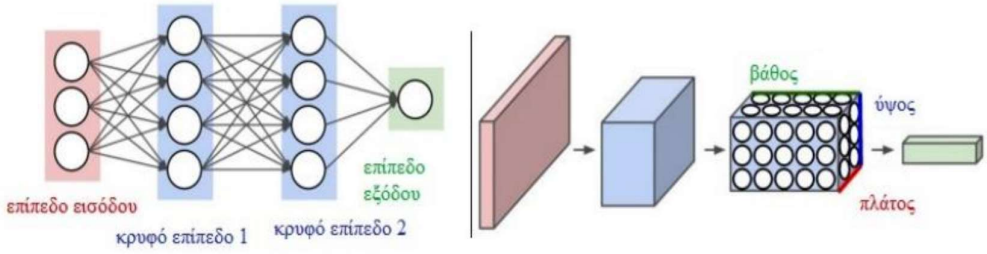
Συνελικτικά Νευρωνικά Δίκτυα

Τα Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ) (CNN:Convolutional Neural Networks) έχουν χρησιμοποιηθεί ευρέως στην αναγνώριση εικόνων και μοιάζουν με τα Τεχνητά Νευρωνικά Δίκτυα πρόσθιας τροφοδότησης (feed-forward networks, FFN). Υπάρχουν διαφορές ανάμεσα σε αυτά τα δύο είδη δικτύων, που καθιστούν τα ΣΝΔ πιο ελκυστικά για χρήση. Πιο συγκεκριμένα στα κλασσικά FFN η κλιμάκωση μεγάλων εικόνων οδηγεί σε μεγάλο μη διαχειρίσιμο πλήθος βαρών. Για παράδειγμα στο σύνολο δεδομένων CIFAR-10 οι εικόνες είναι μεγέθους, μόνο 32 x 32 x 3 (πλάτος × ύψος × χρωματικά κανάλια), και επομένως ένας πλήρως συνδεδεμένος νευρώνας στο πρώτο κρυφό επίπεδο ενός ΤΝΔ θα είχε 32*32*3 = 3072 βάρη. Παρά το γεγονός ότι αυτός ο αριθμός δείχνει διαχειρίσιμος, για είσοδο εικόνας μεγαλύτερων διαστάσεων, π.χ 200 x 200 x 3 θα είχαμε νευρώνες με 200*200*3 = 120.000 βάρη ο καθένας.

Για περισσότερους νευρώνες ο αριθμός των παραμέτρων θα μεγάλωνε ραγδαία. Συνεπώς η πλήρης συνδεσιμότητα είναι σπάταλη, και μπορεί λόγω των πολλών παραμέτρων να οδηγήσει εύκολα σε *υπερπροσαρμογή* (*overfitting*) του δικτύου. Από την άλλη πλευρά, η χρήση των ΣΝΔ εκμεταλλεύεται το γεγονός ότι η είσοδος αποτελείται από εικόνες, και περιορίζει την αρχιτεκτονική του με πιο έξυπνο τρόπο. Γενικότερα τα ΣΝΔ, διαθέτουν νευρώνες οι οποίοι έχουν 3 διαστάσεις (πλάτος-ύψος-βάθος), και έχουν την ιδιαιτερότητα να συνδέονται διαδοχικά με μία μικρή περιοχή του προηγούμενου επιπέδου αντίθετα με ότι γινόταν σε μία πλήρη σύνδεση. Κατ' ουσίαν ο νευρώνας είναι ένα φίλτρο που *συνελίσεται* στα δεδομένα εισόδου του. Θα πρέπει να διευκρινίσουμε εδώ ότι λέγοντας «συνελίσεται» εννοούμε την εφαρμογή της δισδιάστατης συνέλιξης με κάποια διευρυμένη χρήση. Οι όροι νευρώνας, κόμβος (node), φίλτρο και πυρήνας (kernel)χρησιμοποιούνται εναλλακτικά στην βιβλιογραφία των συνελικτικών δικτύων για το ίδιο πράγμα.



Σχήμα _. (Αρχιτεκτονική CNN)

Το παραπάνω σχήμα απεικονίζει στα αριστερά μία αρχιτεκτονική ενός κλασικού νευρωνικού δικτύου και στα δεξιά μία αρχιτεκτονική ενός συνελικτικού νευρωνικού δικτύου. Κάθε επίπεδο ενός ΣΝΔ μετασχηματίζει τον τρισδιάστατο όγκο εισόδου (input volume), σε έναν τρισδιάστατο όγκο εξόδου (output volume. Στο συγκεκριμένο παράδειγμα η έγχρωμη εικόνα του παραπάνω σχήματος αποτελεί το επίπεδο εισόδου επομένως θα έχει βάθος 3, όσα και τα κανάλια μιας RGB εικόνας.

Στα ΣΝΔ εκτελούνται τέσσερις βασικές λειτουργίες:

1. Συνέλιξη (Convolution)
2. Εφαρμογή μη γραμμικότητας (Non Linearity) – ReLU
3. Συγκέντρωση ή Υπο-Δειγματοληψία (Pooling / Sub sampling)
4. Κατηγοριοποίηση με πλήρως συνδεδεμένο επίπεδο (Fully Connected Layer for Classification)

Επίπεδο Συνέλιξης (Convolution)

Θεμελιώδες δομικό επίπεδο ενός Συνελικτικού Νευρωνικού Δικτύου, όπως έχει προαναφερθεί, είναι το επίπεδο στο οποίο συνελίσεται η είσοδος με Ν το πλήθος φίλτρα (νευρώνες βαρών). Το αποτέλεσμα είναι πίνακες που αποτελούν χάρτες χαρακτηριστικών. Το επίπεδο αυτό ονομάζεται συνελικτικό επίπεδο. Τα φίλτρα αυτά στην περίπτωση των έγχρωμων εικόνων είναι τρισδιάστατα, καθώς το μέγεθός τους καθορίζεται από το ύψος *R* (πλήθος γραμμών), το πλάτος (πλήθος στηλών) *C* και το βάθος *D* (πλήθος πινάκων R×C). Στην περίπτωση που η είσοδος είναι μία έγχρωμη εικόνα το βάθος είναι τρία (3) για τα τρία βασικά χρώματα (Red Green Blue: RGB). Για την παραγωγή αυτών των χαρακτηριστικών, σαρώνεται ολόκληρη η εικόνα, πραγματοποιούνται πράξεις εσωτερικού

γινομένου μεταξύ των τιμών του φίλτρου και της υποκείμενης περιοχής του πίνακα και στο τέλος εξάγεται το αποτέλεσμα το οποίο τοποθετείται στον χάρτη χαρακτηριστικών.

Η πράξη της *δισδιάστατης συνέλιξης* για δύο διακριτά σήματα x(r,c), w(r,c),(δισδιάστατες ακολουθίες)ορίζεται τυπικά από την σχέση:

$$x * w = \{x * w\}(r,c) = \sum_k \sum_l x(k,l) \cdot w(r-k, c-l) \text{ με } r,c,k,l \in \mathbb{Z}$$

Η πράξη της *δισδιάστατης συσχέτισης* για δύο διακριτά σήματα x(m,n), y(m,n),(δισδιάστατες ακολουθίες)ορίζεται τυπικά από την σχέση:

$$x \circ w = \{x \circ w\}(r,c) = \sum_k \sum_l x(k,l) \cdot w(k-r, l-c) \text{ με } r,c,k,l \in \mathbb{Z}$$

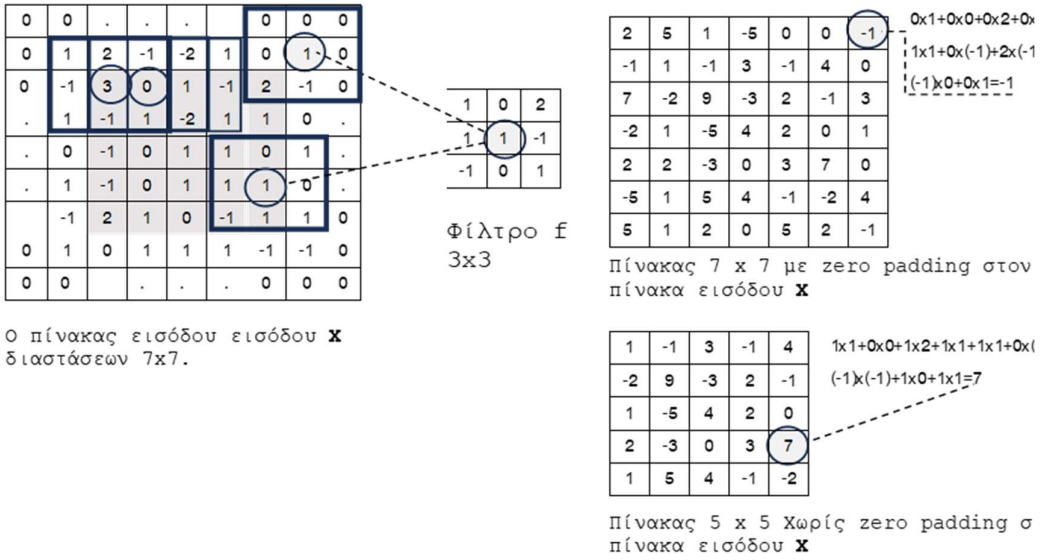
Αν w(k,l)=w(-k,-l) οι δύο πράξεις δίνουν το ίδιο αποτέλεσμα. Για ακολουθίες πεπερασμένου μήκους τα όρια των δεικτών k,l καθορίζονται από τις θέσεις που οι ακολουθίες έχουν όλες τις τιμές τους μηδενικές.

Στα συνελικτικά δίκτυα εφαρμόζεται η πράξη της συσχέτισης. Εν τούτοις ιστορικά επικράτησε να χρησιμοποιείται ο όρος συνέλιξη από την ευρεία χρήση του στην επεξεργασία σημάτων. Θα διατηρήσουμε αυτήν την ορολογία στην συνέχεια.

Ανεξαρτήτωως των ορίων μέσα στα οποία ορίζεται η πράξη της συνέλιξης στο $\mathbb{Z} \times \mathbb{Z}$, εκείνο που μας ενδιαφέρει είναι οι τιμές των ακολουθιών και τις θεωρούμε ως στοιχεία πινάκων α) της εισόδου **X** με διαστάσεις (R_x x C_x) και β) του φίλτρου **W** με διαστάσεις (R_w x C_w).

Στο ακόλουθο παράδειγμα φαίνεται η εφαρμογή της πράξης της συνέλιξης (συσχέτισης). Οι τιμές του του φίλτρου $w(.,.)$ ολισθαίνουν επάνω στις τιμές των γραμμών και των στηλών της $x(.,.)$ και υπολογίζεται το άθροισμα των γινομένων των ομοιόθετων τιμών. Το αποτέλεσμα θα είναι μια ακολουθία πεπερασμένου μήκους με διαστάσεις (R_x+ R_w-1)x(C_x+ C_w-1). Αν θέλουμε οι διαστάσεις της προκύπτουσας ακολουθίας να είναι ίσες με αυτές τις εισόδου, αφαιρούμε R_w-1 γραμμές και C_w-1 στήλες από τα όριά της. Απλούστερα μπορούμε να ολισθαίνουμε το φίλτρο έτσι ώστε η

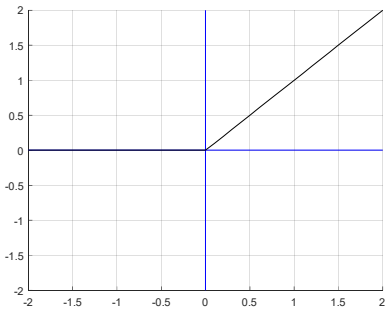
κεντρική τιμή του φίλτρου *w* να φτάνει στα όρια της *x* και αποδίδοντας μηδενικές τιμές στη *x* εκτός των ορίων της (zero padding).



Σχήμα _: Υπολογισμοί στο συνελικτικό επίπεδο

Είναι βολικό το φίλτρο να έχει περιττό αριθμό γραμμών και στηλών π.χ. 3×3, 5×5 ώστε να υπάρχει κεντρική τιμή. Στο τέλος της πράξης της συνέλιξης σε κάθε τιμή που θα προκύψει προσθέτουμε και ένα σταθερό όρο w_0 (bias) η σημασία του οποίου είναι ίδια με αυτήν που αναφέρθηκε στον γραμμικό ταξινομητή.

Η τιμή κάθε στοιχείου του παραγόμενου γίνεται όρισμα μιας συνάρτησης ενεργοποίησης κυρίως της ReLU.



Η συνάρτηση ReLU (Rectified Linear Unit)

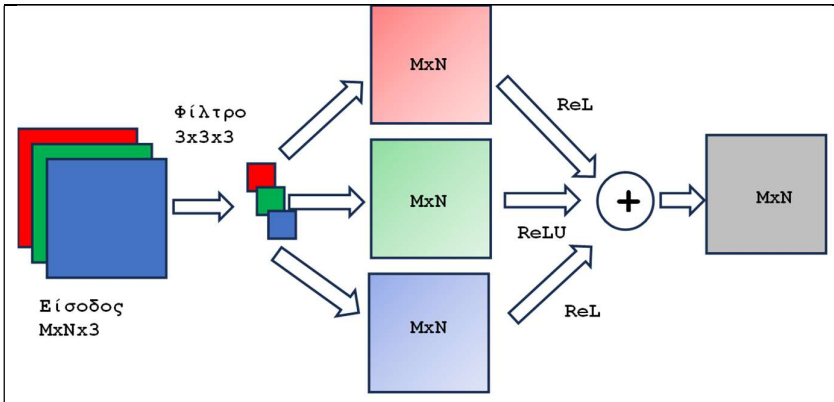
f(x)=max(0,x)

Αν η είσοδος έχει βάθος D δηλαδή αποτελείται από D το πλήθος πίνακες, ονομάζονται και κανάλια (channels), (π.χ. D=3, για είσοδο μία RGB έγχρωμη εικόνα), τότε και κάθε φίλτρο πρέπει να έχει βάθος D, δηλαδή να αποτελείται από D πίνακες. Για παράδειγμα αν η είσοδος είναι R_xxC_x×3 το φίλτρο πρέπει να είναι R_wxC_w×3. Στην περίπτωση αυτή μετά την εφαρμογή του φίλτρου θα προκύψουν και D πίνακες ώς αποτελέσματα. Οι πίνακες αυτοί είναι ίσων διαστάσεων και προστίθενται. Τελικά μετά την άθροιση προκύπτει ένας πίνακας για κάθε φίλτρο. Αν εφαρμοσθούν N φίλτρα θα προκύψουν N πίνακες, όσοι και τα φίλτρα. Στην ορολογία των ΣΝΔ συχνά ένας τρισδιάστατος πίνακας καλείται και *όγκος* (*volume*). Άρα η είσοδος μας είναι ένας όγκος R_xxC_x×3 και με την εφαρμογή N φίλτρων παράγεται ως αποτέλεσμα ένας όγκος R_xxC_x×N.

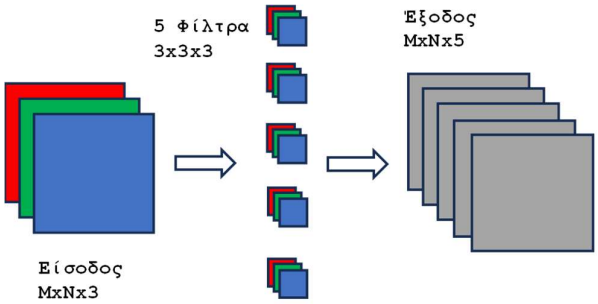
Τα παραπάνω απεικονίζονται στα ακόλουθα σχήματα (Σχ._, Σχ._)

Το φίλτρο κατά την ολίσθηση του είναι αποδεκτό να κινείται και με βήμα (δρασκελιά, stride) μεγαλύτερο του ένα. Αυτό είναι μια παραλλαγή συνέλιξης που οδηγεί σε αποτελέσματα μικρότερων διαστάσεων και ακρίβειας. Έχει χρήση στην περίπτωση μεγάλων εικόνων και φίλτρων (π.χ για φίλτρο 15×15, stride=2).

Τα παραπάνω, δηλαδή η συνέλιξη της εισόδου με τα φίλτρα και η εφαρμογή της μη γραμμικότητας αποτελούν ένα *Επίπεδο Συνέλιξης* (*Convolutional Layer*).

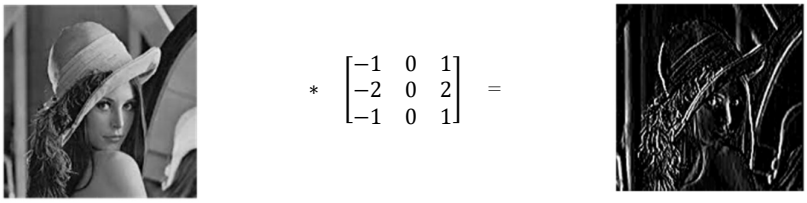


Σχ._ Η εφαρμογή ενός φίλτρου σε είσοδο τριών καναλιών στο συνελικτικό επίπεδο



Σχ΄_ Συνελικτικό επίπεδο με πέντε (5) φίλτρα

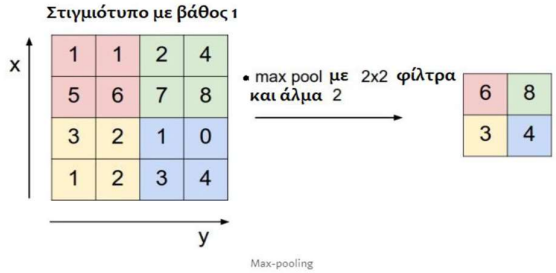
Οι πίνακες που εξάγονται από το επίπεδο συνέλιξης καλούνται χάρτες χαρακτηριστικών. Στο ακόλουθο Σχήμα_ φαίνεται η εφαρμογή του ενός φίλτρου *Sobel* στην εικόνα «lena».



Σχ._

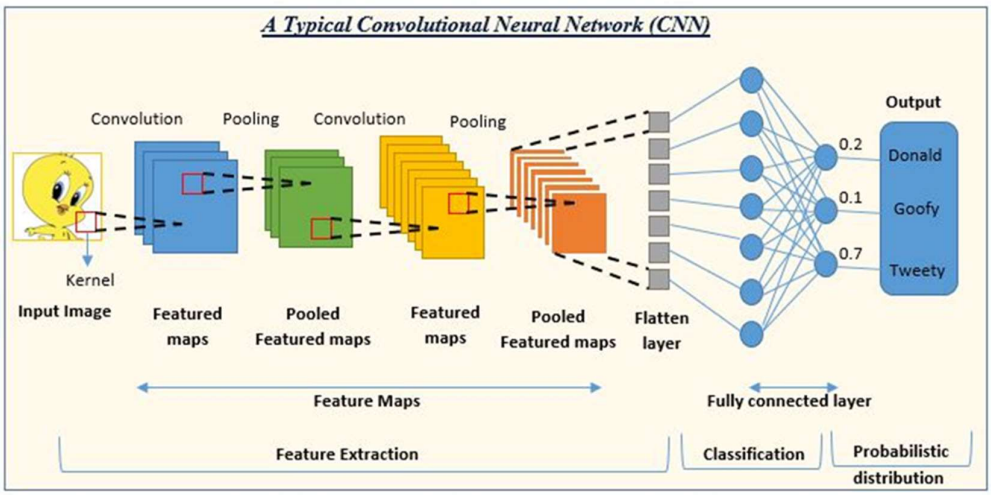
Δειγματοληψία (Pooling)

Τα επίπεδα δειγματοληψίας συναντώνται ανάμεσα στα επίπεδα συνέλιξης των ΣΝΔ. Αυτό που κάνει αυτό το επίπεδο είναι να μειώνει το μέγεθος των πινάκων εισόδων και κατά συνέπεια και τους υπολογισμούς ενός δικτύου και θέτοντας υπό έλεγχο προβλήματα υπερπροσαρμογής (over-fitting). Λαμβάνοντας τις τιμές μιας μικρής περιοχής π.χ. (M×M) ενός πίνακα εισόδου και με βήμα M παίρνουμε την μεγαλύτερη από αυτές (Max Pooling) ή τον μέσο όρο τους (Average Pooling) και δημιουργούμε έναν νέο πίνακα με μικρότερες διαστάσεις. Αν για παράδειγμα ο πίνακας είναι 224×224 για M=2 θα έχουμε ως αποτέλεσμα ένα πίνακα 112×112. Ένα παράδειγμα φαίνεται στο παρακάτω σχήμα:



Σχήμα _. (Επίπεδο Υποδειγματοληψίας)

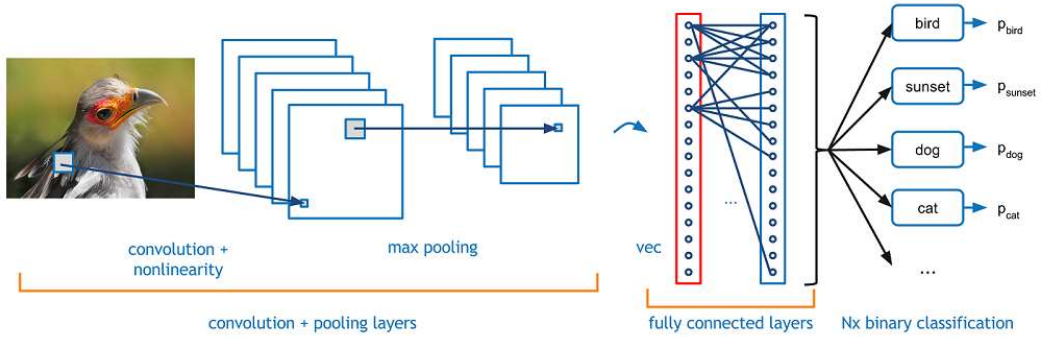
Επίπεδα συνέλιξης και υποδειγματοληψίας μπορούν να τοποθετούνται διαδοχικά καταλήγοντας σε πίνακες μικρότερων διαστάσεων κάθε φορά. Μετά από αυτήν την διαδοχή οι τιμές των τελικών πινάκων τοποθετούνται σε ένα άνυσμα στήλης (*flattening*). Το άνυσμα αυτό θα είναι η είσοδος ενός τελικού επιπέδου ταξινόμησης με πλήρη διασύνδεση.



Σχ._

Το πλήρως συνδεδεμένο Επίπεδο Ταξινόμησης

Σε αυτό το επίπεδο οι νευρώνες έχουν πλήρεις συνδέσεις με όλες τις εξόδους των νευρώνων του προηγούμενου επιπέδου. Το πλήρως συνδεδεμένο επίπεδο (Fully Connected Layer) είναι ένας Perceptron πολλών επιπέδων (MLP), που χρησιμοποιεί ως συνάρτηση κόστους την πολυωνυμική λογιστική (Multinomial Logistic) στο επίπεδο εξόδου . Για την είσοδο των δεδομένων στο επίπεδο αυτό θα πρέπει αυτά να δομήσουν ένα μονοδιάστατο πίνακα (άνυσμα στήλης). Στο ακόλουθο πίνακα παρουσιάζεται μια εκδοχή ενός CNN τεσσάρων κλάσεων.



Σχήμα _.

Υπολογισμός του πλήθους βαρών και εκπαίδευση στα ΣΝΔ

Επίπεδο συνέλιξης : Το συγκεκριμένο επίπεδο είναι αυτό που δημιουργεί τον πίνακα χαρακτηριστικών, οπότε έχουμε πίνακες βαρών. Οι αρχικές τιμές των βαρών είναι τυχαίες. Ο αριθμός των βαρών (παραμέτρων) σε ένα επίπεδο συνέλιξης με N κανάλια και M φίλτρα διαστάσεων r × c θα είναι: r×c×N×M

Επίπεδο δειγματοληψίας : Στο επίπεδο δειγματοληψίας μειώνουμε το μέγεθος των διαστάσεων. Για παράδειγμα αν σε έναν χάρτη χαρακτηριστικών 224×224 στοιχείων εφαρμοστεί δειγματοληψία σε 2×2 υπό-περιοχές τότε αριθμός των στοιχείων του χάρτη χαρακτηριστικών είναι 112×112. Στο επίπεδο δειγματοληψία δεν υπάρχουν βάρη.

Πλήρως συνδεδεμένο επίπεδο (FC) : Το συγκεκριμένο επίπεδο έχει τον υψηλότερο αριθμό βαρών από κάθε άλλο επίπεδο. Για τον υπολογισμό του πλήθους των βαρών υπολογίζεται το γινόμενο του πλήθους των νευρώνων κάθε επιπέδου επί το πλήθος των εξόδων του προηγούμενου επιπέδου. Έτσι το πλήθος των βαρών είναι: πλήθος νευρώνων στο τρέχον επίπεδο × (πλήθος εξόδων στο προηγούμενο επίπεδο+1), το +1 αφορά τον σταθερό όρο. Για παράδειγμα στο ***CNN VGGNET*** το πλήθος των βαρών φαίνεται στο ακόλουθο σχήμα.

```
INPUT: [224x224x3]        memory:  224*224*3=150K   weights: 0
CONV3-64: [224x224x64]  memory:  224*224*64=3.2M   weights: (3*3*3)*64 = 1,728
CONV3-64: [224x224x64]  memory:  224*224*64=3.2M   weights: (3*3*64)*64 = 36,864
POOL2: [112x112x64]  memory:  112*112*64=800K   weights: 0
CONV3-128: [112x112x128]  memory:  112*112*128=1.6M   weights: (3*3*64)*128 = 73,728
CONV3-128: [112x112x128]  memory:  112*112*128=1.6M   weights: (3*3*128)*128 = 147,456
POOL2: [56x56x128]  memory:  56*56*128=400K   weights: 0
CONV3-256: [56x56x256]  memory:  56*56*256=800K   weights: (3*3*128)*256 = 294,912
CONV3-256: [56x56x256]  memory:  56*56*256=800K   weights: (3*3*256)*256 = 589,824
CONV3-256: [56x56x256]  memory:  56*56*256=800K   weights: (3*3*256)*256 = 589,824
POOL2: [28x28x256]  memory:  28*28*256=200K   weights: 0
CONV3-512: [28x28x512]  memory:  28*28*512=400K   weights: (3*3*256)*512 = 1,179,648
CONV3-512: [28x28x512]  memory:  28*28*512=400K   weights: (3*3*512)*512 = 2,359,296
CONV3-512: [28x28x512]  memory:  28*28*512=400K   weights: (3*3*512)*512 = 2,359,296
POOL2: [14x14x512]  memory:  14*14*512=100K   weights: 0
CONV3-512: [14x14x512]  memory:  14*14*512=100K   weights: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512]  memory:  14*14*512=100K   weights: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512]  memory:  14*14*512=100K   weights: (3*3*512)*512 = 2,359,296
POOL2: [7x7x512]  memory:  7*7*512=25K  weights: 0
FC: [1x1x4096]  memory:  4096  weights: 7*7*512*4096 = 102,760,448
FC: [1x1x4096]  memory:  4096  weights: 4096*4096 = 16,777,216
FC: [1x1x1000]  memory:  1000 weights: 4096*1000 = 4,096,000

TOTAL memory: 24M * 4 bytes ~= 93MB / image (only forward! ~*2 for bwd)
TOTAL params: 138M parameters
```

Σχήμα _: Πλήθος των βαρών στο CNN VGGNet 16

Εκπαίδευση των Συνελικτικών Νευρωνικών Δικτύων

Τα ΣΝΔ είναι δίκτυα εμπρόσθιας τροφοδότησης. Εκπαιδεύονται με την ελαχιστοποίηση της συνάρτησης κόστους με την μέθοδο gradient descent και την τεχνική της μείωσης σφάλματος με οπισθοδρόμηση (back error propagation) αυτό περιγράφεται αναλυτικά ακολούθως για κάθε επίπεδο του δικτύου.

Εκπαίδευση στο επίπεδο πλήρους διασύνδεσης

Στο τελικό επίπεδο εξόδου το πλήθος των νευρώνων είναι ίσο με το πλήθος των κλάσεων έστω M. Κάθε κλάση αντιστοιχίζεται σε έναν νευρώνα (τον ίδιο πάντα) και ονοματίζεται με την θέση του m=1...M στο επίπεδο εξόδου. Έστω το $\boldsymbol{x} = [x_1 \dots x_N, 1]$ επαυξημένο διάνυσμα που εισέρχεται στο επίπεδο εξόδου και $\boldsymbol{w_m} = [w_{1m}, \dots w_{nm}, \dots w_{Nm}, w_{0m}]^T$ τα επαυξημένα διανύσματα των βαρών των νευρώνων σ' αυτό.

Αρχικά υπολογίζονται οι ποσότητες $\sigma_m = \boldsymbol{w}_m^T \cdot \boldsymbol{x}$. Οι τιμές αυτές μετασχηματίζονται σύμφωνα την συνάρτηση *softmax* σε $y_m = \frac{e^{w_m^T \cdot x}}{\sum_k e^{w_k^T \cdot x}}$

Οι τιμές y_m είναι θετικές και το άθροισμα τους $\sum_{m=1}^{M} y_m = 1$. Επιθυμούμε οι τιμές αυτές να είναι τέτοιες ώστε αν η είσοδος είναι $\mathbf{x^c}$, $c \in \{1, \dots, M\}$ τότε η έξοδος y_c του c νευρώνα να είναι μεγαλύτερη των άλλων εξόδων και ει δυνατόν ίση με ένα εκφράζοντας έτσι υψηλή πιθανότητα να ανήκει η είσοδος στην επιθυμητή κλάση. Με χρήση των πιθανοτήτων αν $\{1, \dots, M\}$ είναι οι τιμές μιας τυχαίας διακριτής μεταβλητής ω, τότε $y_m = p(\omega = m|\boldsymbol{x^c})$ είναι οι πιθανότητες που αντιστοιχούν στις τιμές τις ω δηλαδή y_1=p(1|$\mathbf{x^c}$),..., y_m=p(m|$\mathbf{x^c}$),..., y_M=p(M|$\mathbf{x^c}$) και θ. Το ιδανικό θα ήταν οι πιθανότητες $p(\omega = m|\boldsymbol{x^c})$ να ήταν ή να συνέκλιναν με πιθανότητες $q(\omega = m|\boldsymbol{x^c}) = \begin{cases} 1 \, \alpha\nu \, m = c \\ 0 \, \alpha\nu \, m \neq c \end{cases}$

Ένα μέτρο σύγκρισης των p(ω) και q(ω) δίνεται από την σχέση της *cross entropy*:

$$H(q,p) = -\sum_{\forall \omega} q(\omega) \cdot \log\big(p(\omega)\big) =$$

$$-\sum_{m=1}^{M} q(\omega = m|\boldsymbol{x}^c) \cdot \log\big(p(\omega = m|\boldsymbol{x}^c)\big) = -\sum_{m=1}^{M} q(\omega = m|\boldsymbol{x}^c) \cdot \log(y_m)$$

Για παράδειγμα σε ένα σύστημα ταξινόμησης τεσσάρων κλάσεων εισέρχεται το άνυσμα $\mathbf{x}^2$ και οι έξοδοι των τεσσάρων νευρώνων είναι

$y_1=0.2,\ y_2=0.5\ y_3=0.1\ y_4=0.2$

και προφανώς

$q(1|\mathbf{x}^2)=0,\ q(2|\mathbf{x}^2)=1,\ q(3|\mathbf{x}^2)=0,\ q(4|\ \mathbf{x}^2)=0$

$H(q,p)=-(0\cdot\log(0.2)+1\cdot\log(0.5)+0\cdot\log(0.1)+\ 0\cdot\log(0.2))=2$

Όσο η y_2 τείνει στην μονάδα η H θα τείνει στο μηδέν.

Κατόπιν αυτού μπορούμε να εισάγουμε ως συνάρτηση κόστους (cost) ή απώλειας (loss) την σχέση

$$L(\boldsymbol{W}) = -\sum_i \sum_m q(\omega = m|\boldsymbol{x}_i^c)\cdot \log(y_{mi}) = -\sum_i \log(y_{ci})$$

Που αθροίζει τις τιμές της *cross entropy* που προκύπτουν από την εισαγωγή στο σύστημα κάθε προτύπου $\mathbf{x}_i$, i=1...I, I το πλήθος των προτύπων.

$$y_{ci} = \frac{e^{W_c^T\cdot x_i}}{\sum_m e^{W_m^T\cdot x_i}}$$

Για τον νευρώνα j στο επίπεδο εξόδου $j \in \{1 \dots M\}$ *και* $\mathbf{w}_j$ *το άνυσμα στήλης των βαρών του ισχύει*

$$\frac{\partial L}{\partial w_j} = \frac{\partial}{\vartheta w_j}(-\sum_i \log(y_{ci})) = -\sum_i \frac{\partial}{\vartheta w_j}\log(y_{ci}) = -\sum_i \frac{\partial}{\vartheta w_j}\log\left(\frac{e^{W_c^T\cdot x_i}}{\sum_m e^{W_m^T\cdot x_i}}\right)$$

Για την παραγώγηση γράφω υπό μορφή σύνθετης συνάρτησης ($y_c=y_{ci}$)

$$z_j = \boldsymbol{w}_j^T\cdot \boldsymbol{x}_i,\quad y_c = \frac{e^{z_c}}{\sum_m e^{z_m}},\quad l_i = \log(y_c)$$

$$\frac{\partial l}{\partial \boldsymbol{w}_j} = \frac{\partial l}{\partial y_c}\frac{\partial y_c}{\partial z_j}\frac{\partial z_j}{\partial \boldsymbol{w}_j}$$

$$\frac{\partial z_j}{\partial \boldsymbol{w}_j} = \boldsymbol{x}_i^T,\ \frac{\partial l_i}{\partial y_c} = \frac{1}{y_c},\frac{\partial y_c}{\partial z_j} = \frac{\partial}{\partial z_j}\left(\frac{e^{z_c}}{\sum_m e^{z_m}}\right) = \begin{cases} \frac{e^{z_c}\cdot\sum_m e^{z_m} - e^{z_c}\cdot e^{z_c}}{(\sum_m e^{z_m})^2} = y_c(1-y_c) & \alpha\nu\ j = c \\ \frac{-e^{z_c}\cdot e^{z_j}}{(\sum_m e^{z_m})^2} = -y_c\cdot y_j & \alpha\nu\ j \neq c \end{cases}$$

άρα $\frac{\partial l_i}{\partial w_j} = \begin{cases}(1-y_c)\cdot \boldsymbol{x}_i^T & \alpha\nu\ j = c \\ -y_j\cdot \boldsymbol{x}_i^T & \alpha\nu\ j \neq c\end{cases}$ *, για κάθε συναπτικό βάρος του j νευρώνα*

$$\frac{\partial l_i}{\partial w_{nj}} = \begin{cases} (1 - y_c) \cdot x_{ni} & \alpha\nu\ j = c \\ -y_j \cdot x_{ni} & \alpha\nu\ j \neq c \end{cases}$$ και τελικά

$$\frac{\partial L}{\partial w_{nj}} = -\sum_i \frac{\partial l_i}{\partial w_{nj}}$$