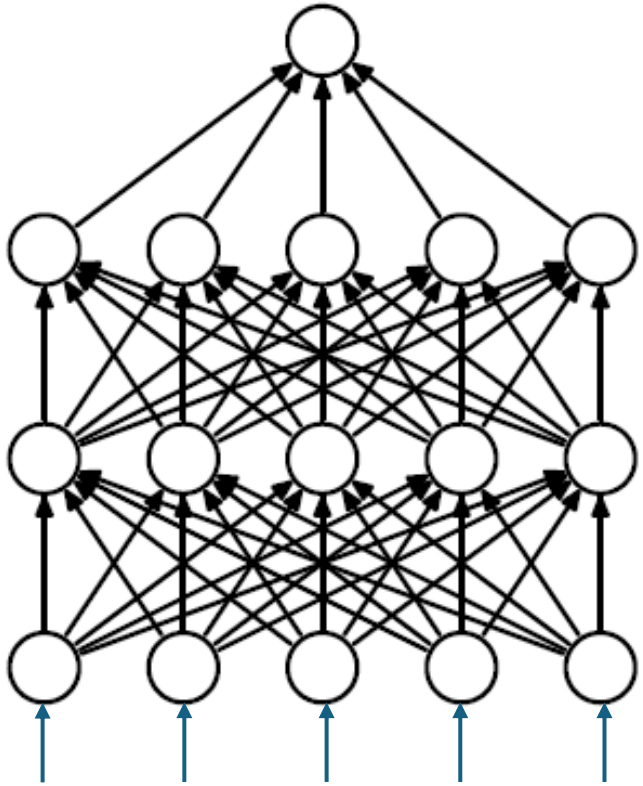
Εισαγωγή στα Συνελικτικά Νευρωνικά Δίκτυα

Αθανάσιος Ροντογιάννης

Αν. Καθηγητής ΣΗΜΜΥ-ΕΜΠ

Δίκτυα MLP

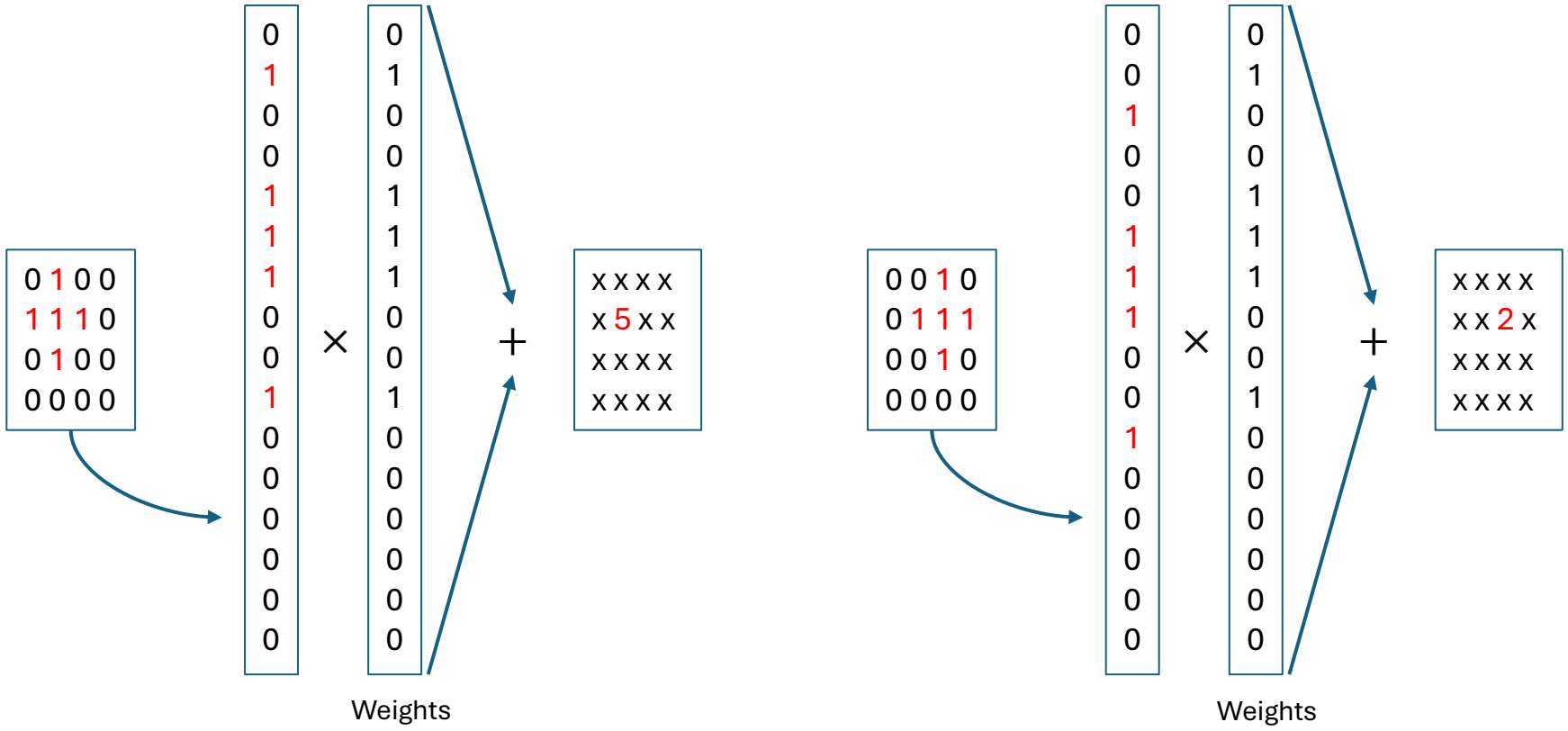


- Η βασική λειτουργία σε κάθε κρυφό στρώμα ενός δικτύου MLP είναι ο υπολογισμός των ενεργοποιήσεων $\mathbf{z} = \sigma(\boldsymbol{W}^T \boldsymbol{x})$, όπου $\boldsymbol{x}$ είναι η είσοδος στο στρώμα, $\boldsymbol{W}$ είναι ο πίνακας βαρών, και $\sigma(\cdot)$ είναι η μη γραμμική συνάρτηση ενεργοποίησης.
- Το i-οστό στοιχείο του κρυφού στρώματος έχει τιμή $z_i = \sigma(\boldsymbol{w}_i^T \boldsymbol{x})$.
- Μπορούμε να θεωρήσουμε αυτή την πράξη εσωτερικού γινομένου ως σύγκριση της εισόδου $\boldsymbol{x}$ με ένα μαθημένο πρότυπο ή μοτίβο $\boldsymbol{w}_i$. Αν η σύγκριση είναι καλή (μεγάλο θετικό εσωτερικό γινόμενο), η ενεργοποίηση αυτής της μονάδας θα είναι υψηλή (υποθέτοντας π.χ., μη γραμμικότητα ReLU), σηματοδοτώντας ότι το i -οστό μοτίβο υπάρχει στην είσοδο.

Περιορισμοί δικτύων MLP

- Τα δίκτυα MLP είναι σχεδιασμένα για εισόδους σταθερού μεγέθους, κατά συνέπεια δεν δουλεύουν για εισόδους μεταβλητού μεγέθους.
- Οι υπολογιστικές απαιτήσεις για την εκπαίδευση των MLPs γίνονται απαγορευτικές όταν οι είσοδοι είναι εικόνες ή, γενικότερα, πολυδιάστατες οντότητες.
- Ένα μοτίβο που εμφανίζεται σε μία θέση της εισόδου μπορεί να μην αναγνωριστεί αν εμφανιστεί σε διαφορετική θέση, δηλαδή, το μοντέλο μπορεί να μην παρουσιάζει μεταφορική αμεταβλητότητα (translational invariance)

Περιορισμοί δικτύων MLP

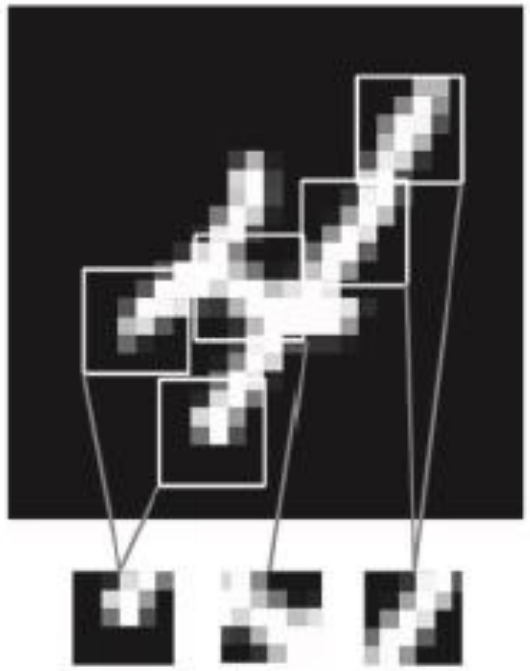


Ανίχνευση μοτίβων σε δισδιάστατες εικόνες με χρήση MLP: Το MLP αποτυγχάνει όταν αλλάζει η θέση του μοτίβου. Το εσωτερικό γινόμενο δεν είναι translational invariant.

Λύση: Συνελικτικά νευρωνικά δίκτυα (CNNs)

- Τα CNNs είναι αποτελεσματικά για την ανάλυση δομημένων δεδομένων, όπως χρονοσειρές και εικόνες.
- Στα CNNs χρησιμοποιείται συνέλιξη αντί για εσωτερικό γινόμενο.
- Τα CNNs **μαθαίνουν** μέσω εκπαίδευσης ένα σύνολο από πίνακες βαρών (φίλτρα ή πυρήνες), που είναι κοινοί για όλες τις εξόδους ενός κρυφού στρώματος.
- Η συνέλιξη εξασφαλίζει μεταφορική αμεταβλητότητα (translational invariance).
- Λόγω του μικρού μεγέθους των πυρήνων, ο αριθμός των συντελεστών και κατ' επέκταση η υπολογιστική πολυπλοκότητα του μοντέλου μειώνονται σημαντικά.

Πυρήνες για ταξινόμηση

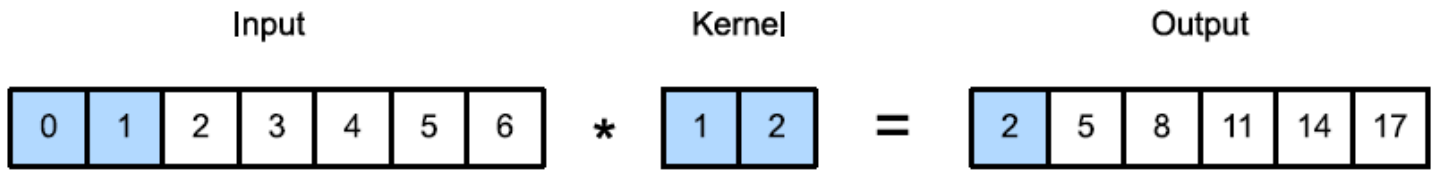


Μπορούμε να ταξινομήσουμε έναν ψηφίο εξετάζοντας συγκεκριμένα διακριτικά χαρακτηριστικά (πρότυπα εικόνας) που εμφανίζονται στις σωστές (σχετικές) θέσεις.

Συνέλιξη – Ετεροσυσχέτιση (1Δ)

Συνέλιξη: $$[\boldsymbol{x} * \boldsymbol{w}](i) = \sum_{l=0}^{L-1} w_l x_{i-l} = \sum_{p=0}^{P-1} x_p w_{i-p}$$

Ετεροσυσχέτιση: $$\boxed{[\boldsymbol{x} \circledast \boldsymbol{w}](i) = \sum_{l=0}^{L-1} w_l x_{i+l}}$$ (Στην ορολογία των CNNs καλείται συνέλιξη)



Ο πυρήνας είναι κοινός για όλες τις εξόδους. Αν είχαμε MLP (εσωτερικά γινόμενα) θα χρειαζόμασταν 35 παραμέτρους για κρυφό στρώμα με 5 νευρώνες, ενώ τώρα μόνο 2 παραμέτρους.

2Δ-Συνέλιξη

Θεωρούμε ένα $H \times W$ φίλτρο (πυρήνα) $\boldsymbol{W}$ και μία 2Δ εικόνα $\boldsymbol{X}$. Ορίζουμε:

$$[\boldsymbol{X} \circledast \boldsymbol{W}](i,j) = \sum_{l=0}^{H-1} \sum_{m=0}^{W-1} w_{l,m} x_{i+l,j+m}$$

$$\mathbf{Y} = \begin{pmatrix} w_1 & w_2 \\ w_3 & w_4 \end{pmatrix} \circledast \begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{pmatrix}$$

$$= \begin{pmatrix} (w_1 x_1 + w_2 x_2 + w_3 x_4 + w_4 x_5) & (w_1 x_2 + w_2 x_3 + w_3 x_5 + w_4 x_6) \\ (w_1 x_4 + w_2 x_5 + w_3 x_7 + w_4 x_8) & (w_1 x_5 + w_2 x_6 + w_3 x_8 + w_4 x_9) \end{pmatrix}$$

Input | Kernel | Output

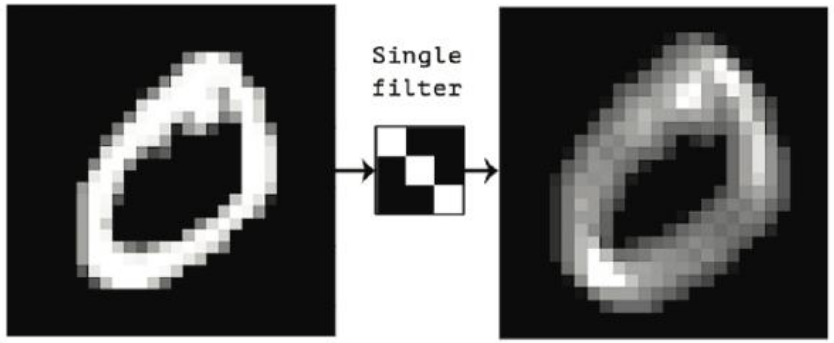
0	1	2
3	4	5
6	7	8

*

0	1
2	3

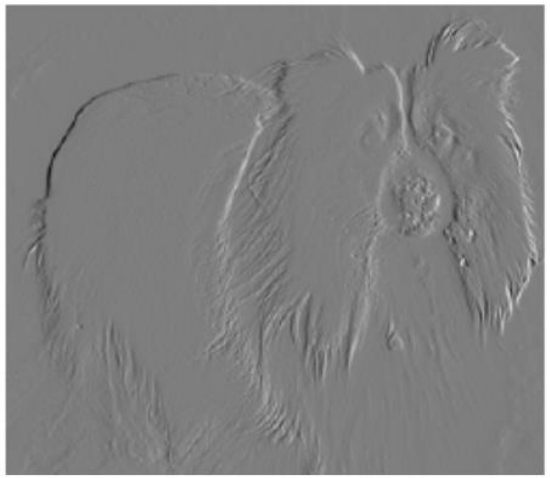
=

19	25
37	43



Ανίχνευση ακμών

$$\boldsymbol{W} = [1 \quad -1]$$



Το πεδίο αποδοχής (receptive field) για κάθε pixel εξόδου είναι η περιοχή της εισόδου που επηρεάζει την έξοδο.

$$\boldsymbol{W} = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Χάρτης χαρακτηριστικών

- Μπορούμε να θεωρήσουμε τη δισδιάστατη συνέλιξη (2D convolution) ως αντιστοίχιση προτύπων (template matching).
- Η έξοδος σε ένα σημείο (i, j) θα είναι μεγάλη, αν το αντίστοιχο τμήμα της εικόνας, που είναι κεντραρισμένο στο (i, j), μοιάζει με το $\boldsymbol{W}$.
- Αν το πρότυπο $\boldsymbol{W}$ αντιστοιχεί σε μια προσανατολισμένη ακμή, τότε η συνέλιξη με αυτό θα "φωτίσει" περιοχές της εικόνας-εξόδου που περιέχουν ακμές που ταιριάζουν με αυτόν τον προσανατολισμό.
- Συνεπώς, μπορούμε να σκεφτούμε τη συνέλιξη ως μια μορφή ανίχνευσης χαρακτηριστικών (feature detection).
- Η προκύπτουσα έξοδος $\boldsymbol{Y} = \boldsymbol{X} \circledast \boldsymbol{W}$ ονομάζεται χάρτης χαρακτηριστικών (feature map).

Η συνέλιξη ως πολλαπλασιασμός πινάκων

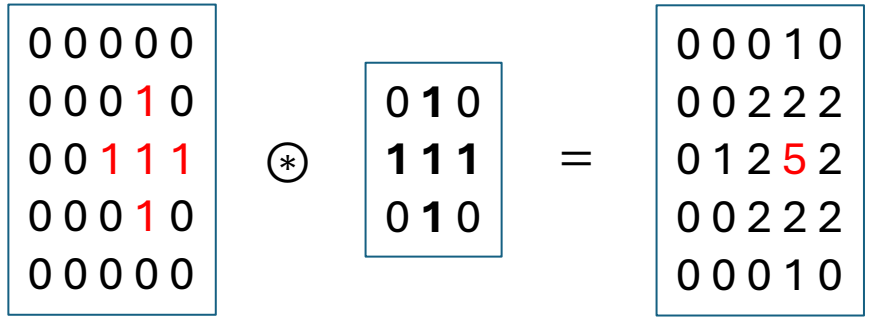
$$\boldsymbol{y} = \mathbf{C}\boldsymbol{x} = \left(\begin{array}{ccc|ccc|ccc} w_1 & w_2 & 0 & w_3 & w_4 & 0 & 0 & 0 & 0 \\ 0 & w_1 & w_2 & 0 & w_3 & w_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & w_1 & w_2 & 0 & w_3 & w_4 & 0 \\ 0 & 0 & 0 & 0 & w_1 & w_2 & 0 & w_3 & w_4 \end{array}\right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{pmatrix}$$

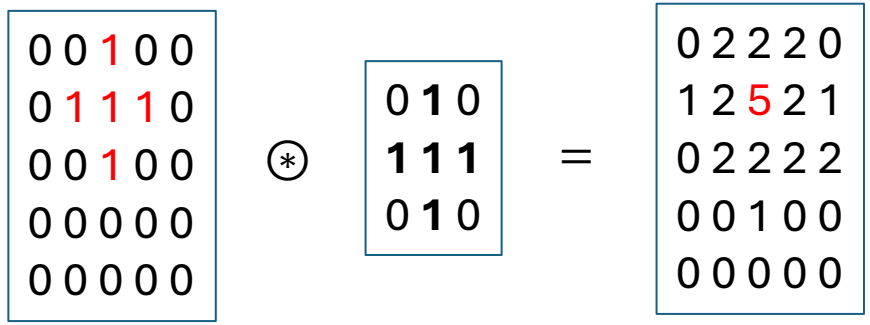
$$= \begin{pmatrix} w_1 x_1 + w_2 x_2 + w_3 x_4 + w_4 x_5 \\ w_1 x_2 + w_2 x_3 + w_3 x_5 + w_4 x_6 \\ w_1 x_4 + w_2 x_5 + w_3 x_7 + w_4 x_8 \\ w_1 x_5 + w_2 x_6 + w_3 x_8 + w_4 x_9 \end{pmatrix}$$

- Παρατηρούμε ότι τα CNNs μπορούν να μετασχηματιστούν σε MLPs, αλλά τώρα οι πίνακες βαρών έχουν μια ειδική αραιή δομή και τα στοιχεία τους είναι κοινά μεταξύ διαφορετικών χωρικών θέσεων.
- Αυτό εξασφαλίζει translation invariance και μειώνει σημαντικά τον αριθμό των παραμέτρων σε σύγκριση με έναν πίνακα βαρών σε ένα πλήρως συνδεδεμένο ή πυκνό στρώμα, όπως αυτά που χρησιμοποιούνται στα MLPs.

Η συνέλιξη είναι μεταφορικά αμετάβλητη

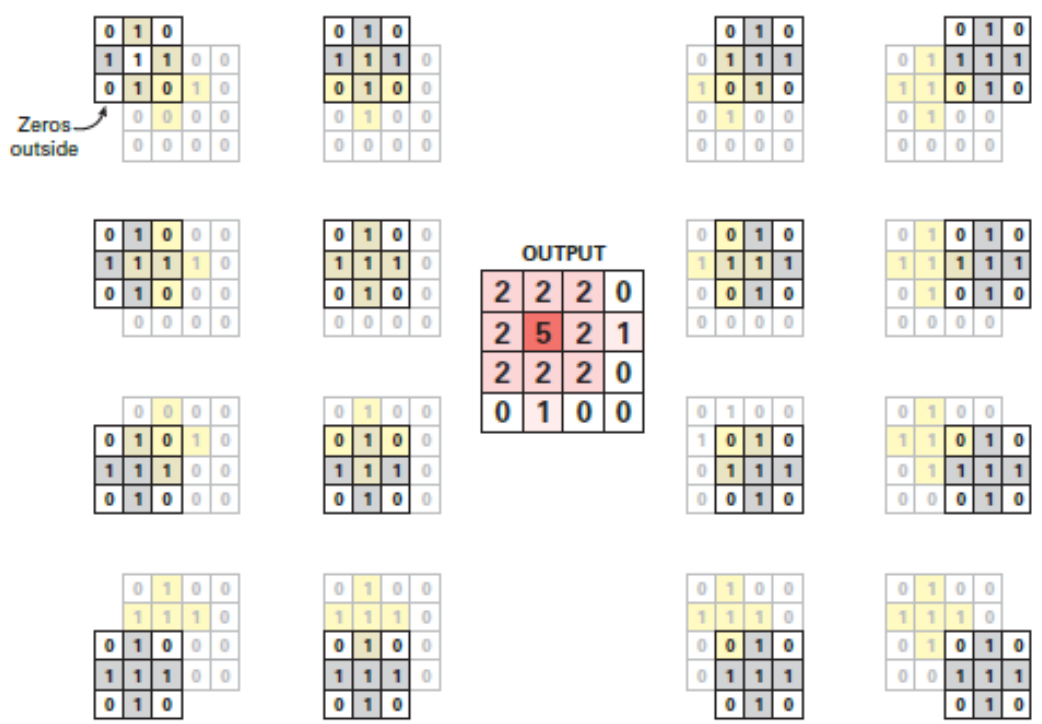
- Με συνέλιξη το μοτίβο εντοπίζεται ανεξάρτητα από τη θέση που βρίσκεται.





Ιδία συνέλιξη (Same convolution)

- Αν ο πυρήνας είναι $f_h \times f_w$ και η εικόνα $x_h \times x_w$, η συνέλιξή τους θα είναι $(x_h - f_h + 1) \times (x_w - f_w + 1)$.
- Αν θέλουμε ο χάρτης χαρακτηριστικών να έχει τις ίδιες διαστάσεις με την αρχική εικόνα θα πρέπει να συμπληρώσουμε μηδενικά στα όρια της εικόνας.
- Η συνέλιξη που προκύπτει ονομάζεται ιδία συνέλιξη.



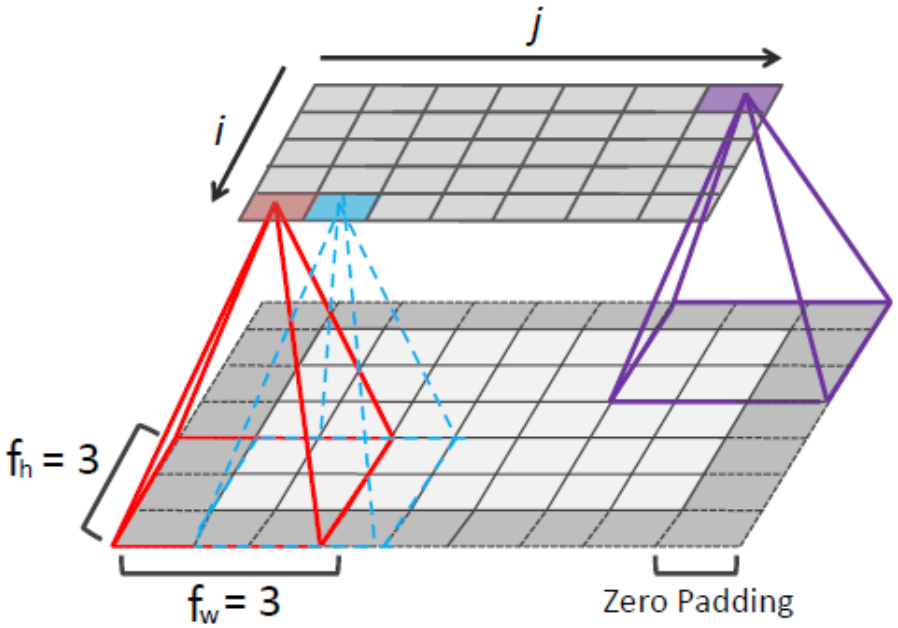
Συμπλήρωμα με μηδενικά (Zero-padding)

- Αν ο πυρήνας είναι $f_h \times f_w$, η εικόνα $x_h \times x_w$ και προσθέτουμε μηδενικά μεγέθους $p_h,\ p_w$ η έξοδος θα είναι:

  $$(x_h - f_h + 2p_h + 1) \times (x_w - f_w + 2p_h + 1).$$

- Στο σχήμα είναι $f_h = f_w = 3, x_h = 5,$ $x_w = 7, p_h, = p_w = 1.$ Άρα οδηγούμαστε σε ιδία συνέλιξη.

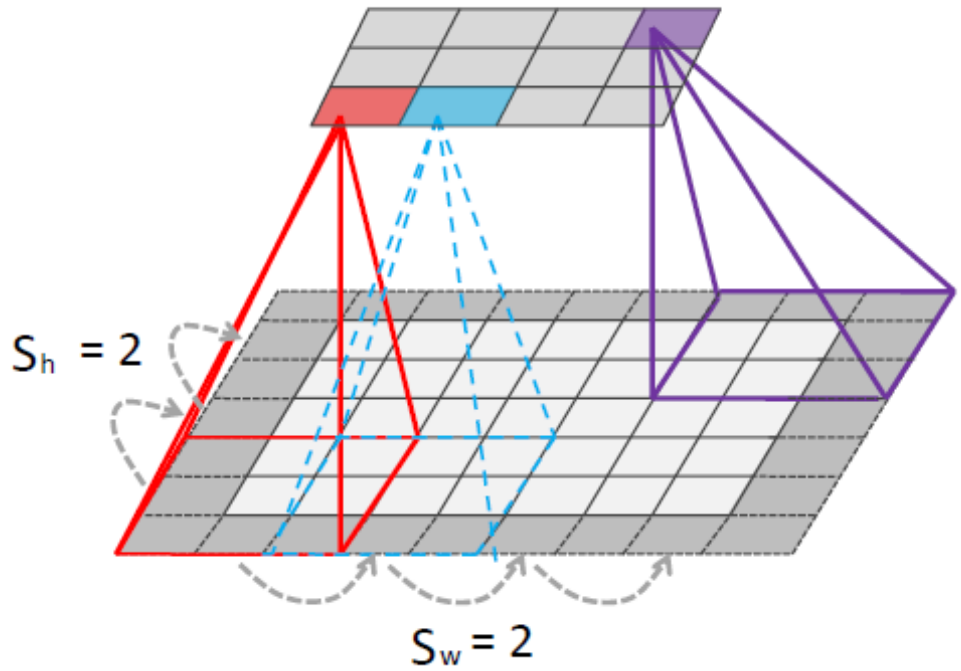
- Γενικά για $2p = f - 1$, η έξοδος θα έχει ίδιες διαστάσεις με την είσοδο.



Strided convolution

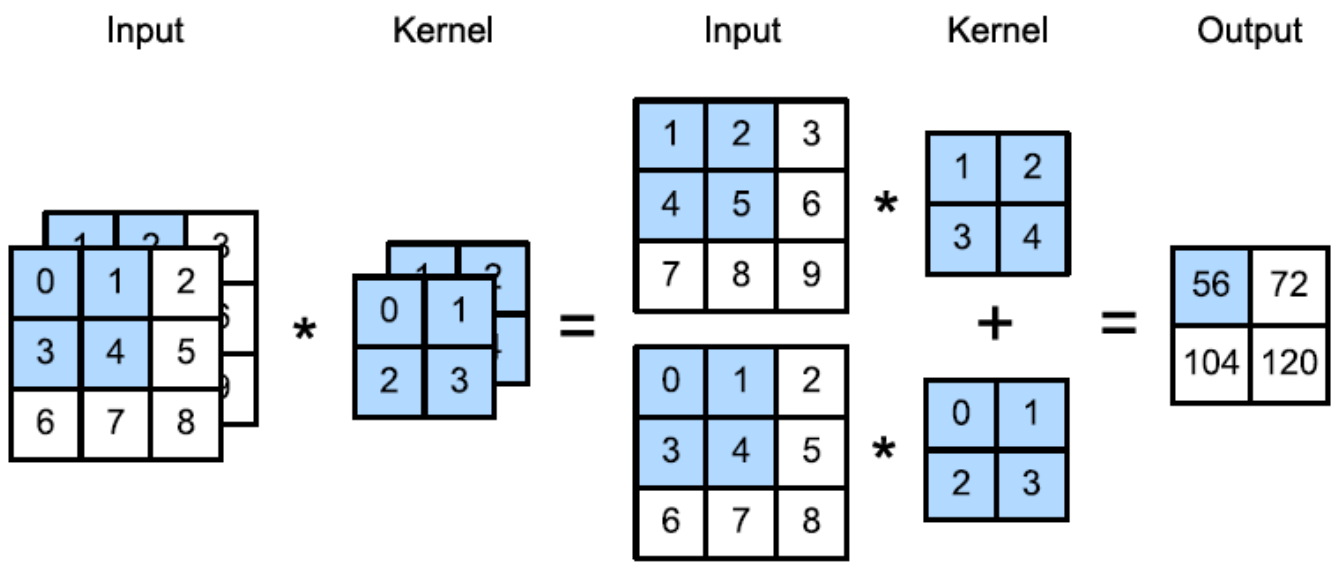
- Είναι μια παραλλαγή της συνέλιξης όπου ο πυρήνας μετατοπίζεται με βήμα (stride) $s > 1$.
- Προκύπτει μια εικόνα μικρότερου μεγέθους (down-sampling), αλλά τα χαρακτηριστικά που μας ενδιαφέρουν διατηρούνται.
- Αν έχουμε και striding, το μέγεθος της εικόνας που προκύπτει είναι:

$$\left\lfloor \frac{x_h - f_h + 2p_h + s_h}{s_h} \right\rfloor \times \left\lfloor \frac{x_w - f_w + 2p_w + s_w}{s_w} \right\rfloor$$



Πολλά κανάλια εισόδου / μία έξοδος

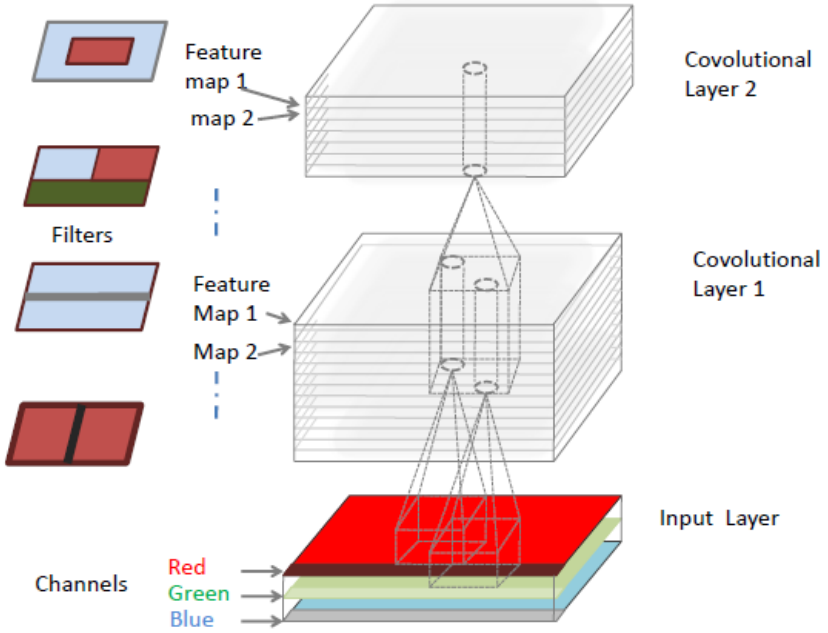
- Τα βάρη συναποτελούν έναν 3Δ τανυστή (tensor).
- Έχουμε έναν πυρήνα ανά κανάλι εισόδου.
- Αθροίζουμε τους χάρτες χαρακτηριστικών που προκύπτουν για κάθε κανάλι.
- Εδώ θεωρούμε το ίδιο stride s και στις δύο διαστάσεις
- b: bias term



Η έξοδος στη θέση (i, j): $y_{i,j} = b + \sum_{l=0}^{H-1} \sum_{m=0}^{W-1} \sum_{c=0}^{C-1} w_{l,m,c} x_{si+l,sj+m,c}$

Πολλαπλά κανάλια εισόδου και εξόδου

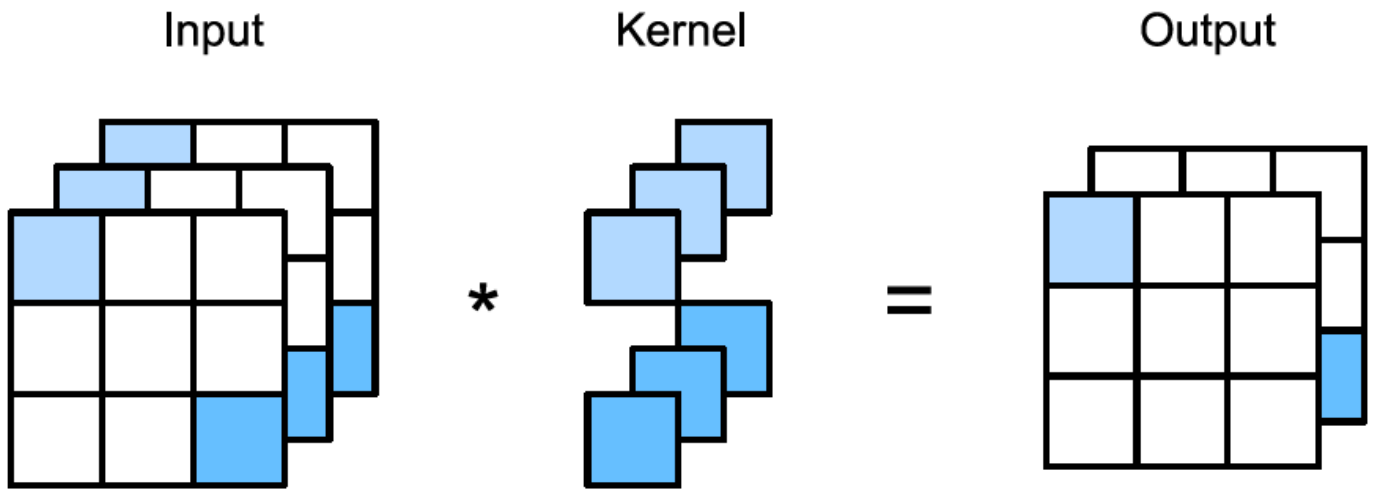
- Για να ανιχνεύσουμε πολλαπλά είδη χαρακτηριστικών χρησιμοποιούμε πολλούς 3Δ πυρήνες ταυτόχρονα
- Τα βάρη δημιουργούν συνολικά έναν 4Δ τανυστή.
- Το φίλτρο που ανιχνεύει ένα χαρακτηριστικό τύπου d στο κανάλι c αποθηκεύεται στο $\boldsymbol{W}_{:,:,c,d}$.
- Στο σχήμα, οι κάθετες στήλες αντιστοιχούν σε ένα σύνολο χαρακτηριστικών εξόδου σε ένα συγκεκριμένο σημείο στο χώρο, $y_{i,j,1:D}$.
- Αυτό πολλές φορές ονομάζεται υπερστήλη (hypercolumn)



Η έξοδος στη θέση (i, j, d): $y_{i,j,d} = b_d + \sum_{l=0}^{H-1} \sum_{m=0}^{W-1} \sum_{c=0}^{C-1} w_{l,m,c,d} x_{si+l,sj+m,c}$

1×1 (σημειακή) συνέλιξη

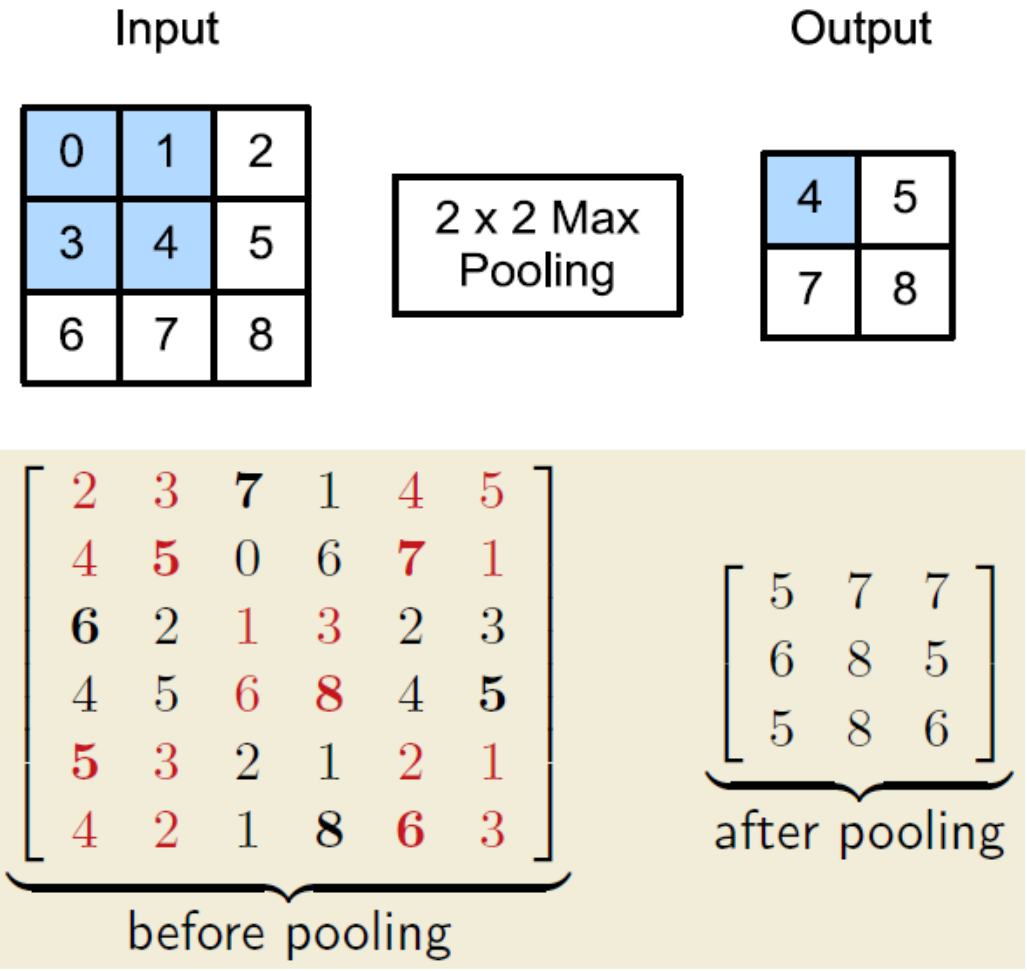
- Μερικές φορές θέλουμε το γραμμικό συνδυασμό των χαρακτηριστικών σε κάθε θέση (pixel).
- Αυτό επιτυγχάνεται με την σημειακή συνέλιξη (pointwise convolution). Οι επιμέρους πυρήνες είναι διανύσματα.
- Με τον τρόπο αυτό ο αριθμός των καναλιών μεταβάλλεται από C σε D, αλλά οι χωρικές διαστάσεις παραμένουν αμετάβλητες.



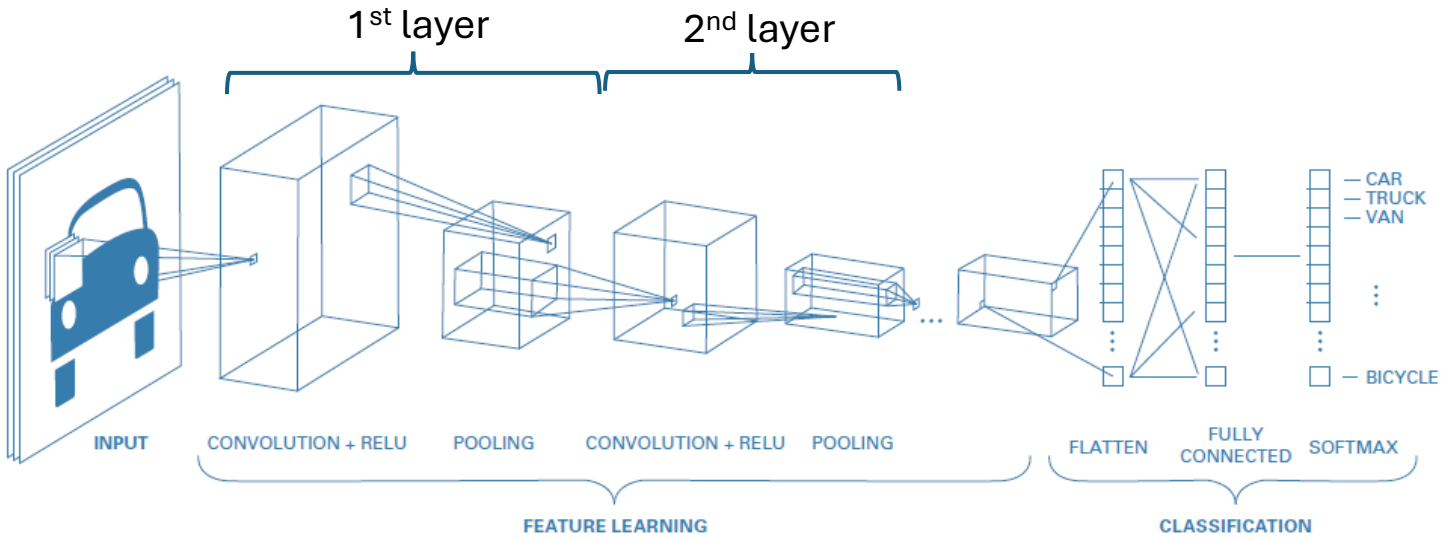
Η έξοδος στη θέση (i, j, d): $y_{i,j,d} = b_d + \sum_{c=0}^{C-1} w_{0,0,c,d} x_{i,j,c}$

Pooling layer

- Το **pooling layer** (στρώμα συγκέντρωσης) μειώνει τη διάσταση των χαρτών χαρακτηριστικών, διατηρώντας παράλληλα τα πιο σημαντικά χαρακτηριστικά.
- Το pooling layer λειτουργεί εφαρμόζοντας μια συνάρτηση σε μικρές υποπεριοχές της εικόνας εισόδου και παράγοντας μια μικρότερη έξοδο (υποδειγματοληψία).
- Δημιουργεί αμεταβλητότητα σε μικρές μετατοπίσεις ή παραμορφώσεις στα δεδομένα εισόδου.
- Τα βασικότερα είδη pooling είναι το max pooling και το average pooling.

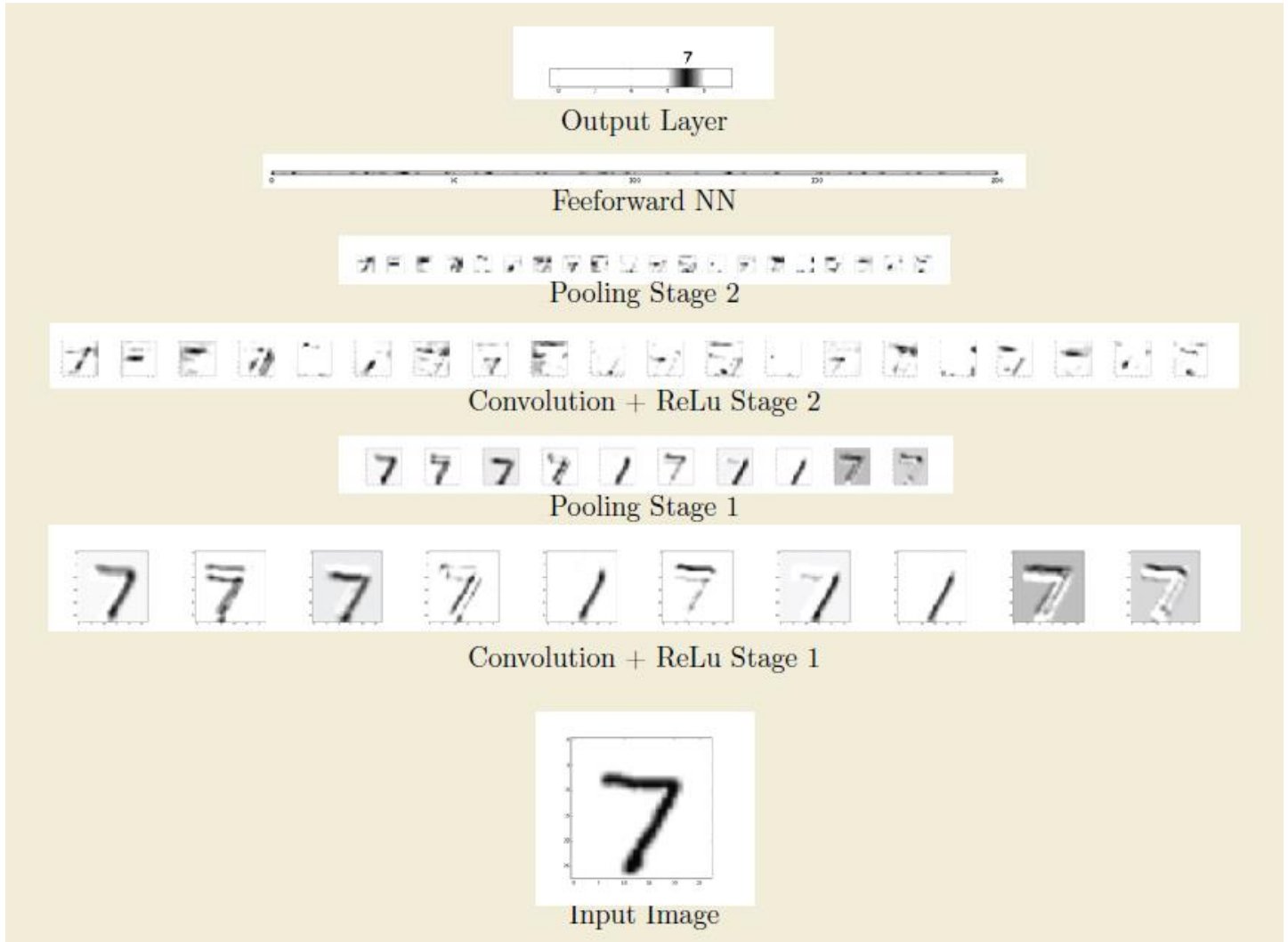


Συνολικό δίκτυο



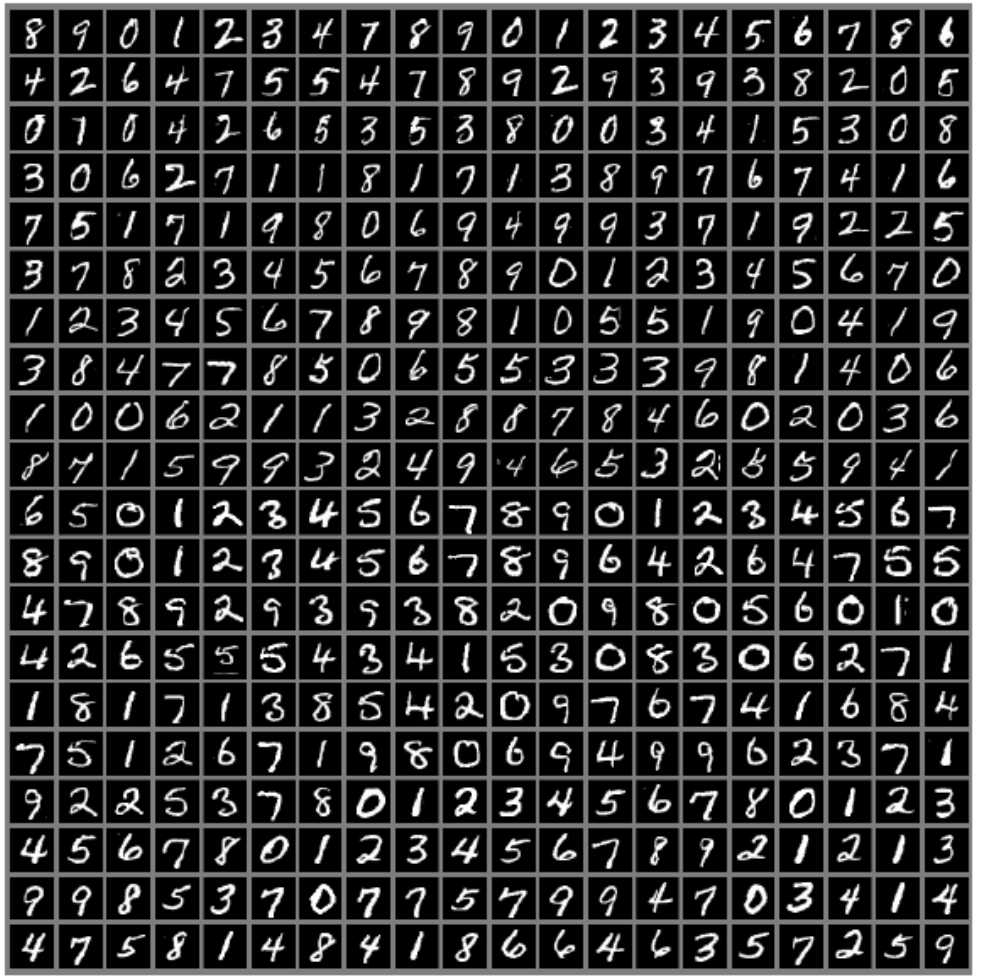
- Κάθε στρώμα του δικτύου αποτελείται από ένα συνελικτικό στάδιο και ένα στάδιο συγκέντρωσης.
- Σε κάθε έξοδο των συνελικτικών σταδίων εφαρμόζεται μια συνάρτηση ενεργοποίησης που στις περισσότερες περιπτώσεις είναι η ReLU.
- Το σύνολο των εξόδων του τελευταίου στρώματος συγκέντρωσης μετατρέπεται σε διανυσματική μορφή (vectorization) και αποτελεί, ως νέο διάνυσμα χαρακτηριστικών, είσοδο σε ένα πλήρως συνδεδεμένο δίκτυο που πραγματοποιεί την ταξινόμηση.
- Συνεπώς το «καθαρό» CNN δημιουργεί μια φειδωλή αναπαράσταση της εισόδου που διατηρεί τα βασικά της χαρακτηριστικά, παράγει δηλαδή τα διανύσματα χαρακτηριστικών με βάση τα οποία πραγματοποιείται η ταξινόμηση.
- Τα βάρη του δικτύου υπολογίζονται μέσω ενός κατάλληλα προσαρμοσμένου στην αρχιτεκτονική αλγόριθμου backpropagation.

Συνολικό δίκτυο στην πράξη

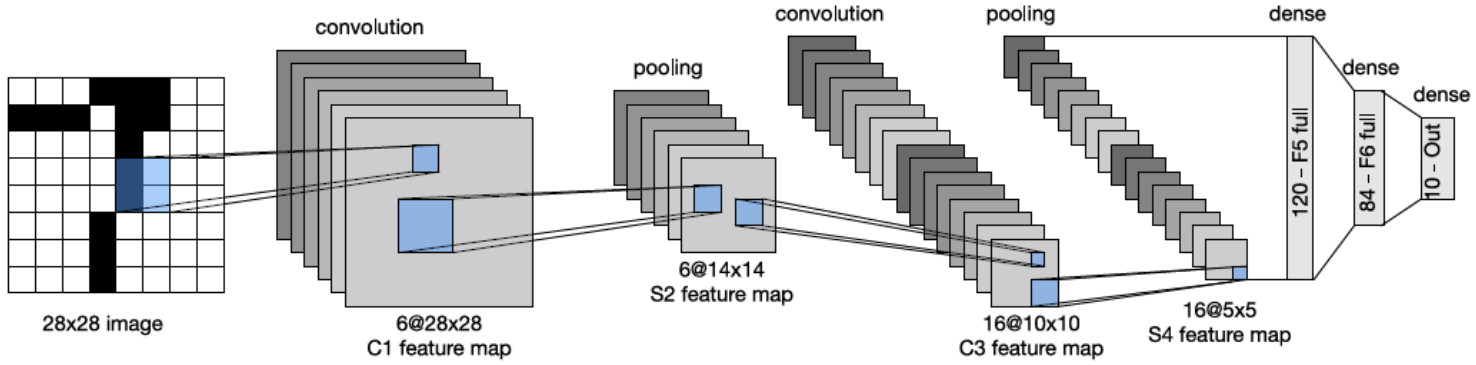


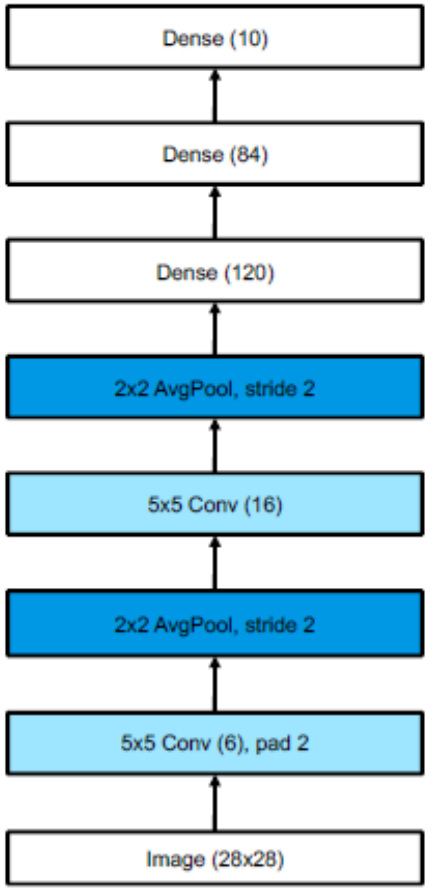
Το σύνολο δεδομένων MNIST

- MNIST (Modified National Institute of Standards and Technology)
- Αποτελείται από εικόνες που αναπαριστούν χειρόγραφα ψηφία.
- Οι εικόνες είναι 28×28 και ο αριθμός των κλάσεων είναι 10.
- Οι αρχικές γκρίζες εικόνες έχουν μετατραπεί σε δυαδικές με κατωφλίωση.
- Υπάρχουν 60000 πρότυπα εκπαίδευσης και 20000 πρότυπα δοκιμής.
- Στο σχήμα φαίνονται 400 εικόνες του MNIST.
- Drosophila of machine learning (Hinton)



LeNet[1]

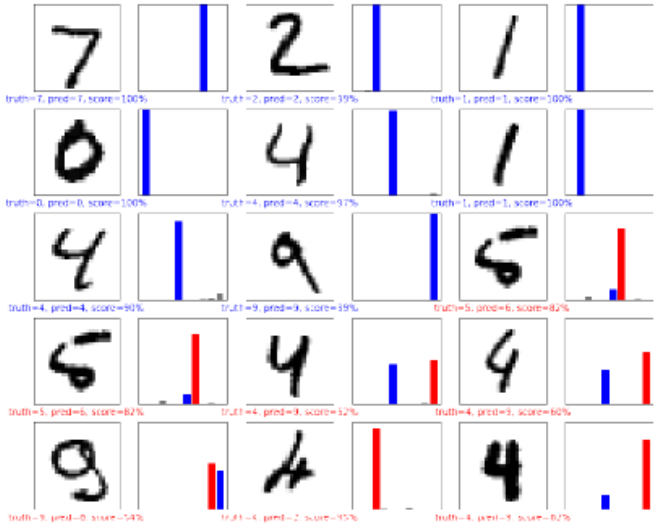


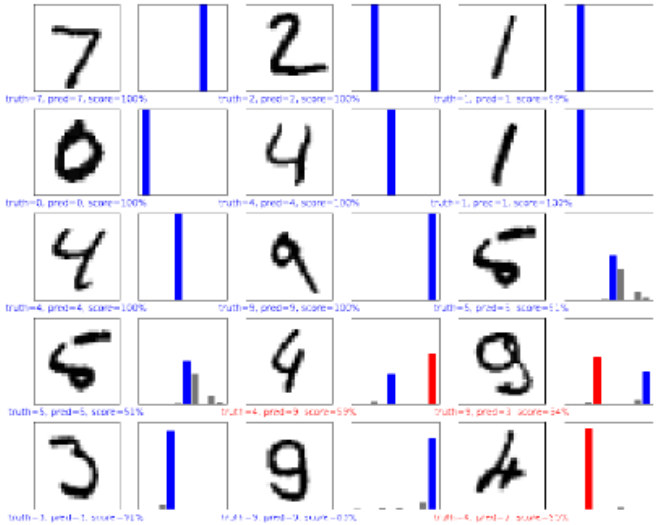


- Συνελικτικό νευρωνικό δίκτυο για την ταξινόμηση εικόνων από το MNIST.
- Αποτελείται από δύο convolution/pooling στρώματα και τρία πλήρως συνδεδεμένα στρώματα. Χρησιμοποιεί μη-γραμμικότητα $tanh$.
- Οι αρχικές γκρίζες εικόνες έχουν μετατραπεί σε δυαδικές με κατωφλίωση.
- Συνδυασμένο με κατάτμηση εικόνων μπορεί να χρησιμοποιηθεί για την αναγνώριση ακολουθιών (χειρόγραφων) ψηφίων ή χαρακτήρων.

[1]Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” In *Proceedings of the IEEE*, 86.11 (1998), pp. 2278-2324.

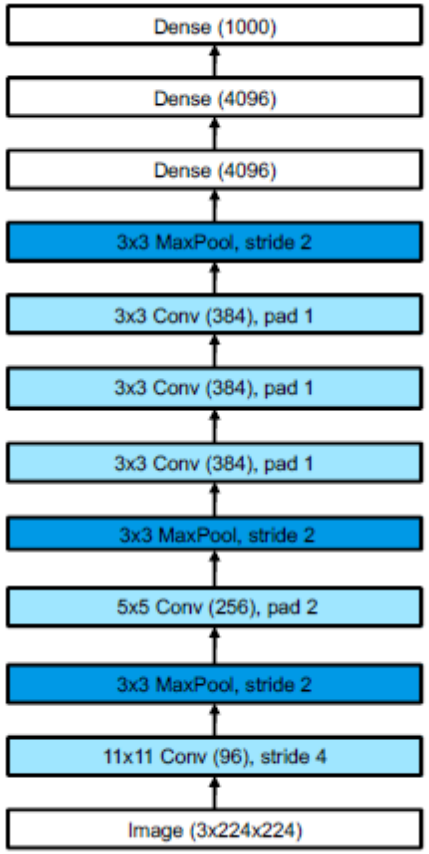
Αποτελέσματα του LeNet στο MNIST

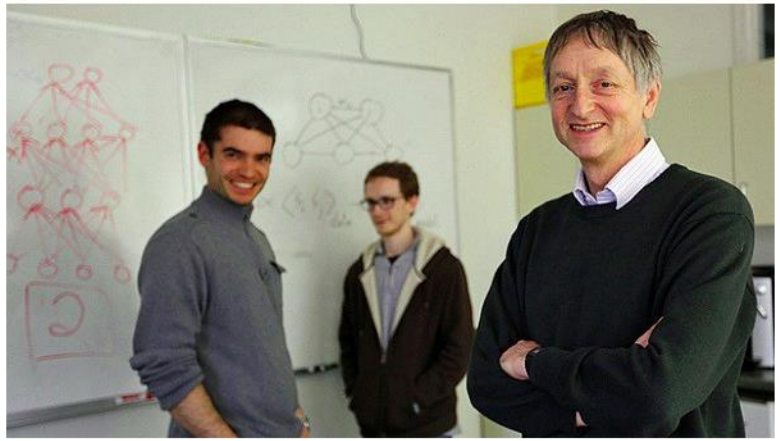




- Αποτελέσματα του LeNet μετά το τέλος της 1ης και 2ης εποχής.
- Ήδη μετά το τέλος της 1ης εποχής το LeNet επιτυγχάνει ακρίβεια 98.7%.

AlexNet[2]

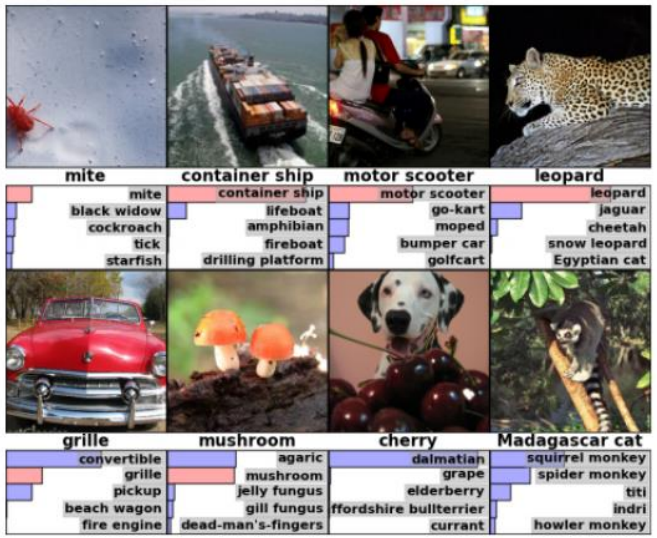


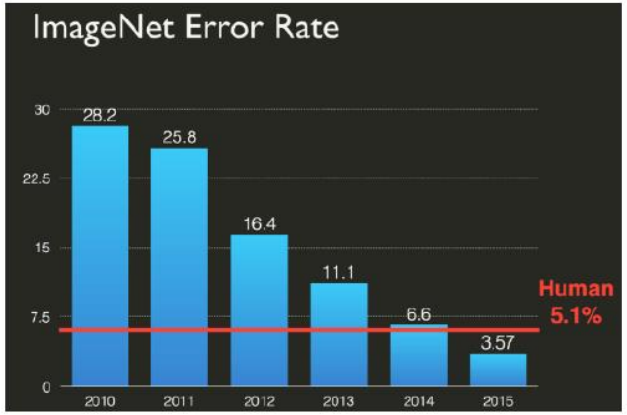


- Το AlexNet διαθέτει περισσότερες από 60M παραμέτρους, που εντοπίζονται κυρίως στα τρία τελευταία πλήρως συνδεδεμένα στρώματα.
- Ως είσοδο δέχεται έγχρωμες εικόνες 224×224.
- Το AlexNet είναι παρόμοιο με το LeNet με τις εξής βασικές διαφορές: α) είναι βαθύτερο, β) χρησιμοποιεί ReLU αντί για tanh, γ) περιέχει συνεχόμενα συνελικτικά στάδια.
- Συνεχόμενα συνελικτικά στάδια δημιουργούν μεγαλύτερα πεδία υποδοχής, π.χ. τρία 3×3 στρώματα οδηγούν σε πεδία υποδοχής 7×7. Ταυτόχρονα όμως εισάγουν και περισσότερες μη-γραμμικότητες σε σχέση με ένα στρώμα 7×7.
- Θεωρείται ένα εξαιρετικό επίτευγμα της μηχανικής.

[2]A. Krizhevsky, I. Sutskever, and G. Hinton. “Imagenet classification with deep convolutional neural networks”. In: NIPS. 2012.

Το ImageNet και ο διαγωνισμός ILSVRC





- Το σύνολο δεδομένων ImageNet περιλαμβάνει περίπου 14M εικόνες «αντικειμένων» από 20000 κλάσεις. Η διαστάσεις των εικόνων είναι $256 \times 256 \times 3$.
- Στον ILSVRC (ImageNet Large Scale Visual Recognition Challenge) χρησιμοποιήθηκε ένα υποσύνολο 1.3M εικόνων από 1000 κλάσεις. Στόχος ήταν η ελαχιστοποίηση του top-5 error rate, δηλαδή να εξασφαλιστεί ότι η σωστή ετικέτα είναι μεταξύ των 5 πιο πιθανών προβλέψεων.
- Το 2012 με το AlexNet το top-5 error rate score μειώθηκε δραματικά από 28.5% σε 16.4%. Το 2015 είναι η πρώτη χρονιά που τα CNNs ξεπέρασαν τον άνθρωπο στον ILSVRC.

Βιβλιογραφία

- K. P. Murphy, Probabilistic Machine Learning: An Introduction, MIT Press, 2022.
- S. Theodoridis, Machine Learning: A Bayesian and Optimization Perspective, 2nd Edition, Academic Press, 2020.
- I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, https://www.deeplearningbook.org/