ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Αναγνώριση Φαγητού σε Εικόνες με Τεχνικές Υπολογιστικής Όρασης και Βαθιάς Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΠΑΛΥΒΟΥ ΧΡΗΣΤΟΥ

Επιβλέπων : Γεώργιος Ματσόπουλος
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2019

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Αναγνώριση Φαγητού σε Εικόνες με Τεχνικές Υπολογιστικής Όρασης και Βαθιάς Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΧΡΗΣΤΟΥ ΠΑΛΥΒΟΥ

Επιβλέπων : Γεώργιος Ματσόπουλος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 8η Ιουλίου 2019.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γ. Ματσόπουλος	Δ. Κουτσούρης	Α. Παναγόπουλος
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αν.Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2019

(Υπογραφή)

.....

ΧΡΗΣΤΟΣ ΠΑΛΥΒΟΣ

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

Περίληψη

Στην εργασία αυτή ερευνούμε την Οπτική Αναγνώριση Φαγητού ακολουθώντας δύο προσεγγίσεις που περιγράφονται παρακάτω. Για την αξιολόγηση κάθε προσέγγισης, χρησιμοποιούμε σύνολα δεδομένων και πειραματικές ρυθμίσεις που ακολουθούνται από τη βιβλιογραφία.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η παρουσίαση και η σύγκριση τεχνικών υπολογιστικής όρασης και τεχνικών βαθιάς μάθησης. Είναι δύο εντελώς διαφορετικές προσεγγίσεις του θέματος, καθεμία με τα δικά της πλεονεκτήματα.

Στην πρώτη προσέγγιση εξάγουμε χαρακτηριστικά από τις εικόνες και ταξινομούμε ένα σύνολο εικόνων με βάση τα χαρακτηριστικά αυτά. Χρησιμοποιούμε χαρακτηριστικά που είναι αμετάβλητα ως προς την κλίμακα, ως προς την περιστροφή καθώς και μία μέθοδο ταχείας εξαγωγής. Δεδομένου ότι ολόκληρη η διαδικασία ανίχνευσης είναι ένα ενιαίο δίκτυο, μπορεί να βελτιστοποιηθεί άμεσα από άκρο σε άκρο. Για την αξιολόγηση των μεθόδων και την εκτέλεση των πειραμάτων χρησιμοποιούμε το σύνολο δεδομένων FOOD-101.

Στην δεύτερη προσέγγιση κάνουμε χρήση Συνελικτικών Νευρωνικών Δικτύων, τα οποία εκπαιδεύονται στο σύνολο δεδομένων FOOD-101. Χρησιμοποιούμε είτε εκπαιδευμένα εκ των προτέρων νευρωνικά δίκτυα είτε όχι για την παραγωγή των μοντέλων. Για την αξιολόγηση των μεθόδων και την εκτέλεση των πειραμάτων χρησιμοποιούμε το σύνολο δεδομένων FOOD-101.

Λέξεις Κλειδιά
Συνελικτικά Νευρωνικά Δίκτυα, Μηχανική Μάθηση, Βαθιά Μάθηση, Μεταφορά Μάθησης,΄Οραση Υπολογιστών, Αναγνώριση Φαγητού, Αναλλοίωτος σε κλίμακα Μετασχγματισμος, Επιτυχαχυνόμενος Μεατασχηματισμός

Abstract

In this work, we investigate on Optical Food Recognition following two approaches that are described below. For the evaluation of each approach, we use datasets and configuration used in bibliography.

The aim of my thesis is to present and compare computer vision and deep learning techniques. They are very different approaches, each one with its own advantages.

In the first approach, we extract features from the images and classify a dataset of images according to these features. We use scale and rotation invariant features as well as a method of fast features extraction. As the whole procedure of classification is a unified network, it can be end-to-end optimized.
For the evaluation of the methods and the experiments we use dataset FOOD-101.

In the second approach, we use Convolutional Neural Networks that are trained on FOOD-101 dataset. We use either pre-trained neural networks or not for the model production. For the methods evaluation and the experiments we use dataset FOOD-101.

Key Words
Convolutional Neural Networks, Machine Learning, Deep Learning, Transfer Learning, Computer Vision, Food Recognition, SIFT, SURF

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον κ. Γεώργιο Ματσόπουλο για την επίβλεψη αυτής της διπλωματικής εργασίας και την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Τεχνολογίας Συστημάτων Πληροφορίας. Επιπλέον ατέρμονες είναι οι ευχαριστίες προς τους γονείς μου για την αγάπη και την υποστήριξη που μου προσέφεραν καθ'όλη την ακαδημαϊκή μου περίοδο. Ειδικότερα, τους ευχαριστώ για την υπευθυνότητα και την δημιουργικότητα, αξίες με τις οποίες με γαλούχησαν.

Τέλος, θα ήθελα να ευχαριστήσω όλους τους φίλους και τους ανθρώπους που μου στάθηκαν στις πιο δύσκολες στιγμές της ζωής μου. Ποτέ δε θα ξεχάσω την διαρκή ενθάρρυνση τους καθώς μοιραζόμουν μαζί τους τα όνειρα και φιλοδοξίες.

Προς τη μητέρα μου, το μοναδικό άνθρωπο που πίστευε,
πιστεύει και θα πιστεύει στις δυνατότητες μου.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο

Η καινοτομία είναι ο κινητήριος παράγοντας για να αλλάξουμε τον τρόπο με τον οποίο ζούμε. Αυτή τη στιγμή βρισκόμαστε στην εποχή της έκρηξης δεδομένων στη χρονική γραμμή της πληροφορικής. Η βελτίωση της τεχνολογίας και η διαθεσιμότητα τέτοιου τεράστιου όγκου δεδομένων μας δίνει ευκαιρίες να το αξιοποιήσουμε με τρόπους που μπορούν να αλλάξουν βαθιά τη ζωή μας.

Αρκετοί ερευνητές εργάζονται σε διάφορους τομείς για να εκμεταλλευτούν την δύναμη που παρέχουν τα μεγάλα δεδομένα και οι υπολογισμοί. Τεράστια ποσότητα δεδομένων αισθητήρων παράγονται καθημερινά. Η ανάλυση αυτών των δεδομένων με παραδοσιακές μεθόδους δεν είναι μόνο προκλητική, αλλά είναι σχεδόν αδύνατη. Οι επιχειρήσεις και τα ιδρύματα απομακρύνονται από τις παραδοσιακές μεθόδους ανάλυσης για να επωφεληθούν από την έκρηξη δεδομένων [7].

Συγκεκριμένα, τα οπτικά δεδομένα από διαφορετικά είδη αισθητήρων έχουν δει τεράστια αύξηση τα τελευταία χρόνια.

Προβλέπεται ότι το οχτώ τοις εκατό της συνολικής κυκλοφορίας Ιστού θα είναι βίντεο μέχρι το 2019 [10]. Για κάθε δευτερόλεπτο δημιουργούνται βίντεο υψηλής ανάλυσης αρκετών ωρών [12]. Αυτό μπορεί να φανεί από το γεγονός ότι όλο και περισσότερα βίντεο μεταδίδονται από κινητές συσκευές στο Periscope, το Facebook κ.λπ. που οδηγούν σε αύξηση των δεδομένων.

Από την άποψη αυτή, η μηχανική μάθηση (ML) γίνεται ο πρωταρχικός μηχανισμός για την εξαγωγή πληροφοριών από δεδομένα. Έχει χρησιμοποιηθεί για διάφορες εφαρμογές, από τη μεταποιητική βιομηχανία έως την ανίχνευση απάτης. Οι εφαρμογές αυτών των τεχνολογιών φαίνεται να είναι ουσιαστικά απεριόριστες αυτή τη στιγμή.

Παραδοσιακά, η εκμάθηση μηχανών περιοριζόταν στη διαδικασία μόνο μικρών συνόλων δεδομένων. Αυτό σήμαινε ότι η εφαρμογή των ιδεών από τον τομέα της Επεξεργασίας Εικόνων στα προβλήματα του πραγματικού κόσμου δεν ήταν εφικτή, κυρίως επειδή δεν είχαμε τα απαραίτητα δεδομένα εικόνας για να εκπαιδεύσουμε τα μηχανήματα ούτε την επαρκή υπολογιστική ισχύ για να τρέξουμε τους αλγορίθμους μάθησης. Ωστόσο, τα τελευταία χρόνια, η υπολογιστική ισχύς έχει αυξηθεί εκθετικά, έχουν καταστεί τεράστιες ποσότητες δεδομένων, ανακαλύφθηκαν προηγμένοι αλγόριθμοι επεξεργασίας δεδομένων και οι γεννήτριες δεδομένων και οι επεξεργαστές έχουν ενσωματωθεί άψογα στην υποδομή δικτύου και στα κέντρα δεδομένων. Αυτές οι εξελίξεις έχουν προετοιμάσει το δρόμο για την ανάπτυξη λογισμικού και υλικού απίστευτης πολυπλοκότητας για να κάνουν τα πάντα, από τα πολύ συνηθισμένα καθήκοντα της καθημερινότητας μέχρι τις πιο προηγμένες προσομοιώσεις.

Ειδικότερα, ως αποτέλεσμα της προόδου στη μηχανική μάθηση και τη βαθιά εκμάθηση, ο τομέας της υπολογιστικής όρασης

σημείωσε τεράστια πρόοδο τον τελευταίο καιρό. Αρκετοί κλάδοι έχουν εξελιχθεί σε μια κατάσταση όπου οι αλγόριθμοι μπορούν να χρησιμοποιηθούν για την αυτόματη αντιστοίχιση συγκεκριμένων καθηκόντων πραγματικού κόσμου που είχαν προηγουμένως γίνει χειρονακτικά από τον άνθρωπο. Η όραση υπολογιστών υπολογιστών γίνεται ένα θεμελιώδες εργαλείο σε αυτά τα προβλήματα αυτοματοποίησης των εργασιών και, ως εκ τούτου, υπάρχει μια τεράστια ζήτηση για καλά λειτουργούντα συστήματα υπολογιστικής όρασης αυτή τη στιγμή. Ένας συγκεκριμένος τομέας τεράστιας δυνατής εφαρμογής είναι το κινητό σύστημα όρασης που μπορεί να λειτουργήσει σε απρόσκοπτα σενάρια καθημερινής ζωής. Η όραση των υπολογιστών είναι ένας κλάδος της Τεχνητής Νοημοσύνης (Artificial Intelligence-AI) που αποσκοπεί στην εξαγωγή πληροφοριών από εικόνες. Τα δεδομένα εικόνας μπορούν να προέρχονται από πολλές πηγές: πλαίσια βίντεο, εικόνες από φωτογραφικές μηχανές ή δεδομένα υψηλής ανάλυσης από εξοπλισμό ιατρικής απεικόνισης κλπ. Η όραση υπολογιστών εφαρμόζει τις θεωρίες και τα μοντέλα που δανείζονται από διάφορους τομείς όπως μηχανική μάθηση, γνωσιακές επιστήμες, ψυχολογία, αναγνώριση προτύπων κλπ., για τη δημιουργία εφαρμογής για την επίλυση πολύ συγκεκριμένων προβλημάτων.

Η όραση υπολογιστή έχει ήδη εφαρμοστεί σε ένα ευρύ φάσμα εφαρμογών όπως η ανίχνευση προσώπου, η ανίχνευση ανωμαλιών στα εργοστάσια παραγωγής κλπ. [6] Σε μια εφαρμογή υπολογιστικής όρασης για την ανίχνευση μη φυσιολογικών προϊόντων στη γραμμή συναρμολόγησης και τον προσδιορισμό της ποιότητας των προϊόντων με χρήση εικόνας επεξεργασία. Ομοίως, [13] χρησιμοποιούν ένα σύστημα υπολογιστικής όρασης για την ανίχνευση ειδών φυτών. Αυτά τα παραδείγματα δείχνουν ότι η όραση υπολογιστή έχει τεράστιες δυνατότητες να εφαρμοστεί σε διαφορετικά περιβάλλοντα όπου μόνο το ανθρώπινο οπτικό σύστημα θα

μπορούσε να κάνει το έργο στο παρελθόν. Από καιρό αναγνωρίζεται ότι μία από τις κύριες εργασίες σε ένα σύστημα υπολογιστικής όρασης είναι να εντοπίσει και να αναγνωρίσει τα αντικείμενα μέσα σε μια εικόνα, αφού μία εικόνα μπορεί να περιέχει πολλά αντικείμενα μέσα σε αυτήν. Για ένα ανθρώπινο ον, το έργο φαίνεται τετριμμένο επειδή η φυσική επιλογή έχει διαμορφώσει το ανθρώπινο οπτικό σύστημα για εκατομμύρια χρόνια για ακριβώς αυτό το είδος εργασιών.

Δεδομένου ότι η λειτουργία του ανθρώπινου οπτικού συστήματος δεν είναι πολύ καλά κατανοητή, παρόμοιες μέθοδοι δεν μπορούν να εφαρμοστούν άμεσα στις μηχανές. Μια εικόνα είναι απλώς μια συλλογή από εικονοστοιχεία από την άποψη του υπολογιστή. Είναι δύσκολο να κατανοήσουν οι μηχανές τι αλλαγές στα εικονοστοιχεία αλλάζουν το αντικείμενο και τι όχι. Παίρνει τεράστια εφευρετικότητα και υπολογιστική ισχύ για μια μηχανή που χειρίζεται ακόμη και απλές εργασίες εντοπισμού και αναγνώρισης.

Ένας τομέας τεράστιας δυνατής εφαρμογής της όρασης υπολογιστή είναι η αναγνώριση τροφίμων. Το φαγητό είναι ένα από τα πιο αναπόσπαστα μέρη ολόκληρης της ανθρωπότητας. Αν και ο κύριος σκοπός του τροφίμου είναι να προμηθεύει τα απαραίτητα θρεπτικά συστατικά στο σώμα, η κατανάλωση φαγητού εξυπηρετεί επίσης έναν ευρύτερο σκοπό, όπως τον κοινωνικό δεσμό. Αυτό μπορεί να επιβεβαιωθεί από το γεγονός ότι σχεδόν όλες οι γιορτές, τα φεστιβάλ και οι διακοπές περιστρέφονται γύρω από κάποια μοναδική λιχουδιά που συνδέεται με αυτά. Επιπλέον, σε πολλούς πολιτισμούς συνηθίζεται ότι όλα τα μέλη μιας οικογένειας έχουν τουλάχιστον ένα γεύμα της ημέρας μαζί. Κάθε ανθρώπινη κοινωνία έχει τα δικά της τρόφιμα και ποτά και πολιτιστικές πρακτικές που συνδέονται με αυτές. Αν κάποιος θέλει να ξανασκεφτεί τα τρόφιμα από διαφορετικό τόπο ή πολιτισμό που

κάποια στιγμή προσπάθησε και άρεσε, τότε μπορούν απλά να τραβήξουν μια εικόνα του φαγητού ως ενθύμιο. Στο μέλλον, η εικόνα μπορεί να χρησιμοποιηθεί σε λογισμικό αυτόματης αναγνώρισης τροφίμων για να αναζητήσει τα συστατικά και τη συνταγή, έτσι ώστε το άτομο να το ετοιμάσει από μόνο του. Η αναζήτηση συστατικών μπορεί επίσης να είναι χρήσιμη εάν, π.χ., κάποιος έχει τροφικές αλλεργίες ή δεν μπορεί να καταναλώσει ορισμένα τρόφιμα για λόγους υγείας ή θρησκείας. Όταν ένας αλλεργικός δεν είναι βέβαιος τι είδους συστατικά έχει ένα φαγητό, το άτομο μπορεί απλά να τραβήξει την εικόνα του φαγητού και να ελέγξει αν υπάρχουν αλλεργιογόνα στο φαγητό. Ωστόσο, ο σύγχρονος κόσμος είχε το μερίδιό του στο πρόβλημα που σχετίζεται με τα τρόφιμα. Ο ΟΗΕ έχει αναφέρει ότι η παιδική και εφηβική παχυσαρκία έχει αυξηθεί κατά 10 φορές τις τελευταίες τέσσερις δεκαετίες [14]. Η κακή συνήθεια των τροφίμων θεωρείται μία από τις κύριες αιτίες αυτού του προβλήματος. Το να είναι σε θέση να σχεδιάσει εφαρμογές που αναγνωρίζουν τα τρόφιμα και στη συνέχεια υπολογίζει τη θρεπτική τους αξία μπορεί να έχει μεγάλη πρακτική εφαρμογή στην καταπολέμηση της παχυσαρκίας, ένα αυξανόμενο πρόβλημα στον αναπτυγμένο κόσμο. Πιστεύουμε ότι ψηφιακές συσκευές όπως η smartphone ή η smartwatch βοηθούν στην αυτόματη εκτίμηση θερμίδων και βοηθούν τους χρήστες να υιοθετήσουν καλές πρακτικές. Οι εργασίες που πραγματοποιήθηκαν κατά τη διάρκεια μιας διατριβής θα ήταν ένα βήμα προς αυτήν την κατεύθυνση.Το έργο αυτής της εργασίας θα είναι μια εφαρμογή της βαθιάς μάθησης και της εξαγωγής χαρακτηριστικών για την αναγνώριση τροφίμων. Εδώ περιορίσαμε το πεδίο εφαρμογής μας μόνο στο «ευρωπαϊκό φαγητό». Σε αυτή την εργασία, θα δημιουργήσουμε ένα σύνολο δεδομένων για την εικόνα των τροφίμων και θα χρησιμοποιήσουμε τις τελευταίες τεχνολογίες για τον εντοπισμό και την αναγνώριση αντικειμένων για να δημιουργήσουμε ένα μοντέλο. Αυτό είναι ένα δύσκολο πρόβλημα επειδή οι εικόνες του ίδιου φαγητού μπορεί να

φαίνονται πολύ διαφορετικές μεταξύ τους. Η πηγή και η θέση του φωτισμού, η θέση του αισθητήρα, τα συστατικά του τροφίμου, η θερμοκρασία των τροφίμων κ.λπ., μπορεί να επηρεάσουν την εμφάνιση του φαγητού. Επιπλέον, η εύρεση ενός σωστά επισημασμένου συνόλου δεδομένων για την εκπαίδευση του μοντέλου μας είναι επίσης μια τεράστια πρόκληση. Επιπλέον, η αναγνώριση της τροφής απλά εξετάζοντας την μπορεί να μην είναι δυνατή, ακόμη και για τον άνθρωπο, πόσο μάλλον για μηχανές, καθώς και άλλα χαρακτηριστικά, όπως η οσμή, η υφή κ.λπ., παίζουν επίσης σημαντικό ρόλο στην ταυτοποίηση των τροφίμων. Έτσι, υπάρχουν πολλές προκλήσεις σε αυτό το πρόβλημα της αναγνώρισης τροφίμων. Μία από τις συνεισφορές μας σε αυτή τη διατριβή είναι να εντοπίσουμε τέτοια προβλήματα και να διερευνήσουμε τις μεθόδους από τη βιβλιογραφία για την επίλυσή τους.

1.2 Δομή αυτής της εργασίας

Εχουμε διαρθρώσει αυτή την εργασία ως εξής

- Στο Κεφάλαιο 1, συζητήσαμε την ανάγκη για δεδομένα, την τάση της εποχής, για την μηχανική εκμάθηση και τις εφαρμογές της. Θέσαμε επίσης τον στόχο της εργασίας μας.
- Στο Κεφάλαιο 2, θα αναφερθούμε σε σχετική εργασία που έχει γίνει πάνω στην αναγνώριση εικόνας
- Στο κεφάλαιο 3, θα συζητήσουμε για τα τεχνητά νευρωνικά δίκτυα, τους τρόπους εκμάθησης, τις μηχανές διανυσματικής υποστήριξης και την μεταφορά μαθησης
- Στο κεφάλαιο 4, θα αναλύσουμε την λειτουργία και τη δομή των συνελικτικών νευρωνικών δικτύων
- Στο κεφάλαιο 5, θα μιλήσουμε για τα δεδομένα που χρησιμοποιήσαμε και θα αναλύσουμε τις μεθόδους που υλοποιήσαμε για την αναγνώριση φαγητού και κάποια

συνοπτικά αποτελέσματα

- Στο κεφάλαιο 6, θα παρουσιάσουμε τα πειράματα και τα αποτελέσματα μας από την αναγνώριση εικόνων φαγητού
- Στο κεφάλαιο 7, θα κάνουμε μια σύνοψη της εργασίας και θα μιλήσουμε για πιθανή μελλοντική έρευνα.

Κεφάλαιο 2

Σχετική Εργασία

Σε αυτό το κεφάλαιο αναφερόμαστε σε σχετική εργασία και τεχνικές που έχουν ακολουθηθεί πάνω στον εντοπισμό χαρακτηριστικών και σε μεθόδους βαθιάς μάθησης σε εικόνες.

2.1 Ανιχνευτής Γωνιών Harris

Οι γωνίες είναι περιοχές μιας εικόνας με μεγάλη ποικιλία στην ένταση σε όλες τις κατευθύνσεις. Ο εντοπιστής γωνιών HARRIS[1] βρίσκει τη διαφορά στην ένταση για μια μετατόπιση (u, v) προς όλες τις κατευθύνσεις. Αυτό εκφράζεται ως εξής:

$$E(u,v) = \sum_{x,y} \underbrace{w(x,y)}_{\text{window function}} [\underbrace{I(x+u,y+v)}_{\text{shifted intensity}} - \underbrace{I(x,y)}_{\text{intensity}}]^2$$

Η λειτουργία παραθύρου είναι είτε ένα ορθογώνιο παράθυρο είτε ένα γκαουσιανό παράθυρο που δίνει βάρη στα από κάτω εικονοστοιχεία.

Πρέπει να μεγιστοποιήσουμε τη συνάρτηση E(u, v) για ανίχνευση γωνιών. Αυτό σημαίνει ότι πρέπει να μεγιστοποιήσουμε τον δεύτερο όρο. Εφαρμόζοντας την επέκταση Taylor στην παραπάνω εξίσωση και χρησιμοποιώντας ορισμένα μαθηματικά βήματα, παίρνουμε την τελική εξίσωση ως:

$$E(u,v) \approx \begin{bmatrix} u & v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix}$$

όπου

$$M = \sum_{x,y} w(x,y) \begin{bmatrix} I_x I_x & I_x I_y \\ I_x I_y & I_y I_y \end{bmatrix}$$

Εδώ τα Ix, Iy είναι οι παράγωγοι της εικόνας στις κατευθύνσεις x, y αντίστοιχα.
Τώρα έρχεται το κύριο μέρος. Μετά από αυτό, δημιουργούμε ένα σκορ, βασικά μια εξίσωση, η οποία θα καθορίσει εάν ένα παράθυρο μπορεί να περιέχει μια γωνία ή όχι.

$$R = det(M) - k(trace(M))^2$$

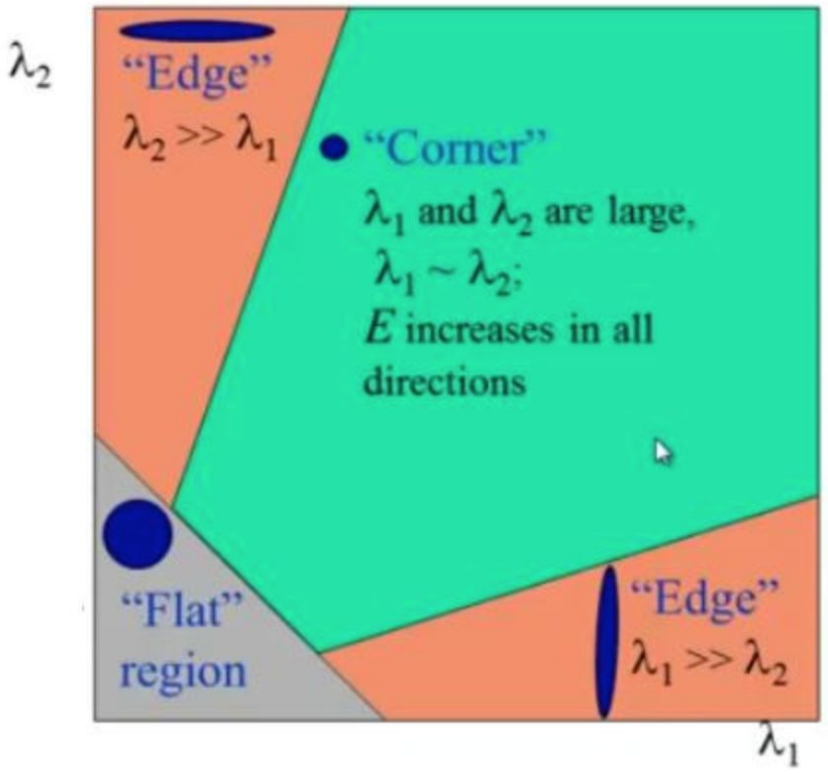
όπου

- $det(M) = \lambda_1 \lambda_2$
- $trace(M) = \lambda_1 + \lambda_2$
- λ_1 and λ_2 are the eigen values of M

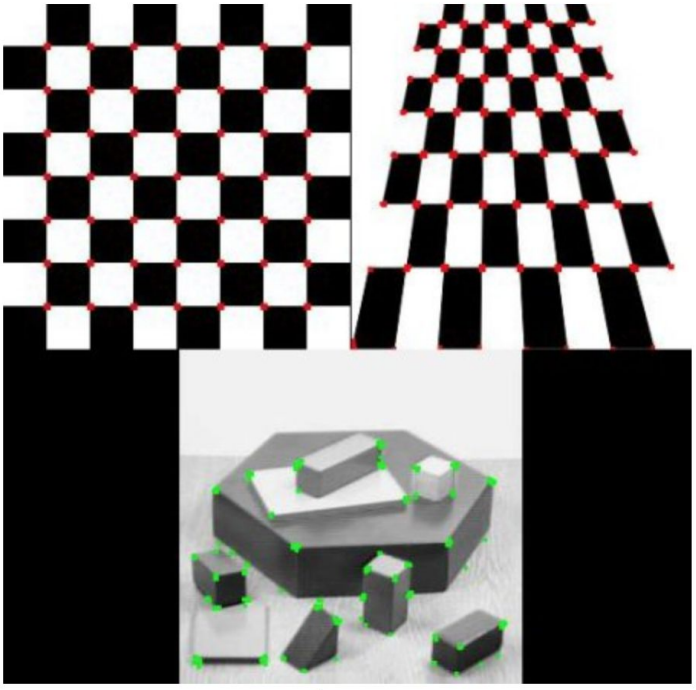
Οπότε οι τιμές αυτών των ιδιοτιμών αποφασίζουν αν μια περιοχή θεωρείται γωνία ή όχι:

- When $|R|$ is small, which happens when λ_1 and λ_2 are small, the region is flat.
- When $R < 0$, which happens when $\lambda_1 >> \lambda_2$ or vice versa, the region is edge.
- When R is large, which happens when λ_1 and λ_2 are large and $\lambda_1 \sim \lambda_2$, the region is a corner.

Μπορεί να αναπαρασταθεί με την παρακάτω εικόνα:

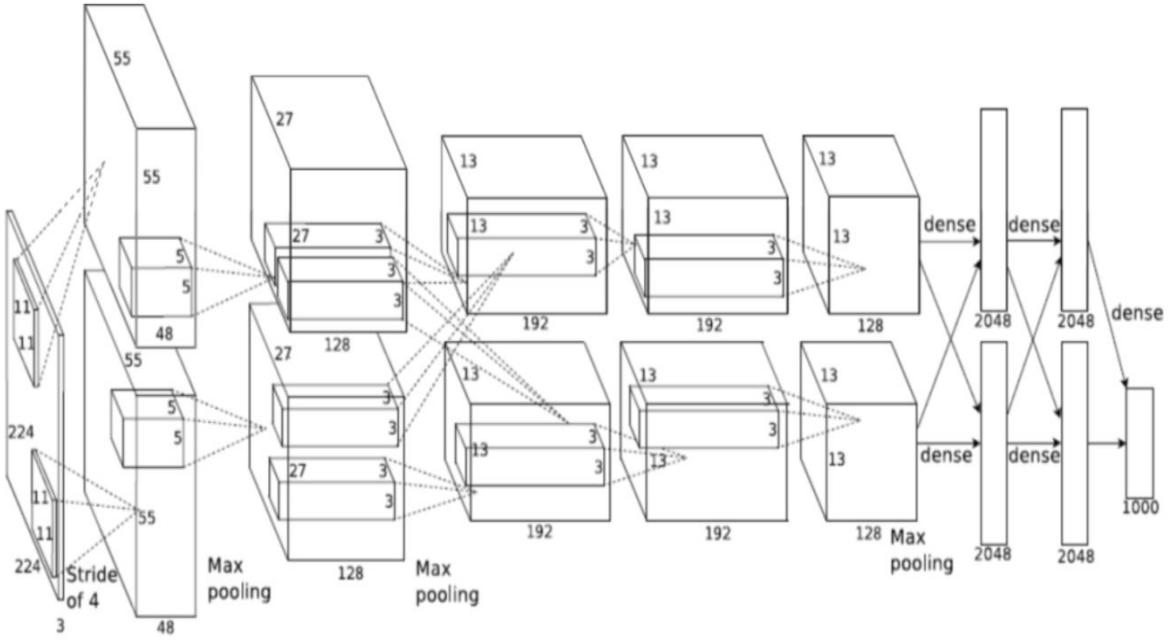


Έτσι, το αποτέλεσμα της ανίχνευσης Harris Corner είναι μια εικόνα γκρι με αυτά τα αποτελέσματα. Κατωφλιώνοντας κατάλληλα παίρνουμε τις γωνίες στην εικόνα. Θα το κάνουμε με μια απλή εικόνα:



2.2 AlexNet

Το Alexanet [15] προτάθηκε από τους Krizhevsky et al. Αυτή η δημοσίευση ήταν ο νικητής του διαγωνισμού ImageNet Large-Scale Visual Recognition για το 2012 (ILSVRC). Πέτυχε top-5 best error 15,4%. Η δεύτερη καλύτερη επίδοση είχε ποσοστό σφάλματος 26,2%. Αυτό σήμαινε τεράστια βελτίωση επιδόσεων.
Παρακάτω φαίνεται η αρχιτεκτονική του δικτύου.



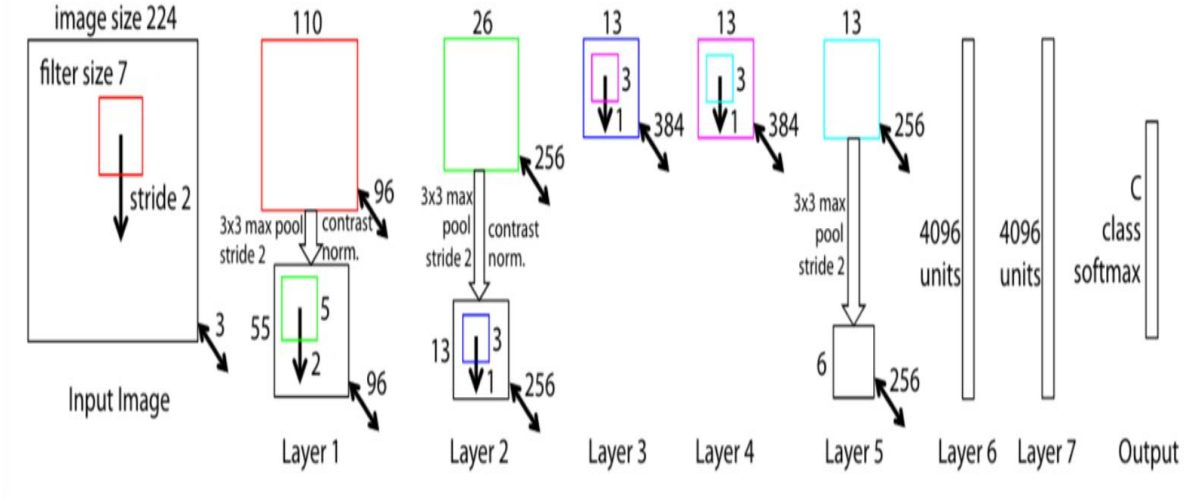
Η διαδικασία εκπαίδευσης ήταν υπολογιστικά ακριβή σε σύγκριση με αυτές του παρόντος. Μπορούμε να δούμε δύο ροές συνελικτικών δικτύων. Ένα άλλο χαρακτηριστικό του AlexNet ήταν ότι η γραμμική μονάδα ανορθωτή χρησιμοποιήθηκε ως μη γραμμική συνάρτηση ενεργοποίησης. Οι ερευνητές χρησιμοποίησαν την ενίσχυση για να λύσουν το πρόβλημα μετασχηματισμού της εικόνας. Είχαν χρησιμοποιήσει την dropout για να αποφύγουν το πρόβλημα του overfitting. Το μοντέλο εκπαιδεύτηκε σε τμήμα του συνόλου δεδομένων 15 εκατομμυρίων εικόνων χρησιμοποιώντας τη μέθοδο στοχαστικής καθόδου κλίσης(stochastic gradient descent-SGD) σε GTX 580 GPU για πέντε έως έξι ημέρες.

Οι επιδόσεις που επιτεύχθηκαν με τη χρήση αυτής της μεθόδου ήταν ένα εξαιρετικό 10,8% στο top 5 error. Αυτό ήταν μια επίδειξη της ικανότητας του CNN να λύσει το ILSVRC, το οποίο θεωρήθηκε δύσκολο πρόβλημα για μεγάλο χρονικό διάστημα. Μετά τη δημοσίευση της παρούσας εργασίας, περισσότεροι ερευνητές στον τομέα της υπολογιστικής όρασης άρχισαν να χρησιμοποιούν CNN στην έρευνά τους [16]. Αυτό υποστηρίζεται επίσης από το γεγονός ότι η παρούσα

δημοσίευση αναφέρθηκε περισσότερες από 20000 φορές.

2.3 ZF Net

Το ZF net [17] προτάθηκε από τους Zeiler και Fergus. Αυτή η δημοσίευση πρότεινε βελτιώσεις σε ορισμένες πτυχές του AlexNet. Ήταν ο νικητής του ILSVRC το 2013, με το κορυφαίο ποσοστό top 5 error 11,2%. Η αρχιτεκτονική του ZF Net παρουσιάζεται παρακάτω. Αν και η αρχιτεκτονική ήταν ελαφρώς διαφορετική από την AlexNet, το μοντέλο αυτό εκπαιδεύτηκε σε μόνο 1,3 εκατομμύρια εικόνες σε σύγκριση με 15 εκατομμύρια εικόνες για το Alexnet. Αλλάζουν επίσης το μέγεθος του φίλτρου από 11 × 11 στο AlexNet σε 7 × 7 στο ZF Net. Μεγαλύτερο φίλτρο χρησιμοποιήθηκε στα ανώτερα στρώματα. Η ReLu χρησιμοποιήθηκε σαν συνάρτηση ενεργοποίησης και η cross-entropy loss σαν συνάρτηση κόστους. Αυτό το μοντέλο εκπαιδεύτηκε επίσης σε GTX 580 GPU αλλά για δώδεκα ημέρες.



2.4 VGG Net

Το VGGNet [18] προτάθηκε το 2014 από τους Simonyan et al. Αντάλλαξαν το φίλτρο 11 × 11 Alexnet και 7 × 7 ZF Net από ένα φίλτρο 3 × 3 στην αρχιτεκτονική του μοντέλου. Έγιναν απλούστερες αλλά αυστηρότερες σχεδιαστικές αποφάσεις για

αυτό το δίκτυο. Το φίλτρο ήταν πάντα 3 × 3 με padding 1, μαζί με ένα 2 × 2 max-pooling στρώμα με stride 2. Συνεπώς, το βάθος του δικτύου αυξάνεται ενώ η χωρική διάσταση συρρικνώνεται. Αυτό απέδωσε καλά όχι μόνο για την αναγνώριση αλλά και για τον εντοπισμό. Αυτό είχε 7,3% στο top 5 error στον διαγωνισμό ILSVRC.

2.5 GoogLeNet/Inception

Το μοντέλο GoogLeNet/Inception ήταν ο νικητής του ILSVRC 2014 με κορυφαίο top 5 error 6,7% [19]. Αυτό αναπτύχθηκε από ερευνητές της Google. Τα CNN που χρησιμοποιήθηκαν σε παλαιότερα μοντέλα όπως το AlexNet, VGG, κλπ. είχαν απλές ενότητες στοίβας συνέλιξης και max-pooling. Ωστόσο, το GoogLeNet περιέχει μια στοίβα από 22 συνελίξεις και μια νέα ενότητα που ονομάζεται Inception Module. Το μοντέλο αυτό αποτελείται από παράλληλους υπολογισμούς συνελίξεων διαφορετικών σχημάτων και max-pooling. Η έξοδος από όλες αυτές τις ενότητες συνδέεται στο τέλος κάθε Inception Module. Το μοντέλο αυτό θα χρησιμοποιηθεί στα πειράματα μας για την αρχικοποίηση των δικών μας μοντέλων εκπαίδευσης.

2.6 ResNet

Το ResNet ήταν δίκτυο βαθιάς αρχιτεκτονικής με 152 στρώματα [20]. Αυτό αναπτύχθηκε από την Microsoft Asia. Ήταν ο νικητής του ILSVRC 2015 με το κορυφαίο top 5 error μόλις 3,6%, το οποίο θεωρείται καλύτερο από την ακρίβεια σε επίπεδο ανθρώπων [21].Οι ερευνητές χρησιμοποίησαν υπολειπόμενες ενότητες για αυτήν την αρχιτεκτονική. Κανονικά στο CNN η είσοδος μετασχηματίζεται σε διαφορετική διάσταση όταν φθάνει στο επίπεδο εξόδου. Από την άλλη πλευρά, τα υπολειπόμενα επίπεδα στο ResNet υπολογίζουν τις αλλαγές στην είσοδο. Αυτό στη συνέχεια προστίθεται στην είσοδο για να παράγει την έξοδο. Οι ερευνητές πιστεύουν ότι αυτή η

υπολειμματική αντιστοίχιση είναι πιο εύκολη στη βελτιστοποίηση. Αυτή η αρχιτεκτονική είναι μια από τις καλύτερες υπερσύγχρονες αρχιτεκτονικές λόγω της σημαντικής βελτίωσης της απόδοσης έναντι των προκατόχων της.

Κεφάλαιο 3

Μηχανική Μάθηση

Σε αυτό και το επόμενο κεφάλαιο, παρουσιάζεται η θεωρία του υπόβαθρου της Μηχανικής Μάθησης που χρησιμοποιήθηκε στην παρούσα εργασία, καθώς και των Συνελικτικών Νευρωνικών Δικτύων, αρχιτεκτονικές και εφαρμογές τους με την πρόθεση να αναλυθεί το σύστημα αναγνώρισης φαγητού σε εικόνες.

3.1 Ορισμός

Μηχανικά μάθηση ονομάζουμε το κομμάτι της επιστήμης υπολογιστών που προσπαθεί να μοντελοποιήσει πολλαπλά επίπεδα αφαιρετικότητας, χρησιμοποιώντας πολλαπλά επίπεδα στις αρχιτεκτονικές της. Τα δίκτυα αυτά είναι εμπνευσμένα από το πώς επεξεργάζεται ο άνθρωπος την πληροφορία και προσπαθούν να προσομοιάσουν την λειτουργία των νευρώνων στο νεοφλοιό του εγκεφάλου (όπου γίνεται περίπου το 80% της ανθρώπινης σκέψης). Η διαδικασία της μάθησης ξεκινά με παρατηρήσεις, που αποτελούν παραδείγματα, ή εμπειρικά αποτελέσματα ή οδηγίες, ούτως ώστε να αναγνωριστούν πρότυπα στα δεδομένα και να ληφθούν καλύτερες αποφάσεις

στο μέλλον, με βάση τα παραδείγματα που διαθέτουμε. Ο πρωταρχικός σκοπός είναι να επιτρέψουμε στους υπολογιστές να μαθαίνουν αυτόματα, χωρίς ανθρώπινη παρέμβαση ή βοήθεια, και να προσαρμόζουν τις πράξεις τους κατάλληλα.

Στη μηχανική μάθηση, τα καθήκοντα ταξινομούνται γενικά σε ευρείες κατηγορίες. Οι κατηγορίες αυτές βασίζονται στον τρόπο με τον οποίο λαμβάνεται η μάθηση ή στον τρόπο με τον οποίο δίνεται ανάδραση στην εκμάθηση στο ανεπτυγμένο σύστημα. Δύο από τις πιο ευρέως υιοθετημένες μεθόδους είναι η επιβλεπόμενη μάθηση, η οποία εκπαιδεύει αλγορίθμους που βασίζονται στα δεδομένα εισόδου και εξόδου τα οποία επισημαίνονται (αποκτούν ετικέτες-labels) από τον άνθρωπο και η μη επιβλεπόμενη μάθηση, η οποία παρέχει τον αλγόριθμο χωρίς επισημασμένα δεδομένα, ούτως ώστε να του επιτρέψει να βρει δομή στα δεδομένα εισόδου του.

3.2 Μάθηση με Επίβλεψη

Η Μάθηση με Επίβλεψη είναι ο τομέας μηχανικής μάθησης όπου μια συνάρτηση μαθαίνει να αντιστοιχίζει δεδομένα εισόδου σε δεδομένα εξόδου χρησιμοποιώντας παραδείγματα ζευγών εισόδου-εξόδου. Αυτό το σύνολο ζευγών ονομάζεται σύνολο εκπαίδευσης και η διαδικασία υπολογισμού μιας τέτοιας συνάρτησης από το παραπάνω σύνολο λέγεται εκπαίδευση. Ο σκοπός της μάθησης με επίβλεψη είναι ο υπολογισμός μιας συνάρτησης που γενικεύει επαρκώς σε δεδομένα εισόδου στα οποία δεν έχει εκπαιδευτεί, αντιστοιχίζοντάς τα σε σωστές εξόδους.

Η μάθηση με επίβλεψη είναι κατάλληλη για εργασίες όπου, ενώ η απεικόνιση εισόδου εξόδου είναι δύσκολο-πολύπλοκο να βρεθεί αναλυτικά, υπάρχει ένα αρκετά μεγάλο σύνολο δεδομένων εκπαίδευσης. Μια τέτοια περίπτωση είναι η ταξινόμηση εικόνων. Από προγραμματιστική άποψη, οι εικόνες αναπαριστώνται από 3D τένσορες με διαστάσεις ύψους,

πλάτους και βάθους (καναλιών, πχ RGB).΄Ενα σύνολο εκπαίδευσης για ταξινόμηση εικόνων περιέχει μια συλλογή εικόνων και αντίστοιχες επισημάνσεις.΄Ενα παράδειγμα είναι εικόνες με αυτοκίνητα και ποδήλατα με τις αντίστοιχες επισημάνσεις "αυτοκίνητο" ή "ποδήλατο" για κάθε μια από αυτές.

Στη μάθηση με επίβλεψη, οι είσοδοι του συνόλου εκπαίδευσης πρέπει να μετατραπούν σε καταλληλότερες δομές ώστε να γίνει περαιτέρω επεξεργασία. Τέτοια είναι η περίπτωση για σύνολα εκπαίδευσης σχετικά με κείμενο και η κλασική προσέγγιση στην όραση υπολογιστών. Πριν τα συνελικτικά νευρωνικά δίκτυα γίνουν δημοφιλή, η διαδικασία ταξινόμησης εικόνων περιλάμβανε την μετατροπή κάθε εικόνας σε ένα σύνολο διανυσμάτων περιγραφής της εικόνας, τα χαρακτηριστικά, τα οποία χρησιμοποιούνταν σαν είσοδος για την εκπαίδευση μοντέλων.΄Ενα τυπικό παράδειγμα χαρακτηριστικών είναι οι γωνίες. Η διαδικασία σχεδιασμού τέτοιων χαρακτηριστικών περιλάμβανε κοπιαστική παραμετροποίηση, ενώ τα αποτελέσματα αυτού του πλαισίου εργασίας φάνηκε να μην βελτιώνεται περαιτέρω προς το τέλος της δεκαετίας του 2000. Με την εισαγωγή των ΣΝΔ, η διαδικασία σχεδιασμού χαρακτηριστικών έγινε μέρος της διαδικασίας εκπαίδευσης. Τα ΣΝΔ υπολογίζουν τα δικά τους, ειδικά χαρακτηριστικά, τα οποία δεν είναι διαισθητικά κατανοητά από ανθρώπους, όπως είναι οι γωνίες.

Μια ακόμα εργασία Μάθησης με Επίβλεψη, εκτός της ταξινόμησης, είναι η παλινδρόμηση, όπου, αντί για κατηγορία ως έξοδο, αναμένεται αριθμητική τιμή. Στο πλαίσιο της ανίχνευσης αντικειμένων από εικόνες, ένα παράδειγμα είναι να βρεθούν οι συντεταγμένες που ορίζουν ένα ορθογώνιο περίγραμμα γύρω από το αυτοκίνητο.

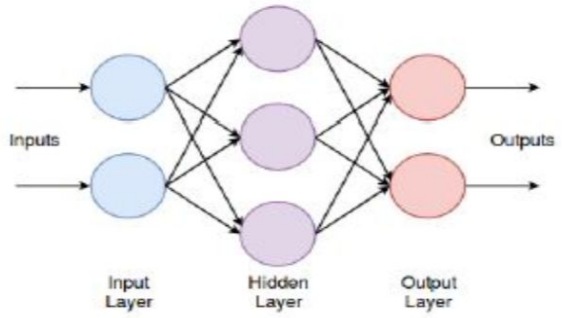
3.3 Μάθηση χωρίς Επίβλεψη

Σε άλλα προβλήματα μηχανικής μάθησης, υπάρχει μια διαφορετική τάξη καθηκόντων που αναφέρεται ως μη επιβλεπόμενη μάθηση. Στα προβλήματα αυτής της κατηγορίας, τα δεδομένα εκπαίδευσης είναι διανύσματα x τα οποία δεν έχουν αντίστοιχες ετικέτες. Επομένως, ο στόχος τη μη επιβλεπόμενης μάθησης είναι να βρίσκει μοτίβα όταν δεν υπάρχουν "σωστές απαντήσεις", ή όταν αυτές είναι αδύνατον να υπολογιστούν. Μία μεγάλη υποκατηγορία μη επιβλεπόμενων τεχνικών είναι το πρόβλημα της ομαδοποίησης (clustering). Η μέθοδος αυτή αναφέρεται στην ομαδοποίηση παρατηρήσεων με τέτοιο τρόπο ούτως ώστε τα μέλη μιας κοινής ομάδας να είναι παρόμοια το ένα με το άλλο, και να διαφέρουν σημαντικά από τα μέλη των άλλων ομάδων. Μια άλλη πολύ ενδιαφέρουσα κατηγορία μη επιβλεπόμενων καθηκόντων είναι οι τα γεννητικά μοντέλα (generative models). Τα μοντέλα αυτά μιμούνται τη διαδικασία δημιουργίας των δεδομένων εκπαίδευσης.΄Ενα καλό γεννητικό μοντέλο θα πρέπει να μπορεί να δημιουργήσει νέα δεδομένα τα οποία, αν και είναι τεχνητά, μοιάζουν με τα αυθεντικά. Αυτός ο τρόπος μάθησης είναι μη επιβλεπόμενος διότι η διαδικασία με την οποία δημιουργούνται ("γεννιούνται") τα δεδομένα δεν είναι άμεσα παρατηρήσιμη - μόνο τα ίδια τα δεδομένα είναι παρατηρήσιμα.

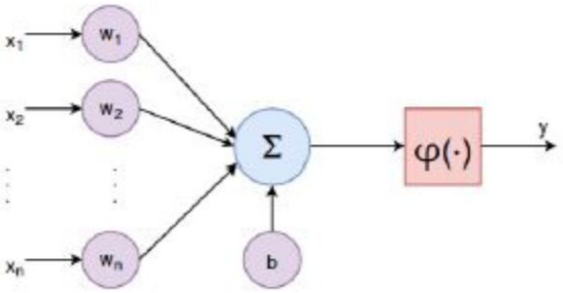
3.4 Νευρωνικά Δίκτυα

Ενα παράδειγμα συνάρτησης Μάθησης με Επίβλεψη είναι τα Νευρωνικά Δίκτυα Πρόσθιας Τροφοδότησης (ΝΔΠΤ). Τα νευρωνικά αυτά δίκτυα αποτελούνται από ένα επίπεδο εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα επίπεδο εξόδου, σειριακά συνδεδεμένα. Το επίπεδο εισόδου παρέχει ένα διάνυσμα εισόδου στο δίκτυο, ενώ το επίπεδο εξόδου παρέχει την πρόβλεψη σαν μια τιμή ή διάνυσμα. Τα κρυφά επίπεδα ορίζουν την πολύπλοκη εσωτερική λειτουργία του δικτύου. Η

γενική όψη τους παρουσιάζεται στο στο παρακάτω σχήμα:



(α′) Απλό Νευρωνικό Δίκτυο Πρόσθιας Τροφοδότησης



(β′) Αρχιτεκτονική ενός Νευρώνα

Η δομή των νευρωνικών δικτύων πρόσθιας τροφοδότησης είναι ένας κατευθυνόμενος ακυκλικός γράφος, όπου κάθε κόμβος, ο Νευρώνας, είναι μια μη γραμμική συνάρτηση Rn → R. Ο Νευρώνας υπολογίζει την έξοδο σε δύο βήματα - υπολογίζει το σταθμισμένο άθροισμα των ν εισόδων του συν έναν όρο πόλωσης και στη συνέχεια εφαρμόζει μια συνάρτηση κανονικοποίησης φ(·) στο άθροισμα. Η τελευταία λέγεται συνάρτηση ενεργοποίησης. Υπάρχουν n + 1 παράμετροι σχετικές του κάθε Νευρώνα, τα n βάρη w1, . . . wn και η πόλωση b, όπως φαίνεται κι από την παρακάτω εξίσωση. Η έξοδος κάθε νευρώνα γίνεται είσοδος σε κάθε Νευρώνα στο επόμενο επίπεδο. Το νευρωνικό δικτυο πρόσθιας τροφοδότησης μπορεί να ρυθμιστεί με εκπαίδευση με προσαρμογή του συνόλου των παραμέτρων των νευρώνων.

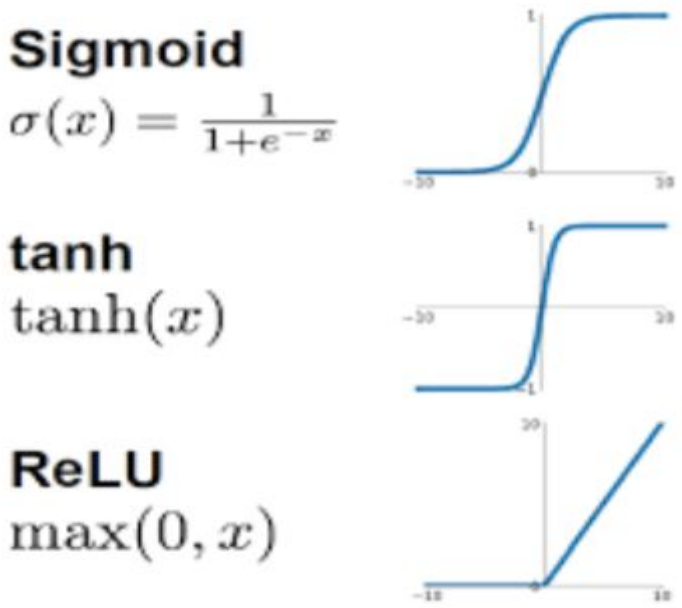
$$y = \varphi\left(\sum_{j=0}^{n} w_j x_j + b\right)$$

Χρησιμοποιώντας μια μη γραμμική συνάρτηση ενεργοποίησης, το νευρωνικό δικτυο πρόσθιας τροφοδότησης που προκύπτει γίνεται επίσης μη γραμμικό. Παραδείγματα συναρτήσεων

ενεργοποίησης είναι η Σιγμοειδής (Sigmoid), η Υπερβολική Εφαπτομένη (Tanh) και η ReLU. Οι μη γραμμικές συναρτήσεις ενεργοποίησης κάνουν τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης καθολικές προσεγγίσεις, δηλαδή μπορούν να προσεγγίσουν επαρκώς συνεχείς συναρτήσεις ορισμένες σε συμπαγή υποσύνολα του Rn επιλέγοντας κατάλληλες παραμέτρους και έναν επαρκή, αλλά πεπερασμένο, αριθμό νευρώνων. Αυτή η ιδιότητα προσθέτει μια θεωρητική δικαιολόγηση χρήσης νευρωνικού δικτυου πρόσθιας τροφοδότησης για εργασίες Μάθησης με Επίβλεψη.

Η εκπαίδευση νευρωνικού δικτυου πρόσθιας τροφοδότησης γίνεται με τον δημοφιλή αλγόριθμο οπισθοδιάδοσης. Η οπισθοδιάδοση είναι ένας αλγόριθμος βελτιστοποίησης που ψάχνει για παραμέτρους (βάρη και πολώσεις) που ελαχιστοποιούν την συνάρτηση σφάλματος. Το σφάλμα ορίζεται ως μια συνάρτηση που μετράει την απόσταση (σφάλμα) μεταξύ των εξόδων του συνόλου εκπαίδευσης και των αντίστοιχων εξόδων υπολογισμένων από το νευρωνικό δικτυο πρόσθιας τροφοδότησης χρησιμοποιώντας τις εισόδους του συνόλου εκπαίδευσης. Η διαισθητική επεξήγηση της οπισθοδιάδοσης είναι ότι το σφάλμα υπολογίζεται στην έξοδο και διανέμεται προς τα προηγούμενα επίπεδα του δικτύου, ανανεώνοντας τις παραμέτρους ανάλογα με την ευαισθησία τους σε αλλαγές του σφάλματος, χρησιμοποιώντας τον κανόνα παραγώγισης αλυσίδας.

Παρακάτω φαίνονται μερικές συναρτήσεις ενεργοποίησης:



3.4.1 Optimizers

Έστω ότι η συνάρτηση που πρέπει να ελαχιστοποιθεί είναι η f(x). Τότε η κλίση της θα είναι η Δf(x). Το μέγεθος του βήματος για την κ επανάληψη είναι t_k. Παρακάτω αναλύονται οι πιο διαδεδομένοι αλγόριθμοι ελαχιστοποίησης της συνάρτησης[27].

Batch Gradient Descent

Ο αλγόριθμος Batch Gradient Descent ενημερώνει τις παραμέτρους x διατρέχωντας όλο το σύνολο δεδομένων ως εξής:

$$x_{k+1} = x_k - t_k \Delta f(x_k)^{(1:n)}$$

Ο αλγόριθμος αυτός εγγυάται σύγκλιση σε ολικό ελάχιστο για κυρτό πρόβλημα(convex problem) και σύγκλιση σε τοπικό ελάχιστο για μη-κυρτό πρόβλημα(non-convex problem). Ομως σε σύγχρονα προβλήματα βαθιάς μάθησης θα πάρει πολυ

χρόνο και μνήμη ο υπολογισμός του ολικού ελαχίστου απο ολόκληρο το σύνολο δεδομένων. Οπότε η μέθοδος αυτή ελαχιστοποίησης χρησιμοποιείται σπάνια σε προβλήματα βαθιάς μάθησης.

Stochastic Gradient Decent

Αντίθετα, ο αλγόριθμος Stochastic Gradient Decent υπολογίζει τις κλίσεις και ενημερώνει τις παραμέτρους για κάθε δεδομενο του συνολου εκπαιδευσης:

$$x_{k+1} = x_k - t_k \Delta f(x_k)^{(i)}$$

Όμως, λόγω της διακύμανσης(variance) των δεδομένων του συνόλου εκπαίδευσης, η αντικειμενική συνάρτηση (objective function) θα διακυμαίνεται έντονα. Παρόλο που το μικρό μέγεθος βήματος επιτρέπει στον αλγόριθμο να συγκλίνει σε ένα καλό σημείο, η εκπαίδευση είναι αργή.

Mini-Batch Gradient Decent

Ο Mini-Batch Gradient Decent συνδυάζει τα πλεονεκτήματα των προηγούμενων δύο αλγορίθμων και ενημερώνει τις παραμέτρους αφού έχει υπολογίσει την κλίση σε ένα μικρό σύνολο (batch) δεδομένων:

$$x_{k+1} = x_k - t_k \Delta f(x_k)^{(i:i+m)}$$

όπου m είναι το μέγεθος του μικρού αυτού συνόλου. ο Mini-Batch Gradient Decent δεν εγγυάται σύγκλιση σε καλό σημείο και το μέγεθος βήματος απαιτεί εμπειρία. Οπότε οι ερευνητές επεκτείνουν τον αλγόριθμο αυτό για να πετύχουν καλύτερο σημείο σύκλισης.

Adagrad

Ο Adagrad[24] προσαρμόζει το μέγεθος του βήματος για κάθε παράμετρο αναλόγα με το ιστορικό ενημερώσεων της παραμέτρου:

$$G_k = G_{k-1} + \Delta f(x_k)^2$$

$$x_{k+1} = x_k - \frac{t}{\sqrt{G_k + \varepsilon}} \Delta f(x_k)$$

όπου G είναι μια συσσώρευση(accumulation) του ιστορικού της παραμέτρου και ε παράμετρος απάλυνσης(smoothing) για αποφυγή διαίρεσης με το μηδέν και είναι της τάξης του 1e-6. Το μέγεθος βήματος είναι μεγαλύτερο για τις παραμέτρους με μικρό G και μεγάλο για παραμέτρους με μεγάλο G. Όμως, όταν το G μεγαλώσει αρκετά, το μέγεθος του βήματος θα τείνει στο μηδέν ύστερα από μεγάλο πλήθος επαναλήψεων. Οπότε έχουν προταθεί οι επόμενες μέθοδοι.

Adadelta

Ο Adadelta[25] προέρχεται από τον Adagrad και βελτιώνει τα κύρια δύο μειονεκτήματά του:

1. Τη συνεχή φθορά(decay) των βαθμών εκμάθησης(learning rates)
2. Την ανάγκη για χειρονακτική επιλογή καθολικού learning rate

Ο Adadelta κλιμακώνει το μέγεθος του βήματος σύμφωνα με την τελευταία ενημέρωση του ιστορικού σε αντίθεση με τον Adagrad που χρησιμοποιεί όλο το ιστορικό. Επίσης χρησιμοποιεί έναν όρο επιτάχυνσης (acceleration term):

$$\mathbb{E}[\Delta f(x)^2]_k = \rho \mathbb{E}[\Delta f(x)^2]_{k-1} + (1-\rho)\Delta f(x_k)^2$$

$$\hat{x}_k = -\frac{\sqrt{\mathbb{E}[\hat{x}^2]_{k-1} + \varepsilon}}{\sqrt{\mathbb{E}[\Delta f(x)^2]_k + \varepsilon}} \Delta f(x_k)$$

$$\mathbb{E}[\hat{x}^2]_k = \rho \mathbb{E}[\hat{x}^2]_{k-1} + (1-\rho)\hat{x}_k^2$$

$$x_{k+1} = x_k + \hat{x}_k$$

όπου ρ είναι μια σταθερά φθοράς(decay constant) και ε παράμετρος σταθερότητας(stability constant)

RMSprop

Ο RMSprop[26] χρησιμοποιείται επίσης για την αντιμετώπιση των προβλημάτων του Adagrad:

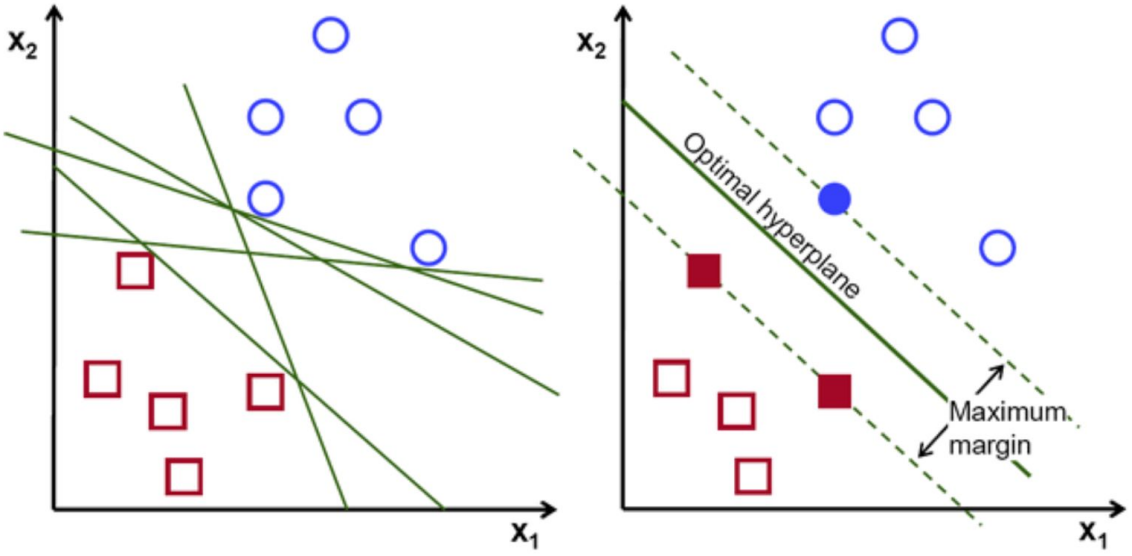
$$\mathbb{E}[\Delta f(x)^2]_k = \rho \mathbb{E}[\Delta f(x)^2]_{k-1} + (1-\rho)\Delta f(x_k)^2$$

$$x_{k+1} = x_k - \frac{t}{\sqrt{\mathbb{E}[\Delta f(x)^2]_k + \varepsilon}} \Delta f(x_k)$$

3.5 Γραμμική Ταξινόμηση

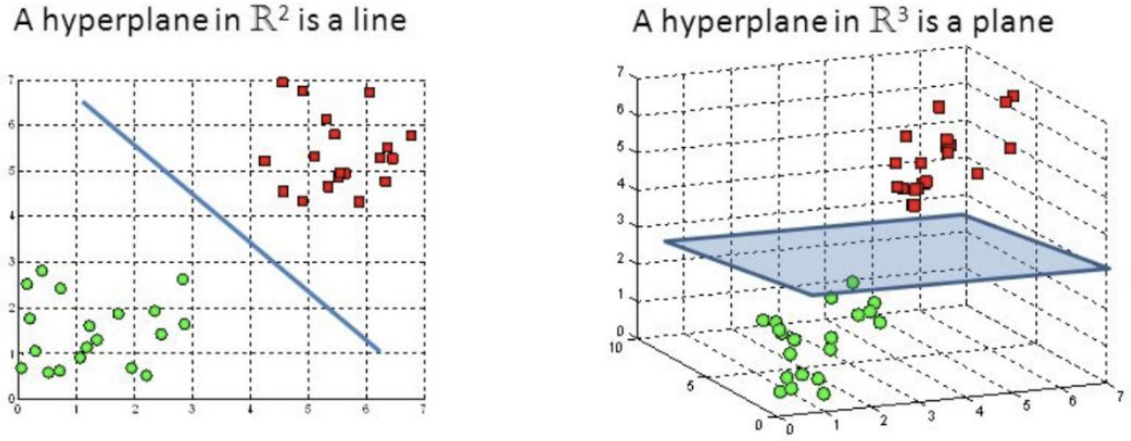
3.5.1 Μηχανές Διανυσμάτων Υποστήριξης

Ο στόχος του αλγόριθμου μηχανής διανυσματικής υποστήριξης[2] είναι να βρεθεί ένα υπερπλάνο σε ένα χώρο N-διαστάσεων (N - ο αριθμός των χαρακτηριστικών) που ταξινομεί τα σημεία δεδομένων.



Possible hyperplanes

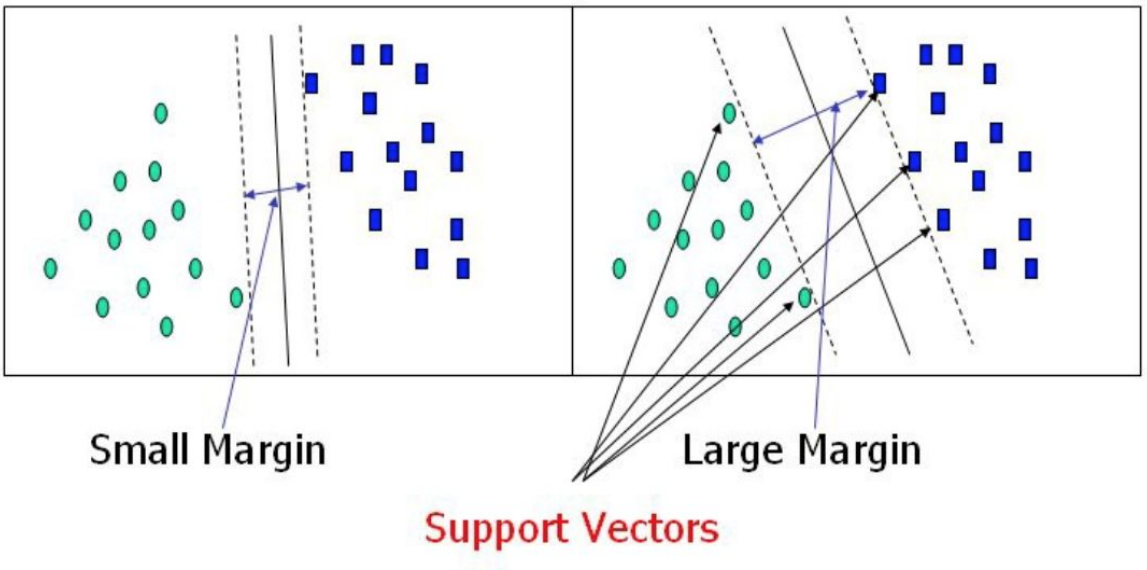
Για να διαχωρίσουμε τις δύο κατηγορίες σημείων δεδομένων, υπάρχουν πολλά πιθανά υπερπλάνα που θα μπορούσαν να επιλεγούν. Στόχος μας είναι να βρούμε ένα υπερπλάνο που έχει το μέγιστο περιθώριο, δηλαδή τη μέγιστη απόσταση μεταξύ σημείων δεδομένων και των δύο κατηγοριών. Η μεγιστοποίηση της απόστασης περιθωρίου παρέχει κάποια ενίσχυση, ώστε τα μελλοντικά σημεία δεδομένων να μπορούν να ταξινομηθούν με μεγαλύτερη αξιοπιστία.



Hyperplanes in 2D and 3D feature space

Τα υπερπλάνα είναι όρια απόφασης που βοηθούν στην ταξινόμηση των σημείων δεδομένων. Τα σημεία δεδομένων

που εμπίπτουν σε κάθε πλευρά του υπερπλάνου μπορούν να αποδοθούν σε διαφορετικές κατηγορίες. Επίσης, η διάσταση του υπερπλάνου εξαρτάται από τον αριθμό των χαρακτηριστικών. Εάν ο αριθμός των χαρακτηριστικών εισόδου είναι 2, τότε το υπερπλάνο είναι μόνο μια γραμμή. Εάν ο αριθμός των χαρακτηριστικών εισόδου είναι 3, τότε το υπερπληρωμή γίνεται ένα δισδιάστατο επίπεδο. Είναι δύσκολο να φανταστεί κανείς πώς θα είναι το υπερπλάνο όταν ο αριθμός των χαρακτηριστικών ξεπερνά τα 3.



Τα διανύσματα υποστήριξης είναι σημεία δεδομένων που είναι πιο κοντά στο υπεπλάνπ και επηρεάζουν τη θέση και τον προσανατολισμό του υπερπλάνου. Χρησιμοποιώντας αυτούς τους φορείς υποστήριξης, μεγιστοποιούμε το περιθώριο του ταξινομητή. Η διαγραφή των διανυσμάτων υποστήριξης θα αλλάξει τη θέση του υπερπλάνου. Αυτά είναι τα σημεία που μας βοηθούν να χτίσουμε το SVM μας.

Συνάρτηση κόστους και ενημέρωση κλίσεων

Στον αλγόριθμο SVM, επιδιώκουμε να μεγιστοποιήσουμε το περιθώριο μεταξύ των σημείων δεδομένων και του υπερπλάνου. Η συνάρτηση κόστους που βοηθά στη μεγιστοποίηση του περιθωρίου είναι:

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases} \qquad c(x, y, f(x)) = (1 - y * f(x))_+$$

Hinge loss function (function on left can be represented as a function on the right)

Το κόστος είναι 0 εάν η προβλεπόμενη τιμή και η πραγματική τιμή είναι του ίδιου σημείου. Εάν δεν είναι, τότε υπολογίζουμε την τιμή της ζημίας. Προσθέτουμε επίσης μια παράμετρο κανονικοποίησης στη συνάρτηση κόστους. Ο στόχος της παραμέτρου κανονικοποίησης είναι η εξισορρόπηση της μεγιστοποίησης περιθωρίου και της απώλειας. Μετά την προσθήκη της παραμέτρου κανονικοποίησης, οι λειτουργίες κόστους φαίνονται όπως παρακάτω.

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

Loss function for SVM

Τώρα που έχουμε τη συνάρτηση κόστους, παίρνουμε μερικές παραγώγου σε σχέση με τα βάρη για να βρούμε τις κλίσεις. Χρησιμοποιώντας τις κλίσεις, μπορούμε να ενημερώσουμε τα βάρη μας.

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 \quad = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} \left(1 - y_i \langle x_i, w \rangle\right)_+ \quad = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

Gradients

Όταν δεν υπάρχει λανθασμένη ταξινόμηση, δηλαδή το μοντέλο μας προβλέπει σωστά την κλάση του σημείου δεδομένων μας,

πρέπει μόνο να ενημερώσουμε την κλίση από την παράμετρο κανονικοποίησης

$$w = w - \alpha \cdot (2\lambda w)$$

Gradient Update—No misclassification

.

Οταν υπάρχει λανθασμένη ταξινόμηση, για παράδειγμα όταν το μοντέλο μας κάνει λάθος στην πρόβλεψη της κλάσσης του σημείου δεδομένου δίνουμε το κόστος με την παράμετρο κανονικοποίησης για να εκτελέσουμε την ενημέρωση των κλίσεων.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Gradient Update—Misclassification

3.5.2 Λογιστική Παλινδρόμηση

Σε προβλήματα ταξινόμησης, θέλουμε να καθορίσουμε την πιθανότητα μια παρατήρηση να ανήκει ή όχι σε μια συγκεκριμένη κλάση. Επομένως, επιθυμούμε να εκφράσουμε την πιθανότητα με μια τιμή μεταξύ του 0 και του 1.΄Ενας απλός αλγόριθμος ταξινόμησης που δημιουργεί τιμές αυτής της μορφης είναι ο ταξινομητής λογιστικής παλινδρόμησης.

Ας υποθέσουμε ότι έχουμε ένα απλό πρόβλημα δυαδικής ταξινόμησης, όπως αυτό που περιγράφηκε νωρίτερα στο ίδιο Κεφάλαιο.΄Εστω x = x1, . . . , xN τα διανύσματα εισόδου όπου yi ∈ 0, 1. Η συνάρτηση ενεργοποίησης του LR ταξινομητή καθορίζεται από την εφαρμογή μιας σιγμοειδούς συνάρτησης πάνω στην γραμμική παλινδρόμηση ούτως έστε να λάβουμε

την τελική απόφαση ταξινόμησης.Όπως περιγράφηκε στα SVMs:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Η συνάρτηση ενεργοποίησης της LR για ένα δοσμένο διάνυσμα x ορίζεται ως εξής:

$$h_w(x) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

Η συνάρτηση κόστους που θέλουμε να ελαχιστοποιηθεί κατά τη διάρκεια της εκπαίδευσης είναι η εξής:

$$\frac{1}{2}\|w\| + C \sum_{i=1}^{N} \log(\exp(-y_i(w^T + b)) + 1)$$

όπου C > 0 και b είναι οι συντελεστές που αναπαριστούν την τιμωρία (penalty) των

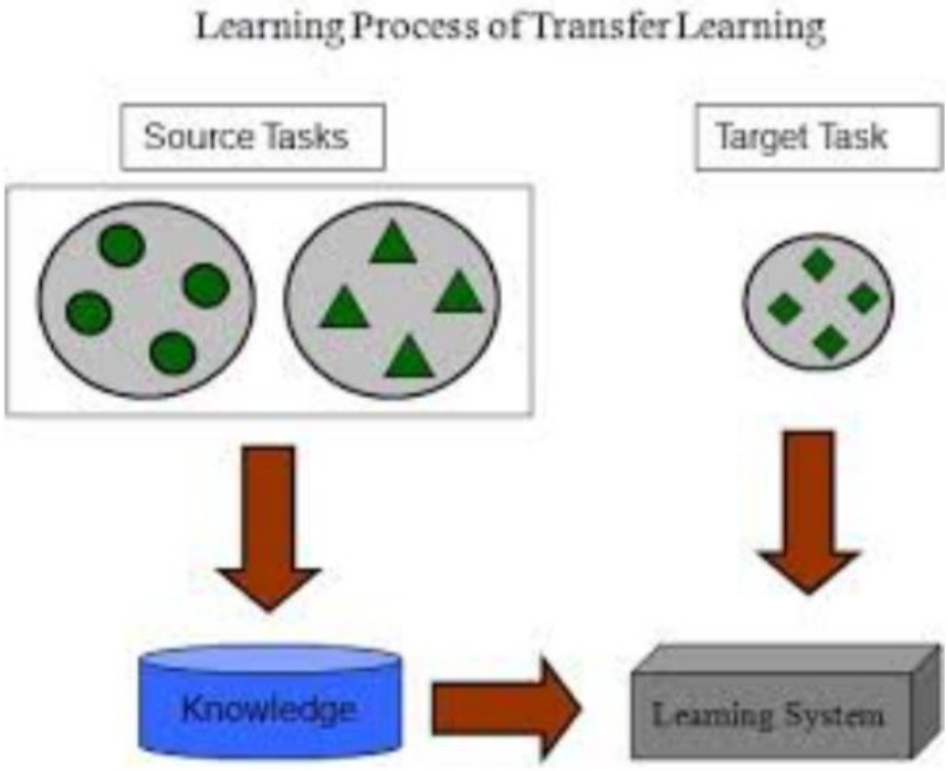
λανθασμένων αποτελεσμάτων ταξινόμησης και την τομή του υπερεπιπέδου αντίστοιχα.

3.6 Μεταφορά Μάθησης

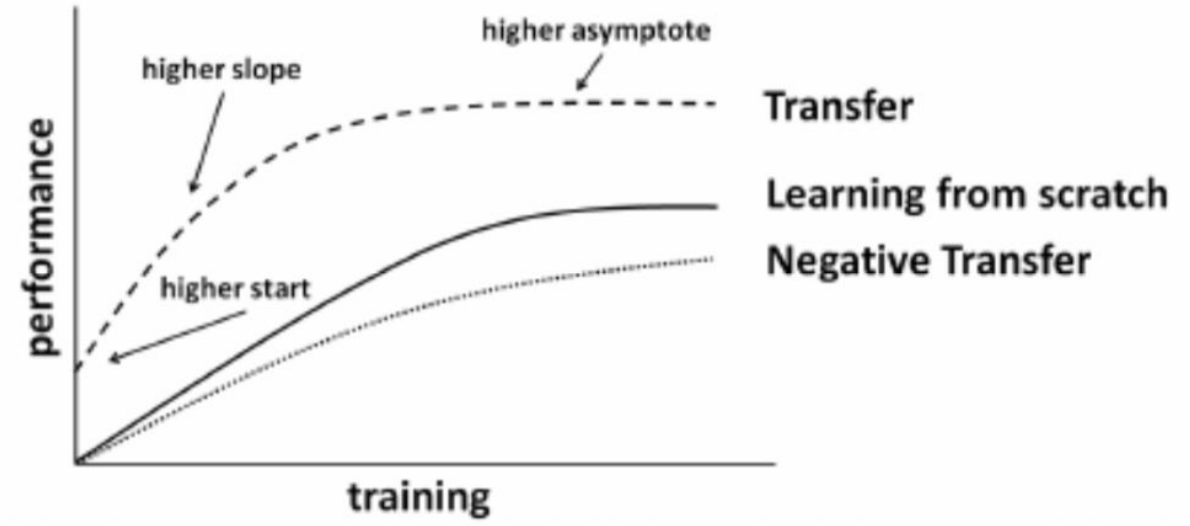
Μεταφορά μάθησης (Transfer Learning)[3] ονομάζουμε το πρόβλημα της μηχανικής μάθησης, στο οποίο προσπαθούμε να αξιοποιήσουμε την γνώση που απέκτησε ένα σύστημα σε ένα πρόβλημα, σε ένα διαφορετικό αλλά σχετικό πρόβλημα. Η μέθοδος αυτή αξιοποιείται πολύ από τα βαθιά νευρωνικά δίκτυα, γιατί αυτά απαιτούν μεγάλο αριθμό δεδομένων για να εκπαιδευτούν. Θεωρητικά, άμα δεν έχουμε αρκετά δεδομένα

για ένα πρόβλημα, μπορούμε να εκπαιδεύσουμε ένα δίκτυο σε ένα σχετικό πρόβλημα, στο οποίο έχουμε περισσότερα δεδομένα και να χρησιμοποιήσουμε τη γνώση που απέκτησε στο αρχικό μας πρόβλημα. Στη μηχανική μάθηση (και ειδικά στη βαθιά μάθηση), αντιμετωπίζουμε ένα πολύ σημαντικό πρόβλημα. Αυτό είναι το γεγονός ότι τα δίκτυα που επιλύουν περίπλοκα προβλήματα απαιτούν τεράστιες ποσότητες δεδομένων. Ωστόσο, η απόκτηση αυτών των δεδομένων για τα επιβλεπόμενα μοντέλα είναι συχνά ανέφικτη λόγω χρονικών ή υπολογιστικών περιορισμών. Επιπλέον, τα μοντέλα που έχουν εκπαιδευτεί σε μικρά, ειδικά σύνολα δεδομένων έχουν χειρότερη απόδοση όταν χρησιμοποιούνται για να αντιμετωπίσουν ένα διαφορετικό πρόβλημα, το οποίο μπορεί να είναι σχετικά παρεμφερές με το πρόβλημα στο οποίο έχουν εκπαιδευτεί.

Ο στόχος της μεταφοράς μάθησης είναι να βελτιώσει την εκμάθηση του προβλήματος-στόχου (target task) αξιοποιώντας γνώση από το πρόβλημα-πηγή (source task), όπως φαίνεται στο παρακάτω σχήμα.



Υπάρχουν τρεις τρόποι με τους οποίους συνήθως η μεταφορά μάθησης βελτιώνει τη διαδικασία εκπαίδευσης, οι οποίοι φαίνονται στο παρακάτω. Πρώτον, η αρχική απόδοση που επιτυγχάνεται στο target task χρησιμοποιώντας μόνο τη γνώση που έχει μεταφερθεί από το source task, προτού εκπαιδευτεί παραπάνω, σε σχέση με την αρχική απόδοση ενός τυχαία αρχικοποιημένου μοντέλου. Δεύτερον, ο χρόνος που χρειάζεται για να εκπαιδευτεί πλήρως το μοντέλο στο target task δεδομένης της γνώσης που έχει μεταφερθεί, σε σχέση με το χρόνο που χρειάζεται για να το μάθει εξαρχής. Τρίτον, το τελικό επίπεδο απόδοσης που επιτυγχάνεται στο target task σε σχέση με το τελικό επίπεδο χωρίς μεταφορά μάθησης [22].



Η περίπτωση της μη-επιβλεπόμενης προεκπαίδευσης (unsupervised pretraining)

Μία συγκεκριμένη περίπτωση μεταφοράς μάθησης είναι όταν το σουρςε τασκ είναι μη επιβλεπόμενο και το target task είναι επιβλεπόμενο. Αυτή η περίπτωση έχει ιδιαίτερο ενδιαφέρον, καθώς πολύ συχνά έχουμε διαθέσιμες μεγάλες ποσότητες μη επιβλεπόμενων δεδομένων εκπαίδευσης, αλλά πολύ λίγα δεδομένα εκπαίδευσης με ετικέτες. Η εκπαίδευση με επιβλεπόμενες τεχνικές στο επισημασμένο υποσύνολο πολλές φορές οδηγεί σε overfitting. Αποκτώντας ποιοτικές αναπαραστάσεις από τα μη επιβλεπόμενα δεδομένα, το μοντέλο μας μπορεί να έχει καλύτερη απόδοση στο πρόβλημα επιβλεπόμενης μάθησης που αντιμετωπίζουμε [23].

Αυτή η περίπτωση μεταφοράς μάθησης ονομάζεται μη-επιβλεπόμενη προεκπαίδευση (unsupervised pretraining). Αυτή η διαδικασία αποτελεί παράδειγμα του πώς μια ανα-παράσταση που έχει δημιουργηθεί από το μοντέλο, όταν αυτό αντιμετωπίζει ένα συγκεκριμένο πρόβλημα (μη επιβλεπόμενο) μπορεί κάποιες φορές να είναι χρήσιμη για ένα άλλο πρόβλημα (επιβλεπόμενο). Ονομάζεται προεκπαίδευση (pretraining),

επειδή αποτελεί μόνο το πρώτο βήμα προτού ένας αλγόριθμος εκπαίδευσης εφαρμοστεί για να προσαρμόσει (fine-tune) όλα τα επίπεδα μαζί. Ως προς το πρόβλημα επιβλεπόμενης μάθησης, μπορεί να θεωρηθεί ένας όρος κανονικοποίησης και αρχικοποίησης των παραμέτρων.

Κανονικοποίηση

Είναι πιθανό ότι η προεκπαίδευση αρχικοποιεί ένα βαθύ νευρωνικό δίκτυο σε μία περιοχή που θα ήταν αλλιώς απροσπέλαστη - για παράδειγμα, μια περιοχή που περιτρι-γυρίζεται από περιοχές όπου η συνάρτηση κόστους εναλλάσσεται τόσο πολύ από το ένα παράδειγμα στο άλλο που μπορεί να υπολογιστεί μόνο μια εκτίμηση του gradient που περιέχει πολύ θόρυβο.

Αρχικοποίηση παραμέτρων

Η προεκπαίδευση, στις περισσότερες περιπτώσεις, βελτιώνει την απόδοση στο επιβλεπόμενο πρόβλημα. Η βασική ιδέα είναι πως κάποια χαρακτηριστικά που είναι χρήσιμα για την επίλυση του μη επιβλεπόμενου προβλήματος είναι επίσης χρήσιμα και για τη ν επίλυση του επιβλεπόμενου προβλήματος.

Κεφάλαιο 4

Συνελικτικα Νευρωνικά Δίκτυα

4.1 Εισαγωγή

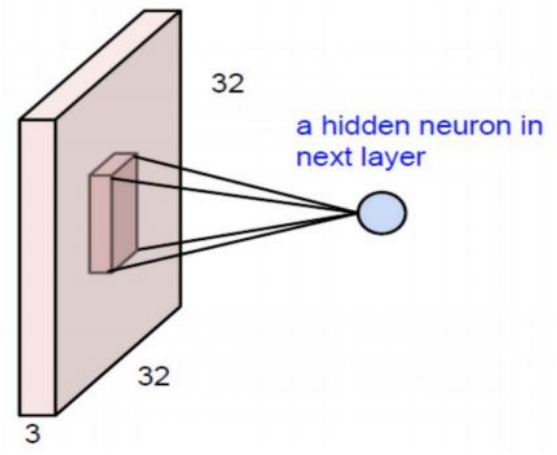
Τα τελευταία χρόνια τα Συνελικτικά Νευρωνικά Δίκτυα(Convolutional Neural Networks-CNNs) έχουν κορυφαίες επιδόσεις στην αναγνώριση προτύπων με κυριότερη εφαρμογή τους την ταξινόμηση εικόνων. Κάτι που απασχόλησε πολύ τους ερευνητές των CNN είναι αν εντοπίζουν μόνο χαρακτηριστικά που έχουν χωρική σχέση μεταξύ τους. Μια ακόμα πολύ σημαντική πλευρά των CNN είναι ο εντοπισμός αυθαίρετων (abstract) χαρακτηριστικών καθώς οι εικόνες τροφοδοτούνται στο δίκτυο. Παρακάτω αναλύονται τα στοιχεία τους[28].

4.2 Στοιχεία Συνελικτικού Νευρωνικού Δικτύου

4.2.1 Συνέλιξη

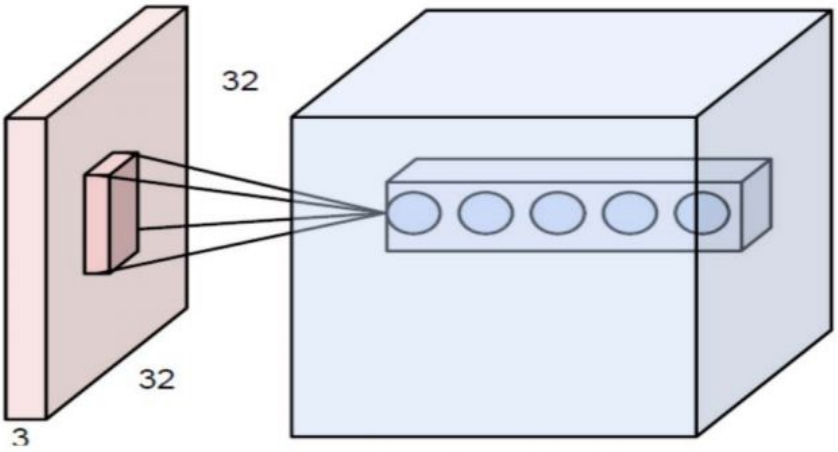
Αντί να στέλνουμε πληροφορία για ολόκληρη την εικόνα στους νευρώνες των επόμενων στοιβάδων, ψάχνουμε για μικρές περιοχές γύρω από τα εικονοστοιχεία, όπως φαίνεται στο παρακάτω σχήμα. Με άλλα λόγια οι νευρώνες των επόμενων

στοιβάδων δέχονται σαν είσοδο μόνο το αντίστοιχο τμήμα εικονας των νευρώνων της προηγούμενης τους στοιβάδας, για παράδειγμα μια 5x5 γειτονιά.



Convolution as alternative for fully connected network.

Με αυτό τον τρόπο μειώνεται πολύ το πλήθος των παραμέτρων στο μοντέλο μας. Επιπλέον, αν ρίξουμε μια ματιά πιο βαθιά, βλέπουμε οτι το σημαντικότερο πλεονέκτημα που παρέχει η συνέλιξη είναι ότι επιτρέπει τον εντοπισμό και την αναγνώριση χαρακτηριστικών ανεξαρτήτως της θέσης τους στην εικόνα. Γι αυτό το λόγο αποκαλούνται συνελίξεις[29,30].
Για να γίνει ακόμα πιο αποδοτική η ιδέα της συνέλιξης, χρησιμοποιούμε πολλές στοιβάδες σε κάθε στρώμα του CNN, ώστε να έχουμε τη δυνατότητα να χρησιμοποιήσουμε πολλαπλά φίλτρα και άρα να εξάγουμε διαφορετικά χαρακτηριστικά απο την εικόνα μας.



Multiple layers which each of them correspond to different filter
but looking at the same region in the given image

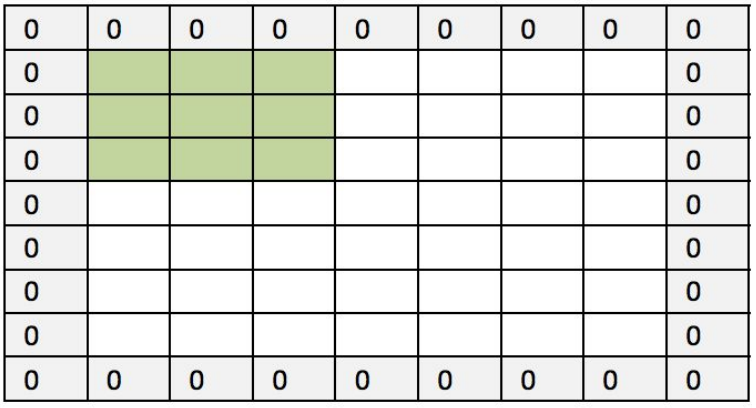
Η ιδέα των πολλαπλών φίλτρων ανά στρώμα φαίνεται στη παραπάνω εικόνα.

4.2.2 Stride

Στην πραγματικότητα το CNN δίνει πολλές επιλογές για μείωση των παραμέτρων του μοντέλου και ταυτόχρονα για την αποφυγή παράπλευρων συνεπειών. Μια από αυτές τις επιλογές είναι το stride. Επειδή υπάρχουν πολλές επικαλύψεις (overlaps) μεταξύ των συνελίξεων των γειτονικών εικονοστοιχείων, το stride μας δίνει τη δυνατότητα να μετακινούμε με βήμα μεγαλύτερο του ενός τον πυρήνα συνέλιξης και άρα να μην επανεξετάζουμε κάποια εικονοστοιχεία. Ετσι, δε μειώνουμε μόνο την επικάλυψη αλλά και τις διαστάσεις της επόμενης στοιβάδας.

4.2.3 Padding

Ένα από τα μειονεκτήματα της συνέλιξης είναι η έλλειψη πληροφορίας για την γειτονική περιοχή των εικονοστοιχείων στα άκρα(borders) της εικόνας. Η λύση σε αυτό το ζήτημα είναι η προσθήκη μηδενικών(zero padding) έξω από τα σύνορα της εικόνας ώστε μπορεί να γίνει η συνέλιξη και στα άκρα της. Συγκεκριμένα, προσθέτουμε τόσες σειρες και στήλες μηδενικών όσο το μισό της διάστασης του πυρήνα του φίλτρου, όπως φαίνεται στη παρακάτω εικόνα.



Οπότε, η διάσταση της εξόδου μιας εικόνας διάστασης W, ύστερα από P γραμμές(ή στήλες) zero padding, stride S και διάσταση φίλτρου K είναι:

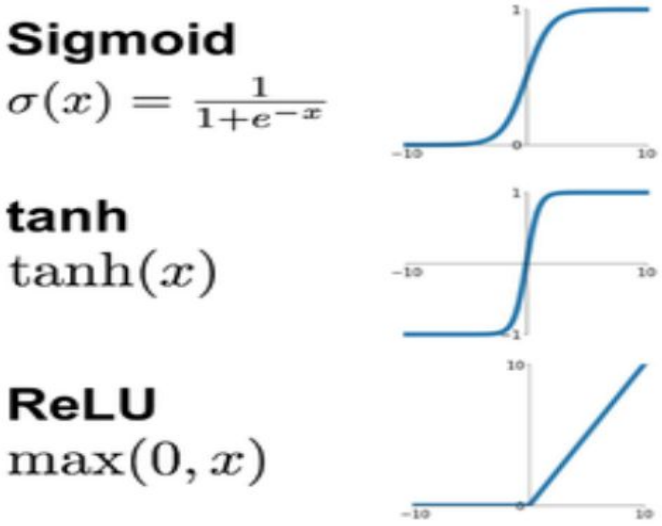
$$O = \frac{(W - K + 2P)}{S} + 1$$

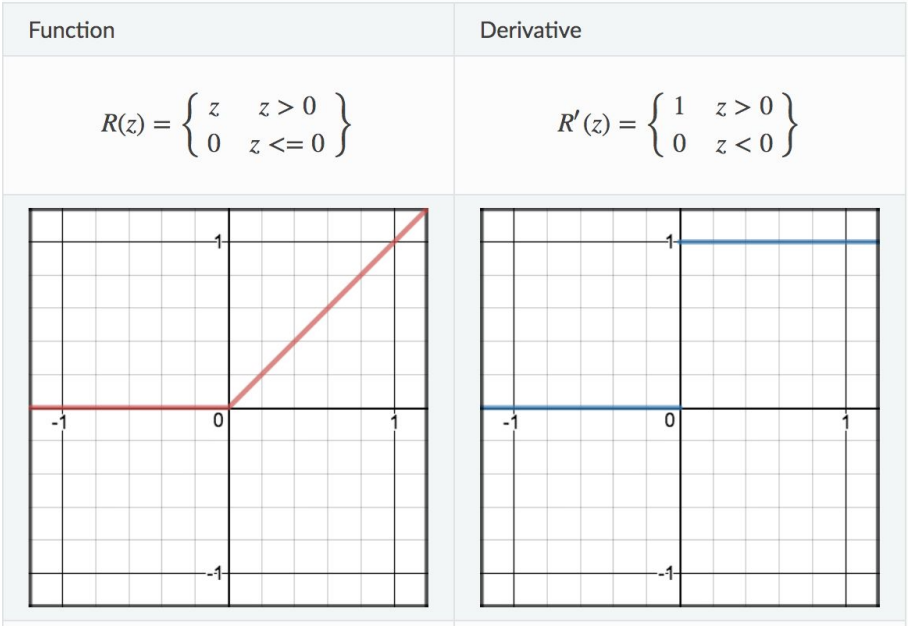
4.3 Μη-Γραμμικότητα

Η επόμενη στοιβάδα ύστερα από την συνελικτική είναι η μη-γραμμικότητα. Χρησιμοποιείται για να προσαρμόσει ή να κόψει (cut-off) την παραγόμενη έξοδο. Ο ρόλος της στοιβάδας αυτής είναι είτε να διαβρέξει(saturate) είτε να περιορίσει την προηγούμενη έξοδο.

Για πολλά χρόνια δημοφιλής ήταν η σιγμοειδής(sigmoid) ή η υπερβολική εφαπτομένη(tanh). Ωστόσο, πρόσφατα χρησιμοποιείται η Rectified Linear Unit (ReLU) για τους εξής λόγους:

- Εχει απλούστερο ορισμό και σαν συνάρτηση και η κλίση της.



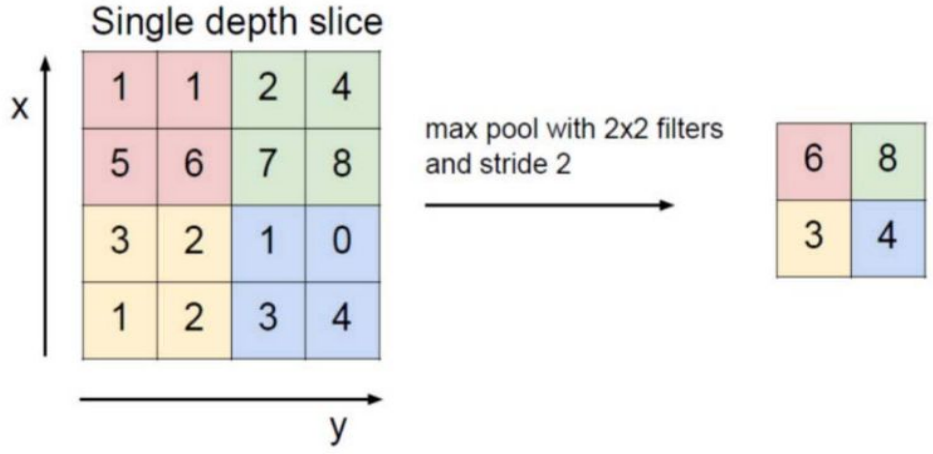


- Η σιγμοειδής και η υπερβολική εφαπτομένη δημιουργούν το λεγόμενο vanishing gradient πρόβλημα όσο πιο βαθιά είναι η αρχιτεκτονική του δικτύου. Αυτό το αποφεύγει η ReLU καθώς έχει σταθερή θετική κλίση.
- Η ReLU δημιουργεί μια πιο αραιή αναπαράσταση(sparse representation) καθώς η μηδενική κλίση οδηγεί σε

απόκτηση ολόκληρων μηδενικών.

4.4 Στοιβάδα Pooling

Η κύρια ιδέα είναι υπο-δειγματοληψία για μείωση της πολυπλοκότητας των χαρακτηριστικών στις στοιβάδες που ακολουθούν. Στις εικόνες, αυτό σημαίνει μείωση της ανάλυσης. Η Pooling δεν επηρεάζει τον αριθμό των φίλτρων. Η Max-Pooling είναι η πιο διαδεδομένη στοιβάδα Pooling. Χωρίζει την εικόνα σε υπο-περιοχές και κρατάει μόνο τη μεγαλύτερη τιμή απο την κάθε υπο-περιοχή, όπως φαίνεται στη παρακάτω εικόνα:



Πρέπει να τονιστεί ότι η υπο-δειγματοληψία δεν μειώνει την χωρική πληροφορία της εικόνας[31].

4.5 Πλήρως Συνδεδεμένη Στοιβάδα

Η πλήρως συνελικτική στοιβάδα(fully connected layer-FC) είναι στοιβάδα που προέρχεται απο τα παραδοσιακά νευρωνικά δίκτυα. Οπότε κάθε νευρώνας στην FC στοιβάδα συνδέται με κάθε νευρώνα της προηγούμενης και της επόμενης στοιβάδας.

Το κύριο μειονέκτημα των πλήρως συνδεδεμένων στοιβάδων είναι η πολυπλοκότητα των παραμέτρων, η οποία κάνει πολύ αργή την εκπαίδευση. Οπότε μειώνουμε όσο μπορούμε το

πλήθος των κόμβων της στοιβάδας αυτής με τη χρήση της τεχνικής dropout.

Κεφάλαιο 5

Προτεινόμενο Σύστημα

Στο κεφάλαιο αυτό αναλύουμε τις μεθόδους που χρησιμοποιήσαμε στις δύο διαφορετικές προσεγγίσεις. Αρχικά κάνουμε μια αναφορά στο σύνολο δεδομένων που χρησιμοποιήθηκε για τη διεξαγωγή των πειραμάτων. Υστερα, προτείνουμε αρχιτεκτονικές υπολογιστικής όρασης και βαθιάς μάθησης για την όσο το δυνατόν καλύτερη προσέγγιση του προβλήματος της αναγνώρισης φαγητού.

5.1 Συλλογή Δεδομένων

Για τη διεξαγωγή των πειραμάτων μας, χρησιμοποιήσαμε ένα νέο μεγάλης κλίμακας σύνολο εικόνων φαγητού: *FOOD-101 Dataset.*
Αυτό το σύνολο δεδομένων περιλαμβάνει 101 κατηγορίες εικόνων φαγητού όπως *apple pie, baklava, cannoli, cheesecake, club-sandwich, donats, fish and chips, greek salad, pizza, omelette, samosa, tacos, sushi* etc.
Κάθε κατηγορία εικόνων περιλαμβάνει 1000 εικόνες, άρα συνολικά το σύνολο δεδομένων μας περιλαμβάνει 101000 εικόνες. Οι εικόνες αυτές έχουν διάσταση 512x384 ή 384x512.

Οι εικόνες αυτού του συνόλου δεδομένων στην εκάστοτε κατηγορία περιλαμβάνουν το φαγητό της κατηγορίας αυτής αλλά και τυχαία και άλλων. Έτσι, ήταν μια αρκετά προκλητική διαδικασία οι μέθοδοι που θα επιλέξουμε. Παρακάτω φαίνονται μερικές εικόνες του συνόλου αυτού δεδομένων.

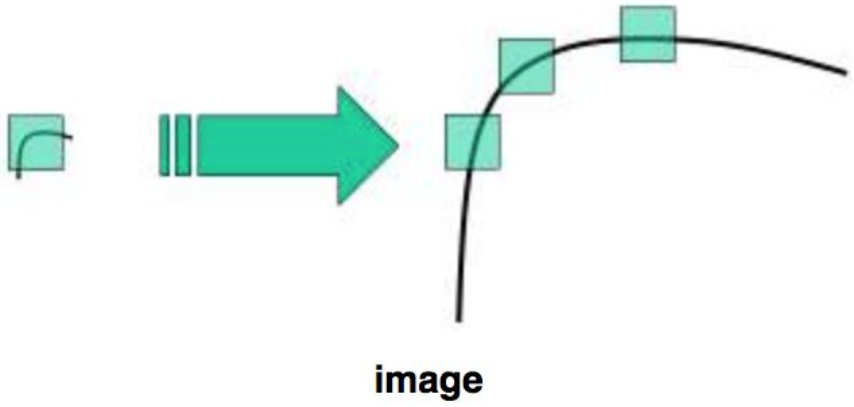



Το σύνολο δεδομένων *FOOD-101* μπορούμε να το κατεβάσουμε από την παρακάτω ηλεκτρονική διεύθυνση:
http://data.vision.ee.ethz.ch/cvl/food-101.tar.gz

5.2 Αναλλοίωτος σε Κλίμακα Μετασχηματιμός

Σε προηγούμενο κεφάλαιο είδαμε τον ανιχνευτή γωνιών Harris. Είναι αμετάβλητος σε περιστροφή, που σημαίνει ότι αν μια

εικόνα περιστραφεί, θα βρούμε τις ίδιες γωνίες. Είναι προφανές καθώς οι γωνίες παραμένουν γωνίες αν περιστρέψουμε την εικόνα. Αλλά σε ότι αφορά την κλίμακα, τι συμβαίνει? Μία γωνία μπορεί να μην αποτελεί γωνία αν αλλάξει η κλίμακα στην εικόνα. Στην παρακάτω εικόνα είναι εμφανές ότι η αλλαγή κλίμακας σε μία εικόνα μπορεί να επηρεάσει τις γωνίες που εντοπίζονται.



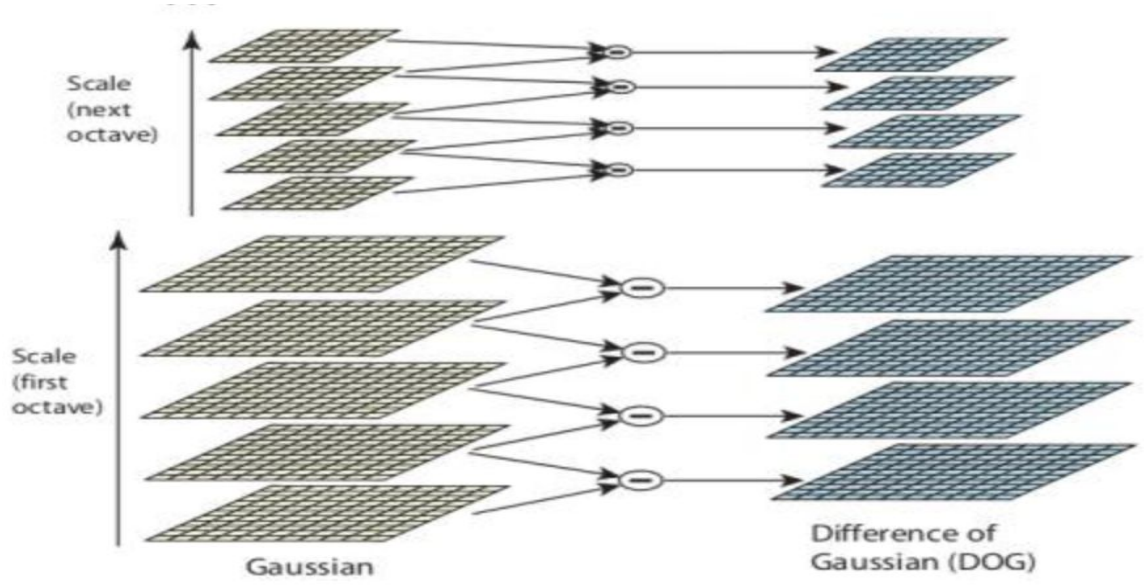
Ακολουθούμε 4 βήματα για τον αναλλοίωτο σε κλίμακα μετασχηματισμό[4].

1. Εύρος ανίχνευσης χώρου κλίμακας

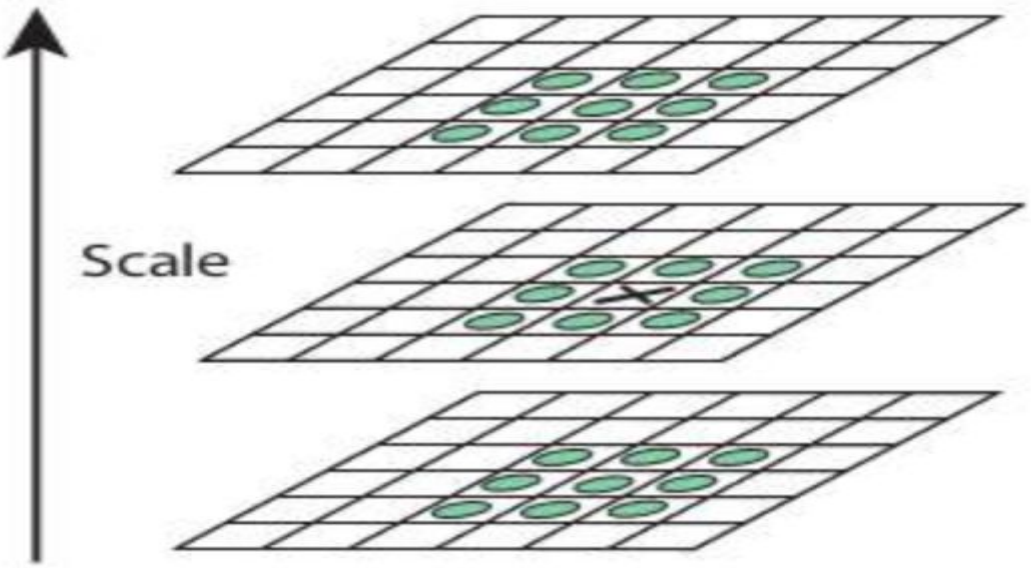
Από την παραπάνω εικόνα, είναι προφανές ότι δεν μπορούμε να χρησιμοποιήσουμε το ίδιο παράθυρο για την ανίχνευση σημείων κλειδιών με διαφορετική κλίμακα. Είναι εντάξει με μικρή γωνία. Αλλά για να ανιχνεύσουμε μεγαλύτερες γωνίες χρειαζόμαστε μεγαλύτερα παράθυρα. Για αυτό, χρησιμοποιείται φιλτράρισμα χώρου κλίμακας. Σε αυτό, βρίσκουμε το Laplacian of Gaussian για την εικόνα με διάφορες τιμές σ. Το LoG ενεργεί ως ανιχνευτής blobs, ο οποίος ανιχνεύει τα blobs σε διάφορα μεγέθη λόγω αλλαγής σ. Ητοι, το σ λειτουργεί ως παράμετρος κλιμάκωσης. Για παράδειγμα, στην παραπάνω εικόνα, ο πυρήνας gaussian με χαμηλό σ δίνει μεγάλη αξία για τη μικρή γωνία ενώ ο πυρήνας guassian με υψηλό σ ταιριάζει καλά για μεγαλύτερη γωνία. Έτσι, μπορούμε να βρούμε τα τοπικά

μέγιστα σε όλη την κλίμακα και το διάστημα που μας δίνει μια λίστα με τιμές (x, y, σ) που σημαίνει ότι υπάρχει ένα πιθανό σημείο κλειδί στα (x, y) στην σ κλίμακα.

Αλλά το LoG είναι υπολογιστικά ακριβό, οπότε ο αλγόριθμος SIFT χρησιμοποιεί τη διαφορά Gaussians που είναι μια προσέγγιση του LoG. Η διαφορά του Gaussian λαμβάνεται ως η διαφορά Gaussian θόλωσης μιας εικόνας με δύο διαφορετικές σ, ας είναι σ και kσ. Αυτή η διαδικασία γίνεται για διαφορετικές οκτάβες της εικόνας στη Γκαουσιανή Πυραμίδα. Αναπαρίσταται στην παρακάτω εικόνα:



Μόλις εντοπιστεί το DoG, στις εικόνες αναζητούνται τοπικά άκρα σε κλίμακα και χώρο. Για παράδειγμα, ένα εικονοστοιχείο σε μια εικόνα συγκρίνεται με τους 8 γείτονές του, καθώς και 9 εικονοστοιχεία στην επόμενη κλίμακα και 9 εικονοστοιχεία σε προηγούμενες κλίμακες. Αν είναι τοπικό ακρότατο, είναι ένα πιθανό σημείο κλειδί. Βασικά αυτό σημαίνει ότι το σημείο κλειδί αντιπροσωπεύεται καλύτερα σε αυτή την κλίμακα. Αυτό φαίνεται παρακάτω:



2. Τοποθέτηση Σημείου κλειδι

Μόλις εντοπιστούν οι πιθανές τοποθεσίες σημείων κλειδιών, πρέπει να αναθεωρηθούν για να επιτευχθούν ακριβέστερα αποτελέσματα. Χρησιμοποιούμε σειρές Taylor επέκτασης κλίμακας χώρου για να πάρουμε ακριβέστερη θέση ακραίων τιμών και εάν η ένταση σε αυτά τα όρια είναι μικρότερη από μια τιμή κατωφλίου, απορρίπτεται.

Το DoG έχει υψηλότερη απόκριση για τις εκμές, επομένως και οι ακμές πρέπει να αφαιρεθούν. Για αυτό, χρησιμοποιείται μια λογική παρόμοια με την ανίχνευση γωνίας Harris. Χρησιμοποιήθηκε ένας 2x2 Hessian matrix (H) για να υπολογιστεί η κύρια καμπυλότητα. Γνωρίζουμε από τον ανιχνευτή γωνιών Harris ότι για τις άκρες, μία ιδιοτιμή είναι μεγαλύτερη από την άλλη. Έτσι, εδώ χρησιμοποιήθηκε μια απλή λειτουργία:
Αν αυτός το κλάσμα είναι μεγαλύτερο από ένα κατώφλι, τότε το σημείο κλειδί απορρίπτεται.

Έτσι εξαλείφει όλα τα σημεία κλειδί χαμηλής αντίθεσης και τα σημεία κλειδί των ακμών και αυτό που παραμένει είναι τα ισχυρά σημεία ενδιαφέροντος.

3. Ανάθεση Προσανατολισμού

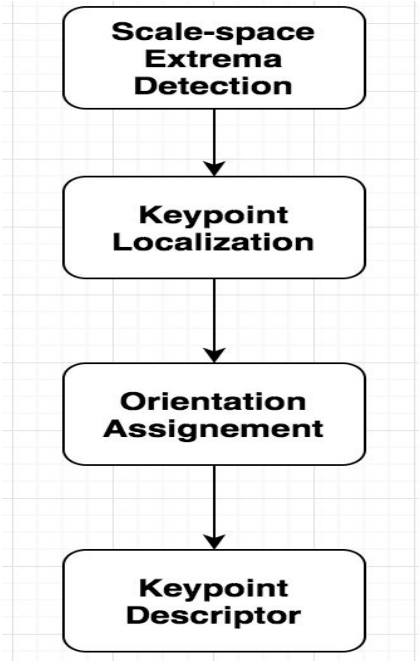
Τώρα ανατίθεται ένας προσανατολισμός σε κάθε βασικό σημείο για να επιτευχθεί η αμεταβλητότητα στην περιστροφή της εικόνας. Μια γειτονιά λαμβάνεται γύρω από τη θέση του βασικού σημείου ανάλογα με την κλίμακα και το μέγεθος και η κατεύθυνση της κλίσης υπολογίζονται σε αυτή την περιοχή. Παρασκευάζεται ιστόγραμμα προσανατολισμού με 36 κουτιά που καλύπτουν 360 μοίρες. Είναι σταθμισμένο από το μέγεθος κλίσης και το κυβικό παράθυρο που έχει σταθμιστεί με Gauss, με σ ίσο με 1,5 φορα την κλίμακα του βασικού σημείου. Λαμβάνεται η υψηλότερη κορυφή στο ιστόγραμμα και θεωρείται επίσης ότι υπολογίζεται ο προσανατολισμός κάθε κορυφής άνω του 80%. Δημιουργεί σημεία κλειδιά με την ίδια θέση και κλίμακα, αλλά διαφορετικές κατευθύνσεις. Συμβάλλει στη σταθερότητα της αντιστοίχισης.

4. Περιγραφητής Σημείου κλειδι

Τώρα δημιουργείται ο περιγραφητής του σημείου κλειδι. Μια γειτονιά 16x16 γύρω από το σημείο έχει ληφθεί. Είναι χωρισμένο σε 16 υπο-τετράγωνα μεγέθους 4x4. Για κάθε υπό-ομάδα δημιουργείται ιστόγραμμα προσανατολισμού 8 κουτιών. Επομένως, είναι διαθέσιμες συνολικά 128 τιμές κουτιού και χρησιμοποιούνται ως άνυσμα για την αναπαράσταση του περιγραφητή του σημείου κλειδί. Επιπλέον, λαμβάνονται διάφορα μέτρα για την επίτευξη ευρωστίας έναντι αλλαγών φωτισμού, περιστροφής κλπ.

Η ανίχνευση των σημείων κλειδί και της κλίμακάς τους φαίνεται στην παρακάτω εικόνα:

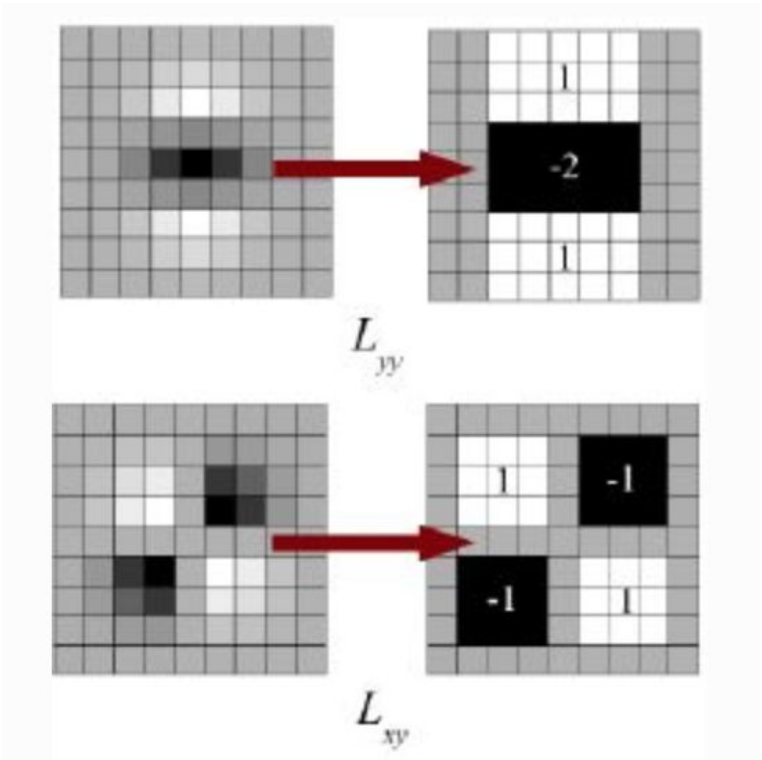
Παρακάτω φαίνεται συνοπτικά η διαδικασία που ακολουθούμε για τον υπολογισμό των SIFT χαρακτηριστικών και των περιγραφητών τους.



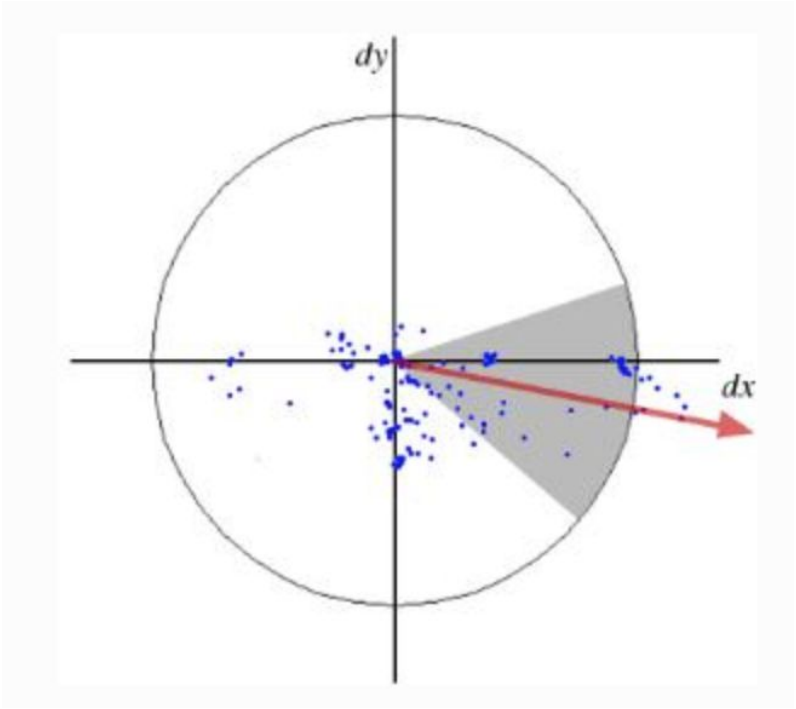
5.3 Επιταχυνόμενος Μετασχηματισμός

Στο κεφάλαιο αυτό, είδαμε το SIFT για την ανίχνευση και περιγραφή βασικών σημείων. Αλλά ήταν συγκριτικά αργή μέθοδος και χρειαζόμαστε πιο γρήγορη έκδοση. Το 2006, τρία άτομα, Bay, H., Tuytelaars, T. και Van Gool, L, δημοσίευσαν ένα άλλο έγγραφο, "SURF: Speed-up Robust Features"[5], που εισήγαγε ένα νέο αλγόριθμο που ονομάζεται SURF. Όπως υποδηλώνει το όνομα, πρόκειται για μια επιταχυνόμενη έκδοση του SIFT.

Στο SIFT, ο Lowe προσέγγισε τον Laplacian του Gaussian με τη διαφορά του Gaussian για την εύρεση κλίμακας χώρου. Ο SURF το διευρύνει λίγο περισσότερο και προσεγγίζει το LoG με το Φίλτρο Κουτιού. Η παρακάτω εικόνα δείχνει μια τέτοια προσέγγιση. Ένα μεγάλο πλεονέκτημα αυτής της προσέγγισης είναι ότι η συνέλιξη με το φίλτρο κουτιού μπορεί εύκολα να υπολογιστεί με τη βοήθεια ολοκληρωμένων εικόνων και μπορεί να γίνει παράλληλα για διαφορετικές κλίμακες. Επίσης, το SURF βασίζεται στον προσδιορισμό του Hessian πίνακα τόσο για την κλίμακα όσο και για την τοποθεσία.



Για ανάθεση προσανατολισμού, ο SURF χρησιμοποιεί αποκρίσεις wavelet σε οριζόντια και κάθετη κατεύθυνση. Επίσης, εφαρμόζονται επαρκή βάρη guassian. Στη συνέχεια, σχεδιάζονται σε ένα χώρο όπως παρουσιάζεται στην παρακάτω εικόνα. Ο κυρίαρχος προσανατολισμός υπολογίζεται με τον υπολογισμό του αθροίσματος όλων των αποκρίσεων μέσα σε ένα παράθυρο ολίσθησης με γωνία 60 μοίρες. Ενδιαφέρον είναι ότι, η απόκριση wavelet μπορεί να βρεθεί χρησιμοποιώντας ολοκληρωμένες εικόνες πολύ εύκολα σε οποιαδήποτε κλίμακα. Για πολλές εφαρμογές, δεν απαιτείται μεταστροφή της περιστροφής, οπότε δεν χρειάζεται να βρεθεί αυτός ο προσανατολισμός, ο οποίος επιταχύνει τη διαδικασία. Το SURF παρέχει μια τέτοια λειτουργικότητα που ονομάζεται Upright-SURF ή U-SURF. Βελτιώνει την ταχύτητα και είναι ανθεκτικός μέχρι και ±15 μοίρες.



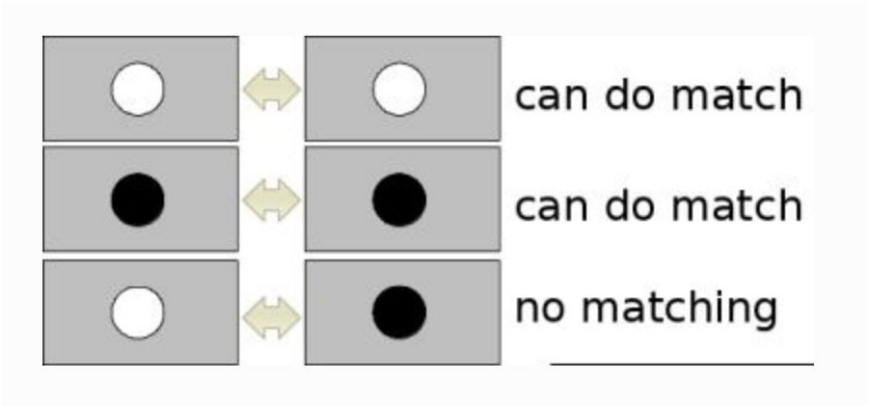
Για την περιγραφή χαρακτηριστικών, ο SURF χρησιμοποιεί αποκρίσεις Wavelet σε οριζόντια και κάθετη κατεύθυνση (και πάλι, η χρήση ολοκληρωμένων εικόνων διευκολύνει τα πράγματα). Μια γειτονιά μεγέθους 20σX20σ λαμβάνεται γύρω από το σημείο κλειδί όπου σ είναι η κλίμακα. Διαιρείται σε υποπεριοχές 4x4. Για κάθε υποπεριοχή, λαμβάνονται οριζόντιες και κάθετες αποκρίσεις wavlet και σχηματίζεται ένα διάνυσμα όπως αυτό:

$$v = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|).$$

Αυτό όταν αντιπροσωπεύεται ως διάνυσμα δίνει τον περιγραφητή χαρακτηριστικών SURF με συνολικά 64 διαστάσεις. Μειώστε τη διάσταση, αυξήστε την ταχύτητα του υπολογισμού και της αντιστοίχισης, αλλά παρέχετε καλύτερη διακριτικότητα των χαρακτηριστικών.

Μια σημαντική βελτίωση είναι η χρήση σημείου Laplacian (ίχνος Hessian Matrix) για το υποκείμενο σημείο ενδιαφέροντος. Δεν προσθέτει κανένα κόστος υπολογισμού, αφού έχει ήδη υπολογιστεί κατά τη διάρκεια της ανίχνευσης.

Το σημάδι του Laplacian διακρίνει τις φωτεινά blobs στο σκοτεινό υπόβαθρο από την αντίστροφη κατάσταση. Στο στάδιο αντιστοίχισης, συγκρίνουμε μόνο τα χαρακτηριστικά εάν έχουν τον ίδιο τύπο αντίθεσης (όπως φαίνεται στην παρακάτω εικόνα). Αυτές οι ελάχιστες πληροφορίες επιτρέπουν την ταχύτερη αντιστοίχιση, χωρίς να μειώνεται η απόδοση του περιγραφέα.

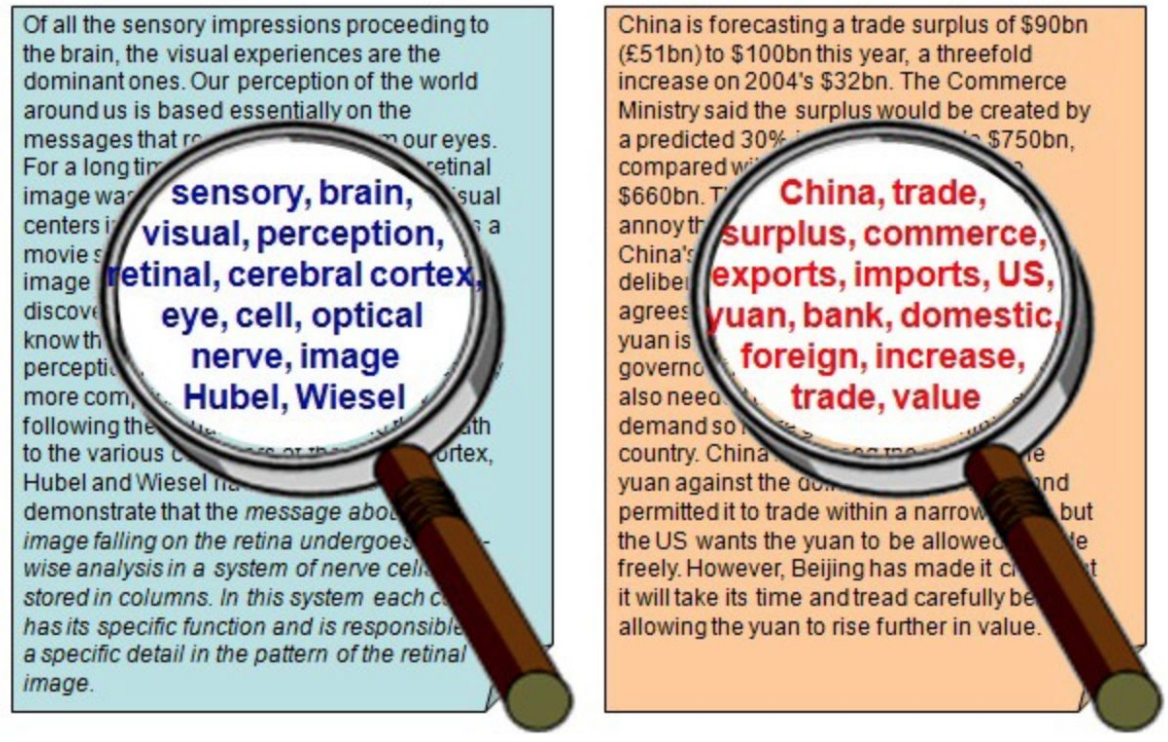


Με λίγα λόγια, ο SURF προσθέτει πολλά χαρακτηριστικά για τη βελτίωση της ταχύτητας σε κάθε βήμα. Η ανάλυση δείχνει ότι είναι 3 φορές ταχύτερη από το SIFT, ενώ η απόδοση είναι συγκρίσιμη με το SIFT. Το SURF είναι καλό στο χειρισμό εικόνων με θόλωση και περιστροφή, αλλά δεν είναι καλό στο χειρισμό της αλλαγής όψεων και της αλλαγής του φωτισμού.

Παρακάτω φαίνεται ένα παράδειγμα υπολογισμού των SURF χαρακτηριστικών με υψηλο Hessian κατωφλι (για λογους αναπαραστασης)

5.4 Τσάντα Οπτικών Λέξεων

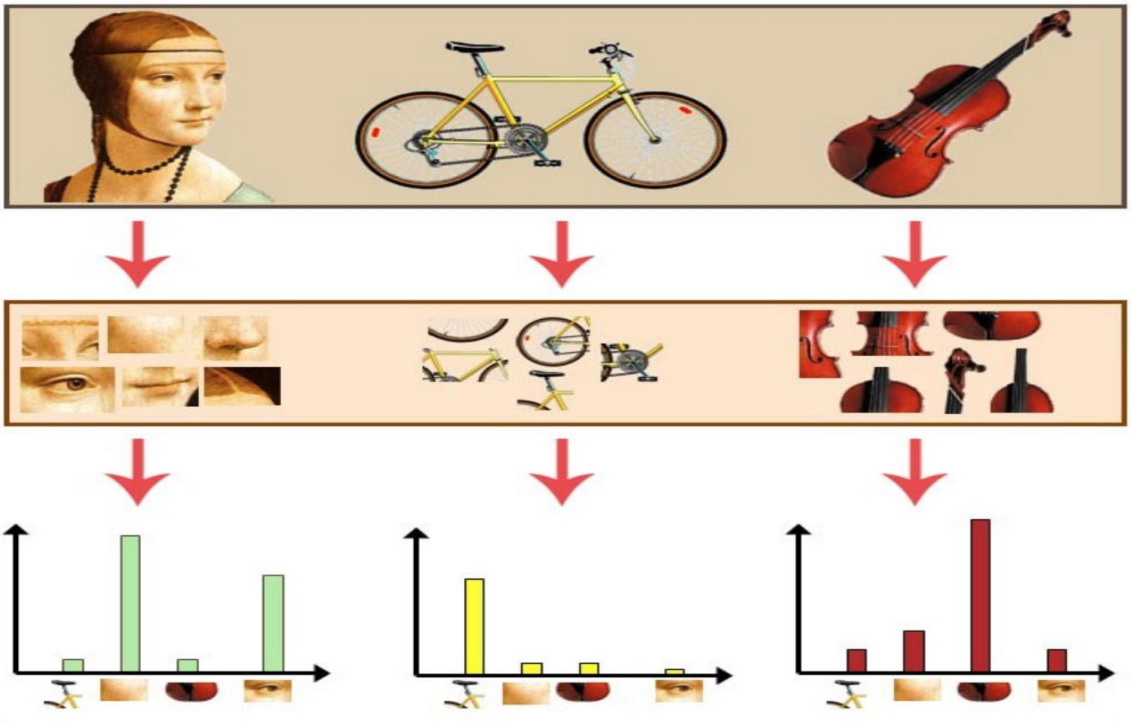
Η τσάντα με οπτικές λέξεις(bag of visual words-BOVW) [8] χρησιμοποιείται συνήθως στην ταξινόμηση εικόνων. Η ιδέα της είναι υιοθετημένη από την ανάκτηση πληροφοριών και την τσάντα λέξεων της επεξεργασίας φυσικής γλώσσας. Στην τσάντα των λέξεων(bag of words-BOW) μετράμε τον αριθμό εμφανίσεων κάθε λέξης που υπάρχει σε ένα έγγραφο. Χρησιμοποιούμε τη συχνότητα κάθε λέξης για να μάθουμε τις λέξεις-κλειδιά του εγγράφου και δημιουργούμε ένα ιστόγραμμα συχνοτήτων από αυτό. Αντιμετωπίζουμε ένα έγγραφο ως BOW. Έχουμε την ίδια ιδέα στο BOVW, αλλά αντί για λέξεις, χρησιμοποιούμε χαρακτηριστικά εικόνας ως λέξεις. Τα χαρακτηριστικά εικόνας είναι μοναδικό πρότυπο που μπορούμε να βρούμε σε μια εικόνα.



Keywords in documents

Η γενική ιδέα της τσάντας των οπτικών λέξεων (BOVW) είναι να αντιπροσωπεύει μια εικόνα ως σύνολο χαρακτηριστικών. Τα χαρακτηριστικά γνωρίσματα αποτελούνται από σημεία-κλειδιά

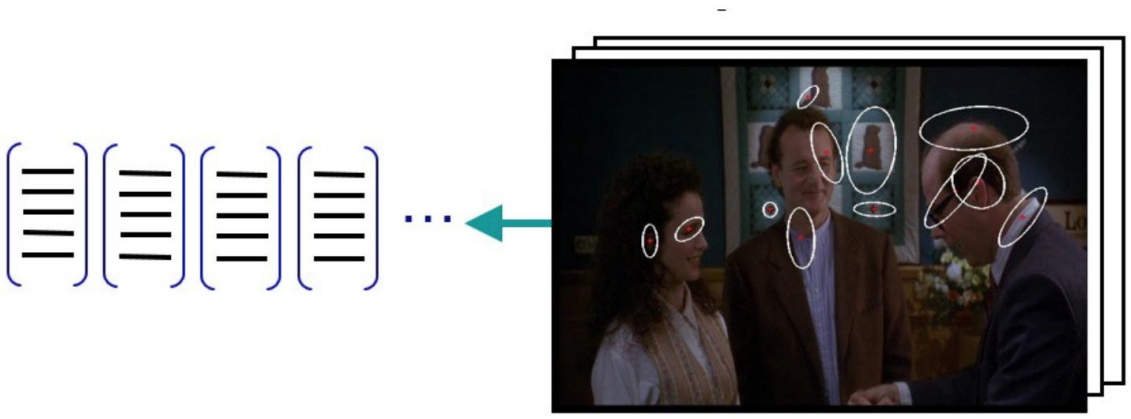
και περιγραφητές. Τα σημεία κλειδιά είναι τα σημεία "stand-out" σε μια εικόνα, οπότε δεν έχει σημασία η εικόνα να περιστρέφεται, να συρρικνώνεται ή να επεκτείνεται, τα σημεία κλειδιά της θα είναι πάντα τα ίδια. Και ο περιγραφητής είναι η περιγραφή του σημείου-κλειδί. Χρησιμοποιούμε τα σημεία κλειδί και τους περιγραφητές για να κατασκευάσουμε λεξιλόγια και να αναπαριστούμε κάθε εικόνα ως ιστογράμματα συχνοτήτων των χαρακτηριστικών που υπάρχουν στην εικόνα. Από το ιστόγραμμα συχνότητας, αργότερα, μπορούμε να βρούμε άλλες παρόμοιες εικόνες ή να προβλέψουμε την κατηγορία της εικόνας.



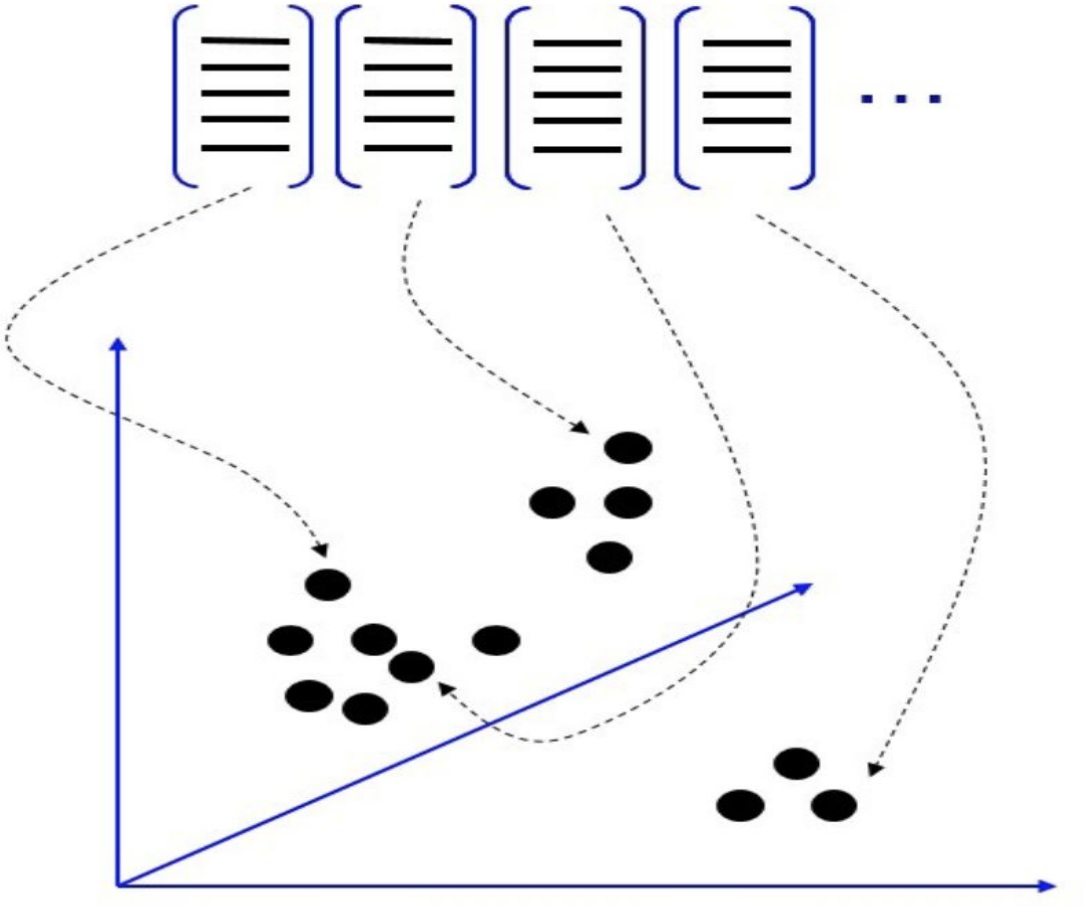
Histogram of visual words

Για να χτίσουμε την τσάντα οπτικών λέξεων ακολουθούμε την εξής διαδικασία:
Εντοπίζουμε χαρακτηριστικά και τους περιγραφητές τους για κάθε εικόνα στο σύνολο εικόνων μας. Υστερα χτίζουμε ένα οπτικό λεξικό για κάθε εικόνα. Ο εντοπισμός των χαρακτηριστικών και ο υπολογισμός των περιγραφητών τους

μπορεί να γίνει με αλγορίθμους που αναφέρθηκαν προηγουμένως(SIFT, SURF).



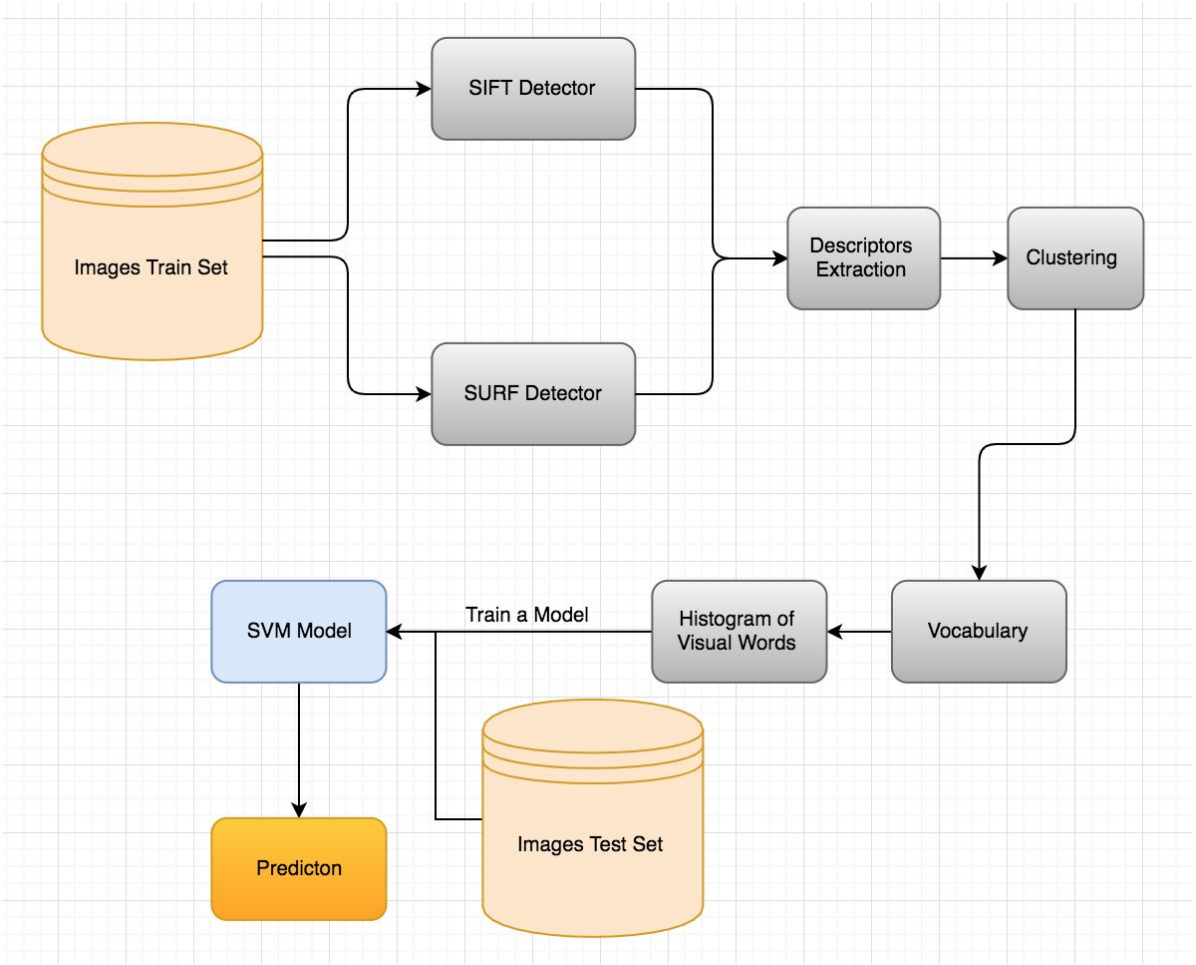
Ύστερα, δημιουργούμε ομάδες από τους περιγραφητές (χρησιμοποιώντας K-Means). Τα κέντρα των ομάδων μπορούν να χρησιμοποιηθούν σαν το λεξιλόγιο των οπτικών λεξικών.



Descriptors clustering

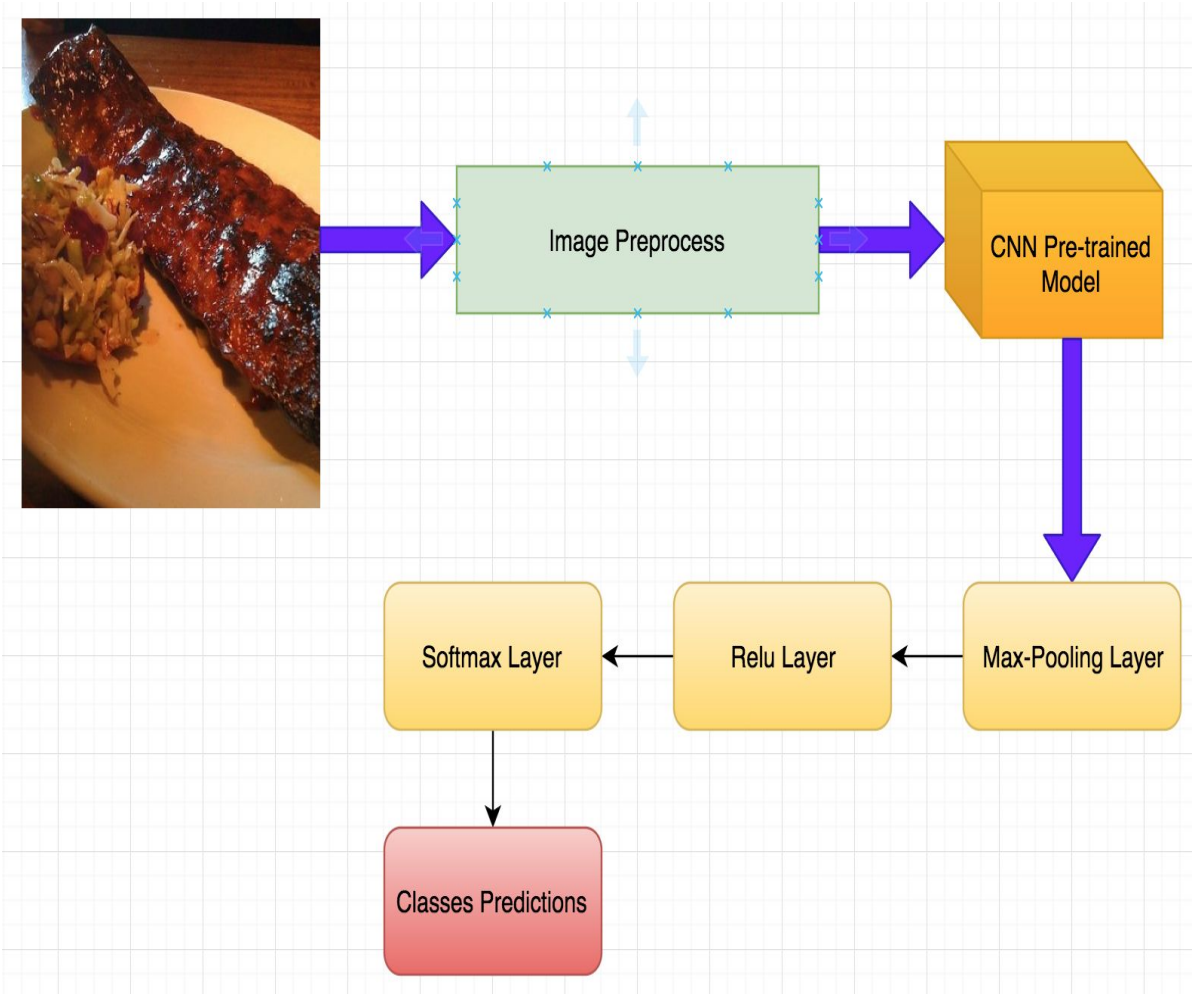
Για την εκπαίδευση, χρησιμοποιούμε Μηχανές Διανυσματικής Υποστήριξης(Support Vector Machines-SVM). Ο SVM εκπαιδεύεται πάνω στα ιστογράμματα των λεξιλογίων των εικόνων προς εκπαίδευση.
Για την ταξινόμηση, υπολογίζουμε αυτά τα κέντρα για τις εικόνες που θέλουμε να ταξινομησουμε και τις κατηγοριοποιούμε ανάλογα με την απόσταση των κέντρων τους από τα κέντρα των εικόνων προς εκπαίδευση.

Παρακάτω φαίνεται ένα διάγραμμα ροής της παραπάνω διαδικασίας:



5.5 Αρχιτεκτονική Συνελικτικού Δικτύου

Τώρα θα συζητήσουμε για τις τεχνικές αναγνώρισης βασισμένες σε συνελικτικά νευρωνικά δίκτυα.



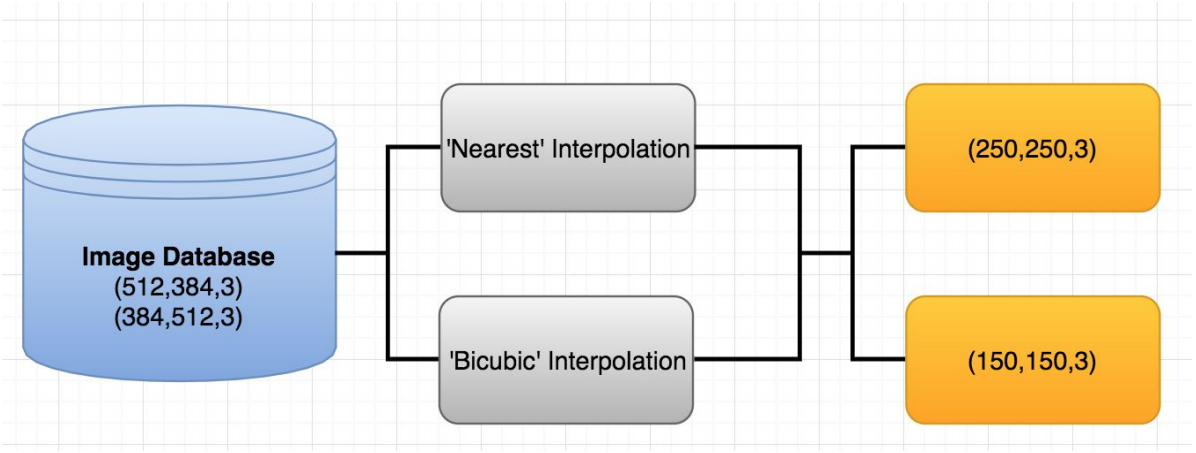
Στην παραπάνω εικόνα βλέπουμε την υψηλού επιπέδου αναπαράσταση της πορείας που ακολουθούμε για την αναγνώριση εικόνας με τη χρήση CNNs.

Η διαδικασία από άκρο σε άκρο έχει ως εξής.

Προεπεξεργασία εικόνας.

Σκοπός της ψηφιακής προ-επεξεργασίας εικόνας πριν αυτή εισέλθει στο νευρωνικό δίκτυο, είναι να διευκολύνει το μοντέλο

να εξάγει τα χαρακτηριστικά που χρειάζεται. Χρειάζεται να εστιάσει συνήθως σε ένα μόνο συγκεκριμένο εύρος εικονοστοιχείων και η προ-επεξεργασία εικόνας βοηθάει το μοντέλο μας να επικεντρωθεί σε αυτά.
Ένας ακόμα πολύ σημαντικός σκοπός της είναι η μείωση της διάστασης της εικόνας για τον περιορισμό των παραμέτρων του μοντέλου, και άρα της υπολογιστικής του πολυπλοκότητας.
Παρακάτω δίνεται ένα διάγραμμα που αφορά την προ-επεξεργασία εικόνας.



Όπως φαίνεται, οι εικόνες της βάσης δεδομένων μας, από 512*384 συρρικνώθηκαν σε 250*250 ή ακόμα και σε 150*150.

Παραπάνω φαίνεται ένα παράδειγμα παρεμβολής εικόνας με:

- Παρεμβολή Κοντινότερου Γείτονα[9]
- Δικυβική Παρεμβολή[9]

Παρεμβολή Κοντινότερου Γείτονα

Πρόκειται για τον απλούστερο τύπο παρεμβολής. Κάθε παρεμβαλόμενο σημείο εξόδου παίρνει τιμή από τον κοντινότερο του γείτονα στην εικόνα εισόδου. Ο πυρήνας συνέλιξης στην παρεμβολή κοντινότερου γείτονα είναι:

$$h(x) = \begin{cases} 0 & |x| > 0 \\ 1 & |x| < 0 \end{cases}$$

Η απόκριση συχνότητας του πυρήνα είναι :

$$H(\omega) = \text{sinc}\,(\omega/2)$$

Δικυβική Παρεμβολή

Η δικυβική παρεμβολή χρησιμοποιεί πολυώνυμα, κυβικά ή κυβικούς αλγορίθμους συνέλιξης. Η κυβική συνελικτική παρεμβολή καθορίζει την τιμή του γκρίζου επιπέδου από το σταθμισμένο μέσο όρο των 16 πλησιέστερων εικονοστοιχείων στις καθορισμένες συντεταγμένες εισόδου και εκχωρεί την τιμή αυτή στις συντεταγμένες εξόδου. Για την Bicubic Interpolation, ο αριθμός των πλεγματικών σημείων που απαιτούνται για την

αξιολόγηση της συνάρτησης παρεμβολής είναι 16, δύο σημεία πλέγματος σε κάθε πλευρά του σημείου κάτω τόσο για την οριζόντια όσο και για την κάθετη κατεύθυνση. Ο πυρήνας παρεμβολής δικυβικής συνέλιξης είναι:

$$W(x)=\begin{cases}(a+2)|x|^3-(a+3)|x|^2+1 & for\ |x|\le 1\\ a|x|^3-5a|x|^2+8a|x|-4a & for\ 1<|x|<2\\ 0 & otherwise\end{cases}$$

όπου η παράμετρος α είναι μεταξύ -0.5 και -0.75

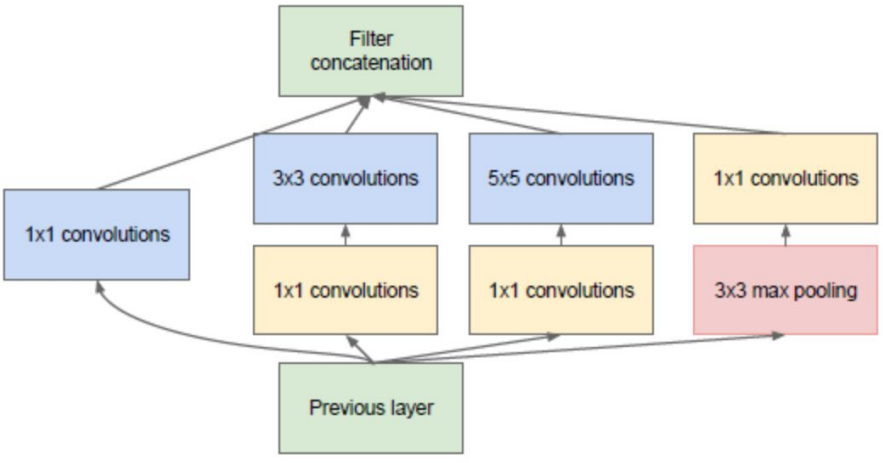
Αρχιτεκτονική CNN

Χρησιμοποιήσαμε προ-εκπαιδευμένα νευρωνικά γιατί είναι βελτιστοποιημένα και ικανά να παράγουν πολύ ισχυρά συνελικτικά χαρακτηριστικά. Συγκεκριμένα χρησιμοποιήσαμε το Inception Version 3 του GoogLeNet καθώς αυτό έδωσε τα καλύτερα αποτελέσματα.
Η αρχιτεκτονική του φαίνεται παρακάτω:

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Details about Parameters of Each Layer in GoogLeNet Network (From Top to Bottom)

Περιλαμβάνει 22 στιβάδες, εκ’ των οποίων εκείνη που κάνει τη διαφορά συγκριτικά με τις αρχιτεκτονικές άλλων προ-εκπαιδευμένων μοντέλων είναι η εξής:



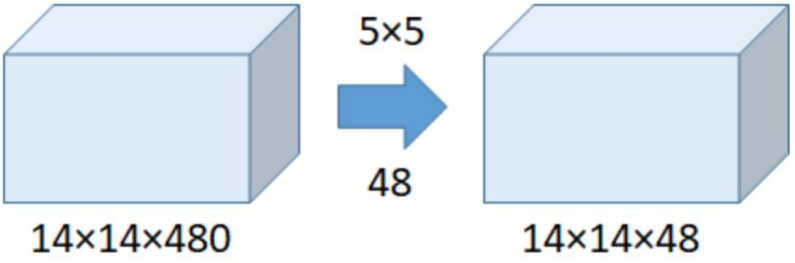
(b) Inception module with dimensionality reduction

Όπως μπορούμε να δούμε, στην είσοδο εφαρμόζονται παράλληλα

- 1X1 Συνελίξεις
- 3X3 Συνελίξεις

- 5X5 Συνελίξεις
- 3X3 Max-pooling

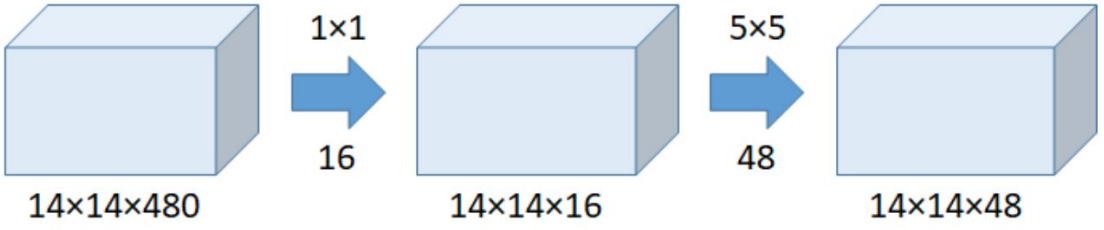
όλες σε συνδυασμό με 1X1 συνελίξεις. Αυτό συμβαίνει γιατί οι 1X1 συνελίξεις μειώνουν την υπολογιστική πολυπλοκότητα, πετυχαίνοντας παράλληλα μη-γραμμική μείωση της διάστασης, που με τη σειρά της οδηγεί σε αποφυγή του overfitting.



Without the Use of 1×1 Convolution

$$\text{Number of operations} = (14\times14\times48)\times(5\times5\times480) = 112.9\text{M}$$

With the use of 1×1 convolution:



With the Use of 1×1 Convolution

$$\text{Number of operations for } 1\times1 = (14\times14\times16)\times(1\times1\times480) = 1.5\text{M}$$

$$\text{Number of operations for } 5\times5 = (14\times14\times48)\times(5\times5\times16) = 3.8\text{M}$$

$$\text{Total number of operations} = 1.5\text{M} + 3.8\text{M} = 5.3\text{M}$$ which is much much smaller than 112.9M !!!!!!!!!!!!!!!

Όπως παρατηρούμε, η εφαρμογή 1X1 συνέλιξης πριν από 5X5 ή οποιαδήποτε άλλη συνέλιξη, βελτιώνει σημαντικά την υπολογιστική πολυπλοκότητα του μοντέλου μας.

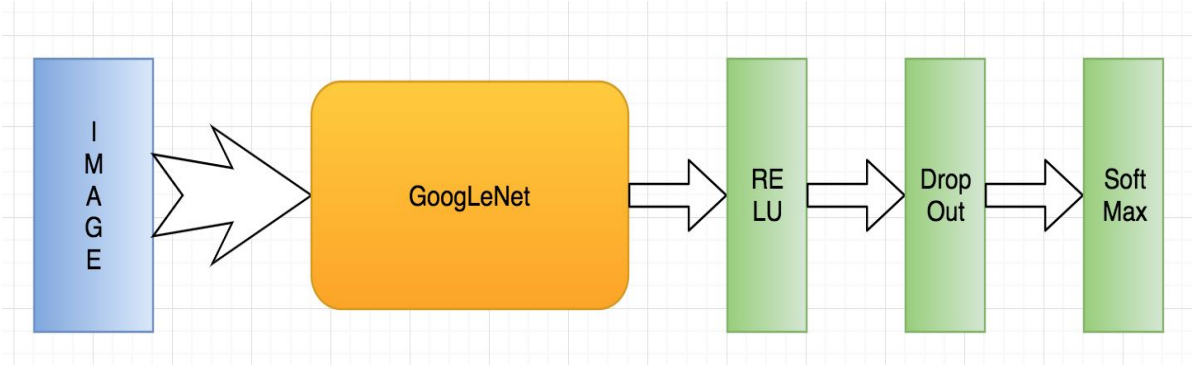
Ύστερα, οι συνελικτικοί χάρτες που προκύπτουν από τις παράλληλες συνελίξεις συνενώνονται, δίνοντας μας 4 διαφορετικά είδη χαρακτηριστικών για την είσοδο.

Στο τέλος εμείς προσθέσαμε 3 ακόμα στιβάδες:

1. ReLu (1 X 1 X 50)
2. Dropout 20%
3. Softmax (1 X 1 X *Num_of_Classes*)

Προσθέσαμε την στιβάδα ReLu για την ομαλότερη μετάβαση (από 1000 νευρώνες σε 50) στην στιβάδα εξόδου.

Παρακάτω φαίνεται η τελική αρχιτεκτονική του δικτύου που υλοποιήσαμε:



Κεφάλαιο 6

Διεξαγωγή Πειραμάτων

6.1 Μετρικές

Στα πειράματα μας εξάγουμε τον *πίνακα σύγχισης*(confusion matrix). Ο Confusion Matrix μας δίνει πληροφορία για τις επιδόσεις της κάθε κλάσσης όχι μόνο σε σχέση με την ίδια, αλλά και σε σχέση με τις υπόλοιπες κλάσσεις των δεδομένων που χρησιμοποίηθηκαν για την αξιολόγηση.
Από τον confusion matrix, μπορούμε λοιπόν να εξάγουμε τις εξής τρεις μετρικές:

- Accuracy
- Precision
- Recall (or Sensitivity)

Accuracy

Η accuracy στα προβλήματα ταξινόμησης είναι το ποσοστό των σωστών προβλέψεων που έκανε το μοντέλο μας πάνω σε όλα τα ειδη των προβλεψεων και ισουται με:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Στον αριθμητή έχουμε τις σωστές προβλέψεις (True Positives & True Negatives) και στον παρονομαστή έχουμε όλες τις προβλέψεις (True Positives & False Positives & False Negatives & True Negatives).
Η accuracy είναι μετρική που χρησιμοποιούμε για την αξιολόγηση των πειραμάτων μας καθώς οι κλάσσεις είναι ισοδύναμες.

Precision

Η precision είναι το ποσοστό των προβλέψεων που ήταν σωστές και ισούται με:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity)

Η recall (or sensitivity) είναι μια μετρική που δίνει το κλάσμα των σωστών προβλέψεων δια των λαθών άλλης κλάσσης και ισούται με:

$$\text{Recall} = \frac{TP}{TP + FN}$$

6.2 Πειραματικές Ρυθμίσεις

Χρησιμοποιήσαμε μηχάνημα με 2,6 GHz Intel Core i5 6 CPU's και 8 GB 1600 MHz DDR3
μνήμη. Η GPU που χρησιμοποιήσαμε ήταν η Intel Iris 1536 MB και ο τύπος του λειτουργικού
συστήματος ήταν MacOS High Sierra. Για τα πειράματά μας χρησιμοποιήσαμε τα εξής frameworks:

- ➔ Tensorflow/Keras
- ➔ Sklearn

Κατα προσέγγιση, κάθε επανάληψη στα πειράματά μας στο μηχάνημα αυτό διαρκούσε μία ώρα.

Στα πειράματά μας χρησιμοποιήσαμε 3 μόνο κλάσσεις απο τις 101. Χωρίσαμε το σύνολο δεδομένων ως εξής:

1. Σύνολο δεδομένων προς εκπαίδευση(train set) (65%)
2. Σύνολο δεδομένων προς εξακρίβωση(validation set)(10%)
3. Σύνολο δεδομένων προς αξιολόγηση(test set)(25%)

6.3 Αποτελέσματα πειραμάτων

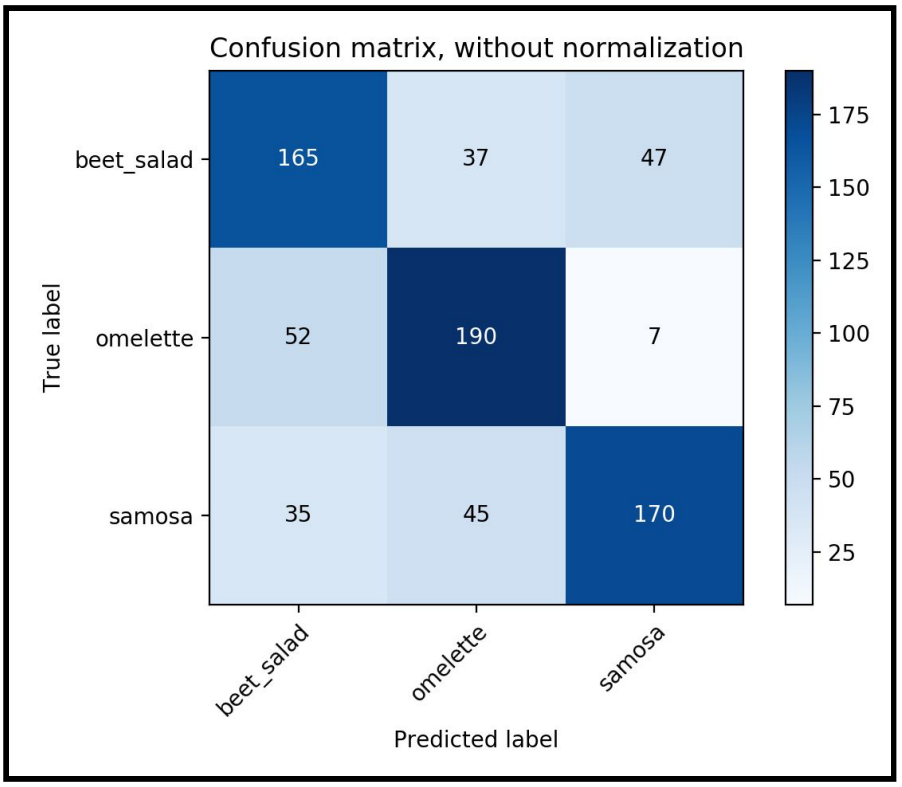
Παρακάτω δίνουμε πίνακες ρυθμίσεων πειράματος για την κάθε μια από τις μεθόδους που χρησιμοποιήσαμε με τα αποτελέσματα που λάβαμε καθώς και πίνακες σύγχισης.

1. Μέθοδος Μετασχηματισμού SIFT

Παρακάτω φαίνεται ο πίνακας ρυθμίσεων ενδεικτικά για κάποια πειράματα.

Exp/Conf	#Sift Features/Image	#Visual Words	Accuracy
exp01	250	700	68%
exp02	700	2000	64%
exp03	**300**	**1000**	**70%**

Για το πείραμα exp03, φαίνεται παρακάτω ο πίνακας σύγχισης καθώς και ο πίνακας accuracy-precision-recall:

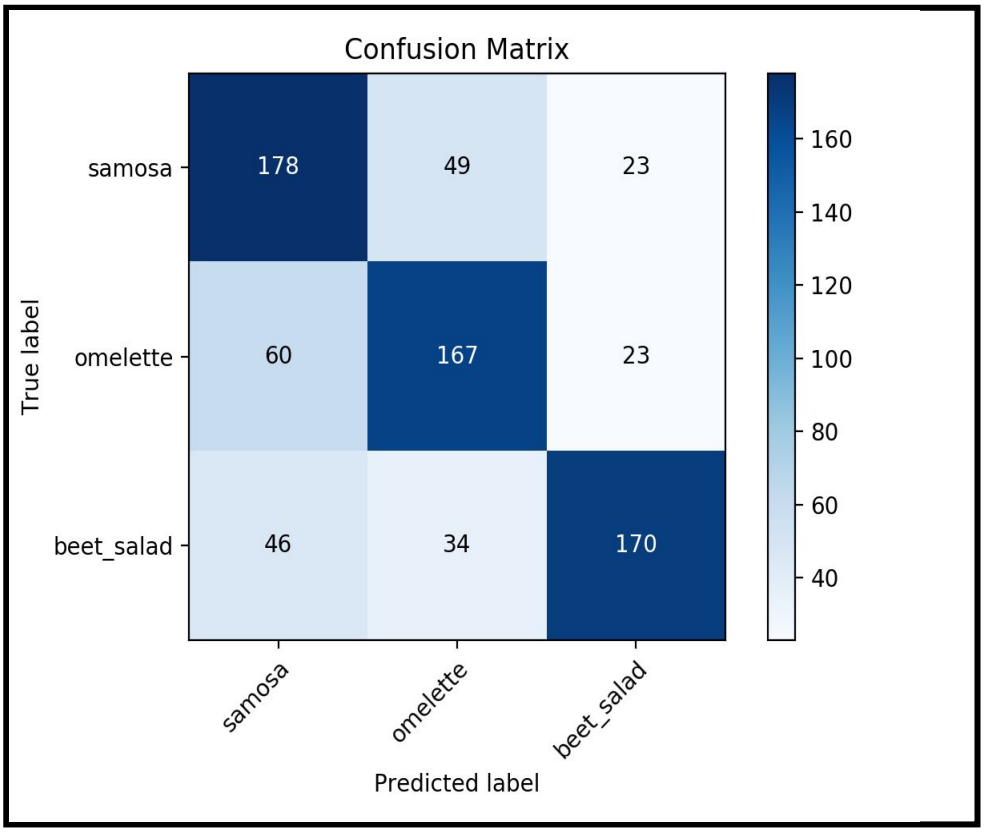


Class / **Metric**		**Precision**	**Recall**
beet_salad		0.66	0.65
omelette		0.76	0.70
samosa		0.68	0.76

2. Μέθοδος Μετασχηματισμού SURF

Exp/Conf	Hessian Threshold	#Visual Words	Accuracy
exp01	500	1000	65%
exp02	**300**	**1300**	**69%**
exp03	100	1300	66%

Για το πείραμα exp02, φαίνεται παρακάτω ο πίνακας σύγχισης καθώς και ο πίνακας precision-recall:



Class/**Metric**		**Precision**	**Recall**
beet_salad		0.71	0.63
omelette		0.60	0.61
samosa		0.68	0.79

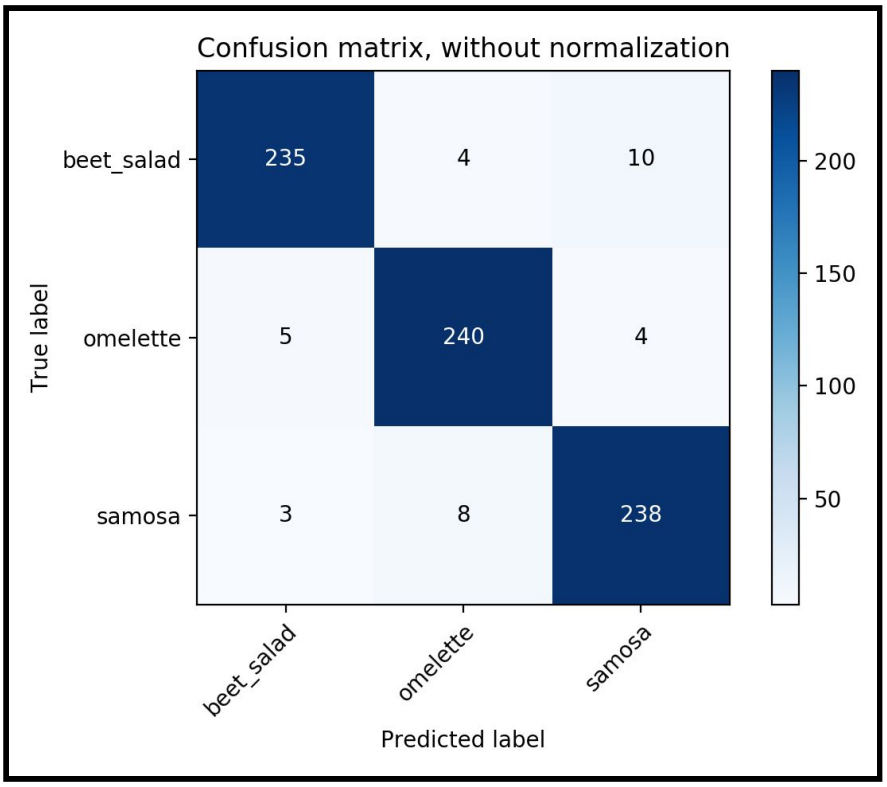
3. Μέθοδος Συνελικτικών Δικτύων

Παρακάτω φαίνεται ο πίνακας πειραμάτων/ρυθμίσεων για την αρχιτεκτονική που χρησιμοποιήσαμε βασισμένη στα συνελικτικά νευρωνικά δίκτυα:

Exp\Conf	Image Preprocess	Epochs	Validation Accuracy	Test Accuracy	Optimizer	Notes
exp01	-->(200,200) (nearest)	7	90%	88%	SGD	
exp02	-->(250,250) (nearest)	4	88%	72%	Adam	
exp03	-->(200,200) (bicubic)	7	86%	73%	SGD	
exp04	-->(150,150) (nearest)	7	88%	89%	RMSprop	Big Batch Size
exp05	-->(150,150) (nearest)	7	91%	91%	RMSProp	Small Batch Size
exp06	**-->(250,250) (nearest)**	**3**	**96%**	**95%**	**RMSProp**	**Small Batch Size**
exp07	-->(150,150) (nearest)	6	87%	88%	RMSprop	Less parameters

Όπως παρατηρούμε, ο RMSprop optimizer με παρεμβολή σε εικόνες 250x250 έδωσε ακρίβεια στο Test Set 95%. Μια ακόμα σημαντική παρατήρηση είναι ότι ο RMSprop optimizer βοήθησε στη γενίκευση του μοντέλου μας καθώς παρατηρούμε μικρές τετριμμένες διαφορές ανάμεσα στο Validation και στο Test accuracy.

Παρακάτω δίνεται ο πίνακας σύγχισης και ο πίνακας που παράχθηκε απο το πείραμα exp06 που έδωσε τα βέλτιστα αποτελέσματα:



Class / **Metric**		**Precision**	**Recall**
beet_salad		0.95	0.97
omelette		0.96	0.95
samosa		0.95	0.94

Κεφάλαιο 7

Μελλοντικές Επεκτάσεις

7.1 Συμπεράσματα

Οι αρχιτεκτονικές συνελικτικών νευρωνικών δικτύων είναι τόσο ισχυρές για την αναπαράσταση μιας εικόνας σαν ένα σύνολο χαρακτηριστικών που οι μη-βασιζόμενες σε αυτά μέθοδοι υστερούν πολύ σε ακρίβεια. Αυτό συμβαίνει γιατί τα CNNs έχουν τη δυνατότητα να χρησιμοποιήσουν πολλών ειδών φίλτρα και να παράγουν χάρτες χαρακτηριστικών(feature maps) οι οποίοι αναπαριστούν πολλά και πολύ ισχυρά χαρακτηριστικά της εικόνας.

7.2 Μελλοντικές Επεκτάσεις

Υπάρχουν αρκετές μελλοντικές προτάσεις για βελτίωση του συστήματος μας. Αρχικά, θα μπορούσαν να δοκιμαστούν και οι άλλες προ-εκπαιδευμένες αρχιτεκτονικές όπως είναι η AlexNet, η ZF Net, η ResNet και η VGG Net. Επιπλέον, μπορεί στη θέση της πλήρως συνελικτικής στοιβάδας στο τελικό επίπεδο του συνελικτικού μας δικτύου να δοκιμαστεί κάποιος ταξινομητής,

όπως SVM. Τέλος, σκοπός μας μετά το τέλος της παρούσας εργασίας είναι να εκπαιδεύσουμε μοντέλα σε ολο FOOD-101 σύνολο δεδομένων και να δούμε τις επιδόσεις σε 101 κατηγορίες φαγητών.

Βιβλιογραφία

[1] Chris Harris & Mike Stephens A COMBINED CORNER AND EDGE DETECTOR, Plessey Research Roke Manor, United Kingdom © The Plessey Company pic. 19

[2]Theodoros Evgeniou and Massimiliano Pontil , SUPPORT VECTOR MACHINES:THEORY AND APPLICATIONS ,Center for Biological and Computational Learning, and Artificial Intelligence Laboratory, MIT, E25-201, Cambridge, MA 02139, USA

[3]Sinno Jialin Pan and Qiang Yang A Survey on Transfer Learning Fellow, IEEE

[4] David G. Lowe Distinctive, Image Features from Scale-Invariant Keypoints, Computer Science Department University of British Columbia Vancouver, B.C., Canada lowe@cs.ubc.ca January 5, 2004

[5] Herbert Bay Tinne Tuytelaars, and Luc Van Gool, SURF: Speeded Up Robust Features ,ETH Zurich {bay, vangool}@vision.ee.ethz.ch 2 Katholieke Universiteit Leuven {Tinne.Tuytelaars, Luc.Vangool}@esat.kuleuven.be

[6] Brosnan, T., and Sun, D.-W. Improving quality inspection of food products by computer vision—-a review. Journal of food engineering 61, 1 (2004), 3–16.

[7] Brynjolfsson, E., and McAfee, A. The big data boom is the innovation story of our time. The Atlantic 21 (2011).

[8]David Aldavert, Marcal Rusinol, Ricardo Toledo, Josep Llad´os, A Study of Bag-of-Visual-Words Representations for

[9] Dianyuan Han, Comparison of Commonly Used Image Interpolation Methods, Dept. of Computer Engineering Wei Fang University Shandong 261061, China wfhdy@163.com

[10] Cisco. Cisco visual networking index: Forecast and methodology, 2016
to 2021 (white paper), 2017.

[11] Piji Li, Optimization Algorithms for Deep Learning, Optimization Algorithms for Deep Learning Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong pjli@se.cuhk.edu.hk

[12] Khoso, M. How much data is produced every day? http://www.northeastern.edu/levelblog/2016/05/13/ how-much-data-produced-every-day, 2016. [Online; accessed 01.06.2018].

[13] Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., and Soares, J. V. Leafsnap: A computer vision system for automatic plant species identification. In Computer vision–ECCV 2012. Springer, 2012, pp. 502–516.

[14] UN. Tenfold increase in childhood and adolescent obesity in four decades: new study by imperial college london and who.
http://www.who.int/en/news-room/detail/11-10-2017-tenfold-increase-in-childhood-and-adolescent-obesity-in-four-decades-n 2017. [Online; accessed 01.06.2018].

[15] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural

networks. In Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[16] Zheng, L., Yang, Y., and Tian, Q. Sift meets cnn: A decade survey of instance retrieval. IEEE transactions on pattern analysis and machine intelligence (2017).

[17] Zeiler, M. D., and Fergus, R. Visualizing and understanding convolutional networks. In European conference on computer vision (2014), Springer, pp. 818–833.

[18] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[19] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al. Going deeper with convolutions. Cvpr.

[20] He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (2015), pp. 1026–1034.

[21] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 3 (2015), 211–252.

[22] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic

optimization. Journal of Machine Learning Research, 12(Jul):2121–2159, 2011.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[24] Lisa Torrey and Jude Shavlik. «Transfer learning». In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global, 2010, pp. 242–264.

[25] Matthew D Zeiler. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.

[26] Geoffrey Hinton, N Srivastava, and Kevin Swersky. Lecture 6a overview of mini–batch gradient descent. Coursera Lecture slides https://class. coursera. org/neuralnets-2012-001/lecture,[Online, 2012.

[27] Piji Li, Optimization Algorithms for Deep Learning, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong pjli@se.cuhk.edu.hk

[28] Saad ALBAWI , Tareq Abed MOHAMMED, Understanding of a Convolutional Neural Network , Department of Computer Engineering Faculty of Engineering and Architecture Istanbul Kemerburgaz University Istanbul, Turkey Saad AL-ZAWI Department of Electronic Engineering Faculty of Engineering Diyala University Diyala , Iraq , ICET2017, Antalya, Turkey 978-1-5386-1949-0/17/$31.00 ©2017 IEEE

[29] I. Kokkinos, E. C. Paris, and G. Group, "Introduction to Deep Learning Convolutional Networks, Dropout, Maxout 1,"

pp. 1–70.

[30] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097- 1105).

[31] D. Stutz and L. Beyer, “Understanding Convolutional Neural Networks,” 2014.