







ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ  
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

# Μοντελοποίηση Αβεβαιότητας σε Βαθιά Νευρωνικά Δίκτυα

Μια Σύγχρονη Προσέγγιση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΛΕΞΑΝΔΡΟΣ Κ. ΜΠΑΡΜΠΕΡΗΣ

Επιβλέπων: Στέφανος Κόλλιας  
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26η Ιουνίου 2022.

.....  
Στέφανος Κόλλιας  
Καθηγητής Ε.Μ.Π.

.....  
Ανδρέας-Γεώργιος  
Σταφυλοπάτης  
Καθηγητής Ε.Μ.Π.

.....  
Γιώργος Στάμου  
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2022



# ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ  
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Copyright © Αλέξανδρος Μπαρμπέρης, 2022. Με επιφύλαξη παντός δικαιώματος. All rights reserved.

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 δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

## ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δε μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(υπογραφή)

.....

**Αλέξανδρος Μπαρμπέρης**

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Αθήνα, Ιούνιος 2022





# Περίληψη

Το

Λέξεις κλειδιά

πρώτη λέξη





# Abstract

Cloud computing is the dominant approach to compute infrastructure,

## Keywords

enwords



# Ευχαριστίες

Για την παρούσα εργασία, που σηματοδοτεί την ολοκλήρωση μίας πορείας ετών, θα ήθελα να ευχαριστήσω



# Περιεχόμενα

Περίληψη	i
Abstract	iii
Ευχαριστίες	v
Περιεχόμενα	viii
<b>1 Εισαγωγή</b>	<b>1</b>
1.1 Επισκόπηση του Κόσμου της Τεχνητής Νοημοσύνης . . . . .	1
1.2 Ιστορική Αναδρομή Τεχνητής Νοημοσύνης . . . . .	4
1.3 Κίνητρο . . . . .	10
1.4 Συνεισφορά Εργασίας . . . . .	11
1.5 Οργάνωση του Τόμου . . . . .	12
<b>2 Θεωρητικό Υπόβαθρο</b>	<b>13</b>
2.1 Τεχνητά Νευρωνικά Δίκτυα . . . . .	13
2.1.1 Μηχανική Μάθηση . . . . .	14
2.1.2 Εκμάθηση Χαρακτηριστικών . . . . .	15
2.1.3 Πολυεπίπεδα Νευρωνικά Δίκτυα . . . . .	17
2.2 Νευρωνικά Δίκτυα με Κάψουλες . . . . .	34
2.2.1 Στοιχεία Έμπνευσης των Νευρωνικών Δικτύων με Κάψουλες . . . . .	34
2.2.2 Θετικά Γνωρίσματα Συνελικτικών Νευρωνικών Δικτύων . . . . .	36
2.2.3 Βασικές Ανεπάρκειες των Συνελικτικών Νευρωνικών Δικτύων . . . . .	37
2.2.4 Αρχές Λειτουργίας Νευρωνικών Δικτύων με Κάψουλες . . . . .	39
2.3 Μετασχηματιστές . . . . .	51
2.4 Χάρτες Αυτο-οργάνωσης . . . . .	51
2.5 Ρεζονστρυσιον ας Ρεγυλαριζατιον . . . . .	51
<b>3 Σχετικές Ερασίες</b>	<b>52</b>
<b>4 Μέθοδος</b>	<b>53</b>
<b>5 Πειράματα</b>	<b>54</b>
<b>6 Επίλογος</b>	<b>55</b>

Βιβλιογραφία	57
Α΄ Ορισμοί Εννοιών	61
Β΄ Απόδοση Ξενόγλωσσων Όρων	66
Γ΄ Συντομογραφίες - Ακρωνύμια	67
Γ΄.1 Ελληνικά . . . . .	67
Γ΄.2 Αγγλικά . . . . .	67

# Κεφάλαιο 1

## Εισαγωγή

«Η επιτυχημένη δημιουργία [γενικευμένης] Τεχνητής Νοημοσύνης θα είναι το μεγαλύτερο γεγονός στην ανθρώπινη ιστορία. Δυστυχώς, ίσως είναι και το τελευταίο εάν δε μάθουμε πώς να αποφεύγουμε τα ρίσκα.» — Stephen Hawking

### 1.1 Επισκόπηση του Κόσμου της Τεχνητής Νοημοσύνης

Είναι πλέον γεγονός, η Τεχνητή Νοημοσύνη (Artificial Intelligence - AI) εντοπίζεται σε πολλές εκφάνσεις της καθημερινότητας των περισσότερων ανθρώπων [1]. Δεν αποτελεί απλά έναν ακόμα μοδάτο όρο που καταχράται στον χώρο της αγοραστικής (marketing). Εν αντιθέσει, διαδραματίζει καθοριστικό ρόλο στην εργασία μας, στη μετακίνησή μας και στην ψυχαγωγία μας. Για παράδειγμα, εντοπίζεται στις μηχανές αναζήτησης όπως η Google, στους ηλεκτρονικούς χάρτες πλοήγησης αλλά και στα συστήματα συστάσεων (Recomender systems) όπως αυτό του YouTube, του Twitter και του Netflix [2] που εξατομικεύουν το προβαλλόμενο περιεχόμενο στα ενδιαφέροντα του χρήστη.

Η επιρροή που έχει η Τεχνητή Νοημοσύνη είναι ακόμα πιο ευδιάκριτη αν τη μελετήσει κανείς υπό μια συλλογική σκοπιά. Για παράδειγμα, στον χώρο του επιχειρείν, η αξιοποίηση τεχνολογιών Τεχνητής Νοημοσύνης έχει αποδειχθεί ότι αυξάνει την επιχειρηματική αξία (business value) μέσα από τη βελτίωση της επίδοσης τόσο στο οικονομικό (financial), αγοραστικό (marketing) και διοικητικό (administrative) επίπεδο όσο και στο επίπεδο επιχειρηματικών διαδικασιών (business process) [3]. Χαρακτηριστικό παράδειγμα αποτελεί η χρήση των συστημάτων συστάσεων αφού με το να φιλτράρουν το περιεχόμενο και να παρουσιάζουν στον χρήστη μόνο αυτό που του είναι οικείο και θεμιτό, αυξάνουν τον βαθμό ενασχόλησής του με κίνδυνο την παγίδευσή του σε μια «φούσκα προκατειλημμένου φιλτραρίσματος» («biased filter bubble») [4]. Άλλωστε, όπως δηλώνει η ομάδα ανάπτυξης του εν λόγω συστήματος για λογαριασμό της συνδρομητικής υπηρεσίας streaming, Netflix, «Πιστεύουμε ότι αθροιστικά, η επίδραση της εξατομικεύσης και των συστάσεων μας εξοικονομεί ένα δισεκατομμύριο δολάρια τον χρόνο» [5].

Στον εργασιακό χώρο, πολλές δουλειές που περιλαμβάνουν επαναλαμβανόμενες, προβλέψιμες εργασίες κυρίως στον τομέα της βιομηχανίας και της γεωργίας αντικαθίστανται από αυτοματισμούς

Τεχνητής Νοημοσύνης (AI automations) εκτοπίζοντας έτσι τον άνθρωπο. Η μείωση των διαθέσιμων θέσεων εργασίας στους τομείς αυτούς δοκιμάζει τα όρια του κοινωνικού οικοδομήματος: τα «εκτοπισμένα» άτομα καλούνται να αποκτήσουν νέες δεξιότητες προκειμένου να βρουν απασχόληση στις πιο δημιουργικές (και συνάμα λιγότερο τυποποιήσιμες) θέσεις του εξελισσόμενου τομέα των υπηρεσιών [6]. Βέβαια, η μείωση των θέσεων εργασίας επαναλαμβανόμενης φύσης είναι μόνο η μια πλευρά του νομίσματος. Σύμφωνα με αναλύσεις, μέχρι την επόμενη δεκαετία οι εφαρμογές της Τεχνητής Νοημοσύνης εν δυνάμει θα αυξήσουν το παγκόσμιο Ακαθάριστο Εθνικό Προϊόν (Gross Domestic Product - GDP) κατά 26% (δεκαπέντε τρισεκατομμύρια δολάρια) [7]. Αυτό, με τη σειρά του, θα οδηγήσει στη δημιουργία πολλών νέων θέσεων εργασίας έτσι ώστε να μην παρατηρηθεί αύξηση στους δείκτες ανεργίας [7, 8].

Τέλος, δε θα μπορούσαμε να παραλείψουμε την επιρροή που έχει η Τεχνητή Νοημοσύνη στον χώρο της υγείας. Οι εφαρμογές είναι ατελείωτες: από συστήματα για πρόωρη διάγνωση ασθενειών μέχρι ρομποτικά συστήματα υποβοήθησης χειρουργείου [9]. Αξιοσημείωτη είναι επίσης η επιτυχημένη εφαρμογή της για την πρόβλεψη της τρισδιάστατης δομής των πρωτεϊνών [10] - ένα θέμα με σημαντικές προεκτάσεις που απασχολούσε την επιστημονική κοινότητα για 50 χρόνια. Παρόλα αυτά, μαζί με την προσπάθεια για αξιοποίηση των νέων τεχνολογιών στην κλινική πράξη προκύπτουν νέες προκλήσεις. Μια πρώτη δυσκολία είναι η ανάπτυξη συστημάτων Τεχνητής Νοημοσύνης που απαιτούν μεγάλο όγκο δεδομένων σε χώρους προβλημάτων όπου αυτά σπανίζουν (όπως για παράδειγμα, στην περίπτωση μιας ασυνήθιστης ασθένειας όπου ο αριθμός των ιατρικών υποθέσεων είναι ελάχιστος). Αυτή η δυσκολία εντείνεται αφενός λόγω της έλλειψης ασφαλών υποδομών για τη συλλογή ιατρικών δεδομένων [11] και αφετέρου λόγω της απόρρητης φύσης αυτών, κάτι που δυσχεραίνει τον ελεύθερο διαμοιρασμό τους. Μια τελευταία δυσκολία αποτελεί το γεγονός ότι πολλά συστήματα Τεχνητής Νοημοσύνης που έχουν αναπτυχθεί σε περιβάλλον εργαστηρίου (Lab setting) δεν παρέχουν αρκετά κίνητρα για μεταστροφή της καθιερωμένης κλινικής πράξης [11]. Για να επιτευχθεί κάτι τέτοιο, μεταξύ άλλων θα πρέπει τα συστήματα να αποδίδουν αποδεδειγμένα τόσο καλά όσο και το καταρτισμένο προσωπικό στο συγκεκριμένο πεδίο εφαρμογής τους και να παρέχουν πληροφορίες που θα τα καθιστούν περισσότερο έμπιστα π.χ. αιτιολογώντας την απόφασή τους (explainability), παρέχοντας μια μετρική αβεβαιότητας (uncertainty) ή δίνοντας τη δυνατότητα αλληλεπίδρασης [12].

Αντιλαμβανόμενοι το εύρος των εφαρμογών της Τεχνητής Νοημοσύνης, η εκτίμηση από την International Data Corporation - IDC πως οι Ευρωπαϊκές δαπάνες σε τέτοιες εφαρμογές θα έχουν σχεδόν τριπλασιαστεί μέσα στα επόμενα τρία χρόνια δε θα πρέπει να μας εκπλήσσει. Ωστόσο, με τη μεγάλη ισχύ έρχεται και μεγάλη ευθύνη. Είναι αλήθεια, η Τεχνητή Νοημοσύνη είναι ήδη εδώ και θα συνεχίζει να επιδρά όλο και εντονότερα στην καθημερινότητα μας και στην κοινωνία. Βέβαια, η σημερινή Τεχνητή Νοημοσύνη είναι εντελώς διαφορετική από αυτό που φαντάζονταν η κοινή γνώμη τις προηγούμενες δεκαετίες (σαφώς επηρεασμένη από ταινίες επιστημονικής φαντασίας όπως το Terminator). Θα λέγαμε, αντίθετα, πως εμφανίζεται περισσότερο με μια περιορισμένη μορφή στην εκάστοτε συγκεκριμένη εφαρμογή (narrow AI). Έτσι, ένα «εφυές» σύστημα για μια εργασία δεν μπορεί να «γενικεύσει» και να εφαρμοστεί σε άλλο χώρο προβλημάτων. Ούτε λόγος δε για αισθήματα και υπαρξιακή συνείδηση: αυτά (ακόμα) ανήκουν στην επιστημονική φαντασία. Αυτό όμως δε σημαίνει ότι η επιπόλαιη χρήση της Τεχνητής Νοημοσύνης δεν ελοχεύει κινδύνους. Σύμφωνα με το περιοδικό Spectrum της IEEE [13] προτού



επιτευχθεί Τεχνητή Νοημοσύνη επιπέδου ανθρώπου (human-like Artificial Intelligence) - αυτή στην οποία αναφέρεται ο Stephen Hawking - υπάρχουν ήδη πολλά σενάρια όπου εφαρμογές της μπορούν να αποβούν μοιραίες. Ενδεικτικά, ένα από αυτά είναι τα deepfakes - ψεύτικα πολυμέσα βίντεο και εικόνες κατασκευασμένα από εφαρμογές Τεχνητής Νοημοσύνης - έχουν υπονομεύσει την εμπιστοσύνη στα συστήματα πληροφόρησης. Επιπρόσθετα, ένα ακόμα καταστροφικό σενάριο σχετίζεται με την ιδιωτικότητα (privacy) και την ελεύθερη βούληση (free will). Με την παραχώρηση ευαίσθητων δεδομένων σε επιχειρήσεις και κυβερνήσεις τους παρέχουμε τη δυνατότητα να μας εποπτεύουν ακόμα και να μας χειραγωγούν. Ένα τελευταίο σενάριο για το οποίο διαδραματίζουν άμεσο ρόλο τα κοινωνικά δίκτυα είναι αυτό του μειωμένου διαστήματος προσοχής (short attention span) ως απόρροια της εκμετάλλευσης του μηχανισμού επιβράβευσης του εγκεφάλου ώστε οι χρήστες να εθίζονται σε αυτά. Το περιοδικό καλεί τον αναγνώστη να αναλογιστεί τις συνέπειες της συνεχόμενης βελτίωσης των μηχανισμών που μας καθηλώνουν από τη νέα τεχνολογία. Συμπερασματικά, η Τεχνητή Νοημοσύνη αν και δεν προσομοιάζει την ανθρώπινη νοημοσύνη δεν παύει να αποτελεί μια πολύ ισχυρή τεχνολογία που μπορεί να αποβεί είτε σωτήρια είτε μοιραία ανάλογα με τον τρόπο αξιοποίησής της.

Είναι λοιπόν απαραίτητη η εξασφάλιση της συνετής χρήσης αυτών των τεχνολογιών μέσω μιας σειράς κανονισμών. Στην Ευρωπαϊκή Ένωση, μια σειρά από διατάξεις επιχειρούν να θέσουν ένα νομοθετικό πλαίσιο ώστε να ωθήσουν στην αξιοποίηση της Τεχνητής Νοημοσύνης διασφαλίζοντας παράλληλα την ασφάλεια των θεμελιωδών δικαιωμάτων [14]. Άλλωστε, σύμφωνα με την von der Lein [15], «Η Τεχνητή Νοημοσύνη πρέπει να εξυπηρετεί τους ανθρώπους και συνεπώς, πρέπει πάντα να συμμορφώνεται με τα δικαιώματά τους. Αυτός είναι ο λόγος που ένα άτομο πρέπει πάντα να έχει τον έλεγχο στην περίπτωση κρίσιμων αποφάσεων[...] Εφαρμογές της Τεχνητής Νοημοσύνης που μπορεί να παρέμβουν στα ανθρώπινα δικαιώματα θα πρέπει να ελέγχονται και να πιστοποιούνται πριν φτάσουν στην Ευρωπαϊκή αγορά. » Αν και οι αυστηροί διακανονισμοί καθυστερούν τη μετάβαση εφαρμογών Τεχνητής Νοημοσύνης από το εργαστήριο στην αγορά, εξασφαλίζουν την ασφάλειά τους συμβάλλοντας στην αξιοπιστία τους.

Με την παραπάνω σύντομη εισαγωγή καλύψαμε εμπεριστατωμένα πολλά από τα μη τεχνικά θέματα που σχετίζονται με την Τεχνητή Νοημοσύνη (Artificial Intelligence). Αναλυτικότερα, αρχίσαμε από παραδείγματα εντοπισμού της στην καθημερινή ζωή σε ατομικό και σε συλλογικό επίπεδο με τα οποία αντιληφθήκαμε τη σημασία της. Συνεχίσαμε με τους κινδύνους που ελοχεύει η απερίσκεπτη εφαρμογή της σε συγκεκριμένες εργασίες επισημαίνοντας ταυτόχρονα ότι η τεχνητή νοημοσύνη απέχει από την ανθρώπινη. Κλείσαμε, με μερικές από τις προσπάθειες που γίνονται σε Ευρωπαϊκό επίπεδο για την αποφυγή αυτών των κινδύνων. Πολλά από τα προαναφερθέντα στοιχεία πιθανότατα να είναι ήδη γνωστά σε έναν έμπειρο αναγνώστη. Εντούτοις, εξυπηρετούν σε μια ομαλή εισαγωγή για τον αρχάριο και σε μια υπενθύμιση για τον έμπειρο αναγνώστη του κόσμου της Τεχνητής Νοημοσύνης. Στην επόμενη υπο-ενότητα αυτού του κεφαλαίου θα παρουσιάσουμε μια ιστορική αναδρομή της τεχνητής νοημοσύνης. Έτσι, ο αναγνώστης θα κατανοήσει σε βάθος την έννοια γύρω από την οποία εκτυλίσσεται η παρούσα διπλωματική προτού εισαχθεί στο συγκεκριμένο τεχνικό της θέμα. Έπειτα, περιγράφεται το κίνητρο που με ωθούσε καθ'όλη τη διάρκεια συγγραφής του έργου. Τέλος, αναφερόμαστε στην τεχνική συνεισφορά της παρούσας εργασίας και στην οργάνωση του τόμου.

## 1.2 Ιστορική Αναδρομή Τεχνητής Νοημοσύνης

Οι εφευρέτες οραματίζονται εδώ και χιλιετίες τη δημιουργία μηχανών που σκέφτονται. Ήδη, γύρω στο 700 π.Χ. αναφέρεται από τον Ησίοδο ο Τάλος: ο μυθικός χάλκινος γίγαντας φτιαγμένος από τον Ήφαιστο με αποστολή να προστατεύσει το νησί της Κρήτης από τους επιδρομείς [16]. Παρόμοια παραδείγματα αποτελούν αυτά της Πανδώρας και της Γαλάτειας. Η μακρόβια επιθυμία για απομίμηση της νοημοσύνης μαρτυρά την αξία που της δίνει ο άνθρωπος. Γεγονός απόλυτα δικαιολογημένο αφού η νοημοσύνη - η νοητική ικανότητα που μας επιτρέπει να σκεπτόμαστε λογικά, να επιλύουμε προβλήματα και να μαθαίνουμε - έχει συμβάλει σημαντικά στην επιβίωση του είδους από τον διαειδικό ανταγωνισμό (interspecific competition)<sup>1</sup>. Σε τελική ανάλυση, ο όρος *homo sapiens* - άνθρωπος ο σοφός - οφείλεται στη σημασία που έχει η νοημοσύνη στη ζωή μας.

Το πρώτο ευρέως αναγνωρισμένο έργο προς την επίτευξη Τεχνητής Νοημοσύνης είναι αυτό της μαθηματικής μοντελοποίησης της λειτουργίας ενός νευρώνα από τους Warren McCulloch και Walter Pitts (1943) [20]. Αναλυτικότερα, βασιζόμενοι στην υπόθεση ότι η κατάσταση λειτουργίας ενός νευρώνα είναι δυαδική («all-or-none») αναπαρέστησαν κάθε νευρώνα ενός δικτύου ως μια πρόταση (proposition) της προτασιακής λογικής (propositional logic). Όπως περιγράφουν, η διέγερση (excitation) του μοντέλου ενός νευρώνα είναι ταυτόσημη με το να είναι η πρόταση του νευρώνα θετική, κάτι που εξαρτάται από την κατάσταση των γειτονικών νευρώνων. Όσο περισσότεροι, διεγερτικά διασυνδεδεμένοι (excitatory connected) προσυναπτικοί νευρώνες (presynaptic neurons) είναι ενεργοποιημένοι τόσο πιο πιθανή είναι η ενεργοποίηση του μετασυναπτικού νευρώνα (postsynaptic neuron). Μεταξύ άλλων, απέδειξαν ότι κάθε συνάρτηση που μπορεί να υπολογιστεί από μια μηχανή Turing μπορεί να υπολογιστεί και από ένα δίκτυο από διασυνδεδεμένους τεχνητούς νευρώνες ύστερα από την κατάλληλη παραμετροποίησή του. Το έργο των Warren McCulloch και Walter Pitts ήταν πρωτοπόρο για την εποχή του αφού έθεσε τις βάσεις όχι μόνο για την σημερινή Τεχνητή Νοημοσύνη αλλά και για την Υπολογιστική Νευροεπιστήμη (Computational Neuroscience). Ωστόσο, οι συγγραφείς δεν παρουσίασαν κανέναν αλγόριθμο για την αλλαγή της τοπολογίας και των παραμέτρων του τεχνητού νευρωνικού δικτύου αν και φαίνεται πως αναγνώριζαν τη σημασία του για τη μάθηση.

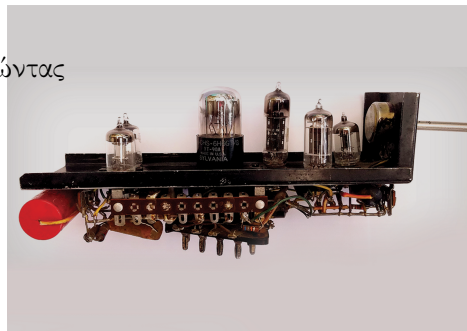
Στο βιβλίο του [21] ο D. Hebb το 1949 επιδίωξε να ενώσει τις αποκλίνουσες θεωρίες της ψυχολογίας και της νευροεπιστήμης θέτοντας κοινές βάσεις για την ερμηνεία της ανθρώπινης συμπεριφοράς. Προηγούμενες θεωρίες απέφευγαν να δώσουν εξήγηση στις διεργασίες του εγκεφάλου ως μεσάζοντα μεταξύ του αισθητηριακού ερεθίσματος (sensory stimuli) και της πιθανής, καθυστερημένης απόκρισης. Αντίθετα, κατέφευγαν στη φιλοσοφία για την ανάλυση των χαρακτηριστικών της ανθρώπινης συμπεριφοράς όπως η προσοχή (attention), το ενδιαφέρον (interest) και η «προσδοκία» (expectancy). Ο Hebb όμως, εργάστηκε στο να αποδείξει ότι η ανθρώπινη συμπεριφορά μπορεί

---

<sup>1</sup>Για την ακρίβεια, ενώ σύμφωνα με τη Δαρβινική θεώρηση η αφηρημένη νοημοσύνη μπορεί να προκύψει άμεσα από τη θεωρία της εξέλιξης των ειδών, νεότερες έρευνες το διαψεύδουν αφού το χαρακτηριστικό της αφηρημένης σκέψης ήταν αχρείαστο στην πραγματιστική παλαιολιθική εποχή. Στην προσπάθειά τους να ερμηνεύσουν την εμφάνιση νοημοσύνης στους ανθρώπους ορίζουν τον όρο «διανοητική βιοθέση» (cognitive niche) για να περιγράψουν όλα τα ζωολογικά ασυνήθιστα χαρακτηριστικά (zoologically unusual traits) που εμφανίζει ο άνθρωπος με τα κυριότερα να είναι η κοινωνικότητα και η λογική αιτίου - αποτελέσματος (cause-and-effect reasoning) [17, 18]. Υποστηρίζουν λοιπόν, ότι η εμφάνιση της διανοητικής βιοθέσης αποτέλεσε τον καταλύτη για την εξέλιξη της ανθρώπινης νοημοσύνης [19]

να γίνει κατανοητή υπό το πρίσμα της φυσιολογίας ("[expectancy] can be a physiologically intelligible process"). Φαινομενικά, ενώ το έργο του απασχολεί μόνο την επιστήμη της Νευροφυσιολογίας (Neurophysiology) συνεισέφερε σημαντικά στο κίνημα του διασυνδετισμού (Connectionism) - το κίνημα μελέτης των διανοητικών διεργασιών με τη χρήση τεχνητών νευρωνικών δικτύων. Για παράδειγμα, η θεώρηση της διαδικασίας μάθησης ως της επαναλαμβανόμενης, ταυτόχρονης πυροδότησης γειτονικών νευρώνων με αποτέλεσμα [την ενδυνάμωση των δεσμών και] τη διαμόρφωση νευρικών συστάδων (cell assemblies) είναι η λογική πίσω από πολλούς αλγόριθμους μάθησης τεχνητών νευρωνικών δικτύων. Επίσης, ένα στοιχείο που θα μας φανεί χρήσιμο στη συνέχεια είναι η παρατήρησή του ότι η κίνηση των ματιών δεν είναι τυχαία αλλά σχετίζεται με τη διαδικασία αντίληψής των θωρούμενων αντικειμένων. Τα παραπάνω δύο έργα έθεσαν το απαραίτητο θεωρητικό υπόβαθρο εμπνέοντας τους ερευνητές στην πειραματική υλοποίηση της Τεχνητής Νοημοσύνης.

Αρκετές ήταν οι απόπειρες δημιουργίας Τεχνητής Νοημοσύνης. Ο πρώτος υπολογιστής τεχνητών νευρωνικών δικτύων ονομάζονταν SNARC (Stochastic Neural Analog Reinforcement Calculator) και κατασκευάστηκε από τους Marvin Minsky και Dean Edmonds το 1950 [22]. Χρησιμοποιώντας 3000 λυχνίες κενού και έναν μηχανισμό αυτόματου πιλότου εξομοίωσαν τη λειτουργία 40 διασυνδεδεμένων νευρώνων. Χρησιμοποιώντας έναν απλό μηχανισμό επιβράβευσης τα «βάρη» του δικτύου - υπο τη μορφή ποτενσιόμετρων - προσαρμόζονταν στο πρόβλημα του λαβυρίνθου στο οποίο δοκιμάστηκε. Ακόμα ένα παράδειγμα τεχνητής νοημοσύνης μπορεί να θεωρηθεί το πρόγραμμα του Christopher Strachey στον υπολογιστή Manchester Mark 1 [23] που αργότερα θεωρήθηκε το πρώτο βιντεοπαιχνίδι. Ήταν ένα παιχνίδι ντάμας που πιθανώς χρησιμοποιούσε κάποιον μη πλήρη (incomplete) αλγόριθμο αναζήτησης στον χώρο των επιτρεπτών ακολουθιών κινήσεων (action sequences). Παρόλα αυτά, η δυνατότητά του να ανταγωνίζεται αποτελεσματικά τον άνθρωπο οδήγησε στη θεώρησή του ως Τεχνητή Νοημοσύνη. Άλλωστε, η σύγχυση για το θέμα ήταν ακόμα μεγαλύτερη την εποχή εκείνη.



Σχήμα 1.1: Ένας από τους 40 νευρώνες του SNARC. Χορηγία της κα. Margaret Minsky [22]

Ο αρχικός ενθουσιασμός για το επιστημονικό πεδίο τράβηξε τα βλέμματα πολλών επιφανών ερευνητών της εποχής. Ο πατέρας της Επιστήμης των Υπολογιστών (Computer Science) και της Τεχνητής Νοημοσύνης, Alan Turing, στην προσπάθειά του να διασαφηνίσει το ερώτημα του «αν οι μηχανές σκέφτονται» επινόησε το επονομαζόμενο Turing Test. Σύμφωνα με τη δημοσίευσή του [24] το 1950, πρόκειται για μια δοκιμασία που εμπλέκει έναν «ανακριτή» ο οποίος διατυπώνει γραπτές ερωτήσεις σε δύο «μάρτυρες»: έναν άνθρωπο και μια μηχανή. Η δοκιμασία θεωρείται επιτυχής όταν ο ανακριτής - χωρίς να έχει οπτική επαφή με τους «μάρτυρες» - δεν μπορεί να ξεχωρίσει τον άνθρωπο από τη μηχανή. Στην ίδια δημοσίευση τόνισε τη σημασία της μάθησης για την ανάπτυξη της Τεχνητής Νοημοσύνης. Υποστήριζε ότι αντί να επιχειρείται η εξονυχιστική συγγραφή ενός προγράμματος που θα μοιάζει με τη σκέψη ενός ώριμου ενήλικα (με αμέτρητες προγραμματισμένες εντολές) είναι προτιμότερη και ταχύτερη η προσομοίωση της νοημοσύνης

ενός παιδιού που μέσα από μηχανισμούς εκπαίδευσης, έμμεσα, αποκτά ώριμη σκέψη. Επίσης, εναπόθεσε τους σπόρους για τους γενετικούς αλγορίθμους, ενώ σε επόμενη δημοσίευσή του [25] μελέτησε τους τρόπους με τους οποίους μια μηχανή με νοημοσύνη θα μπορούσε να λειτουργεί. Ένα ακόμα δημοφιλές όνομα, ο John von Neumann συνεισέφερε στον χώρο αναπτύσσοντας τα «τεχνητά αυτόματα» (artificial automata) [26] ενώ η συμβολή του πιθανώς θα ήταν ακόμα μεγαλύτερη αν προλάβαινε να ολοκληρώσει το βιβλίο του «Ο Υπολογιστής και το Μυαλό» (The Computer and the Brain). Μια τελευταία απόδειξη της προσοχής που έλαβε η Τεχνητή Νοημοσύνη διαφαίνεται στα δέκα μέλη σχετικού σεμιναρίου (workshop) που έλαβε χώρα το καλοκαίρι του 1955 στο Dartmouth College [27]. Ίσως το πιο σημαντικό πόρισμα αυτής της συνάντησης ήταν η ανάπτυξη του Logic Theorist από τους Allen Newell και Herbert Simon, ενός συστήματος για την απόδειξη θεωρημάτων στα μαθηματικά.

Η δεκαετία που ακολούθησε χαρακτηρίζεται από έντονη αισιοδοξία για τις δυνατότητες της Τεχνητής Νοημοσύνης: γενναιόδωρες επενδύσεις σε ερευνητικά προγράμματα ενθάρρυναν τη δημιουργία ποικίλων προγραμμάτων που υποδείκνυαν κάποια μορφή νοημοσύνης. Πιο συγκεκριμένα, αν εξαιρέσουμε την εξέχουσα δουλειά του Arthur Samuel όπου ανέπτυξε ένα παιχνίδι ντάμας χρησιμοποιώντας ενισχυτική μάθηση (Reinforcement Learning), οι περισσότερες προσπάθειες εστίασαν στον χώρο της μίμησης της ανθρώπινης συλλογιστικής (reasoning). Η ιδέα είναι ότι με μια τυπική γλώσσα (formal language) για την αναπαράσταση της γνώσης (knowledge representation) στον υπολογιστή μαζί με την εφαρμογή απλών κανόνων λογικής συμπερασματολογίας (logical inference) σε αυτή καθιστούν δυνατή την εξαγωγή πορισμάτων. Καθοριστικό ρόλο σε αυτή τη «σχολή» είχε η - αναχρονιστική - υπόθεση ότι η νοημοσύνη είναι άρρηκτα συνδεδεμένη με τη δυνατότητα χειρισμού συμβόλων οργανωμένων σε δομές δεδομένων (physical symbol system hypothesis).

Σε αυτήν την κατεύθυνση, δηλαδή της συμβολικής Τεχνητής Νοημοσύνης (Symbolic Artificial Intelligence), εργάστηκαν αρκετοί επιστήμονες της εποχής. Για παράδειγμα, οι δύο ερευνητές πίσω από το Logic Theorist επινόησαν το General Problem Solver. Πρόκειται ουσιαστικά για έναν αλγόριθμο ο οποίος δέχεται σαν είσοδο μια τυποποιημένη περιγραφή του προβλήματος και το επιλύει ακολουθώντας μια στρατηγική ευρετικής αναζήτησης (heuristic search) της λύσης [27]. Στο ίδιο μήκος κύματος εργάστηκε και ο John McCarthy. Εκτός από το ότι ανέπτυξε τη γλώσσα προγραμματισμού Lisp, ειδικά φτιαγμένη για εφαρμογές Τεχνητής Νοημοσύνης, εξέλιξε το πεδίο με το δημοσίευσμά του «Programs with Common Sense» (1958) στο οποίο περιέγραφε το Advice Taker. Αυτό ήταν ένα πρόγραμμα για την επίλυση προβλημάτων μέσω της εφαρμογής «κοινής λογικής» σε προτάσεις διατυπωμένες σε τυπική γλώσσα. Για παράδειγμα, δοθέντος μιας σειράς υποθέσεων σχετικά με το περιβάλλον του προβλήματος διατυπωμένων σε τυπική γλώσσα (π.χ. «Εγώ είμαι στο γραφείο.», «Θέλω να πάω αεροδρόμιο.» κτλ.) ο αλγόριθμος εξήγαγε ένα πλάνο με τα βήματα που πρέπει να ακολουθηθούν για τη μετάβαση στο αεροδρόμιο [28]. Το έργο του συνέχισε ο J. A. Robinson όπου και επινόησε μια πλήρη μέθοδο επίλυσης (complete resolution method) για προβλήματα εκφρασμένα σε λογική πρώτης τάξης. Οι εφαρμογές του ήταν πολλές: από συστήματα μαθηματικού λογισμού (James Slagle's SAINT program [29] και Daniel Bobrow's STUDENT program) μέχρι εφαρμογές ερωταπαντήσεων (Cordell Green's question-answering and planning systems) και ρομποτικής (Shakey Robotics Project). Τέλος, πολλές εφαρμογές της Συμβολικής Τεχνητής Νοημοσύνης αναπτύχθηκαν για το «παιχνίδι» blocks world:

ένα περιβάλλον αποτελούμενο από τουβλάκια που αποσκοπούσε στον πειραματισμό συστημάτων αναπαράστασης γνώσης και συλλογιστικής [30].

Την ίδια εποχή, ειδικά για τον χώρο των νευρωνικών δικτύων υπήρξε σημαντική πρόοδος με τα έργα Perceptron και ADALINE. Το πρώτο συγγράφηκε από τον F. Rosenblatt το 1958 και αποτέλεσε το πρώτο μοντέλο νευρωνικού δικτύου με δυνατότητα επιβλεπόμενης μάθησης supervised learning. Πιο συγκεκριμένα, το Perceptron είναι ένας ταξινομητής γραμμικά διαχωρίσιμων προτύπων με ένα μεμονωμένο τεχνητό νευρώνα του οποίου οι ελεύθεροι παράμετροι - τα προσυναπτικά βάρη (presynaptic weights) και η πόλωση (bias) - προσαρμόζονται στα δεδομένα εισόδου σύμφωνα με έναν αλγόριθμο μάθησης (perceptron rule) [31]. Σε μια εκτενή παρουσίαση του έργου [32], ο Rosenblatt βασίστηκε στη θεωρία του D. Hebb και την επέκτεινε προτείνοντας ένα μοντέλο (το Perceptron) με το οποίο η συμπεριφορά (χαμπύλη εκμάθησης) μπορεί να προβλεφθεί από τη νευροφυσιολογία του συστήματος (τα συναπτικά βάρη). Παρόμοιο ήταν και το έργο ADALINE (Adaptive Linear Neuron) του B. Widrow στο οποίο περιγράφεται και πάλι ένας αλγόριθμος μάθησης για την προσαρμογή των βαρών. Αυτή τη φορά όμως, είναι ο (γνωστός) αλγόριθμος στοχαστικής καθόδου κλίσης stochastic gradient descent που χρησιμοποιείται ακόμα και σήμερα στον αλγόριθμο γραμμικής παλινδρόμησης (linear regression). Μια ακόμα αξιοσημείωτη διαφορά έγκειται στη συνάρτηση ενεργοποίησης όπου ενώ στο πρώτο έργο είναι η βηματική συνάρτηση (step function), στο δεύτερο έργο είναι η γραμμική συνάρτηση (linear activation function - identity function) που καθιστά τον αλγόριθμο κατάλληλο για την πρόβλεψη πραγματικών τιμών [33, 34]. Συνεπώς, ενώ το πρώτο έργο αποτελεί όπως προαναφέρθηκε έναν αλγόριθμο ταξινόμησης, το δεύτερο έργο ανήκει στην κατηγορία αλγορίθμων γραμμικής παλινδρόμησης. Βέβαια, μολονότι και τα δύο έργα είχαν καθοριστική σημασία στην εξέλιξη της Τεχνητής Νοημοσύνης με τη μορφή που τη συναντάμε σήμερα, όπως θα δούμε στη συνέχεια, η έντονη κριτική που ακολούθησε τα επισχίασε για μια ολόκληρη δεκαετία.

Γύρω στο 1970, το επιστημονικό πεδίο της Τεχνητής Νοημοσύνης διήλθε μια εποχή «χειμώνα» (AI winter). Η χρηματοδότηση ερευνητικών προγραμμάτων πάγωσε και έτσι το ενδιαφέρον στράφηκε αλλού. Κοιτώντας πίσω, είναι εύκολο να εντοπίσει κανείς τα αίτια αυτού του «χειμώνα». Καταρχάς, ένας λόγος ήταν (και είναι) το ελλειπές επιστημονικό υπόβαθρο σε ό,τι αφορά την ανθρώπινη νοημοσύνη [35]. Αναμενόμενο, αφού για μια επιτυχημένη μίμηση της ανθρώπινης νοημοσύνης απαιτείται πρώτα η κατανόησή της. Βέβαια, ίσως ο σημαντικότερος λόγος ήταν η απογοήτευση που προκλήθηκε όταν φιλόδοξες υποσχέσεις για τις δυνατότητες της Τεχνητής Νοημοσύνης στο εγγύς μέλλον δεν μπόρεσαν να ικανοποιηθούν. Όπως αποδείχθηκε, η μετάβαση της Τεχνητής Νοημοσύνης από εφαρμογές παιδικών κόσμων όπως το blocks world σε πραγματικά προβλήματα δεν ήταν απλώς ζήτημα γραμμικής αύξησης της υπολογιστικής δύναμης. Για παράδειγμα, στην περίπτωση της συμβολικής Τεχνητής Νοημοσύνης, με την ωρίμανσή της θεωρίας πολυπλοκότητας (computational complexity αναδείχθηκε το θέμα της συνδυαστικής έκρηξης (combinatorial explosion) αποκαλύπτοντας έτσι τη δυσεπίλυτη (intractable) φύση πολλών προβλημάτων του αληθινού κόσμου. Αντίστοιχα εμπόδια έκαναν την εμφάνισή τους στον χώρο των Νευρωνικών Δικτύων. Το σημαντικότερο ήταν αυτό της αδυναμίας ενός μεμονωμένου Perceptron με δύο εισόδους να αναπαραστήσει πολλές συναρτήσεις όπως τη συνάρτηση XOR [27] - περιγράφηκε από το βιβλίο Perceptrons των M. Minsky και S. Papert το 1969. Συνολικά, αν και ορισμένα από τα ανωτέρω αίτια είναι ακόμα σε ισχύ, το ενδιαφέρον σύντομα αναζωπυρώθηκε.

Παρά τις ανωτέρω αδυναμίες της Τεχνητής Νοημοσύνης την εποχή εκείνη, αυτό δεν εμπόδισε την αξιοποίησή της σε εξειδικευμένες εφαρμογές. Πιο συγκεκριμένα, τις δεκαετίες του 1970 και 1980 αναπτύχθηκαν τα «έμπειρα συστήματα» (expert systems). Πρόκειται για προγράμματα που εφαρμόζουν κανόνες συλλογιστικής σε εξειδικευμένη βάση γνώσης (domain-specific knowledge base) μιμούμενοι τη διαδικασία λήψης αποφάσεων ενός εμπειρογνώμονα. Η εξειδικευμένη βάση γνώσης περιόριζε σημαντικά τον χώρο αναζήτησης λύσεων (search space) έτσι ώστε η συνδυαστική έκρηξη να μην αποτελεί πρόβλημα. Αυτό, επέτρεψε να αναπτυχθούν πολλά έμπειρα συστήματα για εμπορική χρήση - κυρίως στον χώρο της υγείας - αποδεικνύοντας για πρώτη φορά έμπρακτα τα οφέλη της Τεχνητής Νοημοσύνης. Ενδεικτικά, δύο δημοφιλή παραδείγματα είναι το MYCIN και το INTERNIST. Το MYCIN αποσκοπούσε στη διάγνωση βακτηριακών μολύνσεων μέσω ενός αλγορίθμου συλλογιστικής που μοντελοποιούσε την αβεβαιότητα των λογικών υποθέσεων και συμπερασμάτων. Το INTERNIST από την άλλη βοηθούσε στη διάγνωση ασθενειών μετά από την περιγραφή των εκδηλούμενων συμπτωμάτων [36]. Αν και τα έμπειρα συστήματα ανανέωσαν το ενδιαφέρον για την Τεχνητή Νοημοσύνη, αυτό δε διήρκεσε πολύ λόγω προβλημάτων που εμφάνιζαν με το κυριότερο να είναι η έλλειψη «κοινής λογικής» [37].

Η έρευνα στον χώρο της Τεχνητής Νοημοσύνης αποκαταστάθηκε στα συνήθη υψηλά επίπεδα σε σύντομο χρονικό διάστημα. Αυτό μπορεί σε μεγάλο βαθμό να αποδοθεί σε ένα μεμονωμένο έργο· το Parallel Distributed Processing που συγγράφηκε από τους David E. Rumelhart et al. και δημοσιεύτηκε το 1986 [38]. Οι συγγραφείς, μεταξύ των οποίων και ο Geoff Hinton εμπνεόμενοι από τα παλαιότερα έργα πάνω στη γνωστική νευροεπιστήμη (cognitive neuroscience) έστρεψαν την έρευνα του χώρου από πειραματικές «αλχημείες» (π.χ. αυτή της συμβολικής λογικής) σε μια πιο επίσημη, φορμαλιστική διαδικασία βασιζόμενη λιγότερο στη φιλοσοφία και περισσότερο στις θετικές επιστήμες<sup>2</sup>. Με αυτόν τον τρόπο, η σχολαστική συγγραφή αναρίθμητων κανόνων προτασιακής λογικής για τη δημιουργία βάσεων γνώσης εγκαταλείφθηκε, μαζί της και η θεωρία της συμβολικής Τεχνητής Νοημοσύνης. Άλλωστε, η τεχνολογία των έμπειρων συστημάτων είχε παρακμάσει αφού όπως φάνηκε από την απουσία «κοινής λογικής» σε αυτά, ήταν εξαιρετικά περιοριστική η χρήση προτασιακής λογικής για την περιγραφή του πραγματικού, αβέβαιου κόσμου [27, 39].

Τη δεκαετία του 1980 τη θέση της συμβολικής Τεχνητής Νοημοσύνης πήρε το κίνημα του κονεκτιβισμού (connectionist movement). Όπως θα δούμε, αυτό συνέβαλλε καθοριστικά στη διαμόρφωση του σημερινού κλάδου των νευρωνικών δικτύων. Τυπικά, ο κονεκτιβισμός είναι το κίνημα της γνωστικής επιστήμης (cognitive science) που επιχειρεί να εξηγήσει τις διανοητικές διεργασίες με τη χρήση ενός δικτύου με βάρη (weighted network) που διασυνδέει απλές μονάδες επεξεργασίας [40]. Σε αυτήν τη θεωρία καταπιάστηκαν και οι συγγραφείς του έργου Parallel Distributed Processing [38] οι οποίοι εδραίωσαν τις ιδέες που σήμερα θεμελιώνουν τη θεωρία των νευρωνικών δικτύων. Η πρώτη σημαντική ιδέα που περιγράφεται λεπτομερώς στο έργο είναι η κατανεμημένη αναπαράσταση (distributed representation) σύμφωνα με την οποία κάθε είσοδος στο σύστημα αναπαρίσταται από πολλά χαρακτηριστικά κατανεμημένα στο δίκτυο και ανάποδα, δηλαδή κάθε μεμονωμένο χαρακτηριστικό μπορεί να αποτελεί μέρος της περιγραφής πολλών, ετερογενών εισόδων. Μια ακόμα σημαντική ιδέα είναι αυτή της μηχανικής μάθησης (machine learning) με την οποία η επίδοση ενός συστήματος βελτιώνεται (μαθαίνει) από την εμπειρία.

---

<sup>2</sup>Στη βιβλιογραφία αυτό το γεγονός αναφέρεται σαν «η νίκη των καθάρων» (victory of the neats) [27].

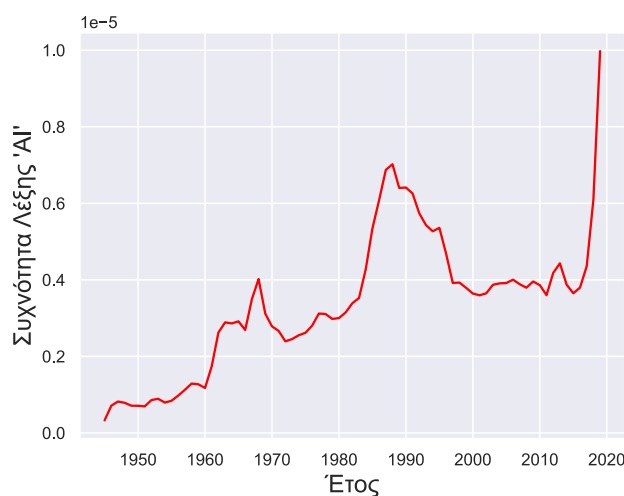
Για τον σκοπό αυτό, παρουσιάζουν έναν επαναστατικό αλγόριθμο ο οποίος αυτοματοποιεί τη διαδικασία μηχανικής μάθησης στα νευρωνικά δίκτυα. Πρόκειται για τον αλγόριθμο ανάστροφης διάδοσης σφάλματος (back propagation) ο οποίος, παραδόξως, ενώ είχε αναπτυχθεί περίπου το 1960 γνώρισε ευρεία χρήση από το 1980 και μετά. Συνεπώς, η σημασία του κονεκτιβισμού είναι καθοριστική αφού αναβίωσε τις ιδέες της γνωστικής επιστήμης επανατοποθετώντας κατά αυτόν τον τρόπο τον κλάδο της Τεχνητής Νοημοσύνης σε μια πιο επιστημονική τροχιά.

Η στροφή σε μια πιο επιστημονική προσέγγιση του κλάδου της Τεχνητής Νοημοσύνης το 1980 δε συνέβαλλε μόνο στην ανάπτυξη του κλάδου των νευρωνικών δικτύων. Για παράδειγμα, ο κλάδος της επεξεργασίας φυσικής γλώσσας επωφελήθηκε σημαντικά από την επιτυχημένη μοντελοποίηση ακολουθιών με τη χρήση κρυφών Μαρκοβιανών μοντέλων (hidden Markov models) και αργότερα, το 1997 με μονάδες μακράς-βραχέας μνήμης (Long-Short Term Memory block - LSTM). Επίσης, ο χώρος της όρασης υπολογιστών επωφελήθηκε από τη σύγκλιση της Τεχνητής Νοημοσύνης με τις θετικές επιστήμες. Την πρόοδο μαρτυρά η εμφάνιση των πρώτων εφαρμογών οπτικής αναγνώρισης χαρακτήρων (optical character recognition) τη δεκαετία του 1980 και ύστερα, των τυποποιημένων βάσεων δεδομένων για ανάπτυξη και μέτρηση απόδοσης οπτικών συστημάτων αναγνώρισης μοτίβων (π.χ. MNIST). Επίσης, σημαντικά αναπτύχθηκε ο χώρος της ταξινόμησης προτύπων με τη δημιουργία ή βελτίωση αρκετών μοντέλων όπως οι μηχανές διανυσματικής υποστήριξης με μέθοδο πυρήνα (Support Vector Machines with kernel trick) και τα δίκτυα ακτινικής βάσης (Radial Basis Networks). Ακόμα και ο χώρος της συλλογιστικής για τον οποίο κάναμε λόγο σε προηγούμενες παραγράφους εμπλουτίστηκε με μια σχολαστική και αποδοτική - αυτή τη φορά - μοντελοποίηση αβεβαιότητας της γνώσης μέσω της ανάπτυξης Μπεϋζιανών δικτύων (Bayesian networks). Τέλος, ο χώρος της στατιστικής γνώρισε πρόοδο αφού στις παραδοσιακές τεχνικές συμπερασματολογίας προστέθηκαν αυτές βασιζόμενες σε μηχανική μάθηση.

Προς τα τέλη της δεκαετίας του 1990 και τις αρχές του 2000 η έρευνα είχε εστιάσει σε κλάδους της Τεχνητής Νοημοσύνης που δε σχετίζονταν με τα νευρωνικά δίκτυα. Η κατάσταση αυτή όμως σύντομα αναστράφηκε. Αρχικά, η μεγάλη υπολογιστική ισχύ που απαιτούσαν αλγόριθμοι εκπαίδευσης βαθιών νευρωνικών δικτύων δε διευκόλυναν τον πειραματισμό [33]. Παρόλα αυτά, χάρη σε τρεις ερευνητές (Geoffrey Hinton, Yann LeCun και Yoshua Bengio) που χρηματοδοτούνταν από το Καναδικό Ινστιτούτο για προηγμένη έρευνα (CIFAR) η ενασχόληση με τα νευρωνικά δίκτυα κρατήθηκε ζωντανή οδηγώντας τελικά σε αξιοσημείωτη πρόοδο. Το πρώτο ορόσημο των βαθιών νευρωνικών δικτύων ήταν το 2006 όπου οι Hinton et al. [41] απέδειξαν ότι ένα είδος βαθύς νευρωνικού δικτύου - τα βαθιά δίκτυα πίστης (deep belief networks) - μπορούν να εκπαιδευτούν αποδοτικά και γρήγορα μέσω ενός άπληστου (greedy) αλγορίθμου. Το δεύτερο ορόσημο που απέδειξε τις προοπτικές των βαθιών νευρωνικών δικτύων ήταν η επιτυχία αυτής της τεχνολογίας σε διαγωνισμούς ταξινόμησης εικόνων της βάσης ImageNet το 2010 και το 2012. Στη δημοσίευσή τους ImageNet Classification with Deep Convolutional Neural Networks, οι A. Krizhevsky et al. [42] περιγράφουν μια νέα αρχιτεκτονική νευρωνικών δικτύων (βαθιά συνελικτικά δίκτυα) αλλά και πρωτοπόρες μεθόδους για αποδοτική εκπαίδευση (dropout, ReLU activation function κλπ.). Έκτοτε, το ενδιαφέρον απογειώθηκε και σε συνδυασμό με την αυξημένη διαθεσιμότητα δεδομένων, εξειδικευμένων συσκευών για παράλληλη επεξεργασία και αποδοτικών αλγορίθμων εκπαίδευσης δημιούργησαν το πιο πρόσφορο έδαφος για την άνθιση της Τεχνητής Νοημοσύνης.

Σήμερα, η εξέλιξη της Τεχνητής Νοημοσύνης και δη των νευρωνικών δικτύων είναι ραγδαία. Για παράδειγμα, με τη χρήση μοντέλων νευρωνικών δικτύων βασισμένων στον μηχανισμό προσοχής (attention-based neural networks) όπως τα λεγόμενα transformers, σημειώθηκε αξιοσημείωτη πρόοδος τόσο στον κλάδο της επεξεργασίας φυσικής γλώσσας (natural language processing) (βλέπε GPT -3) όσο και στον χώρο της όρασης υπολογιστών (βλέπε Vision Transformer και CoAtNet). Αναλυτικότερα για το Generative Pre-trained Transformer 3 - GPT 3, πρόκειται για το γλωσσικό μοντέλο που μπορεί να δημιουργήσει κείμενο σε φυσική γλώσσα ακόμα και να διατηρήσει για εύλογο χρόνο συζήτηση με έναν άνθρωπο. Ακόμα, σημαντική πρόοδος παρατηρείται στη μέθοδο μάθησης με αυτο-επίβλεψη (self-supervision) επιτρέποντας έτσι την εκπαίδευση δικτύων χωρίς να απαιτείται η σχολαστική και χρονοβόρα ανάθεση ετικετών στα δεδομένα εκπαίδευσης. Τέλος, μεταξύ άλλων, σπουδαία εξέλιξη υπήρξε πρόσφατα στις εφαρμογές όπου δοθέντων μερικών εικόνων από ένα αντικείμενο συνθέτουν εικόνες αυτού από νέες γωνίες θέασης (Novel View Synthesis). Ενδεικτικά, πρωτοπόρα έργα στον χώρο αυτό είναι το NeRF και το GIRAFFE [43].

Κλείνοντας το ιστορικό αυτό σημείωμα, δεν μπορώ παρά να αντικρίσω με δέος το μέλλον που επιφέρει η τεχνολογία της Τεχνητής Νοημοσύνης. Ανύπαρκτη πριν έναν αιώνα, σήμερα είναι μέρος της καθημερινότητά μας με τεράστια ορμή που δε φαίνεται να κατευνάζει. Αν μέσα σε μερικές δεκαετίες έχει τόσες δυνατότητες, στο εγγύς μέλλον η δύναμή της θα είναι τεράστια. Δύναμη που θα δημιουργήσει μια επίγεια ουτοπία ή θα αποτελέσει ένα ακόμα τουβλάκι στην οικοδόμηση της κοινωνίας του ρίσκου <sup>3</sup>(;)... ο χρόνος θα δείξει.



Σχήμα 1.2: Γραφική παράσταση της συχνότητας του όρου AI σε βιβλία γραμμένα στην αγγλική γλώσσα ανά έτος (από 1945 μέχρι και 2019). Είναι εμφανείς οι τρεις περίοδοι ακμής του κλάδου. Παράχθηκε από το Google Ngram Viewer.

### 1.3 Κίνητρο

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam

<sup>3</sup>Ο όρος «κοινωνία του ρίσκου» (risk society) είναι δανεισμένος από το ομώνυμο βιβλίο του Ulrich Beck [44] όπου περιγράφεται το χαρακτηριστικό των μοντέρνων κοινωνιών να οργανώνονται γύρω από νέες μορφές, απαρατήρητου ρίσκου (όπως αυτό της κλιματικής αλλαγής).



nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

#### 1.4 Συνεισφορά Εργασίας

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent

blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

## 1.5 Οργάνωση του Τόμου

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

## Κεφάλαιο 2

# Θεωρητικό Υπόβαθρο

Στο παρόν κεφάλαιο θα οικοδομήσουμε την απαραίτητη γνώση στην οποία βασίζεται η έρευνα των επόμενων ενότητων. Αρχικά, θα παρουσιαστούν συνοπτικά τα τεχνητά νευρωνικά δίκτυα <sup>1</sup> υπό μια μαθηματική σκοπιά. Έπειτα, θα αναλυθούν τα νευρωνικά δίκτυα με κάψουλες (capsule networks) τα οποία και αποτελούν το κύριο θέμα της εργασίας. Τέλος, θα γίνει αναφορά σε νέες τεχνικές και αλγόριθμους που χρησιμοποιήθηκαν στο παρόν έργο ώστε η μετέπειτα εισαγωγή των μεθόδων μας για την εξέλιξη των νευρωνικών δικτύων με κάψουλες να είναι περισσότερο ομαλή και κατανοητή.

### 2.1 Τεχνητά Νευρωνικά Δίκτυα

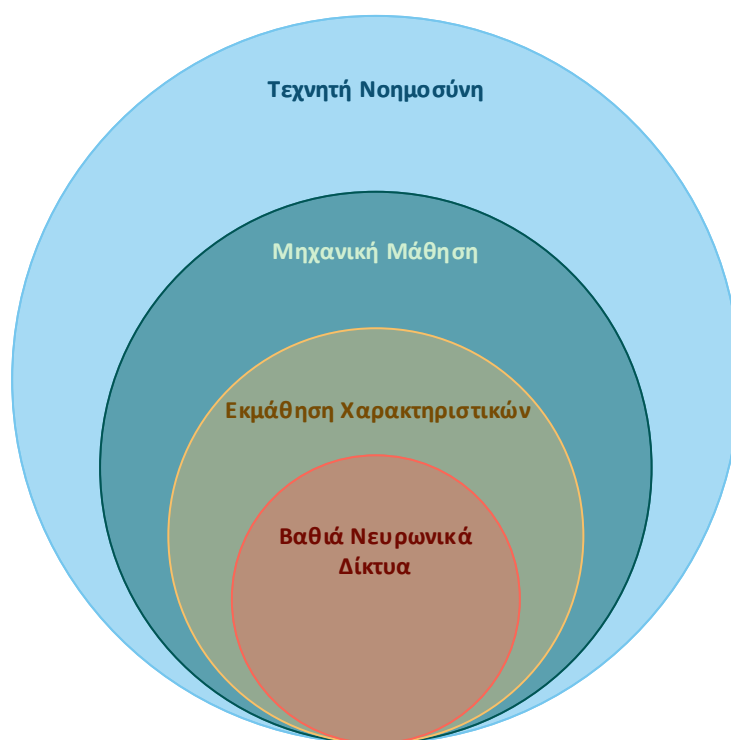
Τα σημερινά τεχνητά νευρωνικά δίκτυα, όπως είναι αναμενόμενο, απέχουν σημαντικά από το πρώτο μοντέλο των Warren McCulloch και Walter Pitts [20] που συζητήσαμε στην ενότητα 1.2. Με την ωρίμανση της τεχνολογίας, αυτή ανεξαρτητοποιήθηκε από την (υπολογιστική) νευροεπιστήμη και εντάχθηκε στην Τεχνητή Νοημοσύνη υπό μια ιεραρχική δομή. Κρίνεται λοιπόν σκόπιμο να παρουσιάσουμε αυτήν την ιεραρχική δομή οργάνωσης της Τεχνητής Νοημοσύνης και μετέπειτα να αναφερθούμε στα επιμέρους στοιχεία της.

Όπως βλέπουμε στο σχήμα 2.1 τα νευρωνικά δίκτυα πολλών επιπέδων (βαθιά νευρωνικά δίκτυα) είναι ένα μέρος του κλάδου της εκμάθησης χαρακτηριστικών (feature learning ή representation learning) που είναι ένα μέρος της μηχανικής μάθησης η οποία με τη σειρά της ανήκει στο ευρύτερο επιστημονικό πεδίο της τεχνητής νοημοσύνης. Φυσικά, η τεχνητή νοημοσύνη περιλαμβάνει αρκετούς άλλους κλάδους εκτός από αυτόν της μηχανικής μάθησης<sup>2</sup>. Μια χρήσιμη παρατήρηση είναι ότι οι σχέσεις υποσύνολου συμπίπτουν με τη χρονική αλληλουχία ανάπτυξης του κάθε κλάδου. Δηλαδή, κάθε υποσύνολο αναπτύχθηκε ταυτόχρονα ή αργότερα από το οποιοδήποτε υπερσύνολό του.

Στη συνέχεια, θα γίνει λόγος για τα στοιχεία εκείνα που περιλαμβάνουν την τεχνολογία των βαθιών νευρωνικών δικτύων προκειμένου ο αναγνώστης να αποκτήσει μια εποπτικότερη εικόνα.

<sup>1</sup> Από εδώ και στο εξής, με τον όρο «νευρωνικά δίκτυα» θα αναφερόμαστε στα «τεχνητά νευρωνικά δίκτυα».

<sup>2</sup> Βέβαια, ο κλάδος της μηχανικής μάθησης είναι σήμερα ο γρηγορότερος αναπτυσσόμενος.



Σχήμα 2.1: Διάγραμμα Venn όπου απεικονίζει τη θέση των νευρωνικών δικτύων στην οργάνωση της τεχνητής νοημοσύνης. Παράχθηκε από το Microsoft Visio™.

### 2.1.1 Μηχανική Μάθηση

Όπως προδίδει ο όρος, σε αδρές γραμμές τα συστήματα μηχανικής μάθησης έχουν τη δυνατότητα να μαθαίνουν μια εργασία χωρίς να έχουν προγραμματιστεί με ρητές εντολές για τη συγκεκριμένη εργασία αυτή<sup>3</sup>. Ίσως, ο πιο πλήρης ορισμός δίνεται από τον Tom M. Mitchell [45] σύμφωνα με τον οποίο, ένα υπολογιστικό πρόγραμμα λέγεται ότι μαθαίνει από μια εμπειρία  $E$ , ως προς ένα σύνολο εργασιών  $T$  και ένα μέτρο απόδοσης  $P$ , εάν η απόδοσή του σε εργασίες του  $T$ , όπως αυτή μετριέται από το  $P$ , βελτιώνεται με την  $E$ .<sup>4</sup>

Σύμφωνα με τον ανωτέρω ορισμό διακρίνουμε τρία βασικά συστατικά ενός συστήματος μηχανικής μάθησης. Αυτά είναι τα παρακάτω:

**Εργασία -  $T$**  Είναι το πρόβλημα το οποίο επιθυμούμε να λύσουμε.

**Μέτρο Απόδοσης -  $P$**  Αποτελεί μια μετρική του στόχου ως ένδειξη ποιότητας της λύσης μας. Από μαθηματική σκοπιά, είναι αυτό που ο αλγόριθμος μάθησης βελτιστοποιεί.

**Εμπειρία -  $E$**  Πρόκειται για τα δεδομένα εισόδου που λαμβάνει το σύστημα υπό τη μορφή

<sup>3</sup>Η δυνατότητα αυτή είναι πολύ σημαντική αφού, όπως διαπιστώσαμε στην ενότητα 1.2 όταν έγινε λόγος για τα έμπειρα συστήματα, για πολλές εργασίες είναι πρακτικός αδύνατο να περιγραφούν ρητά και ντετερμινιστικά οι λύσεις τους.

<sup>4</sup>Ο ορισμός αυτός εξηγεί γιατί για παράδειγμα η λήψη μιας ιστοσελίδας της βικιπédieας και η αποθήκευσή της τοπικά στον υπολογιστή δεν αποτελεί μηχανική μάθηση. Όπως προκύπτει, η «γνώση» αυτή δεν καθιστά καλύτερο τον υπολογιστή σε κάποια εργασία [46].

παραδειγμάτων ή ως ερεθίσματα ανάδρασης από το περιβάλλον. Όπως θα δούμε στη συνέχεια, ο τρόπος απόκτησης αυτών των δεδομένων αλλά και η φύση τους καθορίζει το είδος της μάθησης.

### Βασικά Είδη Συστημάτων Μηχανικής Μάθησης

Τα είδη των συστημάτων μηχανικής μάθησης μπορούν να ταξινομηθούν ανάλογα με το:

- *Αν εκπαιδεύονται με ανθρώπινη επίβλεψη.*  
Ανάλογα με αυτό το κριτήριο έχουμε τις εξής βασικές κατηγορίες: επιβλεπόμενη (supervised), μη-επιβλεπόμενη (un-supervised) και ενισχυτική μάθηση (reinforcement learning).
- *Αν μαθαίνουν σταδιακά (incrementally) και «στον αέρα» (on the fly).*  
Σε αυτήν την περίπτωση χωρίζουμε τα συστήματα μηχανικής μάθησης σε αυτά που πραγματοποιούν μάθηση σε ζωντανό χρόνο (online learning) και σε αυτά που μαθαίνουν κατά δέσμες (batch learning).
- *Αν κατασκευάζουν μοντέλα προσαρμοσμένα στα δεδομένα.*  
Με αυτό το κριτήριο χωρίζονται σε συστήματα βασισμένα σε μοντέλο (model-based) ή σε συστήματα βασισμένα σε παραδείγματα (instance-based). [46]

Προφανώς, κάθε δυνατός συνδυασμός των παραπάνω κριτηρίων είναι αποδεκτός, οδηγώντας έτσι στην ταξινόμηση των συστημάτων μηχανικής μάθησης σε μια πληθώρα από διαφορετικές κατηγορίες. Κρίνεται χρήσιμο, να αναφέρουμε σε όλη την έκταση του έργου τις κατηγορίες στις οποίες ανήκει το κάθε σύστημα που παρουσιάζουμε. Για αυτόν τον λόγο, παροτρύνουμε τον αναγνώστη που δεν είναι εξοικειωμένος με τους ανωτέρω όρους να διαβάσει τους αντίστοιχους ορισμούς στο παράρτημα Α'.

#### 2.1.2 Εκμάθηση Χαρακτηριστικών

Η ανάπτυξη των πρώτων συστημάτων μηχανικής μάθησης απεμπόλησε την ανάγκη των ευφυών εφαρμογών για σχολαστική και ρητή (hard-coded) αναπαράσταση του χώρου του προβλήματος (π.χ. με την χρήση προτασιακής λογικής). Με τα νέα συστήματα, η γνώση για το πρόβλημα μαθαίνονταν αυτοματοποιημένα μέσω αλγορίθμων μάθησης από το σύνολο δεδομένων εκπαίδευσης. Με άλλα λόγια, τα αλγοριθμικά κατασκευάσματα μάθαιναν να αντιστοιχίζουν με αυτοματοποιημένο τρόπο τα δεδομένα εισόδου (κωδικοποιημένα σε μια μορφή αναπαράστασης) σε τιμές εξόδου.

Παρόλα αυτά, τα πρώτα, απλά συστήματα μηχανικής μάθησης δεν έλυσαν όλα τα προβλήματα. Όπως είναι εμφανές από την ανωτέρω περιγραφή, αν και δεν απαιτούνταν η λεπτομερής συγγραφή βάσεων γνώσης, παρέμενε η ανάγκη για αναπαράσταση των δεδομένων εισόδου με μια αποδοτική μορφή. Είναι γεγονός, άλλωστε, ότι η αναπαράσταση σε πολλά συστήματα επηρεάζει καθοριστικά την απόδοση του συστήματος<sup>5</sup>. Για αυτόν τον λόγο, εξελίχθηκαν διαδικασίες «μηχανικής χαρακτηριστικών» (feature engineering) όπου αξιοποιώντας την τεχνική γνώση του χώρου του προβλήματος (domain knowledge) στόχος είναι η αναπαράσταση των ακατέργαστων δεδομένων εκπαίδευσης ως σύνολο

---

<sup>5</sup>Η σημασία της αναπαράστασης δεδομένων στην απόδοση των αλγοριθμικών κατασκευασμάτων δε θα πρέπει να μας εκπλήσσει αφού κάτι αντίστοιχο ισχύει και στους ανθρώπους. Για παράδειγμα, οι περισσότεροι είναι πολύ πιο αποδοτικοί στην αριθμητική χρησιμοποιώντας την αραβική αναπαράσταση αριθμών απ' ό,τι τη λατινική [33].

(συνήθως διάνυσμα) από κατάλληλα χαρακτηριστικά. Η καταλληλότητα έγκειται στο πόσο χρήσιμη πληροφορία παρέχουν τα χαρακτηριστικά υπό τον περιορισμό να είναι όσο το δυνατόν περισσότερο ανεξάρτητα μεταξύ τους ώστε να αποπλέκουν (disentangle) τους παράγοντες διακύμανσης (factors of variation) των δεδομένων που επηρεάζουν την τιμή εξόδου [33].

Οι ανωτέρω έννοιες μπορούν να καταστούν περισσότερο κατανοητές με ένα παράδειγμα συστήματος εκτίμησης τιμών κατοικιών [46] (πρόβλημα παλινδρόμησης, επίλυση με επιβλεπόμενη μάθηση κατά δέσμες). Πιο συγκεκριμένα, δοθέντος ενός συνόλου ακατέργαστων δεδομένων που αφορούν την αγορά σπιτιών σε μια περιοχή, το σύστημα, μέσω μηχανικής μάθησης, θα είναι ικανό να εκτιμήσει την τιμή με την οποία μια κατοικία θα πρέπει να κοστολογηθεί για να βγει στην αγορά. Όπως εξηγήσαμε, προτού τροφοδοτήσουμε το σύστημα με το σύνολο δεδομένων, είναι σκόπιμο να εφαρμόσουμε διαδικασίες μηχανικής χαρακτηριστικών σε αυτά και να δημιουργήσουμε μια νέα αναπαράσταση. Τα ακατέργαστα δεδομένα εκπαίδευσης αποτελούνται από μια λίστα όπου κάθε γραμμή αντιστοιχεί σε μια οικία με όλες τις προδιαγραφές της και την τιμή πώλησής της. Στο πρόβλημα του παραδείγματος:

- Ένας παράγοντας διακύμανσης θα μπορούσε να είναι η ακρίβεια της συγκεκριμένης περιοχής. Εντούτοις, σαν προδιαγραφές ας υποθέσουμε ότι αναφέρονται μόνο το γεωγραφικό πλάτος και γεωγραφικό μήκος με αποτέλεσμα η ακρίβεια της περιοχής να μην είναι άμεσα παρατηρήσιμη (συνηθισμένο φαινόμενο στους παράγοντες διακύμανσης). Θα μπορούσαμε λοιπόν να μετατρέψουμε τις συντεταγμένες σε ένα νέο χαρακτηριστικό: την «κλάση» της περιοχής. Ένας ακόμα παράγοντας διακύμανσης που είναι όμως άμεσα παρατηρήσιμος είναι το εμβαδόν επιφάνειας της κατοικίας.
- Μη χρήσιμη πληροφορία θα μπορούσε να είναι ο προσανατολισμός της οικίας. Σε αυτή την περίπτωση, η δημιουργία μιας νέας αναπαράστασης δεδομένων χωρίς το παρόν προσδιορισμό θα βοηθούσε την επίδοση του συστήματος.
- Δύο αλληλοεξαρτώμενα χαρακτηριστικά (με υψηλή συν—διακύμανση) θα μπορούσαν να είναι ο αριθμός των υπνοδωματίων και ο αριθμός των μπάνιων όπου τότε η επιλογή της συγχώνευσής τους πιθανότατα βελτίωνε την απόδοση.



Σχήμα 2.2: Παράδειγμα εξαγωγής χαρακτηριστικών σε εικόνα ενός αυτοκινήτου με τη μέθοδο του Ιστογράμματος Προσανατολισμένων Κλίσεων (Histogram of Oriented Gradients). Παράχθηκε τοπικά. ΤΟΔΟ

Αν και στο παραπάνω πρόβλημα ήταν σχετικά εύκολη η «χειρονακτική» εξαγωγή χαρακτηριστικών, υπάρχουν πολλοί χώροι προβλημάτων όπου κάτι τέτοιο είναι από πολύ απαιτητικό έως απίθανο. Ενδεικτικά, σε ένα πρόβλημα οπτικής αναγνώρισης ζώων και αντικειμένων (όπως αυτό του CIFAR-10 [47]) είναι εξαιρετικά δύσκολη η περιγραφή χαρακτηριστικών που θα λαμβάνουν μια αναπαράσταση σε εικονοστοιχεία (pixel) και θα παράγουν μια χρήσιμη αναπαράσταση. Μολονότι υπάρχουν γενικευμένες μέθοδοι για την εξαγωγή χαρακτηριστικών σε εικόνες όπως αυτή του σχήματος 2.2, αυτές δεν είναι βέλτιστα προσαρμοσμένες στον χώρο του εκάστοτε προβλήματος. Απόρροια αυτού μεταξύ άλλων είναι η απόρριψη στοιχείων των ακατέργαστων δεδομένων που μπορεί να είναι χρήσιμα (π.χ. σε ένα πρόβλημα αναγνώρισης μάρκας και χρώματος αυτοκινήτου, η μέθοδος Ιστογράμματος Προσανατολισμένων Κλίσεων θα απέρριπτε το χρώμα, στοιχείο χρήσιμο για τον χώρο του προβλήματος).

Η λύση για την αντιμετώπιση των προβλημάτων της χειρονακτικής εξαγωγής χαρακτηριστικών είναι η χρήση των αλγορίθμων μηχανικής μάθησης για την εκμάθηση όχι μόνο για της αντιστοίχισης των δεδομένων εκπαίδευσης στην επιθυμητή έξοδο αλλά και των ίδιων των αναπαραστάσεων των δεδομένων. Αν και συνήθως, οι προκύπτουσες αναπαραστάσεις μετά τον μετασχηματισμό των ακατέργαστων δεδομένων δεν είναι κατανοητές από τον άνθρωπο, εφόσον η εκπαίδευση γίνει επιτυχημένα, αποτελούνται από χρήσιμα χαρακτηριστικά (που κωδικοποιούν τους παράγοντες διακύμανσης). Χαρακτηριστικό παράδειγμα συστήματος για την εκμάθηση χαρακτηριστικών είναι ο Αυτοκωδικοποιητής (Autoencoder).

### 2.1.3 Πολυεπίπεδα Νευρωνικά Δίκτυα

Η εκμάθηση χαρακτηριστικών σε συνδυασμό με τις ιδέες του κονεκτιβισμού περί κατανεμημένης αναπαράστασης (βλ. ενότητα 1.2) μας οδηγεί αναπόφευκτα στη βαθιά μάθηση (deep learning). Υπό μια αφαιρετική σκοπιά, πρόκειται για τα λεγόμενα «πολυεπίπεδα νευρωνικά δίκτυα» τα οποία συνδυάζουν τόσο τον μετασχηματισμό της αναπαράστασης των δεδομένων εισόδου όσο και την αντιστοίχιση αυτών των νέων αναπαραστάσεων στην τιμή εξόδου. Τα συστήματα αυτά, όπως θα δούμε στη συνέχεια, είναι δομημένα από απλές υπολογιστικές μονάδες που τους επιτρέπουν να δημιουργούν σύνθετες αναπαραστάσεις μέσω μιας σειράς από εμφωλευμένες, απλούστερες αναπαραστάσεις. Σημειώνουμε ότι πραγματοποιούν μάθηση με κατασκευή μοντέλου (model-based learning systems) και χρησιμοποιούνται τόσο σε προβλήματα ταξινόμησης όσο και παλινδρόμησης. Στις επόμενες παραγράφους, θα περιγράψουμε με μεγαλύτερη λεπτομέρεια τον χώρο των νευρωνικών δικτύων<sup>6</sup>.

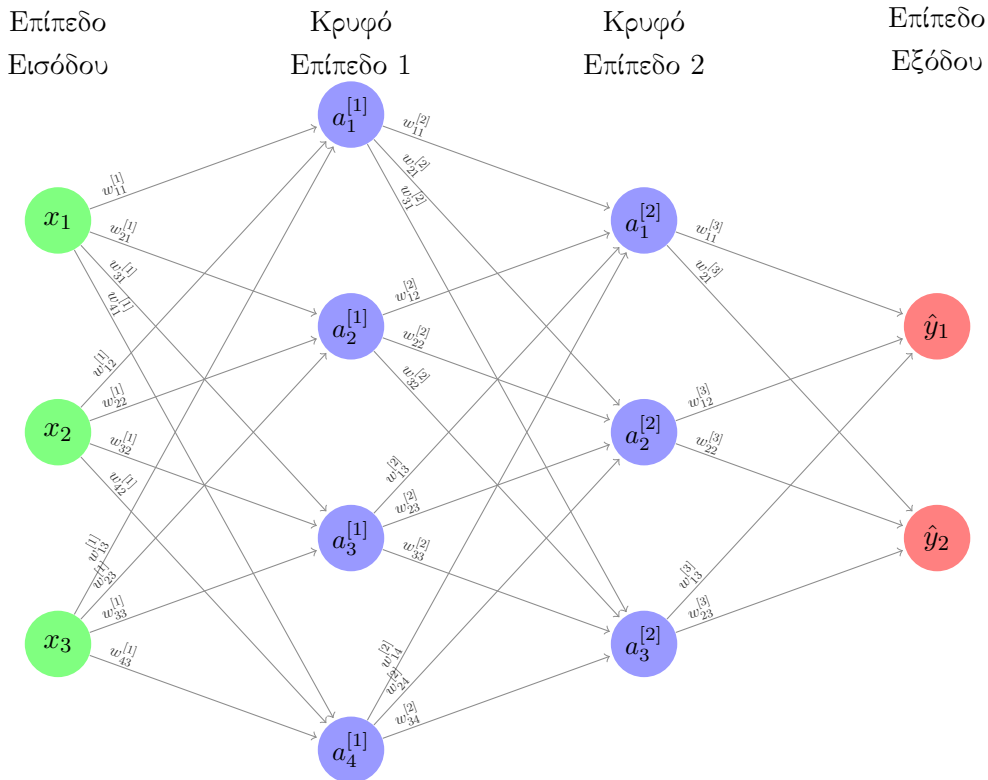
#### Δομή Απλών Νευρωνικών Δικτύων

Τα Νευρωνικά Δίκτυα στην πιο βασική τους μορφή (Feedforward Neural Networks ή Multilayer Perceptron) αποτελούνται από απλούς τεχνητούς νευρώνες διασυνδεδεμένους μεταξύ τους με συνάψεις σχηματίζοντας μια πολυεπίπεδη διάταξη. Το πρώτο επίπεδο ονομάζεται επίπεδο εισόδου (input layer) ενώ το τελευταίο ονομάζεται επίπεδο εξόδου (output layer). Όλα τα ενδιάμεσα επίπεδα λέγονται κρυφά επίπεδα (hidden layers) διότι οι τιμές τους δε δίνονται από τα δεδομένα [33]. Στην απλή περίπτωση που εξετάζουμε, κάθε νευρώνας δέχεται ως είσοδο τιμές από όλους τους νευρώνες του αμέσως προηγούμενου επιπέδου (fully connected layer) και αφού κάνει

---

<sup>6</sup>Για έναν τυπικό ορισμό, παραπέμπουμε τον αναγνώστη στο παράρτημα Α'

υπολογισμούς με αυτές στέλνει το αποτέλεσμα σε όλους τους νευρώνες του αμέσως επόμενου επιπέδου.



Σχήμα 2.3: Διάγραμμα Τεχνητού Νευρωνικού Δικτύου με δύο κρυφά επίπεδα. Οι αγκύλες στους εκθέτες προσδιορίζουν τον αριθμό του επιπέδου. Παράχθηκε από το *LaTeX* πακέτο *neuralnetwork*. Το πακέτο τροποποιήθηκε και επεκτάθηκε τοπικά.

Η αναπαράσταση των νευρωνικών δικτύων γίνεται με έναν γράφο από ακμές (συνάψεις) και κόμβους (νευρώνες ή τιμές εισόδου). Κοιτώντας κανείς το σχήμα 2.3 μπορεί να παρατηρήσει πως οι κόμβοι των επιπέδων εισόδου και εξόδου ξεχωρίζουν από τους κόμβους των κρυφών επιπέδων. Αυτό έχει γίνει για να τονιστεί η ξεχωριστή λειτουργία τους. Πιο συγκεκριμένα, στην περίπτωση του επιπέδου εισόδου, αυτό περιέχει τόσους κόμβους όσος είναι και ο αριθμός των χαρακτηριστικών που περιγράφουν το κάθε παράδειγμα (δηλαδή όσο και το μήκος του διανύσματος εισόδου). Ουσιαστικά, οι κόμβοι εισόδου απλά λαμβάνουν τις τιμές των χαρακτηριστικών και, χωρίς να τις μεταβάλλουν, τις δρομολογούν στους κόμβους του επόμενου επιπέδου (για αυτό και αποφεύγεται η επωνομασία αυτών των κόμβων ως νευρώνες). Στην περίπτωση του επιπέδου εξόδου, ο αριθμός των κόμβων είναι τόσος όσος και ο αριθμός των χαρακτηριστικών για την περιγραφή της πρόβλεψης (τόσο όσο το μήκος του διανύσματος εξόδου). Οι κόμβοι εξόδου συνήθως επιβάλλουν περιορισμούς στις τιμές των χαρακτηριστικών εξόδου ώστε αυτές να ανήκουν σε ένα φραγμένο σύνολο αριθμών (π.χ. το  $[0,1]$ ).

Ένα νευρωνικό δίκτυο χωρίς κρυφά επίπεδα δε διαφέρει από έναν γραμμικό ταξινομητή. Είναι γεγονός ότι οι εκπληκτικές δυνατότητες των νευρωνικών δικτύων αποδίδονται στα κρυφά επίπεδα. Χάρη σε αυτά είναι δυνατή η σταδιακή σύνθεση αφηρημένων αναπαραστάσεων από επίπεδο σε επίπεδο που κωδικοποιούν τους παράγοντες διακύμανσης. Τα κρυφά επίπεδα τα απαρτίζουν οι



κόμβοι κρυφού επιπέδου<sup>7</sup>. Ο κάθε ένας από αυτούς υπολογίζει την έξοδο μιας μη γραμμικής συνάρτησης με είσοδο ένα γραμμικό συνδυασμό των τιμών των κόμβων του προηγούμενου επιπέδου. Αξίζει να αναφερθεί στο σημείο αυτό πως δεν υπάρχει κάποιος συγκεκριμένος περιορισμός για τον αριθμό των κόμβων των κρυφών επιπέδων.

Η φορμαλιστική περιγραφή των παραμέτρων<sup>8</sup> και των υπολογισμών που λαμβάνουν χώρα κατά τη διαδικασία πρόβλεψης ενός νευρωνικού δικτύου περιγράφονται παρακάτω.

Έστω ένα παράδειγμα εισόδου το οποίο περιγράφεται από  $n_x$  χαρακτηριστικά με το διάνυσμα  $X = [x_1, x_2, x_3, \dots, x_{n_x}]^T$ . Όλα τα δεδομένα εκπαίδευσης, έστω  $M$ , μπορούν να ομαδοποιηθούν σε έναν πίνακα  $\mathbf{X}$  ως εξής:

$$\mathbf{X} = \begin{bmatrix} | & | & \dots & | \\ X^{(1)} & X^{(2)} & \dots & X^{(M)} \\ | & | & \dots & | \end{bmatrix}. \quad (2.1)$$

$(n_x \times M)$

Όπου οι παρενθέσεις στους εκθέτες δηλώνουν τον αριθμό του παραδείγματος.

Αφού προσδιορίσαμε μια μαθηματική αναπαράσταση για τα δεδομένα εισόδου, πάμε να προσδιορίσουμε με φορμαλιστικό τρόπο τις παραμέτρους του νευρωνικού δικτύου. Με  $L$  θα συμβολίζουμε τον αριθμό των επιπέδων του νευρωνικού δικτύου (χωρίς να μετράμε το επίπεδο εισόδου). Για παράδειγμα, στο δίκτυο του σχήματος 2.3 ισχύει  $L = 3$ . Επίσης, ο αριθμός των κόμβων ενός επιπέδου, έστω  $l$ , θα συμβολίζεται με  $n^{[l]}$ . Προφανώς, θα ισχύει ότι  $l \in [0, L]$ . Στο παράδειγμα του σχήματος 2.3 ισχύει  $n^{[0]} = n_x = 3, n^{[1]} = 4, n^{[2]} = 3, n^{[3]} = n_y = 2$ . Έχοντας δώσει έναν συμβολισμό ορισμένων βασικών υπερπαραμέτρων<sup>9</sup>, μπορούμε αποδώσουμε φορμαλιστικά τις παραμέτρους του δικτύου οι οποίες είναι:

- Τα βάρη των ακμών (weights) που συνδέουν δύο διαδοχικά επίπεδα.

Τα βάρη μεταξύ διαδοχικών επιπέδων  $l-1$  και  $l$  μπορούμε να τα οργανώσουμε σε μια μορφή πίνακα ως εξής:

$$\mathbf{W}^{[l]} = \begin{bmatrix} w_{11}^{[l]} & w_{12}^{[l]} & \dots & w_{1n^{[l-1]}}^{[l]} \\ w_{21}^{[l]} & w_{22}^{[l]} & \dots & w_{2n^{[l-1]}}^{[l]} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n^{[l]}1}^{[l]} & w_{n^{[l]}2}^{[l]} & \dots & w_{n^{[l]}n^{[l-1]}}^{[l]} \end{bmatrix}. \quad (2.2)$$

$(n^{[l]} \times n^{[l-1]})$

- Τα δυναμικά πόλωσης (biases) του κάθε νευρώνα.

Όπως είναι λογικό, οι κόμβοι του επιπέδου 0 (επίπεδο εισόδου) δε διαθέτουν δυναμικά πόλωσης. Για όλους τους άλλους κόμβους σε κάθε επίπεδο (έστω  $l$ ) έχουμε το εξής

<sup>7</sup>Εφεξής θα αποκαλούνται ως κόμβοι.

<sup>8</sup>Πρόκειται για μεταβλητές των οποίων οι τιμές μαθαίνονται κατά τη διάρκεια της εκπαίδευσης. Έτσι το νευρωνικό δίκτυο λέμε ότι προσαρμόζεται στα δεδομένα.

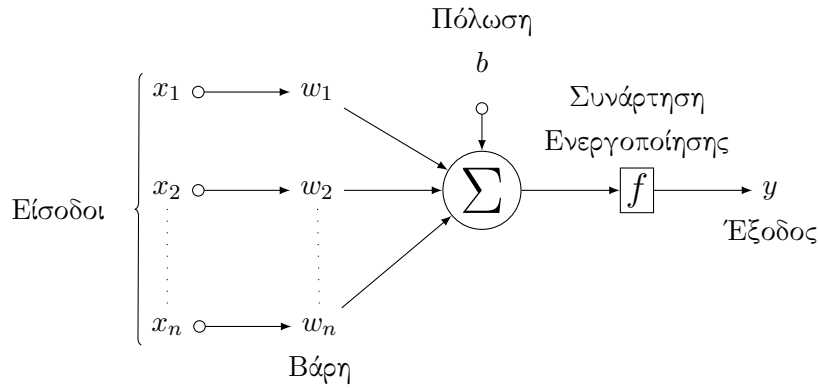
<sup>9</sup>Σε αντίθεση με τις παραμέτρους, οι υπερπαραμέτροι είναι μεταβλητές που ορίζει ο χρήστης και δε μεταβάλλονται κατά την εκπαίδευση. Ονομάζονται έτσι διότι ελέγχουν έμμεσα τις τιμές των παραμέτρων.

διάνυσμα στήλη:

$$\mathbf{b}^{[l]} = \begin{bmatrix} b_1^{[l]} \\ b_2^{[l]} \\ \vdots \\ b_{n^{[l]}}^{[l]} \end{bmatrix}. \quad (2.3)$$

$(n^{[l]} \times 1)$

Τώρα, είμαστε σε θέση να περιγράψουμε τους υπολογισμούς που πραγματοποιεί κάθε νευρώνας. Για τον σκοπό αυτό, παρουσιάζουμε μια αναπαράστασή του στο σχήμα 2.4.



Σχήμα 2.4: Διάγραμμα ενός τεχνητού νευρώνα. Παράχθηκε από το πακέτο *tikz*.

Εσωτερικά ο κάθε νευρώνας δέχεται ως είσοδο τις τιμές των νευρώνων του προηγούμενου επιπέδου ή (στην περίπτωση του δεύτερου επιπέδου) τις τιμές του παραδείγματος εκπαίδευσης και πραγματοποιεί υπολογισμούς με αυτές για να παράξει μια τιμή εξόδου. Την κάθε τιμή εισόδου τη συμβολίζουμε με  $x$  ενώ την τιμή εξόδου του μεμονωμένου νευρώνα τη συμβολίζουμε με  $y$ . Η πράξη που επιτελείται εσωτερικά είναι:

$$y = f\left(\sum_{i=1}^n w_i \times x_i + b\right) \quad (2.4)$$

Όπου  $f(x)$  είναι η συνάρτηση ενεργοποίησης (activation function). Αυτές οι συναρτήσεις, μεταξύ άλλων, επιτρέπουν στο σύστημα να μοντελοποιήσει μη γραμμικές σχέσεις εισόδου–εξόδου. Πρωταρχικά, χρησιμοποιούνταν για να μοντελοποιήσουν τη μη-γραμμική, δύτιμη έξοδο των βιολογικών νευρώνων. Στον χώρο των υπολογιστών όμως, έγιναν γρήγορα αντιληπτά τα πρακτικά πλεονεκτήματα της κατασκευής βιολογικών νευρώνων με έξοδο συνεχείς τιμές που ανήκουν στον χώρο των πραγματικών αριθμών<sup>10</sup>. Έτσι, από τη βηματική συνάρτηση (όπως αυτή στο Perceptron του H. Rosenblat [32]) που έχει ως σύνολο τιμών το 0, 1, άρχισε να γίνεται χρήση άλλων με συνεχές πεδίο τιμών όπως για παράδειγμα η σιγμοειδής ή η υπερβολική συνάρτηση εφαπτομένης (tanh) κ.τ.λ.

Στο σημείο αυτό να αναφέρουμε πως το όρισμα της συνάρτησης ενεργοποίησης, δηλαδή την ποσότητα  $\sum_{i=1}^n w_i \times x_i + b$  τη συμβολίζουμε με  $z$  ενώ την τιμή του  $y$  την ονομάζουμε και τιμή

<sup>10</sup>Αυτός είναι και ο λόγος που ο όρος «ενεργοποίηση» θέλει προσοχή όταν αναφερόμαστε σε τεχνητούς νευρώνες. Στην περίπτωση χρήσης του όρου σε ένα τέτοιο πλαίσιο, θα εννοούμε ότι ο νευρώνας αυτός έχει σχετικά μεγαλύτερη τιμή εξόδου και συνάμα έχει μεγαλύτερη «ευθύνη» στη διαμόρφωση της τελικής εξόδου.

ενεργοποίησης (συμβολίζοντάς τη με  $a$ ). Στην περίπτωση δε που ο νευρώνας ανήκει στο επίπεδο εξόδου, την τιμή  $y$  την ονομάζουμε τιμή πρόβλεψης και την αναπαριστούμε με το γράμμα  $\hat{y}$ .

Κοιτώντας τώρα εποπτικά τη διαδικασία παραγωγής νέων προβλέψεων ενός τυπικού νευρωνικού δικτύου, μπορούμε να την περιγράψουμε χρησιμοποιώντας αναπαράσταση με διανύσματα (vector notation). Αναλυτικότερα, αν συγκεντρώσουμε όλες τις τιμές ενεργοποίησης  $a$  ενός επιπέδου  $l$  σε ένα διάνυσμα  $A^{[l]} = [a_1^{[l]}, a_2^{[l]}, a_3^{[l]}, \dots, a_{n^{[l]}}^{[l]}]^T$  και κάνουμε το ίδιο για τα ορίσματα  $z$  των συναρτήσεων ενεργοποίησης δηλαδή  $Z^{[l]} = [z_1^{[l]}, z_2^{[l]}, z_3^{[l]}, \dots, z_{n^{[l]}}^{[l]}]^T$  τότε συνολικά, γράφουμε:

- Τα στοιχεία  $a$  προκύπτουν από τα στοιχεία  $z$  του ίδιου επιπέδου ως εξής:

$$A^{[l]} = F^{[l]}(Z^{[l]}) \quad (2.5)$$

Όπου η συνάρτηση  $F$  εφαρμόζει την  $f$  σε κάθε στοιχείο του ορίσματος της ξεχωριστά (elementwise).

- Τα στοιχεία  $z$  υπολογίζονται από τα στοιχεία  $a$  του προηγούμενου επιπέδου μέσω της σχέσης:

$$Z^{[l]} = W^{[l]} \times A^{[l-1]} + \mathbf{b}^{[l]} \quad (2.6)$$

Όπου  $A^0 = X$ , το διάνυσμα χαρακτηριστικών ενός παραδείγματος ενώ  $A^{[L]} = \hat{Y}$  το διάνυσμα εξόδου.

Η ανωτέρω ανάλυση πραγματοποιήθηκε για ένα μεμονωμένο παράδειγμα. Θα βοηθήσει για την επεξήγηση του τρόπου εκπαίδευσης των νευρωνικών δικτύων αν παρουσιάσουμε τώρα μια μορφή μαθηματικών τύπων που συγκεντρώνουν τις τιμές όλων των παραδειγμάτων του συνόλου εκπαίδευσης. Θυμηθείτε, ότι κάθε στήλη του πίνακα δεδομένων εισόδου  $\mathbf{X}$  περιέχει ένα διάνυσμα παραδείγματος με αποτέλεσμα ο αριθμός των στηλών να ισούται με τον αριθμό των παραδειγμάτων,  $M$ . Κατά αυτόν τον τρόπο έχουμε:

- Για τα  $z$  ενός επιπέδου  $l$  για το κάθε παράδειγμα εισόδου:

$$\mathbf{Z}^{[l]} = \begin{bmatrix} | & | & \dots & | \\ Z^{[l](1)} & Z^{[l](2)} & \dots & Z^{[l](M)} \\ | & | & & | \end{bmatrix}. \quad (2.7)$$

$(n^{[l]} \times M)$

- Αντίστοιχα, για τα  $a$  ενός επιπέδου  $l$  για το κάθε παράδειγμα εισόδου:

$$\mathbf{A}^{[l]} = \begin{bmatrix} | & | & \dots & | \\ A^{[l](1)} & A^{[l](2)} & \dots & A^{[l](M)} \\ | & | & & | \end{bmatrix}. \quad (2.8)$$

$(n^{[l]} \times M)$

Όπως προκύπτουν από τις σχέσεις

$$\mathbf{Z}^{[l]}_{(n^{[l]} \times M)} = \mathbf{W}^{[l]}_{(n^{[l]} \times n^{[l-1]})} \times \mathbf{A}^{[l-1]}_{(n^{[l-1]} \times M)} + \mathbf{b}^{[l]}_{(n^{[l]} \times 1)} \quad (2.9)$$

και

$$\mathbf{A}^{[l]}_{(n^{[l]} \times M)} = F^{[l]}(\mathbf{Z}^{[l]}_{(n^{[l]} \times M)}) \quad (2.10)$$

αντίστοιχα, όπου η μόνη διαφορά με τις 2.6, 2.5 είναι ότι αντικαταστήσαμε τα διανύσματα στήλες  $Z$  και  $A$  με τους πίνακες  $\mathbf{Z}$  και  $\mathbf{A}$ . Και πάλι,  $\mathbf{A}^0 = \mathbf{X}$ , ο πίνακας όλων των δεδομένων εισόδου ενώ  $\mathbf{A}^{[L]} = \hat{\mathbf{Y}}$  το σύνολο διανυσμάτων εξόδου.

## Εκπαίδευση Νευρωνικών Δικτύων

Στην προηγούμενη παράγραφο διατυπώσαμε τους μαθηματικούς τύπους σύμφωνα με τους οποίους ένα νευρωνικό δίκτυο, δοθέντος ενός συνόλου δεδομένων  $\mathbf{X}$  παράγει ένα σύνολο από προβλέψεις  $\hat{\mathbf{Y}}$ . Η διαδικασία αυτή ονομάζεται και πρόσθια διάδοση (forward propagation). Παρόλα αυτά, δεν αναφερθήκαμε καθόλου στη διαδικασία μάθησης του δικτύου. Υποθέσαμε σιωπηρά ότι αυτό ήταν ήδη εκπαιδευμένο και οι παράμετροί του (τα βάρη και τα δυναμικά πόλωσης) ήταν σταθερά. Με τη μέχρι τώρα παρουσίαση, το νευρωνικό δίκτυο δεν είναι τίποτα άλλο παρά μια μη γραμμική συνάρτηση. Σε αυτήν την παράγραφο όμως, θα κάνουμε τη σύνδεση των νευρωνικών δικτύων με τη μηχανική μάθηση αναλύοντας τον μηχανισμό εκπαίδευσής τους.

Όπως έχουμε αναφέρει, τα νευρωνικά δίκτυα είναι πολυδύναμα συστήματα μηχανικής μάθησης βασισμένα σε μοντέλο ενώ καμία επιπλέον υπόθεση δεν μπορεί να γίνει εκ των προτέρων. Με άλλα λόγια, υπάρχουν νευρωνικά δίκτυα που ανήκουν σε όλες τις υπόλοιπες κατηγορίες που παρατέθηκαν στην ενότητα 2.1.1. Παρόλα αυτά, για τον σκοπό της παρούσας παραγράφου, θα περιορίσουμε αυτά τα πολυδύναμα συστήματα σε αυτά που πραγματοποιούν επιβλεπόμενη μάθηση. Ευτυχώς, αυτή είναι η πιο βασική κατηγορία και οι ιδέες που θα παρουσιαστούν υπό αυτή εύκολα μεταφέρονται και στις υπόλοιπες.

Στο πλαίσιο της επιβλεπόμενης μάθησης, εκτός από το σύνολο δεδομένων εισόδου  $\mathbf{X}$  παρέχεται, και ένα σύνολο δεδομένων εξόδου  $\mathbf{Y}$ . Το τελευταίο αποτελείται από ένα επιθυμητό διάνυσμα (ή τιμή) στόχο για κάθε παράδειγμα του συνόλου δεδομένων εισόδου. Έτσι, (όπως αναφέρουμε και στο παράρτημα Α') σχηματίζονται ζεύγη διανυσμάτων εισόδου–επιθυμητής εξόδου. Στόχος του νευρωνικού δικτύου σε αυτήν την περίπτωση είναι να δημιουργήσει μια συνάρτηση που θα κάνει την αντιστοίχιση από τα παραδείγματα  $X$  στις προβλέψεις του στόχου,  $\hat{Y}$  να είναι όσο πιο πιστή γίνεται στην αντιστοίχιση  $X$  σε  $Y$ . Με μαθηματικούς όρους, έστω η συνάρτηση του νευρωνικού δικτύου:  $\mathcal{F}(X; \bar{W}, \bar{b})$ , όπου με  $\bar{W}$  και  $\bar{b}$  συμβολίζεται αντίστοιχα το σύνολο των βαρών και δυναμικών πόλωσης σε όλα τα επίπεδα<sup>11</sup>. Θέλουμε για τη συνάρτηση  $\mathcal{F}(X; \bar{W}, \bar{b}) : X \rightarrow \hat{Y}$  να ισχύει  $\mathcal{F}(X; \bar{W}^*, \bar{b}^*) \approx \mathcal{G}^*(X)$  όπου  $\mathcal{G}$  η (άγνωστη) συνάρτηση από την οποία (θεωρητικά) παράχθηκαν τα ζεύγη εισόδου–εξόδου και οι δείκτες αστερίσκοι συμβολίζουν τις ιδανικές τιμές των παραμέτρων<sup>12</sup>.

Για να προσαρμόσουμε με βέλτιστο τρόπο τις παραμέτρους του μοντέλου στην εμπειρία απαιτείται

<sup>11</sup>Το σύμβολο «;» διαβάζεται ως «παραμετροποιημένο από».

<sup>12</sup>Ουσιαστικά, το κύριο πρόβλημα που καλούνται να λύσουν τα νευρωνικά δίκτυα (και τα συστήματα μηχανικής μάθησης γενικότερα) πηγάζει από το γεγονός ότι η συνάρτηση  $\mathcal{G}$  είναι άγνωστη. Αντ' αυτής, διατίθεται ένα σύνολο ζευγών  $X - Y$  που αποτελεί υποσύνολο του πληθυσμού όλων των δυνατών εισόδων και το σύστημα επιδιώκει από αυτό το υποσύνολο να μάθει να γενικεύει σε παραδείγματα που δεν έχει δει. Αυτός είναι και ο λόγος που, υπό την αυστηρά μαθηματική έννοια, η διαδικασία της εκπαίδευσης αποκαλείται και συμπερασματολογία (inference). Σε τελική ανάλυση, τα περισσότερα νευρωνικά δίκτυα εκπαιδεύονται χρησιμοποιώντας συμπερασματολογία βασισμένη στη μέγιστη πιθανοφάνεια (maximum likelihood inference) [33]. Δηλαδή, προσπαθούν να μεγιστοποιήσουν την ποσότητα  $p(y|x; \bar{W}, \bar{b})$  και γιαυτό ονομάζονται μοντέλα διάκρισης (discriminative models).

(σύμφωνα με τον ορισμό της μηχανικής μάθησης) μια μετρική η οποία θα μας δείχνει πόσο κατάλληλη είναι η προσέγγισή μας (fitness). Αυτή η μετρική ονομάζεται συνάρτηση απώλειας (loss function) και στη γενική της μορφή είναι  $L(\hat{Y}, Y) = L(\mathcal{F}(\mathbf{X}; \bar{W}, \bar{b}), Y)$ . Ανάλογα με το είδος του προβλήματος, τυπικές συναρτήσεις απώλειας είναι:

- Η (binary cross entropy loss):  $L(\hat{Y}, Y) = -\frac{1}{M} \sum_{i=1}^M (y^{(i)} \log(\hat{y}^{(i)}) + (1-y^{(i)}) \log(1-\hat{y}^{(i)}))$  στην περίπτωση προβλήματος δυαδικής ταξινόμησης (όπου η έξοδος  $\hat{y}$  είναι ίση με την πιθανότητα το παράδειγμα  $x$  να ανήκει στην κλάση 1)
- Η (mean square error loss):  $L(\hat{Y}, Y) = \frac{1}{M} \sum_{i=1}^M \|y^{(i)} - \hat{y}^{(i)}\|^2$  στην περίπτωση προβλήματος παλινδρόμησης (όπου  $\|x\|$  συνήθως είναι η  $L2$  νόρμα).

Να σημειώσουμε ότι ορίσαμε τη συνάρτηση απώλειας ώστε να δέχεται ένα σύνολο από προβλέψεις και επιθυμητές τιμές στόχους. Αντίστοιχα, θα μπορούσαμε να ορίσουμε συνάρτηση απώλειας που συγκρίνει την πρόβλεψη με την έξοδο στόχο ενός συγκεκριμένου παραδείγματος (και να μην αθροίζει όλα τα παραδείγματα). Τότε, θα είχαμε  $L(\hat{y}, y) = L(\mathcal{F}(\mathbf{X}; \bar{W}, \bar{b}), y)$ .

Έχοντας στη διάθεσή μας μια μετρική απόδοσης, το πρόβλημα της εκπαίδευσης του νευρωνικού δικτύου μπορεί να διατυπωθεί ως πρόβλημα βελτιστοποίησης. Με μαθηματικούς όρους έχουμε:

$$\bar{W}^*, \bar{b}^* = \underset{\bar{W}, \bar{b}}{\operatorname{argmax}} (L(\mathcal{F}(\mathbf{X}; \bar{W}, \bar{b}), Y)) \quad (2.11)$$

Εξετάζοντας τη μετρική  $L$  σαν συνάρτηση των  $\bar{W}$  και  $\bar{b}$ , για την επίλυση του ανωτέρω προβλήματος αρκεί να βρούμε το σημείο στον χώρο των παραμέτρων που την ελαχιστοποιεί. Λόγω των μη γραμμικών στοιχείων όμως, δεν υπάρχει κλειστός τύπος (closed form) για την εύρεση του σημείου αυτού. Όπως φαίνεται και από το σχήμα 2.5, η συνάρτηση απώλειας είναι μη κυρτή. Αυτό μας οδηγεί στη χρήση επαναληπτικών μεθόδων για την εύρεση κάποιου (τοπικού) ελάχιστου.

Ο πιο δημοφιλής αλγόριθμος για αυτόν τον σκοπό είναι ο αλγόριθμος καθόδου κλίσης (gradient descent). Σύμφωνα με τον αλγόριθμο αυτό, πραγματοποιούνται επαναλαμβανόμενα βήματα «καθόδου» προς την κατεύθυνση με τη μεγαλύτερη κλίση. Διαισθητικά, φαίνεται λογικό σε κάθε βήμα να υπολογίζουμε σημειακά την κλίση της συνάρτησης που θέλουμε να ελαχιστοποιήσουμε και να κινούμαστε προς την κατεύθυνση με τη μικρότερη κλίση. Πιο συγκεκριμένα, πρώτα αρχικοποιούνται ανεξάρτητα όλες οι παράμετροι σε τυχαίες τιμές  $\bar{W}_0$ ,  $\bar{b}_0$  και έπειτα ξεκινά μια επαναληπτική διαδικασία όπου σε κάθε βήμα αυτής (έστω  $i$ ):

1. Υπολογίζονται οι μερικοί παράγωγοι (η κλίση) της συνάρτησης απώλειας ως προς όλες τις παραμέτρους σημειακά:

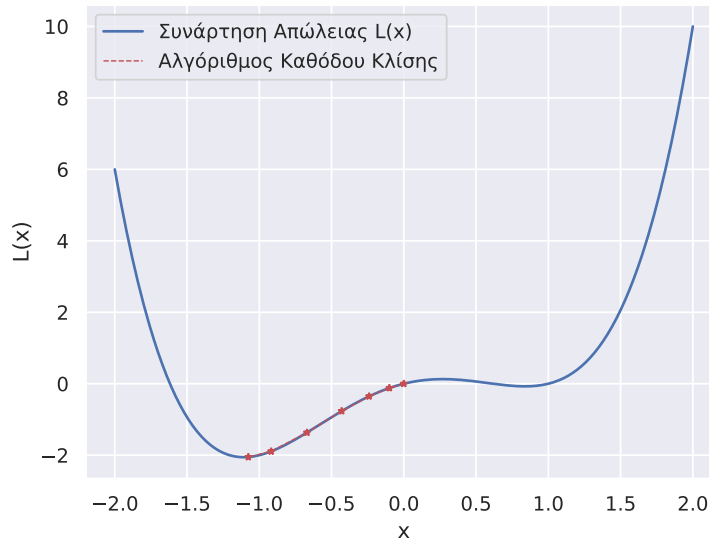
$$dw_i = \left. \frac{\partial L(\bar{W}, \bar{b})}{\partial w} \right|_{(\bar{W}, \bar{b})=(\bar{W}_{i-1}, \bar{b}_{i-1})}, \forall w \quad (2.12)$$

και

$$db_i = \left. \frac{\partial L(\bar{W}, \bar{b})}{\partial b} \right|_{(\bar{W}, \bar{b})=(\bar{W}_{i-1}, \bar{b}_{i-1})}, \forall b \quad (2.13)$$

Όπου  $L(\bar{W}, \bar{b})$  η συνάρτηση απώλειας υπολογισμένη για ένα σύνολο δεδομένων  $\mathbf{X}$  υπό τις παραμέτρους  $\bar{W}, \bar{b}$ . Συνοπτικά, αν συγκεντρώσουμε όλες τις παραμέτρους  $\bar{W}$  και  $\bar{b}$  στο διάνυσμα στήλη  $\bar{W}_{all}$  τότε γράφουμε:

$$d\bar{W}_{all i} = \nabla L(\bar{W}_{all}) \big|_{\bar{W}_{all}=\bar{W}_{all i-1}} \quad (2.14)$$



Σχήμα 2.5: Γραφική παράσταση στην οποία εφαρμόζεται ο αλγόριθμος καθόδου κλίσης σε μια μη κυρτή συνάρτηση με μια παράμετρο (την  $x$ ).

Να προσθέσω link σε κώδικα.

Παράχθηκε τοπικά.

2. Μετακινείται το σημείο στον χώρο παραμέτρων προς την κατεύθυνση της μεγαλύτερης κλίσης σύμφωνα με τον κανόνα ενημέρωσης (update rule) των παραμέτρων<sup>13</sup>. Ο κανόνας είναι ο εξής:

$$w_i = w_{i-1} - \alpha \times dw_i, \forall w \quad (2.15)$$

και

$$b_i = b_{i-1} - \alpha \times db_i, \forall b \quad (2.16)$$

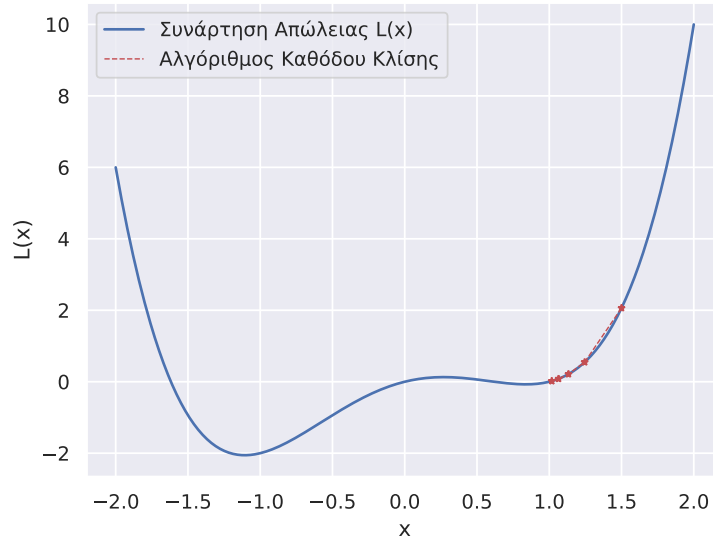
Όπου  $\alpha$  ο ρυθμός μάθησης (learning rate): μια υπερπαραμέτρος που καθορίζει το μέγεθος του βήματος κατά την ενημέρωση των παραμέτρων. Αντίστοιχα με πριν, συνοπτικά, έχουμε:

$$\overline{W}_{all i} = \overline{W}_{all i-1} - \alpha \times d\overline{W}_{all i} \quad (2.17)$$

Ο αλγόριθμος τελειώνει είτε όταν οι ενημερώσεις είναι πλέον αμελητέες και η τιμή της συνάρτησης απώλειας δε μειώνεται άλλο από επανάληψη σε επανάληψη (ο αλγόριθμος έχει βρει ένα τοπικό ελάχιστο) είτε όταν ξεπεραστεί ο μέγιστος αριθμός επαναλήψεων.

Να σημειώσουμε ότι ο αλγόριθμος καθόδου κλίσης δε βρίσκει πάντα το ολικό ελάχιστο της συνάρτησης. Ανάλογα με την αρχικοποίηση των παραμέτρων και τις τιμές των υπερπαραμέτρων (ρυθμός μάθησης, αριθμός επαναλήψεων) οδηγούμαστε κάθε φορά σε διαφορετικά αποτελέσματα. Για παράδειγμα, στο σχήμα 2.6 αρχικοποιήσαμε την τιμή της παραμέτρου  $x$  με την τιμή  $+1.5$  με αποτέλεσμα ο αλγόριθμος να τερματίσει στο τοπικό ελάχιστο της συνάρτησης. Ανεξάρτητα από

<sup>13</sup>Φανταστείτε τις παραγώγους ως «ευθύνες» της κάθε παραμέτρου για τις σωστές ή λάθος προβλέψεις. Όσο πιο μεγάλη η παράγωγος, τόσο πιο καθοριστικό ρόλο παίζει η μεταβλητή στη διαμόρφωση της τιμής της συνάρτησης απώλειας.



Σχήμα 2.6: Γραφική παράσταση στην οποία εφαρμόζεται ο αλγόριθμος καθόδου κλίσης σε μια μη κυρτή συνάρτηση με μια παράμετρο (την  $x$ ) αρχικοποιημένη όμως στην τιμή 1.5 με αποτέλεσμα να βρίσκεται ένα τοπικό ελάχιστο.

τον αριθμό επαναλήψεων, δε θα εντόπιζε ποτέ το ολικό ελάχιστο υπό αυτήν την αρχικοποίηση. Η αδυναμία εγγύησης για την εύρεση του ολικού ελαχίστου αποτελεί τον λόγο για τον οποίο σε αρκετά προβλήματα μηχανικής μάθησης οι αλγόριθμοι εκπαιδεύονται πολλές φορές με διαφορετική όμως αρχικοποίηση των παραμέτρων τους. Ευτυχώς, στα πολυεπίπεδα νευρωνικά δίκτυα δεν ενδιαφέρει η εύρεση του ολικού ελαχίστου<sup>14</sup> αλλά ενός τοπικού ελαχίστου [48].

Ο αλγόριθμος της καθόδου κλίσης δε θα χρῆσιμευε στην εκπαίδευση των νευρωνικών δικτύων αν δεν υπήρχε η δυνατότητα αποδοτικού υπολογισμού των μερικών παραγώγων. Ευτυχώς, τη λειτουργία αυτή την επιτελεί η μέθοδος οπισθοδιάδοσης σφάλματος (back propagation). Με λίγα λόγια, πρόκειται για μια μέθοδο η οποία χρησιμοποιώντας τον κανόνα της αλυσίδας υπολογίζει την παράγωγο της συνάρτησης απώλειας ως προς όλες τις παραμέτρους του δικτύου (σημειακά), ξεκινώντας από αυτές του τελευταίου επιπέδου και τερματίζοντας σε αυτές του πρώτου.

Παρακάτω παρατίθενται οι υπολογισμοί που λαμβάνουν χώρα κατά τη διάρκεια εύρεσης των μερικών παραγώγων ως προς τις παραμέτρους ενός επιπέδου  $L - 1$  μέσω της οπισθοδιάδοσης σφάλματος για την  $i$ -οστή επανάληψη του αλγορίθμου καθόδου κλίσης (με δεδομένα εισόδου ένα σύνολο από  $M$  παραδείγματα). Αν και οι παράγωγοι υπολογίζονται σημειακά για τις τιμές των παραμέτρων  $\bar{W}_i$  και  $\bar{b}_i$ , για λόγους ευκολότερης ανάγνωσης αυτό δε θα απεικονίζεται κατά τη διατύπωση των παρακάτω μερικών παραγώγων. Ξεκινώντας από το επίπεδο  $L$  έχουμε:

- Η παράγωγος της συνάρτησης απώλειας ως προς τα βάρη από τον κόμβο  $k$  του επιπέδου  $L - 1$  στον κόμβο  $j$  του επιπέδου  $L$  είναι:

$$\frac{\partial L(\bar{W}, \bar{b})}{\partial w_{jk}^{[L]}} = \frac{\partial z_j^{[L]}}{\partial w_{jk}^{[L]}} \times \frac{\partial a_j^{[L]}}{\partial z_j^{[L]}} \times \frac{\partial L(\bar{W}, \bar{b})}{\partial a_j^{[L]}} \quad (2.18)$$

<sup>14</sup>Καθώς οδηγεί σε overfitting.

Όπου ο όρος  $\frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{a}_j^{[L]}}$  υπολογίζεται άμεσα από την επιλεγμένη συνάρτηση απώλειας.

Ο όρος  $\frac{\partial \mathbf{a}_j^{[L]}}{\partial z_j^{[L]}}$  υπολογίζεται άμεσα από την επιλεγμένη συνάρτηση ενεργοποίησης.

Τέλος, η μερική παράγωγος  $\frac{\partial z_j^{[L]}}{\partial w_{jk}^{[L]}}$  υπολογίζεται λαμβάνοντας την παράγωγο του γραμμικού συνδυασμού

- Η παράγωγος της συνάρτησης απώλειας ως προς τα δυναμικά πόλωσης είναι:

$$\frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial b_j^{[L]}} = \frac{\partial z_j^{[L]}}{\partial b_j^{[L]}} \times \frac{\partial \mathbf{a}_j^{[L]}}{\partial z_j^{[L]}} \times \frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{a}_j^{[L]}} \quad (2.19)$$

Στην περίπτωση αυτή, η μερική παράγωγος  $\frac{\partial z_j^{[L]}}{\partial b_j^{[L]}} = 1$ .

- Τέλος, η παράγωγος της συνάρτησης απώλειας ως προς τις τιμές ενεργοποίησης του προηγούμενου επιπέδου  $\mathbf{a}_k^{[L-1]}$  είναι:

$$\frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{a}_k^{[L-1]}} = \sum_{j=1}^{n^{[L]}} \frac{\partial z_j^{[L]}}{\partial \mathbf{a}_k^{[L-1]}} \times \frac{\partial \mathbf{a}_j^{[L]}}{\partial z_j^{[L]}} \times \frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{a}_j^{[L]}} \quad (2.20)$$

Παρατηρούμε ότι οι μερικοί παράγωγοι των μεταβλητών ενός επιπέδου  $l - 1$  εξαρτώνται από το επίπεδο  $l$ . Για αυτό και όπως προαναφέραμε, οι υπολογισμοί ξεκινούν από το τελευταίο επίπεδο. Επαγωγικά, με τη χρήση της 2.20 στις 2.18 και 2.19 μπορούμε να βρούμε τις μερικές παραγώγους ως προς τις παραμέτρους όλων των επιπέδων.

Συγκεντρωτικά, χρησιμοποιώντας την αναπαράσταση με χρήση πίνακα που παρουσιάσαμε στην προηγούμενη παράγραφο, οι σχέσεις 2.18, 2.19 και 2.20 γράφονται αντίστοιχα [49]:

$$\frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{W}^{[l]}} = \frac{1}{M} \times \left( \frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{A}^{[l]}} \odot \frac{\partial \mathbf{A}^{[l]}}{\partial \mathbf{Z}^{[l]}} \right) \times \mathbf{A}^{[l-1]T} \quad (2.21)$$

$$\frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{b}^{[l]}} = \frac{1}{M} \times \left( \frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{A}^{[l]}} \odot \frac{\partial \mathbf{A}^{[l]}}{\partial \mathbf{Z}^{[l]}} \right) \times \mathbf{1}_M^T \quad (2.22)$$

$$\frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{A}^{[l-1]}} = \mathbf{W}^{[l]T} \times \left( \frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{A}^{[l]}} \odot \frac{\partial \mathbf{A}^{[l]}}{\partial \mathbf{Z}^{[l]}} \right) \quad (2.23)$$

Όπου ο τελεστής  $\odot$  συμβολίζει το γινόμενο στοιχείο προς στοιχείο (elementwise product) ενώ το  $\mathbf{1}_n = [1, 1, 1, \dots, 1] \in \mathbb{R}^n$ . Για τον όρο στην παρένθεση που συναντάται συχνά στους ανωτέρω τύπους ισχύει  $\left( \frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{A}^{[l]}} \odot \frac{\partial \mathbf{A}^{[l]}}{\partial \mathbf{Z}^{[l]}} \right) = \frac{\partial \mathcal{L}(\bar{\mathbf{W}}, \bar{\mathbf{b}})}{\partial \mathbf{Z}^{[l]}}$ .

Σαν τελικά σχόλια σχετικά με την εκπαίδευση των νευρωνικών δικτύων είναι ωφέλιμο να κάνουμε δύο παρατηρήσεις:

- Κατά την εφαρμογή του αλγορίθμου καθόδου κλίσης σε ένα νευρωνικό δίκτυο, σε κάθε βήμα αυτού γίνονται δύο περάσματα: μια πρόσθια διάδοση που περιγράφεται από τις εξισώσεις 2.9 και 2.10 για τον υπολογισμό της συνάρτησης απώλειας και μια οπισθοδιάδοση που περιγράφεται από τις εξισώσεις 2.21, 2.22 και 2.23 για τον υπολογισμό των παραγώγων που χρησιμοποιούνται στον κανόνα ενημέρωσης.



- Επειδή το σύνολο δεδομένων εισόδου μπορεί να είναι πολύ μεγάλο, αντί να λαμβάνονται όλα τα παραδείγματα  $M$  για τον υπολογισμό του  $d\overline{W}_{all}$  με βάση τη συνάρτησης απώλειας, συνηθίζεται να χωρίζεται σε μικρά πακέτα (mini batches) από  $m$  παραδείγματα το καθένα. Έτσι, πραγματοποιείται ένα βήμα ενημέρωσης για κάθε μικρό πακέτο δεδομένων. Όταν εφαρμόζεται αυτή η τακτική, σιωπηρά γίνεται η υπόθεση ότι το κάθε δείγμα των  $m$  παραδειγμάτων είναι επαρκώς αντιπροσωπευτικό ώστε η συνάρτηση απώλειας υπολογισμένη στα  $m$  παραδείγματα να είναι καλή προσέγγιση της συνάρτησης υπολογισμένης στα  $M$  παραδείγματα. Ακραία μορφή αυτού είναι ο στοχαστικός αλγόριθμος καθόδου κλίσης (stochastic gradient descent) στον οποίο  $m = 1$ . Οι μαθηματικοί τύποι που παραθέσαμε σε αυτό το κεφάλαιο ισχύουν σε κάθε περίπτωση μετά την κατάλληλη ανάθεση της υπερπαραμέτρου  $M$ .

### Συνελικτικά Νευρωνικά Δίκτυα

Έχοντας περιγράψει τη δομή και την εκπαίδευση των απλών νευρωνικών δικτύων πρόσθιας διάδοσης, εύκολα μπορούμε να κατανοήσουμε μερικές από τις παραλλαγές του. Μια από τις σημαντικότερες, είναι αυτή των Συνελικτικών Νευρωνικών Δικτύων (Convolutional Neural Networks) που χρησιμοποιείται συστηματικά στον χώρο της όρασης υπολογιστών. Πρόκειται για την υποκατηγορία των νευρωνικών δικτύων πρόσθιας διάδοσης στην οποία οδηγήθηκε η επιστημονική κοινότητα αφενός επιδιώκοντας να λύσει ορισμένα από τα πρακτικά προβλήματα της εφαρμογής νευρωνικών δικτύων στον χώρο της όρασης υπολογιστών και αφετέρου μελετώντας τη νευρο-φυσιολογία του οπτικού φλοιού (visual cortex).

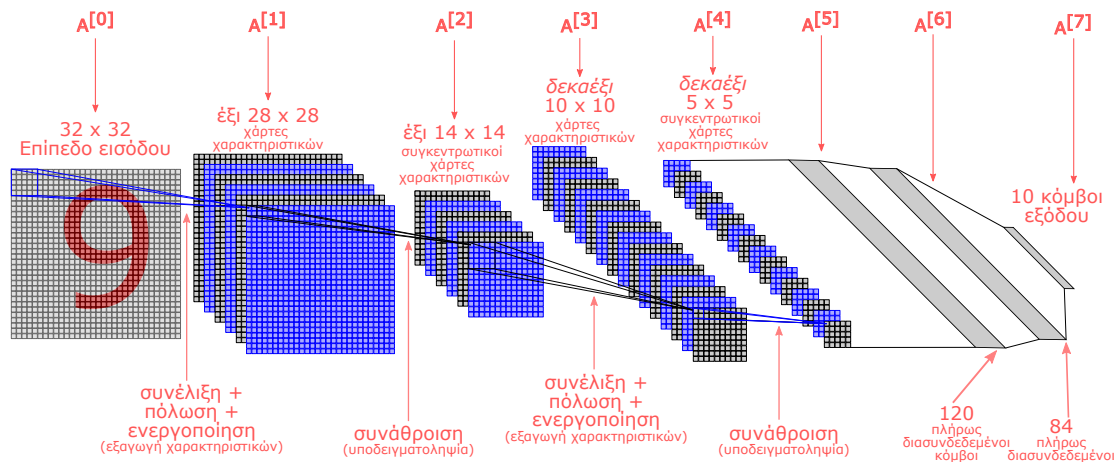
Από τη σκοπιά της νευροεπιστήμης, οι David H. Hubel και Torsten Wiesel μετά από μια σειρά πειραμάτων σε γάτες [50, 51] γύρω στο 1960 και αργότερα, σε πιθήκους [52] έριξαν φως στη δομή του οπτικού φλοιού, εμπνέοντας έτσι το κίνημα του διασυνδετισμού (connectionism). Σύμφωνα με το έργο τους, (για το οποίο τιμήθηκαν με το βραβείο nobel το 1981) πολλοί νευρώνες του οπτικού φλοιού έχουν μικρά, τοπικά πεδία υποδοχής (receptive fields) που μπορεί να επικαλύπτονται μεταξύ τους. Πιο συγκεκριμένα, ο κάθε νευρώνας αφορά ένα περιορισμένο τμήμα του οπτικού πεδίου αλλά όλοι μαζί, καλύπτουν το σύνολό του. Επιπλέον, μετά από πειράματα οπτικής αναγνώρισης σχημάτων (ορθογώνιο παραλληλόγραμμο σε μορφή μπάρας) σε διάφορες γεωμετρίες παρατηρήθηκε ότι διαφορετικοί νευρώνες με το ίδιο πεδίο υποδοχής ενεργοποιούνται ανάλογα με τη γεωμετρία του σχήματος (κάποιοι νευρώνες ενεργοποιούνται κατά τον κάθετο προσανατολισμό της μπάρας ενώ άλλοι με τον οριζόντιο προσανατολισμό της). Τέλος, επισήμαναν ότι ορισμένοι νευρώνες ενεργοποιούνται με την αναγνώριση πιο περίπλοκων μοτίβων όπως προκύπτουν από τη σύνθεση απλών γεωμετριών χαμηλότερου επιπέδου [46].

Από πρακτικής σκοπιάς, η υπολογιστική πολυπλοκότητα που προκύπτει από την τροφοδότηση ενός πλήρως διασυνδεδεμένου νευρωνικού δικτύου με εικόνες είναι απαγορευτική. Για παράδειγμα, έστω ότι διατίθεται ένα σύνολο από ασπρόμαυρες εικόνες μεγέθους  $100 \times 100$  εικονοστοιχεία. Αυτό συνεπάγεται ότι το επίπεδο εισόδου θα διαθέτει τόσους κόμβους όσα είναι και τα χαρακτηριστικά (τα εικονοστοιχεία), δηλαδή  $10.000^{15}$ . Στην κλασσική περίπτωση, ο αριθμός των κόμβων του πρώτου κρυφού επιπέδου θα είναι περίπου ίσος με τον αριθμό των χαρακτηριστικών εισόδου, δηλαδή πάλι 10.000. Αυτό σημαίνει ότι μόνο στο πρώτο επίπεδο του νευρωνικού δικτύου θα

---

<sup>15</sup>Χρησιμοποιούμε κόμμα για τη διάκριση των δεκαδικών (decimal comma separator) και τελεία για τη διάκριση των χιλιάδων.

υπήρχαν  $10.000 \times 10.000$  βάρη και 10.000 δυναμικά πόλωσης, ένας αριθμός, πολύ μεγάλος. Θα μπορούσαμε, φυσικά, αντί να τροφοδοτήσουμε το νευρωνικό δίκτυο με ολόκληρη την εικόνα σε ακατέργαστη μορφή, να εξάγαμε με ντετερμινιστικό τρόπο ορισμένα χαρακτηριστικά ώστε να καταλήξουμε με ένα μικρότερο διάνυσμα χαρακτηριστικών που θα εσωκλείει όλη τη χρήσιμη πληροφορία. Μια τέτοια διαδικασία χειρωνακτικής εξαγωγής χαρακτηριστικών απεικονίζεται στο σχήμα 2.2. Παρόλα αυτά, για τους λόγους που αναφέραμε στην ενότητα 2.1.2 θα επιθυμούσαμε την αυτοματοποιημένη εκμάθησή χαρακτηριστικών.



Σχήμα 2.7: Αρχιτεκτονική Συνελικτικού Νευρωνικού Δικτύου (LeNet-5) [53]. Η αναπαράσταση τους διαφέρει από αυτήν των απλών νευρωνικών δικτύων αφού εδώ δίνεται έμφαση στους πίνακες τιμών ενεργοποίησης  $A^{[l]}$ . Οι χάρτες χαρακτηριστικών αναπαριστώνται με τετράγωνα ενώ τα βάρη με ακμές. Τα δύο τελευταία επίπεδα είναι πλήρως διασυνδεδεμένα. Παράχθηκε από το *Inkscape* τροποποιώντας αυτήν την εικόνα.

Η λύση στο πρόβλημα δόθηκε μέσω της αξιοποίησης της τοπικής χωρικής συνεκτικότητας (local spacial coherence) και της ιεραρχικής δομής (hierarchical structure) των δεδομένων εικόνων. Εμπνευσμένη από τις ανωτέρω επιστημονικές παρατηρήσεις, ενσωματώθηκε η γνώση του χώρου προβλημάτων με εικόνες στη δομή των νευρωνικών δικτύων οδηγώντας έτσι στη δημιουργία των συνελικτικών νευρωνικών δικτύων (βλ. σχήμα 2.7). Οι δομικές διαφορές των συνελικτικών νευρωνικών δικτύων που τους διακρίνουν από τα νευρωνικά δίκτυα που παρουσιάστηκαν στην προηγούμενη ενότητα μπορούν να συνοψιστούν ως εξής:

- Μια πρώτη διαφορά που επιλύει το πρόβλημα της απαγορευτικής υπολογιστικής πολυπλοκότητας έγκειται στο τρόπο διασύνδεσης των κόμβων ενός επιπέδου με τους κόμβους του αμέσως προηγούμενου. Αντί να είναι πλήρως διασυνδεδεμένοι με αυτούς του προηγούμενου επιπέδου όπως στην περίπτωση των απλών νευρωνικών δικτύων, ενώνονται με βάρη μόνο με αυτούς που ανήκουν στο λεγόμενο πεδίο υποδοχής. Με άλλα λόγια, κάθε νευρώνας επιπέδου  $l$  δέχεται σαν είσοδο ένα διαφορετικό και περιορισμένο τμήμα του πίνακα  $A^{[l-1]}$ .
- Στα συνελικτικά νευρωνικά δίκτυα, οι κόμβοι του κάθε επιπέδου είναι οργανωμένοι σε όγκους τριών διαστάσεων με πλάτος, ύψος και βάθος. Με αυτόν τον τρόπο, διατηρείται η τοπική χωρική συνεκτικότητα. Αναλυτικότερα, οι κόμβοι εισόδου, για παράδειγμα,

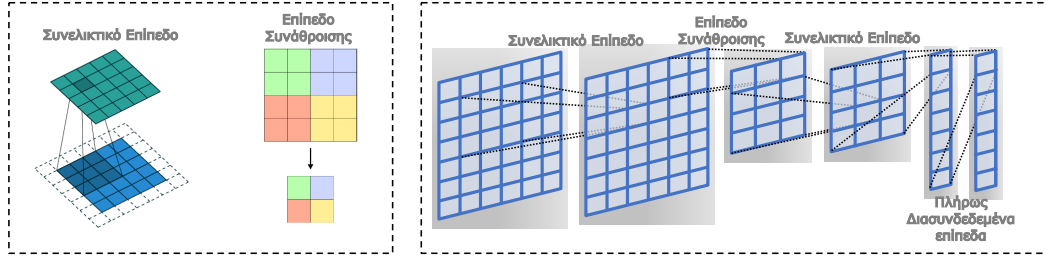
οργανώνονται όπως τα εικονοστοιχεία σε μια εικόνα: το βάθος του επιπέδου αντιστοιχεί στον αριθμό των καναλιών της εικόνας (π.χ. RGB) ενώ το ύψος και το πλάτος του επιπέδου στο ύψος και πλάτος της εικόνας. Έτσι, το νευρωνικό δίκτυο έχει τη δυνατότητα να αντλήσει εύκολα πληροφορία από μια χωρική γειτονιά της εικόνας (το πεδίο υποδοχής κάποιου νευρώνα) αφού οι αποστάσεις μεταξύ των εικονοστοιχείων διατηρούνται αναλλοίωτες. Αν όμως λαμβάναμε την εικόνα και την αναπτύσσαμε σε μια διάσταση (flatten) δημιουργώντας ένα μεγάλο διάνυσμα, τότε οι σχετικές αποστάσεις των στοιχείων εισόδου δε θα διατηρούνταν. Ανάλογες παρατηρήσεις ισχύουν και για τα κρυφά επίπεδα. Δηλαδή, και στα επόμενα επίπεδα οι κόμβοι οργανώνονται σε τρισδιάστατες δομές οι οποίες διατηρούν τοπικό χαρακτήρα. Η διαφορά έγκειται στο ότι η τιμή των κόμβων των κρυφών επιπέδων δεν είναι η τιμή του εκάστοτε εικονοστοιχείου στη θέση αυτή. Αντίθετα, είναι η τιμή ενός (σύνθετου) χαρακτηριστικού της περιοχής που έχει υπολογιστεί από την επεξεργασία απλούστερων χαρακτηριστικών προηγούμενων επιπέδων. Σχετικά με την ιδιότητα της ιεραρχικής δομής των εικόνων του πραγματικού κόσμου, αυτή αξιοποιείται μέσω διαδοχικών κρυφών επιπέδων που σταδιακά διευρύνουν το οπτικό πεδίο και συνθέτουν ολοένα και πιο σύνθετα χαρακτηριστικά. Έχειδειχθεί, ότι το σύστημα μαθαίνει να εξάγει μέσω των πρώτων επιπέδων απλά χαρακτηριστικά (π.χ. οριζόντιες και κάθετες ακμές) τα οποία σε επόμενα επίπεδα συνδυάζει για να εξάγει πιο περίπλοκα χαρακτηριστικά [54]. Αφού οι κόμβοι είναι οργανωμένοι σε όγκους, προκύπτει φυσικά ότι το διάνυσμα ενεργοποίησης  $A^{[l]}$  του κάθε επιπέδου  $l$  που κατασκευάζεται από την έξοδο κάθε κόμβου έχει τη μορφή πίνακα τριών διαστάσεων. Διαισθητικά για τα κρυφά επίπεδα, το ύψος και το πλάτος του διανύσματος ενεργοποίησης κωδικοποιούν αμυδρά τη θέση του χαρακτηριστικού στην εικόνα ενώ το βάθος κωδικοποιεί τα διάφορα χαρακτηριστικά (π.χ. βάθος 1: οριζόντιες ακμές, βάθος δύο: κατακόρυφες ακμές<sup>16</sup>). Στην περίπτωση των δικτύων που εξετάζουμε, το  $A^{[l]}$  λέμε ότι αποτελείται από επίπεδα φύλλα τα οποία στοιβάζονται στη διάσταση  $z$  και ονομάζονται χάρτες χαρακτηριστικών (feature maps).

- Μια ακόμα δομική διαφορά είναι ότι στα συνελικτικά νευρωνικά δίκτυα η εκμάθηση και εξαγωγή των χαρακτηριστικών δε γίνεται ανεξάρτητα σε κάθε περιοχή της εικόνας. Νευρώνες των οποίων τα βάρη προσαρμόζονται ώστε να αναγνωρίζουν και να εξάγουν γενικά χαρακτηριστικά της εικόνας όπως τα χαρακτηριστικά ακμών θα ήταν ασύμφορο να είχαν εφαρμογή μόνο στο πεδίο υποδοχής τους και όχι σε όλη την εικόνα. Έτσι, οδηγούμαστε στην έννοια του διαμοιρασμού παραμέτρων (weight sharing). Σύμφωνα με αυτήν την έννοια, αντί οι κόμβοι ενός επιπέδου να είναι διασυνδεδεμένοι με τους κόμβους του προηγούμενου επιπέδου με ξεχωριστά βάρη και δυναμικά πόλωσης, οι παράμετροι αυτές μοιράζονται μεταξύ των κόμβων. Έτσι, οι κόμβοι ενός επιπέδου επιτελούν τον ίδιο γραμμικό συνδυασμό  $y = f\left(\sum_{i=1}^n w_i \times x_i + b\right)$  αλλά με διαφορετικό διάνυσμα εισόδου  $X$  που εξαρτάται από το οπτικό πεδίο. Σημειώνουμε δε ότι ο διαμοιρασμός βαρών θα ήταν δύσκολο να εφαρμοστεί στην περίπτωση που η είσοδος αποτελούνταν από δομημένα δεδομένα καθώς αυτά μπορεί να είχαν πλήρως ετερογενή χαρακτηριστικά.

Πρακτικά, αν εξαιρέσουμε τα τελευταία, πλήρως διασυνδεδεμένα επίπεδα ενός συνελικτικού

---

<sup>16</sup>Πρακτικά στη διαδικασία εκμάθησης χαρακτηριστικών είναι δύσκολο να εκφράσουμε με σαφήνεια τι αναπαριστά το καθένα.



Σχήμα 2.8: Μεμονωμένο συνελικτικό επίπεδο και επίπεδο υποδειγματοληψίας (αριστερά). Συνδυασμός των επιπέδων για την κατασκευή ενός συνελικτικού νευρωνικού δικτύου (δεξιά). Παράχθηκε από το *Inkscape* τροποποιώντας αυτήν την εικόνα.

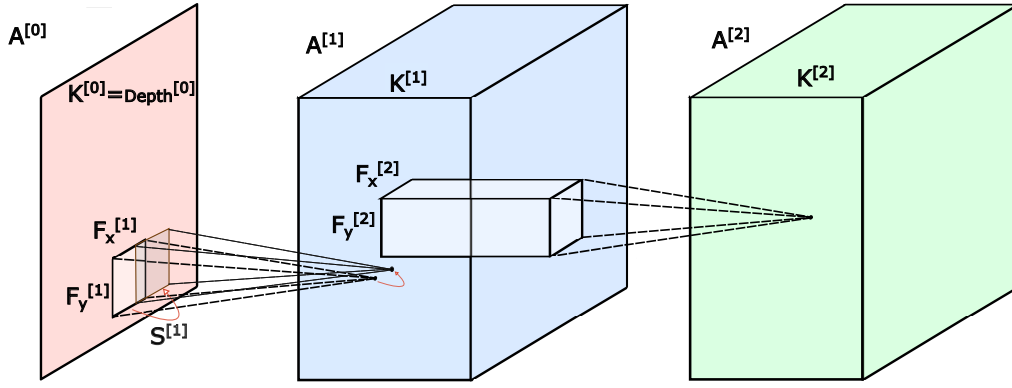
νευρωνικού δικτύου, οι ανωτέρω δομικές διαφορές υλοποιούνται με τη χρήση αφενός των συνελικτικών επιπέδων και αφετέρου των επιπέδων υποδειγματοληψίας. Αναφορικά με τα πρώτα, η εσωτερική τους λειτουργία απεικονίζεται στο αριστερό τμήμα του σχήματος 2.8. Το οπτικό πεδίο αναπαρίσταται με ένα σκούρο παραλληλόγραμμο επάνω στον χάρτη χαρακτηριστικών του προηγούμενου επιπέδου. Τα βάρη, είναι ευκολότερο να τα φανταστεί κανείς σαν ένα παραλληλόγραμμο (ή ένα ορθογωνικό κυβοειδές σε τρεις διαστάσεις) το οποίο έχει τις ίδιες διαστάσεις με το οπτικό πεδίο πάνω στο προηγούμενο επίπεδο. Η τιμή ενεργοποίησης κάθε στοιχείου του τρισδιάστατου πίνακα  $A^{[l]}$  υπολογίζεται ως το αποτέλεσμα της εφαρμογής της συνάρτησης ενεργοποίησης στον γραμμικό συνδυασμό των στοιχείων του πίνακα  $A^{[l-1]}$  που βρίσκονται εντός του οπτικού πεδίου με βάρη τα στοιχεία του  $W^{[l]}$  και το δυναμικό πόλωσης  $b^{[l]}$ . Στην ουσία, τα βάρη υπερτίθενται στο οπτικό πεδίο και επιτελείται γινόμενο μεταξύ των πινάκων στοιχείο προς στοιχείο (elementwise product). Αν υπήρχε ένας πίνακας από βάρη για κάθε στοιχείο του  $A^{[l]}$ , τότε δε θα είχαμε διαμοιρασμό βαρών. Αντιθέτως, ο διαμοιρασμός βαρών έγκειται στην ολίσθηση αυτού του παραλληλόγραμμου (ή ορθογωνικού κυβοειδούς στις τρεις διαστάσεις) στο ύψος και πλάτος της εικόνας, όπως φαίνεται και στο σχήμα 2.9 για την περίπτωση των δύο διαστάσεων. Για την περίπτωση που έχουμε πολλούς χάρτες χαρακτηριστικών, παραπέμπουμε τον αναγνώστη στο σχήμα 2.10. Αυτή η διαδικασία της ολίσθησης του πίνακα βαρών ονομάζεται δισδιάστατη συνέλιξη<sup>17</sup>, ενώ το κυλιόμενο παράθυρο ονομάζεται και φίλτρο (filter) ή πυρήνας (kernel).

Με κάθε συνελικτικό επίπεδο εισάγεται μια σειρά από παραμέτρους πέρα από αυτές που υπήρχαν σε κάθε απλό νευρωνικό δίκτυο. Για κάθε επίπεδο  $l$ , μεταξύ δύο στοιβαγμένων συνόλων από χάρτες χαρακτηριστικών  $A^{[l-1]}$  και  $A^{[l]}$  με διαστάσεις  $Width^{[l-1]} \times Height^{[l-1]} \times Depth^{[l-1]}$  και  $Width^{[l]} \times Height^{[l]} \times Depth^{[l]}$  αντίστοιχα πρέπει να ορίσουμε:

- Το μέγεθος του κυλιόμενου πυρήνα (ή φίλτρου). Αυτό εισαγάγει τις παραμέτρους  $F_x^{[l]}$  και  $F_y^{[l]}$ . Το βάθος του φίλτρου δεν αποτελεί υπερπαραμέτρο και είναι ίσο με  $Depth^{[l-1]}$ .
- Ο αριθμός των φίλτρων,  $K^{[l]}$ . Όπως έχει γίνει σαφές από τα ανωτέρω σχήματα, το κάθε φίλτρο  $k \in [1, K^{[l]}]$  λαμβάνει σαν είσοδο έναν όγκο από χαρακτηριστικά και έχει ως έξοδο έναν αριθμό  $a \in \mathbb{R}$ . Για να έχει ο παραγόμενος χάρτης χαρακτηριστικών βάθος, θα πρέπει να γίνει χρήση πολλαπλών φίλτρων που θα εξάγουν διαφορετικά χαρακτηριστικά. Έτσι,

<sup>17</sup>Τυπικά, η πράξη ονομάζεται διασταυρούμενη συσχέτιση (cross correlation) και είναι ίδια με τη συνέλιξη αν στην πρώτη περίπτωση αναποδογυρίσουμε τον πυρήνα (γυρνώντας τον ως προς την κύρια και δευτερεύουσα διαγώνιο).





Σχήμα 2.10: Συνελικτικά επίπεδα στη σειρά με έμφαση στην περίπτωση όπου οι πίνακες  $A$  είναι τριών διαστάσεων. Στο σχήμα φαίνονται οι υπερπαραμέτροι των συνελικτικών επιπέδων. Παράχθηκε από το *Inkscape* τροποποιώντας αυτήν την εικόνα.

χαρακτηριστικών  $A_k^{[l]}$  με διαστάσεις ίδιες με τον  $Z_k^{[l]}$ , δηλαδή:

$$Width^{[l]} = \frac{Width^{[l-1]} - F_x^{[l]} + 2 \times P_x^{[l]}}{S^{[l]}} + 1, \quad (2.24)$$

και

$$Height^{[l]} = \frac{Height^{[l-1]} - F_y^{[l]} + 2 \times P_y^{[l]}}{S^{[l]}} + 1 \quad (2.25)$$

4. Επανάληψη από το βήμα 1  $K^{[l]}$  φορές, όσο και το βάθος του  $A^{[l]}$ .
5. Στοίβαξη των παραχθέντων χαρτών χαρακτηριστικών ως προς τον άξονα  $z$  ώστε να κατασκευαστεί ο τρισδιάστατος πίνακας  $A^{[l]}$ . Τελικά, το σύνολο των χαρτών χαρακτηριστικών  $A^{[l]}$  έχει διαστάσεις μήκους και πλάτους ίδιες με αυτές του  $A_k^{[l]}$  αλλά το βάθος τώρα, αντί για μονάδα είναι:

$$Depth^{[l]} = K^{[l]} \quad (2.26)$$

Με μαθηματικούς όρους, οι υπολογισμοί που εκτελούνται σε ένα συνελικτικό επίπεδο είναι οι εξής:

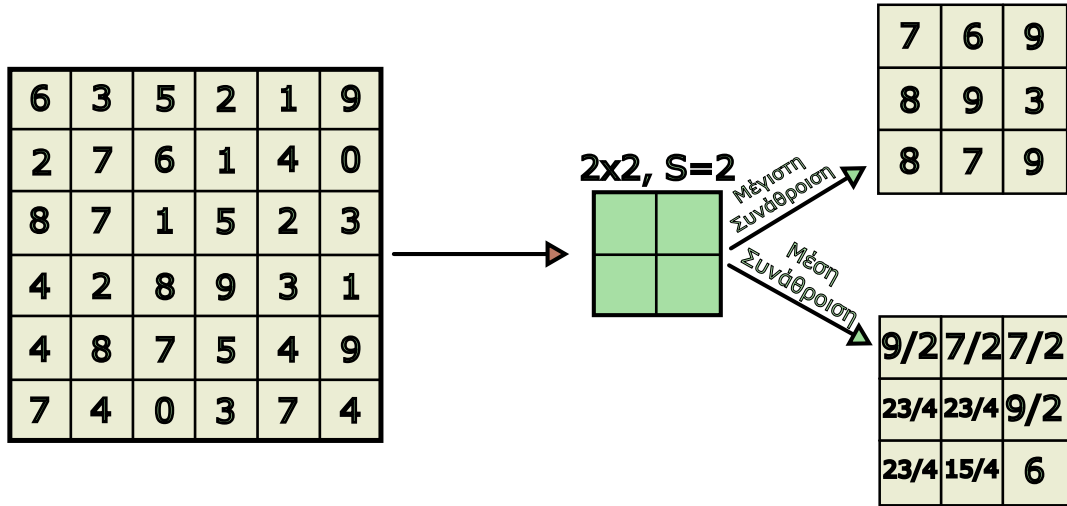
$$Z_k^{[l]} = W_k^{[l]T^T} \underset{step=S^{[l]}}{*} A^{[l-1]} + b_k^{[l]} \quad (2.27)$$

και

$$A_k^{[l]} = F^{[l]}(Z_k^{[l]}). \quad (2.28)$$

Όπου το σύμβολο  $\tau$  στον εκθέτη ενός πίνακα δηλώνει την αναστροφή του πίνακα υπό τη δευτερεύουσα διαγώνιο ενώ το σύμβολο  $\underset{step=S^{[l]}}{*}$  δηλώνει τη συνέλιξη με βήμα ίσο με  $S^{[l]}$ .

Ένα δεύτερο είδος επιπέδου αποτελεί αυτό της υποδειγματοληψίας (ή συνάνθροισης) όπως φαίνεται στο σχήμα 2.11. Το συγκεκριμένο είδος δε διαθέτει καμία παράμετρο αφού η μόνη λειτουργία του είναι να πραγματοποιεί υποδειγματοληψία στο χάρτη χαρακτηριστικών. Ο τρόπος εφαρμογής του είναι παρόμοιος με αυτόν του συνελικτικού επιπέδου. Δηλαδή, και πάλι υπάρχει ένα κυλιόμενο παράθυρο πάνω στον χάρτη χαρακτηριστικών το οποίο συναθροίζει τα στοιχεία στα οποία υπερτίθεται σε ένα στοιχείο υπό μια προκαθορισμένη στρατηγική. Πιο συγκεκριμένα, ανάλογα με το αν επιλέγεται σαν έξοδος το μεγαλύτερο στοιχείο στη γειτονιά συνάνθροισης



Σχήμα 2.11: Σχήμα επιπέδου συνάθροισης. Παράχθηκε από το Inkscape.

(το οπτικό πεδίο) ή μια μέση τιμή αυτών, έχουμε τη μέγιστη συνάθροιση (max pooling) ή τη μέση συνάθροιση (average pooling) αντίστοιχα. Σε κάθε περίπτωση, μια διαφορά με τα συνελικτικά επίπεδα είναι ότι το κυλιόμενο παράθυρο υπερτίθεται σε κάθε «φύλλο» της εισόδου  $A^{[l-1]}$  ξεχωριστά (δηλαδή, σε κάθε  $A_k^{[l-1]}$ ). Με άλλα λόγια, παρόλο που ο πίνακας  $A^{[l-1]}$  μπορεί να είναι τρισδιάστατος και να αποτελείται από πολλούς χάρτες χαρακτηριστικών στοιβαγμένους στον  $z$  άξονα, η γειτονιά συνάθροισης θα είναι πάντα ένα δισδιάστατο παράθυρο. Τέλος, να σημειώσουμε ότι αυτό το επίπεδο συμβάλλει στην ευρωστία του συστήματος καθιστώντας τις τιμές ενεργοποίησης των επόμενων επιπέδων αμετάβλητες σε μικρές διακυμάνσεις της θέσης των αντικειμένων, μια ιδιότητα που θα αναλύσουμε περαιτέρω στην επόμενη ενότητα.

Σε κάθε επίπεδο συνάθροισης  $l$  έχουμε τις εξής υπερπαραμέτρους:

- Το μέγεθος του πυρήνα υποδειγματοληψίας  $Pk_x^{[l]}$  και  $Pk_y^{[l]}$ . Καθορίζει τη γειτονιά συνάθροισης αλλά και το μέγεθος του συναθροισμένου χάρτη χαρακτηριστικών.
- Τη στρατηγική του επιπέδου συνάθροισης. Όπως αναφέραμε, εδώ οι στρατηγικές είναι δύο: μέγιστη συνάθροιση και μέση συνάθροιση.
- Το βηματισμό του κυλιόμενου παραθύρου  $s_x^{[l]}$  κατά τον  $x$  άξονα και  $s_y^{[l]}$  κατά τον  $y$  άξονα (όπως και στα συνελικτικά επίπεδα, συνήθως, οι δύο ποσότητες είναι ίσες και συμβολίζονται ως  $s^{[l]}$ ). Ήθιστε, το βήμα να ισούται με το μέγεθος του πυρήνα.

Αναφορικά με το μέγεθος της εξόδου ενός επιπέδου συνάθροισης  $l$ , με είσοδο έναν χάρτη χαρακτηριστικών  $A^{[l-1]}$  με διαστάσεις  $Width^{[l-1]} \times Height^{[l-1]} \times Depth^{[l-1]}$  ισχύει:

$$Width^{[l]} = \frac{Width^{[l-1]} - Pk_x^{[l]}}{S^{[l]}} + 1, \quad (2.29)$$

$$Height^{[l]} = \frac{Height^{[l-1]} - Pk_y^{[l]}}{S^{[l]}} + 1 \quad (2.30)$$

και

$$Depth^l = Depth^{l-1} \quad (2.31)$$

Έχοντας καλύψει πλήρως τα νευρωνικά δίκτυα και την υποκατηγορία τους η οποία χρησιμοποιείται στην όραση υπολογιστών, είμαστε σε θέση να περιγράψουμε ένα νεότερο είδος νευρωνικών δικτύων για τον ίδιο σκοπό, τα λεγόμενα νευρωνικά δίκτυα με κάψουλες.

## 2.2 Νευρωνικά Δίκτυα με Κάψουλες

Τα τελευταία χρόνια, διερευνάται μια ακόμα παραλλαγή των νευρωνικών δικτύων για εφαρμογές όρασης υπολογιστών: αυτή των νευρωνικών δικτύων με κάψουλες (capsule networks). Η ιδέα πίσω από τη νέα αρχιτεκτονική παρουσιάστηκε από τον Geoffrey Hinton, το ίδιο άτομο που είχε συμβάλει καθοριστικά στην ανάπτυξη και εδραίωση των συνελικτικών δικτύων [42]. Αυτή τη φορά όμως, στα σχετικά έργα του [55–57] τονίζει ορισμένες αδυναμίες της εδραιωμένης, πλέον, τεχνολογίας ενώ προτείνει μια νέα αρχιτεκτονική που θα τις αντιμετωπίζει. Ένας έμπειρος αναγνώστης μπορεί να επισημάνει ότι η σύλληψη της ιδέας των νευρωνικών δικτύων με κάψουλες δεν είναι νέα (2011). Παρόλα αυτά, όπως θα διαπιστώσουμε στο κεφάλαιο 3 μόλις πρόσφατα άρχισε να λαμβάνει πρακτική υπόσταση με την ανάπτυξη σύνθετων αρχιτεκτονικών που την πραγματώνουν.

Στην ενότητα αυτή θα ξεκινήσουμε κάνοντας αναφορά σε ορισμένα στοιχεία του ανθρώπινου μηχανισμού αναγνώρισης προτύπων εικόνων που αποτέλεσαν πηγή έμπνευσης για τα νευρωνικά δίκτυα με κάψουλες. Έπειτα, θα διατυπώσουμε τα ισχυρά και αδύναμα σημεία που παρουσιάζουν τα συνελικτικά νευρωνικά δίκτυα της προηγούμενης ενότητας. Τέλος, βασιζόμενοι στα κύρια έργα του Geoffrey Hinton σχετικά με τα νευρωνικά δίκτυα με κάψουλες στο πλαίσιο επιβλεπόμενης μάθησης [55–57], θα εμβαθύνουμε στις αρχές λειτουργίας τους.

### 2.2.1 Στοιχεία Έμπνευσης των Νευρωνικών Δικτύων με Κάψουλες

Για άλλη μια φορά, πηγή έμπνευσης για αυτήν την υποκατηγορία των νευρωνικών δικτύων με την οποία θα ασχοληθούμε σε μεγάλο βαθμό στην υπόλοιπη έκταση της εργασίας αποτέλεσε η νευροφυσιολογία. Πιο αναλυτικά, όπως έχουμε αναφέρει και στην ενότητα 1.2, οι νευρώνες στον εγκέφαλο οργανώνονται σε ολόενα και μεγαλύτερες δομές ανάλογα με τη λειτουργία τους. Σε γενικές γραμμές, γειτονικοί νευρώνες που επιτελούν παρόμοιες λειτουργίες ενισχύουν τις μεταξύ τους συνδέσεις σχηματίζοντας συστάδες<sup>19</sup>. Προκύπτει λοιπόν η διάθεση πειραματισμού για τη σχεδίαση μιας αρχιτεκτονικής νευρωνικών δικτύων που θα εμπεριέχει ρητά συστάδες από νευρώνες<sup>20</sup>. Αυτές τις συστάδες θα τις ονομάζουμε και κάψουλες (capsules).

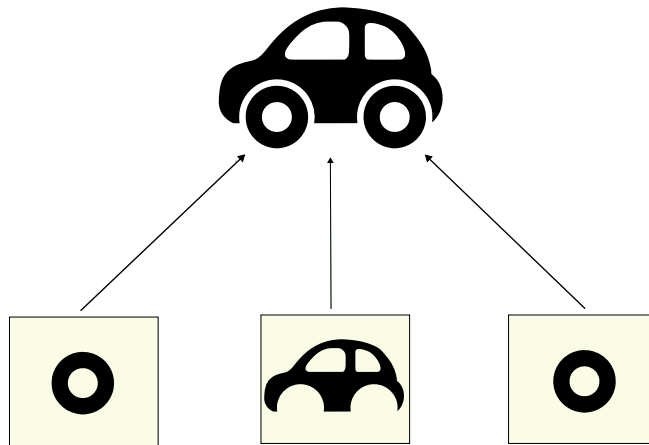
Επιπρόσθετα, έχει παρατηρηθεί ότι ο άνθρωπος αναγνωρίζει ένα αντικείμενο δημιουργώντας δυναμικά ένα ιεραρχικό δέντρο του οποίου η ρίζα εμπεριέχει το αντικείμενο προς αναγνώριση

<sup>19</sup>Παράδειγμα συστάδων με νευρώνες στον άνθρωπο που διαθέτουν κοινή είσοδο και κοινή έξοδο είναι η φλοιική μικρή στήλη (cortical minicolumn)

<sup>20</sup>Στις μέχρι τώρα αρχιτεκτονικές πλήρως διασυνδεδεμένων νευρωνικών δικτύων που έχουμε παρουσιάσει, δεν υπάρχουν τέτοιες δομές. Θα μπορούσαμε να υποθέσουμε ότι κάθε επίπεδο από νευρώνες αποτελεί μια τέτοια οργανωτική δομή. Η υπόθεση αυτή όμως δεν είναι πλήρως ευσταθής καθότι εσωτερικά αυτής οι νευρώνες δεν αλληλεπιδρούν άμεσα μεταξύ τους.



—υπό μια κωδικοποιημένη αναπαράσταση—ενώ τα κλαδιά τα επιμέρους τμήματα (ή χαρακτηριστικά) από τα οποία απαρτίζεται. Εκτός αυτού, θα μπορούσαμε να ισχυριστούμε ότι η ιεραρχική δομή είναι «εμπλουτισμένη», υπό την έννοια ότι τα κλαδιά του δέντρου κωδικοποιούν τη σχετική θέση των επιμέρους τμημάτων [58]. Αυτό προκύπτει από το γεγονός ότι ο άνθρωπος, με την αναγνώριση ενός αντικειμένου είναι πάντα σε θέση να προσδιορίσει τη σχετική θέση των μερών του (βλ. σχήμα 2.12). Η ιεραρχική δομή μεταξύ των αντικειμένων και των αποτελούμενων μερών του φαίνεται να υπάρχει παντού στη φύση. Είναι λογική λοιπόν η επιδίωξη ρητής ενσωμάτωσης μηχανισμών στα νευρωνικά δίκτυα που θα αξιοποιούν την πρότερη γνώση σχετικά με την εμπλουτισμένη ιεραρχική δομή των αντικειμένων του φυσικού κόσμου<sup>21</sup>.



Σχήμα 2.12: Σχήμα ιεραρχικού δέντρου μιας εικόνας αυτοκινήτου. Παράχθηκε από το Inkscape.

Το τρίτο και ίσως πιο σημαντικό στοιχείο από το οποίο εμπνεύστηκαν τα νευρωνικά δίκτυα με κάψουλες προκύπτει από την παρατήρηση ότι οι άνθρωποι πάντα εφαρμόζουν ένα σύστημα συντεταγμένων στα αντικείμενα που αναγνωρίζουν. Με άλλα λόγια, η αναγνώριση ενός αντικειμένου είναι άρρηκτα διασυνδεδεμένη με την αναγνώριση της γεωμετρίας του αντικειμένου. Για παράδειγμα, με τη θόραση ενός αυτοκινήτου αντιλαμβάνομαστε άμεσα και τον προσανατολισμό του. Μάλιστα, όπως φαίνεται στο σχήμα 2.13 ο τρόπος με τον οποίο εφαρμόζεται το σύστημα συντεταγμένων σε μια εικόνα διαδραματίζει πρωτεύοντα ρόλο στην κατανόησή της. Αυτή η λειτουργία του ανθρώπινου οπτικού φλοιού μας προδιαθέτει να δοκιμάσουμε στα τεχνητά νευρωνικά δίκτυα τη ρητή εκμάθηση ενός συστήματος αναφοράς για κάθε αντικείμενο που καλούνται να αναγνωρίσουν και τη σύγκριση κάθε νέας εικόνας εισόδου με αυτό. Όπως θα δούμε στη συνέχεια, μια τέτοια μέθοδος θα οδηγήσει το νευρωνικό δίκτυο σε αποδοτικότερη γενίκευση σχετικά με την εργασία αναγνώρισης αντικειμένων σε νέες γεωμετρίες.

<sup>21</sup>Μπορεί να ισχυριστεί κανείς ότι κάτι τέτοιο ισχύει σε αδρές γραμμές στα κλασσικά είδη βαθιών νευρωνικών δικτύων όπου εξάγοντας απλούστερα χαρακτηριστικά στα πρώτα επίπεδα καθίστανται ικανά να συνθέσουν πιο σύνθετα χαρακτηριστικά στα επόμενα. Μολονότι τα κλασσικά είδη αναλύουν τα δεδομένα μέσω διαδοχικών επιπέδων αξιοποιώντας έτσι την ιεραρχική τους φύση, αδυνατούν εκάτασχευής να ενσωματώσουν με ρητό τρόπο τη γνώση των σχέσεων σύνδεσης μεταξύ μερών του όλου (τμημάτων ενός αντικειμένου) και του όλου (ολόκληρου του αντικειμένου).



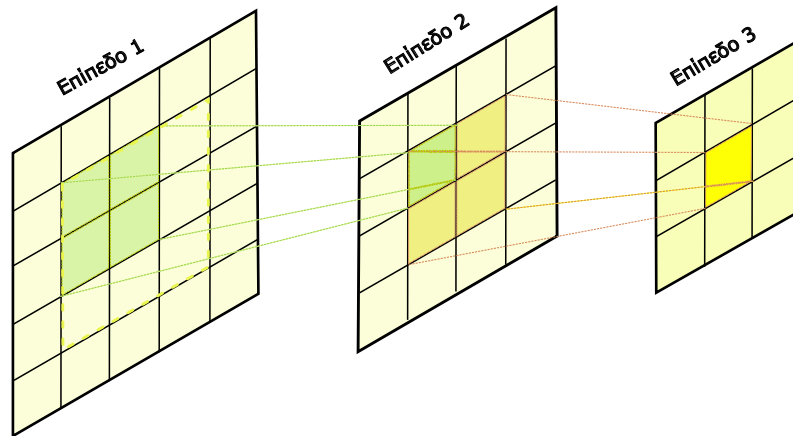
Σχήμα 2.13: Σχήμα όπου απεικονίζεται μια ηλικιωμένη κυρία και μια νεαρή γυναίκα ταυτόχρονα. Ανάλογα με το πιο σύστημα αναφοράς θεωρούμε (προσανατολισμός του κεφαλιού), ο εγκέφαλός μας κάτω από το ίδιο οπτικό ερέθισμα αναγνωρίζει δύο πρόσωπα. *Εγινε λήψη από αυτή την ιστοσελίδα.*

### 2.2.2 Θετικά Γνωρίσματα Συνελικτικών Νευρωνικών Δικτύων

Προτού αναφερθούμε στα μειονεκτήματα των συνελικτικών νευρωνικών δικτύων που θέλουμε να βελτιώσουμε με τα νευρωνικά δίκτυα από κάψουλες, κρίνεται σκόπιμο να αναγνωρίσουμε ορισμένα θετικά στοιχεία τους τα οποία είναι χρήσιμο να κρατήσουμε. Τα βασικά θετικά στοιχεία που έχουμε προαναφέρει συνοπτικά είναι:

- Η αξιοποίηση της χωρικής συνεκτικότητας της εισόδου με τη διατήρηση των σχέσεων απόστασης μεταξύ των χαρακτηριστικών και τη χρήση φίλτρων εξαγωγής χαρακτηριστικών που δρουν τοπικά.
- Η αξιοποίηση της ιεραρχικής δομής των δεδομένων εικόνων με την ενσωμάτωση διαδοχικών συνελικτικών επιπέδων. Ως εκ τούτου, τα φίλτρα των βαθύτερων επιπέδων έχουν μεγαλύτερο οπτικό πεδίο και δύνανται να συνθέσουν πιο σύνθετα χαρακτηριστικά κωδικοποιώντας έτσι πληροφορία ευρύτερου τμήματος της εικόνας εισόδου (βλ. σχήμα 2.14).
- Η ελαχιστοποίηση του υπολογιστικού κόστους (και των απαιτήσεων μνήμης) με την εφαρμογή των τοπικών φίλτρων (δηλαδή όχι πλήρως διασυνδεδεμένων) ως κυλιόμενων παραθύρων στον  $x$  και  $y$  άξονα πάνω στην εικόνα.

Από τα θετικά αυτά δομικά στοιχεία εμμέσως προκύπτει μια πολύ σημαντική ιδιότητα των συνελικτικών δικτύων: αυτή της μεταφοράς των διαχυμάνσεων θέσης των αντικειμένων σε μια εικόνα εισόδου σε κατάλληλες εσωτερικές διαχυμάνσεις των χαρτών χαρακτηριστικών (translation equivariance). Αναλυτικότερα, με τη μετακίνηση ενός αντικειμένου στην εικόνα κατά τον  $x$  ή  $y$  άξονα, λόγω της δισδιάστατης συνέλιξης, αυτή μεταφράζεται σε αντίστοιχη μετακίνηση των εξαχθέντων χαρακτηριστικών. Συνεπώς, θα μπορούσαμε να πούμε ότι ένα συνελικτικό νευρωνικό δίκτυο διαθέτει μηχανισμούς που να μοντελοποιούν τις οριζόντιες και κάθετες μετατοπίσεις της εισόδου ώστε αυτές να γίνονται αντιληπτές από το σύστημα.



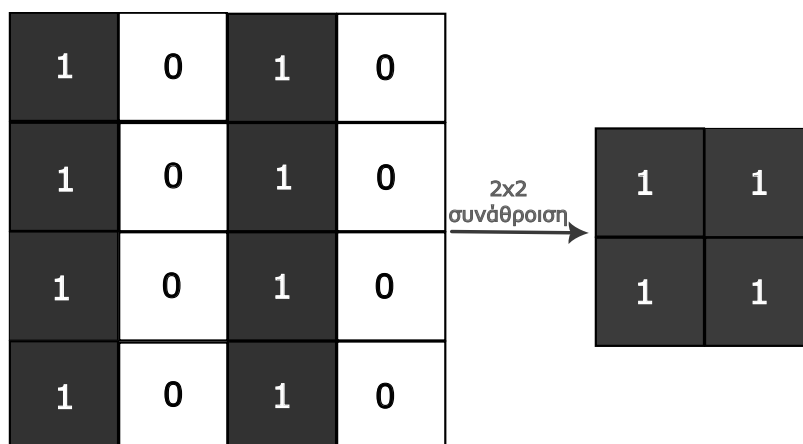
Σχήμα 2.14: Σχήμα τριών διαδοχικών συνελικτικών επιπέδων με μέγεθος φίλτρου  $2 \times 2$  και βήμα 1. Όσο μεγαλύτερη είναι η απόσταση βάθους μεταξύ πρώτου και τελευταίου επιπέδου, τόσο μεγαλύτερο είναι το οπτικό πεδίο από το οποίο εξάγονται τα χαρακτηριστικά του τελικού επιπέδου. Στην εικόνα, το οπτικό πεδίο ενός στοιχείου στο τρίτο επίπεδο χάρτη χαρακτηριστικών σχηματίζει στο πρώτο επίπεδο ένα παραλληλόγραμμο  $3 \times 3$  (απεικονίζεται με διακεκομμένες, κίτρινες γραμμές). Παράχθηκε από το Inkscape.

### 2.2.3 Βασικές Ανεπάρκειες των Συνελικτικών Νευρωνικών Δικτύων

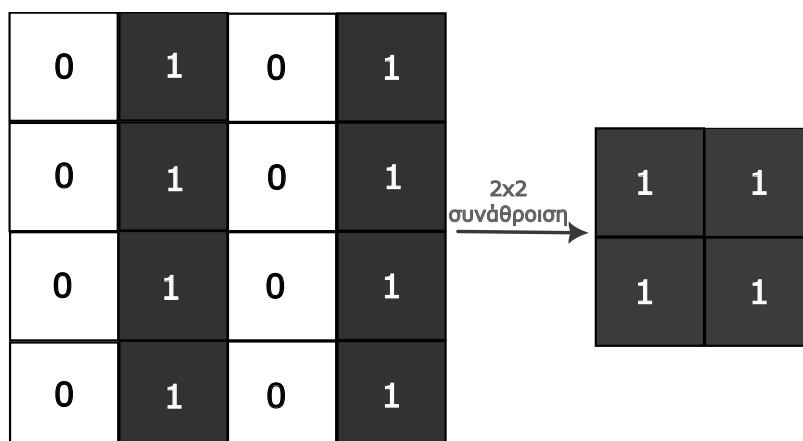
Το βασικό πρόβλημα που αντιμετωπίζουν οι αρχιτεκτονικές συνελικτικών νευρωνικών δικτύων που παρουσιάσαμε είναι η αδυναμία γενίκευσης σε νέες οπτικές γωνίες (novel viewpoints). Με άλλα λόγια, είναι σε θέση να αναγνωρίζουν αντικείμενα μόνο όταν βρίσκονται στον ίδιο προσανατολισμό, κλίμακα, διάτμηση (orientation, scale, shear) κ.τ.λ. με τα στιγμιότυπα αντικειμένων που απεικονίζονται στις εικόνες του συνόλου εκπαίδευσης. Έτσι λοιπόν, οι μόνιμοι αφινικοί μετασχηματισμοί (affine transformations) τους οποίους ένα συνελικτικό νευρωνικό δίκτυο μπορεί να χειριστεί αποδοτικά είναι οι μεταφορές (μεταθέσεις των αντικειμένων της εικόνας) [56].

Για την αναγνώριση αντικειμένων υπό νέες οπτικές γωνίες από τα συνελικτικά νευρωνικά δίκτυα χρησιμοποιούνται μη-αποδοτικές μέθοδοι. Για παράδειγμα, μια μέθοδος είναι ο πολλαπλασιασμός των δεδομένων εισόδου μετά από τυχαία εφαρμογή μετασχηματισμών γνωστή ως «επαύξηση δεδομένων» (data augmentation). Μια άλλη μέθοδος που μπορεί να χρησιμοποιηθεί παράλληλα με την προηγούμενη είναι αυτή της ενσωμάτωσης επιπέδων μέγιστης συνάνθρωσης (max pooling). Όπως έχουμε αναφέρει, τα επίπεδα αυτά αυξάνουν την ευρωστία του συστήματος. Το επιτυγχάνουν, μέσω της υποδειγματοληψίας των χαρτών χαρακτηριστικών έτσι ώστε μικρές μεταβολές στη θέση (ή ακόμα και στον προσανατολισμό [46]) των αντικειμένων να μην αλλάζει τις αποκρίσεις (εξόδους) των φίλτρων των επακόλουθων επιπέδων. Η ιδιότητα αυτή ονομάζεται ανεξαρτησία υπό μεταφορά (translation invariance) και σε αντίθεση με την ιδιότητα των συνελικτικών επιπέδων που περιγράψαμε στην παράγραφο 2.2.2, οι μικρές διακυμάνσεις στην είσοδο του επιπέδου συνάνθρωσης απορρίπτονται και δε μοντελοποιούνται εσωτερικά του συστήματος. Με απλά λόγια, το σύστημα επιδιώκει να πετύχει γενίκευση στους αφινικούς μετασχηματισμούς κάτω από τους οποίους αναγνωρίζει

τα αντικείμενα με το να αχρηστεύει την πληροφορία σχετικά με το συγκεκριμένο στιγμιότυπο εισόδου και να δημιουργεί μια ανεξάρτητη αναπαράσταση (εξαρτώμενη μόνο από το είδος του αντικειμένου) την οποία τα επόμενα επίπεδα θα επεξεργαστούν.



Σχήμα 2.15: Σχήμα όπου εφαρμόζεται μέγιστη συνάθροιση με πυρήνα  $2 \times 2$  και βήμα 2 σε μια δυαδική εικόνα δύο κάθετων ακμών. Παράχθηκε από το *Inkscape*.



Σχήμα 2.16: Σχήμα όπου εφαρμόζεται μέγιστη συνάθροιση με πυρήνα  $2 \times 2$  και βήμα 2 σε μια δυαδική εικόνα δύο κάθετων ακμών, αφού η θέση λήψης μετατοπιστεί. Χάρη στο επίπεδο συνάθροισης, η απόκριση είναι ανεξάρτητη από μικρές μετατοπίσεις της εικόνας εισόδου. Παράχθηκε από το *Inkscape*.

Σύμφωνα με τον G. Hinton [59], τα νευρωνικά δίκτυα θα πρέπει να χειρίζονται όλους τους αφινικούς μετασχηματισμούς με την ίδια λογική που διαχειρίζονται τα συνελικτικά επίπεδα τις κάθετες και οριζόντιες μετατοπίσεις. Δηλαδή, αντί να απορρίπτουν χρήσιμη πληροφορία μέσω των επιπέδων συνάθροισης να διαθέτουν μηχανισμούς που θα μοντελοποιούν εσωτερικά τις διακυμάνσεις στην οπτική γωνία των αντικειμένων. Η πρόταση αυτή βασίζεται στη σημαντική παρατήρηση ότι αλλαγές στη σκοπιά ενός αντικειμένου μεταβάλλουν με σύνθετο, μη γραμμικό τρόπο τα εικονοστοιχεία της εικόνας ενώ τροποποιούν με απλό, γραμμικό τρόπο τη μήτρα πόζας (pose matrix) του αντικειμένου<sup>22</sup>. Συνεπώς, φαίνεται ασύμφορη η προσπάθεια των συνελικτικών

<sup>22</sup>Οι μήτρες πόζας είναι πίνακες που περιγράφουν τη θέση και τον προσανατολισμό ενός αντικειμένου, δύο χαρακτηριστικά τα οποία μεταβάλλονται γραμμικά με την αλλαγή της οπτικής γωνίας θέασης ενός αντικειμένου.

δικτύων να δημιουργούν ανεξάρτητες (υπό την οπτική γωνία) αναπαράστασεις αντικειμένων απευθείας από τον χώρο των εικονοστοιχείων, χωρίς δηλαδή να λαμβάνουν υπόψη τη γραμμική σχέση μεταξύ των διακυμάνσεων της οπτικής γωνίας και των παραμέτρων του στιγμιότυπου (instantiation parameters) του αντικειμένου<sup>23</sup>. Αντίθετα, θα ήταν πιο αποδοτική η μοντελοποίηση αυτής της γραμμικής σχέσης με έναν μηχανισμό ο οποίος θα πραγματοποιούσε ανάστροφα γραφικά (inverse graphics): θα αντιστοιχίζε τον χώρο των εικονοστοιχείων της εικόνας εισόδου σε έναν ιεραρχικό χώρο από μήτρες πόζας για το κάθε απεικονιζόμενο αντικείμενο. Σε αυτήν τη νέα αναπαράσταση, οι αφινικοί μετασχηματισμοί θα άλλαζαν με προβλέψιμο τρόπο τις —απεπλεγμένες από το είδος του αντικειμένου— παραμέτρους των επιμέρους στιγμιότυπων οδηγώντας στην επιθυμητή γενίκευση σε νέες οπτικές γωνίες.

Επιπρόσθετα, ο παρόν τρόπος διαχείρισης αφινικών μετασχηματισμών από τα συνελικτικά νευρωνικά δίκτυα τα καθιστά επιρρεπή σε αντιπαραθετική επίθεση (adversarial attacks). Αυτή τους η αδυναμία, θα μπορούσε να καταπολεμηθεί με την ενσωμάτωση ενός μηχανισμού που θα μοντελοποιούσε τις σχέσεις μεταξύ των τμημάτων ενός αντικειμένου ούτως ώστε, για την αναγνώρισή του, να λαμβάνονταν υπόψη η γεωμετρία των επιμέρους μερών του. Με άλλα λόγια, αν υπήρχε «αποθηκευμένη» στο νευρωνικό δίκτυο η πληροφορία για τον τρόπο σύνδεσης των στοιχείων που απαρτίζουν ένα αντικείμενο τότε θα ήταν περισσότερο εύρωστο σε αυτού του είδους τις επιθέσεις. Στο παράδειγμα του σχήματος 2.17 το συνελικτικό νευρωνικό δίκτυο αφενός δε διαθέτει κάποιο μηχανισμό να αναγνωρίζει την ακριβή θέση του ματιού στην εικόνα (αφού αυτή η πληροφορία απορρίπτεται σε ένα βαθμό μέσω των επιπέδων συνάθροισης) και αφετέρου, ακόμα και αν ήταν διαθέσιμη αυτή η πληροφορία, θα έμενε αναξιοποίητη διότι δεν αποθηκεύεται η γνώση για το ποια θα πρέπει να είναι η θέση του ματιού σε σχέση με τα υπόλοιπα μέρη.

Ένα τελευταίο σημείο αδυναμίας των κλασικών νευρωνικών δικτύων είναι το λεγόμενο πρόβλημα της αποκλειστικής διάζευξης (XOR problem) [60]. Αυτό, προκύπτει από την παρατήρηση ότι η συνάρτηση της αποκλειστικής διάζευξης δεν μπορεί να υλοποιηθεί από έναν μεμονωμένο τεχνητό νευρώνα. Σύμφωνα με την περιγραφή του τεχνητού νευρώνα του σχήματος 2.4, στον πυρήνα του προβλήματος βρίσκεται το γεγονός ότι δεν υπάρχει δυνατότητα σύγκρισης των εισόδων μεταξύ τους. Αντ' αυτού, πραγματοποιείται σύγκριση μεταξύ ενός διανύσματος εισόδων με ένα διάνυσμα (αποθηκευμένων) βαρών. Αυτή η αδυναμία όμως οδηγεί σε μη-αποδοτικές λύσεις του προβλήματος (με την προσθήκη κρυφών επιπέδων). Επαχόλουθη, λοιπόν, είναι η διάθεση πειραματισμού με ένα νέο είδος τεχνητού νευρώνα που θα μπορεί να συγκρίνει τις εισόδους (ή τα διανύσματα εισόδων) επιλύοντας το πρόβλημα της αποκλειστικής διάζευξης και κυρίως επιτρέποντας τον εντοπισμό συνδιακυμάνσεων στα χαρακτηριστικά εισόδου.

#### 2.2.4 Αρχές Λειτουργίας Νευρωνικών Δικτύων με Κάψουλες

Στην προσπάθεια αντιμετώπισης των ανωτέρω μειονεκτημάτων των συνελικτικών νευρωνικών δικτύων σε εργασίες αυτόματης αναγνώρισης αντικειμένων, αναπτύχθηκαν τα νευρωνικά δίκτυα

Χρησιμοποιούνται κατά κόρον στον χώρο των γραφικών με υπολογιστή για την περιγραφή του τρόπου τοποθέτησης των αντικειμένων σε έναν εικονικό κόσμο.

<sup>23</sup>Με τον όρο «παράμετροι στιγμιότυπου» θα αναφερόμαστε κυρίως στην μήτρα πόζας του στιγμιότυπου. Παρόλα αυτά, παράμετροι στιγμιότυπου είναι και άλλοι παράγοντες που δεν εξαρτώνται από την κλάση του αντικειμένου προς αναγνώριση όπως ο φωτισμός, το μέγεθος ή χρώμα του αντικειμένου κ.τ.λ.



Σχήμα 2.17: Σχήμα όπου απεικονίζεται το λεγόμενο πρόβλημα του Picasso στο οποίο η εικόνα έχει όλα τα σωστά μέρη αλλά οι σχέσεις μεταξύ τους είναι λάθος. Ένα υποθετικό συνελικτικό νευρωνικό δίκτυο θα δυσκολεύονταν να αντιληφθεί ότι το σχήμα της εικόνας δεν είναι κανονικό πρόσωπο. (Το παράδειγμα είναι ενδεικτικό αφού δεν έχει αποδειχθεί ότι ένα συνελικτικό δίκτυο θα αναγνώριζε το συγκεκριμένο παράδειγμα ως πρόσωπο.) *Παράχθηκε από το Inkscape τροποποιώντας αυτή την εικόνα.*

με κάψουλες. Αυτά διαμορφώθηκαν:

- αντλώντας στοιχεία από την επιστήμη της νευροφυσιολογίας
- διατηρώντας αρκετά θετικά σημεία των συνελικτικών νευρωνικών δικτύων (π.χ. συνέλιξη)
- εισάγοντας τις ιεραρχικές δομές νευρώνων (κάψουλες) και έναν μηχανισμό για τη διασύνδεσή τους.

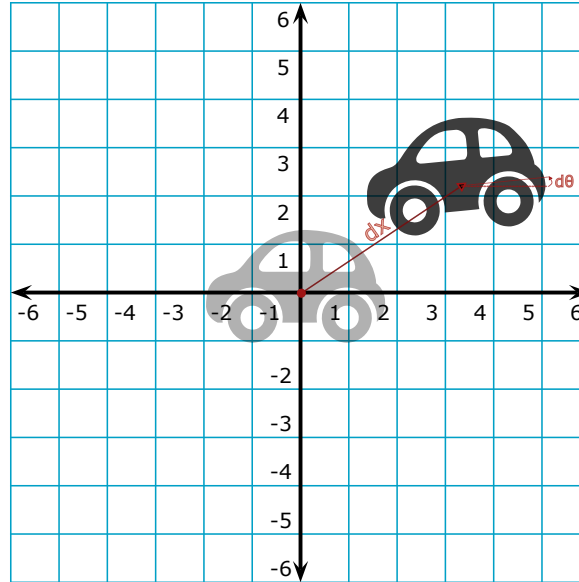
### Ορισμός Κάψουλας και Αρχές Λειτουργίας της

Η κάψουλα δεν είναι τίποτα άλλο παρά μια ομάδα από νευρώνες. Κάθε κάψουλα έχει μια και μοναδική λειτουργία μέσα σε ένα νευρωνικό δίκτυο. Πιο συγκεκριμένα, κάθε μια εκπαιδεύεται να αναπαριστά την πιθανότητα παρουσίας και τις παραμέτρους στιγμιοτύπου μιας συγκεκριμένης οντότητας στο πεδίο υποδοχής της. Για παράδειγμα, σε μια εφαρμογή οπτικής αναγνώρισης του τύπου τροχοφόρων οχημάτων, μια κάψουλα θα μπορούσε να είναι επιφορτισμένη με την αναπαράσταση της οντότητας «ρόδα». Σε περίπτωση που η οντότητα ήταν παρούσα στο πεδίο υποδοχής της αντίστοιχης κάψουλας τότε θα είχε ως έξοδο μεγάλη τιμή πιθανότητας παρουσίας και φυσικά τις τιμές που θα προσδιόριζαν τη θέση, τον προσανατολισμό, το μέγεθος κ.τ.λ. της αναγνωρισμένης ρόδας. Στο σημείο αυτό γίνεται αντιληπτό το πλεονέκτημα της χρήσης συστάδων από νευρώνες έναντι μεμονωμένων νευρώνων (όπως αυτός του σχήματος 2.4 που παράγει μια μονοδιάστατη τιμή εξόδου ή ενεργοποίησης) αφού έτσι γίνεται εφικτή η έξοδος πιο εκφραστικών, πολυδιάστατων αναπαραστάσεων.

Όπως είναι φυσικό, για την παραγωγή των παραμέτρων στιγμιοτύπου μιας οντότητας είναι απαραίτητη η διατήρηση ενός συστήματος αναφοράς για τη συγκεκριμένη οντότητα<sup>24</sup> (βλ. σχήμα

<sup>24</sup>Η αναγκαιότητα ύπαρξης συστήματος αναφοράς της οντότητας γίνεται ακόμα πιο προφανής αν αναλογιστεί

2.18). Αυτό είναι σύμφωνο με τις παρατηρήσεις λειτουργίας του ανθρώπινου οπτικού φλοιού που, όπως διατυπώσαμε, εφαρμόζει συστήματα αναφοράς σε κάθε οπτικό ερέθισμα. Το είδος και η γεωμετρία της οντότητας αναφοράς σχηματίζεται κατά τη διαδικασία εκμάθησης του νευρωνικού δικτύου.



Σχήμα 2.18: Σχήμα όπου με τη βοήθεια μιας οντότητας αναφοράς (απεικονίζεται ως αχνό αυτοκίνητο) και του καρτεσιανού συστήματος συντεταγμένων υπολογίζονται οι παράμετροι στιγμιότυπου (απόσταση από αρχή αξόνων και γωνία περιστροφής). Παράχθηκε από το Inkscape.

Με μαθηματικούς όρους, κάθε κάψουλα  $c_i$  αποτελείται από ένα διάνυσμα  $m_i \in \mathbb{R}^d$  ή πίνακα  $M_i \in \mathbb{R}^{d \times d}$  από παραμέτρους στιγμιότυπου και μια τιμή πιθανότητας παρουσίας  $a_i \in [0, 1]$ . Η εξασφάλιση ότι η τιμή πιθανότητας ανήκει στο διάστημα  $[0, 1]$  γίνεται με την εφαρμογή μιας μη γραμμικής συνάρτησης ενεργοποίησης με σύνολο εξόδου το διάστημα αυτό. Σε ορισμένες υλοποιήσεις, η τιμή πιθανότητας παρουσίας της οντότητας που αναπαριστά η κάψουλα  $c_i$  κωδικοποιείται στο μήκος του διανύσματος  $m_i$ .

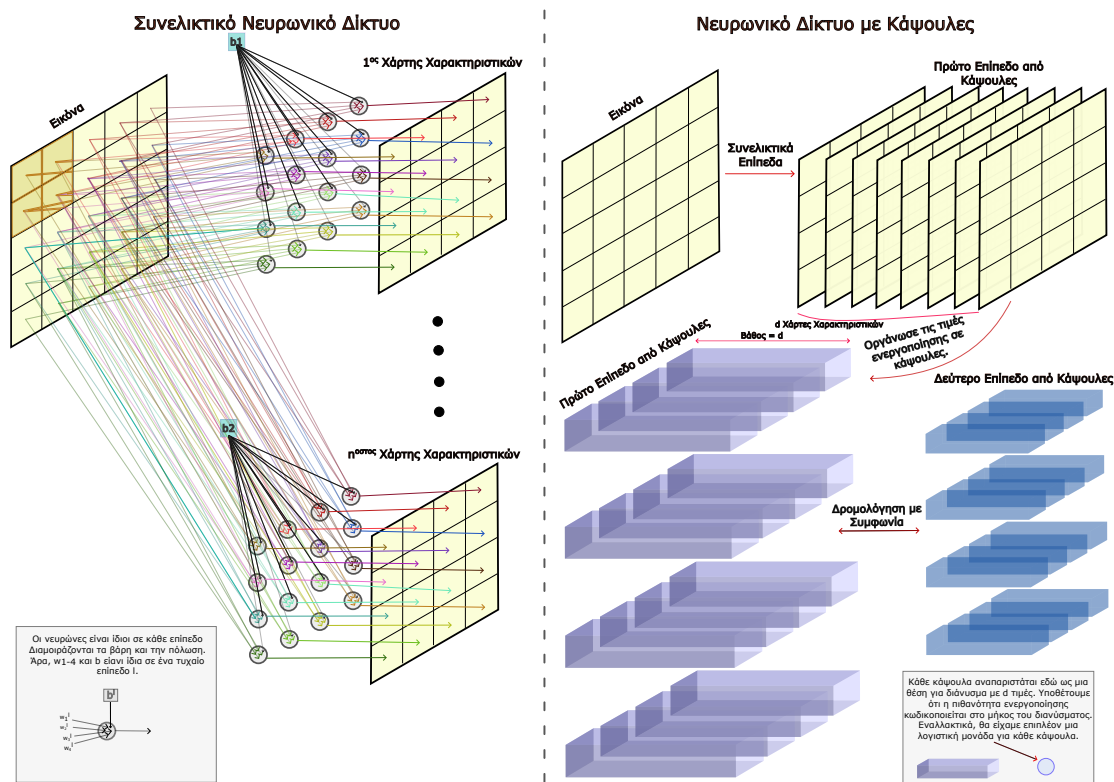
### Οργάνωση των Καψουλών στο Νευρωνικό Δίκτυο

Τα νευρωνικά δίκτυα από κάψουλες αξιοποιούν την τοπική χωρική συνεκτικότητα των εικόνων αφού οι κάψουλες οργάνωνονται σε τρισδιάστατες δομές. Επίσης, αξιοποιούν την ιεραρχική δομή των φυσικών αντικειμένων με το να διατηρούν την πολυεπίπεδη διάταξη που χαρακτηρίζει τα συνελικτικά νευρωνικά δίκτυα (βλ. σχήμα 2.20). Η διαφορά έγκειται στο γεγονός ότι αντί για επίπεδα από χάρτες χαρακτηριστικών υπάρχουν επίπεδα από κάψουλες (συμβολίζονται ως  $C^{[l]}$ )<sup>25</sup>. Τα επίπεδα αυτά μπορεί να είναι πλήρως διασυνδεδεμένα ή να είναι συνελικτικά. Για παράδειγμα, στην περίπτωση των συνελικτικών επιπέδων από κάψουλες, αντί για παραγωγή

κανείς ότι οι παράμετροι στιγμιότυπου ουσιαστικά περιγράφουν τη σχέση μεταξύ μιας οντότητας υπό την οπτική γωνία λήψης της εικόνας και της αντίστοιχης οντότητας αναφοράς.

<sup>25</sup>Ιδιαίτερη προσοχή απαιτείται καθώς οι κάψουλες, αποτελούμενες από ένα σύνολο χαρακτηριστικών της οντότητας που αναγνωρίζουν, μπορούν να αντιπαραβληθούν με τους χάρτες χαρακτηριστικών (δηλαδή τον πίνακα τιμών ενεργοποίησης  $A$ ). Θα ήταν λάθος λοιπόν να αντιπαραβληθούν με νευρώνες που έχουν αποθηκευμένα βάρη (βλ. σχήμα 2.19).

χαρτών χαρακτηριστικών με μοναδιαίο βάθος ο καθένας, θα λέγαμε ότι παράγονται πλούσιοι σε πληροφορία χάρτες από διανύσματα ή πίνακες (πολυδιάστατες αναπαραστάσεις των αναγνωριζόμενων μοτίβων-χαρακτηριστικών) που αποτελούν και το επόμενο επίπεδο από κάψουλες. Η παραγωγή του επόμενου επιπέδου από κάψουλες με βάση το προηγούμενο δε γίνεται με τη χρήση επιπέδων από νευρώνες αλλά πραγματοποιείται μέσω ενός αλγορίθμου «δρομολόγησης μέσω συμφωνίας» ο οποίος θα αναλυθεί στη συνέχεια.



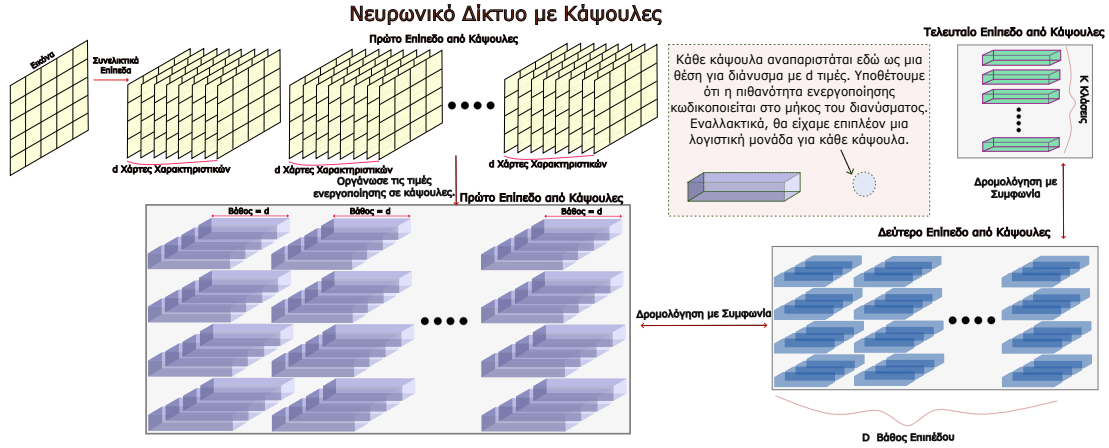
Σχήμα 2.19: Δομικές διαφορές μεταξύ συνελικτικών νευρωνικών δικτύων και δικτύων με κάψουλες. Παρατηρούμε ότι οι κάψουλες δεν έχουν ομοιότητες με τους κλασικούς τεχνητούς νευρώνες. Κυρίως μπορούν να συσχετιστούν με σύνολα από χάρτες χαρακτηριστικών. *Παράχθηκε από το Inkscape.*

## Αλγόριθμος Δρομολόγησης με Συμφωνία

Θα θέλαμε να εστιάσουμε στους υπολογισμούς που λαμβάνουν χώρα μεταξύ δύο διαδοχικών επιπέδων από κάψουλες. Αν και κάθε υλοποίηση είναι διαφορετική, θα παρουσιάσουμε υπό μια αφαιρετική σκοπιά τις βασικές αρχές που διέπουν τον κάθε αλγόριθμο δρομολόγησης. Επειδή σε ένα νευρωνικό δίκτυο από κάψουλες τα πρώτα επίπεδά του επιτελούν ιδιαίτερες λειτουργίες, ας υποθέσουμε χωρίς βλάβη της γενικότητας ότι το έχουμε τροφοδοτήσει με μια εικόνα και διαδοχικά έχουν υπολογιστεί ήδη οι τιμές των επιπέδων του μέχρι και το  $l - 1$ . Θα επιθυμούσαμε στο σημείο αυτό να υπολογίσουμε τις τιμές για το επίπεδο  $l$ .

Με τους μέχρι τώρα υπολογισμούς, έχουμε στη διάθεσή μας μια τρισδιάστατη δομή από κάψουλες όπως φαίνεται στο σχήμα 2.20 σε μώβ χρώμα. Κάθε κάψουλα περιέχει ένα σύνολο από τιμές που περιγράφουν την πόζα του τμήματος του αντικειμένου με το οποίο έχουν ταυτιστεί και





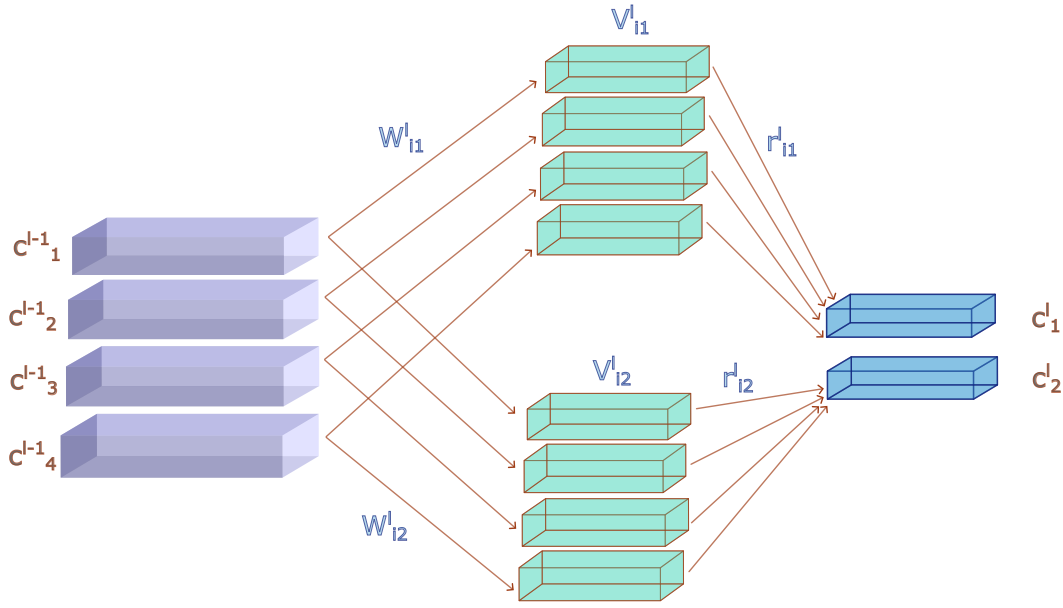
Σχήμα 2.20: Σχήμα όπου φαίνεται μια τυπική οργάνωση ενός νευρωνικού δικτύου με κάψουλες. Παράχθηκε από το *Inkscape*.

αναγνωρίζουν. Επίσης, διαθέτουν και μια τιμή ενεργοποίησης που περιγράφει την πιθανότητα αυτό το τμήμα να είναι παρόν στο οπτικό πεδίο της κάψουλας. Κάψουλες με χαμηλή τιμή ενεργοποίησης θα λέμε ότι είναι ανενεργές και δε θα έχουν μεγάλη βαρύτητα στον υπολογισμό των τιμών του επόμενου επιπέδου.

Όπως είναι γνωστό, στο επίπεδο  $l$  οι κάψουλες αναπαριστούν ανώτερες ιεραρχικά οντότητες σε σχέση με αυτές του επιπέδου  $l-1$ . Συνεπώς, η διαμόρφωση των τιμών του επιπέδου  $l$  ανάγεται στο πρόβλημα της αντιστοίχισης (δρομολόγησης) επιμέρους τμημάτων των αντικειμένων σε μια εικόνα (αναπαριστώνται με κάψουλες  $C[l-1]$ ) στα γενικότερα αντικείμενα που τα περιέχουν (αναπαριστώνται με τις κάψουλες  $C[l]$ ). Αυτή η δρομολόγηση προϋποθέτει την ύπαρξη ενός μηχανισμού που θα προβλέπει τις παραμέτρους που περιγράφουν τις γενικότερες οντότητες που αναγνωρίζουν οι κάψουλες  $C[l]$  με βάση τις παραμέτρους στιγμιοτύπων των επιμέρους οντοτήτων που αναγνωρίζουν οι κάψουλες  $C[l-1]$ <sup>26</sup>. Ο μηχανισμός υλοποιείται με πίνακες από βάρη  $W[l]$  που αναλαμβάνουν να αποθηκεύσουν τις σχέσεις μέρους-όλου (part-whole relationships), δηλαδή τις σχέσεις όλων των δυνατών ζευγών μεταξύ των κάψουλων επιπέδου  $l-1$  και  $l$ . Έτσι αν στο επίπεδο  $l-1$  έχουμε  $n^{[l-1]}$  κάψουλες και στο επίπεδο  $l$ ,  $n^{[l]}$  τότε θα υπάρχουν  $n^{[l-1]} \times n^{[l]}$  πίνακες βαρών μεταξύ των δύο επιπέδων. Μαθηματικά, η πρόβλεψη (ή ψήφο) για μια κάψουλα  $c_j^{[l]}$  με βάση την κάψουλα  $c_i^{[l-1]}$  παράγεται πολλαπλασιάζοντας τον πίνακα βαρών  $W_{ij}^{[l]}$  με το διάνυσμα ή πίνακα παραμέτρων στιγμιότυπου  $M_i^{[l-1]}$ , δηλαδή  $V_{ij}^{[l]} = M_i^{[l-1]} \times W_{ij}^{[l]}$ . Να σημειώσουμε ότι τα βάρη  $W$  μαθαίνονται κατά την εκπαίδευση με τον αλγόριθμο της οπισθοδιάδοσης.

Μέχρι τώρα, έχουμε αναφέρει τον τρόπο με τον οποίο η κάθε μια κάψουλα  $c^{[l]}$  διαθέτει  $n^{[l-1]}$  προβλέψεις (μια από κάθε κάψουλα  $c^{[l-1]}$ ) για το ποιες εκτιμά ότι είναι οι παράμετροι στιγμιότυπου της  $M^{[l]}$ . Δεν έχουμε περιγράψει όμως τον τρόπο με τον οποίο αυτές οι προβλέψεις συγκροτούνται για την τελική διαμόρφωση των  $M^{[l]}$  και  $a^{[l]}$ . Αυτό μας οδηγεί στην έννοια του φιλτραρίσματος μέσω της πολυδιάστατης σύμπτωσης (high dimensional coincidence filtering). Σύμφωνα με την έννοια αυτή, όταν μερικές από τις ψήφους των  $C^{[l-1]}$  συμπίσουν σε μια γειτονιά

<sup>26</sup>π.χ. με βάση τη γεωμετρία της επιμέρους οντότητας «χέρι» να προβλέπεται η γεωμετρία της οντότητας «άνδρας».



Σχήμα 2.21: Τρόπος παραγωγής των ψήφων για την απλή περίπτωση όπου έχουμε τέσσερις κάψουλες στο επίπεδο  $l - 1$  και δύο στο επίπεδο  $l$ . Στο σχήμα,  $i \in [1, 4]$ . Επίσης, η τιμή ενεργοποίησης θεωρήθηκε ότι κωδικοποιείται στο μήκος των διανυσμάτων από παραμέτρους. Παράχθηκε από το *Inkscape*.

του πολυδιάστατου χώρου των αναπαραστάσεων  $\mathbb{R}^d$  τότε αυτή η σύμπτωση δεν μπορεί να είναι τυχαία. Αντιθέτως, το γεγονός ότι μεγάλος αριθμός από κάψουλες  $C^{[l-1]}$  συμφωνούν στο ποιες θα είναι οι τιμές παραμέτρων της εκάστοτε  $c^{[l]}$  σημαίνει ότι πιθανότατα, αυτές είναι οι κατάλληλες τιμές της (και φυσικά ότι η οντότητα που εκπροσωπεί υπάρχει στην εικόνα). Έτσι η διαδικασία δρομολόγησης τελικά ανάγεται σε πρόβλημα εύρεσης συστάδων από ψήφους (clusters of votes) στον χώρο  $\mathbb{R}^d$ .

Με βάση τα ανωτέρω, θα ήταν δυνατή η δρομολόγηση ενός τμήματος μιας οντότητας που αναπαριστάται από μια κάψουλα  $c^{[l-1]}$  σε όλες τις κάψουλες  $C^{[l]}$ . Διαισθητικά, κάτι τέτοιο δεν είναι ορθό αφού ένα τμήμα ενός αντικειμένου δεν είναι δυνατό να αποτελεί μέρος όλων των αντικειμένων που αναπαριστώνται από τις κάψουλες του επόμενου επιπέδου. Επιπλέον, θα δημιουργούσε σύγχυση στον χώρο αναπαραστάσεων κατακλύζοντάς τον με περιττή πληροφορία. Συνεπώς, εισάγουμε τον περιορισμό ότι κάθε κάψουλα μπορεί να δρομολογεί τελικά την ψήφο της μόνο σε μια κάψουλα του επόμενου επιπέδου (single parent assumption). Με αυτόν τον τρόπο προκαλούμε ανταγωνισμό μεταξύ των  $C^{[l]}$  να «εξηγήσουν» όσο το δυνατόν περισσότερες ψήφους των  $C^{[l-1]}$ <sup>27</sup>.

<sup>27</sup> Χρησιμοποιούμε τον όρο «εξηγήσουν» διότι το σύνολο παραμέτρων  $M^{[l]}$  που δημιουργείται για μια κάψουλα  $c^{[l]}$  κατά μια έννοια εκφράζει τις απόψεις των  $C^{[l-1]}$  που την ψήφισαν αναφορικά με το ποιο είναι το στιγμιότυπο του αντικειμένου που αναπαριστά.

Η κάθε κάψουλα  $c^{[l-1]}$  είναι αδύνατο να γνωρίζει εκ των προτέρων (a priori) ποια κάψουλα του επόμενου επιπέδου (κάψουλα πατέρας) θα την εκφράζει καλύτερα αφού αυτό εξαρτάται όπως είπαμε από το αν η ψήφος της συμπίπτει μαζί με άλλες ψήφους στον χώρο αναπαράστασεων (φιλτράρισμα πολυδιάστατης σύμπτωσης). Μονόδρομος λοιπόν είναι η επαναληπτική φύση του αλγορίθμου με συμφωνία κατά την οποία αρχικά οι ψήφοι της κάθε κάψουλας  $c^{[l-1]}$ , βεβαρημένες υπό μια ομοιόμορφη κατανομή διακριτής πιθανότητας, δρομολογούνται σε όλες τις  $C^{[l]}$ . Στη συνέχεια, μέσω του φιλτραρίσματος πολυδιάστατης σύμπτωσης, κάθε κάψουλα  $c^{[l]}$  ανταγωνίζεται για να προσαρτήσει κάψουλες  $c^{[l-1]}$  των οποίων οι ψήφοι σχηματίζουν συστάδες και άρα μπορεί εύκολα να «εξηγήσει»<sup>28</sup>. Χωρίς βλάβη της γενικότητας, κάθε κάψουλα  $c^{[l]}$  «εξηγεί» τις ψήφους προσαρμόζοντας τον πίνακα (ή το διάνυσμα)  $M^{[l]}$  στο κέντρο βάρους των ψήφων. Επιπρόσθετα, προσαρμόζει την τιμή πιθανότητας ενεργοποίησής της ανάλογα με το πόσο καλά εξηγεί τις ψήφους<sup>29</sup>. Αυτές οι δύο προσαρμογές ανατροφοδοτούνται πίσω στις κάψουλες  $c^{[l-1]}$  οι οποίες αλλάζουν, η κάθε μία, την κατανομή των βαρών υπό τα οποία δρομολογούν τις ψήφους τους (coupling coefficients) έτσι ώστε να προτιμούν κάψουλες γονείς που είναι ενεργές και εκφράζουν καλύτερα την ψήφο τους (το διάνυσμα της ψήφου τους είναι πιο κοντά στο διάνυσμα  $M^{[l]}$ ). Όσο εξελίσσονται οι επαναλήψεις, τόσο οι κάψουλες  $c^{[l-1]}$  είναι πιο σίγουρες για το που θα αποστείλουν τις ψήφους τους (η ομοιόμορφη κατανομή εκφυλίζεται σε ένα σημείο-κάψουλα) και οι πίνακες  $M^{[l]}$  συγχλίνουν στο κέντρο των συστάδων από ψήφους.

Συγκεντρωτικά, μπορούμε να παρουσιάσουμε έναν αφαιρετικό αλγόριθμο δρομολόγησης με συμφωνία μεταξύ δύο διαδοχικών επιπέδων από κάψουλες  $l-1$  και  $l$  ως εξής:

1. Για κάθε κάψουλα  $c_i^{[l-1]} \in C^{[l-1]}$  υπολόγισε τις ψήφους-προβλέψεις ως  $V_{ij} = M_i \times W_{ij}$
2. Αρχικοποίησε τις τιμές βαρών δρομολόγησης  $r_{ij} = 1/n^{[l]}, \forall i \in [1, n^{[l-1]}], \forall j \in [1, n^{[l]}]$  έτσι ώστε  $\sum_{j=1}^{n^{[l]}} r_{ij} = 1$ .
3. Επανάλαβε για προκαθορισμένο αριθμό επαναλήψεων:
  - (α') Για καθεμία κάψουλα  $c_j^{[l]}$  εντόπισε τις συστάδες από βεβαρημένες ψήφους στον χώρο αναπαράστασης  $\mathcal{R}^d$  και με βάση αυτές υπολόγισε τις παραμέτρους στιγμιοτύπου  $M_j^{[l]}$  που θα τις εκφράζει,  $\forall j \in [1, n^{[l]}]$ .
  - (β') Υπολόγισε για καθεμία κάψουλα  $c_j^{[l]}$  την τιμή ενεργοποίησής της με βάση το πόσο καλά εξηγεί τα δεδομένα.
  - (γ') Για κάθε κάψουλα  $c_i^{[l-1]}$  ενημέρωσε τα βάρη δρομολόγησης  $r_{ij}, \forall j \in [1, n^{[l]}]$  ώστε να δίνεται μεγαλύτερο βάρος σε κάψουλες γονείς των οποίων οι παράμετροι στιγμιοτύπου εξηγούν καλύτερα τις ψήφους.
4. Τερμάτισε με έξοδο  $M^{[l]}$  και  $a^{[l]}$ .

<sup>28</sup>Στον σχηματισμό συστάδων βοηθάει το γεγονός ότι οι κάψουλες του συνόλου  $C^{[l]}$  βλέπουν διαφορετικές ψήφους της κάθε κάψουλας του συνόλου  $C^{[l-1]}$  αφού παράγονται μετά από πολλαπλασιασμό με διαφορετικό πίνακα βαρών. Έτσι, κάψουλες οι οποίες σχηματίζουν συστάδες από ψήφους στον χώρο αναπαράστασης μιας κάψουλας  $c^{[l]}$  μπορεί στον χώρο αναπαράστασης μιας διαφορετικής κάψουλας να σχηματίζουν απόμακρες προβλέψεις.

<sup>29</sup>Ο μηχανισμός υπολογισμού πιθανότητας ενεργοποίησης  $a^{[l]}$  διαφέρει σημαντικά από υλοποίηση σε υλοποίηση αλλά γενικά είναι ανάλογο του αριθμού από κάψουλες  $C^{[l-1]}$  που προτιμούν την κάψουλα  $c^{[l]}$  και της πυκνότητας των ψήφων τους.

Ας φανταστούμε τώρα ότι έχουμε πολλά διαδοχικά επίπεδα από κάψουλες. Με τον αλγόριθμο δρομολόγησης (και λόγω της υπόθεσης μοναδικού πατέρα), δημιουργείται δυναμικά κατά την πρόσθια τροφοδότηση του δικτύου ένα ιεραρχικό δέντρο από ενεργές κάψουλες όπου η κάθε μια αναπαριστά τις οντότητες που βρίσκονται στην εικόνα. Ο σχηματισμός του ιεραρχικού δέντρου θα μπορούσε να παρομοιαστεί με το σκάλισμα ενός γλυπτού από ένα κομμάτι μαρμάρου [56]. Το μάρμαρο είναι όλες οι κάψουλες σε κάθε επίπεδο ενώ η διαδικασία σκαλίσματος πραγματοποιείται από τον αλγόριθμο δρομολόγησης με συμφωνία που επιλεκτικά συνδέει και ενεργοποιεί ορισμένες κάψουλες. Επειδή η κάθε κάψουλα αναπαριστά όχι μόνο την πιθανότητα ύπαρξης της οντότητας αλλά και τις παραμέτρους στιγμιοτύπου της, μπορούμε να υποθέσουμε (χρησιμοποιώντας την ορολογία της προηγούμενης υποενότητας) ότι η ιεραρχική δομή είναι εμπλουτισμένη. Με άλλα λόγια, η δενδροειδής δομή από τις ενεργές κάψουλες που σχηματίζεται δυναμικά κατά την πορόσθια τροφοδότηση δε μοντελοποιεί μόνο την ιεραρχία μεταξύ ενός αντικειμένου και των μερών του (με σχέσεις τύπου κόμβοι γονέων  $\supseteq$  κόμβοι παιδιών) αλλά και τις γεωμετρικές σχέσεις μεταξύ αυτών (π.χ. σε ποια θέση τοποθετούνται τα επιμέρους τμήματα για να συνθέσουν το όλο).

### Πρώτα Επίπεδα ενός Νευρωνικού Δικτύου με Κάψουλες

Τα πρώτα επίπεδα ενός νευρωνικού δικτύου, όπως προκύπτει από τη μέχρι τώρα ανάλυση, είναι επιφορτισμένα με τον μετασχηματισμό από τον χώρο των εικονοστοιχείων στον χώρο αναπαράστασης των παραμέτρων στιγμιοτύπων<sup>30</sup> ώστε να μπορούν μετά σε αυτό να δράσουν οι κάψουλες. Πρακτικά, πρόκειται για συνελικτικά επίπεδα από κλασσικούς νευρώνες που παράγουν χάρτες χαρακτηριστικών που αποτελούνται από βαθμωτά μεγέθη. Στη συνέχεια αυτά τα βαθμωτά μεγέθη ομαδοποιούνται σε διανύσματα ή πίνακες ούτως ώστε κάθε ομάδα να ενθυλακώνει τις παραμέτρους μιας κάψουλας. Μέσω εκπαίδευσης, τα πρώτα επίπεδα μαθαίνουν να πραγματοποιούν ανάστροφα γραφικά (derendering)<sup>31</sup> και να παράγουν αυτό που ονομάζουμε «αρχείο σκηνής» (βλ. παράρτημα Α'). Αυτή τη λειτουργία δε διαθέτουν τα κλασσικά συνελικτικά νευρωνικά δίκτυα διότι οφείλεται στους δομικούς και λειτουργικούς περιορισμούς που επιβάλλει το νευρωνικό δίκτυο από κάψουλες.

### Τελευταίο Επίπεδο ενός Νευρωνικού Δικτύου με Κάψουλες

Συνήθως, το τελευταίο επίπεδο ενός τέτοιου δικτύου είναι ένα επίπεδο από κάψουλες (όπως φαίνεται στο σχήμα 2.20). Στις περισσότερες περιπτώσεις για εφαρμογές ταξινόμησης, υπάρχει μία κάψουλα ανά κατηγορία. Η πρόβλεψη  $\hat{y}$  του νευρωνικού δικτύου λαμβάνεται ως η οντότητα που εκπροσωπείται από την κάψουλα του τελευταίου επιπέδου που έχει την μεγαλύτερη τιμή ενεργοποίησης (για την συγκεκριμένη είσοδο).

### Πως τα Νευρωνικά Δίκτυα με Κάψουλες Γενικεύουν σε Νέες Οπτικές Γωνίες

Ο κύριος λόγος της αποδοτικής γενίκευσης των νευρωνικών δικτύων με κάψουλες αποδίδεται στο ότι εργάζονται στον χώρο παραμέτρων των στιγμιοτύπων μιας εικόνας (ονομάζεται και χώρος αναπαράστασης γραφικών) όπου οι αλλαγές στην οπτική γωνία προκαλούν γραμμικές μεταβολές

<sup>30</sup>Υπενθυμίζουμε ότι στον χώρο αυτό, οι αλλαγές στην οπτική γωνία προκαλούν γραμμικές μεταβολές στις παραμέτρους (στα χαρακτηριστικά).

<sup>31</sup>Ουσιαστικά πραγματοποιούν μη παραμετρικό μετασχηματισμό Hough.

στις παραμέτρους. Έχοντας διαχωρίσει τους παράγοντες διακύμανσης της κάθε οντότητας (την πιθανότητα ύπαρξης από τις παραμέτρους στιγμιοτύπου της), και έχοντας μεταβεί σε έναν χώρο όπου αλλαγές στην οπτική γωνία αλλάζουν με γραμμικό τρόπο τα χαρακτηριστικά στιγμιοτύπου, η γενίκευση σε νέες οπτικές γωνίες έγκειται απλά στη γραμμική παρεμβολή των χαρακτηριστικών αυτών. Έτσι, στο απλό παράδειγμα που το δίκτυο έχει εκπαιδευτεί να αναγνωρίζει ένα ψηφίο στραμμένο με τυχαίο τρόπο κατά  $\theta^\circ \in [-5, +5]$  τότε θα διαθέτει κάποια κάψουλα  $c_i^{[L]}$  που αναγνωρίζει την ύπαρξη αυτού του ψηφίου με πιθανότητα  $a_i^{[L]}$  και κωδικοποιεί τον προσανατολισμό του (μεταξύ άλλων χαρακτηριστικών) στον πίνακα  $M_i^{[L]}$ . Έτσι, μπορεί εύκολα να προβλέψει μέσω παρεμβολής τι επίδραση θα έχει η στρέψη του ψηφίου κατά  $+10^\circ$ .<sup>32</sup>

Για να επιτευχθεί γενίκευση σε νέες οπτικές γωνίες, όλες οι σχεδιαστικές αποφάσεις των νευρωνικών δικτύων με κάψουλες εξυπηρετούν έμμεσα ή άμεσα την εσωτερική μοντελοποίηση των αλλαγών στις διακυμάνσεις της. Για παράδειγμα, δε χρησιμοποιούνται επίπεδα συνάντησης καθώς αυτά όπως έχουμε αναφέρει οδηγούν σε αναπαραστάσεις, ανεξάρτητες της οπτικής γωνίας. Επιπλέον, χρησιμοποιούν κάψουλες και όχι χάρτες από βαθμωτά χαρακτηριστικά καθώς μέσω των πρώτων μπορούν να αναπαριστούν σε ένα διάνυσμα πλούσια πληροφορία σχετικά με τη γεωμετρία του αντικειμένου που αναγνωρίζουν (η πληροφορία αυτή θα ήταν αδύνατο να κωδικοποιηθεί σε μια μεμονωμένη τιμή). Οι παράμετροι στιγμιοτύπου μιας κάψουλας που αφορούν το αντικείμενο που αναπαριστούν αλλάζουν με προβλέψιμο τρόπο καθώς το αντικείμενο μετακινείται στην πολλαπλότητα των δυνατών απεικονίσεων (manifold of possible appearances). Συνεπώς, οι αλλαγές στην οπτική γωνία μεταφέρονται με αποδοτικό τρόπο μέσα από το σύστημα. Αντίθετα, η τιμή πιθανότητας ύπαρξης του αντικειμένου που αναγνωρίζει η κάθε κάψουλα στο πεδίο υποδοχής της επιθυμούμε να είναι όσο το δυνατό ανεξάρτητη από τον τρόπο απεικόνισης του αντικειμένου<sup>33</sup>. Επίσης, ανεξάρτητοι επιβάλλουμε να είναι οι πίνακες  $\mathbf{W}$  που αποθηκεύουν τις σχέσεις μεταξύ μερών και του όλου<sup>34</sup>. Ένα τελευταίο παράδειγμα που συμβάλλει έμμεσα στην επίτευξη γενίκευσης σε μεταβολές οπτικής γωνίας είναι η ενσωμάτωση του αλγορίθμου δρομολόγησης με συμφωνία. Εκτός από τον ρόλο που περιγράψαμε στο να διαμορφώνει τις κάψουλες του επόμενου επιπέδου, είναι πολύ σημαντικό ότι εντοπίζει συνδιακυμάνσεις μεταξύ των αναπαραστάσεων εισόδου<sup>35</sup> συγκρίνοντας τα διανύσματα ψήφων μεταξύ τους μέσω του φιλτραρίσματος πολυδιάστατης σύμπτωσης.

### Απλό Παράδειγμα Εφαρμογής Αλγορίθμου Δρομολόγησης με Συμφωνία

Στο απλό παράδειγμα που εξετάζουμε [46], ας υποθέσουμε ότι κάθε κάψουλα έχει ως παραμέτρους στιγμιοτύπου τις τιμές που προσδιορίζουν τη θέση του αντικειμένου  $(x, y)$  και την τιμή  $\theta$  του προσανατολισμού. Όπως αναφέραμε, στόχος είναι να γίνουν ανάστροφα γραφικά όπως φαίνεται και στο σχήμα 2.22. Η αναπαράσταση των τιμών μιας κάψουλας στο παράδειγμά μας θα γίνεται με ένα διάνυσμα του οποίου ο προσανατολισμός θα κωδικοποιεί τις παραμέτρους στιγμιοτύπου και το μήκος του την τιμή ενεργοποίησης. Όσο μεγαλύτερο το μήκος, τόσο πιο σίγουρη είναι η

<sup>32</sup>Η στρέψη του ψηφίου κατά  $+10$  μοίρες είναι ισοδύναμη με τη στρέψη της κάμερας κατά  $-10$  μοίρες. Συνεπώς, αποτελεί μια νέα οπτική γωνία για την οποία το δίκτυο δεν έχει εκπαιδευτεί.

<sup>33</sup>Στις μέχρι τώρα κύριες υλοποιήσεις, δεν υπάρχει πλήρης ανεξαρτησία σε όλο το φάσμα των πιθανών απεικονίσεων.

<sup>34</sup>Προφανώς, όπως και στα γραφικά υπολογιστή, ο πίνακας μετασχηματισμού που εκφράζει τις σχέσεις μέρους-όλου είναι ανεξάρτητος από την εκάστοτε οπτική γωνία του στιγμιοτύπου.

<sup>35</sup>Ισοδύναμα, δεν υποφέρει από το πρόβλημα της αποκλειστικής διάζευξης.

κάψουλα για την ύπαρξη της οντότητας που αναγνωρίζει. Ας υποθέσουμε ότι έχουμε δύο είδη από κάψουλες στο πρώτο επίπεδο (σχηματίζονται συνήθως από συνελκτικά επίπεδα): αυτές που αναγνωρίζουν ορθογώνιο και αυτές που αναγνωρίζουν τρίγωνο (στο σχήμα 2.23 συμβολίζονται με πράσινα και μπλε βέλη αντίστοιχα). Τα δύο είδη από κάψουλες διαμοιράζονται στον χώρο όπως τα φίλτρα στα συνελκτικά επίπεδα. Στο σχήμα 2.23 παρατηρούμε ότι όλες οι κάψουλες έχουν μικρά διανύσματα εκτός από τις δύο που έχουν πεδίο υποδοχής το μέρος της εικόνας όπου τοποθετείται το τρίγωνο και το τετράγωνο.

Τώρα, καλούμαστε να υπολογίσουμε τις κάψουλες του επόμενου επιπέδου γνωρίζοντας τις τιμές του προηγούμενου επιπέδου, μια διαδικασία που έχουμε ονομάσει δρομολόγηση μέσω συμφωνίας. Με αυτόν τον τρόπο, θα υπολογίσουμε τις παραμέτρους στιγμιότυπου πιο σύνθετων αντικειμένων (βλ. σχήμα 2.24). Για τον σκοπό αυτό, κάθε κάψουλα του πρώτου επιπέδου, με βάση τις τιμές της, παράγει τόσες προβλέψεις όσες είναι οι κάψουλες του επόμενου επιπέδου που βλέπει. Ας υποθέσουμε ότι υπάρχουν δύο κάψουλες στο επόμενο επίπεδο: μια που αναπαριστά την οντότητα σπίτι και μια που αναπαριστά την οντότητα βάρκα. Όπως γίνεται κατανοητό στο σχήμα 2.25 η κάθε κάψουλα του πρώτου επιπέδου προβλέπει το διάνυσμα της κάψουλας που αναπαριστά το σπίτι και τη βάρκα με το να πολλαπλασιάζει τις τιμές της με τον αντίστοιχο πίνακα μετασχηματισμού  $W_{ij}$ .

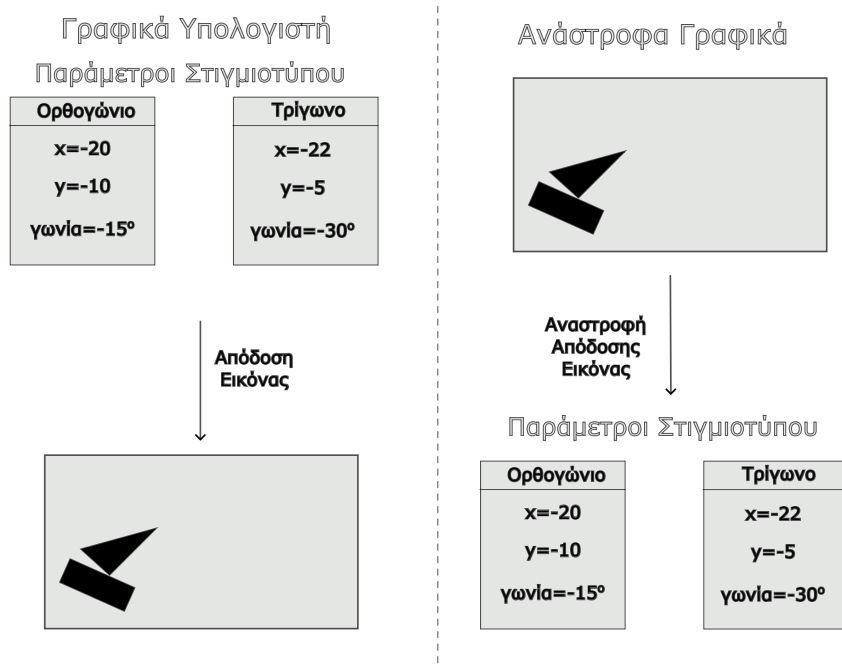
Στην πρώτη επανάληψη του αλγορίθμου συμφωνίας, κάθε κάψουλα δρομολογεί στις κάψουλες ανώτερου επιπέδου τις προβλέψεις της ισάξια. Όμως, σύντομα αναγνωρίζεται ότι υπάρχει μεγάλη συμφωνία μεταξύ των προβλέψεων για βάρκα (βλ. σχήμα 2.25). Λόγω της συμφωνίας, το διάνυσμα της κάψουλας για τη βάρκα που σχηματίζεται από τις σύμφωνες συστάδες προβλέψεων αποκτά μεγάλο μήκος. Έτσι, επαναληπτικά, οι κάψουλες προσαρμόζουν τα βάρη δρομολόγησης ώστε τελικά να δρομολογούν όλη την ψήφο τους στην οντότητα που τους εκφράζει καλύτερα (στην περίπτωσή μας, τη βάρκα).

## Υποθέσεις Νευρωνικών Δικτύων με Κάψουλες

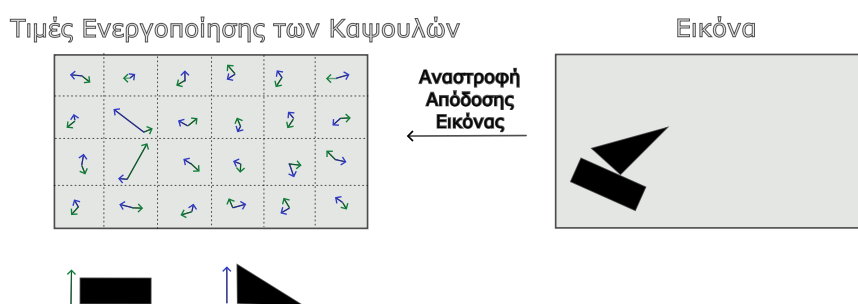
Οι υποθέσεις στις οποίες βασίζονται τα νευρωνικά δίκτυα με κάψουλες είναι οι εξής:

- Οι τιμές των καψουλών  $(M, a)$  εξηγούν πιστά τις όποιες μεταβολές της εικόνας εισόδου και των αντικειμένων που αυτή περιέχει (capturing equivariance). Αντίθετα, οι πίνακες βαρών  $W$  κωδικοποιούν την ανεξάρτητη (από την είσοδο) γνώση (capturing invariance).
- Οι πολυδιάστατες συμπτώσεις (high-dimensional coincidences) αποτελούν ένα κατάλληλο φίλτρο για εξαγωγή χαρακτηριστικών.
- Αλλαγές στην οπτική γωνία προκαλούν μη γραμμικές μεταβολές στα εικονοστοιχεία και γραμμικές στις σχέσεις μεταξύ αντικειμένων (ή μερών του) και της κάμερας.
- Κάθε τμήμα ενός αντικειμένου ανήκει σε ένα γενικότερο αντικείμενο (single parent assumption) και κάθε περιοχή περιέχει το πολύ ένα στιγμιότυπο του ίδιου αντικειμένου<sup>36</sup>.

<sup>36</sup>Η τελευταία υπόθεση είναι αναγκαία διότι στο ίδιο οπτικό πεδίο υπάρχει μια κάψουλα για κάθε οντότητα. Αυτό είναι και το τίμημα της χρήσης της θέσης των καψουλών μέσα στο δίκτυο για να προσδιοριστεί η ακριβής θέση των οντοτήτων που αναπαριστούν (όπως στα συνελκτικά νευρωνικά δίκτυα χωρίς επίπεδα συνάθροισης).

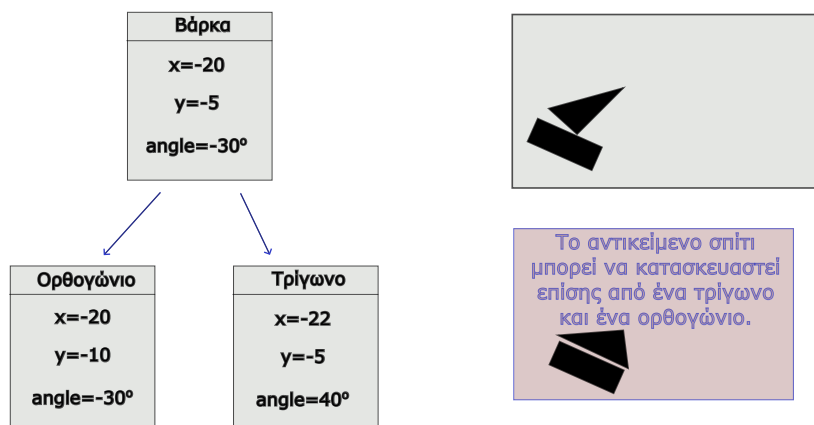


Σχήμα 2.22: Η διαδικασία ανάστροφων γραφικών που επιχειρείται από τα νευρωνικά δίκτυα από κάψουλες. Στο σχήμα, αντιπαραβάλλεται με τη διεργασία της απόδοσης εικόνας (rendering). Παράχθηκε από το *Inkscape*.

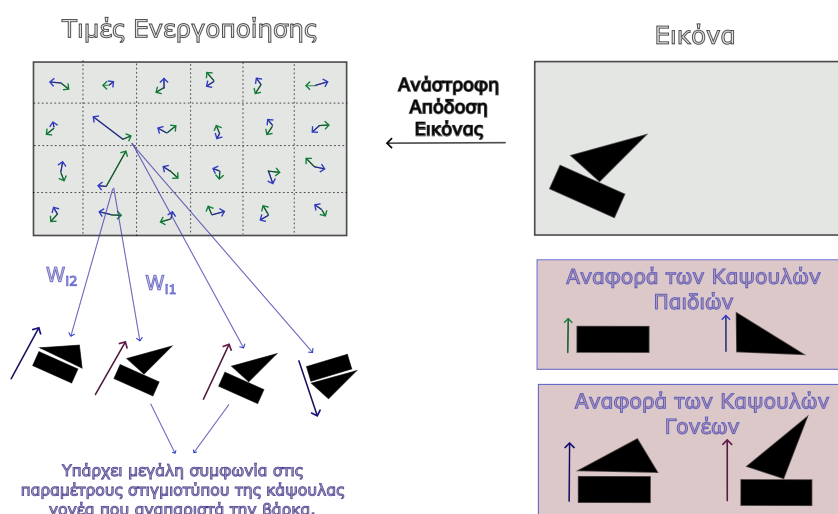


Σχήμα 2.23: Τιμές ενεργοποίησης όπως προκύπτουν από ένα συνελικτικό επίπεδο καψουλών που αναγνωρίζουν δύο οντότητες: ορθογώνιο και τρίγωνο. Οι κάψουλες αυτού του επιπέδου θα μπορούσε να είναι το αποτέλεσμα συνελικτικών επιπέδων από νευρώνες (στην περίπτωση αρχικών επιπέδων) ή το αποτέλεσμα δρομολόγησης μέσω συμφωνίας από προηγούμενο επίπεδο καψουλών. Παράχθηκε από το *Inkscape*.

### Ιεραρχία Μερών Αντικειμένου



Σχήμα 2.24: Παρατηρούμε ότι με τα ίδια τμήματα μπορούμε να φτιάξουμε δύο διαφορετικά αντικείμενα (βάρκα και σπίτι). Παίζει μεγάλο ρόλο λοιπόν η σωστή αναγνώριση της γεωμετρίας των μερών και το πως συνδέονται αυτά μεταξύ τους. Παράχθηκε από το *Inkscape*.



Σχήμα 2.25: Στο ανωτέρω σχήμα απεικονίζονται ενδεικτικά οι δύο προβλέψεις για τις δύο πιο ενεργές κάψουλες παιδιά (κάψουλες του προηγούμενου επιπέδου). Κάθε μια κάψουλα προσπαθεί να προβλέψει τη γεωμετρία των αντικειμένων του ανώτερου επιπέδου με βάση τη γεωμετρία της οντότητας που αναγνωρίζει. Βλέπουμε λοιπόν ότι το φιλτράρισμα συμπτώσεων υψηλής διάστασης είναι αποτελεσματικό αφού και οι δύο κάψουλες συμφωνούν στην οντότητα βάρκα. Παράχθηκε από το *Inkscape*.



### Αδυναμίες των Νευρωνικών Δικτύων με Κάψουλες

Αυτού του είδους των νευρωνικών δικτύων, ακόμα και στην πιο πρόσφατη υλοποίησή του από τον G. Hinton, παρουσιάζει ορισμένα προβλήματα. Αυτά, είναι τα εξής:

- Δεν κλιμακώνονται εύκολα σε πιο μεγάλα και σύνθετα σύνολα δεδομένων λόγω υψηλών απαιτήσεων μνήμης και μη της έλλειψης αποδοτικών αλγορίθμων βελτιστοποιημένων ως προς τους υπολογισμούς ενός δικτύου με κάψουλες.
- Δεν είναι δυνατή η διαμόρφωση των πινάκων βαρών  $W$  με μη-επιβλεπόμενη μάθηση (Κάτι τέτοιο θα συνέπτυσσε όλες τις ψήφους σε ένα σημείο.). ;;;; ΤΟΔΟ
- Κάψουλες που αναπαριστούν αντικείμενα τα οποία λόγω της γεωμετρίας τους έχουν ακαθόριστη πόζα, δεν μπορούν να προβλέψουν τις παραμέτρους στιγμιοτύπου των επόμενων καψουλών. Για παράδειγμα, μια κάψουλα που αναπαριστά μια ρόδα δεν μπορεί να προβλέψει τη γεωμετρία του αυτοκινήτου.
- Είναι δύσκολη η παραμετροποίηση του αλγορίθμου εύρεσης συστάδων ώστε να επιτυγχάνεται υψηλή επίδοση. Η ρύθμιση του αλγορίθμου πρέπει να είναι τέτοια ώστε να πετυχαίνει μια ισορροπία μεταξύ της πυκνότητας των συστάδων και τον αριθμό των ψήφων που περιέχουν.

## 2.3 Μετασχηματιστές

Σε αυτή την ενότητα θα αναφερθούμε σε μια αναδυόμενη αρχιτεκτονική νευρωνικών δικτύων, αυτή των Μετασχηματιστών (Transformers). Αν και αρχικά αναπτύχθηκε για εφαρμογές ακολουθιακών δεδομένων [61], η μεγάλη του επιτυχία οδήγησε σύντομα στον πειραματισμό του σε μη-ακολουθιακά δεδομένα όπως αυτά των (στατικών) εικόνων [62, 63]. Σε αυτήν την ενότητα θα κάνουμε μια σύντομη εισαγωγή στην τεχνολογία των επαναλαμβανόμενων νευρωνικών δικτύων (Recurrent Neural Networks) [60] και σε ορισμένα προβλήματά τους [61, 64, 65]. Έπειτα, θα αναφερθούμε στις διαδοχικές βελτιώσεις —με κυριότερη αυτή της προσοχής (attention) [65]— οι οποίες τελικά διαμόρφωσαν την αρχιτεκτονική των μετασχηματιστών. Τέλος, συνοπτικά θα γίνει λόγος σε ορισμένες εφαρμογές της νέας αρχιτεκτονικής σε εικόνες [62, 63], προκειμένου αργότερα να είναι ευκολότερα κατανοητή μιας από τις μεθόδους μας η οποία αφορά την ενσωμάτωση μηχανισμού προσοχής πολλών κεφαλών (multihead attention) στα νευρωνικά δίκτυα με κάψουλες.

## 2.4 Χάρτες Αυτο-οργάνωσης

## 2.5 Ρεσονστρυκτιον ας Ρεγυλαριζατιον

## Κεφάλαιο 3

### Σχετικές Ερασίες

## Κεφάλαιο 4

### Μέθοδος

## Κεφάλαιο 5

### Πειράματα

Κεφάλαιο 6

Επίλογος



# Βιβλιογραφία

- [1] C. S. Smith, “A.I. Here, There, Everywhere,” *The New York Times*, Feb. 2021.
- [2] B. Marr, “The 10 Best Examples Of How AI Is Already Used In Our Everyday Life,” Dec. 2019.
- [3] S.-L. Wamba-Taguimdje, S. F. Wamba, J. R. K. Kamdjoug, and C. E. T. Wanko, “Influence of artificial intelligence (ai) on firm performance: the business value of ai-based transformation projects,” *Business Process Management Journal*, 2020.
- [4] M. Fernández, A. Bellogín, and I. Cantador, “Analysing the effect of recommendation algorithms on the amplification of misinformation,” *arXiv preprint arXiv:2103.14748*, 2021.
- [5] C. A. Gomez-Urbe and N. Hunt, “The netflix recommender system: Algorithms, business value, and innovation,” *ACM Transactions on Management Information Systems (TMIS)*, vol. 6, no. 4, pp. 1–19, 2015.
- [6] S. Makridakis, “The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms,” *Futures*, vol. 90, pp. 46–60, 2017.
- [7] J. Hawksworth, R. Berriman, and S. Goel, “Will robots really steal our jobs? an international analysis of the potential long term impact of automation,” *PricewaterhouseCoopers*, <http://pwc.co.uk/economics>, access, vol. 13, 2018.
- [8] W. E. Forum, “The future of jobs report 2020,” 2020.
- [9] M. C.-T. Tai, “The impact of artificial intelligence on human society and bioethics,” *Tzu-Chi Medical Journal*, vol. 32, no. 4, p. 339, 2020.
- [10] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, *et al.*, “Improved protein structure prediction using potentials from deep learning,” *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.
- [11] T. Panch, H. Mattie, and L. A. Celi, “The “inconvenient truth” about ai in healthcare,” *NPJ digital medicine*, vol. 2, no. 1, pp. 1–3, 2019.
- [12] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.
- [13] N. Bajema, “AI’s 6 Worst-Case Scenarios,” Jan. 2022.

- [14] European Comission, “A European approach to artificial intelligence | Shaping Europe’s digital future,” 2021.
- [15] U. von der Leyen, “President von der Leyen on the Commission’s new strategy: Shaping Europe’s Digital Future,” 2019.
- [16] A. Mayor, *Gods and robots: myths, machines, and ancient dreams of technology*. Princeton University Press, 2020.
- [17] S. Pinker, “The cognitive niche: Coevolution of intelligence, sociality, and language,” *Proceedings of the National Academy of Sciences*, vol. 107, no. supplement\_2, pp. 8993–8999, 2010.
- [18] J. Tooby and I. DeVore, “The reconstruction of hominid behavioral evolution through strategic modeling,” *The evolution of human behavior: Primate models*, pp. 183–237, 1987.
- [19] J. C. Avise and F. J. Ayala, *In the Light of Evolution: Volume IV: The Human Condition*. National Academies Press, 2010.
- [20] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.
- [21] D. Hebb, *The organization of behavior; a neuropsychological theory*. Wiley, 1949.
- [22] J. Akst, “Machine, learning, 1951,” May 2019.
- [23] J. A. Lee, *Computer pioneers*. IEEE Computer Society Press, 1995.
- [24] A. M. TURING, “I.—COMPUTING MACHINERY AND INTELLIGENCE,” *Mind*, vol. LIX, pp. 433–460, 10 1950.
- [25] A. M. Turing, “Intelligent machinery,” 1948.
- [26] H. Muehlenbein, “Artificial intelligence and neural networks the legacy of alan turing and john von neumann,” *Int J Comput*, vol. 5, no. 3, pp. 10–20, 2014.
- [27] S. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Pearson, 2020.
- [28] J. McCarthy, “Programs with common sense,” in *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, RLE and MIT computation center Cambridge, MA, USA, 1960.
- [29] J. R. Slagle, “A heuristic program that solves symbolic integration problems in freshman calculus,” *J. ACM*, vol. 10, p. 507–520, oct 1963.
- [30] J. Slaney and S. Thiébaux, “Blocks world revisited,” *Artificial Intelligence*, vol. 125, no. 1-2, pp. 119–153, 2001.
- [31] S. S. Haykin, *Neural networks and learning machines*. Upper Saddle River, NJ: Pearson Education, third ed., 2009.



- [32] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain.,” *Psychological review*, vol. 65, no. 6, p. 386, 1958.
- [33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [34] A. Vinhas, “Adaline neural networks: The origin of gradient descent,” Mar 2021.
- [35] M. Mitchell, “Why ai is harder than we think,” *arXiv preprint arXiv:2104.12871*, 2021.
- [36] B. S. Todd, *An introduction to expert systems*. Oxford University Computing Laboratory, Programming Research Group, 1992.
- [37] M. Z. Bell, “Why expert systems fail,” *Journal of the Operational Research Society*, vol. 36, no. 7, pp. 613–619, 1985.
- [38] D. E. Rumelhart, J. L. McClelland, P. R. Group, *et al.*, *Parallel distributed processing*, vol. 1. IEEE New York, 1988.
- [39] R. Singh, “Rise and fall of symbolic ai,” Sep 2019.
- [40] E. N. Zalta, “The stanford encyclopedia of philosophy,” 2019.
- [41] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.
- [43] N. Benaich and I. Hogarth, “State of ai report 2021,” Oct 2021.
- [44] U. Beck, *Risk society*. Sage Publications Ltd, 1992.
- [45] T. M. Mitchell *et al.*, “Machine learning,” 1997.
- [46] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.", 2019.
- [47] A. Krizhevsky, “Learning multiple layers of features from tiny images,” pp. 32–33, 2009.
- [48] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, “The loss surfaces of multilayer networks,” in *Artificial intelligence and statistics*, pp. 192–204, PMLR, 2015.
- [49] A. Ng, “Forward and backward propagation (c1w4l06).”
- [50] D. H. Hubel, “Single unit activity in striate cortex of unrestrained cats,” *The Journal of physiology*, vol. 147, no. 2, p. 226, 1959.
- [51] D. H. Hubel and T. N. Wiesel, “Receptive fields of single neurones in the cat’s striate cortex,” *The Journal of physiology*, vol. 148, no. 3, p. 574, 1959.
- [52] D. H. Hubel and T. N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex,” *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968.

- [53] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [54] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, pp. 818–833, Springer, 2014.
- [55] A. Kosioerek, S. Sabour, Y. W. Teh, and G. E. Hinton, “Stacked capsule autoencoders,” *Advances in neural information processing systems*, vol. 32, 2019.
- [56] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” *Advances in neural information processing systems*, vol. 30, 2017.
- [57] G. E. Hinton, S. Sabour, and N. Frosst, “Matrix capsules with em routing,” in *International conference on learning representations*, 2018.
- [58] G. Hinton, “How to represent part-whole hierarchies in a neural network,” *arXiv preprint arXiv:2102.12627*, 2021.
- [59] G. E. Hinton, “Capsule networks.”
- [60] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
- [61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.
- [62] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [63] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*, pp. 213–229, Springer, 2020.
- [64] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [65] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.
- [66] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

## Παράρτημα Α΄

### Ορισμοί Εννοιών

Το παρόν παράρτημα περιέχει ορισμούς εννοιών που εισάγονται κατά τη διάρκεια της παρούσας εργασίας. Κατά αυτόν τον τρόπο, δε διακόπτεται η ροή του κυρίως κειμένου. [27, 33, 46, 66]

#### **Τεχνητή Νοημοσύνη**

Έχουν υπάρξει πολλοί διαφορετικοί ορισμοί της Τεχνητής Νοημοσύνης: Μερικοί την περιγράφουν σαν εσωτερική διαδικασία της σκέψης που προσομοιάζει αυτή του ανθρώπου ενώ άλλοι ως εξωτερική διαδικασία μαθηματικά βέλτιστης συμπεριφοράς. Σύμφωνα με το κυρίαρχο μοντέλο, η Τεχνητή Νοημοσύνη ασχολείται κυρίως με τη λογική δράση. Ένας ιδανικός ευφυής πράκτορας δρα βέλτιστα σε κάθε περίπτωση. Έτσι λοιπόν, η μελέτη της δημιουργίας ευφύων πρακτόρων μπορεί να τεθεί ως ορισμός της Τεχνητής Νοημοσύνης.

#### **Μηχανική Μάθηση**

Με λίγα λόγια, πρόκειται για τον κλάδο της τεχνητής νοημοσύνης ο οποίος ασχολείται με την ανάπτυξη υπολογιστικών συστημάτων ικανών να μαθαίνουν από παραδείγματα. Αναλυτικότερα, μπορούν και προσαρμόζονται χωρίς να ακολουθούν ρητές εντολές αλλά μέσω αλγορίθμων και στατιστικών μοντέλων που τους επιτρέπουν να αναλύουν και να εξάγουν συμπεράσματα από μοτίβα σε δεδομένα. Χαρακτηριστικό γνώρισμα των συστημάτων μηχανικής μάθησης είναι η ικανότητά τους να βελτιώνουν την απόδοσή τους σε μια εργασία (όπως αυτή μετράται με κάποια κατάλληλη μετρική) όσο η «εμπειρία» τους σε αυτήν αυξάνεται [45].

#### **Τεχνητά Νευρωνικά Δίκτυα**

Τα τεχνητά νευρωνικά δίκτυα αποτελούν ένα αλγοριθμικό κατασκεύασμα από απλούς υπολογιστικούς κόμβους διασυνδεδεμένους μεταξύ τους μέσω ακμών κάτω από μια συγκεκριμένη τοπολογία (συνήθως οργανώνονται σε επίπεδα, βλ. 2.1.3). Εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα, οι κόμβοι μπορούν να παρομοιαστούν με κύτταρα νευρώνων ενώ οι ακμές με νευρικές συνάψεις.

Τα τεχνητά νευρωνικά δίκτυα είναι παράδειγμα συστήματος μηχανικής μάθησης αφού μετά την κατάλληλη εκπαίδευσή τους, γενικεύουν από τα δεδομένα (inference). Τελικά, μετά την ανάπτυξή τους, υπό μια αφαιρετική σκοπιά αποτελεί το καθένα μια συνάρτηση που αντιστοιχίζει δεδομένα από τον χώρο εισόδου σε «προβλέψεις» του χώρου εξόδου.

### **Βαθιά Μάθηση**

Αποτελεί μια υποκατηγορία μηχανικής μάθησης όπου χρησιμοποιούνται πολυεπίπεδα νευρωνικά δίκτυα. Τα πολλαπλά επίπεδα που διαθέτουν τους επιτρέπουν να μαθαίνουν και να αναγνωρίζουν εσωτερικά, γενικευμένα χαρακτηριστικά των δεδομένων εισόδου.

### **Υπολογιστική Νευροεπιστήμη (Computational Neuroscience)**

Πρόκειται για τον κλάδο της Νευροεπιστήμης που χρησιμοποιεί μαθηματικά μοντέλα, μαθηματική ανάλυση και προσεγγιστικά προς τον εγκέφαλο συστήματα για να κατανοήσει τις αρχές ανάπτυξης, δομής, φυσιολογίας καθώς και των γνωστικών (cognitive) ικανοτήτων του νευρικού συστήματος.

### **Επιβλεπόμενη Μάθηση**

Στους αλγορίθμους επιβλεπόμενης μάθησης, ως είσοδος παρέχεται ένα σύνολο δεδομένων μαζί με τους επιθυμητούς στόχους. Δηλαδή, τα δεδομένα δίνονται σε ζεύγη (παράδειγμα εισόδου—επιθυμητή τιμή εξόδου). Με βάση αυτά, το σύστημα καλείται να εξάγει μια συνάρτηση η οποία θα έχει μάθει να μοντελοποιεί τη σχέση εισόδου–εξόδου μέσα από τα παραδείγματα και τελικά θα είναι ικανή να προβλέψει την τιμή εξόδου σε νέα παραδείγματα για τα οποία η τιμή στόχος είναι άγνωστη. Συνήθως, εκτός από τα δεδομένα για την εκπαίδευση υπάρχουν και άλλα σύνολα δεδομένων για τον έλεγχο της απόδοσης του συστήματος πρόβλεψης.

Ανάλογα με το αν η τιμή στόχος είναι διακριτή η συνεχής, έχουμε αντίστοιχα το πρόβλημα ταξινόμησης (classification) ή της παλινδρόμησης (regression). Παράδειγμα συστήματος ταξινόμησης επιβλεπόμενης μάθησης είναι το φίλτρο ανεπιθύμητης αλληλογραφίας το οποίο αφού εκπαιδεύτηκε με ένα σύνολο επισημασμένων αλληλογραφιών ως ανεπιθύμητων ή επιθυμητών έμαθε να εντοπίζει νέα εισερχόμενη ανεπιθύμητη αλληλογραφία. Ένα παράδειγμα συστήματος παλινδρόμησης επιβλεπόμενης μάθησης είναι αυτό της πρόβλεψης τιμών μετοχών καθώς ο στόχος (κόστος μετοχής) είναι συνεχής αριθμός.

### **Μη-επιβλεπόμενη Μάθηση**

Οι αλγόριθμοι μη-επιβλεπόμενης μάθησης, σε αντιδιαστολή με τους αλγορίθμους επιβλεπόμενης μάθησης, δέχονται ως είσοδο ένα σύνολο δεδομένων που περιλαμβάνει παραδείγματα, χωρίς όμως να συνοδεύονται από αντίστοιχες τιμές–στόχους. Στην περίπτωση αυτή, το υπό εκπαίδευση σύστημα επιχειρεί να μάθει πρότυπα στα δεδομένα εισόδου χωρίς κάποιο μηχανισμό ανατροφοδότησης. Συνήθεις εφαρμογές μη-επιβλεπόμενης μάθησης είναι αυτές της ομαδοποίησης των δεδομένων σε συστάδες ή της αναπαράστασής τους με ένα γράφημα.

### **Ενισχυτική Μάθηση**

Στην ενισχυτική μάθηση, στο σύστημα (το οποίο καλείται «ευφυής πράκτορας» στο πλαίσιο αυτό) δεν παρέχεται κάποιο σύνολο δεδομένων αλλά η όποια εμπειρία αποκτάται μέσω της αλληλεπίδρασής του με το περιβάλλον. Ο πράκτορας έχει τη δυνατότητα να παρατηρήσει το περιβάλλον του και τη (πιθανή) κατάστασή του και ανάλογα με μια στρατηγική (policy) να δράσει σε αυτό. Το περιβάλλον του, με κάθε δράση (και ανάλογα την κατάσταση) παρέχει την απαραίτητη εμπειρία υπό τη μορφή επιβράβευσης (reward) ή ποινής (punishment). Έτσι, ο πράκτορας μαθαίνει από την εμπειρία προσαρμόζοντας τη στρατηγική του ώστε να μεγιστοποιεί την επιβράβευση την οποία λαμβάνει και τελικά να πετυχαίνει τον στόχο του.

---

Παράδειγμα ενός τέτοιου πράκτορα είναι ένα σύστημα το οποίο παίζει σκάκι.

### **Μάθηση Κατά Δέσμες**

Αφορά το είδος συστημάτων μηχανικής μάθησης που δεν έχουν τη δυνατότητα να μαθαίνουν σταδιακά αλλά εκπαιδεύονται μονομιάς χρησιμοποιώντας όλο το σύνολο δεδομένων στην είσοδό τους. Σε περίπτωση που προστεθούν νέα δεδομένα στα οποία θα επιθυμούσαμε το σύστημα να προσαρμοστεί, απαιτείται εκ νέου εκπαίδευση στο καινούριο σύνολο δεδομένων το οποίο θα περιέχει τόσο τα παλαιά όσο και τα επιπρόσθετα δεδομένα (διαδικασία χρονοβόρα και υπολογιστικά κοστοβόρα). Συνήθως, σε τέτοιες περιπτώσεις το σύστημα πρέπει να σταματήσει να λειτουργεί και να μεταβεί στη φάση σχεδιασμού. Παραδείγματα αυτών των μεθόδων αποτελούν ο αλγόριθμος Expectation Maximization και ο Self-organizing map όπως περιγράφονται στην ενότητα ;; και 2.5.

### **Μάθηση σε Ζωντανό Χρόνο**

Πρόκειται για τα συστήματα μηχανικής μάθησης που, σε αντίθεση με αυτά που μαθαίνουν κατά δέσμες, είναι ικανά να εκπαιδεύονται σταδιακά, είτε με ένα παράδειγμα τη φορά είτε με μικρές δέσμες παραδειγμάτων στην είσοδό τους. Το θετικό σε αυτά τα συστήματα είναι η δυνατότητα προσαρμογής τους σε νέα δεδομένα με πολύ μικρό χρονικό και υπολογιστικό κόστος. Αποτέλεσμα αυτού είναι ότι υπάρχει (συνήθως) η δυνατότητα η εκπαίδευσή τους να γίνει ζωντανά (online) χωρίς να σταματήσει η λειτουργία του συστήματος. Παράδειγμα αποτελούν οι εφαρμογές πρόβλεψης τιμών μετοχών όπου απαιτείται συνεχής προσαρμογή του συστήματος στα νέα δεδομένα της αγοράς.

### **Μάθηση Βασισμένη σε Παραδείγματα**

Είναι μια οικογένεια απλών συστημάτων μηχανικής μάθησης που αφορά τον τρόπο με τον οποίο ένα σύστημα γενικεύει από τα παραδείγματα του συνόλου εισόδου. Στα συγκεκριμένα, όταν τα τροφοδοτούμε με κάποιο νέο παράδειγμα, το συγκρίνουν με τα δεδομένα εισόδου (ή ένα υποσύνολο αυτών) τα οποία έχουν αποθηκευθεί στη μνήμη τους κατά την εκπαίδευση. Ένα χαρακτηριστικό μειονέκτημα αυτών των συστημάτων είναι ότι ο χώρος που απαιτείται για την αποθήκευση του μοντέλου (του συστήματος μάθησης μετά την εκπαίδευσή του) αυξάνεται με το μέγεθος του συνόλου εισόδου (συνήθως με γραμμικό τρόπο). Ενδεικτικά, ένα σύστημα που γενικεύει κατά αυτόν τον τρόπο είναι το K-nearest neighbors.

### **Μάθηση Βασισμένη σε μοντέλο**

Είναι μια άλλη οικογένεια συστημάτων όπου η μηχανική μάθηση γίνεται μέσω της προσαρμογής (fitting) ενός μοντέλου στα δεδομένα εισόδου. Έχοντας εκφράσει το σύνολο των δεδομένων εκπαίδευσης (ή τη σχέση αυτών με την επιθυμητή έξοδο) χρησιμοποιώντας ένα κατάλληλα εκφραστικό (expressive) μοντέλο, λέμε ότι το σύστημα μαθαίνει να «γενικεύει» από τα παραδείγματα. Έτσι, για να παράξει προβλέψεις σε νέα δεδομένα, δεν απαιτείται η αποθήκευση όλων των δεδομένων εκπαίδευσης αλλά μόνο των παραμέτρων του μοντέλου που εκφράζει.

### **Γνωστική Νευροεπιστήμη**

Η Γνωστική νευροεπιστήμη είναι το πεδίο μελέτης που ασχολείται με τα νευρωνικά υποστρώματα των διανοητικών διεργασιών. Είναι η τομή της ψυχολογίας με τη νευροεπιστήμη. Συνδυάζει τις θεωρίες της γνωσιακής ψυχολογίας και της υπολογιστικής μοντελοποίησης με πειραματικά δεδομένα του κεφάλου.

### **Αναγνώριση Προτύπων**

Είναι ένα επιστημονικό πεδίο με στόχο την ανάπτυξη αλγορίθμων για την αυτοματοποιημένη απόδοση κάποιας τιμής (παλινδρόμησης) ή διακριτικού στοιχείου (ταξινόμηση) με βάση μοτίβα/χαρακτηριστικά που παρατηρούνται στα εισαγόμενα δεδομένα, συνήθως κωδικοποιημένα ως αλληλουχίες αριθμών.

### **Γραμμικά Διαχωρίσιμες Κλάσεις**

Λέμε ότι ένα σύνολο δεδομένων για ταξινόμηση που περιέχει δύο κλάσεις είναι γραμμικά διαχωρίσιμο αν και μόνο αν μπορούμε να διαχωρίσουμε τις δύο κλάσεις στον πολυδιάστατο χώρο χαρακτηριστικών εισόδου χρησιμοποιώντας ένα υπερεπίπεδο. Στην περίπτωση όπου ο χώρος χαρακτηριστικών είναι διδιάστατος, αρκεί να μπορούμε να χαράξουμε μια ευθεία γραμμή στο καρτεσιανό επίπεδο που να διαχωρίζει τις δύο κλάσεις.

### **Γραμμικά Μοντέλα**

Τα γραμμικά μοντέλα περιγράφουν τη σχέση μεταξύ ενός ή περισσότερων μεταβλητών εισόδου (μεταβλητές πρόβλεψης) και μιας συνεχούς τιμής εξόδου (απόκρισης). Η χρήση των μοντέλων αυτών ενδείκνυται όταν οι σχέσεις μεταξύ εισόδου–εξόδου είναι (σχεδόν) γραμμικές στο διάστημα μελέτης. Μια στατιστική μέθοδος για την παραγωγή γραμμικών μοντέλων που μοντελοποιούν αυτές τις σχέσεις από σύνολα δεδομένων εισόδου–εξόδου είναι η γραμμική παλινδρόμηση.

### **Γενετικοί Αλγόριθμοι**

Οι Γενετικοί αλγόριθμοι ανήκουν στο κλάδο της επιστήμης υπολογιστών και αποτελούν μια μέθοδο αναζήτησης βέλτιστων λύσεων σε προβλήματα βελτιστοποίησης. Είναι χρήσιμοι σε περιπτώσεις όπου ο χώρος αναζήτησης λύσης είναι πολύ μεγάλος και δεν υπάρχει αναλυτική μέθοδος που να μπορεί να βρει το βέλτιστο συνδυασμό τιμών των μεταβλητών του προβλήματος ώστε το υπό εξέταση σύστημα να αντιδρά με βέλτιστο τρόπο. Ο τρόπος λειτουργίας των Γενετικών Αλγορίθμων είναι εμπνευσμένος από τη βιολογία. Χρησιμοποιεί δηλαδή την ιδέα της εξέλιξης μέσω γενετικής μετάλλαξης, φυσικής επιλογής και διασταύρωσης. Για να αξιοποιήσουμε αυτές τις ιδέες, κωδικοποιήσουμε κάθε πιθανή λύση του προβλήματος σαν ένα συγκεκριμένο γονιδίωμα και ξεκινήσαμε από έναν τυχαίο πληθυσμό τέτοιων λύσεων/γονιδιωμάτων. Έπειτα, ορίζοντας μια συνάρτηση ικανότητας (fitness function) που περιγράφει την ποιότητα της λύσης είμαστε σε θέση να αφήσουμε τον μηχανισμό εξέλιξης να δράσει για ορισμένες γενιές ώστε τελικά να έχουν απομείνει και πολλαπλασιαστεί γονιδιώματα που περιγράφουν (σχεδόν) βέλτιστες λύσεις. Οι γενετικοί αλγόριθμοι δεν εγγυούνται την εύρεση της βέλτιστης λύσης.

### **Νευρωνικά Δίκτυα με Κάψουλες (Capsule Networks)**

Πρόκειται για βαθιά νευρωνικά δίκτυα που επιδιώκουν να πραγματοποιήσουν ανάστροφα γραφικά για να λύσουν κυρίως προβλήματα αναγνώρισης αντικειμένων σε εικόνες. Αποτελούνται από επίπεδα από κάψουλες. Κάθε κάψουλα είναι σαν μια συνάρτηση η οποία προσπαθεί να προβλέψει τις παραμέτρους στιγμιότυπου (π.χ. προσανατολισμός, θέση κ.τ.λ.) ενός συγκεκριμένου αντικειμένου και την πιθανότητα ύπαρξής του σε μια περιοχή της εικόνας (δηλαδή στο πεδίο υποδοχής της κάψουλας).

### **Γραφικά Υπολογιστή**

---

Αφορά τον κλάδο της επιστήμης υπολογιστών που μελετά μεθόδους για ψηφιακή σύνθεση και χειρισμό οπτικού περιεχομένου. Εμπεριέχει μια δόση τέχνης αφού σχετίζεται με τον σχεδιασμό του περιεχομένου αυτού.

**Απόδοση Εικόνας (Rendering)** Είναι η διεργασία δημιουργίας εικόνας από ένα μοντέλο δύο ή τριών διαστάσεων με τη χρήση ενός προγράμματος υπολογιστή. Πολλά μοντέλα ορίζονται σε ένα αρχείο σκηνής (scene file) το οποίο περιγράφει όλη την πληροφορία της οπτικής σκηνής που θα παραχθεί με την απόδοση εικόνας. Συνήθως, το αρχείο σκηνής περιέχει πληροφορία για τη γεωμετρία, την οπτική γωνία, την υφή, τον φωτισμό και τη σκίαση των αντικειμένων.

### **Ανάστροφα Γραφικά**

Πρόκειται για την ανάστροφη διαδικασία της απόδοσης εικόνας. Δηλαδή, δοθείσης μιας οπτικής εικόνας, να προσδιοριστεί το αρχείο σκηνής από το οποίο δημιουργήθηκε.

## Παράρτημα Β΄

# Απόδοση Ξενόγλωσσων Όρων

<u>Ξενόγλωσσος όρος</u>	<u>Ελληνική απόδοση</u>
batch learning	μάθηση κατά δέσμες
online learning	μάθηση σε ζωντανό χρόνο
supervised learning	επιβλεπόμενη μάθηση
unsupervised learning	μη-επιβλεπόμενη μάθηση
reinforcement learning	ενισχυτική μάθηση
capsule networks	νευρωνικά δίκτυα με κάψουλες
instance based	βασισμένο σε παραδείγματα
model based	βασισμένο σε μοντέλο



## Παράρτημα Γ'

# Συντομογραφίες - Ακρωνύμια

### Γ'.1 Ελληνικά

<u>Συντομογραφία ή Ακρωνύμιο</u>	<u>Πλήρης όρος</u>
δφθφεωργερρεωγτε	γσδφ εργωεγφεωρφτωεγ
δφθφεωργεργεργρεωγτε	γσγφεωρφτωεγ
δφθφεωργεγτε	γσδφ εωρφτωεγ
δφθφεωργερρεγεργεργτεργγεργρεωγτε	γσδφ εργωεγφεωρφτωεγεγεργεργερωγεργφ

### Γ'.2 Αγγλικά

<u>Συντομογραφία ή Ακρωνύμιο</u>	<u>Πλήρης όρος</u>
δφθφεωργερρεωγτε	γσδφ εργωεγφεωρφτωεγ
δφθφεωργεργεργρεωγτε	γσγφεωρφτωεγ
δφθφεωργεγτε	γσδφ εωρφτωεγ
δφθφεωργερρεγεργεργτεργγεργρεωγτε	γσδφ εργωεγφεωρφτωεγεγεργεργερωγεργφ