Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Επεξεργασίας Εικόνων, Βίντεο
και Συστημάτων Πολυμέσων

Τεχνικές Βαθιάς Μάθησης και Εφαρμογές στην Ανίχνευση Προσώπων σε Εικόνες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΑΝΑΓΙΩΤΗΣ Χ. ΜΕΛΕΤΗΣ

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Επεξεργασίας Εικόνων, Βίντεο
και Συστημάτων Πολυμέσων

Τεχνικές Βαθιάς Μάθησης και Εφαρμογές στην Ανίχνευση Προσώπων σε Εικόνες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΑΝΑΓΙΩΤΗΣ Χ. ΜΕΛΕΤΗΣ

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Οκτωβρίου 2015.

.....	.....	.....
Στέφανος Κόλλιας	Ανδρέας-Γεώργιος Σταφυλοπάτης	Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

Παναγιώτης Χ. Μελέτης

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

Περίληψη

Τα τελευταία χρόνια η αφθονία της οπτικοακουστικής πληροφορίας και η ταχεία αύξηση των υπολογιστικών δυνατοτήτων των μηχανών έστρεψαν το ενδιαφέρον πολλών ερευνητών σε μεθόδους αντιμετώπισης προβλημάτων οδηγούμενες από μεγάλες ποσότητες δεδομένων. Ο προσανατολισμός αυτός, οδήγησε σε μεγάλη ανάπτυξη των Τεχνικών Βαθιάς Μάθησης και ιδιαίτερα του κλάδου των Συνελικτικών Νευρωνικών Δικτύων (ΣΝΔ). Τα Δίκτυα αυτά εμπνέονται από τη δομή και τη λειτουργικότητα του ανθρώπινου εγκεφάλου και έχουν τη δυνατότητα να αυτορρυθμίζουν τη βαρύτητα των συνδέσεών τους μέσω επιβλεπόμενης εκπαίδευσης σε πολύ μεγάλα σύνολα δεδομένων. Στόχοι της Διπλωματικής Εργασίας είναι η ανάλυση των σύγχρονων ΣΝΔ και η παρουσίαση των τελευταίων εξελίξεων, αλλά και η εφαρμογή ενός, τελευταίας τεχνολογίας, ΣΝΔ για τον εντοπισμό προσώπων σε εικόνες.

Αρχικά, γίνεται αναφορά των βασικών στοιχείων της Μηχανικής Μάθησης, που είναι απαραίτητα για την ανάπτυξη των ΣΝΔ, ενώ μεγάλο μέρος της Εργασίας αφιερώνεται στην ανάλυση της αρχιτεκτονικής, των ιδιοτήτων και του τρόπου εκπαίδευσης των σύγχρονων ΣΝΔ. Η παρουσίαση γίνεται με υπόβαθρο το θεμελιώδες πρόβλημα της ταξινόμησης εικόνων και ως εφαρμογή υλοποιείται ένα σύστημα εντοπισμού προσώπων ανεξαρτήτου γωνίας λήψης σε εικόνες, με τη χρήση του επιτυχημένου ΣΝΔ ταξινόμησης AlexNet. Κατασκευάζεται, με έξυπνες τεχνικές επαύξησης, ένα σύνολο δεδομένων 1 εκατομμυρίου εικόνων από τις βάσεις δεδομένων AFLW και FaceScrub, που έχουν συνολικά 100,000 εικόνες. Το σύνολο αυτό χρησιμοποιείται για την ειδική προσαρμογή του προεκπαιδευμένου AlexNet από το σύνολο ILSVRC. Για τον εντοπισμό χρησιμοποιείται πολυκλιμακωτή ανάλυση εικόνων και το εξειδικευμένο AlexNet, ώστε να προκύψουν οι τοποθεσίες των προσώπων σε οποιαδήποτε κλίμακα.

Λέξεις κλειδιά

Τεχνικές Βαθιάς Μάθησης, Συνελικτικά Νευρωνικά Δίκτυα, Εντοπισμός Προσώπων, Ειδική Προσαρμογή, Πυκνή Ταξινόμηση, Επαύξηση Συνόλου Δεδομένων, Πολυκλιμακωτή Ανάλυση

Abstract

In recent years the availability of abundant data and the increase of computational capacity and capability of machines, lead researchers to tackle computer vision problems through data-driven approaches. This trend fostered Deep Learning Techniques and particularly the branch of Convolutional Neural Networks (CNN), which is inspired by the structure and functionality of human brain. CNN are trained on large image datasets and take advantage of the deep hierachical structure of images. A first goal of this Diploma Thesis is to describe modern CNN architectures thoroughly and present recent developments in the field. A second goal is to employ a state-of-the-art CNN, originally designed for image classification, for multiview detection of faces in images.

In the first part of the Thesis, essential background tools of Machine Learning, necessary to the deployment of CNNs, are discussed. In the next part, an extensive analysis of the architecture and the training procedure of CNNs is performed. Throughout the text, concepts are developed having the fundamental problem of classification on the background, as the outmost goal is to use the AlexNet CNN in order to achieve face localization. Clever dataset augmentation techniques are applied to AFLW and FaceScrub databases (total 100,000 images), to generate a training set of 1 million images, which is used to fine-tune a pretrained AlexNet model on ILSVRC. To achieve detection, a multiscale approach is adopted, so the fine-tuned AlexNet can infer the existence of faces in multiple scales.

Key words

Deep Learning, Convolutional Neural Networks, Multiview Face Detection, Fine-tuning, Dense Classification, Dataset Augmentation, Multiscale Analysis

Ευχαριστίες

Θα ήθελα αρχικά, να ευχαριστήσω τον καθηγητή κ. Στέφανο Κόλλια για την ευκαιρία που μου έδωσε να εκπονήσω αυτή τη Διπλωματική Εργασία στο Εργαστήριο Επεξεργασίας Εικόνας, Βίντεο και Πολυμέσων και για το χρόνο που αφιέρωσε στην επίβλεψή της. Επίσης, ευχαριστώ τον Κώστα Ραπαντζίκο που με ενέπνευσε να ασχοληθώ με τον τομέα των Βαθέων Νευρωνικών Δικτύων και για την καθοδήγηση κατά τη διάρκεια της Διπλωματικής, αλλά και το Χρήστο Βαρυτιμίδη για τις πολύτιμες συμβουλές του κατά τη διάρκεια της υλοποίησης. Τέλος, θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που με στήριξαν αυτά τα χρόνια και ιδιαίτερα του φίλους μου και την οικογένειά μου για την αγάπη και υπομονή καθ' όλη τη διάρκεια των σπουδών μου.

Παναγιώτης Χ. Μελέτης,

Αθήνα, 27η Οκτωβρίου 2015

Περιεχόμενα

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Αναγνώριση Αντικειμένων σε Εικόνες

Η Όραση Υπολογιστών και η Αναγνώριση Προτύπων είναι δύο στενά συσχετιζόμενα πεδία που αναπτύσσονται ταχύτατα τα τελευταία χρόνια και ενώ ξεκίνησαν ως πεδία των γενικότερων τομέων της Τεχνητής Νοημοσύνης και Επεξεργασίας Σημάτων, σήμερα δανείζονται ιδέες από πολλές άλλες επιστήμες (βιολογία, φυσική, εφαρμοσμένα μαθηματικά) και εξελίσσονται αυτόνομα, βρίσκοντας πολλές εφαρμογές στην καθημερινότητα.

Ένα από τα αντικείμενα της Όρασης Υπολογιστών είναι η ανάπτυξη τεχνικών για την εύρεση και ταυτοποίηση αντικειμένων σε εικόνες, που αναφέρεται γενικότερα με τον όρο Αναγνώριση Αντικειμένων. Τα ερωτήματα που προσπαθεί να απαντήσει είναι: *Πόσα «διακριτά» αντικείμενα υπάρχουν σε μία εικόνα (ύπαρξη), πού βρίσκονται (θέση) και σε ποιες από τις γνωστές κατηγορίες αντικειμένων μπορεί να ανήκουν (ταυτότητα)?* Τα μέρη του ερωτήματος αναλύονται στα παρακάτω προβλήματα:

- Ανίχνευση/Ύπαρξη (Detection): Εξακρίβωση αν ένα συγκεκριμένο αντικείμενο βρίσκεται στην εικόνα, χωρίς να αναζητείται η θέση, κλίμακα ή η πόζα του.

- Εντοπισμός/Εύρεση (Localization) της ακριβούς θέσης ενός αντικειμένου, το οποίο είναι (πιθανώς) γνωστό ότι υπάρχει στην εικόνα· γενικεύει το προηγούμενο πρόβλημα.

- Ταξινόμηση/Κατηγοριοποίηση (Classification/Identification): Η ταξινόμηση μπορεί να γίνει εφόσον έχουν εντοπιστεί επιμέρους «διακριτά» αντικείμενα, αλλά και συνολικά για μία εικόνα, στην/στις κατηγορίες που ανήκουν.

Τα προβλήματα αυτά μπορούν να αντιμετωπιστούν ανεξάρτητα ή από κοινού με πλεονεκτήματα και μειονεκτήματα και στους δύο τρόπους αντιμετώπισης. Στο πλαίσιο της Διπλωματικής Εργασίας οι όροι Ανίχνευση και Εντοπισμός, θα χρησιμοποιούνται εναλλάξ και θα αφορούν στην αναζήτηση ενός ή περισσότερων στιγμιοτύπων του επιθυμητού αντικειμένου και στον προαιρετικό προσδιορισμό της τοποθεσίας του.

Αν και οι παραπάνω στόχοι έχουν επιτευχθεί σε μεγάλο βαθμό σε ελεγχόμενα περιβάλλοντα, οι εικόνες από την καθημερινότητα εμφανίζουν πολλές προκλήσεις που οφείλονται κυρίως στους εξής παράγοντες:

- μεγάλες αποκλίσεις στο σχήμα και στη μορφή αντικειμένων (κυρίως των παραμορφώσιμων), που ανήκουν ακόμα και στην ίδια σημασιολογική κατηγορία (intra-class variations),

- παραμορφώσεις που οφείλονται στη γωνία θέασης ή εισάγονται από το οπτικό σύστημα (φακοί) και την αποθήκευση και μοντελοποιούνται μαθηματικά με αφινικούς, προβολικούς, ή άλλους γενικούς μετασχηματισμούς,
- επιλεκτικές αλλοιώσεις στη ομοιομορφία της φωτεινότητας, λόγω σκιών ή φωτισμού,
- επικαλύψεις αντικειμένων στις πραγματικές σκηνές (occlusions).

1.2 Το κεντρικό πρόβλημα της Ταξινόμησης Εικόνων

Αναπαράσταση Ψηφιακών Εικόνων στον Υπολογιστή

Οι εικόνες που προέρχονται από τον φυσικό κόσμο για να αποθηκευτούν σε ένα ψηφιακό σύστημα πρέπει να δειγματοληπτηθούν στο πεδίο του χώρου και να κβαντιστούν στο πεδίο τιμών τους, ώστε να μπορούν να αποθηκευτούν σε πεπερασμένο χώρο. Οι συνήθεις εικόνες που προέρχονται από εμπορικό εξοπλισμό έχουν D κανάλια και χωρικές διαστάσεις μήκους Πλάτος $W \times$ Ύψος H, π.χ. $W \times H \times 1$ για ασπρόμαυρες ή $D = 3$ για έγχρωμες. Κατ' αυτόν τον τρόπο αποθηκεύονται στην μνήμη ως τριδιάστατοι πίνακες. Εναλλακτικά, αλλάζοντας τη χωρική κατανομή των τιμών των πίξελ μπορούν να αναπαρασταθούν ως ένα μονοδιάστατο διάνυσμα μήκους $W \cdot H \cdot D$ ή D μονοδιάστατα διανύσματα μήκους $W \cdot H$. Η μετατροπή, στη χωρική οργάνωση, αυτού του τύπου, ονομάζεται «διανυσματοποίηση» (vectorization) και είναι χρήσιμη σε πράξεις εικόνων, όπως το φιλτράρισμα και η συνέλιξη (βλ. και Ενότητα 3.2.1).

Ταξινόμηση Εικόνων

Στόχος ενός συστήματος ταξινόμησης είναι να αναθέσει στην εικόνα εισόδου ετικέτες μίας η περισσότερων κατηγοριών από ένα σύνολο προκαθορισμένων. Η ανάθεση μπορεί να είναι απόλυτη (αμοιβαία αποκλειόμενες ή όχι κατηγορίες) ή πιθανοτική (κατανομή για το διάνυσμα κατηγοριών). Στο γενικότερο πλαίσιο της ταξινόμησης μπορεί να συμπεριληφθούν και άλλα προβλήματα όπως η ανίχνευση και ο εντοπισμός γνωστών αντικειμένων σε εικόνες, η επιλογή κατηγοριών στις οποίες ανήκει μία εικόνα ή περιεχόμενά της, η κατάτμηση, κ.τ.λ.

Η ταξινόμηση εικόνων σε προκαθορισμένες κατηγορίες υπονοεί τη χρήση επιβλεπόμενης μάθησης. Κατά την επιβλεπόμενη μάθηση δίνεται ένα επισημασμένο σύνολο δεδομένων, από το οποίο προκύπτει (π.χ. με τυχαία διαμέριση) ένα σύνολο εκπαίδευσης D (training dataset) N παραδειγμάτων. Αυτό αποτελείται από ζεύγη εικόνων $\mathbf{x}_i$ και ετικετών $\mathbf{y}_i$:

$$\mathsf{D} = \{(\mathbf{x}_i, \mathbf{y}_i),\ i = 1, ..., N\}$$

Πριν την αντιμετώπιση του προβλήματος, απαραίτητη είναι, η επιλογή των αναπαραστάσεων των εικόνων, που αναλύθηκε παραπάνω, και των ετικετών. Στο πρόβλημα δύο κατηγοριών συχνές επιλογές για τις ετικέτες είναι οι: $y_i \in \{0, 1\}$ και $y_i \in \{-1, 1\}$. Στο πρόβλημα k κατηγοριών οι

συνηθέστερες επιλογές είναι οι $y_i \in \{1, ..., k\}$ και $\mathbf{y}_i = [0,\ ...,\ 0,\ 1,\ 0,\ ...,\ 0]^T \in \mathbb{R}^k$, όπου το 1 βρίσκεται μόνο στη θέση που αντιστοιχεί στην κατηγορία που ανήκει το $\mathbf{x}_i$.

Είναι φανερό ότι για την ποσοτικοποίηση του αποτελέσματος και μέτρηση της επίδοσης του συστήματος πρέπει να γίνουν συγκρίσεις μεταξύ των πραγματικών ετικετών και των ετικετών που προκύπτουν από τον αλγόριθμο. Επομένως, η ονοματοδοσία από κάποιο σημείο της διαδικασίας και μετά έχει σημαντικό ρόλο. Γενικά, είναι ένας καλός αλγόριθμος ταξινόμησης είναι ανεξάρτητος από την ονοματοδοσία ετικετών, ωστόσο αν αυτή λαμβάνει μέρος στη διαδικασία, υπάρχουν αλγόριθμοι που μπορούν να διαχειριστούν τις εμφανιζόμενες δυσκολίες. Ένας τέτοιος αλγόριθμος είναι ο Ho-Kashyap [Ho65], που αποτελεί ένα σχήμα επίλυσης, το οποίο λύνει το πρόβλημα και βρίσκει παράλληλα και τις βέλτιστες τιμές y_i.

Το πρόβλημα της ταξινόμησης έχει πολλά κοινά με την προσέγγιση συναρτήσεων. Η προσέγγισης μίας άγνωστης καμπύλης (curve fitting) χρησιμοποιώντας πεπερασμένο πλήθος σημείων είναι γνωστή στην Αριθμητική Ανάλυση ως παρεμβολή συναρτήσεων (function approximation) και στην Στατιστική ως παλινδρόμηση (regression) και έχει στόχο να βρει μία συνάρτηση που περιγράφει - παρεμβάλει τα δεδομένα σημεία καλύτερα. Στην ταξινόμηση αναζητείται η διαχωριστική υπερεπιφάνεια ανάμεσα στις κατηγορίες.

1.3 Οργάνωση Κειμένου

Το κείμενο οργανώνεται σε 6 κεφάλαια, στα οποία παρουσιάζονται πτυχές της Μηχανικής Μάθησης με έμφαση στα Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ) και περιγράφονται τεχνικές εντοπισμού προσώπων σε εικόνες, ενώ υλοποιείται και ένας σύστημα ανίχνευσης τελευταίας τεχνολογίας.

Στο Κεφάλαιο 2 αναφέρονται τα τρία κύρια συστατικά της Μηχανικής Μάθησης: ο στόχος, το μέτρο επίδοσης και ο τρόπος μάθησης. Παραθέτονται τα κυρίως χρησιμοποιούμενα μέτρα επίδοσης, με φόντο τα ΣΝΔ και γίνεται μία εισαγωγή στη βελτιστοποίησή τους. Ως παράδειγμα συνδυασμού των τριών στοιχείων εξετάζονται οι Γραμμικοί Ταξινομητές, που αποτελούν μία απλή περίπτωση μοντελοποίησης δεδομένων και μάθησης, τα στοιχεία των οποίων χρησιμοποιούνται όμως σε συνθετότερους ταξινομητές, όπως τα ΣΝΔ. Επίσης, εισάγονται τα Νευρωνικά Δίκτυα και γίνεται μία ανασκόπηση αυτών και των δυνατοτήτων τους, δίνοντας έμφαση στα Βαθέα (ΒΝΔ) και Συνελικτικά Νευρωνικά Δίκτυα.

Στο Κεφάλαιο 3 περιγράφονται εκτενώς τα σύγχρονα Συνελικτικά Νευρωνικά Δίκτυα. Αναλύεται η δομή τους, ο τρόπος εκπαίδευσης και χρήσεις τους και θίγονται πρακτικά ζητήματα που αναφέρονται κυρίως στα ΣΝΔ, αλλά βρίσκουν εφαρμογή γενικότερα στους τομείς των Νευρωνικών Δικτύων και της Μηχανικής Μάθησης. Επίσης, περιγράφεται η αρχιτεκτονική του ΣΝΔ AlexNet, που έθεσε το 2012 τα θεμέλια για την σύγχρονη ανάπτυξη των ΣΝΔ, και αναφέρονται άλλα παρόμοια δίκτυα που έχουν εξαιρετικά αποτελέσματα στον τομέα της Ταξινόμησης και όχι μόνο.

Στο Κεφάλαιο 4 παρουσιάζονται συνοπτικά σύγχρονα εργαλεία για την ανάπτυξη ΒΝΔ, που χρησιμοποιήθηκαν και στην παρούσα Διπλωματική. Τα εργαλεία αυτά λόγω του σύνθετου έργου που επιτελούν, έχουν αναπτυχθεί με πρότυπο και εύρωστο τρόπο για να επιτυγχάνουν ταχύτητα, διαφάνεια και αξιοπιστία.

Στο Κεφάλαιο 5 γίνεται μία σύντομη ανασκόπηση στις Τεχνικές Εντοπισμού Προσώπων σε εικόνες. Επίσης, περιγράφεται αναλυτικά και υλοποιείται ένα επιτυχημένο σύστημα εντοπισμού προσώπων που προτάθηκε το 2015 και έχει ως βάση ένα ΣΝΔ παρόμοιο με το AlexNet, ενώ προτείνονται και κάποιες βελτιώσεις στο υπάρχον σύστημα.

Στο Κεφάλαιο 6 παρέχονται κατευθύνσεις για μελλοντική έρευνα στον τομέα των ΒΝΔ και του εντοπισμού προσώπων.

Κεφάλαιο 2

Μηχανική Μάθηση και Νευρωνικά Δίκτυα

Η Μηχανική Μάθηση προσφέρει τεχνικές επίλυσης προβλημάτων προσανατολισμένες στα δεδομένα (data-driven), που βασίζονται περισσότερο στην προσαρμοστικότητα των συστημάτων που αναπτύσσονται και λιγότερο στη διαίσθηση των μηχανικών. Προβλήματα υψηλότερου επιπέδου, όπως η ταξινόμηση, δεν μπορούν να αντιμετωπιστούν εύκολα με αλγοριθμικό τρόπο και γι' αυτό συχνά καταφεύγουμε σε τεχνικές Μηχανικής Μάθησης. Από τους πιο σημαντικούς εκπροσώπους αυτού του πεδίου είναι τα Νευρωνικά Δίκτυα, που εκπαιδεύονται μέσω μίας προσέγγισης δοκιμής, σφάλματος και αναπροσαρμογής. Η μάθηση στα ΝΔ βασίζεται στην αυτόματη προσαρμογή των παραμέτρων τους και η επιτυχία τους στην ιδιότητα που έχουν να συλλαμβάνουν πολύπλοκες και μη γραμμικές σχέσεις των δεδομένων.

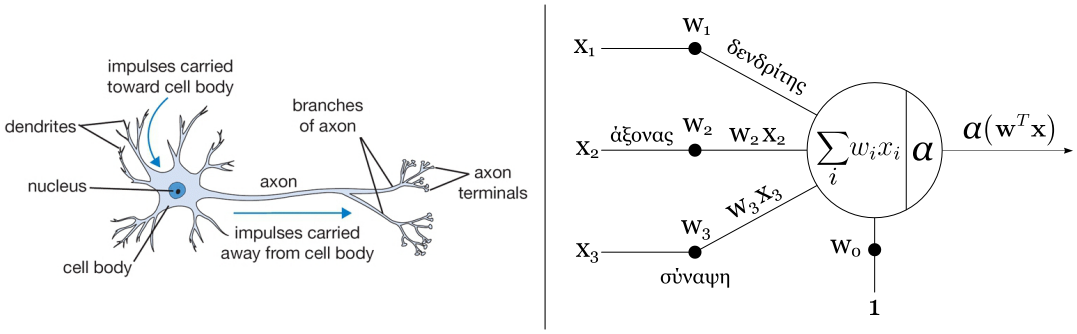
Από τη στιγμή που επιλεγεί η αρχιτεκτονική και οι υπερπαράμετροι ενός ΝΔ και γίνει η εκπαίδευση αυτό μπορεί να θεωρηθεί ως «μαύρο κουτί» που δέχεται μία είσοδο και εξάγει αποφάσεις. Ως προς μία άποψη τα ΝΔ αφαιρούν φορτίο από το σχεδιαστή του συστήματος και επιφορτώνουν υπολογιστικά τη μηχανή. Πλεονέκτημά τους είναι ότι μπορούν να εκπαιδευτούν σε πολύ μεγάλα σύνολα δεδομένων, ώστε να μπορούν να μάθουν σύνθετες αλληλεπιδράσεις των δεδομένων, ωστόσο είναι επιρρεπή στην υπερπροσαρμογή σε αυτά.

2.1 Η έννοια του Τεχνητού Νευρώνα και σύνδεση με τον Βιολογικό

Τα Τεχνητά Νευρωνικά Δίκτυα (ΝΔ) ξεκίνησαν ως μαθηματικά μοντέλα του ανθρώπινου εγκεφάλου. Το ανθρώπινο νευρικό σύστημα έχει ως δομική μονάδα τον νευρώνα και υπολογίζεται ότι περιλαμβάνει περίπου 86 δισεκατομμύρια νευρώνες, που συνδέονται με 100 τρισεκατομμύρια συνάψεις.

Ο απλοποιημένος νευρώνας του Σχήματος 2.1 λαμβάνει τα σήματα εισόδου από τους δενδρίτες, τα επεξεργάζεται και παράγει το σήμα εισόδου, που μεταφέρεται μέσω του άξονά του σε άλλους νευρώνες. Στο αντίστοιχο μαθηματικό μοντέλο τα σήματα (π.χ. x_0) που μεταδίδονται από το προηγούμενο νευρώνα συνδέονται πολλαπλασιαστικά (αναλογικά: $w_0 x_0$) μέσω των συνάψεων (βάρη) στους δενδρίτες του επόμενου νευρώνα. Η ιδέα πίσω από αυτό το μοντέλο είναι ότι τα συναπτικά βάρη είναι εκπαιδεύσιμα και ελέγχουν την δύναμη της επιρροής ενός νευρώνα σε έναν άλλο, αλλά και την κατεύθυνση επιρροής: αρνητικά βάρη – ανασταλτικός (inhibitory), θετικά βάρη – διεγερτικός (excitatory). Η επεξεργασία των σημάτων γίνεται στον πυρήνα του νευρώνα και στη συνέχεια ενεργοποιείται στέλνοντας έναν παλμό μέσω του άξονά του ή μένει ανενεργός. Η λειτουργία αυτή μοντελοποιείται με την

πράξη της πρόσθεσης των σταθμισμένων σημάτων εισόδου και η πυροδότηση μέσω της συνήθως μη γραμμικής συνάρτησης ενεργοποίησης.



Σχήμα 2.1: Αριστερά: Απεικόνιση μερών και λειτουργίας βιολογικού νευρώνα. Δεξιά: Το προσεγγιστικό μαθηματικό μοντέλο, που αποτελεί έναν τεχνητό νευρώνα.

Ένας νευρώνας έχει m εισόδους $\{x_1, ..., x_m\}$ και m συναπτικά βάρη $\{w_1, ..., w_m\}$, που πολλαπλασιάζουν τις εισόδους και ένα κατώφλι b για την πόλωση. Υπολογίζει το σταθμισμένο άθροισμα των εισόδων (γραμμικός συνδυασμός) και εφαρμόζει σε αυτό τη συνάρτηση ενεργοποίησής του (μη-γραμμικότητα) για να δημιουργήσει την απόκρισή του α (Σχέση 2.1). Συνήθως, συμφέρει να ορίσουμε το κατώφλι ως ακόμα ένα βάρος $w_0 = b$ με σταθερή είσοδο $x_0 = 1$, ώστε το μοντέλο να απλοποιηθεί.

$$\alpha\left(\sum_{i=1}^{m} w_i x_i + b\right) = \alpha\left(\sum_{i=0}^{m} w_i x_i\right) = \alpha\left(\mathbf{w}^T \mathbf{x}\right) \tag{2.1}$$

όπου $\mathbf{x} = [1,\ x_1,\ ...,\ x_m]^T$ το επαυξημένο διάνυσμα εισόδου και $\mathbf{w} = [w_0,\ w_1,\ ...,\ w_m]^T$ το επαυξημένο διάνυσμα βαρών.

Σχέση Βιολογικών και Τεχνητών Νευρωνικών Δικτύων

Αν και τα Τεχνητά Νευρωνικά Δίκτυα ξεκίνησαν ως προσπάθεια μοντελοποίησης της συμπεριφοράς του ανθρώπινου εγκεφάλου, σήμερα η εξέλιξή τους είναι σχετικά ανεξάρτητη, χωρίς να ακολουθεί πιστά την πορεία της νευροβιολογίας· παρ' όλα αυτά, αναλογίες μπορούν ακόμα να εντοπιστούν.

Το παραπάνω μοντέλο νευρώνα που παρουσιάστηκε θα μπορούσε να ονομαστεί νευρώνας γραμμικού συνδυασμού. Νέες έρευνες έχουν παρατηρήσει ότι ο ανθρώπινος εγκέφαλος έχει πολλά είδη νευρώνων με διαφορετικές λειτουργίες. Στο απλό μοντέλο οι συνάψεις αποτελούν ένα απλό βάρος, και οι δενδρίτες επιτελούν μόνο πολλαπλασιασμό. Στην πραγματικότητα, οι συνάψεις αποτελούν ένα μή γραμμικό δυναμικό σύστημα και οι δενδρίτες επιτελούν σύνθετους μή γραμμικούς υπολογισμούς. Επίσης, στους πραγματικούς νευρώνες ο ακριβής χρόνος πυροδότησης δεν είναι σταθερός, με αποτέλεσμα να μην αντιλαμβάνονται επόμενοι νευρώνες τις εξόδους των προηγούμενων την ίδια στιγμή[1].

[1] Σχόλιο: Η τεχνική αποκοπής συνδέσεων (dropout), που αναλύεται στην Ενότητα 3.3.5, μοντελοποιεί κατά κάποιο τρόπο αυτή τη συμπεριφορά.

2.2 Μοντελοποίηση ενεργοποίησης νευρώνα

Αν και οι παρακάτω συναρτήσεις αναφέρονται συνήθως ως ενεργοποιήσεις νευρώνων, μπορούν να ειδωθούν γενικότερα ως μία οικογένεια συναρτήσεων για τη μοντελοποίηση σχέσεων εισόδου-εξόδου και παραμετροποίησης ενός δικτύου, δηλαδή ως δομικά στοιχεία-συναρτήσεις που συνθέτουν πολύπλοκες, μή γραμμικές συναρτήσεις απεικόνισης ενός Βαθέος ΝΔ.

Οι συναρτήσεις ενεργοποίησης $\alpha\left(t\right)$, στο πλαίσιο μοντελοποίησης ενός νευρώνα δέχονται πάντα έναν αριθμό, το άθροισμα του ηλεκτρικού φορτίου που δέχεται ο νευρώνας, εφαρμόζουν τη μηγραμμικότητα και δίνουν αποτέλεσμα έναν άλλο αριθμό, την απόκριση/ενεργοποίηση. Παρακάτω παραθέτονται οι πιο κοινές επιλογές και αναφέρονται θετικά και αρνητικά στοιχεία τους με ορίζοντα εφαρμογής τα Συνελικτικά Νευρωνικά Δίκτυα.

Σιγμοειδής (Sigmoid)

$$s\left(t\right)=\frac{L}{1+e^{-k(t-t_0)}}$$

όπου L η μέγιστη τιμή της, t_0 το κεντρικό σημείο της καμπύλης και k η παράμετρος που καθορίζει την καμπυλότητα. Η συνάρτηση απεικονίζει την είσοδό της στο διάστημα εξόδου $[0,L]$, άρα για $L=1$ μπορεί να έχει ερμηνεία πιθανότητας για την είσοδό της t. Ειδική περίπτωση που χρησιμοποιείται ειδικά στα ΝΔ είναι για $L=1,\ x_0=0,\ k=1$ και φαίνεται στο Σχήμα 2.2:

$$\sigma\left(t\right)=\frac{1}{1+e^{-t}}=\frac{e^t}{1+e^t}$$

Ιστορικά είναι η πιο συνήθης χρησιμοποιούμενη, καθώς ερμηνεύεται άμεσα ως ρυθμός πυροδότησης ενός νευρώνα· 0 όταν είναι πλήρως απενεργοποιημένος και 1 για κορεσμένη ενεργοποίηση στη μέγιστη συχνότητα. Η έξοδος δεν είναι κεντραρισμένη γύρω από το 0, γεγονός που μπορεί να οδηγήσει σε κάποια προβλήματα. Τα δύο σημαντικότερα μειονεκτήματά της είναι τα εξής:

- Η σιγμοειδής έχει όρια κορεσμού τα 0, 1, επομένως για πολύ μικρές ή πολύ μεγάλες τιμές εισόδους η κλίση σε αυτές τις τιμές θα είναι σχεδόν μηδενική. Κατά την μάθηση με οπισθοδρόμηση, η τοπική κλίση του νευρώνα θα είναι σχεδόν μηδενική, μηδενίζοντας την «μεταφερόμενη» κλίση από τον κανόνα της αλυσίδας, με αποτέλεσμα να μην ρέει καθόλου σήμα από τον νευρώνα προς τα πίσω, δηλαδή να σταματάει τη μάθηση (βλ. και Ενότητα 2.4.4).

- Η αρχικοποίηση των βαρών σε νευρώνες με τέτοια συνάρτηση ενεργοποίησης πρέπει να είναι προσεκτική, καθώς αν αρχικά καθοριστούν τυχαία τα βάρη σε μεγάλες τιμές (θετικές ή αρνητικές), τότε ο νευρώνας μπορεί να οδηγηθεί σε κορεσμό.

Υπερβολική εφαπτομένη (tanh)

$$\tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}} = 2\sigma(2t) - 1$$

Προτιμάται (Σχήμα 2.2) σε σχέση με τη σιγμοειδή, κυρίως γιατί έχει κεντραρισμένη έξοδο γύρω από το 0 έξοδο, αλλά το γεγονός ότι κορένεται, και η διερεύνηση της χρήσης της στην ταχύτητας μάθησης (Σχήμα 3.14) σε σχέση με την επόμενη, έχουν οδηγήσει στην σταδιακή εγκατάλειψή της.

Ανορθωμένη Γραμμική ή Θετικό Μέρος ή Συνάρτηση Ράμπας (Rectified Linear Unit – ReLU)

$$\alpha(t) = \max(0, t) \tag{2.2}$$

Η συνάρτηση αυτή φαίνεται στο Σχήμα 2.2 και επιταχύνει την εκπαίδευση σε σχέση με τις προηγούμενες (Σχήμα 3.14), γεγονός που οφείλεται στην γραμμική και μη κορεσμένη έξοδό της. Επιπλέον, η υλοποίησή της είναι εύκολη και γρήγορη αφού χρειάζεται μόνο μία σύγκριση με το 0, αντίθετα με τον υπολογισμό σύνθετων συναρτήσεων (tanh, exp).

Ωστόσο, έχει και ένα μειονέκτημα: κατά τη μάθηση η ανανέωση βαρών μπορεί να οδηγήσει τα βάρη ενός νευρώνα σε έναν συνδυασμό, λόγω του οποίου ο νευρώνας δεν θα μπορεί να ενεργοποιηθεί ποτέ (άθροισμα εξόδου αρνητικό). Σε αυτή την περίπτωση η τοπική κλίση θα μηδενιστεί και ο νευρώνας δεν θα μαθαίνει εφεξής, θα είναι πρακτικά «νεκρός», δηλαδή δεν θα ενεργοποιείται ποτέ για οποιοδήποτε παράδειγμα του συνόλου δεδομένων. Αυτό μπορεί να συμβεί, με μεγαλύτερη πιθανότητα, αν ο ρυθμός μάθησης είναι μεγάλος και μπορεί να παρατηρηθεί μέχρι και το 40% των νευρώνων ενός δικτύου να είναι «νεκροί».

Διαρρέουσα Ανορθωμένη Γραμμική (Leaky ReLU) και Παραμετροποιήσιμη ReLU (PReLU)

$$\lambda(t) = \begin{cases} at & t < 0 \\ t & t \geqslant 0 \end{cases} = \max(0,\ t) + a\min(0,\ t) \tag{2.3}$$

Η συνάρτηση αυτή προσπαθεί να διορθώσει το πρόβλημα των απενεργοποιημένων («νεκρών») νευρώνων, προσθέτοντας μία μικρή κλίση για αρνητικά αθροίσματα και φαίνεται στο Σχήμα 2.2. Προτάθηκε στο [Maas13] με σταθερά επιλεγμένο a και στο [He15] βελτιώθηκε με το a να αποτελεί παράμετρο προς μάθηση. Η δεύτερη επιλογή, σύμφωνα με τους συγγραφείς, βελτιώνει κατά πολύ τα αποτελέσματα της μάθησης, ωστόσο εισάγει επιπλέον ένα μεγάλο σύνολο παραμέτρων προς μάθηση.

Παραλλαγές των προαναφερόμενων

Λιγότερο χρησιμοποιούμενες συναρτήσεις ενεργοποίησης είναι οι: απόλυτη τιμή $|t|$, αυτούσια ή επιβαλλόμενη σε κάποια από τις προηγούμενες, η γραμμική (σε όλο το $\mathbb{R}$) t, η βηματική με πεδίο τιμών $[0, 1]$, είτε $[-1, 1]$ και η γραμμική με κορεσμό στις τιμές $L_+,\ L_-$ ή κατωφλιοποιημένη γραμμική ή «σκληρή» υπερβολική εφαπτομένη (hard tanh) $\max\left(L_-,\ \min\left(t,\ L_+\right)\right)$. Τέλος, μία ομαλοποιημένη, διαφορίσιμη έκδοση της ReLU είναι η Softplus $\log\left(1 + e^t\right)$.

Μονάδα καλύτερων παραμέτρων (Maxout)

Μία διαφορετική προσέγγιση (δεν θεωρείται συνάρτηση ενεργοποίησης), είναι οι «νευρώνες βέλτιστων παραμέτρων» ή μονάδα maxout, που εισήχθησαν στο [Good13] και γενικεύουν πολλές από τις παραπάνω συναρτήσεις (π.χ. την οικογένεια ReLU και την $|\cdot|$). Η έξοδος μιας μονάδας με εισόδους $\mathbf{x}$ δίνεται από την 2.4. Διαισθητικά μια μονάδα maxout μπορεί να θεωρηθεί ως μία ομάδα «υπο-νευρώνων», οι οποίοι εκπαιδεύονται παράλληλα, αλλά η έξοδος της ομάδας είναι κάθε φορά η μέγιστη των εξόδων των μελών της.[2]

$$\alpha = \max_i \mathbf{w}_i^T \mathbf{x} \tag{2.4}$$

Το πλήθος των «υπο-νευρώνων» i μπορεί να θεωρηθεί ως υπερ-παράμετρος. Αν επιλεγεί π.χ. $i = 2$, τότε με έναν τέτοιο νευρώνα κατασκευάζονται η ReLU ($\mathbf{w}_1 = \mathbf{0}$) και η $|\cdot|$. Αυτή η συνάρτηση έχει όλα τα πλεονεκτήματα αυτών που γενικεύει, αλλά σε αντίτιμο εισάγει $i - 1$ φορές παραπάνω βάρη. Επίσης, η διαδικασία εκπαίδευσης μίας τέτοιας μονάδας μπορεί να θεωρηθεί ως μάθηση της καλύτερης συνάρτησης ενεργοποίησης, και έχει αποδειχθεί ότι με τη χρήση ικανού αριθμού μονάδων μπορεί να επιτευχθεί προσέγγιση οποιασδήποτε κυρτής συνάρτησης σε οποιαδήποτε επιλεγμένη πιστότητα.

Άλλες επιλογές

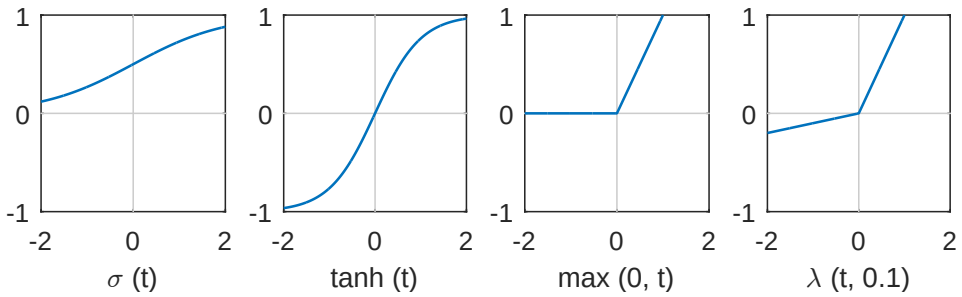
Στο πλαίσιο των δομικών συναρτήσεων μπορούν να αναφερθούν και άλλες συναρτήσεις, που δεν εφαρμόζονται σε γραμμικό συνδυασμό της εισόδου αλλά απευθείας σε αυτή όπως:

- Συνάρτηση Softmax. Απεικόνιση διανύσματος εισόδου σε διάνυσμα εξόδου με στοιχεία στο διάστημα $[0, 1]$ και με ερμηνεία πιθανότητας.

$$\sigma_i\left(\mathbf{x}\right) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2.5}$$

- Μονάδα RBF. Η έξοδος ενός νευρώνα είναι κυκλικά συμμετρική $\alpha\left(\|\mathbf{x} - c\|, \sigma\right)$.

[2] Εναλλακτικά, μπορεί να θεωρηθεί ως εκπαίδευση νευρώνων ενοποιημένη με τοπικό, προεπιλεγμένο max-pooling.



Σχήμα 2.2: Τέσσερις συνηθέστερες συναρτήσεις ενεργοποίησης.

2.3 Τεχνικές Μηχανικής Μάθησης με και χωρίς Επίβλεψη

Απαραίτητο στοιχείο της μάθησης, είτε είναι βιολογική, είτε είναι μηχανική είναι τα δεδομένα. Από αυτά, χρησιμοποιώντας στοιχεία νοημοσύνης, ο άνθρωπος ή μια μηχανή μπορεί να μάθει, να εξάγει συμπεράσματα και να κάνει προβλέψεις. Τα δεδομένα έχουν άφθονη και συσχετισμένη πληροφορία, που είναι πολλές φορές περιττή για τη μάθηση. Οι αλγόριθμοι μάθησης έχουν στόχο να μειώσουν τη διάσταση αυτής της πληροφορίας, κρατώντας τα απαραίτητα διακριτικά χαρακτηριστικά. Στο χώρο των εξαγόμενων χαρακτηριστικών συγγενείς κατηγορίες δεδομένων θα πρέπει να είναι εγγύτερα.

Σε πολλά προβλήματα, όπως στην ταξινόμηση, είναι διαθέσιμες ή εύκολα αποκτίσιμες μεγάλες βάσεις δεδομένων, κάνοντας εφικτή τη χρήση μεθόδων μάθησης για την αντιμετώπισή τους. Ωστόσο, αυτά τα δεδομένα, είτε λόγω του μεγάλου πλήθους τους, είτε λόγω της μορφής τους δεν είναι επισημασμένα ή κατηγοριοποιημένα. Σε αυτή την περίπτωση εφαρμόζονται τεχνικές μάθησης χωρίς επίβλεψη (unsupervised learning), που στόχο έχουν την ομαδοποίηση/συσταδοποίηση (clustering) των δεδομένων, δηλαδή την εύρεση της δομής και των αλληλοσυσχετίσεων των δεδομένων.

Όταν τα δεδομένα είναι επισημασμένα τότε μαζί με κάθε στοιχείο τους ακολουθούν και πληροφορίες/μετα-δεδομένα, που περιγράφουν το περιεχόμενό τους. Για παράδειγμα, για το πρόβλημα της ταξινόμησης εικόνων αρκεί μαζί με την εικόνα να υπάρχει μία ετικέτα για το κύριο αντικείμενο που απεικονίζει. Για το συνθετότερο πρόβλημα της κατάτμησης κάθε πίξελ πρέπει να έχει μία ετικέτα για την περιοχή που ανήκει.

Σήμερα το internet και οι νέες τεχνολογίες έχουν διευκολύνει τη συλλογή δεδομένων και κυρίως μέσω τεχνικών πληθοπορισμού (crowdsourcing) έχει επιτευχθεί μεγάλη προσπάθεια για τη δημιουργία πλήρως επισημασμένων βάσεων. Οι βάσεις αυτές έδωσαν και το έναυσμα για την άνθιση της επιβλεπόμενης μάθησης (supervised learning). Σε αυτή τη Διπλωματική Εργασία θα επικεντρωθούμε σε τεχνικές επιβλεπόμενης μάθησης.

2.4 Συστατικά Επιβλεπόμενης Μηχανικής Μάθησης

Η Μηχανική Μάθηση, ως βασικό συστατικό της Τεχνητής Νοημοσύνης, αναπτύσσει μεθόδους για να λύσει μία ευρεία γκάμα προβλημάτων και αποτελείται από τα επόμενα τρία κύρια μέρη [Beng15, Karp15].

Ο Στόχος / Το Έργο

Ο στόχος ενός συστήματος μάθησης είναι να αποκτήσει «γνώση» από τα δεδομένα που του παρέχονται, ώστε να μπορεί να παίρνει αποφάσεις και να εξάγει προβλέψεις, σχετικά το έργο που του έχει ανατεθεί. Στη συνέχεια, συγκεκριμενοποιούμε τις έννοιες αυτές στο κεντρικό πρόβλημα της ταξινόμησης σε εικόνες, που είναι και το θέμα της Διπλωματικής Εργασίας.

Ένα σύστημα ταξινόμησης εικόνων έχει στόχο την κατηγοριοποίηση μίας εικόνας σε μία ομάδα με γνωστά παραδείγματα. Μαθηματικά αυτό μπορεί να περιγραφεί ως μία παραμετροποιήσιμη απεικόνιση της εισόδου (εικόνας) σε έξοδο που αντιπροσωπεύει κατηγορίες ή/και αντικείμενα. Η συνάρτηση *απεικόνισης/βαθμολόγησης (mapping/score function)* μπορεί να έχει διανυσματική (πολλαπλές κατηγορίες/αντικείμενα) ή βαθμωτή έξοδο. Στην πιο απλή περίπτωση είναι μία αφινική συνάρτηση (π.χ. Γραμμικοί Ταξινομητές – Ενότητα 2.4.2), ενώ τις πλέον σύνθετες απεικονίσεις αποτελούν τα Νευρωνικά Δίκτυα.

Το Μέτρο Επίδοσης

Για να ποσοτικοποιηθεί η επίδοση του συστήματος χρειάζεται ένα μέτρο επίδοσης/επιτυχίας του αποτελέσματος. Η μέτρηση της ποιότητας αποτελεσμάτων είναι αρκετές φορές δύσκολο να γίνει «κατηγορηματικά» και η ποσοτικοποίηση είναι υποκειμενική. Συγκεκριμένα, για το πρόβλημα της ταξινόμησης πρέπει να οριστεί η αποτελεσματικότητα (accuracy) μέσω μίας *συνάρτησης κόστους (απωλειών ή εμπειρικού ρίσκου) (cost/loss/risk function)*, η οποία ποσοτικοποιεί την απόκλιση της προβλεπόμενης από την επιθυμητή έξοδο, δηλαδή το πόσο καλά έχουν επιλεγεί οι παράμετροι της συνάρτησης απεικόνισης, για παράδειγμα κόστος SVM και Softmax (Σχέσεις 2.15 και 2.18 αντίστοιχα).

Η Μάθηση

Η διαδικασία μάθησης είναι αυτή που δίνει σε ένα σύστημα Μηχανικής Μάθησης τη δυνατότητα να επιτελεί τους σκοπούς του και χρειάζεται την επιλογή της *μεθόδου μάθησης* και τα *δεδομένα* ή αλλιώς την «εμπειρία». Έχοντας ένα μέτρο της επίδοσης, που εκφράζεται συναρτήσει των παραμέτρων της απεικόνισης μπορούμε να το βελτιστοποιήσουμε, βρίσκοντας έτσι την παραμετροποίηση της απεικόνισης που δίνει καλύτερα αποτελέσματα. Συχνά, αλλά όχι πάντα, η μάθηση προέρχεται από την ελαχιστοποίηση της συνάρτησης κόστους.

Η *βελτιστοποίηση (optimization)* στην μηχανική μάθηση διαφέρει από τη συνήθη βελτιστοποίηση. Συγκεκριμένα δρα έμμεσα, γιατί δεν βελτιστοποιεί άμεσα την συνάρτηση απεικόνισης, που είναι και ο στόχος, αλλά μέσω της μείωσης ενός ποιοτικού κόστους που καθορίζεται από εμάς. Λόγω αυτής της έμμεσης διαδικασίας είναι αναγκαία η χρήση των παραδειγμάτων μάθησης (σύνολο εκπαίδευσης), τα οποία συνήθως παρουσιάζονται επαναληπτικά στο σύστημα. Η μάθηση μπορεί να είναι με επίβλεψη ή χωρίς όπως αναφέρθηκε και στην Ενότητα 2.3.

2.4.1 Ροή πληροφορίας κατά την Επιβλεπόμενη Μάθηση

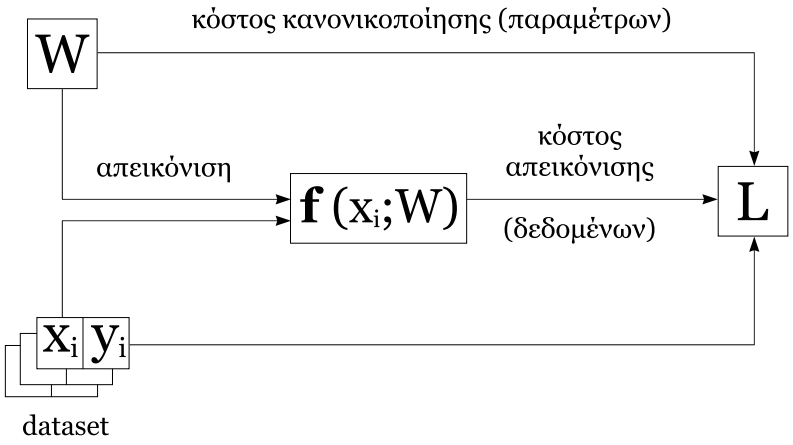
Για την αντιμετώπιση ενός προβλήματος μέσω της Μηχανικής Μάθησης (π.χ. ταξινόμηση εικόνων), αρχικώς πρέπει να γίνουν επιλογές για τα 3 στοιχεία που περιγράφονται στην Ενότητα 2.4: η

παραμετρική απεικόνιση των εικόνων στο σύνολο των επιθυμητών εξόδων, το μέτρο επίδοσης της απεικόνισης και ο τρόπος μάθησης.

Η επίδοση της απεικόνισης εικόνων σε αριθμούς (σκορ) συνήθως ποσοτικοποιείται μέσω μία συνάρτησης κόστους, που έχει τη γενική μορφή της Σχέσης 2.6. Το κόστος είναι συνάρτηση των παραδειγμάτων $\mathbf{x}_i$ και των παραμέτρων W. Τα $\mathbf{x}_i$ συνήθως είναι γνωστά και σταθερά, ενώ τα W μεταβλητά. Στο Σχήμα 2.3 απεικονίζεται η ροή πληροφορίας από το σύνολο δεδομένων μέχρι το κόστος L.

$$L(W) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} L_i(W)}_{\text{κόστος δεδομένων}} + \underbrace{\lambda R(W)}_{\text{κόστος παραμέτρων}} \tag{2.6}$$

Η άθροιση στην σχέση 2.6 γίνεται ως προς όλα τα $i = 1, ..., N$ παραδείγματα εκπαίδευσης και L_i είναι το κόστος δεδομένων για κάθε παράδειγμα. Το λ αποτελεί υπερ-παράμετρο και συνήθως εκλέγεται με πειραματισμό (π.χ. στο σύνολο επαλήθευσης). Καθορίζοντας τις L_i και R μπορούμε να ελαχιστοποιήσουμε το L ως προς τις παραμέτρους, ώστε να βρούμε τις βέλτιστες τιμές τους που ελαχιστοποιούν το επιλεγέν κόστος.



Σχήμα 2.3: Διάγραμμα γενικής ροής πληροφορίας κατά την επιβλεπόμενη μάθηση.

Στο Σχήμα 2.3 φαίνεται η γενική διαδικασία που ακολουθείται κατά την επιβλεπόμενη μάθηση. Το σύνολο δεδομένων εκπαίδευσης D θεωρείται σταθερό και δεδομένο[3] και οι παράμετροι (βάρη) W μεταβλητές. Με τη χρήση αυτών η παραμετρική συνάρτηση απεικόνισης **f** μετατρέπει την είσοδο (εικόνα) σε μία έξοδο (σκορ), την οποία συγκρίνει η συνάρτηση L με την επιθυμητή ετικέτα και προκύπτει το κόστος. Η συνάρτηση L εκτός από το κόστος απεικόνισης συχνά περιέχει και έναν όρο κόστους κανονικοποίησης βαρών. Η βελτιστοποίηση γίνεται ελαχιστοποιώντας τις μερικές παραγώγους της L ως προς τα βάρη W. Η επιλογή του τρόπου μάθησης είναι γενικά δεδομένη και γίνεται με κάποιον αλγόριθμο κατάβασης δυναμικού.

[3] Μπορούμε να θεωρήσουμε τα $\mathbf{x}_i$ μεταβλητά π.χ. στα ΣΝΔ όταν εκτελούμε οπισθοδιάδοση με παραγώγους ως προς $\mathbf{x}_i$ για βρούμε ποια πρότυπα στην εικόνα εισόδου ενεργοποιούν ένα χάρτη χαρακτηριστικών.

2.4.2 Μοντέλο απεικόνισης

Επικεντρωνόμαστε στο πρόβλημα ταξινόμησης εικόνων, όπου έχουμε ένα επισημασμένο σύνολο εκπαίδευσης N εικόνων $\mathsf{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$, όπου η κάθε εικόνα θεωρείται ως ένα διάνυσμα $\mathbf{x}_i \in \mathbb{R}^m$ και ανήκει σε μία ή περισσότερες από k κατηγορίες. Στόχος είναι να βρεθεί μία διανυσματική συνάρτηση που θα δέχεται ως είσοδο την εικόνα και θα έχει έξοδο βαθμολογίες (σκορ) για κάθε μία από τις κατηγορίες. Όσο μεγαλύτερο είναι το σκορ για κάποια κατηγορία τόσο πιθανότερο είναι η εικόνα να ανήκει σε αυτήν.

Αναζητούμε δηλαδή, k συναρτήσεις $f : \mathbb{R}^m \mapsto \mathbb{R}$ που απεικονίζουν τα πίξελ της εικόνας σε σκορ για την αντίστοιχη κατηγορία. Η πιο απλή επιλογή περιγραφής δεδομένων είναι το γραμμικό μοντέλο, το οποίο απεικονίζει ένα διάνυσμα εικόνας $\mathbf{x}$ μέσω των παραμέτρων/βαρών $\mathbf{w}$ σε έναν αριθμό (σκορ):

$$f\left(\mathbf{x}; \mathbf{w}\right) = \sum_{j=1}^{m} w_j x_j + b = \sum_{j=0}^{m} w_j x_j = \mathbf{w}^T \mathbf{x} \tag{2.7}$$

όπου $\mathbf{x} = [1,\ x_1,\ ...,\ x_m]^T \in \mathbb{R}^n$ το επαυξημένο διάνυσμα εισόδου και $\mathbf{w} = [w_0,\ w_1,\ ...,\ w_m]^T \in \mathbb{R}^n$ το επαυξημένο διάνυσμα βαρών. Το κάθε βάρος καθορίζει πόσο συμβάλει το κάθε στοιχείο της εισόδου στον υπολογισμό του σκορ. Ορίζοντας k τέτοιες συναρτήσεις, κατασκευάζουμε k ανεξάρτητα γραμμικά μοντέλα με τα αντίστοιχα σύνολα βαρών $\mathbf{w}_i$, όπου το i μοντέλο εξάγει ένα σκορ για την i κατηγορία.

$$f_i\left(\mathbf{x}; \mathbf{w}_i\right) = \mathbf{w}_i^T \mathbf{x} \tag{2.8}$$

Αν ομαδοποιήσουμε αυτές τις συναρτήσεις καταλήγουμε σε μία διανυσματική συνάρτηση $\mathbf{f} = [f_1,\ f_2,\ ...,\ f_k]^T : \mathbb{R}^n \mapsto \mathbb{R}^k$, που απεικονίζει την είσοδο σε k σκορ ανεξάρτητα για κάθε κατηγορία:

$$\mathbf{f}(\mathbf{x}; W) = W\mathbf{x} \tag{2.9}$$

όπου W είναι ο συγκεντρωτικός πίνακας βαρών μεγέθους $k \times n$ με τα βάρη για όλες τις κατηγορίες:

$$W = \begin{bmatrix} — & \mathbf{w}_1^T & — \\ — & \mathbf{w}_2^T & — \\ & \vdots & \\ — & \mathbf{w}_k^T & — \end{bmatrix}$$

Αυτή η επιλογή συνάρτησης απεικόνισης οδηγεί στους Γραμμικούς Ταξινομητές (Linear Classifiers), οι οποίοι ως βάση χρησιμοποιούνται και σήμερα στα ΣΝΔ, παρά την απλότητά τους. Για παράδειγμα, ένας νευρώνας υπολογίζει μία γραμμική απεικόνιση των εισόδων και την περνάει από τη συνάρτηση ενεργοποίησής του. Ομοίως στα ΣΝΔ, που εξετάζονται στο Κεφάλαιο 3, το επίπεδο συνέλιξης και το πλήρως συνδεδεμένο επίπεδο έχουν ως βάση αυτή τη γραμμική απεικόνιση.

Μέχρι στιγμής τα σκορ για κάθε κατηγορία δεν έχουν κάποια ιδιαίτερη ερμηνεία (π.χ. πιθανότη-

τας)· είναι αυθαίρετοι αριθμοί και μία εικόνα ανήκει στην κατηγορία που έχει το μεγαλύτερο σκορ. Στην επόμενη Ενότητα περιγράφονται οι πιο συνηθισμένες συναρτήσεις κόστους, μέσω των οποίων ορίζεται ένα κριτήριο επιτυχίας της απεικόνισης. Τέλος, με την ελαχιστοποίηση της συνάρτησης κόστους βρίσκονται τα βέλτιστα βάρη.

2.4.3 Συναρτήσεις Κόστους

Η επόμενη επιλογή που χρειάζεται να γίνει είναι η συνάρτηση κόστους. Σε αρκετές περιπτώσεις στη βιβλιογραφία συνηθίζεται το «επιθυμητό» σκορ της συνάρτησης απεικόνισης να ταυτίζεται με την ετικέτα της κατηγορίας. Αν και αυτή η επιλογή δεν δημιουργεί προβλήματα για τη δυαδική ταξινόμηση με συμμετρικές ετικέτες, γενικά επιβάλλει περιορισμούς στη διαδικασία μοντελοποίησης και ταξινόμησης και δεν είναι καλή στρατηγική για να χρησιμοποιηθεί εδώ. Στη συνέχεια, παρουσιάζονται τέσσερις δημοφιλείς συναρτήσεις κόστους που μπορούν να χρησιμοποιηθούν και σε άλλα προβλήματα εκτός της ταξινόμησης, αλλά και με οποιαδήποτε συνάρτηση απεικόνισης.

Κόστος Perceptron

Το κόστος Perceptron αναφέρεται για πληρότητα και δε χρησιμοποιείται σήμερα, αφού η εκπαίδευση με αυτό συγκλίνει μόνο για γραμμικά διαχωρίσιμα προβλήματα και επιβάλλει αρκετούς περιορισμούς. Για δύο κατηγορίες με ετικέτες $y_i = \pm 1$ χρειάζεται ένα σύνολο βαρών $\mathbf{w}$ και το κόστος γράφεται:

$$L = \sum_{\text{sign}(\mathbf{w}^T\mathbf{x}_i) \neq y_i} \text{sign}\left(\mathbf{w}^T\mathbf{x}_i\right)\mathbf{w}^T\mathbf{x}_i = \sum_{\text{sign}(\mathbf{w}^T\mathbf{x}_i) \neq y_i} -y_i\mathbf{w}^T\mathbf{x}_i \tag{2.10}$$

όπου η άθροιση γίνεται για τα λάθος ταξινομημένα παραδείγματα.

Μέσο Τετραγωνικό Σφάλμα (MSE)

Στην περίπτωση των πολλών κατηγοριών για το παράδειγμα $(\mathbf{x}_i, \mathbf{y}_i)$ οι ετικέτες έχουν τη μορφή $\mathbf{y}_i = [0,\ ...,\ 0,\ 1,\ 0,\ ...,\ 0]^T$ και το συνολικό κόστος εκφράζεται ως:

$$L = E\left[\|\mathbf{y} - \mathbf{f}\left(\mathbf{x}; W\right)\|^2\right] \tag{2.11}$$

Ο υπολογισμός αυτού του κριτηρίου προϋποθέτει να έχουμε διαθέσιμη στατιστική πληροφορία και να υπολογίσουμε τον πίνακα συνδιασποράς όλων των παραδειγμάτων. Επειδή στην πράξη τα στατιστικά στοιχεία είναι άγνωστα ή δύσκολα υπολογίσιμα (π.χ. λόγω μεγέθους συνόλου δεδομένων), από τη θεωρία στοχαστικής προσέγγισης προκύπτει ο αλγόριθμος Ελαχίστων Μέσων Τετραγώνων (LMS) ή αλγόριθμος Widrow-Hoff, που συγκλίνει ασυμπτωτικά στην MSE λύση ανεξάρτητα από τη γραμμική διαχωρισιμότητα του προβλήματος. Η μορφή του αλγορίθμου μπορεί να αντιστοιχηθεί στον αλγόριθμο εκπαίδευσης ενός γραμμικού νευρώνα, που είναι γνωστός ως Adaline.

Η ίδια μορφή κόστους μπορεί να χρησιμοποιηθεί και σε προβλήματα προσέγγισης συναρτήσεων (παλινδρόμησης). Σε αυτό το περιβάλλον τα ζεύγη $(\mathbf{x}_i, \mathbf{y}_i)$ αποτελούν τις δεδομένες συντεταγμένες της συνάρτησης που επιθυμούμε να προσεγγίσουμε.

Μία προσέγγιση του κόστους MSE παίρνουμε αγνοώντας τη μέση τιμή και θεωρώντας το επόμενο κόστος:

$$L_i = \|\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i; W)\|^2 = \sum_{j=1}^{k} \left([\mathbf{y}_i]_j - f_j(\mathbf{x}_i; \mathbf{w}_j) \right)^2 \tag{2.12}$$

δηλαδή προσεγγίζουμε τη μέση τιμή με το μέσο όρο[4]. Με αυτή την επιλογή η τελική συνάρτηση κόστους 2.6 για όλα τα παραδείγματα θα είναι:

$$L = \frac{1}{N} \sum_{i=1}^{N} L_i + \lambda R = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} \left([\mathbf{y}_i]_j - f_j(\mathbf{x}_i; \mathbf{w}_j) \right)^2 + \lambda R(W) \tag{2.13}$$

Αυτή είναι η Μέθοδος Ελαχίστων Τετραγώνων.

Κόστος τύπου SVM

Μία άλλη επιλογή που απευθύνεται κυρίως στο πρόβλημα της ταξινόμησης είναι το κόστος SVM που προτάθηκε στο [West99], χρησιμοποιεί το Hinge loss και έχει τη μορφή:

$$L_i = \sum_{j \neq y_i} \max(0,\ f_j(\mathbf{x}_i; \mathbf{w}_j) - f_{y_i}(\mathbf{x}_i; \mathbf{w}_{y_i}) + \Delta) \tag{2.14}$$

όπου οι ετικέτες μπορούν να ονομαστούν $y_i \in \{1, ..., k\}$ και η άθροιση γίνεται για $j = 1, ..., k \wedge j \neq y_i$.

Σύμφωνα με αυτή την επιλογή, για να μην συσσωρεύεται κόστος από το παράδειγμα $\mathbf{x}_i$ πρέπει το σκορ f_{y_i} που αντιστοιχεί στη σωστή κατηγορία y_i να είναι τουλάχιστον μεγαλύτερο από Δ από όλα τα υπόλοιπα σκορ. Το κόστος για όλα τα παραδείγματα και για γραμμική μοντελοποίηση από τις 2.6, 2.14 είναι:

$$L = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq y_i} \max\left(0,\ \mathbf{w}_j^T \mathbf{x}_i - \mathbf{w}_{y_i}^T \mathbf{x}_i + \Delta\right) + \lambda R(W) \tag{2.15}$$

Εναλλακτικά, μπορεί να χρησιμοποιηθεί το τετραγωνικό Hinge loss $\left(\max(0,\ \cdot)\right)^2$, που ονομάζεται L2-SVM. Τέλος, μία άλλη επιλογή είναι και το κόστος του δομημένου SVM (Structured SVM). Πειράματα που δημοσιεύονται στο [Tang13] δείχνουν βελτίωση της επίδοσης ταξινομητών με ΣΝΔ και χρήση κόστους τύπου SVM ή αλλιώς μεγίστου περιθωρίου (max-margin based loss).

Με αυτή τη διατύπωση του κόστους δεν χρειάζονται περιορισμοί στην επιλογή των ετικετών (όπως στα πρώτα SVM) αφού δεν εμφανίζονται στο κόστος. Επίσης, για την ταξινόμηση σε πολ-

[4] Υπό ήπιες προϋποθέσεις το άθροισμα των τετραγωνικών σφαλμάτων τείνει στη λύση MSE για μεγάλες τιμές του N.

λές κατηγορίες με τα κλασικά SVM πρέπει να υιοθετηθούν προσεγγίσεις του τύπου μία-έναντι-μίας ή μία-έναντι-όλων, που δεν είναι τόσο εύρωστες όσο η παραπάνω επιλογή.

Κόστος Διεντροπίας

Μία τελευταία επιλογή που κερδίζει έδαφος τα τελευταία χρόνια, είναι η χρήση του κόστους διεντροπίας (cross-entropy loss), που προέρχεται από τη Θεωρία της Πληροφορίας και «δουλεύει» με πιθανότητες. Για το λόγο αυτό, τα σκορ που προκύπτουν από τη συνάρτηση απεικόνισης πρέπει πρώτα να αποκτήσουν μία ερμηνεία πιθανότητας. Αυτό μπορεί να γίνει με πολλούς τρόπους, όπως διαιρώντας με το διάνυσμα **f** με το μέγιστο στοιχείο του. Συνήθως όμως, αυτό επιτυγχάνεται καλύτερα περνώντας το **f** από τη συνάρτηση Softmax (βλ. Ενότητα 2.2). Οι ετικέτες ακολουθούν την ονοματοδοσία $y_i \in \{1, ..., k\}$. Το κόστος παραδείγματος είναι:

$$L_i = -\log\left(\sigma_{y_i}\left(\mathbf{f}\right)\right) = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) = -f_{y_i} + \log\sum_{j=1}^{k} e^{f_j} \tag{2.16}$$

Έτσι το συνολικό κόστος 2.6 για όλα τα παραδείγματα γίνεται:

$$\begin{aligned} L &= -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) + \lambda R \\ &= -\frac{1}{N}\sum_{i=1}^{N} f_{y_i} + \frac{1}{N}\sum_{i=1}^{N}\left(\log\sum_{j=1}^{k} e^{f_j}\right) + \lambda R \end{aligned} \tag{2.17}$$

όπου οι εξαρτήσεις από τα **w** και τα $\mathbf{x}_i$ έχουν παραληφθεί για απλότητα. Στην περίπτωση της γραμμικής μοντελοποίησης το γενικό κόστος παίρνει τη μορφή:

$$\begin{aligned} L &= -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{\mathbf{w}_{y_i}^T\mathbf{x}_i}}{\sum_j e^{\mathbf{w}_j^T\mathbf{x}_i}}\right) + \lambda R\left(W\right) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_{y_i}^T\mathbf{x}_i + \frac{1}{N}\sum_{i=1}^{N}\left(\log\sum_{j=1}^{k} e^{\mathbf{w}_j^T\mathbf{x}_i}\right) + \lambda R\left(W\right) \end{aligned} \tag{2.18}$$

Η ταξινόμηση με αυτή τη συνάρτηση κόστους ονομάζεται και παλινδρόμηση softmax ή πολυωνυμική λογιστική παλινδρόμηση (multinomial logistic regression).

Το κόστος για το παράδειγμα $\mathbf{x}_i$ μπορεί να ερμηνευτεί ως ο αρνητικός λογάριθμος της εκ των υστέρων πιθανότητας της κατηγορίας y_i που ανήκει:

$$P\left(y_i|\mathbf{x}_i; W\right) = \frac{e^{f_{y_i}}}{\sum_j e^{f_j}} \tag{2.19}$$

Με αυτή την πιθανοτική ερμηνεία η ελαχιστοποίηση του συνολικού κόστους μπορεί να θεωρηθεί ως

ελαχιστοποίηση του αρνητικού λογαρίθμου της πιθανοφάνειας των κατηγοριών y_i, δηλαδή ακολουθείται μία προσέγγιση εκτίμησης μέγιστης πιθανοφάνειας (MLE). Σε αυτό το πλαίσιο ο όρος κανονικοποίησης R μπορεί να θεωρηθεί ότι επιβάλλει τα βάρη να ακολουθούν μία gaussian κατανομή (gaussian prior) και αντίστοιχα να εκτελούμε εκτίμηση με τη μέθοδο μέγιστης εκ των υστέρων πιθανοφάνειας (MAP).

Από τη Θεωρία της Πληροφορίας η εντροπία μεταξύ της «πραγματικής» k-διάστατης κατανομής $\mathbf{y}_i = [0,\ ...,\ 0,\ 1,\ 0,\ ...,\ 0]$ και της εκτιμώμενης $\boldsymbol{\sigma}\left(\mathbf{f}\right)$ είναι:

$$H\left(\mathbf{y}_i, \boldsymbol{\sigma}\right) = -\sum_{j=1}^{k} \left[\mathbf{y}_i\right]_j \log\left(\sigma_j\left(\mathbf{f}\right)\right) = -\log\left(\sigma_{arg1(\mathbf{y}_i)}\left(\mathbf{f}\right)\right) = L_i \tag{2.20}$$

όπου η συνάρτηση $arg1\left(\cdot\right)$ επιστρέφει το δείκτη της μονάδας στο διάνυσμα $\mathbf{y}_i$.

Ελαχιστοποιώντας το κόστος, ελαχιστοποιούμε τη διεντροπία ή ισοδύναμα την απόσταση μεταξύ των δύο κατανομών, καθώς η εντροπία μπορεί να γραφεί και συναρτήσει της απόκλισης Kullback-Leibler (K-L divergence) ως εξής:

$$H\left(\mathbf{y}_i, \boldsymbol{\sigma}\right) = H\left(\mathbf{y}_i\right) + D_{KL}\left(\mathbf{y}_i \| \boldsymbol{\sigma}\right) = D_{KL}\left(\mathbf{y}_i \| \boldsymbol{\sigma}\right) \tag{2.21}$$

αφού η εντροπία μίας διακριτής δ κατανομής είναι 0.

2.4.4 Βελτιστοποίηση Συναρτήσεων Κόστους

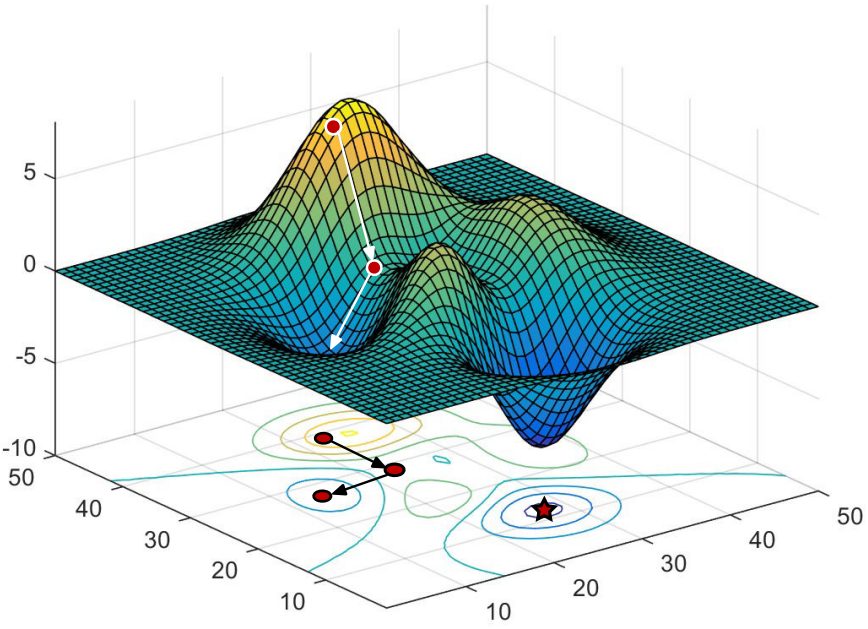
Σε αυτή την Ενότητα θα αναπτυχθεί το τρίτο και σημαντικότερο στοιχείο της Μηχανικής Μάθησης, η Βελτιστοποίηση των συναρτήσεων κόστους, μέσω της οποίας επιτυγχάνεται η μάθηση. Η συνάρτηση L εκφράζει το κόστος που έχουμε επιλέξει αν τα αποτελέσματα της ταξινόμησης δεν είναι τα επιθυμητά· επομένως, με κάποιο τρόπο πρέπει να ελαχιστοποιηθεί ως προς τα ορίσματά της, δηλαδή τα βάρη/παραμέτρους.

Μία πρώτη αντιμετώπιση είναι να δοκιμάσουμε τυχαίες τιμές βαρών και να επιλέξουμε τα βάρη που δίνουν την ελάχιστη τιμή της L. Επέκταση αυτής της σκέψης είναι να επιλέξουμε αρχικά μία τυχαία τιμή των βαρών και στη συνέχεια να δοκιμάζουμε μικρές αθροιστικές διακυμάνσεις, κρατώντας τις νέες τιμές βαρών, που έχουν μικρότερο κόστος. Αν και αυτές οι τεχνικές είναι αρκετά χρήσιμες[5], η διαστασιμότητα του προβλήματος είναι συνήθως πάρα πολύ μεγάλη (μέχρι και τάξης 10^6), γεγονός που καθιστά απαγορευτικές τέτοιες μεθόδους.

Η βιβλιογραφία πάνω στη Βελτιστοποίηση είναι μεγάλη, γι' αυτό θα επικεντρωθούμε στην πιο συχνά χρησιμοποιούμενη μέθοδο στα ΝΔ, αυτή της Κατάβασης Δυναμικού/Κλίσης (Gradient Descent). Η ανάπτυξη του βασικού αλγορίθμου, προεκτάσεις του και αναφορά άλλων μεθόδων γίνεται στην Ενότητα 3.3.7.

[5] Η τυχαία αναζήτηση (Random Search) είναι χρήσιμη σε περιπτώσεις, όπου το πρόβλημα είναι μικρής διάστασης ή η συνάρτηση κόστους είναι είτε μη διαφορίσιμη, είτε εξαιρετικά μη κυρτή, έχει δηλαδή πάρα πολλές κορυφές και κοιλάδες, και στις οποίες «εγκλωβίζεται» μία τεχνική καθόδου δυναμικού.

Διαισθητικά, έχουμε την πολυδιάστατη επιφάνεια της συνάρτησης κόστους και ξεκινώντας από ένα σημείο πάνω στην επιφάνεια ακολουθούμε την κατεύθυνση στην οποία η κλίση είναι μεγάλη. Με αυτόν τον τρόπο θα καταλήξουμε σε κάποιο ελάχιστο της συνάρτησης, το οποίο όμως σε περιπτώσεις μη κυρτές θα είναι μόνο ένα τοπικό ελάχιστο.



Σχήμα 2.4: Εφαρμογή της μεθόδου Κατάβασης Κλίσης για εύρεση ελαχίστου για ένα πρόβλημα 2 διαστάσεων. Η αρχικοποίηση έχει μεγάλη σημασία στο αποτέλεσμα της βελτιστοποίησης, όπως εδώ, η μέθοδος καταλήγει σε ένα τοπικό ελάχιστο.

Η κλίση μιας συνάρτησης μπορεί να υπολογιστεί αναλυτικά, όταν η συνάρτηση είναι διαφορίσιμη και απλή. Τα ελάχιστα μπορούν να βρεθούν μηδενίζοντας την κλίση[6], τα οποία επίσης να υπολογιστούν αναλυτικά όταν οι εξισώσεις είναι αναλυτικά επιλύσιμες. Τις περισσότερες φορές όμως, ειδικά στη Μηχανική Μάθηση και τα Νευρωνικά Δίκτυα καμία από τις τρεις προϋποθέσεις δεν ισχύουν επιβάλλοντας την εισαγωγή των ασθενών παραγώγων και υποπαραγώγων[7], των επαναληπτικών σχημάτων και της οπισθοδιάδοσης (backpropagation). Τέλος, λόγω της διακριτής φύσης των υπολογισμών στους ηλεκτρονικούς υπολογιστές, εισάγονται προσεγγίσεις στον υπολογισμό της κλίσης, που γίνεται συνήθως με πεπερασμένες διαφορές.

Υποπαράγωγοι (Subderivatives)

Οι υποπαράγωγοι γενικεύουν την έννοια της παραγωγισιμότητας σε μη διαφορίσιμες συναρτήσεις, με το συνήθη ορισμό, όπως οι συναρτήσεις μεγίστου $\max(\cdot)$ και απόλυτης τιμής $|\cdot|$. Από εδώ και στο εξής ο όρος παράγωγος θα χρησιμοποιείται για αναφορά σε όλους τους τύπους παραγώγων.

[6] Αυτή η συνθήκη είναι αναγκαία και όχι ικανή, δηλαδή βρίσκει γενικά κρίσιμα σημεία (ελάχιστα, μέγιστα, σαγματικά).

[7] Οι δύο έννοιες δεν ταυτίζονται, αλλά οι ασθενείς παράγωγοι είναι πιο γενικοί. Οι υποπαράγωγοι εμφανίζονται κυρίως σε περιβάλλοντα κυρτής βελτιστοποίησης.

Διακριτοποίηση – Επαναληπτικά Σχήματα (Numerical Gradient – Iterative Schemes)

Για να επιλυθεί το πρόβλημα βελτιστοποίησης με υπολογιστή πρέπει να διακριτοποιηθεί. Συνήθως η προσέγγιση των μερικών παραγώγων γίνεται με κάποιων σχήμα πεπερασμένων διαφορών (π.χ. κεντρικές διαφορές), διαδικασία που εισάγει το μήκος βήματος και την ανάγκη εφαρμογής μίας επαναληπτικής μεθόδου.

Οπισθοδιάδοση (Backpropagation)

Σε προβλήματα όπως η γραμμική ταξινόμηση, που αναφέρθηκε στην ενότητα 2.4.2, η συνάρτηση κόστους είναι πολύ απλή και ο απευθείας υπολογισμός της κλίσης, για την εξαγωγή αναλυτικού τύπου, είναι εφικτός. Στην περίπτωση των ΒΝΔ, η συνάρτηση κόστους από την είσοδο μέχρι την έξοδο γίνεται πολύ πολύπλοκη και χρειάζεται διαφορετική τεχνική υπολογισμού, με χρήση του κανόνα της αλυσίδας.

Η οπισθοδιάδοση είναι ένας γρήγορος τρόπος υπολογισμού των παραγώγων σύνθετων συναρτήσεων. Χωρίζοντας τη συνάρτηση σε συνθέσεις δομικών συναρτήσεων και υπολογίζοντας τις τοπικές κλίσεις αυτών, μπορούμε στη συνέχεια να συνδυάσουμε αυτές τις παραγώγους μέσω του κανόνα της αλυσίδας και να καταλήξουμε στην αρχικά ζητούμενη παράγωγο.

Στο επόμενο παράδειγμα εφαρμόζεται η οπισθοδιάδοση στο απλό μοντέλο ενός νευρώνα[8], εξηγείται η ανάγκη εφαρμογής της και τα πλεονεκτήματα που έχει ως προς την ταχύτητα υπολογισμού παραγώγων σε σχέση με τον αναλυτικό υπολογισμό της παραγώγου από την αρχή ως το τέλος του δικτύου. Κάνουμε τις εξής επιλογές:

- Ο νευρώνας έχει 2 εισόδους, δηλαδή

$$\mathbf{x} = [1,\ x_1,\ x_2]^T, \quad \mathbf{w} = [b,\ w_1,\ w_2]^T \tag{2.22}$$

- Η συνάρτηση ενεργοποίησης να είναι η ανορθωμένη γραμμική (ReLU), δηλαδή η έξοδος του νευρώνα είναι

$$\alpha\left(\mathbf{w}^T\mathbf{x}\right) = \max\left(0,\ w_1x_1 + w_2x_2 + b\right) \tag{2.23}$$

Στο παραπάνω νευρώνα λαμβάνουν μέρος 3 βασικές πράξεις, η άθροιση, ο πολλαπλασιασμός και το μέγιστο. Θεωρώντας αυτές τις πράξεις ως δομικά στοιχεία, ορίζονται (μόνο για αυτό το παράδειγμα) ως εξής: $add\left(y,z\right) = y + z$, $mul\left(y,z\right) = y \cdot z$ και $max\left(y,z\right) = y \cdot \mathbb{1}_{y \geqslant z} + z \cdot \mathbb{1}_{z \geqslant y}$ αντίστοιχα. Οι μερικές (υπο)παράγωγοι ως προς τα ορίσματά τους είναι:

$$\frac{\partial add}{\partial y} = 1, \quad \frac{\partial add}{\partial z} = 1 \tag{2.24}$$

[8] Από τη σκοπιά των ΣΝΔ αυτός ο νευρώνας μπορεί να είναι μέρος ενός Συνελικτικού επιπέδου με συνάρτηση ενεργοποίησης την ReLU.

$$\frac{\partial mul}{\partial y} = z, \quad \frac{\partial mul}{\partial z} = y \tag{2.25}$$

$$\frac{\partial max}{\partial y} = \mathbb{1}_{y \geqslant z}, \quad \frac{\partial max}{\partial z} = \mathbb{1}_{z \geqslant y} \tag{2.26}$$

Οι Σχέσεις 2.24 - 2.26 υπολογίζουν την κλίση της εξόδου ως προς κάθε είσοδο για κάθε δομική μονάδα. Για παράδειγμα σύμφωνα με την τελευταία Σχέση 2.26, οι μερικές υποπαράγωγοι της εξόδου της μονάδας $max\,(\cdot,\cdot)$ ως προς τις εισόδους είναι 1 για τη μεγαλύτερη είσοδο και 0 για τη μικρότερη. Η μάθηση γίνεται ως προς τα βάρη **w**, επομένως αυτά θεωρούνται ως μεταβλητές. Οι μερικές παράγωγοι των πρώτων μονάδων ως προς τα βάρη είναι:

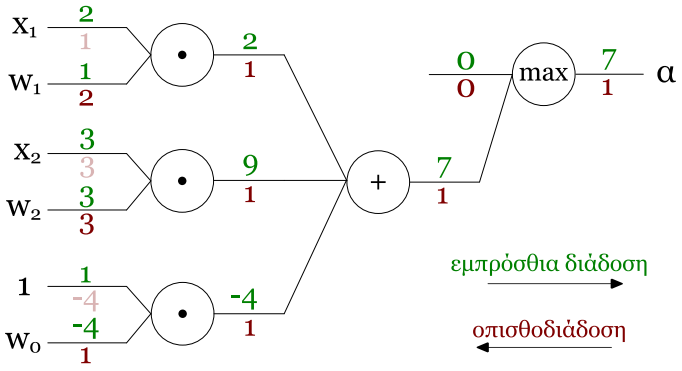
$$\frac{\partial mul_j}{\partial w_i} = \begin{cases} x_i, & j = i \\ 0, & j \neq i \end{cases} \quad i, j = 0, 1, 2 \tag{2.27}$$

Σύμφωνα με τον κανόνα της αλυσίδας, και με τη βοήθεια των Σχέσεων 2.24 - 2.27, οι μερικές παράγωγοι της εξόδου του νευρώνα ως προς τα βάρη του είναι:

$$\frac{\partial \alpha}{\partial w_i} = \frac{\partial max}{\partial w_i} = \frac{\partial max}{\partial add} \cdot \frac{\partial add}{\partial mul_i} \cdot \frac{\partial mul_i}{\partial w_i} = \mathbb{1}_{add \geqslant 0} \cdot 1 \cdot x_i, \quad i = 0, 1, 2 \tag{2.28}$$

Τελικά, οι μερικές παράγωγοι της εξόδου ως προς κάθε μεταβλητή είναι:

$$\frac{\partial \alpha}{\partial w_i} = x_i \cdot \mathbb{1}_{\mathbf{w}^T \mathbf{x} \geqslant 0}, \quad i = 0, 1, 2 \tag{2.29}$$



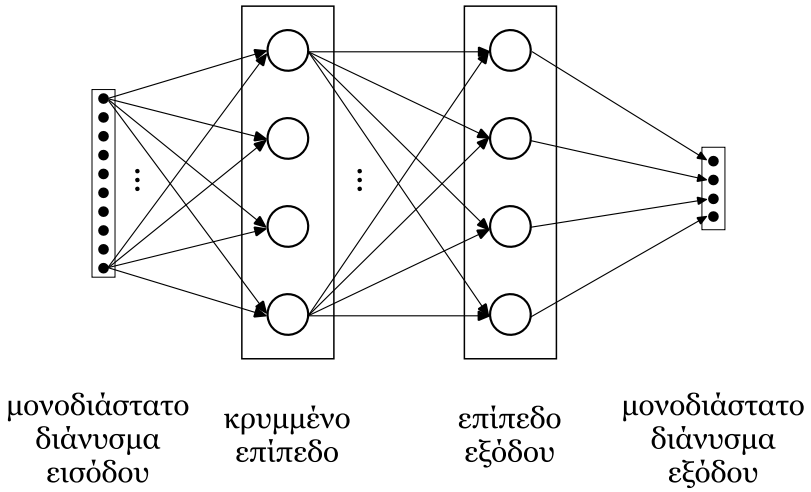
Σχήμα 2.5: Λειτουργία απλού νευρώνα κατά την οπισθοδιάδοση. Ο νευρώνας έχει είσοδο $\mathbf{x} = [1,\ 2,\ 3]$ και τα βάρη αρχικοποιούνται στις τιμές $[-4,\ 1,\ 3]$. Κατά την εμπρόσθια διάδοση ο νευρώνας ενεργοποιείται, αφού το εσωτερικό γινόμενο ξεπερνάει το κατώφλι 0.

Κάθε δομική μονάδα αφού λάβει τις εισόδου της μπορεί να υπολογίσει την έξοδό της και την τοπική κλίση της εξόδου της ως προς τις εισόδους της ανεξάρτητα από τις άλλες. Η λειτουργία των τριών δομικών μονάδων κατά την οπισθοδιάδοση είναι η εξής:

- Μονάδα add: Σύμφωνα με τις 2.24 κατά την οπισθοδιάδοση μεταφέρουν την κλίση που εμφανίζεται στην έξοδό τους, σε κάθε μία από τις εισόδους αυτούσια (πολλαπλασιασμένη με 1).
- Μονάδα max: Σύμφωνα με τις 2.26 μεταφέρουν την κλίση που δέχονται στην είσοδο που είχε τη μέγιστη τιμή κατά την εμπρόσθια διάδοση, ενώ στην άλλη είσοδο το 0.
- Μονάδα mul: Σύμφωνα με τις 2.25 μεταφέρουν σε κάθε είσοδο την κλίση που δέχονται πολλαπλασιασμένη με την τιμή της άλλης εισόδου κατά την εμπρόσθια διάδοση, π.χ. στο Σχήμα 2.5 για την μονάδα mul_0: $1 \cdot w_0 = -4$ και $1 \cdot x_0 = 1$.

2.5 Κλασικά Νευρωνικά Δίκτυα και δύναμη αναπαράστασης

Συνδυάζοντας πολλούς Νευρώνες, όπως αυτούς του Σχήματος 2.1, σε επίπεδα κατασκευάζεται ένα Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ ή ΝΔ) (Artifial Neural Network), που ονομάζεται κατά παράδοση και Multi-Layer Perceptron (MLP). Οι νευρώνες οργανώνονται σε δομή ακυκλικού γράφου, όλοι οι νευρώνες ενός επιπέδου συνδέονται με κάθε νευρώνα του προηγούμενου και επόμενου επιπέδου, ενώ εντός κάθε επιπέδου δεν υπάρχουν συνδέσεις (Σχήμα 2.6). Κάθε επίπεδο περιγράφεται από το επαυξημένο πίνακα των βαρών W διάστασης $K \times D$, που σε κάθε γραμμή έχει τα βάρη καθενός από τους K νευρώνες και ένα διάνυσμα κατωφλιών διάστασης D. Οι νευρώνες του σταδίου εξόδου συνήθως δεν έχουν συνάρτηση ενεργοποίησης.



Σχήμα 2.6: Ένα πλήρως συνδεδεμένο δίκτυο εμπρόσθιας τροφοδότησης MLP.

Λειτουργία – Ανάκληση/Εμπρόσθια Διάδοση (Feed-Forward computation)

Η είσοδος σε κάθε επίπεδο είναι ένα διάνυσμα D στοιχείων και κάθε επίπεδο υπολογίζει την έξοδό του εκτελώντας τον πολλαπλασιασμό $W\mathbf{x}$ εφαρμόζοντας σε κάθε στοιχείο του αποτελέσματος τη μη-γραμμικότητα.

ΝΔ ως Καθολικοί Προσεγγιστές

Από μία σκοπιά τα ΝΔ μπορεί να θεωρηθούν ως προσεγγιστές συναρτήσεων, συνδυάζοντας συναρτήσεις βάσης με κατάλληλα βάρη (εκπαιδεύσιμα). Στο πλαίσιο της μηχανικής μάθησης μπορούν να θεωρηθούν ότι κατασκευάζουν μία υπερεπιφάνεια διαχωρισμού των κατηγοριών. Σύμφωνα με το θεώρημα Καθολικής Προσέγγισης (Cybenko 1989 - για σιγμοειδείς συναρτήσεις ενεργοποίησης - [Cybe89]) ένα ΝΔ εμπρόσθιας διάδοσης (feed-forward NN) με πεπερασμένο πλήθος νευρώνων (π.χ. MLP) μπορεί να προσεγγίσει μία συνεχή συνάρτηση σε ένα συμπαγές υποσύνολο του $\mathbb{R}^n$ με όση ακρίβεια επιθυμείται.

Το θεώρημα αυτό δηλώνει ότι οι συναρτήσεις μπορούν να προσεγγιστούν μαθηματικά από αθροίσματα και συνθέσεις απλών/δομικών συναρτήσεων. Αυτό αποτελεί μία κατασκευαστική μέθοδο που αντιστοιχεί στην αρχιτεκτονική των ΝΔ, ωστόσο δεν εξετάζει ζητήματα της εκπαιδευσιμότητας, π.χ. το κατά πόσο μπορούν να μαθευτούν όντως τα απαραίτητα βάρη για την προσέγγιση.

Αν και ένα δίκτυο με τουλάχιστον 1 κρυμμένο επίπεδο μπορεί να αναπαραστήσει οποιαδήποτε συνάρτηση διαχωρισμού, έχει παρατηρηθεί εμπειρικά ότι δίκτυα με περισσότερα κρυμμένα επίπεδα λειτουργούν καλύτερα, γεγονός που δεν έχει αποδειχθεί πλήρως μαθηματικά. Φυσικό επακόλουθο είναι να αναρωτηθούμε αν αυξάνοντας συνεχώς το βάθος του δικτύου θα βελτιώνονται και τα αποτελέσματα. Η απάντηση δίνεται και πάλι εμπειρικά και φαίνεται ότι ένα ΝΔ με 2 ή 3 κρυμμένα επίπεδα θα είναι καλύτερο από ένα δίκτυο με μόνο 1 κρυμμένο επίπεδο, αλλά εισάγοντας επιπλέον επίπεδα σπάνια βελτιώνει τα αποτελέσματα. Αυτές οι παρατηρήσεις καθυστέρησαν την ανάπτυξη των ΝΔ για μία δεκαετία, και για να διαχωρίσουμε αυτά τα ΝΔ από τα σύγχρονα τα ονομάζουμε κλασικά ΝΔ.

Αντίθετα, προ-αναφέρουμε ότι στα Βαθέα Νευρωνικά Δίκτυα το βάθος φαίνεται να παίζει σημαντικό ρόλο. Ένα σύγχρονο δίκτυο ταξινόμησης εικόνων μπορεί να έχει μέχρι και 30 επίπεδα. Οι εικόνες έχουν βαθιά ιεραρχική δομή (π.χ. πρόσωπο, αποτελούμενο από μέρη (μάτια, στόμα, ...) που αποτελούνται από ακμές, κτλ.), επομένως η ύπαρξη και οι λειτουργίες πολλών επιπέδων βρίσκουν αντιστοιχίες με τη δομή τους. Αυτό φαίνεται και πειραματικά κατά την οπτικοποίηση των ενεργοποιήσεων των επιπέδων (Ενότητα 5.4).

Παρόμοια και γενικευμένα θεωρήματα για την προσέγγιση συναρτήσεων από δίκτυα εμπρόσθιας τροφοδότησης παρουσιάστηκαν από τους, Hornik κ.ά. (1989), Sun, Cheney και Light (1992).

2.6 Ιστορική Ανασκόπηση: Βαθέα και Συνελικτικά Νευρωνικά Δίκτυα

Το πεδίο των Νευρωνικών Δικτύων εμπνέεται από τη δομή, τις λειτουργίες και τις δυνατότητες του ανθρώπινου εγκεφάλου και προσπαθεί να τις μεταφέρει στους υπολογιστές, για την επίλυση προβλημάτων όπως η ταξινόμηση, ο εντοπισμός αντικειμένων και η κατάτμηση σκηνών τα οποία ο άνθρωπος επιλύει εύκολα και γρήγορα. Η ανάπτυξη αυτών των δικτύων συνδέεται στενά με την πρόοδο των επιστημών της βιολογίας και της νευρολογίας, απ' όπου αντλούνται δομικά και λειτουργικά χαρακτηριστικά του ανθρώπινου νευρικού συστήματος. Από νωρίς είχε παρατηρηθεί ότι ο ανθρώπινος εγκέφαλος εξάγει συμπαγείς και συνοπτικές περιγραφές για τις πληροφορίες που δέχεται, καθώς ο χώρος αποθήκευσής τους είναι περιορισμένος και οι διαστάσεις της πληροφορίας πολλές. Αυτό ονομάζεται «κατάρα των πολλών διαστάσεων» (curse of dimensionality) (η δυσκολία εκμάθησης αυ-

ξάνεται εκθετικά με την γραμμική αύξηση των διαστάσεων) και οδηγεί στη θεμελιώδη ανάγκη για εξαγωγή «περιληπτικών» περιγραφών, που ονομάζεται εξαγωγή χαρακτηριστικών.

Τα Νευρωνικά Δίκτυα είναι ένα σχετικά νέο πεδίο στον ευρύτερο τομέα της Τεχνητής Νοημοσύνης και έχει τις απαρχές στο 1960 ή και νωρίτερα, όπου καθορίστηκαν οι ιδιότητες και η δομή των νευρώνων μέσα σε ολοκληρωμένα δίκτυα, αλλά και μία διαδικασία μάθησης με τη βοήθεια της διαδικασίας της οπισθοδιάδοσης (backpropagation). Για αρκετά χρόνια η Μηχανική Μάθηση βασιζόταν σε «ρηχές» (shallow) τεχνικές για μοντελοποίηση δεδομένων και εξαγωγή χαρακτηριστικών, όπως τα Γκαουσιανά Μίγματα Κατανομών (GMM), διάφορα γραμμικά και μη συστήματα, Μαρκοβιανά Μοντέλα (MM, HMM), Conditional Random Fields (CRF), Μηχανές Διανυσμάτων Υποστήριξης (SVM), αλλά και τα κλασικά δίκτυα εμπρόσθιας διάδοσης (MLP), που αποτελούνταν από 2 έως 6 το πολύ επίπεδα.

Οι ενδείξεις από τις νευροβιολογικές επιστήμες, ωστόσο, ήταν ότι το ανθρώπινο νευρικό σύστημα αποτελείται από πολλαπλά επίπεδα επεξεργασίας της σύνθετης πληροφορίας εισόδου από τους αισθητήρες του. Η αλληλουχία αυτών των επιπέδων είναι που ξετυλίγει τη σύνθετη δομή των δεδομένων και βοηθά στην κατασκευή μίας πλούσιας και συμπαγούς αναπαράστασης. Οι ιδέες αυτές οδήγησαν το 1980 να προταθεί, από τον Kunihiko Fukushima, ένα ιεραρχικό, τεχνητό ΝΔ, το Neocognitron που θεωρείται ο πρόγονος των σημερινών Βαθέων (Εμβριθών) Νευρωνικών Δικτύων (ΒΝΔ) (Deep Neural Networks). Αν και η δομή αυτού του δικτύου μοιάζει πολύ με τα σημερινά, ο τρόπος μάθησης ήταν διαφορετικός, γεγονός που οδήγησε, σε συνδυασμό με άλλους λόγους, σε όχι τόσο επιτυχή αποτελέσματα.

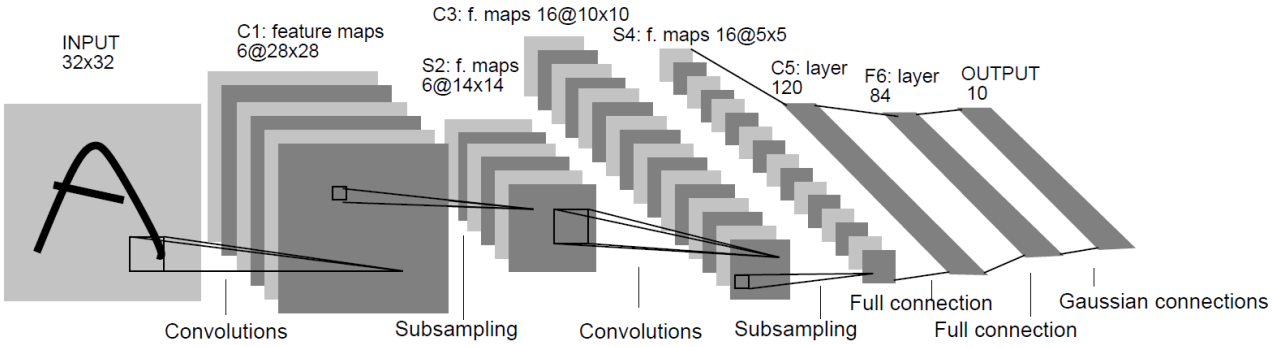
Από τότε έχουν προταθεί πολλά μοντέλα ΒΝΔ, με επιβλεπόμενη ή μη μάθηση, διακριτικά (discriminative) ή αναγεννητικά (generative), αναδρομικά ή ακυκλικά, όπως τα Deep Belief Networks, Stacked Auto-Encoders, Hierarchical Temporal Memory, Long Short-Term Memory, Restricted Boltzmann Machines, Recurrent Neural Networks. Τη μεγαλύτερη επιτυχία και ανάπτυξη έχουν όμως τα Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ) (Convolutional Neural Networks – CNN), τα οποία μετά το Neocognitron επανήλθαν στο προσκήνιο, κυρίως με το μεγάλο βήμα που επετεύχθη στην αναγνώριση/ταξινόμηση γραπτών ψηφίων (0-9) (MNIST Database) [9] από το ΣΝΔ LeNet [LeCu98], με ποσοστό λάθους μικρότερο του 3% [10], που προτάθηκε το 1998 από τους Lecun, Bottou, Bengio και Haffner. Η δομή του δικτύου φαίνεται στο Σχήμα 2.7.

Τα επόμενα χρόνια υπήρξε μία μείωση της δραστηριότητας στο πεδίο των ΣΝΔ, μέχρι το 2012, που παρουσιάστηκε το ΣΝΔ AlexNet [Kriz12] από τους Krizhevsky, Sutskever και Hinton. Στόχος του ήταν να ταξινομεί καθημερινές εικόνες σε 1000 κατηγορίες, κάτι που κατάφερε με σφάλμα μόλις 16%, δεδομένου του μεγάλου πλήθους κατηγοριών. Από τότε η ανάπτυξη των ΣΝΔ είναι ραγδαία και ο τομέας των ΣΝΔ είναι μόλις στην αρχή του. Ποιοι ήταν όμως οι κύριοι λόγοι που οδήγησαν στην ανάπτυξη αυτών των δικτύων σε αυτή την περίοδο, ενώ οι βασικές ιδέες είχαν εμφανιστεί πολύ νωρίτερα;

Τα τελευταία χρόνια δύο είναι οι κύριοι παράγοντες που έχουν οδηγήσει σε ραγδαία ανάπτυξη των ΒΝΔ. Αρχικά, η ανάπτυξη των δικτύων (internet, κοινωνικά δίκτυα) και η εκλαΐκευση της τεχνολογίας οδήγησαν σε πολύ μεγάλη συσσώρευση οπτικοακουστικών δεδομένων παγκοσμίως και εύκολης

[9] Η βάση γραπτών ψηφίων MNIST περιέχει 60,000 εικόνες εκπαίδευσης και 10,000 εικόνες δοκιμής, διαστάσεων 28×28.

[10] Το σφάλμα σήμερα είναι μικρότερο του 0.23% (δηλαδή 23 στα 10,000 ψηφία ταξινομούνται λανθασμένα) και επιτυγχάνεται από επιτροπή 35 ΣΝΔ, που εκπαιδεύτηκαν λαμβάνοντας υπόψη ελαστικές παραμορφώσεις των ψηφίων.



Σχήμα 2.7: Η αρχιτεκτονική του ΣΝΔ LeNet για την αναγνώριση ψηφίων.

αρχειοθέτησης. Τα δεδομένα αυτά είτε μέσω αυτόματων τρόπων, είτε χειροκίνητα (υπηρεσίες πληθοπορισμού (crowdsourcing)) ήταν εύκολο να επισημειωθούν με πληθώρα χαρακτηριστικών. Αυτό δημιούργησε πολύ μεγάλες βάσης δεδομένων, που είναι αναγκαίες για την εκπαίδευση ΒΝΔ.

Ένα δεύτερο απαραίτητο στοιχείο είναι η υπολογιστική ισχύς και οι χώροι ψηφιακής αποθήκευσης. Τα τελευταία χρόνια, μέσω της ανάπτυξης πολυπύρηνων επεξεργαστών, ισχυρών καρτών γραφικών και σύννεφων διαμοιραζόμενων υπολογιστικών πόρων (servers υπολογιστών απομακρυσμένης πρόσβασης (clusters)), έχει γίνει εφικτή η εκπαίδευση μεγάλων δικτύων σε εύλογο χρονικό διάστημα.

Σε αυτή τη Διπλωματική Εργασία θα επικεντρωθούμε στην ανάλυση και υλοποίηση των ΣΝΔ που επεξεργάζονται εικόνες, τα οποία έχουν αναπτυχθεί περισσότερο και έχουν επιτυχίες σε πολλούς τομείς τα τελευταία χρόνια. Τρεις σημαντικές διαφορές τους από τα κλασικά ΝΔ είναι:

- Μέγεθος Δικτύων. Τα σύγχρονα ΣΝΔ είναι μεγαλύτερα τόσο σε βάθος, από 2-3 επίπεδα σε τουλάχιστον 8, όσο και σε πλάτος, από τάξης 10^3 νευρώνες σε κάθε επίπεδο σε $10^5 - 10^6$.
- Περισσότερα δεδομένα εκπαίδευσης. Τα σύγχρονα σύνολα δεδομένων περιέχουν πλήθος εικόνων της τάξης $10^6 - 10^8$.
- Η μάθηση είναι γρηγορότερη, χάρη στην άφθονη υπολογιστική ισχύ, καλύτερη χάρη στους νέους αλγορίθμους βελτιστοποίησης και αποδοτικότερη, λόγω πρόσφατα ανεπτυγμένων τεχνικών κανονικοποίησης (regularization). Επίσης, ένα πολύ βασικό καινούριο χαρακτηριστικό των ΣΝΔ είναι ότι εκμεταλλεύονται στο έπακρο τη χωρική δομή μιας εικόνας, κάτι που δεν γινόταν στα MLP.

Διαγωνισμός ImageNet Large Scale Visual Recognition (ILSVR)

Ο ετήσιος διαγωνισμός ILSVR [Russ14] αποτελεί αναφορά (benchmark) στην κατηγοριοποίηση και εντοπισμό αντικειμένων σε βάσεις εκατομμυρίων εικόνων και διενεργείται από το 2010. Κάθε χρόνο ο διαγωνισμός περιλαμβάνει διάφορες κατηγορίες, όπως εντοπισμό και ταυτοποίηση αντικειμένων, κατηγοριοποίηση εικόνων, κ.ά. και προσελκύει ομάδες από όλο το κόσμο. Τρία ενδιαφέροντα σημεία είναι τα εξής:

- Η συλλογή εικόνων και η πλήρης επισημείωσή τους για τις κατηγορίες, την ύπαρξη και την θέση των αντικειμένων αποτελεί επίτευγμα από μόνο του, δεδομένου των μεγεθών της βάσης: 1000

κατηγορίες αντικειμένων, περίπου 1.3 εκατομμύρια εικόνες εκπαίδευσης, 50 χιλιάδες εικόνες επαλήθευσης και 100 χιλιάδες εικόνες δοκιμής.

- Μέχρι και το 2011 στο διαγωνισμό δεν υπήρχαν συμμετοχές με Νευρωνικά Δίκτυα και οι νικητήριες ομάδες χρησιμοποιούσαν κλασσικές τεχνικές (SHIFT, SVM, Fisher Vectors). Η εμφάνιση του ΣΝΔ AlexNet το 2012, άλλαξε το τοπίο και από τότε οι περισσότερες συμμετοχές περιλαμβάνουν ΣΝΔ (2010-2012: 21 ομάδες συνολικά - 1 ΣΝΔ, 2013: 23 ομάδες - σχεδόν όλες ΣΝΔ, 2014: 36 ομάδες - σχεδόν όλες ΣΝΔ) Μία περιγραφή των δικτύων των πρώτων θέσεων για την ταξινόμηση εικόνων από το 2012 και μετά δίνεται στην Ενότητα 3.4.

- Τα νικητήρια ποσοστά top-5 σφάλματος στην ταξινόμηση εικόνων έχουν μειωθεί δραματικά από το 2012 και το 2015 ξεπέρασαν και το ανθρώπινο σφάλμα[11]: 2010-2015: 28%, 26%, 16% ([Kriz12]), 12%, 7%, 5% ([Ioff15]). Από το 2015 ο διαγωνισμός για την κατηγοριοποίηση εικόνων σταματά, ενώ συνεχίζονται οι διαγωνισμοί για τον εντοπισμό και την ανίχνευση αντικειμένων σε εικόνες, ανίχνευση αντικειμένων σε βίντεο και κατηγοριοποίηση σκηνών.

[11] Σχόλιο: Το ανθρώπινο σφάλμα είναι κάπως «τεχνητά» αυξημένο· καθώς αν και ο άνθρωπος μπορεί να αντιλαμβάνεται πλήρως τι υπάρχει σε μία εικόνα, είναι δύσκολο να αποδώσει μία μοναδική κατηγορία στην εικόνα, ίδια με την επισημειωμένη.

Κεφάλαιο 3

Συνελικτικά Νευρωνικά Δίκτυα

Παρά την πρόσφατη εμφάνιση και ανάπτυξή τους, τα ΣΝΔ έχουν στρέψει το ενδιαφέρον πολλών ερευνητών στην προσαρμογή πολλών υπαρχόντων προβλημάτων σε αυτά, ώστε να μπορούν να επιλυθούν από μία διαδικασία μάθησης. Οι επιδόσεις των ΣΝΔ, μέσα σε λίγα χρόνια, σε προβλήματα «χαμηλού» επιπέδου, όπως η αποθορυβοποίηση (denoising, compression artifact reduction), η όξυνση (deblurring, sharpening), η υπερ-ανάλυση (super-resolution), η ενδοσυμπλήρωση (inpainting) και η εύρεση ακμών (edge detection) έχουν προσεγγίσει και αρκετές φορές ξεπεράσει τεχνικές, που αναπτύσσονται από μηχανικούς για πολλά χρόνια.

Σε ότι αφορά προβλήματα «υψηλότερου» επιπέδου, που θεωρούνται γενικά πιο πολύπλοκα, όπως η ταξινόμηση (classification), ο εντοπισμός/αναγνώριση αντικειμένων, σκηνών, κινήσεων, πόζας (detection/recognition), η (σημασιολογική) κατάτμηση (semantic segmentation), αλλά και η εκτίμηση βάθους (depth estimation), τα ΣΝΔ έχουν σημειώσει τεράστια πρόοδο. Ενδεικτικά αναφέρουμε μερικά άλλα πολύπλοκα προβλήματα στα οποία τα ΣΝΔ έχουν βρει εφαρμογή: εκτίμηση οπτικής ροής (tloptical flow estimation), ταυτόχρονος εντοπισμός και χαρτογράφηση (Simultaneous Localization and Mapping), πρόβλεψη σημείων κλειδιών σε βίντεο (keypoint prediction), ανακατασκευή και ανάλυση τριδιάστατων σχημάτων και σταθερότητα χρώματος (color constancy). Τέλος, αναφέρουμε ότι τα ΣΝΔ έχουν έμφυτη την δυνατότητα εξαγωγής «καλών» χαρακτηριστικών και γι’ αυτό χρησιμοποιούνται όλο και περισσότερο για την ανάκτηση, την αναπαράσταση και το ταίριασμα (image registration, matching).

Τα δίκτυα αυτά έχουν μεγάλη επιτυχία σε επεξεργασία δεδομένων με έντονη ιεραρχία, όπως οι εικόνες, σε σχέση με άλλες μεθόδους και χαρακτηριστικά, που εξάγονται από σχεδιασμένες διαδικασίες από μηχανικούς, κυρίως λόγω των επόμενων δύο σημείων:

- Ο τρόπος εξαγωγής χαρακτηριστικών (συντελεστές φίλτρων) υπαγορεύεται σε μεγάλο βαθμό από τη διαδικασία βελτιστοποίησης, με αποτέλεσμα τα συνελικτικά χαρακτηριστικά που δημιουργεί ένα ΣΝΔ να είναι τα «καλύτερα» για το αντίστοιχο πρόβλημα (πάντα ως προς κάποιο επιλεγμένο κριτήριο).

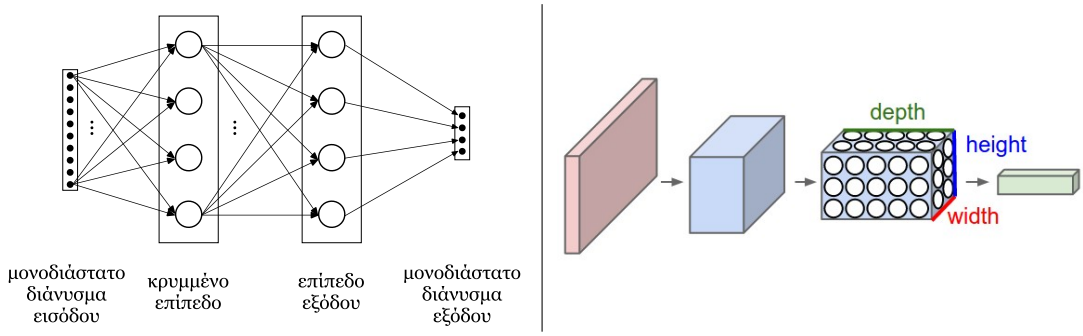
- Οι λειτουργίες του δικτύου είναι κατάλληλες και αντιστοιχούν άμεσα με τη δομή διδιάστατων εικόνων. Ένα κλασικό ΝΔ που δέχεται ένα διάνυσμα εισόδου σταθερού μήκους δεν μπορεί για παράδειγμα να χρησιμοποιηθεί άμεσα για την ταξινόμηση εικόνων οποιουδήποτε μεγέθους.

3.1 Δομή ΣΝΔ και σύγκριση με τα κλασικά ΝΔ

Η βασική δομή ενός ΣΝΔ είναι αυτή ενός Κατευθυνόμενου Ακυκλικού Γράφου (Directed Acyclic Graph) από επίπεδα, που επιτελούν συγκεκριμένες λειτουργίες. Τα ΣΝΔ αποτελούνται γενικά από δύο κύρια μέρη: το μέρος που είναι υπεύθυνο για την εξαγωγή των συνελικτικών χαρακτηριστικών (convolutional features) και το κυρίως ΝΔ, το οποίο εκπαιδεύεται χρησιμοποιώντας αυτά τα χαρακτηριστικά. Το δεύτερο μέρος μπορεί να είναι ένας ταξινομητής (γραμμικός ή μη), ένα Μπεϋσιανό Δίκτυο (π.χ. Belief Network), κτλ.

Μία διαφορά από τα κλασικά ΝΔ είναι ότι υποθέτουν ότι οι είσοδοί τους είναι εικόνες ή έχουν μορφή και χαρακτηριστικά εικόνων (π.χ. τοπικότητα), κωδικοποιώντας στην αρχιτεκτονική τους αυτό το δεδομένο. Στα κλασικά ΝΔ η τοπικότητα και η γειτνίαση δεν λαμβάνονται υπόψη, αντίθετα στα ΣΝΔ η οργάνωση στο χώρο, η διάταξη και η σειρά (π.χ. για χρονικά δεδομένα) έχουν καθοριστική σημασία.

Μία δεύτερη σημαντική διαφορά είναι ότι τα ΣΝΔ βασίζονται στη πράξη της συνέλιξης, ενώ τα κλασικά ΝΔ στην πράξη του εσωτερικού γινομένου. Αν και στην ουσία οι δύο πράξεις είναι παρόμοιες, η συνέλιξη δεν περιορίζει το μέγεθος της εισόδου, αφού μπορεί να εφαρμοστεί επαναληπτικά σε κάθε περιοχή της εισόδου. Επιπλέον, τα ΣΝΔ έχουν μια πληθώρα επιπέδων, που εκτός από πράξεις, κάνουν μείωση της διάστασης των χαρακτηριστικών, κανονικοποίηση και άλλες λειτουργίες, κάτι που τα εφοδιάζει με επιπλέον δυνατότητες. Οι δομικές διαφορές φαίνονται στο Σχήμα 3.1.



Σχήμα 3.1: Αριστερά: Ένα συμβατικό ΝΔ 2 επιπέδων, η είσοδος είναι ένα διάνυσμα με D_{in} στοιχεία (ή $1 \times 1 \times D_{in}$) και η έξοδος όμοια ένα διάνυσμα με D_{out} στοιχεία. Δεξιά: Ένα ΣΝΔ αποτελείται από επίπεδα σε αλληλουχία τα οποία επικοινωνούν με όγκους δεδομένων. Τα επίπεδα με νευρώνες οργανώνουν σε 3 διαστάσεις. Στο παράδειγμα μία RGB εικόνα εισόδου έχει διαστάσεις $W \times H \times 3$ και η έξοδος του ΣΝΔ είναι ένα D-διάστατο διάνυσμα (ή $1 \times 1 \times D_{out}$ στην ορολογία των ΣΝΔ).

Για το βασικό πρόβλημα της ταξινόμησης η αρχιτεκτονική ενός ΣΝΔ αποτελείται από κάποια επίπεδα που εξάγουν χαρακτηριστικά και ένα συμβατικό ΝΔ - ταξινομητή που δέχεται αυτά τα χαρακτηριστικά και εκτελεί την ταξινόμηση. Η αρχιτεκτονική αυτών των δικτύων εκμεταλλεύεται την «εικονική» μορφή της εισόδου, κάτι που οδηγεί σε συγκεκριμενοποίηση και περιορισμό των λειτουργιών των επιπέδων. Τα δεδομένα εισόδου έχουν συνήθως 3 διαστάσεις· για εικόνες (πλάτος, ύψος, βάθος) ($W \times H \times D$), όπου D ο αριθμός των καναλιών: $D = 1$ (ασπρόμαυρες), $D = 3$ (έγχρωμες) ή πολυκαναλικές, αλλά μπορεί να είναι οποιαδήποτε δεδομένα μπορούν να αναπαρασταθούν σε

παρόμοια μορφή.

3.2 Βασικά Επίπεδα των ΣΝΔ (Layers)

Τα ΣΝΔ είναι πλήρως αρθρωτά δίκτυα που αποτελούνται από θεμελιώδη επίπεδα/στρώματα. Τα κυριότερα επίπεδα μίας κοινής αρχιτεκτονικής είναι τα εξής:

- Επίπεδο Συνέλιξης (Convolutional): Το επίπεδο αυτό υλοποιεί την γνωστή πράξη της συνέλιξης σε όγκους εισόδου και εξάγει όγκους εξόδου, στους οποίους κάθε φέτα ονομάζεται και χάρτης χαρακτηριστικών. Περιλαμβάνει νευρώνες οργανωμένους σε παράλληλα στρώματα, που φιλτράρουν όλα τα κανάλια της εισόδου. Κάθε νευρώνας έχει ένα περιορισμένο οπτικό/δεκτικό πεδίο (receptive field), άρα η έξοδός του αποτελεί έναν τοπικό περιγραφητή. Μέσω της εκπαίδευσης το δίκτυο μαθαίνει τον τρόπο να εξάγει μία αποτελεσματική και χρήσιμη τοπική περιγραφή. Η τοπικότητα και η «βελτιστότητα» των εξαγόμενων χαρακτηριστικών είναι οι κύριες αιτίες που οδήγησαν στην μεγάλη ανάπτυξη και επιτυχία αυτών των δικτύων.

- Επίπεδο Συγκέντρωσης (Pooling): Εκτός από τα επίπεδα του φιλτραρίσματος, αναγκαία είναι η περιοδική ύπαρξη επιπέδων, τα οποία θα μειώνουν τις χωρικές διαστάσεις της αναπαράστασης, ελαττώνοντας το πλήθος παραμέτρων προς μάθηση και τους υπολογισμούς σε ένα δίκτυο, ελαχιστοποιώντας ταυτόχρονα τις πιθανότητες για υπερ-εκπαίδευση.(βάλε πληροφορίες από [Bour10]). Αυτά τα επίπεδα παρεμβάλλονται περιοδικά σε ένα δίκτυο με σκοπό να «συνοψίσουν» τα αποτελέσματα γειτονικών νευρώνων στους χάρτες χαρακτηριστικών.

- Επίπεδο Κανονικοποίησης (Normalization): Αρχικά αυτά τα επίπεδα εφαρμόστηκαν για να προσομοιάσουν ανασταλτικές διαδικασίες των βιολογικών νευρώνων. Ουσιαστικά κανονικοποιούν τις σχετικές διαφορές γειτονικών τιμών χαρακτηριστικών, ώστε να έχουν χρησιμοποιήσιμες τιμές από τα επόμενα επίπεδα. Η λειτουργία που επιτελούν θα μπορούσε να παραλληλιστεί με τη βελτίωση του contrast (contrast normalization) σε φωτογραφίες.

- Πλήρως συνδεδεμένο Επίπεδο (Fully-Connected): Αυτά τα επίπεδα είναι ίδια με τα επίπεδα που χρησιμοποιούνται στα κλασικά νευρωνικά δίκτυα. Περιλαμβάνουν νευρώνες, που δέχονται ένα διάνυσμα χαρακτηριστικών εισόδου και εξάγουν μία απόκριση. Συνήθως, τοποθετούνται σε αλληλουχία, δημιουργώντας ένα κλασσικό ΝΔ.

Από αυτά τα επίπεδα μόνο τα Συνελικτικά και τα Πλήρως συνδεδεμένα έχουν νευρώνες με τη συνήθη έννοια, επομένως μόνο αυτά εισάγουν βάρη προς εκπαίδευση στο δίκτυο. Τα υπόλοιπα επίπεδα επιτελούν μία προκαθορισμένη (σταθερή) λειτουργία.

Το κάθε επίπεδο από τα προαναφερόμενα έχει μία συγκεκριμένη λειτουργία, επομένως η σειρά τοποθέτηση και οργάνωσής τους σε ένα ΣΝΔ πρέπει να ακολουθεί συγκεκριμένα μοτίβα (layer patterns). Η βάση κάθε ΣΝΔ είναι μία αλληλουχία από συνελικτικά και πλήρως συνδεδεμένα επίπεδα, όπου τα συνελικτικά τοποθετούνται στην αρχή λόγω της τοπικότητας των ιδιοτήτων τους, ενώ τα πλήρως συνδεδεμένα τοποθετούνται συνήθως στο τέλος, καθώς δεν «αντιλαμβάνονται» κάποια τοπικότητα. Και δύο αυτά επίπεδα μπορεί να ακολουθούνται από επίπεδα συγκέντρωσης, τα οποία συνοψίζουν

και μειώνουν τις χωρικές διαστάσεις των χαρακτηριστικών. Τα επίπεδα κανονικοποίησης μπορούν να προστεθούν προαιρετικά πριν ή μετά από τα επίπεδα συγκέντρωσης.

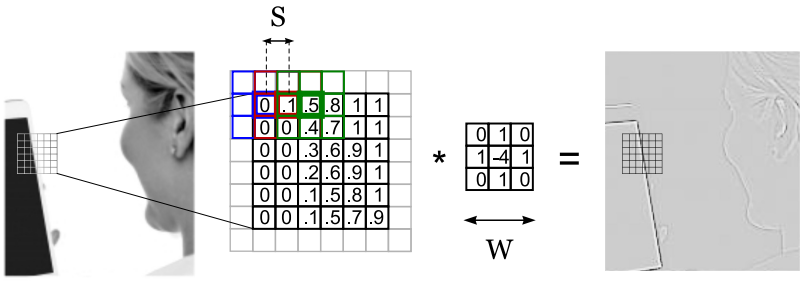
3.2.1 Το βασικό επίπεδο της Συνέλιξης

Εισαγωγή στην πράξη της Συνέλιξης (Convolution)

Η συνέλιξη αποτελεί βασική μαθηματική πράξη και απαντάται σε πάρα πολλούς επιστημονικούς τομείς. Η ανάλυση που ακολουθεί θα βασιστεί στη χρήση της στο περιβάλλον των ΣΝΔ. Με υπόβαθρο τις διακριτές εικόνες που αποθηκεύονται ως πίνακες από τιμές φωτεινότητας (πίξελ) η συνέλιξη ορίζει ένα τοπικό, γραμμικό φιλτράρισμα μιας εικόνας x με ένα φίλτρο w και στο διακριτό περιβάλλον τα στοιχεία της εικόνας εξόδου y υπολογίζονται ως εξής:

$$y\left(i,j\right)=\left(x*w\right)\left(i,j\right)=\sum_{m,n}x\left(m,n\right)\ w\left(i-m,j-n\right)$$

Αν τα πίξελ της περιοχής που φιλτράρεται κάθε φορά αναδιαταχθούν, ώστε να δημιουργούν ένα διάνυσμα, και γίνει το ίδιο και στους συντελεστές του φίλτρου[1], τότε η συνέλιξη ισοδυναμεί με το εσωτερικό γινόμενο των δύο διανυσμάτων με αναδιάταξη του αποτελέσματος στο χώρο της εικόνας. Στο σχήμα 3.2 φαίνεται γραφικά η συνέλιξη μιας περιοχής της εικόνας x με ένα φίλτρο w ενίσχυσης ακμών. Αν και συνέλιξη θα μπορούσε να γίνει με τυχαίο σχήμα φίλτρου, στα ΝΔ επιλέγεται συνήθως τετραγωνικό φίλτρο για ευκολία ανάλυσης και υλοποίησης. Στο Σχήμα 3.3 φαίνεται η «διανυσματοποίηση» περιοχών της εικόνας για αποδοτική υλοποίηση της συνέλιξης ως εσωτερικό γινόμενο.



Σχήμα 3.2: Γραφική απεικόνιση συνέλιξης ασπρόμαυρης (1 κανάλι) εικόνας με φίλτρο ανίχνευσης ακμών. Το φίλτρο είναι συμμετρικό, επομένως δεν χρειάζεται κατοπτρισμός. Γίνεται επέκταση της εικόνας έξω από τα όριά της, ώστε η φιλτραρισμένη να έχει τις ίδιες διαστάσεις με την αρχική. Οι τιμές της επέκτασης δεν φαίνονται, αλλά συνήθως χρησιμοποιούνται οι τιμές των συνόρων.

Παράμετροι και τεχνικά ζητήματα της συνέλιξης είναι:

- το μέγεθος F του φίλτρου w που καθορίζει το οπτικό πεδίο της συνέλιξης, συνήθεις τιμές είναι από 3×3 μέχρι και 100×100 ανάλογα με το μέγεθος της εικόνας, αν και η τάση στα ΣΝΔ είναι οι διαστάσεις του φίλτρου να μειώνονται για φιλτράρισμα υψηλής ευκρίνειας,

[1] Σύμφωνα με τον ορισμό το φίλτρο πρέπει να αναστραφεί.

- η επικάλυψη των παραθύρων ή το βήμα του φιλτραρίσματος (stride) S, που καθορίζει κάθε πόσα πίξελ θα εφαρμοστεί ο πυρήνας και πρέπει να ισχύει $S \leqslant F$, ώστε να μην μένουν πίξελ χωρίς φιλτράρισμα,

- η μορφή της, κανονική ή επεκτεταμένη, στο Σχήμα 3.2 γίνεται επεκτεταμένη συνέλιξη, δηλαδή πρώτα επεκτείνεται η εικόνα με κάποια τεχνική έξω από τα όριά της και μετά εκτελείται η πράξη κανονικά. Το ζήτημα αυτό προκύπτει από τις επιθυμητές διαστάσεις της εικόνας εξόδου· αν θέλουμε η φιλτραρισμένη εικόνα να έχει ίδια διάσταση τότε επεκτείνουμε όσο χρειάζεται την εικόνα, σε διαφορετική περίπτωση μετά από διαδοχικές συνελίξεις η εικόνα θα μικραίνει,

- κανονικοποίηση τιμών: οι τιμές της εικόνας πρέπει να έχουν ένα γνωστό εύρος, ώστε να μπορούν να απεικονιστούν και να επεξεργαστούν σωστά. Αν γίνει συνέλιξη με ένα τυχαίο, αστάθμητο φίλτρο οι τιμές της φιλτραρισμένης εικόνας θα έχουν άγνωστο εύρος. Γι' αυτό απαραίτητο είναι οι τιμές του φίλτρου να κανονικοποιηθούν στο εύρος των τιμών της εικόνας. Στο Σχήμα 3.2 το φίλτρο πρέπει να πολλαπλασιαστεί με το συντελεστή $1/8$ (άθροισμα απόλυτων τιμών συντελεστών).

Η συνέλιξη έχει κάποιες ιδιότητες, οι οποίες χρησιμοποιούνται για την επιτάχυνση των υπολογισμών στα ΣΝΔ. Οι πιο σημαντικές από αυτές είναι:

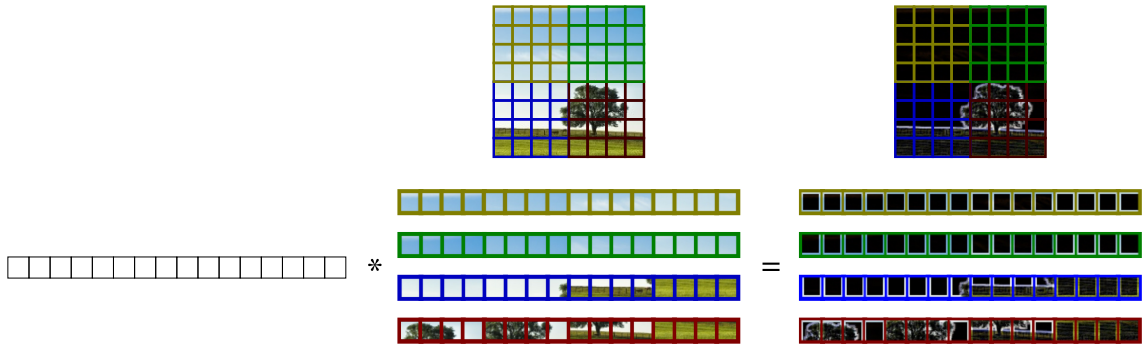
- Γραμμικότητα: Λόγω της γραμμικότητας, το φιλτράρισμα κάθε περιοχή μπορεί να υπολογιστεί ανεξάρτητα και επομένως παράλληλα. Αυτό αξιοποιείται σε πολυπύρηνες αρχιτεκτονικές και ο χρόνος συνέλιξης μπορεί να μειωθεί δραματικά.

- Αντιμεταθετικότητα: $x * w = w * x$. Μέσω αυτής της ιδιότητας για τον υπολογισμό του αποτελέσματος μπορεί να «κυλίεται» το φίλτρο πάνω στην εικόνα ή το αντίστροφο, κάτι που επίσης λαμβάνεται υπόψη για επιτάχυνση των υπολογισμών.

- Παραγώγιση: $\frac{d}{dt}(x * w) = \frac{dx}{dt} * w = x * \frac{dw}{dt}$. Η ιδιότητα αυτή χρησιμοποιείται κατά την οπισθοδιάδοση κλίσεων.

Μία διαφορετική θεώρηση της κλασικής 2D συνέλιξης

Σύμφωνα με το κλασσικό γραμμικό φιλτράρισμα το φίλτρο κυλίεται πάνω στην εικόνα εφαρμόζεται όσες φορές χρειάζεται και προκύπτει η φιλτραρισμένη εικόνα. Έστω ότι αντί για κύλιση του ίδιου φίλτρου σε όλες τις περιοχές έχουμε ένα φίλτρο αφιερωμένο κάθε περιοχή. Το φιλτράρισμα αποτελείται από πολλαπλασιασμούς και αθροίσεις και εν τέλει αντιστοιχεί σε ένα εσωτερικό γινόμενο, λειτουργίες που επιτελεί ένας κλασικός νευρώνας, όπως αυτός του Σχήματος 2.1. Αντικαθιστώντας κάθε τοπικό φίλτρο με έναν τοπικό νευρώνα έχουμε ένα στρώμα νευρώνων, όπου ο καθένας βλέπει σε μία συγκεκριμένη περιοχή της εικόνας (έχει περιορισμένο οπτικό πεδίο). Εκπαιδεύοντας αυτούς τους νευρώνες, το στρώμα εφαρμόζει ένα «βέλτιστο» φιλτράρισμα ως προς ένα «υποκειμενικό» κόστος που έχει επιλεγεί από εμάς.



Σχήμα 3.3: Οι συντελεστές του φίλτρου και οι τιμές των πίξελ μπορούν να οργανωθούν σε μονοδιάστατα διανύσματα και η συνέλιξη να υλοποιηθεί μέσω εσωτερικού γινομένου. Στο συγκεκριμένο παράδειγμα $\mathcal{F} = 4,\ \mathcal{S} = 4$. Μετά τις πράξεις τα διανύσματα οργανώνονται ξανά στη χωρική δομή της εικόνας.

Επίπεδο Συνέλιξης

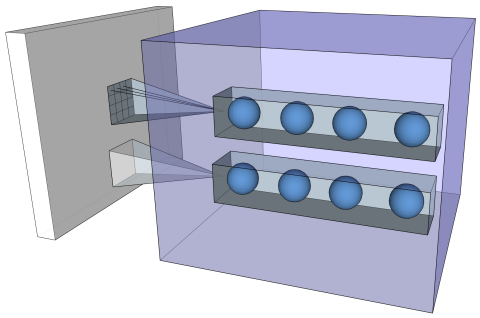
Το επίπεδο αυτό βρίσκεται σε κάθε ΣΝΔ, καθώς αποτελεί τη βάση για την εξαγωγή χαρακτηριστικών από τις εικόνες. Στα πλήρως συνδεδεμένα δίκτυα της Ενότητας 2.5 κάθε νευρώνας ενός επιπέδου συνδεόταν με όλες τις εξόδους των νευρώνων του προηγούμενου επιπέδου. Αντίθετα, στα ΣΝΔ κάθε νευρώνας «βλέπει» μόνο μία συγκεκριμένη περιοχή της εικόνας εισόδου, συνήθως μία τετραγωνική περιοχή μερικών πίξελ, και αγνοεί τελείως την υπόλοιπη εικόνα.

Επομένως, ενώ προηγουμένως είχαμε K νευρώνες ενός επιπέδου που έβλεπαν σε όλη την είσοδο, τώρα έχουμε D ομάδες/στρώματα (φέτες στο Σχήμα 3.4), από $W \times H$ νευρώνες η κάθε μία, που ο καθένας τους βλέπει σε περιορισμένη περιοχή τη εισόδου. Οι νευρώνες κάθε εγκάρσιου στρώματος του όγκου νευρώνων μπορεί να θεωρηθούν ότι αποτελούν ένα εκπαιδεύσιμο, τοπικό φίλτρο πάνω στην εικόνα. Η χωροθέτηση των νευρών αλλάζει σε σχέση με τα κλασικά ΝΔ και σε αυτή αυτή την περίπτωση, ένα ολόκληρο στρώμα νευρώνων βλέπει όλη την είσοδο και όχι κάθε νευρώνας ξεχωριστά. Η τρίτη διάσταση (βάθος – D), δηλώνει πόσο φίλτρα εφαρμόζονται στην εικόνα και αντιστοιχεί στη μία και μοναδική διάσταση των κρυμμένων επιπέδων των κλασικών ΝΔ.

Κάθε νευρώνας υλοποιεί την πράξη της συνέλιξης, που αναλύεται σε πολλαπλασιασμούς βαρών και χαρακτηριστικών και αθροίσεις. Μετά το συνελικτικό φιλτράρισμα για να ολοκληρωθεί η λειτουργία του κάθε νευρώνα το αποτέλεσμα της συνέλιξης διέρχεται από τη συνάρτηση ενεργοποίησής του. Η δημοφιλέστερη επιλογή είναι η ReLU, ωστόσο πολλές άλλες έχουν προταθεί με εξίσου καλά αποτελέσματα. Οι συναρτήσεις ενεργοποίησης περιγράφονται στην Ενότητα 2.2.

Υπερπαράμετροι του επιπέδου συνέλιξης

1. Βάθος επιπέδου (depth) $\mathcal{D}$ ή πλήθος φίλτρων $\mathcal{K}$: Καθορίζει το πλήθος των νευρώνων που κοιτούν στην ίδια περιοχή του όγκου εισόδου (αντιστοιχεί στο πλήθος των νευρώνων ενός επιπέδου στο MLP). Καθένας από αυτούς τους νευρώνες θα ενεργοποιείται για διαφορετικά χαρακτηριστικά της εισόδου, για παράδειγμα ένας μπορεί να ενεργοποιείται (έχει μεγάλη τιμή εξόδου) για το πρόσωπο και άλλος για το σώμα.



Σχήμα 3.4: Απεικόνιση Συνελικτικού Επιπέδου. Κάθε στρώμα νευρώνων ως σύνολο βλέπει όλη την εικόνα εισόδου, και κάθε στήλη νευρώνων έχει ίδιο οπτικό πεδίο.

2. Οπτικό πεδίο νευρώνων (receptive field) $\mathcal{F}$: Καθορίζει πόσα πίξελ (το εμβαδόν) του όγκου εισόδου «βλέπουν» οι νευρώνες κάθε στήλης. Τα πρώτα συνελικτικά επίπεδα ενός δικτύου έχουν μικρά οπτικά πεδία πάνω στην εικόνα εισόδου και εξάγουν τοπικά χαρακτηριστικά. Τα βαθύτερα επίπεδα έχουν πολύ μεγαλύτερο οπτικά πεδία πάνω στην εικόνα και εξάγουν ευρύτερα σημασιολογικά χαρακτηριστικά. Χάρη στο περιορισμένο οπτικό πεδίο τα ΣΝΔ έχουν πολύ μικρότερο αριθμό παραμέτρων και μεγαλύτερο βάθος, από το αν υλοποιούνταν ως πλήρως συνδεδεμένα δίκτυα.

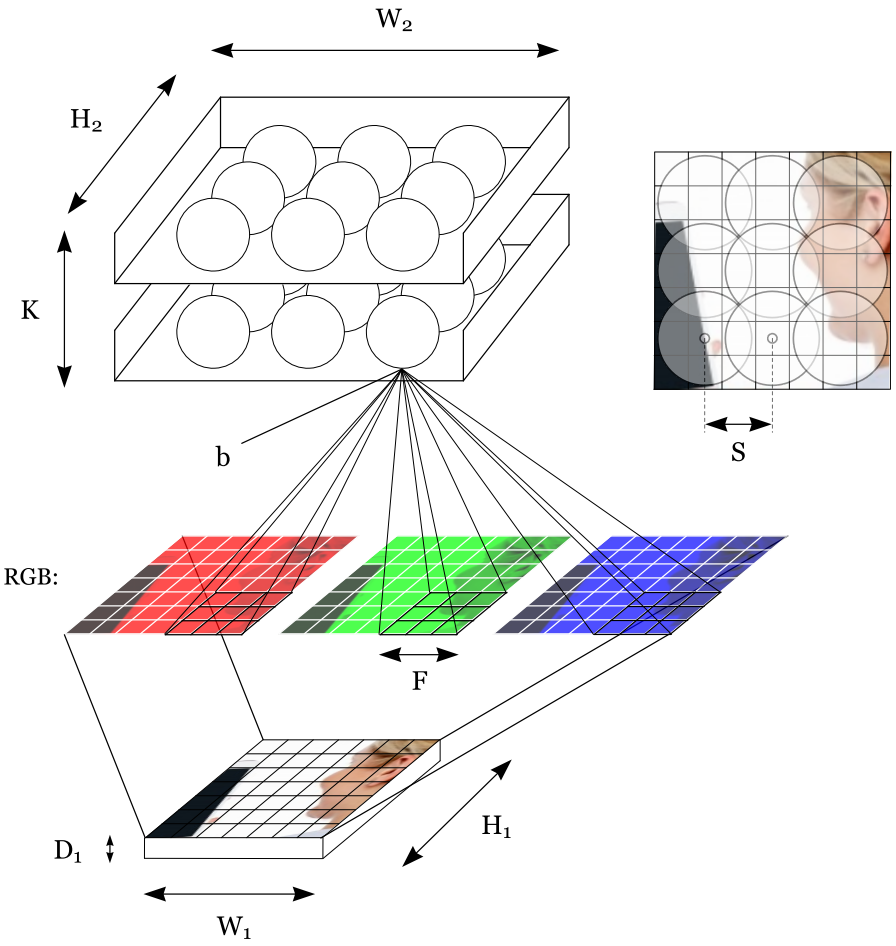
3. Βήμα φιλτραρίσματος (stride) $\mathcal{S}$: Το βήμα φιλτραρίσματος καθορίζει κάθε πόσα πίξελ θα γίνεται φιλτράρισμα, δηλαδή καθορίζει την επικάλυψη των δεκτικών πεδίων. Για παράδειγμα όταν $\mathcal{S} = 1$ τα οπτικά πεδία κάθε στήλης νευρώνων έχουν μεγάλη επικάλυψη και ο όγκος εξόδου θα έχει μεγάλες διαστάσεις μήκους (πλάτος, ύψος: $W,\ H$). Αντίθετα, αν επιλέξουμε μεγάλο βήμα τα οπτικά πεδία θα επικαλύπτονται λιγότερο ή καθόλου και ο όγκος εξόδου θα έχει μικρότερες διαστάσεις μήκους.

4. Επέκταση εισόδου για υπολογισμό συνέλιξης (padding) $\mathcal{P}$: Όπως εξηγήθηκε παραπάνω μέσω της επέκτασης της εισόδου μπορούμε να καθορίζουμε τις διαστάσεις μήκους του όγκου εξόδου. Ο τρόπος επέκτασης επηρεάζει την εξαγωγή χαρακτηριστικών και πρέπει να λαμβάνεται υπόψη.

Οι υπερπαράμετροι αν και δεν υπάρχει περιορισμός να είναι ίδιοι για κάθε νευρώνα, στην πράξη επιλέγονται ίδιοι, ώστε κάθε επίπεδο του δικτύου να περιγράφεται από μόνο 4 παραμέτρους και με σκοπό την γενική απλοποίηση περιγραφής και υλοποίησης. Στο Σχήμα 3.5 φαίνεται πως τροφοδοτείται κάθε νευρώνας ενός συνελικτικού επιπέδου από τον όγκο εισόδου και όλοι οι υπερπαράμετροι που παίρνουν μέρος.

Διαμοιρασμός Βαρών

Με το παραπάνω σκεπτικό ο αριθμός των βαρών σε κάθε επίπεδο γίνεται πολύ μεγάλος, αυτό θα φανεί με ένα παράδειγμα, στο οποίο χρησιμοποιούμε το υπάρχον δίκτυο AlexNet που αναλύεται στην Ενότητα 3.4.1.



Σχήμα 3.5: Απεικόνιση Όγκου Εισόδου, Νευρώνων Συνελικτικού Επιπέδου και οπτικού πεδίου. Η RGB εικόνα εισόδου έχει διαστάσεις $7 \times 7 \times 3$, το επίπεδο της συνέλιξης έχει υπερπαραμέτρους $\mathcal{K} = 2$, $\mathcal{F} = 3$, $\mathcal{S} = 2$, $\mathcal{P} = 0$ και ο όγκος εξόδου θα έχει διαστάσεις $3 \times 3 \times 2$.

Όπως διαπιστώνεται και στο [Karp15], το πρώτο συνελικτικό επίπεδο (Σχήμα 3.13) έχει διαστάσεις $55 \times 55 \times 96$, δηλαδή περιλαμβάνει $55 \times 55 = 3025$ νευρώνες σε καθένα από τα 96 στρώματα, δηλαδή συνολικά $290,400$ νευρώνες. Ο καθένας από αυτούς τους νευρώνες έχει $11 \times 11 \times 3 + 1 = 364$ βάρη και ένα κατώφλι, καταλήγοντας έτσι μόνο το πρώτο επίπεδο να έχει $290,400 \times 364 = 105,705,600$ βάρη και κατώφλια, ένα τεράστιο νούμερο για οποιαδήποτε επεξεργασία.

Αντί αυτού στην πράξη περιορίζονται οι νευρώνες που ανήκουν σε ένα φίλτρο να έχουν τα ίδια βάρη, δηλαδή ένα στρώμα με 3025 νευρώνες, αντί να έχει 3025×364 βάρη έχει μόνο 364. Με αυτόν τον τρόπο τα βάρη μειώνονται σε $96 \times (11 \times 11 \times 3 + 1) = 34,944$ ένα πιο εύκολα διαχειρίσιμο νούμερο.

Ο περιορισμός αυτός, ωστόσο, δεν είναι αυθαίρετος, και βασίζεται στην λογική υπόθεση ότι η εφαρμογή ενός καλού φίλτρου θα έχει χρήσιμα αποτελέσματα (εξαγόμενα χαρακτηριστικά) σε όποια θέση και αν εφαρμοστεί. Αυτή η υπόθεση δεν ισχύει πάντα όμως, για παράδειγμα αν μια εικόνα έχει ένα πρόσωπο ιδανικά επιθυμούμε από κάθε περιοχή να εξάγονται ειδικά χαρακτηριστικά που αντιστοιχούν στα μάτια, στο στόμα, κτλ. Όταν θα το επιτρέψουν οι υπολογιστικοί πόροι αξίζει να

διερευνηθεί, αν αμελώντας το διαμοιρασμό παραμέτρων, τα αποτελέσματα βελτιώνονται αισθητά.

Κατά την εκπαίδευση και το backpropagation για κάθε νευρώνα υπολογίζεται ξεχωριστά η κλίση, αλλά κατά την φάση ενημέρωσης των βαρών όλες οι αλλαγές προστίθενται και ενημερώνεται μόνο ένα κοινό σετ βαρών για κάθε στρώμα.

Σημειώνεται ότι μέσω του διαμοιρασμού βαρών κάθε στρώματα αποτελεί ένα κανονικό, συμβατικό φίλτρο με το οποίο συνελίσσεται η είσοδος, δηλαδή κανονικό γραμμικό φιλτράρισμα. Δεδομένου ότι η είσοδος έχει συνήθως πολλά κανάλια το καθένα από αυτά τα φίλτρα έχουν και αυτά αντίστοιχο αριθμό καναλιών (Σχήμα 3.15).[2]

Συνοπτικά το επίπεδο συνέλιξης χαρακτηρίζεται από τα εξής στοιχεία:

- Είσοδος: όγκος διαστάσεων $W_1 \times H_1 \times D_1$ (πολυκαναλική εικόνα ή όγκος χαρακτηριστικών).
- Υπερπαράμετροι: 4 παράμετροι που καθορίζουν τη συνέλιξη: ο αριθμός των φίλτρων K, το μέγεθος των φίλτρων F, το βήμα του φιλτραρίσματος S και ο αριθμός των πίξελ για την επέκταση της εικόνας συμμετρικά P.
- Έξοδος: όγκος διαστάσεων $W_2 \times H_2 \times D_2$, για τον οποίον ισχύουν οι σχέσεις:
  - $W_2 = (W_1 - F + 2P)/S + 1$,
  - $H_2 = (H_1 - F + 2P)/S + 1$,
  - $D_2 = K$.
- Βάρη: εισάγει γενικά $\underbrace{(W_2 \cdot H_2 \cdot D_2)}_{\text{πλήθος νευρώνων}} \cdot \underbrace{(F \cdot F \cdot D_1 + 1)}_{\text{βάρη νευρώνα}}$ βάρη και κατώφλια ή $D_2 \cdot (F \cdot F \cdot D_1 + 1)$ στην περίπτωση του διαμοιρασμού βαρών.

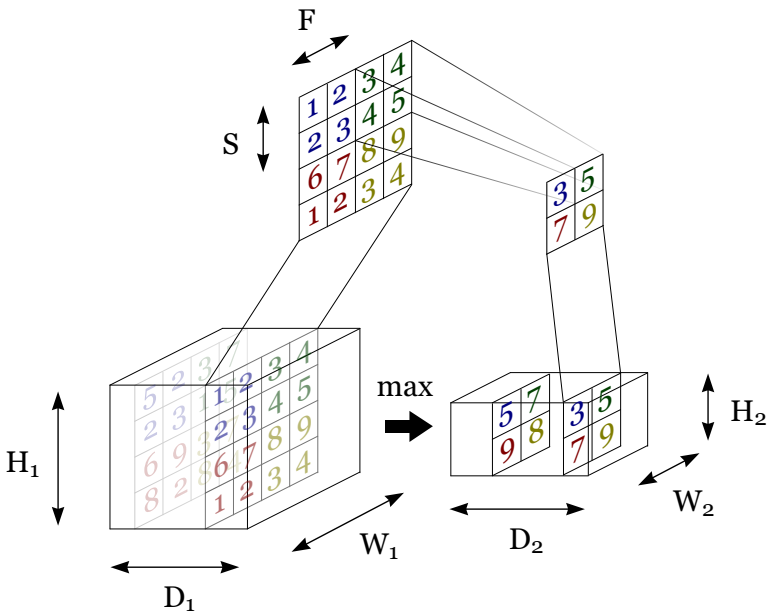
Οπισθοδιάδοση

Η συνέλιξη αποτελείται από τις θεμελιώδεις πράξεις του πολλαπλασιασμού και της πρόσθεσης, οι οποίες είναι διαφορίσιμες, επομένως ο κανόνας της αλυσίδας είναι απλός και η οπισθοδιάδοση δεν θέλει κάποια ιδιαίτερη τεχνική.

3.2.2 Επίπεδο Συγκέντρωσης/Υποδειγματοληψίας

Το επίπεδο αυτό μειώνει τις διαστάσεις του όγκου χαρακτηριστικών, και κατ' επέκταση το αναγκαίο πλήθος παραμέτρων του δικτύου και διαισθητικά τα συνοψίζει, ενώ προσφέρει και την ιδιότητα του αναλλοίωτου στη μετατόπιση. Η λειτουργία του επιπέδου αυτού εφαρμόζεται ανεξάρτητα σε κάθε στρώμα του όγκου/χάρτη χαρακτηριστικών, μειώνοντας τις διαστάσεις μήκους, χωρίς να επηρεάζεται το βάθος τους. Υπάρχουν πολλά διαφορετικές τεχνικές συγκέντρωσης, αλλά οι δύο κυρίως χρησιμοποιούμενες είναι η συγκέντρωση μεγίστου (max pooling) και η συγκέντρωση μέσου όρου (average pooling). Στο Σχήμα 3.6 φαίνεται γραφικά ένα παράδειγμα συγκέντρωσης μεγίστου.

[2] Όταν δεν γίνεται διαμοιρασμός βαρών θα αναφέρουν κάθε εγκάρσια φέτα βαρών του συνελικτικού επιπέδου ως στρώμα, όταν γίνεται διαμοιρασμός όμως το κάθε στρώμα έχει λειτουργία ισοδύναμη με αυτή ενός κλασικού φίλτρου και επομένως θα αναφέρεται έτσι.



Σχήμα 3.6: Απεικόνιση δράσης επιπέδου συγκέντρωσης μεγίστου σε όγκο εισόδου διαστάσεων $4 \times 4 \times D_1$. Οι υπερπαράμετροι για αυτό το παράδειγμα είναι $\mathcal{F} = 2,\ \mathcal{S} = 2$ και ο όγκος εξόδου έχει διαστάσεις $2 \times 2 \times D_1$.

Συνοπτικά το επίπεδο της συγκέντρωσης χαρακτηρίζεται από:

- Είσοδος: όγκος διαστάσεων $W_1 \times H_1 \times D_1$.
- Υπερπαράμετροι: 2 παράμετροι που καθορίζουν τη συγκέντρωση: το μέγεθος περιοχής $\mathcal{F}$ και το βήμα συγκέντρωσης $\mathcal{S}$.
- Έξοδος: όγκος διαστάσεων $W_2 \times H_2 \times D_2$, για τον οποίον ισχύουν οι σχέσεις:
  - $W_2 = (W_1 - F)/S + 1$,
  - $H_2 = (H_1 - F)/S + 1$,
  - $D_2 = D_1$.
- Βάρη: δεν εισάγει βάρη, εκτελεί προκαθορισμένη λειτουργία.

Το επίπεδο αυτό δεν περιλαμβάνει νευρώνες με τη συμβατική λειτουργία και στην βιβλιογραφία δεν αναφέρεται σαν επίπεδο νευρώνων, αλλά θα μπορούσε να αντιστοιχηθεί σε μη-γραμμικούς νευρώνες, οι οποίοι έχουν έξοδο την είσοδο με τη μέγιστη τιμή. Επίσης, αν προηγείται ένα συνελικτικό επίπεδο, τα δύο επίπεδα μπορούν να θεωρηθούν ότι λειτουργούν ως ένας συμβατικός νευρώνας.

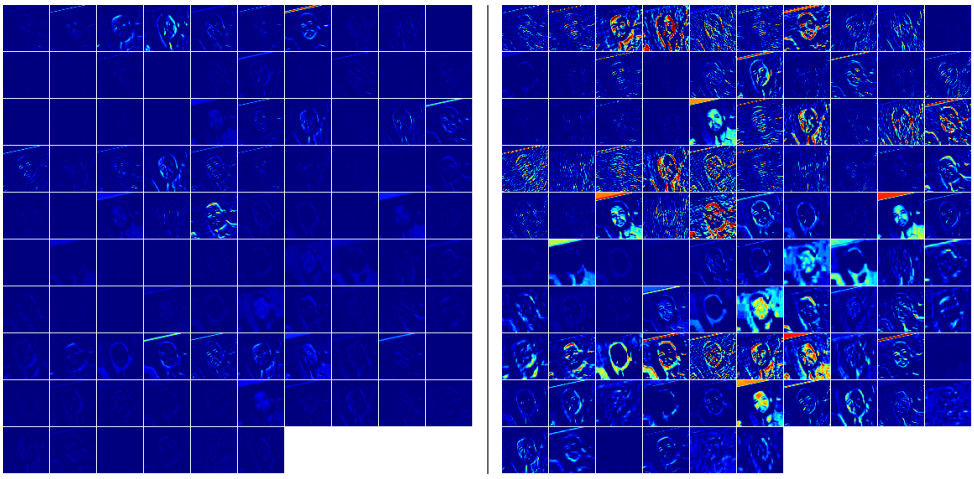
Οπισθοδιάδοση

Η οπισθοδιάδοση στην περίπτωση της συγκέντρωσης μεγίστου πρέπει να υλοποιηθεί με προσοχή. Όπως αναφέρθηκε στην Ενότητα 2.4.4 η οπισθοδιάδοση στην πράξη $\max(\cdot)$ γίνεται μεταφέροντας την κλίση εξόδου στην είσοδο, με τη μεγαλύτερη τιμή κατά την εμπρόσθια διάδοση. Επομένως, για

κάθε οπτικό πεδίο πρέπει να κρατηθεί η θέση του μεγίστου, ώστε σε αυτή να προωθηθεί η κλίση, ενώ στις υπόλοιπες θέσεις να μεταφερθεί μηδενική κλίση.

3.2.3 Επίπεδο Κανονικοποίησης

Οι αρχικές ιδέες για την εισαγωγή των επιπέδων κανονικοποίησης προέρχονται από το πεδίο της υπολογιστικής νευροβιολογίας ([DiCa08], [Lyu08]). Τα επίπεδα αυτά χρησιμοποιήθηκαν στα πρώτα ΣΝΔ με πολλές παραλλαγές και πιο επιτυχημένη εφαρμογή στο AlexNet [Kriz12] (βλ. Ενότητα 3.4.1), ωστόσο σιγά-σιγά εκλείπουν, καθώς πρακτικά η συνεισφορά τους είναι πολύ μικρή. Μπορούν να χρησιμοποιηθούν σε περιπτώσεις, όπου οι ενεργοποιήσεις των νευρώνων δεν είναι φραγμένες (π.χ. ReLU), καταπιέζοντας τοπικά ομοιόμορφες περιοχές με πολύ υψηλές τιμές χαρακτηριστικών, οι οποίες σε διαφορετική περίπτωση θα επηρέαζαν γειτονικές περιοχές με πολύ χαμηλές τιμές, ενώ ταυτόχρονα ενισχύει τις χαμηλές ενεργοποιήσεις. Με αυτόν τον τρόπο εισάγεται ανταγωνισμός ανάμεσα στους νευρώνες κατά τη διάρκεια της εκπαίδευσης. Στο Σχήμα 3.7 φαίνονται οι όγκοι εισόδου και εξόδου αυτού του επιπέδου από την αρχιτεκτονική AlexNet (Ενότητα 3.4.1).



Σχήμα 3.7: Αριστερά: Τα 96 στρώματα ενεργοποιήσεων του πρώτου συνελικτικού επιπέδου του AlexNet, με διαστάσεις όγκου: $55 \times 55 \times 96$. Δεξιά: Τα 96 στρώματα (όγκος εξόδου) μετά το επίπεδο κανονικοποίησης. Η έγχρωμη απεικόνιση αντιστοιχεί σε κλίμακα έντασης των ενεργοποιήσεων: μπλε - χαμηλές τιμές, κόκκινο - υψηλές.

Το ενδιαφέρον σημείο αυτού του επιπέδου είναι ότι η κανονικοποίηση μπορεί να γίνει αποκλειστικά εντός ενός χάρτη χαρακτηριστικών, αλλά και εγκάρσια στον όγκο χαρακτηριστικών, δηλαδή ως προς πολλούς γειτονικούς χάρτες. Τέλος, το επίπεδο αυτό, δεν επηρεάζει καμία διάσταση του όγκου χαρακτηριστικών παρά μόνο τις τιμές τους.

3.2.4 Πλήρως συνδεδεμένο Επίπεδο

Τα τρία προηγούμενα επίπεδα χρησιμοποιούνται για την εξαγωγή χαρακτηριστικών. Η κύρια λειτουργία του ΝΔ (π.χ. ταξινόμηση) επιτυγχάνεται από αυτά τα επίπεδα, τα οποία αν τοποθετηθούν σε αλληλουχία αποτελούν ένα κλασικό πολυστρωματικό ΝΔ (π.χ. MLP). Όπως και στα απλά ΝΔ τα

επίπεδα αυτά αποτελούνται από κλασικούς νευρώνες. Η είσοδος σε ένα τέτοιο επίπεδο είναι ένα διάνυσμα χαρακτηριστικών, όπως αναφέρθηκε στην Ενότητα 2.5. Έτσι ο όγκος εισόδου μετατρέπεται σε ένα μονοδιάστατο διάνυσμα «διανυσματοποιείται» χάνοντας κάθε ιδιότητα τοπικότητας. Οι νευρώνες μεταξύ δύο τέτοιων επιπέδων συνδέονται πλήρως (όλοι με όλους) και η μάθηση γίνεται, όπως στην περίπτωση των κλασικών δικτύων.

Απέχει πολύ ένα πλήρως συνδεδεμένο από ένα συνελικτικό επίπεδο; - Μετατροπή Πλήρως συνδεδεμένου σε Συνελικτικό Επίπεδο

Και τα δύο επίπεδα αποτελούνται από νευρώνες, που δέχονται δεδομένα εισόδου και παράγουν ενεργοποιήσεις. Κάθε νευρώνας ενός πλήρως συνδεδεμένου επιπέδου υπολογίζει το εσωτερικό γινόμενο του διανύσματος εισόδου με τα βάρη του και εφαρμόζει τη μη γραμμικότητά του, παράγοντας μία απόκριση, όπως αναλύθηκε στην Ενότητα 2.5. Στην περίπτωση του συνελικτικού επιπέδου η λειτουργία είναι ακριβώς η ίδια, αν ο όγκος εισόδου «διανυσματοποιηθεί», όπως στο Σχήμα 3.3, με τη μόνη διαφορά ότι οι νευρώνες έχουν περιορισμένο οπτικό πεδίο πάνω στην είσοδο. Με αυτό το σκεπτικό οι νευρώνες του πλήρως συνδεδεμένου επιπέδου, μπορεί να θεωρηθούν ότι έχουν οπτικό πεδίο που καλύπτει όλη την είσοδο.

Επομένως, η μόνη διαφορά ανάμεσα στα δύο επίπεδα είναι η οργάνωση στο χώρο, όπως αναφέρεται και στο [Karp15]. Οι νευρώνες ενός συνελικτικού επιπέδου οργανώνονται σε τριδιάστατους όγκους, διαστάσεων $W \times H \times D$, όπου κάθε στήλη (της τρίτης διάστασης) έχει ίδιο οπτικό πεδίο. Οι K νευρώνες του πλήρως συνδεδεμένου οργανώνονται σε μία διάσταση και έχουν όλοι το ίδιο οπτικό πεδίο. Θεωρώντας λοιπόν, αυτούς να αποτελούν μία στήλη $1 \times 1 \times K$, το επίπεδο ισοδυναμεί με ένα συνελικτικό με υπερπαραμέτρους $\mathcal{D} = K,\ \mathcal{F} = W_1 = H_1,\ \mathcal{S} = 1,\ \mathcal{P} = 0$. Αν μετά ακολουθούν κι άλλα πλήρως συνδεδεμένα επίπεδα, η διαδικασία είναι ίδια με μόνη αλλαγή το μέγεθος του φίλτρου, που είναι $\mathcal{F} = 1$. Παρόμοια, ένα συνελικτικό μπορεί να μετατραπεί σε ένα πλήρως συνδεδεμένο επίπεδο.

Η παρατήρηση αυτή προσφέρει επιτάχυνση στους υπολογισμούς καθώς οι λειτουργίες του πλήρως συνδεδεμένου επιπέδου μπορούν να επωφεληθούν από τους γρήγορους αλγορίθμους για τη συνέλιξη. Επίσης, μετατρέποντας τα πλήρως συνδεδεμένα επίπεδα σε συνελικτικά, ο περιορισμός του μεγέθους της εισόδου εξαλείφεται και τα επίπεδα γίνονται «αγνωστικά» (agnostic) ως προς τις διαστάσεις μήκους της εισόδου.

Με αυτό τον τρόπο, αν σε ένα ΣΝΔ ταξινόμησης υπάρχουν πλήρως συνδεδεμένα επίπεδα και άρα επιβάλλεται περιορισμός στη διάσταση εισόδου (π.χ. 227×227 στο AlexNet), το δίκτυο μπορεί να μετατραπεί σε δίκτυο πυκνής εξόδου, ανεξάρτητα από τη διάσταση της εισόδου 3.3.9. Στην περίπτωση αυτή, αν χρησιμοποιείται κάποιο προεκπαιδευμένο δίκτυο θα πρέπει να γίνει προσεκτικά η μεταφορά βαρών από μορφή κατάλληλη για εσωτερικό γινόμενο σε μορφή κατάλληλη για συνέλιξη.

3.3 Εκπαίδευση και Ειδική Προσαρμογή ΣΝΔ

3.3.1 Μάθηση και Γενίκευση

Το σύνολο δεδομένων χωρίζεται συνήθως σε τρία υποσύνολα: το υποσύνολο εκπαίδευσης (training-set), το σύνολο επαλήθευσης (validation-set) και το σύνολο ελέγχου (testing-set). Από αυτά τα δύο πρώτα χρησιμοποιούνται στην εκπαίδευση του δικτύου, ενώ το τρίτο στην αξιολόγηση της επίδοσής του. Το μέγεθος του δικτύου και κατ' επέκταση ο αριθμός των ελεύθερων παραμέτρων (βαρών), πρέπει να επιλεγεί προσεκτικά και λαμβάνοντας υπόψη τα χαρακτηριστικά του συνόλου δεδομένων, ώστε να έχει μέγιστη ικανότητα γενίκευσης.

Υπο/Υπερπροσαρμογή

Για να αποφευχθεί ο κίνδυνος της υπο-προσαρμογής (underfitting), ο αριθμός των ελεύθερων παραμέτρων θα πρέπει να είναι αρκετά μεγάλος, ώστε να μπορεί το δίκτυο να μάθει να διαφοροποιεί κατηγορίες μεταξύ τους, αλλά και να καλύπτει παραλλαγές αντικειμένων μέσα στην ίδια κατηγορία.

Ωστόσο, αν ο αριθμός των παραμέτρων είναι άσκοπα μεγάλος, το δίκτυο θα αρχίσει να υπερπροσαρμόζεται (overfitting) στις ιδιαίτερες λεπτομέρειες του συνόλου εκπαίδευσης. Αυτό μπορεί να ανιχνευθεί από το σφάλμα επαλήθευσης, όπως φαίνεται στο 3.8. Επομένως, ο αριθμός των ελεύθερων παραμέτρων θα πρέπει να μην είναι τόσο μικρός, ώστε το δίκτυο να μην έχει τη δυνατότητα να μάθει τις υποκείμενες διαφορές ανάμεσα στα δεδομένα της ίδιας κατηγορίας.

Ο βαθμός της υπερπροσαρμογής φαίνεται από την απόσταση των καμπυλών του σφάλματος εκπαίδευσης και επαλήθευσης. Όταν λαμβάνει χώρα ασθενής υπερπροσαρμογή, τα σφάλματα επαλήθευσης και εκπαίδευσης είναι κοντά (Σχήμα 3.8) και αυτό δηλώνει ότι η «χωρητικότητα» του μοντέλου πρέπει να αυξηθεί, αυξάνοντας τον αριθμό των βαρών.

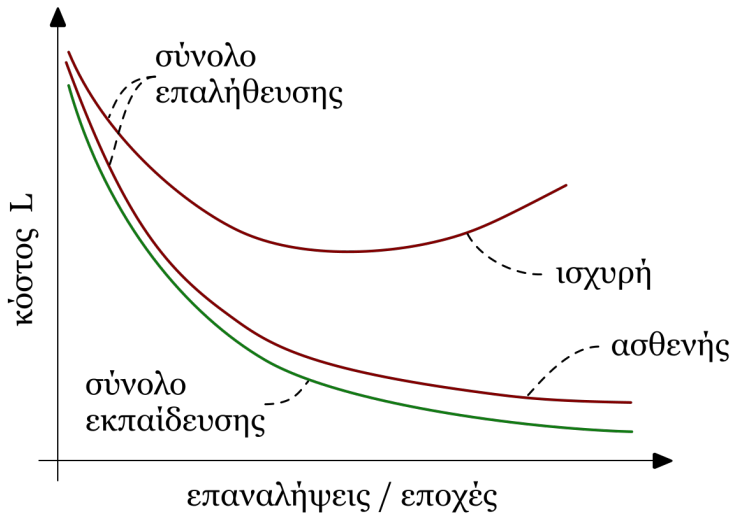
Αντίθετα, όταν η υπερπροσαρμογή είναι ισχυρή πρέπει να αυξηθεί ο βαθμός της κανονικοποίησης, είτε αυξάνοντας την υπερπαράμετρο λ του συνολικού κόστους 2.6, που επιβάλλει ομαλότητα, είτε εφαρμόζοντας μεγαλύτερη πιθανότητα αποκοπής συνδέσεων (dropout). Εναλλακτικά, πρέπει να επεκταθεί το σύνολο εκπαίδευσης, με επιπλέον παραδείγματα, ιδανικά, ασυσχέτιστα από τα υπάρχοντα, ή με τεχνικές επαύξησης δεδομένων.

Υπο/Υπερεκπαίδευση

Η υπερπροσαρμογή των παραμέτρων στα δεδομένα μπορεί να προκύψει και ως αποτέλεσμα υπερεκπαίδευσης και το σφάλμα/κόστος να αρχίζει να αυξάνεται, όπως φαίνεται στο Σχήμα 3.8. Για την αποφυγή του κινδύνου υπερεκπαίδευσης μπορεί να χρησιμοποιηθεί το σύνολο επαλήθευσης, που είναι συνήθως πολύ μικρότερο από το σύνολο εκπαίδευσης. Η εκπαίδευση πρέπει να σταματήσει όταν το σφάλμα στο σύνολο επαλήθευσης αρχίζει να αυξάνεται μετά το πρώτο ελάχιστο (early stopping).

3.3.2 Προεπεξεργασία δεδομένων

Η αναπαράσταση και η αποθήκευση των ψηφιακών εικόνων μπορεί να διαφέρει από σύστημα σε σύστημα· το ίδιο συμβαίνει με οποιαδήποτε μορφή δεδομένων, όμως εδώ επικεντρωνόμαστε στις



Σχήμα 3.8: Σφάλματα εξόδου των συνόλων επαλήθευσης και εκπαίδευσης κατά την υπερεκπαίδευση. Η υπερεκπαίδευση μπορεί να είναι ισχυρή ή ασθενής. Η μάθηση πρέπει να σταματήσει στην επανάληψη, όπου το σφάλμα έχει ελάχιστο.

εικόνες. Η συνήθης μορφή των εικόνων στους υπολογιστές είναι πολυδιάστατοι πίνακες με τιμές (πίξελ), κυρίως στο διάστημα $[0, 255]$. Ένα σύστημα μάθησης πρέπει να έχει τα ίδια αποτελέσματα ανεξάρτητα από την αναπαράσταση, γι' αυτό πολλές φορές είναι αναγκαία η προ-επεξεργασία των εικόνων.

Μια εικόνα εισόδου μπορεί, εκτός από ακατέργαστες τιμές πίξελ, να θεωρηθεί ότι περιέχει τιμές χαρακτηριστικών χωρικά κατανεμημένων, που περιγράφουν το περιεχόμενό της. Για να χρησιμοποιηθούν αυτά τα χαρακτηριστικά αποδοτικά πρέπει να υποστούν κάποια προ-επεξεργασία, αλλιώς η επίδρασή τους στη συνάρτηση κόστους θα είναι ασύμμετρη. Μερικές από τις πιο συχνά χρησιμοποιούμενες τεχνικές είναι οι εξής:

1. Κανονικοποίηση ως προς τη μέση τιμή ή/και τη διασπορά (z-normalization) των δεδομένων εκπαίδευσης. Τόσο οι εικόνες πριν την μάθηση, όσο και οι εικόνες δοκιμής σχεδόν πάντα πριν την εκπαίδευση ΣΝΔ κανονικοποιούνται ως προς τη μέση τιμή, δηλαδή αφαιρείται από όλες, η μέση εικόνα του συνόλου εκπαίδευσης, ούτως ώστε να έχουν μηδενική μέση τιμή. Επιπλέον, μπορεί να γίνει και κανονικοποίηση ως προς τη διασπορά, ώστε να έχουν όλα τα δεδομένα μοναδιαία διασπορά.

   Αν δεν υπάρχουν πολλά δεδομένα, το καλύτερο είναι να εφαρμοστεί ο μετασχηματισμός λεύκανσης (whitening transform), που μετασχηματίζει γραμμικά τα παραδείγματα ώστε να έχουν μηδενική μέση τιμή, μοναδιαία διασπορά και να είναι ασυσχέτιστα. Επίσης, μπορεί να εφαρμοστεί και Ανάλυση Κυρίων Συνιστωσών (PCA) για μείωση των διαστάσεων, ή οποιοσδήποτε άλλος παρόμοιος Μ/Σ.

   Η κανονικοποίηση των δεδομένων συμβάλλει στην επιτάχυνση της μάθησης [Orr03, Beng15]. Αν για παράδειγμα, όλα τα πίξελ της εικόνας εισόδου έχουν θετικές τιμές, το ποσό της ανανέωσης από τον κανόνα μάθησης θα έχει ίδιο πρόσημο για όλα τα βάρη, με αποτέλεσμα να

μετατοπίζονται όλα προς την ίδια κατεύθυνση. Με αυτόν τον τρόπο τα βάρη θα αυξάνονται ή θα μειώνονται όλα μαζί και η εκπαίδευση θα γίνεται πάνω σε μία ευθεία. Γι' αυτό το λόγο είναι πολύ σημαντικό να κεντράρονται τα δεδομένα ως προς τη μέση τιμή τους. Παρόμοια, είναι επιθυμητό τα παραδείγματα να είναι όσο το δυνατόν ασυσχέτιστα και να έχουν μικρή διασπορά εντός των κατηγοριών.

2. Κλιμάκωση των τιμών των χαρακτηριστικών σε ένα προκαθορισμένο διάστημα π.χ. $[-1, 1]$ ή $[0, 1]$ μέσω γραμμικών ή μη γραμμικών (όταν η κατανομή γύρω από τη μέση τιμή δεν είναι ομοιόμορφη, π.χ. softmax) μετασχηματισμών.

3. Αποκοπή ακραίων τιμών (outlier removal). Αν γνωρίζουμε, για παράδειγμα, ότι σε περιοχές των εικόνων υπάρχει κορεσμός στη φωτεινότητα καλό θα ήταν να αντιμετωπίσουμε διαφορετικά αυτές τις περιοχές, είτε θεωρώντας τες ως ελλιπή δεδομένα, είτε αποκόπτοντάς τες.

3.3.3 Καθορισμός υπερπαραμέτρων – Μέθοδος Cross-Validation

Οι βασικές υπερπαράμετροι ενός ΣΝΔ, που πρέπει να επιλεγούν εξαρχής από το σχεδιαστή του δικτύου είναι ο γενικός ρυθμός μάθησης και το χρονοδιάγραμμά του, ενδεχομένως η ορμή, ο αριθμός των εποχών εκπαίδευσης, οι επιλογές για τα μεγέθη του δικτύου (αριθμός επιπέδων, νευρώνων, υπερπαράμετροι επιπέδων), κ.ά. Μερικές από τις υπερπαραμέτρους είναι πιο ευαίσθητες από άλλες, οπότε αν υπάρχει επαρκώς μεγάλο σύνολο δεδομένων, υπολογιστική ισχύς και χρόνος, αυτές μπορούν να επιλεγούν μέσω της μεθοδου cross-validation.

Σύμφωνα με αυτή, έστω ότι θέλουμε να επιλέξουμε την καλύτερη τιμή μιας υπερπαραμέτρου από m υποψήφιες τιμές. Χωρίζουμε το αρχικό σύνολο εκπαίδευσης σε k μέρη (folds) και για κάθε μία από τις m επιλογές, εκπαιδεύουμε το δίκτυο k φορές χρησιμοποιώντας για σύνολα εκπαίδευσης και επαλήθευσης, όλους τους δυνατούς συνδυασμούς από $k-1$ και 1 μέρη. Καταλήγουμε έτσι για κάθε τιμή από τις m να έχουμε k τιμές κόστους ή αποτελεσματικότητας για το σύνολο επαλήθευσης. Βρίσκοντας το μέσο όρο για κάθε m, επιλέγουμε την τιμή της υπερπαραμέτρου που έχει χειρότερο ή καλύτερο μέσο όρο αντίστοιχα.

3.3.4 Τεχνικές Επαύξησης Συνόλου δεδομένων (dataset augmentation)

Η εκπαίδευση των ΒΝΔ, λόγω της βαθιάς και ιεραρχικής δομής τους, απαιτεί μεγάλο πλήθος δεδομένων. Αν ένα σύνολο δεδομένων δεν επαρκεί για την εκπαίδευση, τότε είναι δυνατόν να επαυξηθεί με τεχνητές μεθόδους. Αυτές περιλαμβάνουν οποιοδήποτε μετασχηματισμό, γραμμικό ή μη, π.χ. αφινικούς μετασχηματισμούς (μετατόπιση, περιστροφή, κλιμάκωση), προσθήκη θορύβου με τη γενικότερη έννοια και τυχαίες περικοπές. Η επαύξηση δεδομένων ωφελεί περισσότερο τα προβλήματα ταξινόμησης και ανίχνευσης, αλλά εφαρμόζεται σε άλλα. Κατά τη διαδικασία επαύξησης θα πρέπει να εφαρμόζονται μετασχηματισμοί που δεν θα αλλοιώνουν το περιεχόμενο της εικόνας και κυρίως την κατηγορία που ανήκει. Για παράδειγμα στην αναγνώριση ψηφίων περιστροφές 180° δεν επιτρέπονται γιατί το «6» γίνεται «9».

3.3.5 Κανονικοποίηση (Regularization) στα ΣΝΔ

Μία διαφορετική μέθοδος αντιμετώπισης της υπερπροσαρμογής είναι η κανονικοποίηση, η οποία παρουσιάζεται στην παρούσα Ενότητα με ιδέες από τα [Beng15, Karp15]. Η θεωρία της Κανονικοποίησης προτάθηκε το 1963 από τον Tikhonov για την επίλυση κακώς ορισμένων (ill-posed) προβλημάτων. Τέτοια είναι τα προβλήματα που είτε δεν έχουν λύση, είτε έχουν πολλές, είτε αυτές είναι ευαίσθητες σε σχέση με την επιλογή παραμέτρων. Μέσω της κανονικοποίησης, δηλαδή της αλλαγής του τρόπου μάθησης, στοχεύουμε στην αύξηση της ικανότητας γενίκευσης του συστήματος, με κόστος την αύξηση του σφάλματος εκπαίδευσης.

Η «χωρητικότητα» και μαθησιακή ικανότητα ενός δικτύου είναι άλλες φορές μικρή και άλλες φορές μεγάλη ανάλογα με το εκάστοτε πρόβλημα. Μία λύση για να προσαρμόσουμε τη χωρητικότητα στο πρόβλημα είναι να προσθέσουμε ή αφαιρέσουμε επίπεδα και νευρώνες από το δίκτυο. Τις περισσότερες φορές όμως δεν θέλουμε να καταφύγουμε σε μία τέτοια στρατηγική, γι' αυτό εφαρμόζουμε τις τεχνικές κανονικοποίησης. Επίσης, τα περισσότερα προβλήματα έχουν απειρία λύσεων και τα βάρη μπορούν να πάρουν αυθαίρετα υψηλές τιμές, κάτι που δεν είναι επιθυμητό. Για να επιβάλλουμε μικρές τιμές και ομαλότητα στα βάρη επίσης καταφεύγουμε στην κανονικοποίηση.

Ένας τρόπος για να επιτευχθεί είναι η προσθήκη του όρου κανονικοποίησης στη γενική συνάρτηση κόστους 2.6 και την αντίστοιχης υπερπαραμέτρου λ, η οποία ζυγίζει την επιρροή του όρου στο συνολικό κόστος. Η προσθήκη αυτή δημιουργεί προτίμηση σε μικρές τιμές βαρών και τιμωρεί την πολυπλοκότητα των απεικονίσεων. Ωστόσο, η τιμή της δεν πρέπει να είναι πολύ μεγάλη, γιατί τότε οι κλίσεις σε έναν αλγόριθμο κατάβασης δυναμικού, θα προέρχονται κυρίως από τον όρο κανονικοποίησης, ενώ οι κλίσεις από τον όρο δεδομένων θα επικαλύπτονται. Μερικοί από τους τρόπους κανονικοποίησης είναι οι εξής:

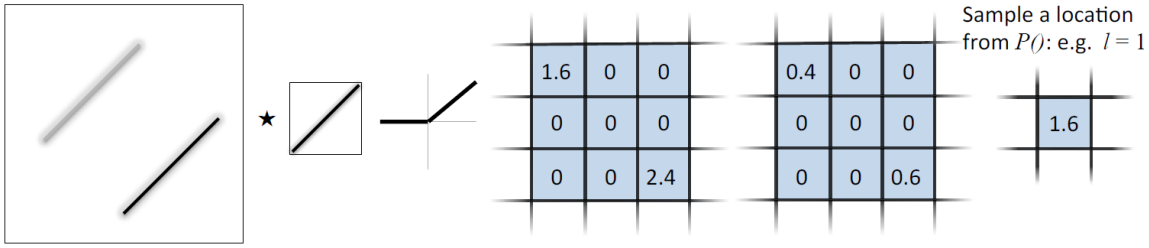
- Όροι κανονικοποίησης στη γενική συνάρτηση κόστους. Συνήθως είναι κάποια νόρμα των βαρών π.χ. L_2 ή L_1. Αυτή η κανονικοποίηση ερμηνεύεται και ως βελτιστοποίηση με περιορισμούς. Από τη Μπεϋσιανή σκοπιά αντιστοιχεί στην επιβολή εκ των προτέρων κατανομών στα βάρη του δικτύου. Για παράδειγμα, η απλή επιλογή $\|W\|^2$ ισοδυναμεί με επιβολή κανονικής κατανομής βαρών με μέση τιμή 0, τείνει να μειώσει το πλήθος των ελεύθερων παραμέτρων, οδηγεί σε απλούστερο δίκτυο και ονομάζεται εξασθένιση βαρών.

- Τυχαία αποκοπή συνδέσεων dropout. Προτάθηκε πρόσφατα στο [Sriv14] και χρησιμοποιείται με μεγάλη επιτυχία σήμερα στα ΣΝΔ και σε άλλα ΒΝΔ. Εφαρμόζεται στα πλήρως συνδεδεμένα επίπεδα και η ιδέα είναι να μεταφέρονται οι ενεργοποιήσεις των νευρώνων στο επόμενο επίπεδο, με κάποια πιθανότητα, ή αλλιώς να αποκόπτονται κάποιες συνδέσεις με τυχαίο τρόπο, μόνο κατά την εκπαίδευση.

  Το dropout μπορεί να θεωρηθεί και ως μία μέθοδος για την υλοποίηση του bagging σε μεγάλα ΝΔ. Το bagging έγκειται στην εκπαίδευση πολλών ίδιων και ασυσχέτιστων μοντέλων και συνδυασμό των τελικών αποτελεσμάτων τους, όπως μία επιτροπή. Μπορούμε να ισχυριστούμε ότι τα μοντέλα είναι ασυσχέτιστα αν εκπαιδευτούν σε διαφορετικά υποσύνολα δεδομένων. Μέσω του dropout προσεγγίζουμε την εκπαίδευση εκθετικά πολλών ΝΔ (απόδειξη στο [Beng15]), αφού σε κάθε πέρασμα εκπαιδεύεται το υποσύνολο του δικτύου στο οποίο δεν έχουν αποκοπεί οι συνδέσεις.

Το dropout ωστόσο, δεν βοηθά σε πολύ μεγάλα ή πολύ μικρά σύνολα δεδομένων. Αν έχουμε ένα πολύ μεγάλο σύνολο δεδομένων και κατ' επέκταση ένα μεγάλο δίκτυο, η κανονικοποίηση που προσφέρει συμβάλλει ελάχιστα στο σφάλμα γενίκευσης. Παρομοίως, σε πολύ μικρά επισημειωμένα σύνολα επίσης δεν βοηθάει, τότε όμως χρησιμοποιούμε και πολύ μικρότερο δίκτυο.

- Στοχαστική συγκέντρωση (stochastic pooling). Τεχνική που εισήχθη στο [Zeil13] και ακολουθεί παρόμοιο σκεπτικό με το dropout, αλλά εφαρμόζεται στα επίπεδα συγκέντρωσης, όπου επιλέγει τυχαία τις τοποθεσίες στις οποίες θα γίνει η συγκέντρωση. Πιο συγκεκριμένα, σε έναν χάρτη χαρακτηριστικών γίνεται κανονικοποίηση όλων των ενεργοποιήσεων, ώστε να αθροίζονται στο 1, και αντιμετωπίζονται ως πιθανότητες. Στη συνέχεια, ως έξοδος μιας περιοχής επιλέγεται η θέση που προκύπτει από την πολυωνυμική κατανομή με πιθανότητες αυτές που υπολογίστηκαν από την κανονικοποίηση των χαρακτηριστικών.[3] Η διαδικασία απεικονίζεται στο Σχήμα 3.9.



Σχήμα 3.9: Μία εικόνα με δύο ακμές συνελίσσεται με ένα φίλτρο και το αποτέλεσμα υφίσταται μέγιστη συγκέντρωση για να προκύψει ένας χάρτης χαρακτηριστικών. Στη συνέχεια, ο χάρτης κανονικοποιείται και ως αποτέλεσμα της στοχαστικής συγκέντρωσης στην περιοχή 3×3 λαμβάνεται το χαρακτηριστικό του οποίου η θέση προκύπτει από την πολυωνυμική κατανομή $\mathcal{P}\left(p_1, ..., p_9\right)$. Πηγή: [Zeil13].

Γενικότερα ως τρόποι κανονικοποίησης μπορούν να θεωρηθούν ([Beng15]) η προσθήκη θορύβου στα βάρη, η επαύξηση του συνόλου δεδομένων, το πρόωρο σταμάτημα της εκπαίδευσης (early stopping) και ο διαμοιρασμός βαρών (π.χ. στα συνελικτικά επίπεδα των ΣΝΔ).

3.3.6 Μάθηση δια Μεταφοράς (Transfer learning) και Ειδική Προσαρμογή (Fine-tuning)

Η εκπαίδευση των ΣΝΔ με τους σημερινούς υπολογιστικούς πόρους είναι χρονοβόρα και δαπανηρή. Λόγω των τεράστιων ποσοτήτων δεδομένων και χρόνου που χρειάζεται η μάθηση ακόμα και ενός όχι πολύ βαθέος δικτύου, συνήθως χρησιμοποιείται ένα προεκπαιδευμένο δίκτυο και η δυνατότητα της μάθησης δια μεταφοράς. Τα κύρια σενάρια χρήσης ενός προεκπαιδευμένου δικτύου είναι τα εξής:

1. *Χρήση του ΣΝΔ ως εξαγωγέα συνελικτικών χαρακτηριστικών.* Τα χαρακτηριστικά που εξάγονται από ΣΝΔ έχουν μεγάλη δύναμη αναπαράστασης και μπορούν να χρησιμοποιηθούν και

[3] Αυτό προϋποθέτει την πλήρη ανεξαρτησία των θέσεων, κάτι που δεν ισχύει πάντα. Για παράδειγμα, όσο μεγαλώνει η επικάλυψη (stride), τόσο η υπόθεση ανεξαρτησίας γίνεται ασθενέστερη.

για διαφορετικούς σκοπούς από αυτόν του δικτύου από το οποίο εξήχθησαν. Παίρνοντας για παράδειγμα το AlexNet (Σχήμα 3.13) και απομακρύνοντας το τελικό επίπεδο ταξινόμησης σε κατηγορίες, μπορούμε να τροφοδοτήσουμε το δίκτυο με μία εικόνα και μετά την εμπρόσθια διάδοση να εξάγουμε τους όγκους χαρακτηριστικών από οποιοδήποτε από τα 7 επίπεδα. Τα δύο τελευταία επίπεδα δίνουν μονοδιάστατες περιγραφές, ενώ τα πέντε προηγούμενα τριδιάστατους όγκους χαρακτηριστικών.

Η επιλογή από πιο στάδιο θα εξαχθούν τα χαρακτηριστικά στηρίζεται σε δοκιμές, διαίσθηση αλλά και στη γνώση του νέου προβλήματος στο οποίο θέλουμε να χρησιμοποιήσουμε τα χαρακτηριστικά. Αν το νέο σύνολο δεδομένων μοιάζει με το σύνολο εκπαίδευσης του δικτύου, τότε επιλέγουμε χαρακτηριστικά από υψηλότερα επίπεδα που δίνουν πιο γενικές, σημασιολογικές περιγραφές. Αν το νέο σύνολο δεν μοιάζει τόσο πολύ, καλύτερα επιλέγουμε χαμηλότερες, τοπικές περιγραφές. Και στις δύο περιπτώσεις μπορούμε να εκπαιδεύσουμε κάποιον ανεξάρτητο ταξινομητή (π.χ. γραμμικό ή SVM), ανάλογα με το πλήθος παραδειγμάτων που υπάρχουν στο νέο σύνολο δεδομένων.

2. *Ειδική Προσαρμογή του δικτύου στο νέο σύνολο δεδομένων*. Η συνηθέστερη περίπτωση είναι το νέο πρόβλημα να είναι ίδιο ή παρόμοιο με το στόχο του ΣΝΔ και να έχουμε ένα νέο σύνολο δεδομένων. Τότε χάρη στη μεταφερσιμότητα των συνελικτικών χαρακτηριστικών μπορούμε να εκπαιδεύσουμε για λίγες εποχές το δίκτυο στο νέο σύνολο δεδομένων και να επιτύχουμε πολύ καλά αποτελέσματα. Αυτό ονομάζεται fine-tuning και ουσιαστικά ρυθμίζει/προσαρμόζει τις γενικευμένες καλές δυνατότητες του δικτύου στις λεπτομέρειες του νέου συνόλου. Χρειάζεται πολύ λιγότερα δεδομένα από την αρχική εκπαίδευση, λιγότερο χρόνο μάθησης, αλλά και μείωση του ρυθμού μάθησης κατά 10-100 φορές.

   Αν ωστόσο, το νέο σύνολο δεδομένων είναι μικρό σχετικά με το μέγεθος του δικτύου, δεν πρέπει να ακολουθείται αυτή η διαδικασία, γιατί υπάρχει αυξημένος κίνδυνος υπερπροσαρμογής· εναλλακτικά μπορεί να ακολουθηθεί κάποιος τρόπος τεχνικής επαύξησης δεδομένων (βλ. Ενότητα 3.3.4). Σε αυτή την περίπτωση είναι καλύτερο να χρησιμοποιούνται μόνο τα συνελικτικά χαρακτηριστικά από το δίκτυο και να εκπαιδεύεται ένας εξωτερικός ταξινομητής. Αν το νέο σύνολο είναι αρκετά μεγάλο, μπορούμε να ακολουθήσουμε την τεχνική της ειδικευμένης προσαρμογής, είτε να εκπαιδεύσουμε το δίκτυο εξ' αρχής, όμως ακόμα και τότε η αρχικοποίηση των βαρών από ένα ήδη εκπαιδευμένο δίκτυο βοηθάει στο χρόνο και στην αποδοτικότητα της μάθησης.

Μεταφερσιμότητα συνελικτικών χαρακτηριστικών και Ειδική Προσαρμογή

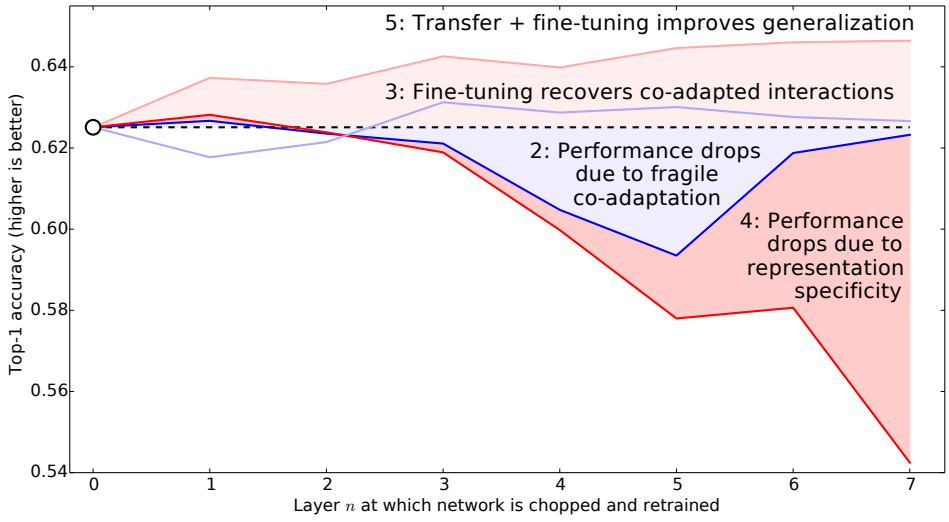
Στο [Yosi14] παρουσιάζεται μία σημαντική μελέτη των εξαγόμενων χαρακτηριστικών από ΣΝΔ, του fine-tuning και της σχέσης τους· τα βασικά αποτελέσματα φαίνονται στο Σχήμα 3.10. Το κεντρικό πείραμα περιλαμβάνει δύο δίκτυα Alexnet A και B που είναι εκπαιδευμένα σε δύο ξένα υποσύνολα του ImageNet, έστω A και B. Στις καμπύλες 2 και 3 εξετάζονται οι δυνατότητες των συνελικτικών χαρακτηριστικών χρησιμοποιώντας μόνο το δίκτυο B. Στις καμπύλες 4 και 5 εξετάζονται οι δυνατότητες της ειδικής προσαρμογής χρησιμοποιώντας ένα υβριδικό δίκτυο, όπου το πρώτο μέρος του

περιλαμβάνει επίπεδα από το δίκτυο A, ενώ το δεύτερο μέρος επίπεδα από το B. Στις περιπτώσεις 2 και 4 διατηρούνται σταθερά τα βάρη μέχρι κάποιο από τα 7 επίπεδα και τα δίκτυα εκπαιδεύονται από εκεί και πέρα στα σύνολα δεδομένων. Στις περιπτώσεις 3 και 5 επιλέγονται τα βάρη μέχρι κάποιο επίπεδο από κάποιο από τα δίκτυα A ή B, αλλά κατά την εκπαίδευση επιτρέπεται η μάθηση όλων των επιπέδων.

Έστω η καμπύλη 2 που αντιστοιχεί στην περίπτωση BnB, δηλαδή επιλέγονται n επίπεδα από το δίκτυο B και κρατούνται σταθερά, και εκπαιδεύονται τα υπόλοιπα $7 - n$ επίπεδα. Παρατηρούμε ότι τα δίκτυα B3B έως και B6B έχουν μικρότερη επιτυχία από το δίκτυο αναφοράς B. Αυτό συμβαίνει γιατί τα επίπεδα 3-6 εξάγουν εξαρτώμενα, συν-προσαρμοζόμενα (co-adaptive) χαρακτηριστικά και αν κοπεί αυτή η αλυσίδα τα τελικά επίπεδα δεν μπορούν να εξάγουν συμπεράσματα.

Για την καμπύλη 4 χρησιμοποιείται το μοντέλο AnB (fine-tuning), όπου επιλέγονται n επίπεδα από το δίκτυο A (κρατούνται σταθερά) και $7 - n$ επίπεδα από το δίκτυο B, στα οποία τους δίνεται η δυνατότητα μάθησης. Η αποτελεσματικότητα της αναπαράστασης μειώνεται συνεχώς λόγω της εξειδίκευσης του πρώτου μέρους του δικτύου στο σύνολο A.

Από τις καμπύλες 3 και 5 που εξετάζουν τα μοντέλα BnB και AnB αντίστοιχα, με επίτρεψη μάθησης σε όλα τα επίπεδα, βλέπουμε ότι σε όλες τις περιπτώσεις η επίδοση των νέων δικτύων είναι καλύτερη από το δίκτυο αναφοράς B.



Σχήμα 3.10: Μεταφερσιμότητα χαρακτηριστικών και σχέση της με την ειδική προσαρμογή συναρτήσει των επιπέδων. Πηγή: [Yosi14].

3.3.7 Βελτιστοποίηση Συναρτήσεων Κόστους

Στην Ενότητα 2.4.4 εξηγήθηκε πώς υπολογίζονται αποδοτικά οι μερικές παράγωγοι μίας σύνθετης συνάρτησης. Η τεχνική της οπισθοδιάδοσης (backpropagation), προσφέρει ένα γρήγορο τρόπο υπολογισμού της κλίσης της συνάρτησης κόστους ως προς τις παραμέτρους του δικτύου, αλλά δεν περιγράφει καθόλου πώς θα βρεθούν αυτοί οι βέλτιστοι παράμετροι[4]. Αυτό είναι ο στόχος των αλ-

[4] Στη βιβλιογραφία πολλές φορές ο όρος backpropagation περιλαμβάνει τόσο τον τρόπο υπολογισμού κλίσεων, όσο και τον αλγόριθμο εκπαίδευσης, εδώ θα γίνει διαχωρισμός των δύο εννοιών.

γορίθμων βελτιστοποίησης. Οι τελικές συναρτήσεις κόστους των ΣΝΔ είναι εξαιρετικά μη κυρτές συναρτήσεις.

Σκοπός της εκπαίδευσης ενός δικτύου είναι η ελαχιστοποίηση του επιλεγμένου κόστους, που επιτυγχάνεται έμμεσα με τη βελτιστοποίηση των παραμέτρων του δικτύου. Το κόστος περιγράφεται από την 2.6, η οποία επαναλαμβάνεται εδώ για την επισήμανση της εξάρτησης του κόστους από τα βάρη:

$$L(W) \approx \frac{1}{|\mathsf{I}|} \sum_{i \in \mathsf{I}} L_i(W) + \lambda R(W) \tag{3.1}$$

Το συνολικό πλήθος των παραδειγμάτων του συνόλου δεδομένων για τα ΒΝΔ στην πράξη είναι πολύ μεγάλο, επομένως η άθροιση κοστών για κάθε εικόνα πρέπει να γίνει σε ένα υποσύνολό του, που ονομάζεται πακέτο εκπαίδευσης (mini-batch). Το $\mathsf{I} \subseteq \mathsf{D}$ είναι το υποσύνολο των παραδειγμάτων εκπαίδευσης που επιλέγονται (συνήθως τυχαία) από το σύνολο δεδομένων για να γίνει η ανανέωση βαρών: $\mathsf{I} \subseteq \{1, ..., N\}$. Στην πράξη το πακέτο περιλαμβάνει μέχρι μερικές εκατοντάδες εικόνες (32 - 256), κυρίως λόγω περιορισμών μνήμης.

Επομένως, ως προς το πλήθος παραδειγμάτων που χρησιμοποιούν σε κάθε επανάληψη οι αλγόριθμοι μπορούν να κατηγοριοποιηθούν σε [Beng15]:

- Αλγόριθμοι εκπαίδευσης συνόλου (Batch algorithms). Σε κάθε επανάληψη χρησιμοποιούνται όλα τα παραδείγματα του συνόλου δεδομένων.
- Αλγόριθμοι στοχαστικής εκπαίδευσης (Stochastic or Online algorithms). Σε κάθε επανάληψη χρησιμοποιείται μόνο ένα παράδειγμα εκπαίδευσης για την ανανέωση βαρών.

Στην πράξη οι αλγόριθμοι εφαρμόζουν μία μέση λύση ανάμεσα στις δύο, καθώς η πρώτη μέθοδος αν και καλύτερη χρειάζεται πολύ μνήμη και υπολογιστική ισχύ, ενώ η δεύτερη κάνει στατιστικές υποθέσεις για τα παραδείγματα και τη σειρά τροφοδότησής τους. Οι αλγόριθμοι αυτοί ονομάζονται στοχαστικοί με χρήση πακέτου εκπαίδευσης (mini-batch) (και εφεξής θα αναφέρονται απλά ως στοχαστικοί) και επιβάλλουν περιορισμούς στις ιδιότητες των εικόνων που θα περιλαμβάνει κάθε πακέτο. Για παράδειγμα, στο πρόβλημα της ταξινόμησης, ένα πακέτο δεν θα πρέπει να περιλαμβάνει εικόνες μόνο από μία κατηγορία ή εικόνες που έχουν μεγάλη συσχέτιση μεταξύ τους, γιατί η εκπαίδευση θα είναι μονομερής ή ελλιπής.

Σε ότι αφορά στην τάξη παραγώγων που χρησιμοποιούν, οι αλγόριθμοι μπορούν να κατηγοριοποιηθούν σε:

- 1ης τάξης. Χρησιμοποιούν για την ανανέωση των βαρών μόνο τις πρώτες μερικές παραγώγους της L. Αυτοί οι αλγόριθμοι είναι και οι πιο συνήθεις χρησιμοποιούμενοι, είτε γιατί οι δεύτερες παράγωγοι είναι δύσκολες να υπολογιστούν, είτε δεν υπάρχουν.
- 2ης τάξης. Προσφέρουν σε πολλές περιπτώσεις[5] γρηγορότερη σύγκλιση στο ελάχιστο, ωστόσο χρειάζονται τον υπολογισμό των δευτέρων παραγώγων, είναι πιο χρονοβόροι και μπορεί να είναι ασταθείς. Επίσης, μπορεί να χρησιμοποιούν και τις παραγώγους 1ης τάξης.

[5] Μερικοί αλγόριθμοι δεύτερης τάξης μπορούν να αποτύχουν κοντά σε ένα σημείο σέλης, ενώ κάποιοι 1ης τάξης μπορούν (θεωρητικά) να ξεφύγουν (βλ. Σχήμα 3.12).

Ο τελευταίος διαχωρισμός σχετίζεται με το μήκος βήματος των αλγορίθμων ή όπως αναφέρεται στο πεδίο των ΝΔ, ο ρυθμός μάθησης. Το βήμα αποτελεί μία πολύ σημαντική και ευαίσθητη υπερπαράμετρο του δικτύου και πρέπει να επιλέγεται με προσοχή. Για το λόγο αυτό αναπτύχθηκαν αλγόριθμοι που μπορούν να καθορίζουν το βήμα εκπαίδευσης δυναμικά, σε κάθε επανάληψη, με ευρετικές ή αναλυτικές μεθόδους. Έτσι οι αλγόριθμοι μπορούν να διαχωριστούν σε:

- Σταθερού/προκαθορισμένου ρυθμού μάθησης, αρκετά μικρού, ώστε να μην δημιουργούνται ταλαντώσεις και άλματα στη βελτιστοποίηση, αλλά και αρκετά μεγάλου, ώστε να είναι γρήγορη.
- Προσαρμοζόμενου ρυθμού μάθησης. Υπάρχει ένα μικρό υποσύνολο των παραμέτρων ενός δικτύου, που επιβάλλει πολύ μικρές μεταβολές στο βήμα, γιατί η ευαισθησία της συνάρτησης κόστους ως προς αυτές είναι πολύ μεγάλη (μεγάλες κλίσεις). Σε αυτές τις περιπτώσεις το βήμα πρέπει να επιλέγεται πολύ μικρό για να τις ικανοποιήσει. Αν σε κάποιο στάδιο η βελτιστοποίηση δεν γίνεται ως προς το ευαίσθητο υποσύνολο, τότε μπορεί να επιταχυνθεί και αυτό επιτυγχάνεται με τον προσαρμοζόμενο ρυθμό μάθησης.

Στη συνέχεια, παρουσιάζονται 6 δημοφιλείς στοχαστικοί αλγόριθμοι 1ης τάξης για τη βελτιστοποίηση της $L(W)$ ως προς τα βάρη W [Beng15, Karp15]. Στα Σχήματα 3.11, 3.12 φαίνονται στιγμιότυπα από τη βελτιστοποίηση για 6 από τους κυρίως χρησιμοποιούμενους αλγορίθμους.

Στοχαστική Κατάβαση Δυναμικού (Stochastic Gradient Descent – SGD)

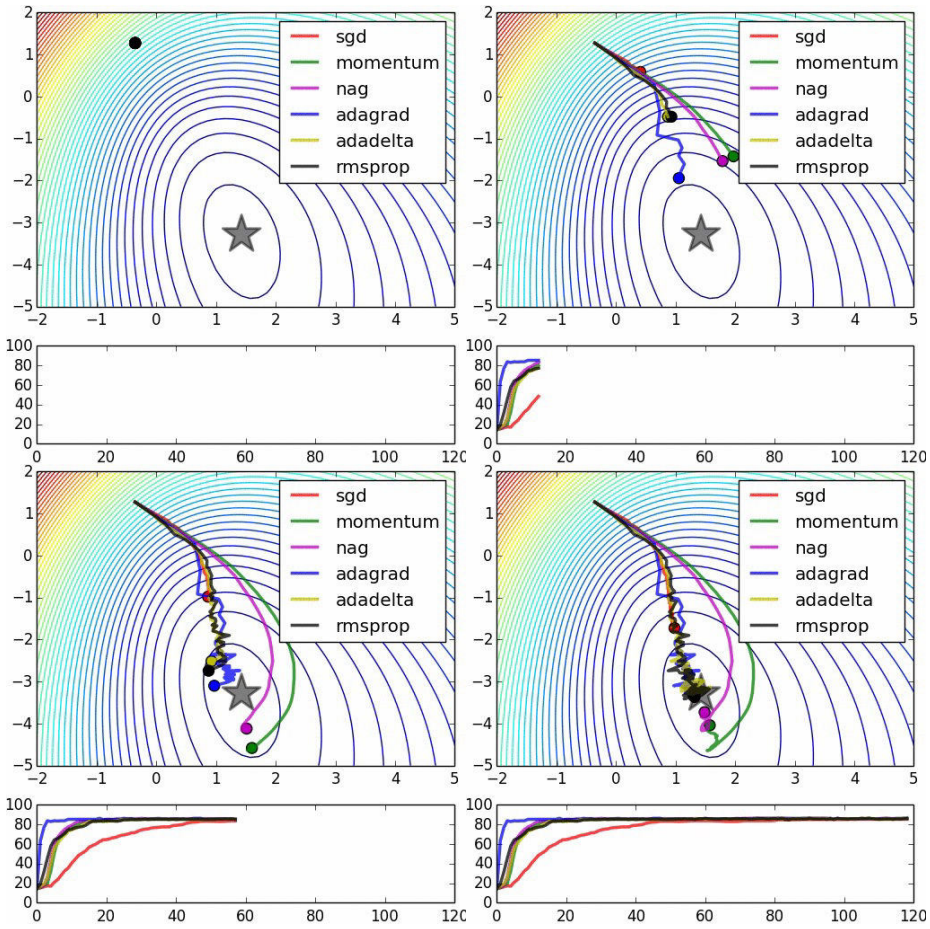
Μέθοδος ελαχιστοποίησης 1ης τάξης, σύμφωνα με την οποία για να βρεθεί το ελάχιστο συνάρτησης κινούμαστε με βήματα στην κατεύθυνση της αρνητική κλίσης. Η ανανέωση βαρών γίνεται από την 3.2. Χρειάζεται να παρέχεται εξαρχής ο ρυθμός μάθησης η, που διατηρείται σταθερός καθ' όλη τη διαδικασία. Στα Σχήματα 3.11, 3.12 φαίνονται οι εφαρμογές του SGD σε διάφορες συναρτήσεις και παραθέτονται συγκριτικά σχόλια.

Σε ένα γενικό περιβάλλον, η «στοχαστικότητα» αναφέρεται στη χρήση μία στοχαστικής (θορυβώδους) εκτίμησης της κλίσης στην ανανέωση. Στην περίπτωση των ΝΔ αυτό επιτυγχάνεται χρησιμοποιώντας ένα πακέτο εκπαίδευσης του συνόλου δεδομένων. Έτσι όταν όλα τα παραδείγματα του πακέτου, θεωρούμενα ως πολυδιάστατες τυχαίες μεταβλητές, είναι ανεξάρτητα και ισόνομα, τότε υπολογίζοντας τη μέση κλίση έχουμε έναν αμερόληπτο εκτιμητή της πραγματικής κλίσης[6].

$$W_{t+1} = W_t + \eta \nabla L(W_t) \tag{3.2}$$

Πολλές φορές στην πράξη αντί για σταθερό ρυθμό επιλέγεται ένα σχήμα μειουμένου ρυθμού μάθησης, συνήθως με μερικά σταθερά επίπεδα. Αυτό είναι υποχρεωτικό καθώς η τυχαία δειγματοληψία των πακέτων οδηγεί στο γεγονός η κλίση να μην είναι πολύ κοντά στο 0 σε ένα τοπικό ελάχιστο, με αποτέλεσμα ο αλγόριθμος να ταλαντώνεται γύρω από το ελάχιστο.

[6] Από αυτή την ανάλυση φαίνεται ότι τα παραδείγματα κάθε πακέτου πρέπει να είναι πολύ προσεκτικά επιλεγμένα.



Σχήμα 3.11: Σε αυτή την απλή κυρτή συνάρτηση παρατηρείται ο SGD να ακολουθεί σταθερή και σωστή πορεία, αλλά να είναι αργός. Αντίθετα, οι γρηγορότεροι προσαρμοζόμενου ρυθμού μάθησης AdaGrad, AdaDelta, RMSprop εμφανίζουν σοβαρές ταλαντώσεις, ωστόσο ακολουθούν γενικά σωστή πορεία προς το ελάχιστο. Τέλος, οι SGD+momentum, NAG λόγω της κατασκευής τους προσεγγίζουν το ελάχιστο περιφερειακά, αλλά μόλις βρεθούν σε άξονα (ιδιοτιμή) του παραβολοειδούς επιταχύνουν. Στο διάγραμμα χρόνου φαίνεται ότι ο AdaGrad είναι ο γρηγορότερος στην προσέγγιση του ελαχίστου, και ο SGD ο πιο αργός. Πηγή: Στιγμιότυπα από animation του Alec Radford (`https://twitter.com/AlecRad`).

Στοχαστική Κατάβαση Δυναμικού με παράγοντα ορμής (SGD with momentum)

Γενικεύει τον προηγούμενο αλγόριθμο, ώστε να έχει καλύτερη συμπεριφορά σε συγκεκριμένες περιπτώσεις, όπως σε συνθήκες συνεχόμενης μικρής τιμής κλίσης, με το αντίτιμο της προσθήκης μίας επιπλέον υπερπαραμέτρου.[7] Το σχήμα ανανέωσης των βαρών δίνεται από τις εξισώσεις 3.3 και στοχεύει να προσδώσει στον αλγόριθμο την ιδιότητα της φυσικής επιτάχυνσης που οφείλεται στη ροπή ενός αντικειμένου (π.χ. όταν κατέρχεται ένα κεκλιμένο επίπεδο μικρής κλίσης). Πιο συγκεκριμένα, η ανανέωση των βαρών γίνεται με έναν γραμμικό συνδυασμό του αρνητικού της κλίσης $-\nabla_W L\left(W_t\right)$

[7] Αποτελεί γενίκευση (ως προς τη συνάρτηση κόστους) του παλαιότερου αλγορίθμου ADALINE, που εξάγεται ισοδύναμα με τον LMS για συνάρτηση κόστους το Μέσο Τετραγωνικό Σφάλμα (MSE).

και της τιμής της προηγούμενης ανανέωσης V_t. Η υπερπαράμετρος η (βάρος κλίσης του γραμμικού συνδυασμού) είναι ο γνωστός ρυθμός μάθησης και η μ (βάρος προηγούμενης ανανέωσης) αντιπροσωπεύει την ορμή. Η ενδιάμεση μεταβλητή V_t έχει το ρόλο της ταχύτητας και συσσωρεύει την κλίση. Στα Σχήματα 3.11, 3.12 φαίνονται οι εφαρμογές του SGD with momentum και συγκρίσεις με άλλους αλγορίθμους.

$$\begin{aligned} W_{t+1} &= W_t + V_{t+1} \\ V_{t+1} &= \mu V_t - \eta \nabla L\left(W_t\right) \end{aligned} \tag{3.3}$$

Όμοια με τον προηγούμενο αλγόριθμο, άμεση επέκταση αυτής της μεθόδου είναι η προσθήκη χρονοδιαγράμματος, πιο συνηθισμένα, στο ρυθμό μάθησης και λιγότερο στην ορμή. Ο ρυθμός μάθησης γίνεται συνάρτηση των επαναλήψεων η_t. Ένα συχνά χρησιμοποιούμενο χρονοδιάγραμμα είναι να μειώνεται και να παραμένει σταθερός για 3-4 διαστήματα κατά τη διάρκεια της εκπαίδευσης. Το σκεπτικό πίσω από αυτή την επιλογή είναι αρχικά το ΝΔ να μαθαίνει γρήγορα, ενώ προς το τέλος, που θα έχει βρεθεί η περιοχή ενός τοπικού ελαχίστου της συνάρτησης κόστους, να γίνονται μικρές προσαρμογές.

Επιταχυνόμενη Κατάβαση Δυναμικού (Nesterov's Accelerated Gradient – NAG)

Η μέθοδος αυτή είναι παρόμοια με την προηγούμενη με μοναδική διαφορά το σημείο υπολογισμού της κλίσης σε κάθε επανάληψη. Όπως φαίνεται στις 3.4, η κλίση υπολογίζεται αφού έχει εφαρμοστεί η μετατόπιση λόγω της συσσωρευμένης ταχύτητας. Αυτός ο βελτιωτικός παράγοντας επιταχύνει αρκετά τη σύγκλιση, όπως φαίνεται στα Σχήματα 3.11, 3.12.

$$\begin{aligned} W_{t+1} &= W_t + V_{t+1} \\ V_{t+1} &= \mu V_t - \eta \nabla L\left(W_t + \mu V_t\right) \end{aligned} \tag{3.4}$$

Προσαρμοστική Κατάβαση Δυναμικού (Adaptive Gradient Descent – AdaGrad)

Ο αλγόριθμος AdaGrad είναι 1ης τάξης με προσαρμοζόμενο ρυθμό μάθησης. Απαιτεί την αποθήκευση πληροφορίας για προηγούμενες τιμές βαρών και προσαρμόζει το ρυθμό μάθησης κάθε βάρους ξεχωριστά. Το i στοιχείο των βαρών ανανεώνεται σύμφωνα με την 3.5, χρησιμοποιώντας τις ανανεώσεις βαρών από όλες τις προηγούμενες επαναλήψεις $\nabla L\left(W_\tau\right),\ \tau \in \{1, ..., t\}$. Οι συντελεστές μάθησης στη συγκεκριμένη περίπτωση είναι συνάρτηση του ιστορικού ανανέωσης των βαρών.

Έχει παρατηρηθεί ότι στα ΝΔ η συσσώρευση των τετραγώνων των κλίσεων από την αρχή της βελτιστοποίησης οδηγεί σε αργότερη σύγκλιση και η ποιότητα αποτελεσμάτων δεν απέχει πολύ από αυτά των προηγούμενων αλγορίθμων (Σχήματα 3.11, 3.12), αν και η επίδοσή του σε άλλους τομείς είναι βελτιωμένη.

$$[W_{t+1}]_i = [W_t]_i - \frac{\eta}{\sqrt{\sum_{\tau=1}^{t} [\nabla L(W_\tau)]_i^2}} [\nabla L(W_t)]_i \tag{3.5}$$

RMSprop, AdaDelta

Και οι δύο αλγόριθμοι προσπαθούν να βελτιώσουν ελαττώματα του AdaGrad και συγκρίνονται με τους προηγούμενους στα Σχήματα 3.11, 3.12. Ο μεν πρώτος αντικαθιστά την τετραγωνική συσσώρευση όλων των κλίσεων με έναν εκθετικά σταθμισμένο κινούμενο μέσο όρο, εισάγοντας μία επιπλέον υπερπαράμετρο, τον ρυθμό απόσβεσης. Ο δεύτερος απαιτεί δύο υπερπαραμέτρους και προσαρμόζει το ρυθμό μάθησης κάθε βάρους χρησιμοποιώντας πληροφορία από δεύτερες μερικές παραγώγους.

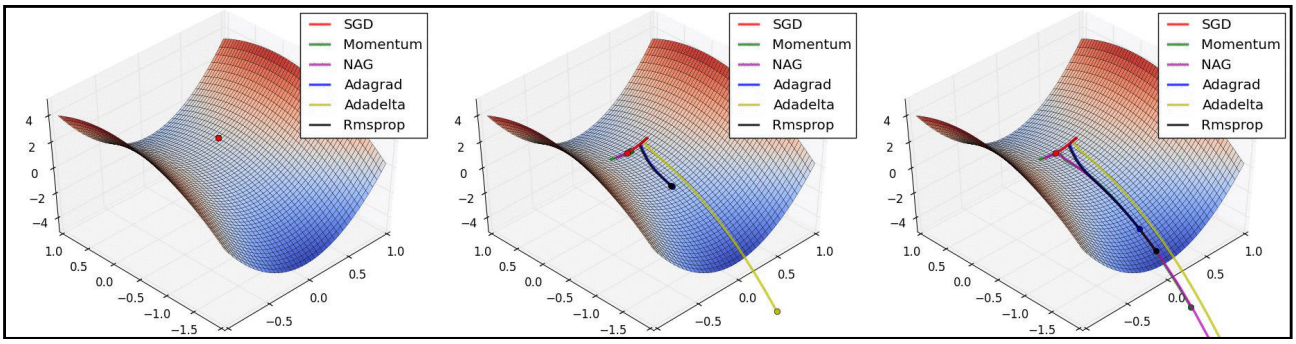
3.3.8 Αρχικοποίηση Βαρών

Η εκπαίδευση των ΝΔ γίνεται σύμφωνα με ένα επαναληπτικό σχήμα, όπως αυτά που αναφέρθηκαν στην Ενότητα 2.4.4. Κάθε τέτοιο σχήμα ξεκινάει από μία αρχική εκτίμηση των μεταβλητών προς βελτιστοποίηση, η οποία πρέπει να βρίσκεται αρκετά κοντά σε κάποιο τοπικό ή καλύτερα και γενικό ελάχιστο της συνάρτησης κόστους, ώστε η διαδικασία της βελτιστοποίησης να είναι γρήγορη και αποτελεσματική. Λόγω αυτού συμπεραίνουμε ότι η αρχικοποίηση καθορίζει σε σημαντικό βαθμό το αποτέλεσμα της μάθησης.

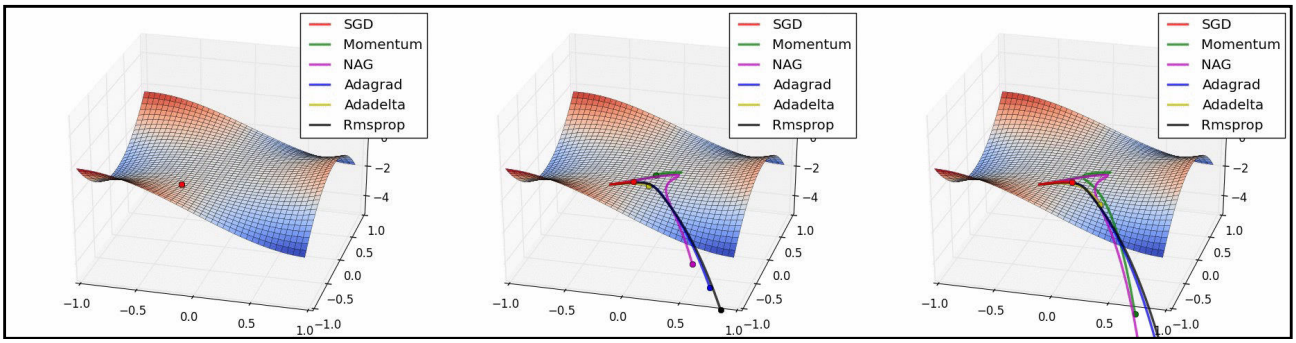
Μια πρώτη λογική προσέγγιση θα ήταν να αρχικοποιήσουμε όλα τα βάρη στο 0, δεδομένου της εφαρμογής κανονικοποίησης στα δεδομένα (αφαίρεση μέση τιμής κτλ.), και αναμένοντας το πλήθος των αρνητικών και θετικών βαρών να είναι περίπου ίσο, ενώ να αθροίζονται ιδανικά στο 0. Αυτή η επιλογή αποδεικνύεται ότι είναι η χειρότερη, καθώς επιβάλλοντας αυτή την απόλυτη συμμετρία στο δίκτυο, κάθε νευρώνας έχει την ίδια έξοδο, υπολογίζει τις ίδιες κλίσεις και τελικά το δίκτυο δεν μαθαίνει.

Η επόμενη επιλογή είναι η τυχαία αρχικοποίηση με πολύ μικρούς αριθμούς, π.χ. από μία πολυδιάστατη gaussian κατανομή. Οι μικροί, τυχαίοι αριθμοί ικανοποιούν τις παραπάνω υποθέσεις, ωστόσο αν το μέτρο τους είναι πολύ μικρό, τότε κατά την οπισθοδιάδοση οι κλίσεις που θα μεταδίδονται θα είναι επίσης πολύ μικρές οδηγώντας το δίκτυο να μαθαίνει αργά. Επιπλέον, με αυτή την επιλογή η διασπορά των τιμών της εξόδου ενός νευρώνα (πριν την ενεργοποίηση) εξαρτάται από το πλήθος των εισόδων του. Αυτό φαίνεται στην 3.6 αν θεωρήσουμε ότι $w_i,\ x_i$ ανεξάρτητες, ισόνομες τυχαίες μεταβλητές για κάθε i, με $E[w_i] = E[x_i] = 0$. Σημειώνουμε ότι αυτή η απλοποιημένη ανάλυση δεν ισχύει σε πολλές περιπτώσεις, όπως για ενεργοποίηση τύπου ReLU, όπου η μέση τιμή είναι σίγουρα θετική.

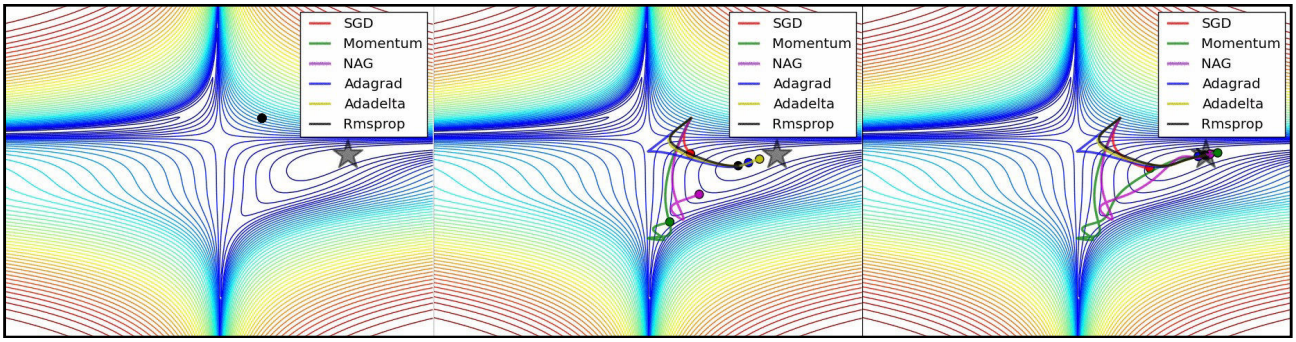
$$Var\left[\mathbf{w}^T\mathbf{x}\right] = \sum_{i=1}^{n} Var[w_i]\,Var[x_i] = \left(nVar[\mathbf{w}]\right)Var[\mathbf{x}] \tag{3.6}$$



(a) Η συμμετρία της συνάρτησης κόστους δυσκολεύει τους μη προσαρμοστικούς αλγορίθμους. Ο SGD αφού ταλαντώνεται αρχικά, παγιδεύεται. Οι SGD+momentum, NAG, ενώ στην αρχή ταλαντώνονται καταφέρνουν να αναπτύξουν ταχύτητα στην σωστή κατεύθυνση και μάλιστα ο NAG, μόλις τη βρίσκει επιταχύνει πολύ γρήγορα, ξεπερνώντας και τους προσαρμοζόμενους. Από τους αλγόριθμους με προσαρμοζόμενο ρυθμό μάθησης ο AdaDelta είναι ο αποδοτικότερος (καθώς χρησιμοποιεί και πληροφορία δευτέρων παραγώγων), οι RMSprop, AdaGrad ξεκινούν πολύ καλά και δεν εγκλωβίζονται, αλλά συνεχίζουν με μικρή ταχύτητα.



(b) Στην περίπτωση σημείου σέλης ο SGD μένει πάλι στάσιμος, το ζεύγος SGD+momentum, NAG πάει αρκετά καλά, από διαφορετική όμως διαδρομή, αλλά ο AdaDelta σε αυτή την περίπτωση είναι αρκετά αργός.



(c) Στη συνάρτηση του Beale οι SGD+momentum, NAG, AdaGrad μπερδεύονται αρχικά, λόγω της μεγάλης κλίσης δημιουργείται ταχύτητα προς τη λάθος κατεύθυνση. Οι AdaDelta, RMSprop ακολουθούν σωστή πορεία και χειρίζονται τις μεγάλες αρχικές κλίσεις σωστά. Έκπληξη είναι ο SGD, που λόγω της απλότητάς του οδεύει αργά, αλλά σταθερά προς το ελάχιστο.

Σχήμα 3.12: Συγκρίσεις αλγορίθμων Βελτιστοποίησης σε διάφορες συναρτήσεις για εύρεση ελαχίστου. Πηγή: Στιγμιότυπα από animation του Alec Radford (`https://twitter.com/AlecRad`).

Για να είναι λοιπόν η διασπορά της εξόδου ανεξάρτητη από τον αριθμό των εισόδων και ίση με τη διασπορά των εισόδων πρέπει τα w_i να επιλεγούν από την $\mathcal{N}(0,1)$ και μετά να διαιρεθούν με την τιμή $\sqrt{1/n}$, ώστε να έχουν τελικά διασπορά $1/n$ (σύμφωνα με την ιδιότητα $Var[\alpha \cdot] = \alpha^2 Var[\cdot]$). Από μία αντίστοιχη, ενδιαφέρουσα ανάλυση στο [Glor10] οι συγγραφείς καταλήγουν ότι καλύτερα αποτελέσματα δίνει η επιλογή

$$Var[\mathbf{w}] = \frac{2}{n_{in} + n_{out}}$$

όπου n_{in} και n_{out} το πλήθος των εισόδων και εξόδων αντίστοιχα και με το πλήθος των εξόδων να εννοείται το πλήθος των συνδέσεων που τροφοδοτεί ο νευρώνας. Τέλος, στο [He15] από μία ανάλυση συγκεκριμένα για νευρώνες με ReLU ενεργοποίηση, εξάγεται ότι καλύτερη τιμή της διασποράς των βαρών είναι η $2/n_{out}$, που χρησιμοποιείται συνήθως σήμερα. Υπάρχουν πολλές αναλύσεις σαν τις προαναφερόμενες, ωστόσο πολλές φορές στην πράξη θα πρέπει να προσαρμόζονται στο πρόβλημα μετά από πειραματισμό.

Στο [Ioff15] οι συγγραφείς ξεπερνούν τη μεγάλη ευαισθησία στην αρχικοποίηση εισάγοντας την τεχνική κανονικοποίησης πακέτων εκπαίδευσης και χαρτών χαρακτηριστικών (batch normalization), η οποία βελτιώνει τα στατιστικά χαρακτηριστικά της εισόδου κάθε επιπέδου.

3.3.9 Πρακτικά ζητήματα

- Διαφορισιμότητα συναρτήσεων επιπέδων. Για να χρησιμοποιηθεί ένας αλγόριθμος κατάβασης δυναμικού, ακόμα και 1ης τάξης, η σύνθετη συνάρτηση κόστους πρέπει να είναι παραγωγίσιμη. Στα ΣΝΔ (αλλά και στη Μηχανική Μάθηση γενικότερα) η προϋπόθεση αυτή πολύ συχνά δεν ισχύει. Τότε πρέπει να χρησιμοποιηθούν γεικευμένες παράγωγοι (π.χ. ασθενείς) όπως αναφέρθηκε στην Ενότητα 2.4.4.

- Σε πολλά σύνολα δεδομένων όλες οι κατηγορίες δεν έχουν τον ίδιο αριθμό παραδειγμάτων (class imbalance). Αυτό επηρεάζει αρνητικά τη μάθηση καθώς σε ένα πακέτο εκπαίδευσης μπορεί να υπάρχουν πολύ λίγα έως καθόλου παραδείγματα για κάποια κατηγορία και το ΣΝΔ να μην μπορεί να μάθει να τη διακρίνει ικανοποιητικά. Ενεργώντας μόνο στο σύνολο δεδομένων, κάποιες από τις επιλογές που έχουμε είναι να αγνοήσουμε κάποιο πλήθος παραδειγμάτων από υπεράριθμες κατηγορίες, ή να εφαρμόσουμε τεχνικές επαύξησης στις λειψές κατηγορίες. Αντί ή παράλληλα αυτών μπορούμε να τροποποιήσουμε και τα άλλα στοιχεία της Μηχανικής Μάθησης.

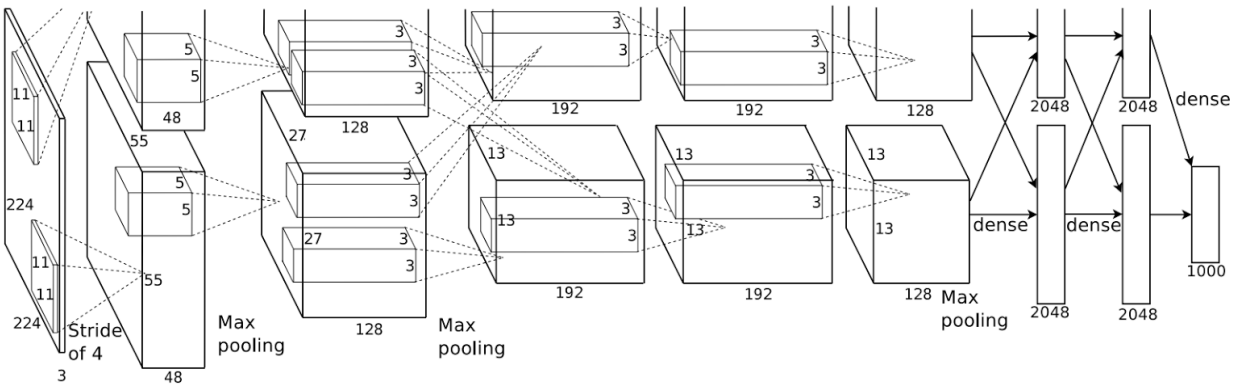
- Υπάρχουν σύνολα δεδομένων που έχουν πολλές κατηγορίες αντικειμένων και λίγα παραδείγματα σε κάθε κατηγορία, κάτι που δυσχεραίνει τη μάθηση στα ΒΝΔ. Σε αυτές τις περιπτώσεις μπορεί να εφαρμοστεί ιεραρχική ταξινόμηση, δηλαδή να οργανωθούν οι κατηγορίες σε ομάδες και για ταξινομητής του ΣΝΔ να επιλεγεί ο hierarchical softmax. Σε αυτή την περίπτωση, δημιουργείται ένα δυαδικό δένδρο ετικετών και σε κάθε κόμβο του δέντρου εκπαιδεύεται ένας δυαδικός softmax ταξινομητής, που διαχωρίζει τα δύο κλαδιά. Η ταξινόμηση εξαρτάται σε μεγάλο βαθμό από την οργάνωση του δέντρου ετικετών.

3.4 Σύγχρονα ΣΝΔ

Μετά την επιτυχία του δικτύου LeNet [LeCu98] το 1998, το 2012 εμφανίστηκε το δίκτυο AlexNet [Kriz12] που αποτέλεσε ορόσημο στην ανάπτυξη των ΣΝΔ με αντικείμενο επεξεργασίας τις εικόνες. Από τότε έχουν διερευνηθεί και βελτιωθεί αρκετά ζητήματα εκπαίδευσης και απόδοσης, με τις σημαντικότερες αλλαγές να έχουν γίνει στη δομή του και συγκεκριμένα στο βάθος του. Παρατηρείται ότι όσο το βάθος ενός ΝΔ αυξάνεται τόσο καλύτερη είναι η επίδοσή του, π.χ. το νεότερο δίκτυο GoogLeNet. Σε αυτή την Ενότητα αναλύεται διεξοδικά το AlexNet, το οποίο και χρησιμοποιείται στο Κεφάλαιο 5 ως βασικό μέρος μίας εφαρμογής τον εντοπισμού προσώπων. Επίσης, αναφέρονται τα σημαντικότερα χαρακτηριστικά νεότερων δικτύων, που τους δίνουν τη δυνατότητα να καλύτερων επιδόσεων.

3.4.1 Ανάλυση ολοκληρωμένης αρχιτεκτονικής AlexNet (2012)

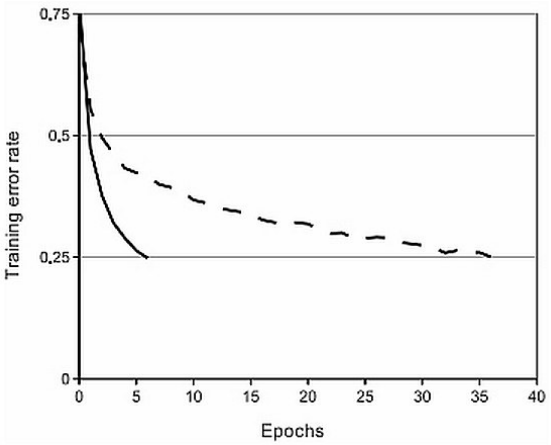
Σε αυτή την ενότητα θα αναλυθεί το ΣΝΔ AlexNet [Kriz12], το οποίο εκπαιδεύτηκε με 1.2 εκατομμύρια υψηλής ευκρίνειας εικόνες από το σύνολο δεδομένων του διαγωνισμού ταξινόμησης ImageNet, για την κατηγοριοποίηση εικόνων σε 1000 κατηγορίες. Το δίκτυο έχει 8 επίπεδα (5 εξαγωγής χαρακτηριστικών και 3 πλήρως συνδεδεμένα για ταξινόμηση), 650,000 νευρώνες και 60,000,000 βάρη νευρώνων. Το 2012 στο διαγωνισμό ImageNet LSVRC-2012 challenge πέτυχε top-5 σφάλμα ταξινόμησης 15.3%, με μεγάλη διαφορά από τα συνδιαγωνιζόμενα συστήματα (στην πλειονότητά τους ΝΔ), ενώ το δεύτερο καλύτερο υποψήφιο σύστημα πέτυχε 26.2% για ένα πρόβλημα το οποίο επί χρόνια τα συστήματα έδιναν μεγάλα ποσοστά λάθους. Λόγω των περιορισμένων υπολογιστικών πόρων (κυρίως σε μνήμη GPU) που ήταν διαθέσιμοι το 2012, αυτό το ΣΝΔ αναπτύχθηκε σε δύο κάρτες γραφικών, κάτι που απαιτούσε μηχανισμούς επικοινωνίας μεταξύ τους. Εδώ θα αναλυθεί σαν ένα δίκτυο, που θεωρείται γενικά ισοδύναμο. Η μόνη διαφορά είναι ότι οι δύο κάρτες γραφικών του AlexNet επικοινωνούν μόνο σε συγκεκριμένα επίπεδα, αλλά κατά τα άλλα η αρχιτεκτονική είναι ίδια. Η ενιαία υλοποίηση ονομάζεται και CaffeNet και έχει κάποιες μικρές διαφορές, με σημαντικότερη την τοποθέτηση των επιπέδων κανονικοποίησης μετά τα επίπεδα συγκέντρωσης.



Σχήμα 3.13: Η αρχιτεκτονική CNN AlexNet. Στο Σχήμα απεικονίζονται ο όγκος εισόδου του δικτύου (εικόνα $227 \times 227 \times 3$) και οι όγκοι εξόδου των 8 επιπέδων. Πηγή [Kriz12].

Τα βασικά στοιχεία της αρχιτεκτονικής και οι υπερπαράμετροι του δικτύου φαίνονται στο Σχήμα 3.13. Αυτό το ΣΝΔ χρησιμοποιεί συνελικτικά, συγκέντρωσης μεγίστου, κανονικοποίησης και πλήρως συνδεδεμένα επίπεδα με τελικό ταξινομητή Softmax 1000 κατηγοριών. Σε αυτό το δίκτυο εισάγεται η πολύ επιτυχημένη τεχνική της τυχαίας αποκοπής συνδέσεων (dropout) [Hint12] για την αποφυγή της υπερεκπαίδευσης (overfitting).

ReLU μη γραμμικότητα Για τη μοντελοποίηση της ενεργοποίησης των νευρώνων χρησιμοποιείται η ReLU (βλ. Ενότητα 2.2). Σύμφωνα με τους συγγραφείς η χρήση αυτής της συνάρτησης επιταχύνει κατά πολύ την εκπαίδευση (Σχήμα 3.14). Στο δίκτυο χρησιμοποιείται αμέσως μετά από κάθε συνελικτικό επίπεδο και κάθε πλήρως συνδεδεμένο επίπεδο (εκτός από το τελευταίο που εξάγει αποτελέσματα της ταξινόμησης και δεν έχει νόημα να εφαρμοστεί), συνολικά 7 φορές.



Σχήμα 3.14: Σφάλμα εκπαίδευσης ΣΝΔ 4 επιπέδων με ReLU (συνεχόμενη γραμμή) και tanh (διακεκομμένη γραμμή). Πηγή [Kriz12].

Τοπική κανονικοποίηση απόκρισης Αν και οι ανορθωμένες γραμμικές συναρτήσεις ενεργοποίησης δεν χρειάζονται κανονικοποίηση της εισόδου (Ενότητα 2.2), οι συγγραφείς παρατήρησαν ότι η κανονικοποίηση βοηθάει στη γενίκευση, ρίχνοντας το σφάλμα δοκιμής (test error) περίπου 2%. Έτσι χρησιμοποιούν τη Σχέση 3.7, όπου με $a^i_{x,y}$ συμβολίζουν την ενεργοποίηση του νευρώνα, που υπολογίζεται εφαρμόζοντας τη μη γραμμικότητα, μετά από εφαρμογή του φίλτρου i (δηλαδή μιας εγκάρσιας φέτας), στη θέση (x, y). Εν τέλει, η κανονικοποίηση γίνεται ως προς n γειτονικά φίλτρα (πυρήνες) από τα συνολικά N φίλτρα.

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=i-n/2}^{i+n/2} \left(a^j_{x,y}\right)^2 \right)^{\beta} \tag{3.7}$$

Στην Σχέση 3.7 το άθροισμα αναφέρεται σε n συμμετρικά γειτονικές φέτες αποκρίσεων και εννοείται ότι το j δεν παίρνει τιμές κάτω από 0 ή πάνω από $N-1$. Πιο αυστηρά μπορεί να γραφεί ότι $j \in [\max(0, i - \lceil n/2 \rceil), \min(N-1, i + \lceil n/2 \rceil)]$.

Οι υπερπαράμετροι που εισάγονται από αυτό το επίπεδο είναι $\{k, n, \alpha, \beta\}$ και χρησιμοποιώντας το σύνολο επαλήθευσης (validation set) το άρθρο βρίσκει ότι καλύτερες τιμές είναι οι $\{2, 5, 10^{-4}, 0.75\}$. Υπάρχει ομοιότητα αυτής της λειτουργίας με την πλευρική αναστολή των βιολογικών νευρώνων, με την έννοια ότι δημιουργείται ανταγωνισμός μεταξύ γειτονικών νευρώνων και αυτοί με την ισχυρότερη ενεργοποίηση παρεμποδίζουν τους ασθενέστερους. Η τεχνική αυτή εφαρμόζεται μετά τα δύο πρώτα συνελικτικά επίπεδα μόνο, αφού έχει εννοείται εφαρμοστεί η μη γραμμικότητα και έχει γίνει συγκέντρωση χαρακτηριστικών.

Επικαλυπτόμενη συγκέντρωση χαρακτηριστικών Τα 3 επίπεδα συγκέντρωσης αυτού του δικτύου εφαρμόζουν συγκέντρωση χαρακτηριστικών με επικαλυπτόμενα παράθυρα και υπερπαραμέτρους $\mathcal{F} = 3$, δηλαδή παράθυρα 3×3 με επικάλυψη $\mathcal{S} = 2$ θέσεων.

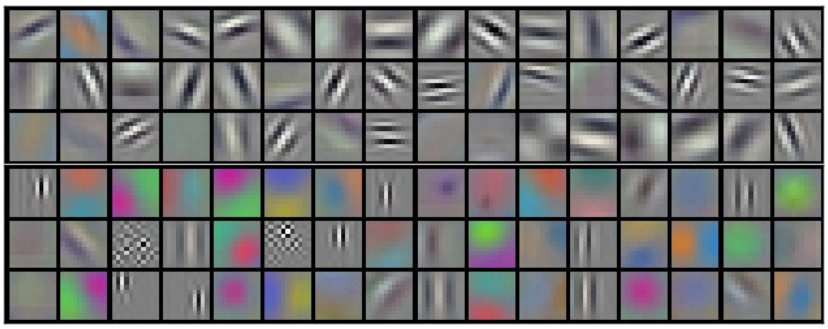
Αντιμετώπιση κινδύνου υπερεκπαίδευσης Οι παρακάτω τεχνικές εφαρμόζονται μόνο κατά την εκπαίδευση.

Τυχαία αποκοπή συνδέσεων (Dropout) Σύμφωνα με αυτή την τεχνική η απόκριση κάθε κρυφού νευρώνα μηδενίζεται με πιθανότητα 0.5, επιβαρύνοντας βέβαια το χρόνο εκπαίδευσης του δικτύου περίπου διπλασιάζοντάς τον. Σύμφωνα με το [Hint12] η αποκοπή μερικών των συνδέσεων (δηλαδή ο μηδενισμός τους), οδηγεί κάθε νευρώνα να εκπαιδευτεί πιο εύρωστα και να στηρίζει την απόφασή του περισσότερο στον εαυτό του παρά στους γείτονες νευρώνες.

Επαύξηση συνόλου δεδομένων Χρησιμοποιήθηκαν 2 μέθοδοι αύξησης του πλήθους των εικόνων. Σύμφωνα με την πρώτη, η οποία δεν επιβαρύνει υπολογιστικά το σύστημα, εφαρμόστηκαν γεωμετρικοί μετασχηματισμοί (τυχαίες μετατοπίσεις και οριζόντιος κατοπτρισμός) δίνοντας δείγματα εκπαίδευσης διαστάσεων 224×224. Λογικό είναι αυτή η μέθοδος να δημιουργεί μεγάλη συσχέτιση στο σύνολο εκπαίδευσης, αφού υπάρχουν μεγάλες εξαρτήσεις μεταξύ των εικόνων. Η δεύτερη μέθοδος βασίζεται στην ιδέα ότι η ταυτότητα των αντικειμένων σε μια εικόνα δεν θα πρέπει να επηρεάζεται από μικρές αλλαγές στο χρώμα και στην ποσότητα του φωτισμού. Σε κάθε εικόνα κάθε φορά που δίνεται ως δείγμα προς εκπαίδευση προστίθεται, και στα τρία κανάλια (RGB), μία ποσότητα που εξαρτάται από τα ιδιοδιανύσματα και τις ιδιοτιμές του πίνακα συνδιακύμανσης 3×3 (covariance matrix) της εικόνας, ενώ περιλαμβάνει και μία τυχαιότητα (λεπτομέρειες στο [Kriz12]). Υλοποιώντας όλες αυτές τις μεθόδους το σφάλμα ταξινόμησης πέφτει επιπλέον παραπάνω από 1%.

Εκπαίδευση και Δοκιμή του Δικτύου Ο τελικός ταξινομητής έχει συνάρτηση κόστους την Softmax και υλοποιεί την Πολυωνυμική Λογιστική Παλινδρόμηση (Ενότητα 2.4.3). Για την εκπαίδευση το σύνολο δεδομένων αυξάνεται τεχνητά (Ενότητα 3.4.1) και οι τελικές εικόνες εκπαίδευσης έχουν μέγεθος 224×224. Παράμετροι της εκπαίδευσης: batch size: 128 εικόνες, momentum: 0.9 και weight decay: 0.0005, learning rate: 0.01 και μειώθηκε κατά βήματα 3 φορές, 90 εποχές. Κατά τη φάση της δοκιμής, το δίκτυο εξάγει 10 κομμάτια (4 γωνιακά, 1 κεντρικό και τα οριζόντια συμμετρικά τους) από την αρχική εικόνα προς ταξινόμηση. Η τελική απόφαση λαμβάνεται από το μέσο όρο των 10 ταξινομήσεων.

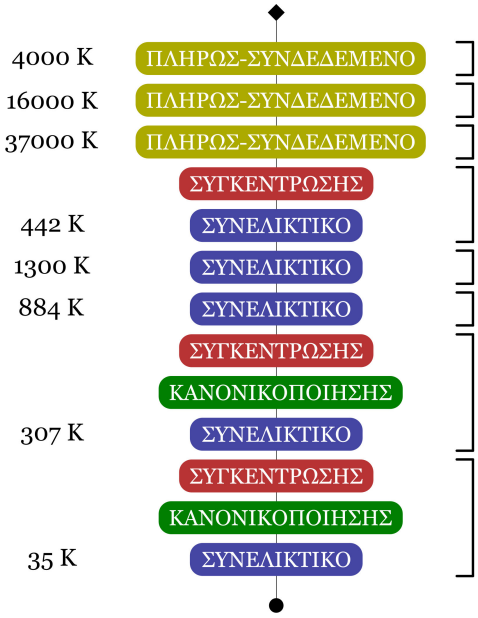
Στο Σχήμα 3.15 φαίνονται τα 96 φίλτρα που μαθαίνονται στο πρώτο συνελικτικό επίπεδο. Οι διαστάσεις τους είναι $11 \times 11 \times 3$ διότι το πρώτο επίπεδο «βλέπει» όγκο εισόδου με βάθος 3 (3 κανάλια



Σχήμα 3.15: Τα 96 φίλτρα διαστάσεων $11 \times 11 \times 3$ του πρώτου συνελικτικού επιπέδου του AlexNet, που διαμοιράζονται από τους 55×55 νευρώνες σε κάθε στρώμα νευρώνων.

των έγχρωμων εικόνων) και γι' αυτό το λόγο η οπτικοποίηση των φίλτρων είναι έγχρωμη. Φαίνεται ότι ενσωματώνουν στοιχεία διαφορετικών συχνοτήτων και κατευθύνσεων, ενώ έχουν συγκεκριμένη προτίμηση προς τη μάθηση ακμών και ομάδων χρώματος. Τα πάνω 48 φίλτρα εκπαιδεύονται στην GPU 1, ενώ τα κάτω 48 στην GPU 2 και οι δομικές διαφορές τους οφείλονται στην επικοινωνία των δύο καρτών μόνο σε συγκεκριμένα επίπεδα (που δεν θα σχολιαστεί περαιτέρω εδώ).

Σχόλια Στο Σχήμα 3.16 φαίνονται όλα τα επίπεδα του δικτύου Alexnet και δίπλα σε καθένα αναγράφεται ο αριθμός των βαρών (παραμέτρων) που εισάγει. Συνολικά το δίκτυο έχει 60 εκατομμύρια παραμέτρους προς βελτιστοποίηση.



Σχήμα 3.16: Η αλληλουχία επιπέδων στο δίκτυο AlexNet. Βάρη προς μάθηση εισάγουν μόνο τα επίπεδα που περιέχουν νευρώνες (Συνελικτικά και Συγκέντρωσης). Οι αγκύλες δείχνουν τη συνήθη ομαδοποίηση σε 8 λειτουργικά επίπεδα.

Η έξοδος του 5ου συνελικτικού επιπέδου της αρχιτεκτονικής είναι ένας όγκος χαρακτηριστικών, που διανυσματοποιείται για να αποτελέσει το διάνυσμα εισόδου του 1ου πλήρως συνδεδεμένου επιπέδου. Τα χαρακτηριστικά που εξάγονται από το δίκτυο είναι πολύ χρήσιμα και είναι δυνατόν να

αποθηκευτούν και να χρησιμοποιηθούν ως περιγραφή της εισόδου (εικόνας) σε οποιοδήποτε άλλο περιβάλλον (π.χ. ταξινόμηση με SVM). Η είσοδος του δικτύου έχει διαστάσεις $224 \times 224 \times 3$, δηλαδή η πληροφορία της εικόνας περιγράφεται από $150,528$ τιμές (πίξελ). Η αναπαράσταση που προκύπτει (συνελικτικά χαρακτηριστικά) από το δίκτυο έχει διαστάσεις $7 \times 7 \times 512$, δηλαδή 512 χάρτες χαρακτηριστικών 7×7 και έχει $25,088$ στοιχεία (6 φορές λιγότερα).

3.4.2 ZF Net, GoogLeNet, VGG Net

Το AlexNet το 2012 στο διαγωνισμό ImageNet-LSVRC μείωσε το top-5 σφάλμα ταξινόμησης εικόνων, από όλες τις προηγούμενες τεχνικές σε 15.3% και μόλις ένα χρόνο αργότερα το ZF Net από τους Zeiler και Fergus πέτυχε top-5 σφάλμα περίπου 12%[8]. Οι διαφορές από το AlexNet ήταν λίγες· κυρίως η αύξηση του βάθους και η τροποποίηση μερικών κρίσιμων υπερπαραμέτρων. Το 2014 στον ίδιο διαγωνισμό ταξινόμησης το GoogLeNet ρίχνει το σφάλμα στο μισό 6.7%. Οι διαφορές είναι κυρίως στο βάθος του δικτύου (22 επίπεδα), αλλά και στο πλήθος των παραμέτρων προς βελτιστοποίηση, μόλις 4 εκατομμύρια, αντί για 20 εκατομμύρια του AlexNet. Το δεύτερο δίκτυο στην κατάταξη, το VGG Net, επιτυγχάνει λίγο μεγαλύτερο σφάλμα 7.3% με 140 εκατομμύρια παραμέτρους, αλλά έχει δύο πολύ θετικά στοιχεία. Πρώτον, είναι αρκετά ομοιόμορφο και τα 16 επίπεδά του έχουν ίδιες τιμές υπερπαραμέτρων ($\mathcal{F} = 3,\ \mathcal{S} = 1$ για το συνελικτικό επίπεδο και $\mathcal{F} = 2,\ \mathcal{S} = 2$ για το επίπεδο συγκέντρωσης). Δεύτερον, τα συνελικτικά χαρακτηριστικά που εξάγονται από αυτό το δίκτυο έχουν μεγάλη δύναμη αναπαράστασης, ακόμα και από το GoogLeNet, και χρησιμοποιούνται στην πράξη περισσότερο.

[8] Το σφάλμα αυτό προέρχεται από μία επιτροπή παρόμοιων ΣΝΔ.

Κεφάλαιο 4

Σύγχρονα Εργαλεία και πλαίσιο ανάπτυξης ΣΝΔ

Βασικό εργαλείο για την μοντελοποίηση Βαθέων Νευρωνικών Δικτύων είναι το Caffe [Jia 14], που αναπτύχθηκε το 2014 και συντηρείται από το Berkeley Vision and Learning Center (BVLC). Αποτελεί ένα ολοκληρωμένο πλαίσιο ανάπτυξης, επαλήθευσης και προτυποποίησης ΒΝΔ, που έχουν διάταξη Κατευθυνόμενου Ακυκλικού Γράφου. Το Caffe έχει υλοποιηθεί σε σύγχρονες γλώσσες προγραμματισμού (C++, Python) και πλαισιώνεται από νέες τεχνολογίες και μηχανισμούς (Protocol Buffers, LMDB, CUDA, DIGITS), που του προσδίδουν φορητότητα, επεκτασιμότητα και ευκολία υλοποίησης.

Μερικά σημεία που το Caffe εμφανίζει βελτιώσεις συγκριτικά με τους προκατόχους του (Theano/ Pylearn2, cuda-convnet, Torch7, Decaf) είναι τα εξής:

- αρθρωτή ανάπτυξη (modularity): προσφέρει ευκολία επέκτασης και ανάπτυξης νέων επιπέδων, αλλά και τροποποίηση υπαρχόντων,
- διαχωρισμός της υλοποίηση της βιβλιοθήκης και των προγραμμάτων από την αναπαράσταση και τις δυνατότητες των Νευρωνικών Δικτύων που κατασκευάζει,
- υπολογιστική υποστήριξη τόσο σε CPU, όσο και σε GPU: Οι Γραφικές Μονάδες Επεξεργασίας επιτρέπουν μαζική παραλληλία, λόγω των πολλαπλών πυρήνων (τάξης χιλιάδων σε σχέση με μερικών δεκάδων στη CPU). Αυτό οδήγησε στην ανάπτυξη ΝΔ πολύ μεγάλης κλίμακας και στην άνθιση των ΣΝΔ, τα οποία μπορούν να εκμεταλλευθούν σε μεγάλο βαθμό υπολογιστικά συστήματα που υποστηρίζουν παραλληλία.
- προσφέρει επικοινωνία και υλοποιεί διεπαφές με μοντέρνα προγραμματιστικά περιβάλλοντα (MATLAB, Python και terminal) για γρήγορη προτυποποίηση και ενσωμάτωση σε υπάρχων κώδικα,
- ύπαρξη προ-εκπαιδευμένων, επιτυχημένων μοντέλων αναφοράς για γρήγορο πειραματισμό, χωρίς να χρειάζεται χρονοβόρα εκπαίδευση κάθε φορά και εύκολη σύγκριση με baseline μοντέλα για νέες μεθόδους και αποτελέσματα,
- υλοποίηση του πυρήνα τη βιβλιοθήκης σε C++, που δίνει τη δυνατότητα ενσωμάτωσης σε υπάρχοντα συστήματα, χωρίς να υπάρχει η ανάγκη εξειδικευμένων απαιτήσεων σε υλικό ή λειτουργικό.

4.1 Αρχιτεκτονική και δυνατότητες του Caffe

Τέσσερα είναι τα βασικά στοιχεία του μοντέλου ανάπτυξης του Caffe:

1. Δίκτυο: Ο ορισμός της δομής του δικτύου γίνεται επίπεδο προς επίπεδο σε κατανοητή από τον άνθρωπο γλώσσα (plaintext protocol buffer schema (prototxt)). Τα εκπαιδευμένα δίκτυα και οποιαδήποτε άλλα δεδομένα χρειάζεται να μεταφερθούν σειριοποιούνται, σύμφωνα με τη δομή που καθορίζεται από το μοντέλο (binary protocol buffer (binaryproto)).

2. Μάζες αποθήκευσης (Blobs): Τα δεδομένα (π.χ. πακέτα εικόνων, παράμετροι μοντέλου, παράγωγοι για βελτιστοποίηση) μετακινούνται μεταξύ των επιπέδων σε μία ενιαία ορισμένη δομή τα Blobs. Στη γενικότερη μορφή τους είναι 4D πίνακες. Για δεδομένα, π.χ. για εικόνες έχουν τη μορφή: N(μέγεθος πακέτου) $\times$ K(κανάλια) $\times$ H(ύψος) $\times$ W(πλάτος), ενώ για τα βάρη και τα κατώφλια του επιπέδου συνέλιξης: N(έξοδοι) $\times$ K(είσοδοι) $\times$ H(ύψος) $\times$ W(πλάτος) και N(κατώφλια) $\times 1 \times 1 \times 1$, κ.ο.κ. Μέσω αυτής της οργάνωσης το Caffe χειρίζεται το συγχρονισμό μεταξύ CPU - GPU αποκρύπτοντας τις δυσκολίες από τον σχεδιαστή.

3. Επίπεδα: Το βασικό δομικό στοιχείο των αρχιτεκτονικών του Caffe είναι τα επίπεδα. Οποιαδήποτε ενέργεια υποστηρίζεται, μπορεί να γίνει από το αντίστοιχο επίπεδο, π.χ. φόρτωση και προεπεξεργασία εικόνων, φιλτράρισμα, συνέλιξη, συγκέντρωση, κανονικοποίηση, υπολογισμοί κόστους. Εκτός από τα υλοποιημένα επίπεδα, ο σχεδιαστής μπορεί να δημιουργήσει και άλλα εξειδικευμένα, ορίζοντας τις 3 βασικές λειτουργίες τους (εγκατάσταση, εμπρόσθια και προς τα πίσω διάδοση). Κατά την εγκατάσταση το επίπεδο αρχικοποιείται και ελέγχει τις συνδέσεις του με προηγούμενα και επόμενα επίπεδα. Η εμπρόσθια διάδοση καθορίζει τους υπολογισμούς που γίνονται στον όγκο εισόδου για να προκύψει ο όγκος εξόδου, λαμβάνει μέρος βασικά κατά την ανάκληση ενός δικτύου, αλλά και κατά τη μάθηση. Η οπισθοδιάδοση καθορίζει πώς θα μεταφερθούν κατά την βελτιστοποίηση οι κλίσεις ως προς το επόμενο επίπεδο στο προηγούμενο επίπεδο, δηλαδή υλοποιεί ένα βήμα του κανόνα της αλυσίδας.

4. Η εκπαίδευση, η δοκιμή και η ειδική προσαρμογή (Training, Testing, Fine-tuning, Deploying) γίνονται από κοινό πρόγραμμα, παρέχοντάς του τα απαραίτητα δεδομένα. Για παράδειγμα η εκπαίδευση γίνεται σε περιβάλλον τερματικού με την εντολή caffe -train και παρέχοντας το αρχείο ορισμού του μοντέλου (π.χ. `train-val.prototxt`), που περιλαμβάνει την αρχιτεκτονική του, και το αρχείο επίλυσης (π.χ. `solver.prototxt`), που περιέχει υπερπαραμέτρους της εκπαίδευσης και άλλες ρυθμίσεις. Στην περίπτωση της ειδικής προσαρμογής πρέπει επιπλέον να παρέχονται, τα βάρη και η δομή του ήδη εκπαιδευμένου μοντέλου (π.χ. `alexnet.caffemodel`).

Στη συνέχεια, αναφέρουμε μερικές παρατηρήσεις πάνω στο Caffe που χρησιμοποιήθηκαν στην ανάπτυξη και εκπαίδευση του μοντέλου για τον εντοπισμό προσώπων στην παρούσα Διπλωματική Εργασία:

- Το Caffe δεν μετράει τις ανακυκλώσεις πάνω στο σύνολο δεδομένων με εποχές, αλλά με επαναλήψεις.

$$\frac{\text{μέγεθος συνόλου δεδομένων}}{\text{μέγεθος πακέτων}} = \frac{\text{επαναλήψεις}}{\text{εποχές}} \tag{4.1}$$

Επομένως, για ένα συχνά χρησιμοποιούμενο μέγεθος πακέτων βελτιστοποίησης 128 εικόνων, από ένα σύνολο 50,000 εικόνων, αν επιλέξουμε 20 εποχές μάθησης θα γίνουν 1,600,000 επαναλήψεις.

- Ο διαμοιρασμός παραμέτρων (π.χ. σε συνελικτικά ή πλήρως συνδεδεμένα επίπεδα) γίνεται πολύ εύκολα δίνοντας το ίδιο όνομα στην επιλογή `param` π.χ. `param: 'sharedweights'` μέσα στον ορισμό ενός επιπέδου.

- Στο Caffe κάθε επίπεδο έχει δυνατότητα να έχει το δικό του ρυθμό μάθησης μέσα από τις υπερπαραμέτρους πολλαπλασιασμού `lr_mult` του γενικού ρυθμού μάθησης που καθορίζεται στο αρχείο επίλυσης.

- Ειδική προσαρμογή. Όταν παρέχεται κατά την εκπαίδευση ένα μοντέλο με βάρη, τότε το Caffe μπαίνει σεmode fine-tuning, αντιγράφοντας όλα τα επίπεδα από τον ορισμό του μοντέλου και όλες τις παραμέτρους από το ήδη εκπαιδευμένο μοντέλο, εκτός από τα επίπεδα που έχουν διαφορετικό όνομα. Αυτά τα αρχικοποιεί με τυχαία βάρη. Με αυτόν τον τρόπο μπορεί να γίνει ειδική προσαρμογή από ένα πρόβλημα 1000 κατηγοριών (π.χ. ImageNet) σε ένα νέο με μόνο 20 κατηγορίες, εκπαιδεύοντας εξ' αρχής μόνο το τελευταίο επίπεδο, αλλάζοντας το όνομά του. Αν επιθυμούμε η μάθηση να μην συνεχίζει στα υπόλοιπα επίπεδα θέτουμε απλά την παράμετρο πολλαπλασιασμού του ρυθμού μάθησης σε `lr_mult:0`. Επίσης, αν θέλουμε το τελευταίο/νέο επίπεδο να μαθαίνει γρήγορα, ενώ τα προηγούμενα/προ-εκπαιδευμένα να μαθαίνουν αργά. Αυτό καθορίζεται στο αρχείο ορισμού του μοντέλου και από τις υπερπαραμέτρους `lr_mult` των επιπέδων.

4.2 Απλό δίκτυο για δυαδική λογιστική παλινδρόμηση

Οι 3 επιλογές έχουν ως εξής:

1. Μοντέλο δεδομένων: γραμμικό, από την 2.7 έχουμε:

$$\mathbf{g}\left(\mathbf{x}; W\right) = \begin{bmatrix} \mathbf{w}_1^T \mathbf{x} \\ \mathbf{w}_2^T \mathbf{x} \end{bmatrix}$$

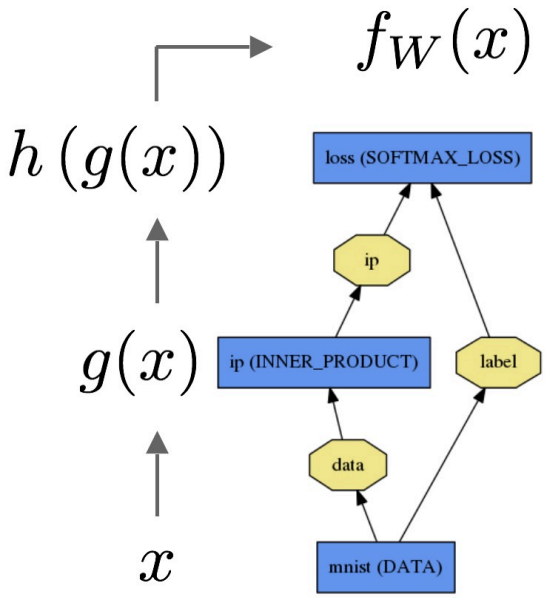
2. Συνάρτηση κοστους: διεντροπίας, από την 2.18 έχουμε[1]:

$$L = -\sum_{i=1}^{64} \log \left(\frac{e^{g_{y_i}}}{e^{g_1} + e^{g_2}} \right)$$

   Η πιθανότητα να ανήκει το διάνυσμα εισόδου $\mathbf{x}$ στις δύο κατηγορίες δίνεται από το διάνυσμα $\boldsymbol{\sigma}\left(\mathbf{g}\right)$ (συνάρτηση softmax – Ενότητα 2.2).

[1] Εφόσον επιλέγουμε να μην προσθέσουμε παράγοντα κανονικοποίησης, ο πολλαπλασιαστικός παράγοντας $1/64$ μπορεί να παραληφθεί για την ελαχιστοποίηση.

3. Ελαχιστοποίηση κόστους με κάποιον από τους αλγορίθμους της Ενότητας 3.3.7, συνήθως με τον SGD.




```
name: ”LogReg”
layer {
name: ”mnist”
type: ”Data”
top: ”data”
top: ”label”
data_param {
source: ”input_leveldb”
batch_size: 64 }
}
layer {
name: ”ip”
type: ”InnerProduct”
bottom: ”data”
top: ”ip”
inner_product_param {
num_output: 2 }
}
layer {
name: ”loss”
type: ”SoftmaxWithLoss”
bottom: ”ip”
bottom: ”label”
top: ”loss”
}
```

Κεφάλαιο 5

Εντοπισμός Προσώπων με χρήση ΣΝΔ

Τα προβλήματα σχετικά με την αναγνώριση ανθρώπινων προσώπων σε εικόνες και βίντεο ήταν από τα πρώτα που διατυπώθηκαν στην Όραση Υπολογιστών και στην Αναγνώριση Προτύπων. Τα κυριότερα είναι τα εξής:

- Διαπίστωση ύπαρξης (detection) ή όχι προσώπων στην εικόνα, δηλαδή αν στη σκηνή που απεικονίζεται υπάρχουν εμφανή πρόσωπα. Δοθέντος μιας εικόνας το πρόβλημα αυτό είναι το ίδιο με την ταξινόμησή της σε δύο κατηγορίες, εικόνα προσώπου ή όχι.
- Εντοπισμός (localization) των «αρκετά» φανερών προσώπων.
- Εντοπισμός μερών/χαρακτηριστικών του προσώπου, που συνήθως χρησιμοποιούνται στην καταχώρηση (image registration) και ευθυγράμμιση (alignment) προσώπων μέσω συγκρίσεων και μετασχηματισμών.
- Αναγνώριση (Recognition) της ταυτότητας των προσώπων με χρήση ή όχι αποτελεσμάτων από το προηγούμενο πρόβλημα.

Επιπρόσθετα με τους παράγοντες που αναφέρθηκαν στην Ενότητα 1.1 η Αναγνώριση Προσώπων αντιμετωπίζει τις εξής δυσκολίες:

- Ένα πρόσωπο μπορεί να έχει πολλές διαφορετικές εκφράσεις, διακυμάνσεις στο σχήμα, χρώμα και φωτισμό (π.χ. σκιάσεις).
- Η γωνία λήψης δίνει στην προβολή του προσώπου πολλές διαφορετικές μορφές.
- Είναι πολύ σύνηθες ένα πρόσωπο να επικαλύπτεται μερικώς, συνήθως από χέρια, γυαλιά, μαλλιά ή άλλα αντικείμενα που βρίσκονται μπροστά.

5.1 Σύντομη Ιστορική Ανασκόπηση Τεχνικών Εντοπισμού Προσώπων

Σχεδόν κάθε τεχνική για εντοπισμό και αναγνώριση που έχει εμφανιστεί στα πεδία της Όρασης Υπολογιστών και της Επεξεργασίας Σημάτων έχει μεταφερθεί και εφαρμοστεί στον εντοπισμό και την αναγνώριση προσώπων. Οι κυριότερες ανασκοπήσεις είναι οι [Zhan10, Zhao03, Tolb06, Sach15, Hjel01] απ' όπου αναφέρονται ενδεικτικά, οι πιο επικρατούσες και θεμελιώδεις μέθοδοι για τον εντοπισμό προσώπων:

1. Βασισμένοι στον αλγόριθμο Viola-Jones. Το 2001 οι P. Viola και M. Jones σχεδίασαν έναν ανιχνευτή προσώπων πραγματικού χρόνου, που χρησιμοποιεί χαρακτηριστικά τύπου Haar, τον αλγόριθμο εκπαίδευσης Adaboost και ένα σύνολο από ταξινομητές σε αλληλουχία [Viol01]. Η εξαγωγή χαρακτηριστικών τύπου Haar γίνεται αθροίζοντας τιμές πίξελ προκαθορισμένων περιοχών και αφαιρώντας γειτονικά αθροίσματα, σύμφωνα με προκαθορισμένα πρότυπα. Είναι πολύ γρήγορα στον υπολογισμό τους και αντί να περιγράφουν συστατικά στοιχεία του προσώπου (π.χ. μάτια, στόμα), περιγράφουν περιοχές σε επίπεδο σχετικών διαφορών (π.χ. φωτεινές περιοχές που είναι δίπλα ή περιβάλλονται από σκοτεινές περιοχές).

   Ο αλγόριθμος χρειάζεται ένα σύνολο δεδομένων πάνω στο οποίο θα εκπαιδευτεί και χρησιμοποιεί τον Adaboost για βρει τα καλύτερα χαρακτηριστικά να εξάγει ισχυρά συμπεράσματα από μία αλληλουχία από αδύναμους ταξινομητές. Επεκτάσεις, που χρησιμοποιούν στην βάση τους αυτόν τον αλγόριθμο και βελτιώνουν κάθε φορά κάποιο τμήμα του, δημοσιεύονται ακόμα, ωστόσο τα συστήματα που βασίζονται στον Viola-Jones αντιμετωπίζουν τα κάποιες δυσκολίες. Πρώτον, κάνουν υποθέσεις σχετικά με τη γωνία θέασης του προσώπου, π.χ. ο βασικός αλγόριθμος εντόπιζε μόνο όρθια πρόσωπα, που κοιτάζουν κατά μέτωπο την κάμερα και δεν είναι επικαλυμμένα. Επίσης, αν και πολλές επεκτάσεις βελτιώνουν αυτόν τον περιορισμό χρειάζονται επιπλέον επισημάνσεις για την κατεύθυνση και άλλα στοιχεία του προσώπου. Το πιο σύγχρονο σύστημα με αυτή τη φιλοσοφία ονομάζεται HeadHunter [Math14] και παρουσιάστηκε το 2014.

2. Βασισμένοι σε Μοντέλα Παραμορφώσιμων Μερών (DPM). Ένα πρόσωπο μοντελοποιείται από τα μέρη του, που καθορίζονται από επιβλεπόμενη ή μη μάθηση και ένας ταξινομητής, συνήθως ένας λανθάνων SVM, εκπαιδεύεται για να τα εντοπίζει και να βρίσκει τις γεωμετρικές συσχετίσεις μεταξύ τους. Αν και αυτές οι τεχνικές είναι αρκετά εύρωστες σε ό,τι αφορά επικαλύψεις, είναι υπολογιστικά δαπανηρές. Επίσης, σε πολλές περιπτώσεις είναι αναγκαία η εκπαίδευση πολλαπλών μοντέλων για επίτευξη αποτελεσμάτων state-of-the-art και χρειάζονται επιπλέον επισημάνσεις για τα μέρη του προσώπου κατά την εκπαίδευση. Την καλύτερη επίδοση σήμερα, ακόμα και από σύνθετα συστήματα DPM, έχει το σύστημα που προτάθηκε το 2014 στο [Math14] και αποτελείται από ένα απλό DPM (vanilla DPM), που εκπαιδεύεται όμως με «βέλτιστες υπερπαραμέτρους» (αριθμός μερών, πλήθος δεδομένων, LUV χρωματικά κανάλια, κτλ.).

3. Βασισμένοι σε εξαγωγή περιγραφητών (HOG, SHIFT) και SVM.

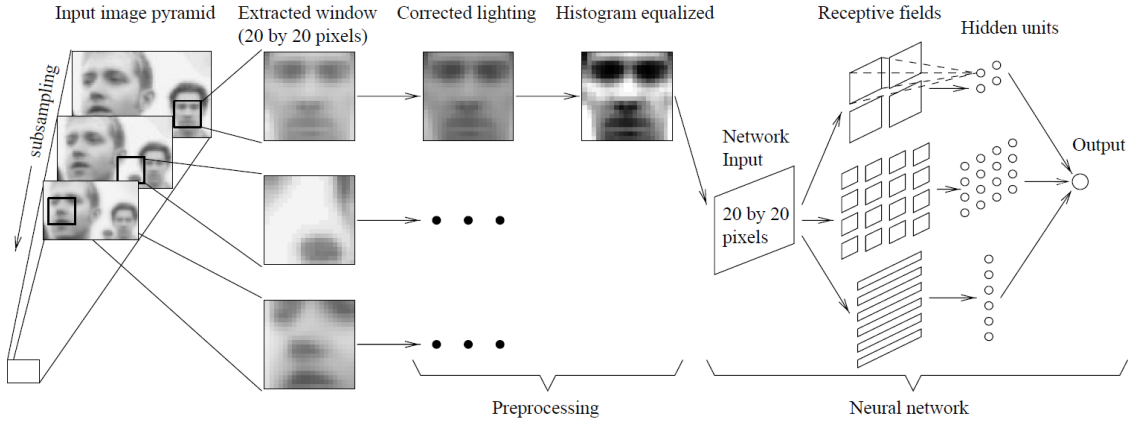
4. Βασισμένοι σε Νευρωνικά Δίκτυα. Έχουν γίνει πολλές προσπάθειες εφαρμογής των νευρωνικών δικτύων στο παρελθόν, για τον εντοπισμό προσώπων με μέτρια ποιότητα αποτελεσμάτων, που δεν οφειλόταν στην αποτελεσματικότητα των μεθόδων, αλλά στο μικρό μέγεθος των δεδομένων και στις χαμηλές υπολογιστικές ικανότητες. Για παράδειγμα, στο [Vail94] (1994) αναπτύσσεται ένα σύστημα δύο επιπέδων με ρηχά ΣΝΔ. Συγκεκριμένα το πρώτο δίκτυο εντοπίζει χονδρικά τις θέσεις των προσώπων και το δεύτερο επαληθεύει τον εντοπισμό και προτείνει λεπτομερέστερες θέσεις. Για σύγκριση με τα σημερινά δεδομένα παραθέτονται κάποια μεγέθη του συστήματος: εικόνες συνόλου εκπαίδευσης 20×20 πίξελ, 1157 βάρη δικτύου, 4 επίπεδα: 2 συνελικτικά, 1 συγκέντρωσης και 1 πλήρως συνδεδεμένο. Στο [Rowl98] (1998) εκπαιδεύονται πολλαπλά δίκτυα που συνδυάζονται για βελτίωση της απόδοσης, ενώ χρησιμοποιείται και ανάλυση σε πολλαπλές κλίμακες. Το σύστημα φαίνεται στο Σχήμα 5.1. Τα μεγέθη του συστήματος

δεν συγκρίνονται με τα σημερινά, ενώ είχε αναπτυχθεί κυρίως για όρθια πρόσωπα στον άξονα της κάμερας. Στο [Garc04] (2004) εκπαιδεύεται ένα αυτοτελές ΣΝΔ δίκτυο με περισσότερα επίπεδα από προηγούμενες προσπάθειες, αλλά με λιγότερα από τα σημερινά, ενώ λαμβάνονται μερικώς υπόψη στροφές των προσώπων στο επίπεδο της κάμερας ή αλλαγές πόζας. Τέλος, στο [Osad07] (2007) προτείνεται η εκπαίδευση ενός δικτύου, που επιτυγχάνει συνεργατικά εντοπισμό και ανίχνευση πόζας των προσώπων. Η από κοινού μάθηση έχει καλύτερα αποτελέσματα και στα δύο επιμέρους προβλήματα.

Κατά τη διάρκεια συγγραφής της Διπλωματικής παρουσιάστηκαν συστήματα και από τις τρεις γενικές κατηγορίες που ξεπερνούν σε αποτελεσματικότητα τον ανιχνευτή προσώπων DDFD που υλοποιείται στη συνέχεια. Αυτά είναι:

- Faceness-Net [Yang15] (Σεπτ. 2015): Το σύστημα εκπαιδεύει ένα Βαθύ Συνελικτικό Δίκτυο (ΕΣΝ), μία παραλλαγή των ΣΝΔ που προτείνουν οι συγγραφείς, και εντοπίζει πολλά μέρη του κεφαλιού για να ψηφίσει και να αποφανθεί αν το μέρος που τα περιέχει είναι πρόσωπο. Σύμφωνα με τα δημοσιευμένα αποτελέσματα ξεπερνάει όλα τα διαθέσιμα συστήματα.

- DenseBox [Huan15] (Σεπτ. 2015): Ένα εξαιρετικό σύστημα που γενικεύει τον εντοπισμό αντικειμένων (πρόσωπα, αυτοκίνητα ή ο,τιδήποτε άλλο) σε ένα πλαίσιο μάθησης Πλήρως Συνελικτικών Δικτύων (FCN), που μπορεί να επωφεληθεί επιπλέον από τυχόν διαθέσιμα σημεία ενδιαφέροντος (landmarks). Ξεπερνάει στον εντοπισμό προσώπων την αποτελεσματικότητα του DDFD και όλων των άλλων διαθέσιμων μεθόδων κατά ένα μεγάλο ποσοστό. Το μόνο αρνητικό στοιχεία, όπως διαπιστώνουμε και από τον DDFD είναι ο χρόνος συμπεράσματος (inference time), ο οποίος είναι αρκετά δευτερόλεπτα για μία εικόνα.[1]



Σχήμα 5.1: Τα στάδια του πολυκλιμακωτού αλγορίθμου εντοπισμού προσώπων που προτάθηκε στο [Rowl98]. Πηγή: [Rowl98].

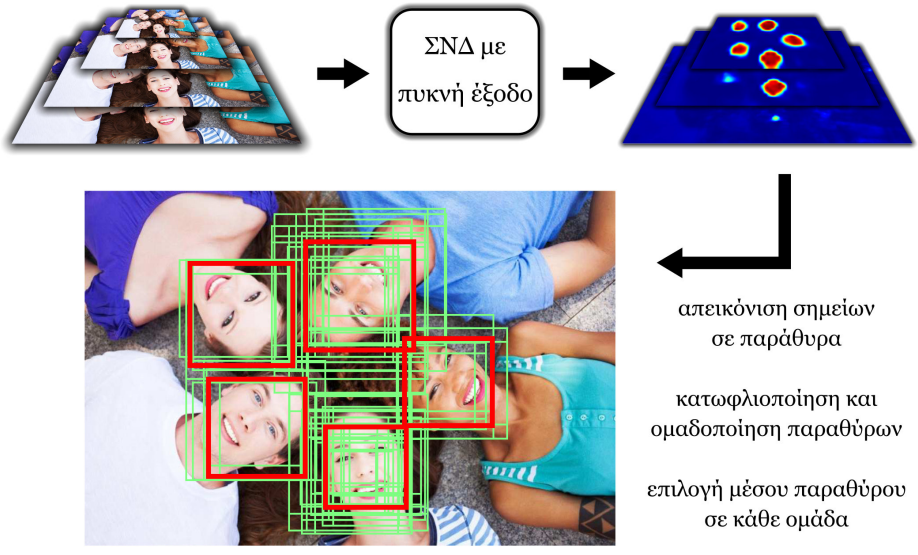
[1] Οι συγγραφείς σημειώνουν ότι αυτό το πρόβλημα έχει ξεπεραστεί, ενώ θα γίνει και σχετική δημοσίευση στο μέλλον.

5.2 Ανίχνευση προσώπων ανεξαρτήτου γωνίας θέασης με τον Deep Dense Face Detector (DDFD)

5.2.1 Περιγραφή του Ανιχνευτή DDFD

Στην ενότητα αυτή περιγράφεται ο Ανιχνευτής προσώπων που προτείνεται στο [Sach15] και βασίζεται στα [Garc04, Osad07]. Καθώς δεν υπάρχει κάποια υλοποίηση του DDFD ή μερών αυτού διαθέσιμη δημόσια, το συνολικό σύστημα πρέπει να αναπτυχθεί από την αρχή. Ο Ανιχνευτής αποτελεί ένα αυτοτελές σύστημα με βάση ένα ΣΝΔ, που είναι σε θέση να εντοπίζει πρόσωπα σε εικόνες σε μεγάλο εύρος προσανατολισμών, με διακυμάνσεις στο φωτισμό και με μερικές επικαλύψεις, ενώ για την εκπαίδευση δεν απαιτεί επιπλέον επισημάνσεις (π.χ. επισημάνσεις πόζας, πλήθος προσώπων, διακριτικά σημεία) όπως άλλοι μέθοδοι, παρά μόνο τη θέση του προσώπου. Οι διαφορές από τις προηγούμενες μεθόδους με ΝΔ είναι δύο· πρώτον το νευρωνικό αυτό είναι αρκετά πιο βαθύ από τα προηγούμενα και δεύτερον εκπαιδεύεται πάνω σε ένα πολύ μεγαλύτερο σύνολο εκπαίδευσης.

Ο DDFD υιοθετεί τη προσέγγιση κυλιόμενων παραθύρων για τον εντοπισμό προσώπων. Αντίπαλος αυτής της τεχνικής είναι ο εντοπισμός σε περιοχές ενδιαφέροντος με χρήση κάποιας μεθόδου εξαγωγής τέτοιων περιοχών (π.χ. selective search). Ο κύριος λόγος της πρώτης επιλογής είναι ότι πρόσθετες μονάδες κάνουν πιο πολύπλοκο και πιο αργό το συνολικό σύστημα, ενώ επιπλέον δημιουργούν εξάρτηση των αποτελεσμάτων του ΣΝΔ από την αποτελεσματικότητα της εξαγωγής σημαντικών περιοχών. Στο Σχήμα 5.2 παρουσιάζεται περιληπτικά η λειτουργία του συστήματος και στα Σχήματα A.1 - A.3 του Παραρτήματος παραδείγματα εντοπισμού προσώπων.



Σχήμα 5.2: Περιγραφή του DDFD με NMS μέσου όρου. Η πυραμίδα εικόνων διέρχεται από το ΣΝΔ που εξάγει πιθανότητες για το αν συγκεκριμένα παράθυρα είναι πρόσωπα ή όχι. Οι περιοχές μαζί με τις αντίστοιχες πιθανότητες διέρχονται από το υποσύστημα NMS μέσου όρου που εξάγει τις ακριβείς θέσεις των προσώπων με τη μορφή παραθύρου.

Ανάλυση του συστήματος

Τη βάση του συστήματος αποτελεί ένα ΣΝΔ, παραπλήσιας αρχιτεκτονικής με το AlexNet, που έχει στόχο τη δυαδική ταξινόμηση ενός παραθύρου της εικόνας ως πρόσωπο ή όχι. Μία εικόνα της καθημερινότητας μπορεί να περιέχει πρόσωπα σε πολλές κλίμακες και γι' αυτό χρησιμοποιείται πολυκλιμακωτός εντοπισμός. Κατασκευάζεται μία πυραμίδα κλιμάκων της εικόνας και για κάθε παράθυρο σε κάθε εικόνα το ΣΝΔ αποφασίζει αν είναι πρόσωπο ή όχι. Το πλήθος επιπέδων της πυραμίδας είναι μία υπερπαράμετρος που καθορίζεται με πειραματισμό και το βήμα εξαγωγής παραθύρων υπαγορεύεται από την χωρική ανάλυση του ΣΝΔ (εξετάζονται παρακάτω).

Το επίπεδο εξόδου του τροποποιημένου AlexNet είναι ένας δυαδικός softmax ταξινομητής (βλ. Ενότητα 4.2, 2.4.3), επομένως το ΣΝΔ έχει έξοδο μία πιθανότητα για το κατά πόσο θεωρεί ένα παράθυρο πρόσωπο ή όχι. Σημαντικό είναι να τονίσουμε ότι το δίκτυο αποφασίζει αν το παράθυρο εισόδου 227×227 [2] είναι πρόσωπο και όχι αν περιέχει πρόσωπο/α, ή είναι τμήμα προσώπου/ων. Αυτό δεν αποτελεί υπόθεση ή περιορισμό, αλλά βασίζεται στην επιλογή της πολυκλιμακωτής ανάλυσης (λεπτομέρειες στην Ενότητα 5.2.4), και επηρεάζει το περιεχόμενο των παραδειγμάτων εκπαίδευσης.

Στέλνοντας κάθε παράθυρο της πυραμίδας στο ΣΝΔ καταλήγουμε σε μία πυραμίδα «θερμικών χαρτών» (heat-map), δηλαδή χαρτών πιθανοτήτων για την ύπαρξη προσώπων. Οι διαστάσεις του κάθε επιπέδου εξαρτώνται (ανάλογα) από τις διαστάσεις του αντίστοιχου επιπέδου στην πυραμίδα εικόνων. Για παράδειγμα, μία θέση σε κάποιο επίπεδο, δίνει την πιθανότητα προσώπου στο παράθυρο 227×227, που αντιστοιχεί σε αυτή τη θέση, στο αντίστοιχο επίπεδο της πυραμίδας εικόνων. Για την αντιστοίχιση θέσεων, υπάρχει μία απεικόνιση των θέσεων της πυραμίδας χαρτών στην πυραμίδα εικόνων, που προκύπτει από το βήμα εξαγωγής παραθύρων (εξετάζεται παρακάτω).

Στη συνέχεια, η πυραμίδα θερμικών χαρτών επεξεργάζεται από έναν αλγόριθμο non maximum suppresion (NMS) για να αγνοηθούν οι θέσεις (και κατ' επέκταση τα αντίστοιχα παράθυρα), που έχουν σφάλμα ή επικαλυπτόμενα παράθυρα. Ο αλγόριθμος προσπαθεί να βρει ποιες είναι οι σωστές περιοχές των προσώπων και πόσα πρόσωπα υπάρχουν. Δεν υπάρχει κάποιος «βέλτιστος» τρόπος για να γίνει αυτό και οι τεχνικές είναι ευρετικές. Από αυτό αντιλαμβανόμαστε ότι η ευκρίνεια των αποτελεσμάτων του ΣΝΔ επηρεάζουν άμεσα τον αλγόριθμο. Οι μέθοδοι που χρησιμοποιούν οι συγγραφείς είναι 2:

- NMS μεγίστου. Σύμφωνα με αυτή την τεχνική επιλέγεται το παράθυρο με τη μέγιστη πιθανότητα προσώπου και απομακρύνονται όλα τα γειτονικά του με τα οποία επικαλύπτεται πάνω από ένα κατώφλι.

- NMS μέσου όρου. Αρχικά, απομακρύνονται όλα τα παράθυρα με πιθανότητα προσώπου κάτω από κάποιο κατώφλι (π.χ. 0.2). Στη συνέχεια, ομαδοποιούνται τα παράθυρα βάσει κάποιου κατωφλιού επικάλυψης και από κάθε ομάδα αφαιρούνται τα παράθυρα με πιθανότητα προσώπου μικρότερη από κάποιο κατώφλι (π.χ. 0.9) της μέγιστης πιθανότητας κάθε ομάδας. Βρίσκεται ο μέσος όρος των τεσσάρων συντεταγμένων των παραθύρων για κάθε ομάδα και για πιθανότητα του δημιουργημένου παραθύρου θεωρείται η μέγιστη της ομάδας.

[2] Το αυθεντικό δίκτυο έχει διαστάσεις εισόδου 224×224, εδώ χρησιμοποιείται αντ' αυτού το ισοδύναμο δίκτυο που παρέχεται από το Caffe.

Αποδοτική Υλοποίηση της αναζήτησης με κυλιόμενο παράθυρο

Στην Ενότητα 3.2.4 περιγράφηκε πως μετατρέπεται ένα πλήρως συνδεδεμένο επίπεδο σε ένα συνελικτικό επίπεδο. Συνοπτικά, οι νευρώνες του πλήρως συνδεδεμένου επιπέδου δέχονται ένα μονοδιάστατο διάνυσμα χαρακτηριστικών και εκτελούν την πράξη του εσωτερικού γινομένου με τα βάρη τους, περνώντας το αποτέλεσμα από τη συνάρτηση ενεργοποίησης. Κατά τη λειτουργία αυτή, το πλήθος των στοιχείων του διανύσματος εισόδου περιορίζεται να είναι ίδιο με το πλήθος των βαρών του κάθε νευρώνα. Μετά τη μετατροπή σε συνελικτικό επίπεδο η λειτουργία του επιπέδου παραμένει ίδια (εσωτερικό γινόμενο και ενεργοποίηση), ενώ εξαλείφεται ο περιορισμός της διάστασης εισόδου.

Μετά τη μετατροπή, το ΣΝΔ αγνοεί τη διάσταση της εικόνας εισόδου, αφού αποτελείται μόνο από συνελικτικά επίπεδα, επίπεδα συγκέντρωσης και επίπεδα κανονικοποίησης. Επίσης, αυτή η μετατροπή δίνει την δυνατότητα να χρησιμοποιηθούν υπεραποδοτικές υλοποιήσεις της συνέλιξης σε πολυπύρηνες αρχιτεκτονικές, χωρίς να περιορίζεται η διάσταση της εικόνας εισόδου σε 227×227.

Κατ' αυτόν τον τρόπο περνώντας μία εικόνα τυχαίων διαστάσεων από το ΣΝΔ, καταλήγουμε σε μία «πυκνή έξοδο», μικρότερης διάστασης από την είσοδο, όπου η κάθε θέση αντιστοιχεί σε ένα παράθυρο 227×227 στην αρχική εικόνα. Μέχρι στιγμής δεν έχουμε θίξει τη σημαντική λεπτομέρεια του βήματος εξαγωγής. Από την Ενότητα 3.4.1 υπενθυμίζουμε ότι το πέμπτο επίπεδο AlexNet έχει οπτικό πεδίο 32×32 στην εικόνα εισόδου διαστάσεων 227×227. Δεδομένου αυτού, συμπεραίνουμε ότι κάθε θέση στην πυκνή έξοδο υπολογίζεται με βήμα 32 πίξελ στην αρχική εικόνα.

Η διαδικασία που περιγράφηκε παραπάνω αντιστοιχεί ακριβώς στο σκεπτικό κυλιόμενου παραθύρου με βήμα 32 πίξελ. Επομένως, αντί να στέλνουμε κάθε παράθυρο ανεξάρτητα σε ένα κλασικό AlexNet, στέλνουμε όλη την εικόνα σε ένα τροποποιημένο AlexNet πυκνής εξόδου. Λόγω του βήματος, πρόσωπα που είναι κοντύτερα από 32 πίξελ δεν μπορούν να εντοπιστούν ως διαφορετικά.

5.2.2 Κατασκευή Συνόλου Δεδομένων για την Εκπαίδευση

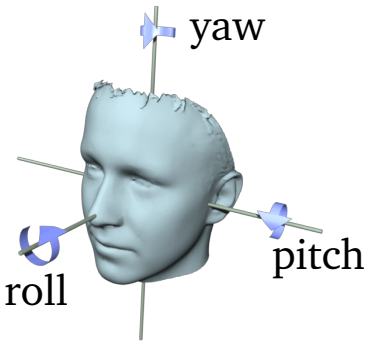
Ως σύνολο δεδομένων εκπαίδευσης χρησιμοποιείται η βάση δεδομένων AFLW [Koes11], που αποτελείται από 21,000 εικόνες με 24,000 επισημάνσεις προσώπων, που δίνονται ως συντεταγμένες παραθύρου που τα περιέχει (4 σημεία). Για να αυξηθούν τα παραδείγματα προσώπων οι συγγραφείς εφαρμόζουν τις εξής τεχνικές επαύξησης:

1. Τυχαία απομόνωση/κόψιμο (crop) παραθύρων από τις αρχικές εικόνες. Για να θεωρηθεί ένα παράθυρο ότι περιέχει πρόσωπο χρησιμοποιείται το κριτήριο (Intersection over Union – IoU) με κατώφλι 50%.

2. Επιπλέον, γίνεται αναστροφή (καθρεπτισμός) στις εικόνες από το βήμα 1 για διπλασιασμό των παραδειγμάτων.

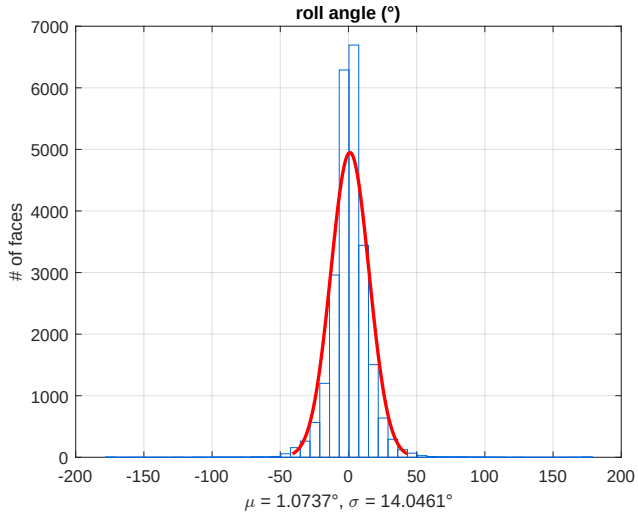
Εφαρμόζοντας αυτές τις τεχνικές, το σύνολο δεδομένων καταλήγει να έχει 200,000 εικόνες με πρόσωπα (θετικά παραδείγματα) και 20,000,000 εικόνες αρνητικών παραδειγμάτων, σε διάφορες διαστάσεις. Επειδή το AlexNet εκπαιδεύεται με συγκεκριμένο μέγεθος εισόδου 227×227, αυτές αναπροσαρμόζονται στις αναγκαίες διαστάσεις.

Όπως αναφέρθηκε στην Ενότητα 3.3.4 η συσχέτιση των εικόνων του συνόλου δεδομένων πρέπει να διερευνάται ενδελεχώς και ιδιαίτερα στην περίπτωση που γίνεται τεχνητή επαύξηση. Επιπλέον, οι



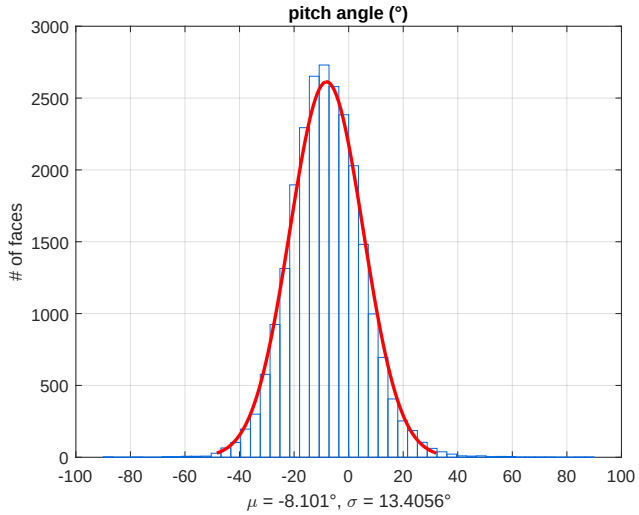
Σχήμα 5.3: Γωνίες θέασης προσώπου: γωνία στροφής στο επίπεδο της εικόνας (roll), γωνία πρόνευσης (pitch) και γωνία εκτροπής (yaw). Πηγή: [Koes11].

δύο κατηγορίες του συνόλου εικόνων δεν είναι ισορροπημένες· υπάρχουν 100 φορές περισσότερες εικόνες χωρίς πρόσωπα. Η αναλογία των εικόνων σε κάθε πακέτο εκπαίδευσης θα πρέπει επίσης να επιλέγεται προσεκτικά, ώστε να εκπαιδευτεί σωστά το δίκτυο.

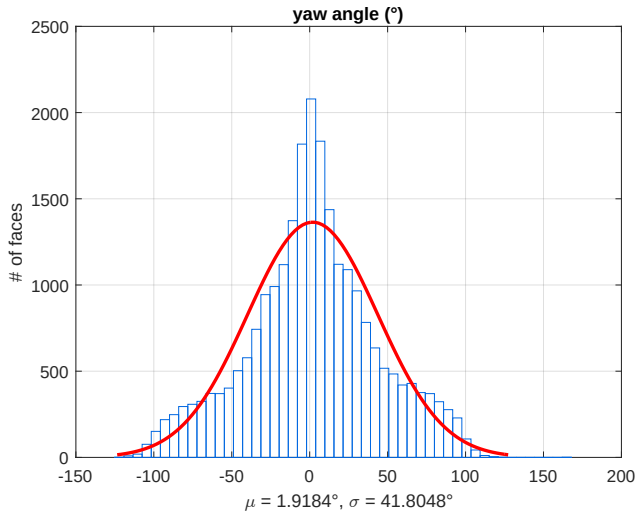


Σχήμα 5.4: Ιστόγραμμα γωνίας στροφής προσώπων στο επίπεδο της εικόνας.

Στο Σχήμα 5.4 φαίνεται η κατανομή των προσώπων ως προς τη γωνία, για περιστροφή στο επίπεδο της εικόνας. Παρατηρούμε ότι πρόσωπα σε γωνίες μεγαλύτερες από 30° υπάρχουν πολύ λίγα, περίπου 50 και με την επαύξηση γύρω στα 500. Είναι επομένως αναμενόμενο τέτοια πρόσωπα να είναι δυσκολότερο να ανιχνευθούν, ενώ σε περίπτωση που βρεθούν, ο δείκτης εμπιστοσύνης (confidence) θα είναι χαμηλός. Παρομοίως στα Σχήματα 5.5 και 5.6 φαίνονται οι κατανομές των προσώπων ως προς τη γωνία, για περιστροφή εκτός επιπέδου εικόνας, πάνω-κάτω και δεξιά-αριστερά, αντίστοιχα. Χρησιμοποιώντας ένα ασύμμετρο σύνολο δεδομένων σαν κι αυτό αναμένουμε το ΣΝΔ να είναι προκατειλημμένο προς συγκεκριμένες αποφάσεις και να αγνοεί κάποια πρόσωπα.



Σχήμα 5.5: Ιστόγραμμα γωνίας πρόνευσης (πάνω-κάτω) προσώπων.



Σχήμα 5.6: Ιστόγραμμα γωνίας εκτροπής (δεξιά-αριστερά) προσώπων.

5.2.3 Ειδική Προσαρμογή του AlexNet και μετατροπή για «πυκνή έξοδο»

Αρχικά, τροποποιείται ο softmax ταξινομητής του σταδίου εξόδου του AlexNet, ώστε αντί για 1000 κατηγορίες να έχει 2 (ύπαρξη ή όχι προσώπου). Το νέο δίκτυο επανεκπαιδεύεται για 50,000 επαναλήψεις (σύμφωνα με την ορολογία του Caffe) που υπολογίζεται σε περίπου μία εποχή (ή κάτι λιγότερο).[3]. Η εκπαίδευση γίνεται με τον αλγόριθμο Στοχαστικής Κατάβασης Δυναμικού (SGD) (βλ. Ενότητα 3.3.7), ο οποίος χρησιμοποιεί πακέτα εικόνων(mini-batches) για τη βελτιστοποίηση. Το σύνολο δεδομένων είναι πολωμένο τόσο ως προς τις γωνίες των προσώπων, όσο και ως προς το πλήθος των παραδειγμάτων για τις δύο κατηγορίες. Για τον δεύτερο λόγο οι συγγραφείς επιλέγουν να χρησιμοποιήσουν σε κάθε πακέτο 128 εικόνες, εκ των οποίων 32 με πρόσωπα (θετικά δείγματα) και 96 χωρίς (αρνητικά).

Στη συνέχεια για αποδοτική υλοποίηση της ανίχνευσης με κυλιόμενο παράθυρο, μετατρέπονται

[3] Για τον τύπο υπολογισμού βλ. Κεφάλαιο 4.1: $e = (b \cdot i)/t$

τα 3 τελευταία πλήρως-συνδεδεμένα επίπεδα, σε συνελικτικά, σύμφωνα με την Ενότητα 3.2.4, απλώς αλλάζοντας την πράξη του εσωτερικού γινομένου με τη γενίκευσή της, τη συνέλιξη. Με αυτήν την αλλαγή είναι δυνατόν να δώσουμε στο ΣΝΔ εικόνα οποιασδήποτε διάστασης, και να δώσει μία έξοδο για κάθε περιοχή 227×227 πίξελ με βήμα 32 πίξελ.

5.2.4 Ανίχνευση προσώπων

Μετά από τις παραπάνω αλλαγές τροφοδοτώντας στο ΣΝΔ μία εικόνα με τυχαίες διαστάσεις, το δίκτυο εξάγει αποτελέσματα για παράθυρα 227×227 πίξελ, με βήμα 32 πίξελ στις δύο χωρικές διαστάσεις εισόδου και έχει έξοδο ένα «θερμικό χάρτη» ή «χάρτη εμπιστοσύνης» με πιθανότητες για την ύπαρξη προσώπου.

Είναι επιθυμητό να ανιχνεύονται τα πρόσωπα σε όποια κλίμακα και αν υπάρχουν στις εικόνες. Μπορεί αυτά να καλύπτουν μικρή περιοχή μερικών πίξελ (π.χ. 45×45), αλλά μπορεί και να καταλαμβάνουν ολόκληρη την εικόνα (π.χ. 1000×1000) για ένα πορτρέτο. Το ΣΝΔ «βλέπει» μόνο περιοχές 227×227 πίξελ και αποφασίζει αν αυτές είναι πρόσωπα ή όχι. Για να γίνεται η ανίχνευση στην αρχική εικόνα σε οποιαδήποτε κλίμακα, είναι απαραίτητο να μεταβάλλουμε το οπτικό πεδίο του δικτύου. Αυτό μπορεί να γίνει κλιμακώνοντας τις εικόνες και κατασκευάζοντας μία πυραμίδα εικόνων. Οι συγγραφείς μετά από πειραματισμό προτείνουν να μεγεθυνθεί μία εικόνα 5 φορές, ώστε να είναι δυνατός ο εντοπισμός προσώπων από 45×45 πίξελ ($\approx 227/5$) και στη συνέχεια, να σμικρύνεται κατά παράγοντα 0.7937 μέχρι η ελάχιστη διάσταση της εικόνας να γίνει μικρότερη από 227. Μάλιστα, αυτές οι επιλογές δίνουν καλύτερα αποτελέσματα από άλλες που δημιουργούν περισσότερες κλίμακες.

Εφαρμόζοντας την παραπάνω μέθοδο δημιουργούμε μία πυραμίδα εικόνων εισόδων, η οποία στέλνεται στο ΣΝΔ, που εξάγει μία αντίστοιχη πυραμίδα θερμικών χαρτών (Σχήμα 5.2). Το επόμενο βήμα είναι να επεξεργαστεί κατάλληλα ο κάθε χάρτης, ώστε να εντοπιστούν με κάποιον τρόπο τα πρόσωπα. [4] Ο εντοπισμός πρέπει να γίνει στην αρχική εικόνα και όχι στις μεγεθυμένες ή σμικρυμένες εκδόσεις της, επομένως οι υποψήφιες θέσεις από την πυραμίδα θερμικών χαρτών, δηλαδή τα αντίστοιχα παράθυρα, πρέπει να αντιστοιχιστούν στις διαστάσεις της αρχικής εικόνας. Τέλος, λανθασμένα ή επαναλαμβανόμενα παράθυρα απομακρύνονται με χρήση μίας τεχνικής μη μέγιστης συμπίεσης (NMS) παραθύρων.

5.3 Αλλαγές στην Υλοποίηση και Βελτιώσεις στον DDFD

Στη δημοσίευση δεν αναφέρονται αποφάσεις σχεδίασης και αρκετοί παράμετροι του συστήματος, επομένως έπρεπε να καθοριστούν μέσω πειραματισμού. Οι σημαντικότερες από αυτές είναι:

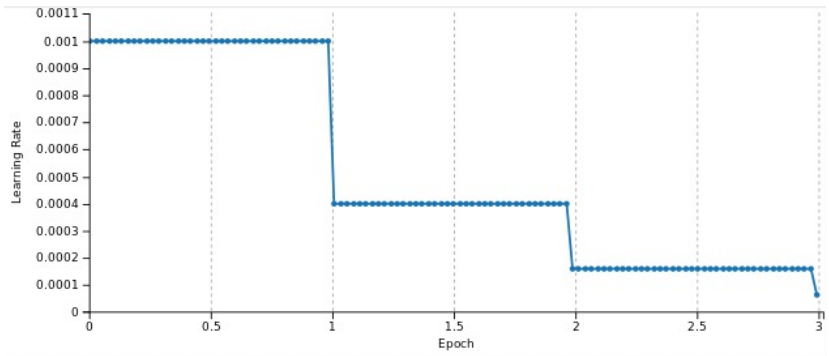
[4] Σχόλιο: Η πυκνή έξοδος του θερμικού χάρτη δίνει πολύ καλή πληροφορία για τη θέση, το μέγεθος και τα πίξελ δέρματος του προσώπου. Θα μπορούσε με κάποιο κατώφλι πιθανότητας να περιορίζονταν οι περιοχές και να προέκυπταν τα πρόσωπα. Ο στόχος του συστήματος είναι να εντοπίζει πρόσωπα και η μορφή εξόδου είναι ένας θερμικός χάρτης. Ωστόσο, για λόγους ποσοτικοποίησης και σύγκρισης αποτελεσμάτων με άλλες μεθόδους γίνεται εξαγωγή ευρεθέντων προσώπων με τη μορφή κουτιού που τα περιβάλλει (bounding box).

- Το σύνολο δεδομένων κατασκευάζεται από τυχαία crops από τις βάσεις δεδομένων AFLW [Koes11] και FaceScrub [Ng14] με συνολικά 100,000 εικόνες και με επαύξηση καταλήγουμε σε 1,000,000 εικόνες όπου το 1/3 περίπου είναι πρόσωπα. Για ευκολότερη ανίχνευση σε πολλαπλές γωνίες γίνεται καθρεπτισμός ως προς τους άξονες y (με πιθανότητα 1) και επιπλέον x (με πιθανότητα 0.3). Πρόσωπο θεωρείται όποιο crop έχει $IoU > 50\%$ και όχι πρόσωπο ότι έχει $5\% < IoU < 40\%$.

- Ο γενικός ρυθμός μάθησης του δικτύου υποδεκαπλασιάζεται (0.01 σε 0.001), ενώ ο τοπικός πολλαπλασιαστής του ρυθμού μάθησης του τελευταίου επιπέδου δεκαπλασιάζεται (1 σε 10), σε σχέση με τις αντίστοιχες τιμές του AlexNet (βλ. Κεφάλαιο 4.1 για την ορολογία και Παράρτημα C). Θέλουμε να κάνουμε ειδική προσαρμογή στο δίκτυο, επομένως ο ρυθμός μάθησης δεν θα πρέπει να είναι μεγάλος γιατί στις πρώτες επαναλήψεις, που θα παρουσιαστούν σχετικά διαφορετικά παραδείγματα στο δίκτυο, τα βάρη θα υποστούν σοβαρές μεταβολές. Ωστόσο, στο τελευταίο επίπεδο τα βάρη αρχικοποιούνται από την αρχή για να γίνει κατηγοριοποίηση ως προς τις 2 νέες κατηγορίες, επομένως ο τοπικός ρυθμός μάθησης πρέπει να μείνει υψηλός. Στο 5.7 φαίνεται το χρονοδιάγραμμα του γενικού ρυθμού μάθησης.

- Η ειδική προσαρμογή έγινε στο DIGITS για 3 εποχές ή 23280 επαναλήψεις με μέγεθος πακέτου 128 και αναλογία παραδειγμάτων 1/3 πρόσωπα – 2/3 όχι πρόσωπα. Το κόστος στα σύνολα εκπαίδευσης και επαλήθευσης, αλλά και η αποτελεσματικότητα της ταξινόμησης συναρτήσει των εποχών φαίνονται στο Σχήμα 5.8.

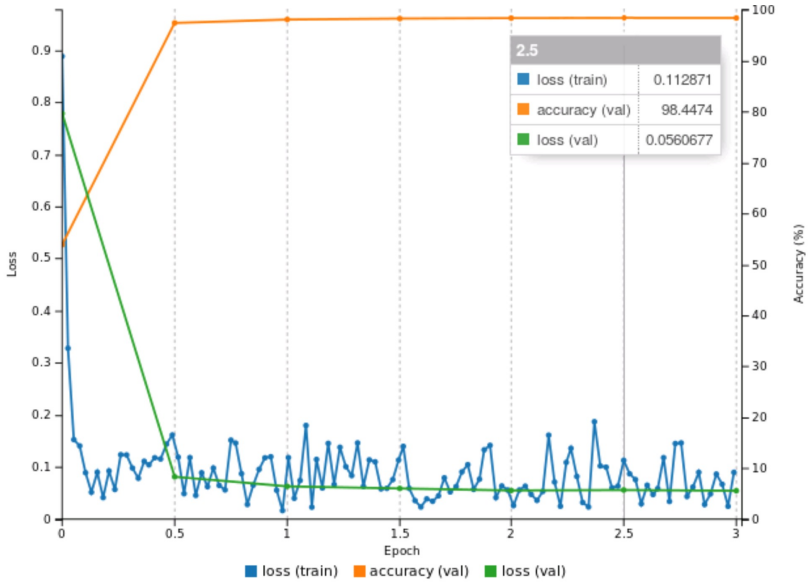
- Υλοποιείται το υποσύστημα NMS μέσου όρου με χρήση της συνάρτησης groupRectangles() του openCV. Πριν από οποιαδήποτε επεξεργασία καταστέλλονται παράθυρα με πιθανότητα προσώπου μικρότερη του 0.9. Η groupRectangles() αφού ομαδοποιήσει τα παράθυρα, καταργεί ομάδες με λιγότερα από 3 παράθυρα και εξάγει τα μέσα παράθυρα κάθε ομάδας.



Σχήμα 5.7: Χρονοδιάγραμμα γενικού ρυθμού μάθησης κατά την ειδική προσαρμογή.

Στη συνέχεια, θα αναφέρουμε κάποιες προτεινόμενες μελλοντικές τροποποιήσεις που θα βελτιώσουν την απόδοσή του και μπορούν να εφαρμοστούν στο παραπάνω σύστημα· γενικότερες κατευθύνσεις μελλοντικής έρευνας αναφέρονται στο Κεφάλαιο 6:

- *Βελτίωση των Τεχνικών Επαύξησης Δεδομένων.* Στα ΒΝΔ η δυσκολία δεν έγκειται στον καλό σχεδιασμό του συστήματος, αλλά στην κατασκευή ενός πλούσιου συνόλου δεδομένων με όσο



Σχήμα 5.8: Κόστος συνόλων εκπαίδευσης και επαλήθευσης και επίδοση στο σύνολο επαλήθευσης συναρτήσει των εποχών κατά τη διάρκεια ειδικής προσαρμογής του ΣΝΔ για την ταξινόμηση εικόνας ως πρόσωπο ή όχι.

το δυνατόν καλύτερες ιδιότητες. Για τη βελτίωση της ποικιλομορφίας του συνόλου εκπαίδευσης μπορούν (εκτός από κατοπτρισμοί) να χρησιμοποιηθούν τυχαίοι αφινικοί μετασχηματισμοί (περιστροφές, κλιμακώσεις), ελαστικοί μετασχηματισμοί (όπως στο LeNet) και τεχνητές επικαλύψεις. Επίσης, αν υπάρχουν πληροφορίες για τις γωνίες περιστροφής των προσώπων, τότε η καλύτερη επιλογή είναι να χρησιμοποιηθεί εξειδικευμένη πληθυσμιακή επαύξηση, δηλαδή να αυξηθούν περισσότερο πρόσωπα σε σπάνιες γωνίες, δηλαδή να γίνουν πιο ομοιόμορφες οι κατανομές 5.4 - 5.6.

- *Βελτίωση της αναλογίας παραδειγμάτων στο πακέτο εκπαίδευσης.* Κατά την εκπαίδευση το minibatch θα πρέπει να περιέχει όσο περισσότερες διακυμάνσεις στην πόζα του προσώπου και στην ποικιλία εικόνων (ασυσχέτιστες), ενώ η αναλογία προσώπων - όχι προσώπων πρέπει να είναι προσεκτικά επιλεγμένη.

- *Επανεκπαίδευση με λάθος κατηγοριοποιημένα παραδείγματα.* Αφότου εκπαιδευτεί το δίκτυο, είναι δυνατόν να το δοκιμάσουμε σε όλα τα δεδομένα (εκπαίδευσης, επαλήθευσης και δοκιμής) και να χρησιμοποιήσουμε αυτά που ταξινομεί λανθασμένα για να το ξανά εκπαιδεύσουμε.

- *Εκπαίδευση με εικόνες που «απατούν» το ΣΝΔ.* Χρησιμοποιώντας τεχνικές που αναπτύχθηκαν στα [Nguy14, Good14, Szeg13], π.χ. ελαχιστοποίηση κόστους και οπισθοδιάδοση ως προς τα δεδομένα μπορούμε να δημιουργήσουμε τεχνητές εικόνες που ξεγελούν το ΣΝΔ και να το εκπαιδεύσουμε με αυτές.

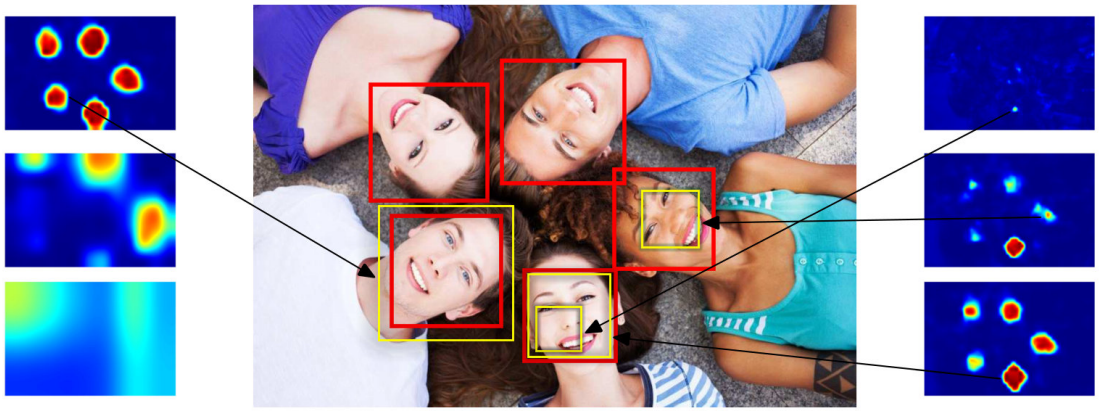
- *Αλλαγές στα επίπεδα του ΣΝΔ.* Η επίδοση της ταξινόμησης μπορεί να βελτιωθεί χρησιμοποιώντας παραμετροποιήσιμες ReLU [He15], ώστε να αποφύγουμε το πρόβλημα των «νεκρών» νευρώνων και μονάδες maxout [Good13] για πιο σύνθετες συναρτήσεις ενεργοποίησης (βλ. Ενότητα 2.2). Επιπλέον, η στοχαστική συγκέντρωση [Zeil13], που υλοποιήθηκε βελτιώνει όντως

τα αποτελέσματα, εφαρμόζοντας κανονικοποίηση στα επίπεδα συγκέντρωσης (εκτός από το dropout) (βλ. Ενότητα 3.3.5).

- *Χρήση άλλων τεχνικών NMS.* Οι ευρετικές τεχνικές, για την καταπίεση παραθύρων, που μπορούν να κατασκευαστούν είναι σίγουρα πάρα πολλές. Καλές τεχνικές θεωρούνται αυτές που είναι γρήγορες, έχουν λίγες και όχι ευαίσθητες παραμέτρους. Μία τέτοια νέα τεχνική εισάγεται στο [Roth15], η οποία κατά βάση στηρίζεται στην μεταφορά των παραθύρων σε έναν χώρο ομοιότητας, στην μεταφορά μηνυμάτων και εν τέλει στην ομαδοποίηση των παραθύρων.

5.4 Αποτελέσματα και Ανάλυση του ΣΝΔ

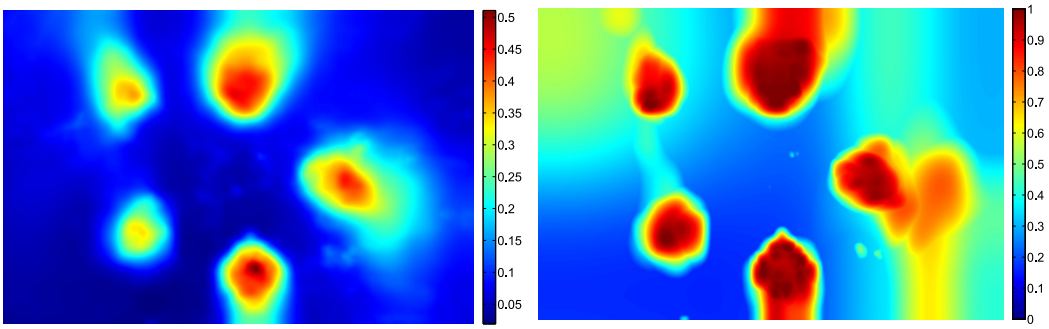
Ιδανικά από το ΣΝΔ ταξινόμησης παραθύρων προσώπων αναμένουμε να δίνει μεγάλες πιθανότητες σε παράθυρα που περικλείουν ακριβώς ένα πρόσωπο και πολύ μικρές πιθανότητες σε οποιαδήποτε άλλα παράθυρα. Κατ' αυτόν τον τρόπο θα γίνεται ανίχνευση ενός προσώπου μόνο στις κλίμακες που αντιστοιχούν με την κλίμακά του, ενώ στις υπόλοιπες κλίμακες δεν θα πρέπει να επιστρέφονται μεγάλες πιθανότητες. Επίσης, επιθυμούμε οι ακμές των περιοχών ανίχνευσης να είναι όσο το δυνατόν λιγότερο θολές, ώστε τα παράθυρα να είναι διακριτά, ώστε να μην δημιουργούν επιπλοκές στο σύστημα NMS μέσου όρου (όπως στο Σχήμα A.2). Στο Σχήμα 5.9 φαίνεται η διαδικασία εντοπισμού και συμπεραίνουμε ότι οι παραπάνω προδιαγραφές επιτυγχάνονται σε σημαντικό βαθμό.



Σχήμα 5.9: Αποτελέσματα εντοπισμού σε μία απαιτητική εικόνα. Φαίνονται επίσης 6 από τις 12 κλίμακες της πυραμίδας θερμικών χαρτών και η κλίμακα παραθύρων στην οποία αντιστοιχεί ένα πίξελ από κάθε επίπεδο της πυραμίδας.

Για να εξεταστεί η ευκρίνεια και η οξύτητα των χαρτών χαρακτηριστικών επιλέγονται οι διαστάσεις ενός επιπέδου της πυραμίδας (συνήθως του μεσαίου) και μεγεθύνονται ή σμικρύνονται οι υπόλοιποι χάρτες σε αυτό το μέγεθος. Στη συνέχεια, μπορούν να υπολογιστούν στατιστικά στοιχεία, όπως ο μέσος όρος και το μέγιστο της πυραμίδας χαρτών. Στο Σχήμα 5.10 φαίνονται οι δύο χάρτες μέσου όρου και μεγίστου.

Το δυαδικό πρόβλημα ταξινόμησης προσώπου έχει κάποια ιδιαίτερα χαρακτηριστικά. Επίσης, η επιλογή να γίνει ειδική προσαρμογή του AlexNet και όχι εκπαίδευση εξ αρχής σε μικρότερο ίσως



Σχήμα 5.10: Αριστερά: Μέσος όρος θερμικών χαρτών. Δεξιά: Θερμικός χάρτης μεγίστων.

δίκτυο, λόγω των δύο μόνο κατηγοριών, έχουν επιδράσεις, όχι απαραίτητα αρνητικές, στα αποτελέσματα. Μία από αυτές είναι η μορφή της καμπύλης κόστους ανά εποχή (Σχήμα 5.8). Το κόστος ξεκινάει από μία λογική τιμή 0.9 και πέφτει πολύ γρήγορα γύρω στο 0.1 στις πρώτες 1000 επαναλήψεις, ενώ κατά τη διάρκεια της εκπαίδευσης ταλαντεύεται γύρω από αυτή την τιμή και μειώνεται ελάχιστα. Η συμπεριφορά αυτή οφείλεται ως επί το πλείστον στο πλήθος των κατηγοριών (δύο)· στις πρώτες 100 επαναλήψεις στο δίκτυο παρουσιάζονται 12,800 εικόνες με περίπου 4,300 πρόσωπα, τα οποία είναι αρκετά σύμφωνα με το διάγραμμα για τη δυαδική απόφαση. Αν και η ειδική προσαρμογή θα μπορούσε να σταματήσει στις δύο εποχές, επιλέγουμε να σταματήσει στην 3η ώστε το κάθε παράδειγμα να περάσει τουλάχιστον 2 φορές από το δίκτυο.

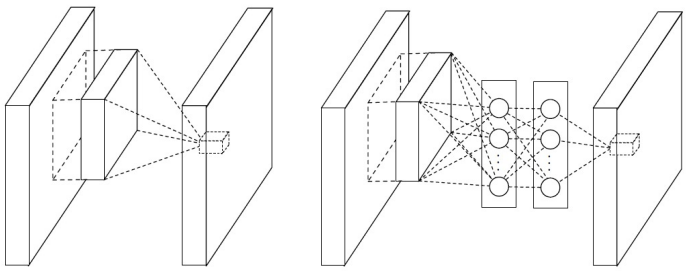
Λόγω της φύσης του προβλήματος, επίσης παρατηρούμε από τα πρώτα επίπεδα το δίκτυο να εξάγει γενικευμένες σημασιολογικές περιγραφές, κάτι που στο AlexNet γινόταν προς τα τελευταία συνελικτικά επίπεδα. Οι πρώτοι όγκοι χαρακτηριστικών (Σχήματα B.2, B.3) έχουν πολλούς χάρτες ανίχνευσης ακμών σε διάφορες διευθύνσεις και άλλων χαρακτηριστικών χαμηλού επιπέδου (π.χ. γωνίες), αλλά υπάρχουν και χάρτες που περιγράφουν το προσκήνιο και το παρασκήνιο της εικόνας, περιοχές όπου θα πρέπει να αναζητηθεί ή όχι πρόσωπο και χάρτες που ξεχωρίζουν όμοια αντικείμενα από το περιβάλλον τους (π.χ. άνθρωποι, μέρη προσώπων, δέρμα, ρούχα (με ομοιόμορφη υφή))· τρεις επιλογές φαίνονται στο Σχήμα 5.11. Τέλος, σε υψηλότερα επίπεδα παρατηρούμε αρκετά εμφανές το πρόβλημα των ανενεργών νευρώνων που αναλύθηκε στην παράγραφο ReLU της Ενότητας 2.2.

Σχήμα 5.11: Επιλεγμένοι χάρτες χαρακτηριστικών από το πρώτο επίπεδο κανονικοποίησης.

Κεφάλαιο 6

Μελλοντική Έρευνα

Κατά το σχεδιασμό μίας βαθιάς αρχιτεκτονικής πρέπει να προεπιλεγούν πολλές υπερπαράμετροι, όπως το πλήθος των επιπέδων, οι περιοχές των φίλτρων, η επικάλυψη, το πλήθος των νευρώνων, κτλ. Αντίθετα, σε ένα «ρηχό» αλγόριθμο, η επιλογή των καλύτερων τέτοιων παραμέτρων είναι πολύ πιο εύκολη απ' ότι στα ΒΝΔ. Για να αντιμετωπιστεί η δυσκολία επιλογής υπερπαραμέτρων στο [MinL13] προτείνεται η εισαγωγή «μικρο-δικτύων» μετά τα συνελικτικά επίπεδα (Network in Network – NiN), δηλαδή γίνεται αντικατάσταση του γενικευμένου γραμμικού φίλτρου με ένα μη γραμμικό φίλτρο. Αυτό βελτιώνει την αφαιρετική ικανότητα των συνελικτικών επιπέδων και μειώνει τις απαιτούμενες υπερπαραμέτρους, αν και πρέπει να καθοριστεί ένα νέο σύνολο υπερπαραμέτρων για το νέο δίκτυο.



Σχήμα 6.1: Αρχιτεκτονική Network in Network.

Ο DDFD είναι ότι αποτελείται από δύο υποσυστήματα που είναι αλληλεξαρτώμενα. Για να αντιμετωπιστεί αυτή η εξάρτηση, αλλά και να μειωθεί η ανάγκη πολυκλιμακωτής ανάλυσης, μπορεί να χρησιμοποιηθεί ένα Πλήρως Συνελικτικό Δίκτυο (FCN) [Long14], το οποίο εκπαιδεύεται από την αρχή έως το τέλος (end-to-end) με εικόνες. Βέβαια, σε αυτή την περίπτωση πρέπει να δημιουργηθούν οι απαιτούμενοι στόχοι εξόδου, οι οποίοι αντί για ετικέτες θα είναι εικόνες (δυαδικές μάσκες). Κατ' αυτό τον τρόπο, επιτυγχάνεται ακόμα πυκνότερη ταξινόμηση, στις διαστάσεις εισόδου.

Μία άλλη επιλογή, που έχει χρησιμοποιηθεί πολλές φορές στο παρελθόν είναι η εκπαίδευση πολλαπλών δικτύων και ο συνδυασμός των αποτελεσμάτων τους με τεχνικές συγχώνευσης (fusion, model ensembles). Στα ΣΝΔ αυτό δεν είναι ίσως εφικτό ακόμα λόγω του μεγάλου χρόνου εκπαίδευσης και δοκιμής τους. Τέλος, τα ΣΝΔ αποτελούν καλούς εξαγωγείς χαρακτηριστικών με σημαντικές ιδιότητες αναλλοίωτου κυρίως σε μεταφορά, αλλά και περιστροφή ή κλιμάκωση, επομένως, εξάγοντας τα συνελικτικά χαρακτηριστικά με χρήση κάποιου δικτύου (π.χ. VGG Net) μπορούν να χρησιμοποιηθούν σε άλλους ταξινομητές, π.χ. οι SVM, οι οποίοι, όπως υποστηρίζεται στο [Tang13], έχουν καλύτερα αποτελέσματα σε σχέση με τον softmax.

Βιβλιογραφία

[Beng15] Yoshua Bengio, Ian J. Goodfellow and Aaron Courville, "Deep Learning". Book in preparation for MIT Press, 2015. 24, 54, 56, 57, 60, 61

[Bour10] Y-Lan Boureau, Jean Ponce and Yann Lecun, "A Theoretical Analysis of Feature Pooling in Visual Recognition", in *27th International Conference on Machine Learning, Haifa, Israel*, 2010. 43

[Cybe89] George Cybenko, "Approximation by superpositions of a sigmoidal function", *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989. 36

[DiCa08] James J DiCarlo, Nicolas Pinto and David Daniel Cox, "Why is Real-World Visual Object Recognition Hard?", 2008. 51

[Garc04] Christophe Garcia and Manolis Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 11, pp. 1408–1423, 2004. 79, 80

[Glor10] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks", in *International conference on artificial intelligence and statistics*, pp. 249–256, 2010. 66

[Good13] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville and Yoshua Bengio, "Maxout networks", *arXiv preprint arXiv:1302.4389*, 2013. 23, 87

[Good14] Ian J Goodfellow, Jonathon Shlens and Christian Szegedy, "Explaining and harnessing adversarial examples", *arXiv preprint arXiv:1412.6572*, 2014. 87

[He15] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", *arXiv preprint arXiv:1502.01852*, 2015. 22, 66, 87

[Hint12] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors", *arXiv preprint arXiv:1207.0580*, 2012. 68, 69

[Hjel01] Erik Hjelmas and Boon Kee Low, "Face detection: A survey", *Computer vision and image understanding*, vol. 83, no. 3, pp. 236–274, 2001. 77

[Ho65] Yu-Chi Ho and RL Kashyap, "An algorithm for linear inequalities and its applications", *IEEE Transactions on Electronic Computers*, vol. 5, no. EC-14, pp. 683–688, 1965. 17

[Huan15] Lichao Huang, Yi Yang, Yafeng Deng and Yinan Yu, "DenseBox: Unifying Landmark Localization with End to End Object Detection", *arXiv preprint arXiv:1509.04874*, 2015. 79

[Ioff15] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", *arXiv preprint arXiv:1502.03167*, 2015. 39, 66

[Jia 14] Jia, Yangqing and Shelhamer, Evan and Donahue, Jeff and Karayev, Sergey and Long, Jonathan and Girshick, Ross and Guadarrama, Sergio and Darrell, Trevor, "Caffe: Convolutional Architecture for Fast Feature Embedding", *arXiv preprint arXiv:1408.5093*, 2014. 73

[Karp15] Andrej Karpathy and Fei-Fei Li, "Stanford University, CS231n Course Notes: Convolutional Neural Networks for Visual Recognition", January - March, 2015. 24, 48, 52, 56, 61

[Koes11] Martin Koestinger, Paul Wohlhart, Peter M. Roth and Horst Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization", in *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 82, 83, 86

[Kriz12] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in neural information processing systems*, pp. 1097–1105, 2012. 37, 39, 51, 67, 68, 69

[LeCu98] Yann LeCun, Léon Bottou, Yoshua Bengio and Patrick Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 37, 67

[Long14] Jonathan Long, Evan Shelhamer and Trevor Darrell, "Fully convolutional networks for semantic segmentation", *arXiv preprint arXiv:1411.4038*, 2014. 91

[Lyu08] Siwei Lyu and Eero P Simoncelli, "Nonlinear image representation using divisive normalization", in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008. 51

[Maas13] Andrew L Maas, Awni Y Hannun and Andrew Y Ng, "Rectifier nonlinearities improve neural network acoustic models", in *Proc. ICML*, vol. 30, 2013. 22

[Math14] M. Mathias, R. Benenson, M. Pedersoli and L. Van Gool, "Face detection without bells and whistles", in *ECCV*, 2014. 78

[MinL13] Qiang Chen Min Lin and Shuicheng Yan, "Network In Network", in *International Conference on Learning Representations*, 2013. 91

[Ng14] Hong-Wei Ng and Stefan Winkler, "A data-driven approach to cleaning large face datasets", in *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 343–347, IEEE, 2014. 86

[Nguy14] Anh Nguyen, Jason Yosinski and Jeff Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images", *arXiv preprint arXiv:1412.1897*, 2014. 87

[Orr03] Genevieve B Orr and Klaus-Robert Müller, *Neural networks: tricks of the trade*, Springer, 2003. 54

[Osad07] Margarita Osadchy, Yann Le Cun and Matthew L Miller, "Synergistic face detection and pose estimation with energy-based models", *The Journal of Machine Learning Research*, vol. 8, pp. 1197–1215, 2007. 79, 80

[Roth15] Rasmus Rothe, Matthieu Guillaumin and Luc Van Gool, *Non-maximum suppression for object detection by passing messages between windows*, Springer, 2015. 88

[Rowl98] Henry Rowley, Shumeet Baluja, Takeo Kanade et al., "Neural network-based face detection", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 1, pp. 23–38, 1998. 14, 78, 79

[Russ14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein et al., "Imagenet large scale visual recognition challenge", *International Journal of Computer Vision*, pp. 1–42, 2014. 38

[Sach15] Sachin Sudhakar Farfade and Mohammad Saberian and Li-Jia Li, "Multi-view Face Detection Using Deep Convolutional Neural Networks", in *International Conference on Multimedia Retrieval (ICMR)*, Shanghai, China, June 2015. 77, 80

[Sriv14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. 56

[Szeg13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow and Rob Fergus, "Intriguing properties of neural networks", *arXiv preprint arXiv:1312.6199*, 2013. 87

[Tang13] Yichuan Tang, "Deep learning using linear support vector machines", *arXiv preprint arXiv:1306.0239*, 2013. 29, 91

[Tolb06] AS Tolba, AH El-Baz and AA El-Harby, "Face recognition: A literature review", *International Journal of Signal Processing*, vol. 2, no. 2, pp. 88–103, 2006. 77

[Vail94] Régis Vaillant, Christophe Monrocq and Yann Le Cun, "Original approach for the localisation of objects in images", *IEE Proceedings-Vision, Image and Signal Processing*, vol. 141, no. 4, pp. 245–250, 1994. 78

[Viol01] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features", in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–511, IEEE, 2001. 78

[West99] Jason Weston, Chris Watkins et al., "Support vector machines for multi-class pattern recognition.", in *ESANN*, vol. 99, pp. 219–224, 1999. 29

[Yang15] Shuo Yang, Ping Luo, Chen Change Loy and Xiaoou Tang, "From Facial Parts Responses to Face Detection: A Deep Learning Approach", *arXiv preprint arXiv:1509.06451*, 2015. 79

[Yosi14] Jason Yosinski, Jeff Clune, Yoshua Bengio and Hod Lipson, "How transferable are features in deep neural networks?", in Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 3320–3328, Curran Associates, Inc., 2014. 58, 59

[Zeil13] Matthew D Zeiler and Rob Fergus, "Stochastic pooling for regularization of deep convolutional neural networks", *arXiv preprint arXiv:1301.3557*, 2013. 57, 87

[Zhan10] Cha Zhang and Zhengyou Zhang, "A survey of recent advances in face detection", Technical report, Tech. rep., Microsoft Research, 2010. 77

[Zhao03] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips and Azriel Rosenfeld, "Face recognition: A literature survey", *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003. 77

Παράρτημα Α

Παραδείγματα Εντοπισμού

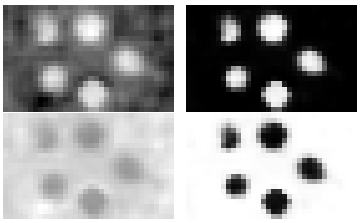
Σχήμα Α.1: Τα πρόσωπα εντοπίζονται ανεξαρτήτου γωνίας στροφής στο επίπεδο της κάμερας ή εκτός, με μερικές επικαλύψεις και με μεγάλη γωνίας απόκλισης από το επίπεδο της κάμερας.

Σχήμα Α.2: Σύγκριση υποσυστημάτων NMS. Αριστερή στήλη: NMS μέσου όρου: αγνοεί επιτυχώς λανθασμένες ομάδες παραθύρων, εντοπίζει όλα τα πρόσωπα, αλλά υστερεί στην ομαδοποίηση όταν τα πρόσωπα είναι κοντά. Δεξιά στήλη: NMS μεγίστου: έχει μικρότερη επιτυχία στον εντοπισμό προσώπων, αλλά ξεχωρίζει πρόσωπα που είναι κοντά.

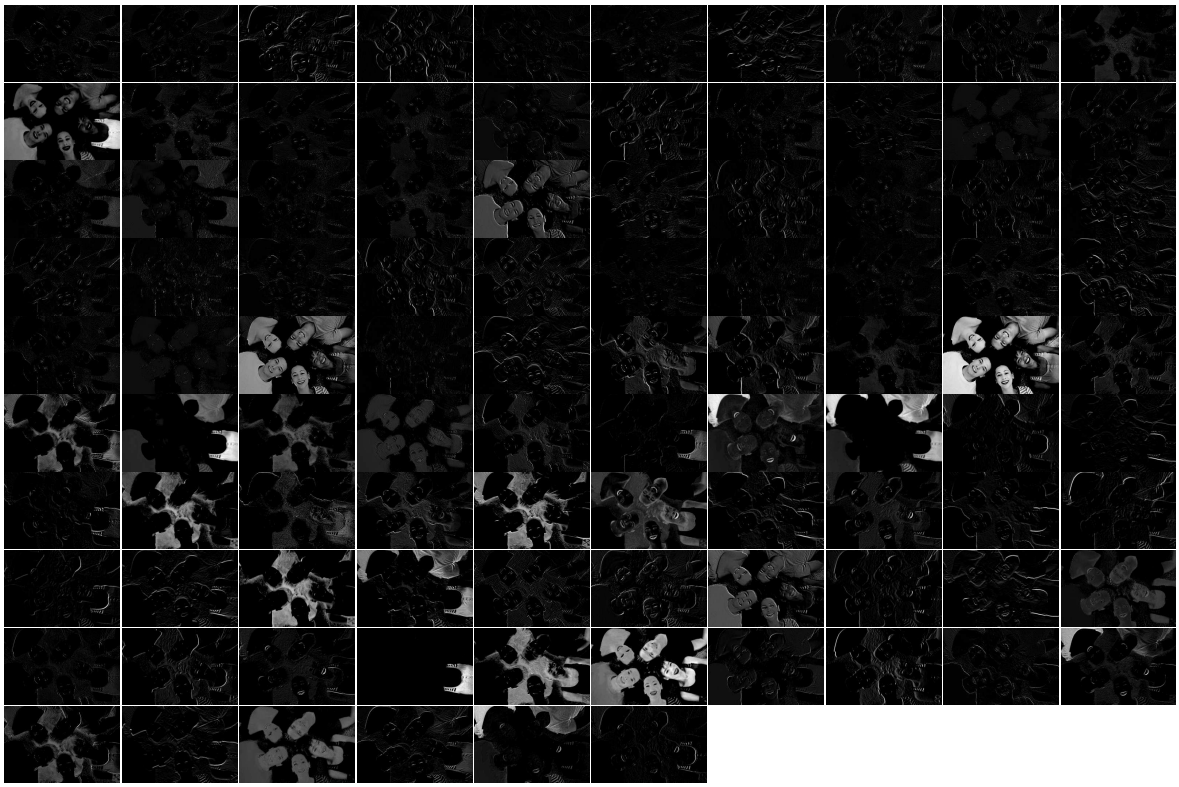
Σχήμα Α.3: Περιπτώσεις αποτυχίας. Αριστερά: [Μερική] αποτυχία του ανιχνευτή καθώς θεωρεί τις μάσκες πρόσωπα [που απεικονίζουν βέβαια πρόσωπα]. Δεξιά: Ο ανιχνευτής χάνει ένα πρόσωπο, γιατί το ΣΝΔ δίνει μικρή πιθανότητα ύπαρξης προσώπου στην περιοχή (< 0.6), που αποκόπτεται από το κατώφλι.

Παράρτημα Β

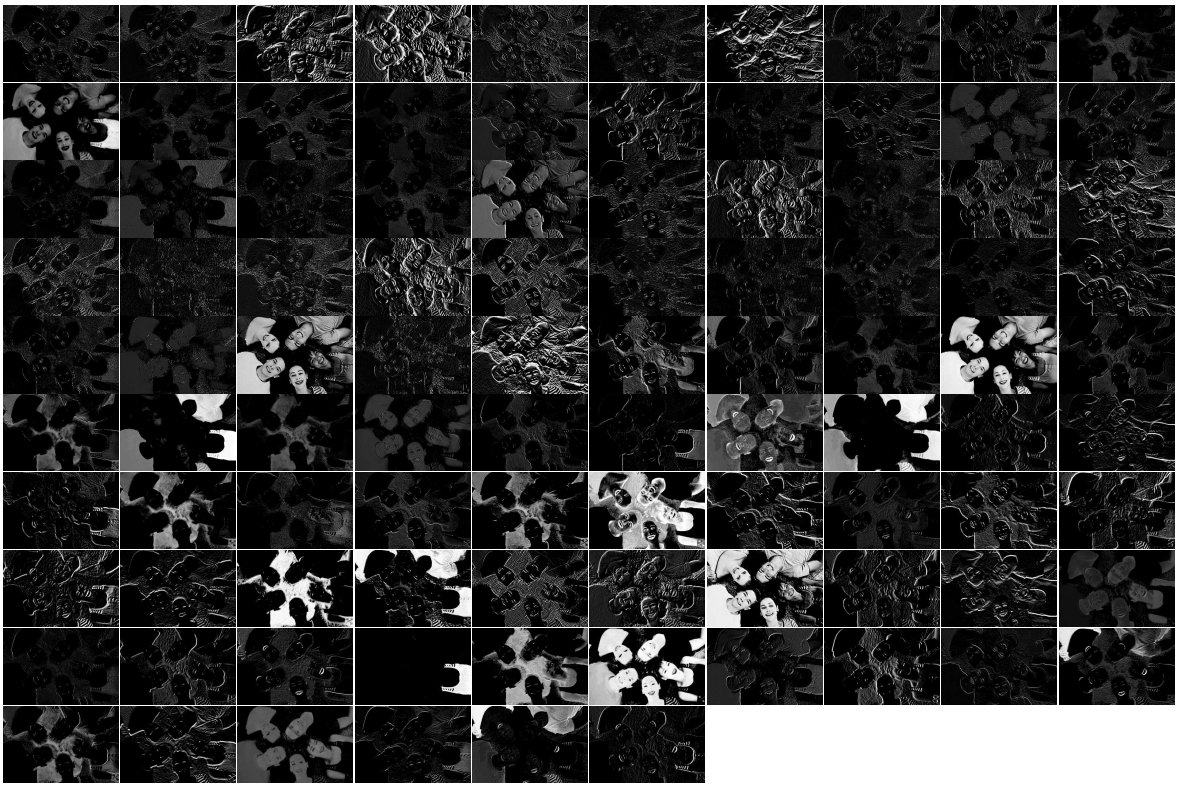
Όγκοι Χαρακτηριστικών



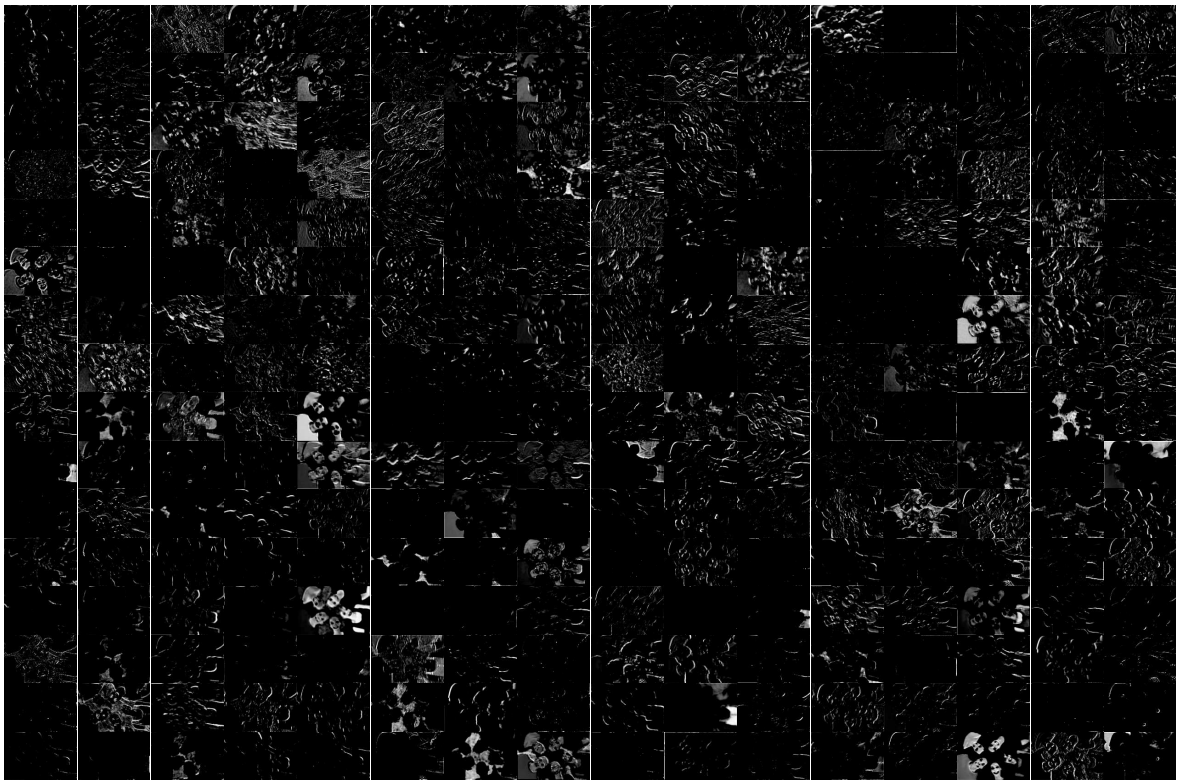
Σχήμα Β.1: Πυκνή έξοδος προτελευταίου και τελευταίου επιπέδου για εικόνα εισόδου 1283×867. Τα δύο κανάλια αντιστοιχούν στις δύο κατηγορίες πρόσωπο και όχι πρόσωπο. Φαίνεται πως επιτυγχάνεται ευκρίνεια και το αποτέλεσμα της softmax.



Σχήμα Β.2: Οι 96 χάρτες χαρακτηριστικών ακριβώς μετά το πρώτο συνελικτικό επίπεδο για εικόνα εισόδου 1283×867.



Σχήμα Β.3: Οι 96 χάρτες χαρακτηριστικών ακριβώς μετά το πρώτο επίπεδο κανονικοποίησης για εικόνα εισόδου 1283×867.



Σχήμα Β.4: Οι 256 χάρτες χαρακτηριστικών ακριβώς μετά το δεύτερο επίπεδο κανονικοποίησης για εικόνα εισόδου 1283×867.

Παράρτημα C

DDFD CNN deploy prototxt

```
name: ”FaceNet”
input: ”data”
input_dim: 1 # batch_size
input_dim: 3 # channels RGB = 3
input_dim: 300 # image height = width in caffe
input_dim: 400 # image width = height in caffe
layer {
  name: ”conv1”
  type: ”Convolution”
  bottom: ”data”
  top: ”conv1”
  param {
    lr_mult: 1.0
    decay_mult: 1.0
  }
  param {
    lr_mult: 2.0
    decay_mult: 0.0
  }
  convolution_param {
    num_output: 96
    kernel_size: 11
    stride: 4
    weight_filler {
      type: ”gaussian”
      std: 0.01
    }
    bias_filler {
      type: ”constant”
      value: 0.0
    }
  }
}
layer {
  name: ”relu1”
  type: ”ReLU”
  bottom: ”conv1”
  top: ”conv1”
  relu_param {
        negative_slope: 0.05
  }
}
layer {
  name: ”norm1”
  type: ”LRN”
  bottom: ”conv1”
  top: ”norm1”
  lrn_param {
    local_size: 5
    alpha: 0.0001
    beta: 0.75
  }
}
layer {
  name: ”pool1”
  type: ”Pooling”
  bottom: ”norm1”
  top: ”pool1”
  pooling_param {
    pool: MAX
    kernel_size: 3
    stride: 2
  }
}
layer {
  name: ”conv2”
  type: ”Convolution”
  bottom: ”pool1”
  top: ”conv2”
  param {
    lr_mult: 1.0
    decay_mult: 1.0
  }
  param {
    lr_mult: 2.0
    decay_mult: 0.0
  }
  convolution_param {
    num_output: 256
    pad: 2
    kernel_size: 5
    group: 2
    weight_filler {
      type: ”gaussian”
      std: 0.01
    }
    bias_filler {
      type: ”constant”
      value: 0.1
    }
  }
}
layer {
  name: ”relu2”
  type: ”ReLU”
  bottom: ”conv2”
  top: ”conv2”
  relu_param {
    negative_slope: 0.05
  }
}
layer {
  name: ”norm2”
  type: ”LRN”
  bottom: ”conv2”
  top: ”norm2”
  lrn_param {
    local_size: 5
    alpha: 0.0001
    beta: 0.75
  }
}
layer {
  name: ”pool2”
  type: ”Pooling”
  bottom: ”norm2”
```

```
  top: "pool2"
  pooling_param {
    pool: MAX
    kernel_size: 3
    stride: 2
  }
}
layer {
  name: "conv3"
  type: "Convolution"
  bottom: "pool2"
  top: "conv3"
  param {
    lr_mult: 1.0
    decay_mult: 1.0
  }
  param {
    lr_mult: 2.0
    decay_mult: 0.0
  }
  convolution_param {
    num_output: 384
    pad: 1
    kernel_size: 3
    weight_filler {
      type: "gaussian"
      std: 0.01
    }
    bias_filler {
      type: "constant"
      value: 0.0
    }
  }
}
layer {
  name: "relu3"
  type: "ReLU"
  bottom: "conv3"
  top: "conv3"
  relu_param {
    negative_slope: 0.05
  }
}
layer {
  name: "conv4"
  type: "Convolution"
  bottom: "conv3"
  top: "conv4"
  param {
    lr_mult: 1.0
    decay_mult: 1.0
  }
  param {
    lr_mult: 2.0
    decay_mult: 0.0
  }
  convolution_param {
    num_output: 384
    pad: 1
    kernel_size: 3
    group: 2
    weight_filler {
      type: "gaussian"
      std: 0.01
    }
    bias_filler {
      type: "constant"
      value: 0.1
    }
  }
}
layer {
  name: "relu4"
  type: "ReLU"
  bottom: "conv4"
  top: "conv4"
  relu_param {
    negative_slope: 0.05
  }
}
layer {
  name: "conv5"
  type: "Convolution"
  bottom: "conv4"
  top: "conv5"
  param {
    lr_mult: 1.0
    decay_mult: 1.0
  }
  param {
    lr_mult: 2.0
    decay_mult: 0.0
  }
  convolution_param {
    num_output: 256
    pad: 1
    kernel_size: 3
    group: 2
    weight_filler {
      type: "gaussian"
      std: 0.01
    }
    bias_filler {
      type: "constant"
      value: 0.1
    }
  }
}
layer {
  name: "relu5"
  type: "ReLU"
  bottom: "conv5"
  top: "conv5"
  relu_param {
    negative_slope: 0.05
  }
}
layer {
  name: "pool5"
  type: "Pooling"
  bottom: "conv5"
  top: "pool5"
  pooling_param {
    pool: MAX
    kernel_size: 3
    stride: 2
  }
}
layer {
  name: "fc6_conv"
  type: "Convolution"
  bottom: "pool5"
  top: "fc6_conv"
  param {
    lr_mult: 1.0
    decay_mult: 1.0
  }
  param {
    lr_mult: 2.0
    decay_mult: 0.0
  }
  convolution_param {
    num_output: 4096
    kernel_size: 6
    weight_filler {
      type: "gaussian"
      std: 0.005
    }
    bias_filler {
      type: "constant"
      value: 0.1
    }
  }
}
layer {
  name: "relu6"
  type: "ReLU"
  bottom: "fc6_conv"
```

```
  top: ”fc6_conv”
  relu_param {
    negative_slope: 0.05
  }
}
layer {
  name: ”drop6”
  type: ”Dropout”
  bottom: ”fc6_conv”
  top: ”fc6_conv”
  dropout_param {
    dropout_ratio: 0.5
  }
}
layer {
  name: ”fc7_conv”
  type: ”Convolution”
  bottom: ”fc6_conv”
  top: ”fc7_conv”
  param {
    lr_mult: 1.0
    decay_mult: 1.0
  }
  param {
    lr_mult: 2.0
    decay_mult: 0.0
  }
  convolution_param {
    num_output: 4096
    kernel_size: 1
    weight_filler {
      type: ”gaussian”
      std: 0.005
    }
    bias_filler {
      type: ”constant”
      value: 0.1
    }
  }
}
layer {
  name: ”relu7”
  type: ”ReLU”
  bottom: ”fc7_conv”
  top: ”fc7_conv”
  relu_param {
    negative_slope: 0.05
  }
}
layer {
  name: ”drop7”
  type: ”Dropout”
  bottom: ”fc7_conv”
  top: ”fc7_conv”
  dropout_param {
    dropout_ratio: 0.5
  }
}
layer {
  name: ”fc8_facenet_conv”
  type: ”Convolution”
  bottom: ”fc7_conv”
  top: ”fc8_facenet_conv”
  param {
    lr_mult: 10.0
    decay_mult: 1.0
  }
  param {
    lr_mult: 20.0
    decay_mult: 0.0
  }
  convolution_param {
    num_output: 2
    kernel_size: 1
    weight_filler {
      type: ”gaussian”
      std: 0.01
    }
    bias_filler {
      type: ”constant”
      value: 0.0
    }
  }
}
layer {
  name: ”prob”
  type: ”Softmax”
  bottom: ”fc8_facenet_conv”
  top: ”prob”
}
```
