," 2023.

[36] P. Baheti, "Activation functions in neural networks [12 types & use cases]," 2021.

[37] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[38] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.