Slides created by Sofia Eleftheriou

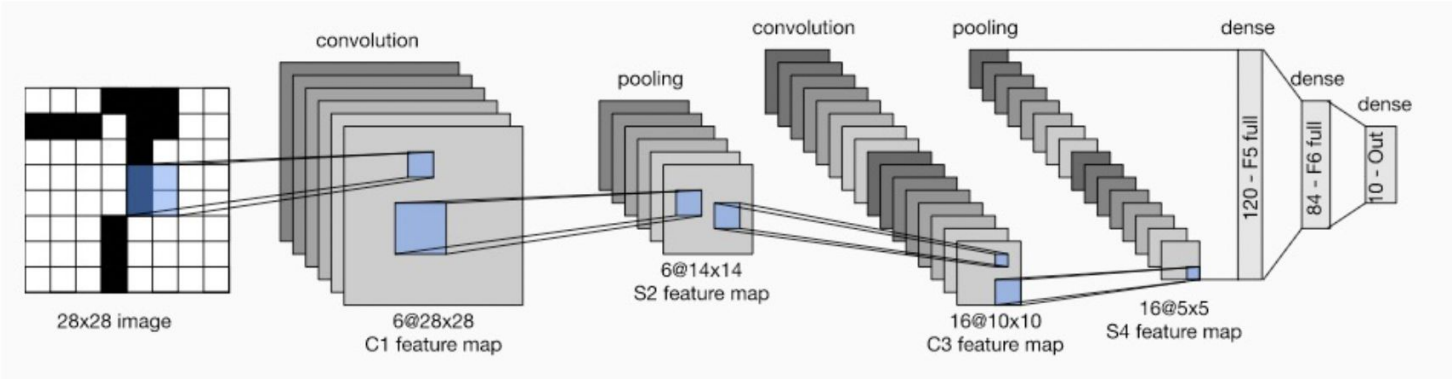
Ασκήσεις μελέτης Β7 + Β8
Lab 9

Human-Computer Interaction, AUEB
Εαρινό εξάμηνο 2023-2024

Lab Assistant: Sofia Eleftheriou

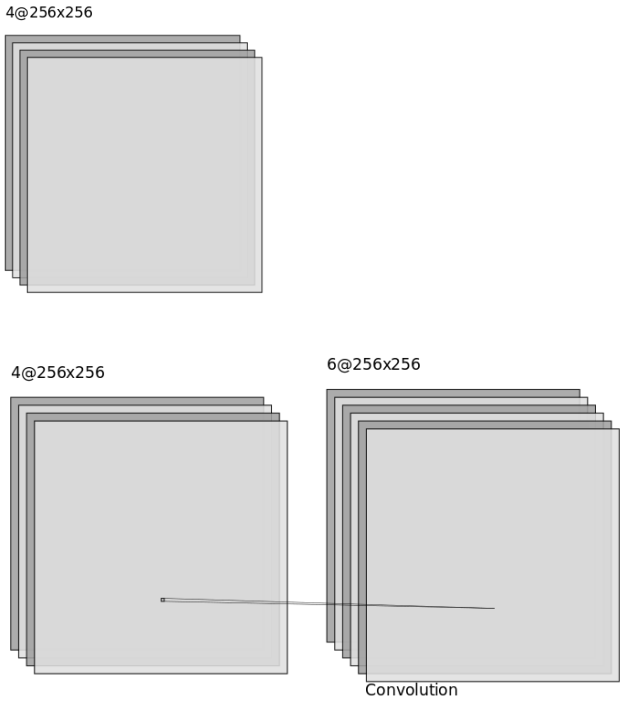
Άσκηση B7.1.

Θέλουμε να χρησιμοποιήσουμε μια τροποποιημένη μορφή του συνελικτικού νευρωνικού δικτύου της διαφάνειας 23 (LeNet), για να εντοπίζουμε τις συντεταγμένες (x, y) του κέντρου του κεφαλιού και των δύο ώμων σε εικόνες (ή video frames) που περιλαμβάνουν έναν μόνο άνθρωπο μπροστά από μια κονσόλα ηλεκτρονικών παιχνιδιών εφοδιασμένη με έγχρωμη κάμερα και κάμερα βάθους. Η κάθε εικόνα έχει ανάλυση 256x256 και τέσσερα κανάλια (RGB και βάθος), δηλαδή είναι ένας τανυστής (tensor) τριών αξόνων, με σχήμα (shape) (256, 256, 4). Όπως στο σχήμα της διαφάνειας 23, υπάρχουν δύο συνελικτικά στρώματα (convolutional layers) που παράγουν 6 και 16 χάρτες χαρακτηριστικών (feature maps) αντίστοιχα αλλά οι συνελίξεις χρησιμοποιούν πυρήνες (kernels) με παράθυρο 3x3 και είναι ευρείες (wide, same), δηλαδή χρησιμοποιούν padding και διατηρούν την ανάλυση της αρχικής εικόνας σε κάθε κανάλι (βλ. και διαφάνεια 10). Τα δύο στρώματα υπο- δειγματοληψίας (pooling) χρησιμοποιούν max-pooling με παράθυρο 4x4 και βήμα (stride) 4 και στους δύο άξονες. Τα δύο πρώτα (τα κρυφά) πυκνά (dense) στρώματα του τελικού MLP εξακολουθούν να έχουν 120 και 84 νευρώνες αντίστοιχα.



α) Πόσους πυρήνες θα χρησιμοποιεί το πρώτο συνελικτικό στρώμα και τι σχήμα θα έχει ο καθένας;

Το πρώτο συνελικτικό στρώμα θα χρησιμοποιεί 6 πυρήνες, ώστε να προκύπτουν 6 χάρτες χαρακτηριστικών, όπως φαίνεται στο σχήμα της διαφάνειας 23. Ο κάθε πυρήνας θα έχει 4 φέτες (slices), αφού η είσοδος έχει τώρα 4 κανάλια. Γνωρίζουμε από την εκφώνηση ότι κάθε πυρήνας εφαρμόζει σε κάθε κανάλι της εισόδου παράθυρο 3x3. Επομένως κάθε ένας από τους 6 πυρήνες θα είναι ένας τανυστής (tensor) τριών αξόνων, με σχήμα (shape) (3, 3, 4).



Cheat sheet

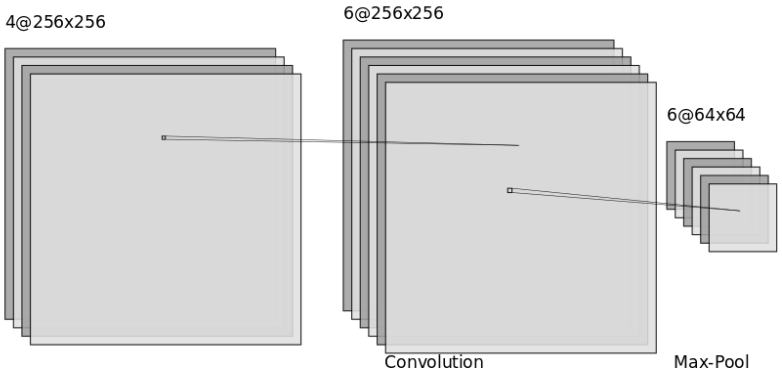
- Input: n X n X nc
- Padding: p
- Stride: s
- Filter size: f X f X nc'
- Output: [(n+2p-f)/s+1] X [(n+2p-f)/s+1] X nc'

[1]Here, nc is the number of channels in the input and filter, while nc' is the number of filters.

[2]**Same**: Here, we apply padding so that the output size is the same as the input size, i.e., s=1, n+2p-f+1 = n. So, p = (f-1)/2

β) Τι ανάλυση θα έχουν οι χάρτες χαρακτηριστικών που θα προκύπτουν από το πρώτο στρώμα max-pooling;

Αφού τα συνελικτικά στρώματα που χρησιμοποιούμε διατηρούν την ανάλυση σε κάθε κανάλι, κάθε ένας από τους 6 χάρτες χαρακτηριστικών (κανάλια) που προκύπτουν από το πρώτο συνελικτικό στρώμα, δηλαδή κάθε κανάλι στην είσοδο του πρώτου στρώματος max-pooling θα εξακολουθεί να έχει ανάλυση 256x256. Αφού κάθε στρώμα max-pooling χρησιμοποιεί παράθυρο 4x4 με βήμα (stride) 4 και στους δύο άξονες, ο κάθε ένας από τους 6 χάρτες που εξέρχονται από το πρώτο στρώμα max-pooling θα έχει ανάλυση (256/4)x(256/4), δηλαδή 64x64.



Cheat sheet

- Filter size: f X f
- Stride: s
- Max or average pooling

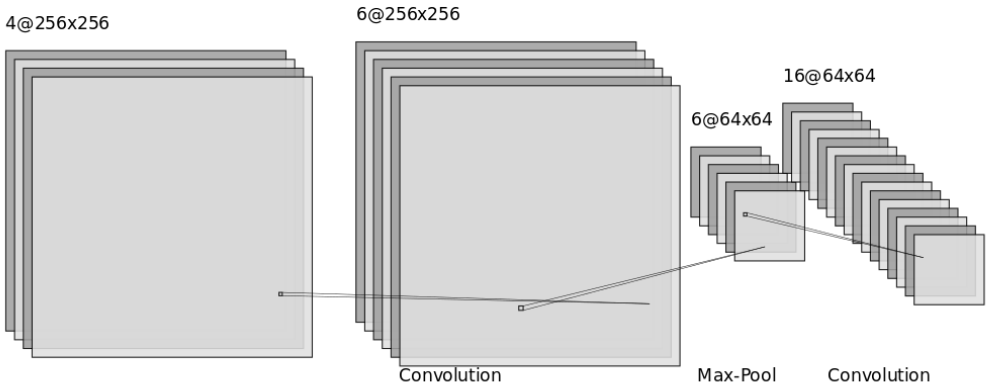
If the input of the pooling layer is nh X nw X nc,

then the output will be:

[{(nh – f) / s + 1} X {(nw – f) / s + 1} X nc].

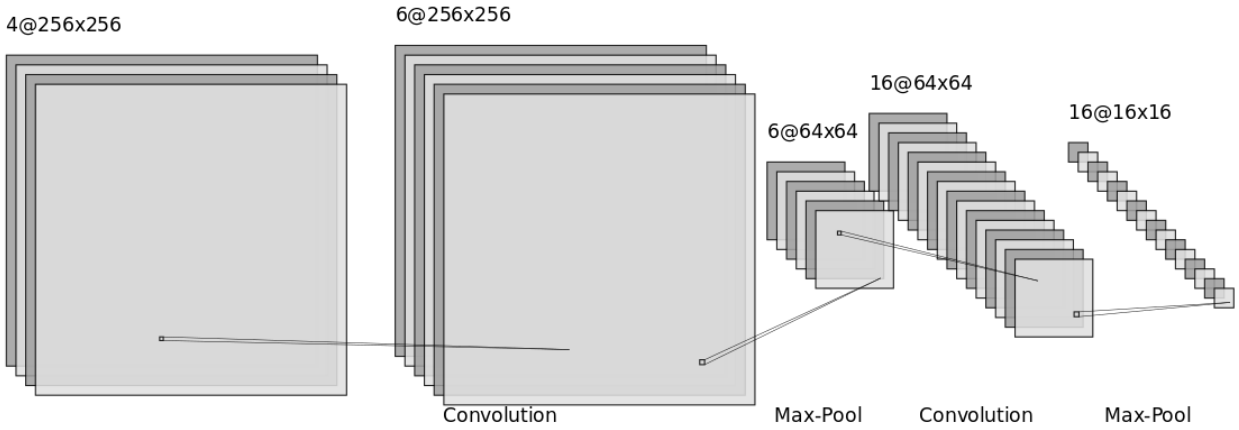
γ) Πόσους πυρήνες θα χρησιμοποιεί το δεύτερο συνελικτικό στρώμα και τι σχήμα θα έχει ο καθένας;

Το δεύτερο συνελικτικό στρώμα θα χρησιμοποιεί 16 πυρήνες, ώστε να προκύπτουν 16 χάρτες χαρακτηριστικών, όπως φαίνεται στο σχήμα της διαφάνειας 23. Ο κάθε πυρήνας θα έχει 6 φέτες (slices), αφού η είσοδος του συνελικτικού στρώματος (η έξοδος του πρώτου στρώματος max-pooling) έχει 6 κανάλια (χάρτες). Γνωρίζουμε από την εκφώνηση ότι κάθε πυρήνας εφαρμόζει σε κάθε κανάλι της εισόδου του παράθυρο 3x3. Επομένως κάθε ένας από τους 16 πυρήνες θα είναι ένας τανυστής (tensor) τριών αξόνων, με σχήμα (shape) (3, 3, 6).



δ) Τι ανάλυση θα έχουν οι χάρτες χαρακτηριστικών που θα προκύπτουν από το δεύτερο στρώμα max-pooling;

Αφού τα συνελικτικά στρώματα που χρησιμοποιούμε διατηρούν την ανάλυση σε κάθε κανάλι, κάθε ένας από τους 16 χάρτες χαρακτηριστικών (κανάλια) που προκύπτουν από το δεύτερο συνελικτικό στρώμα, δηλαδή κάθε κανάλι στην είσοδο του δεύτερου στρώματος max-pooling θα εξακολουθεί να έχει ανάλυση 64x64 (όπως στην έξοδο του πρώτου στρώματος max-pooling). Αφού κάθε στρώμα max-pooling χρησιμοποιεί παράθυρο 4x4 με βήμα (stride) 4 και στους δύο άξονες, ο κάθε ένας από τους 16 χάρτες που εξέρχονται από το δεύτερο στρώμα max-pooling θα έχει ανάλυση (64/4)x(64/4), δηλαδή 16x16.

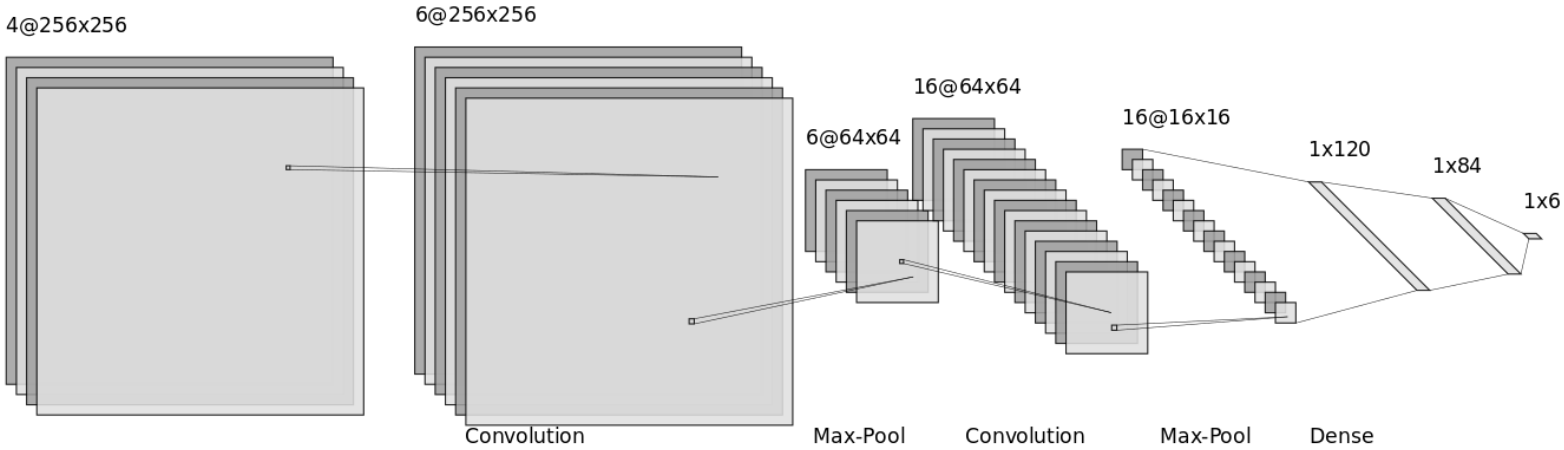


ε) Πόσους νευρώνες θα έχει η είσοδος του τελικού MLP;

Οι 16 χάρτες ανάλυσης 16x16 που εξέρχονται από το δεύτερο στρώμα max- pooling θα συνενώνονται σε ένα διάνυσμα 16x16x16 = 4096 χαρακτηριστικών, που θα δίνεται ως είσοδος στο τελικό MLP (τρία πυκνά στρώματα) του σχήματος της διαφάνειας 23.

στ) Πόσους νευρώνες θα έχει το τελικό στρώμα εξόδου του MLP; Τι συνάρτηση ενεργοποίησης θα έχουν;

Το στρώμα εξόδου του MLP θα έχει 6 νευρώνες, δύο για τις συντεταγμένες (x, y) του κεφαλιού και τέσσερις για τις συντεταγμένες των δύο ώμων. Οι νευρώνες αυτοί δεν θα έχουν συνάρτηση ενεργοποίησης, ώστε να μπορούν να παράγουν οποιονδήποτε πραγματικό αριθμό ο καθένας.



CNN tutorial: https://www.analyticsvidhya.com/blog/2018/12/guide-convolutional-neural-network-cnn/

Άσκηση Β7.2.

Μια εταιρεία κατασκευής οικιακών συσκευών ετοιμάζει έναν νέο τύπο (μοντέλο) φούρνου μικροκυμάτων που θα διαθέτει κάμερα. Η εταιρεία θέλει ο φούρνος να έχει τη δυνατότητα να αναγνωρίζει μέσω της κάμερας τον χρήστη που στέκεται μπροστά του, ώστε να προσαρμόζονται οι ρυθμίσεις του φούρνου στις προτιμήσεις του συγκεκριμένου χρήστη. Η εταιρεία σχεδιάζει να χρησιμοποιήσει ένα συνελικτικό νευρωνικό δίκτυο (CNN), το οποίο θα τροφοδοτείται με μια φωτογραφία του χρήστη που στέκεται μπροστά στη συσκευή. Το CNN θα έχει 10 νευρώνες εξόδου, γιατί η εταιρεία θεωρεί ότι κάθε συσκευή του συγκεκριμένου τύπου θα χρησιμοποιείται σε ένα σπίτι ή γραφείο όπου οι χρήστες θα είναι το πολύ δέκα. Η εταιρεία διαθέτει 1.000 φωτογραφίες 50 ενδεικτικών χρηστών (20 από κάθε ενδεικτικό χρήστη) που έχουν τραβηχτεί με την κάμερα του νέου φούρνου. Κάθε μία από τις 1.000 φωτογραφίες είναι επισημειωμένη με τον κωδικό (id, 1–50) του αντίστοιχου ενδεικτικού χρήστη. Αλλά η εταιρεία δεν διαθέτει εκ των προτέρων φωτογραφίες όλων των χρηστών (σε κάθε σπίτι, γραφείο) που θα χρησιμοποιήσουν την κάθε μία συσκευή του συγκεκριμένου νέου τύπου. Όταν μία συσκευή του συγκεκριμένου τύπου εγκαθίσταται σε ένα σπίτι ή γραφείο, θα ζητείται από κάθε έναν από τους (το πολύ 10) χρήστες της να τραβήξει 5-10 φωτογραφίες του με την κάμερα της συσκευής, χρησιμοποιώντας ειδική επιλογή της διεπαφής χρήστη. Εξηγήστε πώς θα μπορούσε η εταιρεία να χρησιμοποιήσει τις 1.000 φωτογραφίες ενδεικτικών χρηστών που διαθέτει, καθώς και μια γενική συλλογή εκατομμυρίων επισημειωμένων εικόνων (π.χ. εικόνες ζώων, τοπίων κ.λπ., όπως στο ImageNet), ώστε να προ-εκπαιδεύσει (από το εργοστάσιο) το CNN του νέου τύπου φούρνου και να καταφέρει η κάθε συσκευή του νέου τύπου να αναγνωρίζει (με ελάχιστη πρόσθετη εκπαίδευση) τους συγκεκριμένους χρήστες της (σε συγκεκριμένο σπίτι ή γραφείο) έχοντας στη διάθεσή της μόνο 5-10 φωτογραφίες του καθενός.

Απάντηση:

Η εταιρεία θα μπορούσε να χρησιμοποιήσει έναν κωδικοποιητή **CNN προ-εκπαιδευμένο στη συλλογή των εκατομμυρίων επισημειωμένων εικόνων** (π.χ. προ-εκπαιδευμένο στο ImageNet).

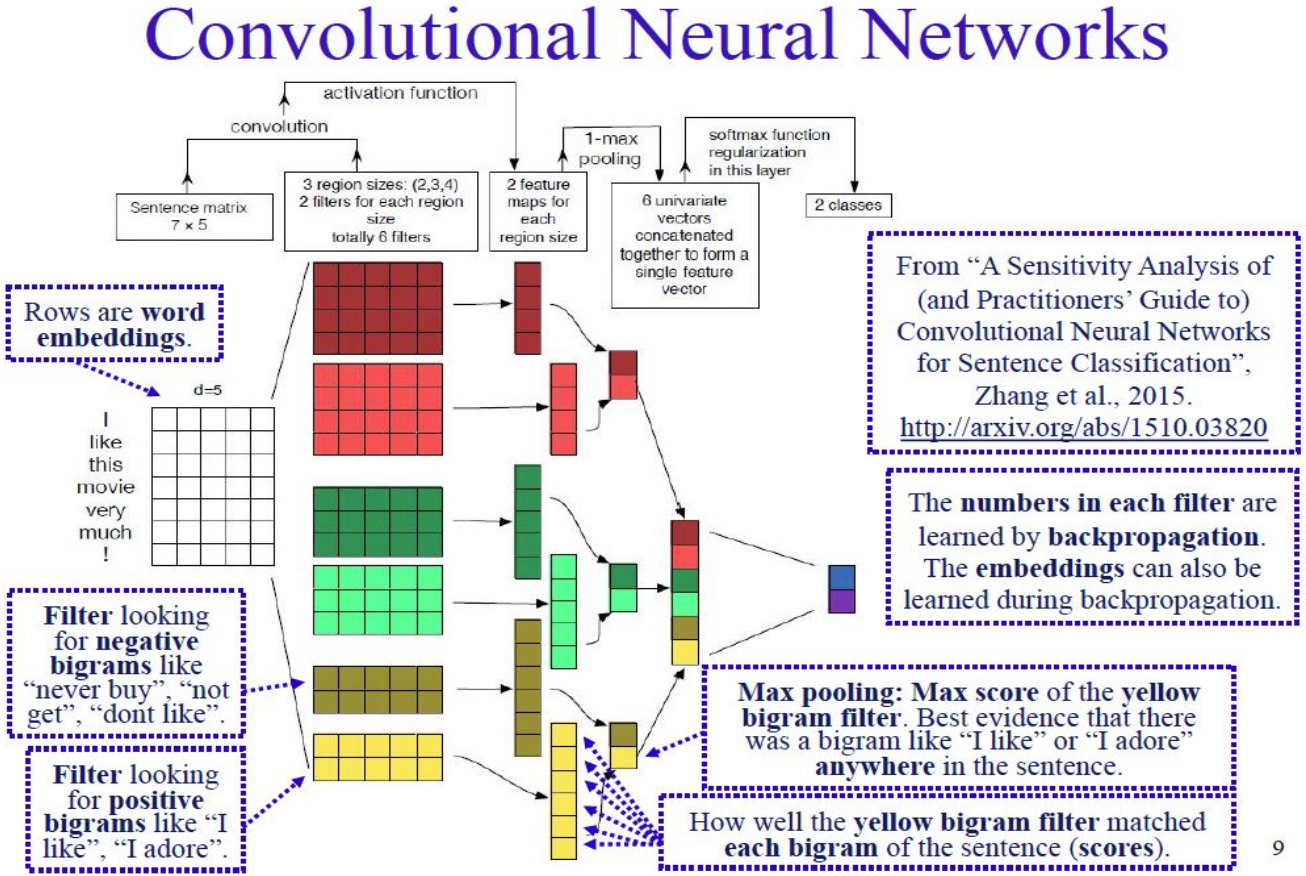
Από τον προ-εκπαιδευμένο κωδικοποιητή, θα κρατούσε **μόνο τα συνελικτικά επίπεδα** (και τα επίπεδα υπο-δειγματοληψίας), όπως στις διαφάνειες 25–26. Πάνω από αυτά θα **πρόσθετε ένα MLP με 50 νευρώνες εξόδου** (έναν νευρώνα εξόδου για κάθε χρήστη του συνόλου των 1.000 φωτογραφιών ενδεικτικών χρηστών, με softmax συνάρτηση ενεργοποίησης στο επίπεδο εξόδου). Θα **εκπαίδευε (fine-tuning)** το συνολικό σύστημα **στις 1.000 φωτογραφίες** ενδεικτικών χρηστών, εφαρμόζοντας και **επαύξηση δεδομένων** (data augmentation, διαφάνεια 27), **ξεπαγώνοντας σταδιακά τα τελευταία συνελικτικά επίπεδα** (όπως στη διαφάνεια 26), ώστε να προσαρμοστούν στο πρόβλημα της αναγνώρισης προσώπων.

Κατόπιν θα αντικαθιστούσε το MLP με ένα **νέο MLP με 10 μόνο νευρώνες εξόδου** (έναν για κάθε πιθανό χρήστη ενός συγκεκριμένου σπιτιού ή γραφείου, πάλι με softmax στο επίπεδο εξόδου), χωρίς να εκπαιδεύσει το νέο MLP.

Κατά την εγκατάσταση του φούρνου σε ένα νέο σπίτι ή γραφείο, το σύστημα με το νέο MLP (και τα συνελικτικά επίπεδα και επίπεδα δειγματοληψίας) θα **εκπαιδευόταν (πρόσθετο fine-tuning) με τις φωτογραφίες των χρηστών του συγκεκριμένου σπιτιού** ή γραφείου (5–10 φωτογραφίες για τον καθένα), εφαρμόζοντας **πάλι και επαύξηση δεδομένων**. Στο τελευταίο αυτό στάδιο εκπαίδευσης, ενδέχεται να ήταν προτιμότερο να κρατηθούν **παγωμένα (αμετάβλητα) τα συνελικτικά επίπεδα**, λόγω των σχετικά λίγων δεδομένων (φωτογραφιών) εκπαίδευσης που θα είχαμε ανά σπίτι ή γραφείο. Θα μπορούσε, όμως, η εταιρεία να διερευνήσει και την περίπτωση να ξεπαγώνει πάλι τα τελευταία συνελικτικά επίπεδα.

Άσκηση B8.1.

Γράψτε (όπως στις διαφάνειες 10–14) τις εξισώσεις του CNN της διαφάνειας 9.
Προσδιορίστε επίσης τις διαστάσεις όλων των εμπλεκομένων πινάκων και διανυσμάτων.



Απάντηση: The dimensionality of the word embeddings is $d = 5$. We can think of the two bigram filters as a matrix $W^{(2)} \in \mathbb{R}^{2\times 2d} = \mathbb{R}^{2\times 10}$ and a bias terms vector $b^{(2)} = \mathbb{R}^2$ (similarly to slide 12, where we have three bigram filters). Similarly, we can think of the two trigram filters as a matrix $W^{(3)} \in \mathbb{R}^{2\times 3d} = \mathbb{R}^{2\times 15}$ and a bias terms vector $b^{(3)} = \mathbb{R}^2$; and the two 4-gram filters as a matrix $W^{(4)} \in \mathbb{R}^{2\times 4d} = \mathbb{R}^{2\times 20}$ and a bias terms vector $b^{(4)} = \mathbb{R}^2$.

The embeddings of each bigram of the input text can be thought of as a vector $x^{(2)} \in \mathbb{R}^{2d}$. Applying the two bigram filters to the i-th bigram $x_i^{(2)}$ of the input text produces:

$$h_i^{(2)} = \text{ReLU}\left(W^{(2)}x_i^{(2)} + b^{(2)}\right) \in \mathbb{R}^2, \qquad i = 1, \dots, 6$$

where we assumed that we use 'narrow convolutions', i.e., that the filters do not move out of the words of the input text (to partially overlap with padding tokens).

Max-pooling over $h_1^{(2)}, \dots, h_6^{(2)}$ produces a vector:

$$h^{(2)} = \langle \max_i h_{i,1}^{(2)}, \max_i h_{i,2}^{(2)} \rangle^T \in \mathbb{R}^2$$

ΔΙΑΣΤΑΣΗ $h^{(2)}$

- Input: n (number of words=7)
- Padding: p (=0)
- Stride: s (=1)
- Filter size: f (bi-gram=2)
- Output: [(n+2p-f)/s+1]

Similarly, applying the two trigram filters to the i-th trigram $x_i^{(3)} \in \mathbb{R}^{3d}$ of the input text and the two 4-gram filters to the i-th 4-gram $x_i^{(4)} \in \mathbb{R}^{4d}$ produces:

$$h_i^{(3)} = \text{ReLU}\left(W^{(3)} x_i^{(3)} + b^{(3)}\right) \in \mathbb{R}^2, \qquad i = 1, \dots, 5$$
$$h_i^{(4)} = \text{ReLU}\left(W^{(4)} x_i^{(4)} + b^{(4)}\right) \in \mathbb{R}^2, \qquad i = 1, \dots, 4$$

Max-pooling over $h_1^{(3)}, \dots, h_5^{(3)}$ and over $h_1^{(4)}, \dots, h_4^{(4)}$ produces:

$$h^{(3)} = \langle \max_i h_{i,1}^{(3)}, \max_i h_{i,2}^{(3)} \rangle^T \in \mathbb{R}^2$$
$$h^{(4)} = \langle \max_i h_{i,1}^{(4)}, \max_i h_{i,2}^{(4)} \rangle^T \in \mathbb{R}^2$$

The feature vector of the input text is the concatenation $h = \left[h^{(2)}; h^{(3)}; h^{(4)}\right]^T \in \mathbb{R}^6$.

We pass on h to a classifier, e.g., a logistic regression layer, i.e., a dense layer $W^{(P)} \in \mathbb{R}^{|C| \times 6}$ with a bias vector $b^{(P)} \in \mathbb{R}^{|C|}$ and a softmax activation function, to obtain a probability distribution $\vec{o}$ over the classes $c_1, \dots, c_{|C|} \in C$:

$$\vec{o} = \langle P(c_1), \dots, P(c_{|C|}) \rangle^{\mathrm{T}} = \text{softmax}(W^{(P)} h + b^{(P)})$$

ΔΙΑΣΤΑΣΗ $h^{(3)}$

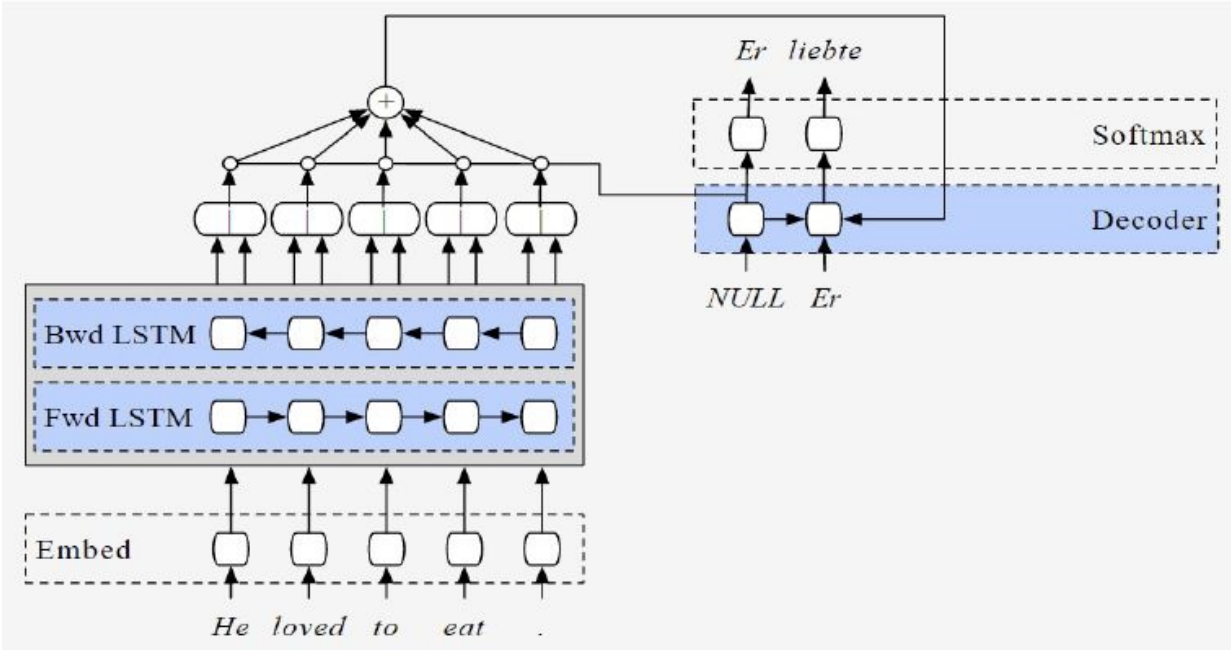
- Input: n (number of words=7)
- Padding: p (=0)
- Stride: s (=1)
- Filter size: f (tri-gram=3)
- Output: [(n+2p-f)/s+1]

ΔΙΑΣΤΑΣΗ $h^{(4)}$

- Input: n (number of words=7)
- Padding: p (=0)
- Stride: s (=1)
- Filter size: f (4-gram=4)
- Output: [(n+2p-f)/s+1]

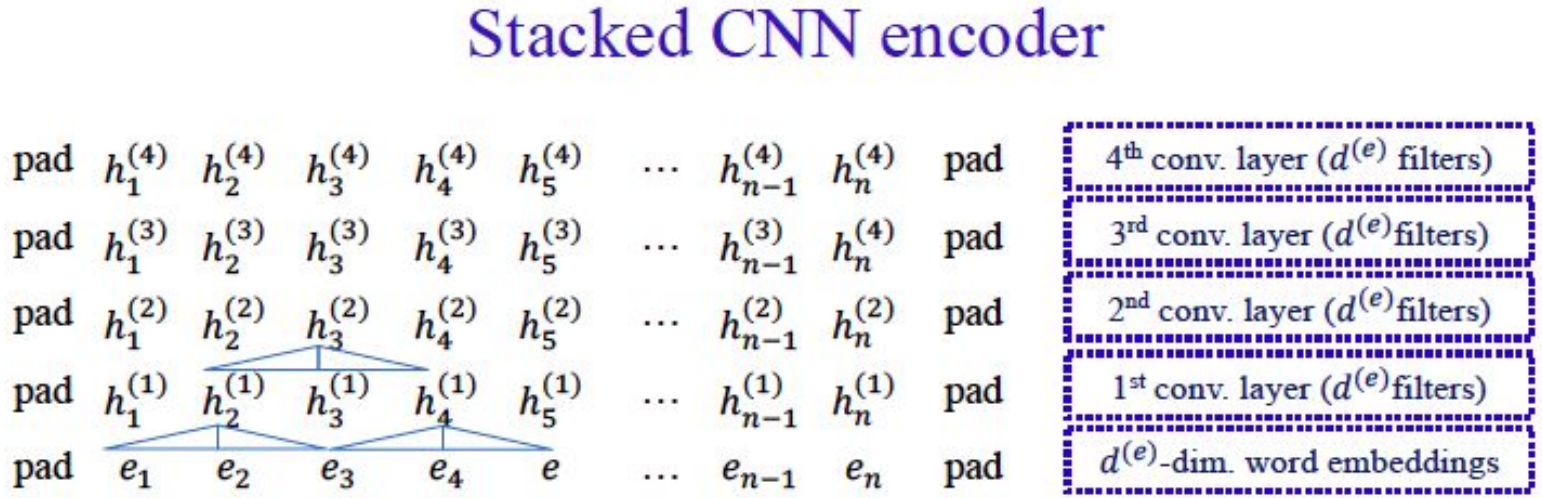
Άσκηση B8.2.

Consider the following LSTM-based machine translation model (see also exercise 4 of section B6).



Άσκηση B8.2.

We wish to replace the BiLSTM encoder of the model above by the stacked CNN-based encoder with trigram filters illustrated below, retaining the encoder-decoder attention and the LSTM decoder of the original model.



Άσκηση B8.2.

Let V, V' be the vocabularies of the source language (English) and target language (German), respectively. Each training instance is a pair consisting of (i) a sequence of one-hot vectors:

$$x_1, x_2, x_3, \dots, x_n \in \{0,1\}^{|V|}$$

corresponding to an English sentence (each vector shows the position of the corresponding word in V) and (ii) a sequence of one-hot vectors:

$$y_1, y_2, y_3, \dots, y_m \in \{0,1\}^{|V'|}$$

corresponding to a German sentence that is the correct (gold) translation of the English one (each vector shows the position of the corresponding word in V'). For simplicity, we assume all the English sentences are n words long, and all the German sentences are m words long.

Let $E \in \mathbb{R}^{d^{(e)} \times |V|}$ and $E' \in \mathbb{R}^{d^{(e)} \times |V'|}$ contain the word embeddings of the source and target language, respectively. Notice that word embeddings have $d^{(e)}$ dimensions in both languages, and that all the convolution layers of the CNN encoder also use $d^{(e)}$ filters.

The following formulae describe how the new model works and how the loss (L) is computed, given a training instance. **Fill in the blanks (they have been filled in in red in the solution).** The notation $[\dots;\dots]$ denotes concatenation and f, g denote activation functions.

Απάντηση:

Encoder: ($i \in \{1, 2, 3, \ldots, n\}$, $l \in \{2, 3, 4\}$)

$e_i = E\, x_i \in \mathbb{R}^{d^{(e)}}$ (Το embedding της σωστής αγγλικής λέξης στη θέση i.)

(Assume that $e_0 = e_{n+1}$ is always an all-zeros embedding of the padding token.)

$$h_i^{(1)} = \mathrm{ReLU}\left(W^{(1)}[e_{i-1}; e_i; e_{i+1}] + b^{(1)}\right) + e_i \in \mathbb{R}^{d^{(e)}}$$

$$W^{(1)} \in \mathbb{R}^{d^{(e)} \times 3 \cdot d^{(e)}}$$
$$b^{(1)} \in \mathbb{R}^{d^{(e)}}$$

$$h_i^{(l)} = \mathrm{ReLU}\left(W^{(j)}\left[h_{i-1}^{(l-1)}; h_i^{(l-1)}; h_{i+1}^{(l-1)}\right] + b^{(l)}\right) + h_i^{(l-1)} \in \mathbb{R}^{d^{(e)}}$$

$$W^{(j)} \in \mathbb{R}^{d^{(e)} \times 3 \cdot d^{(e)}}$$
$$b^{(j)} \in \mathbb{R}^{d^{(e)}}$$

Απάντηση:

Decoder: $(i \in \{1, 2, 3, \ldots, \mathrm{n}\},\ j \in \{1, 2, 3, \ldots, m\})$

$t_j = E' y_j \in \mathbb{R}^{d^{(e)}}$ (Το embedding της σωστής γερμανικής λέξης στη θέση j.)

$z_j = \mathrm{LSTM}(z_{j-1}, [t_{j-1}; c_j]) \in \mathbb{R}^{d^{(e)}}$ $\quad z_0 \in \mathbb{R}^{d^{(e)}}, t_0 \in \mathbb{R}^{d^{(e)}}$

$\tilde{a}_{i,j} = v^T \cdot f(W^{(a)} [h_i^{(4)}; z_{j-1}] + b^{(a)}) \in \mathbb{R}$ $\quad W^{(a)} \in \mathbb{R}^{d^{(a)} \times 2 \cdot d^{(e)}}$

$b^{(a)} \in \mathbb{R}^{d^{(a)}}, v \in \mathbb{R}^{d^{(a)}}$

$a_{i,j} = \dfrac{\exp(\tilde{a}_{i,j})}{\sum_{i'} \exp(\tilde{a}_{i',j})}$

$c_j = g(\sum_i a_{i,j} h_i^{(4)} + b^{(c)}) \in \mathbb{R}^{d^{(e)}}$ $\quad b^{(c)} \in \mathbb{R}^{d^{(e)}}$

Απάντηση:

$$\tilde{o}_j = W^{(o)} z_j + b^{(o)} \in \mathbb{R}^{|V'|}$$

$$W^{(o)} \in \mathbb{R}^{|V'| \times d^{(e)}}$$
$$b^{(o)} \in \mathbb{R}^{|V'|}$$

$$o_{j,k} = \frac{\exp(\tilde{o}_{j,k})}{\sum_{k=1}^{|V'|} \exp(\tilde{o}_{j,k})}$$

(Πόσο πιθανό θεωρεί το μοντέλο η k-στή λέξη του γερμανικού λεξιλογίου να είναι η σωστή για την j-στή θέση της μετάφρασης.)

$$r_j = \text{argmax}_l \; y_{j,l}$$

(Σύμφωνα με το 1-hot y_j, η σωστή λέξη στην j-στή θέση της μετάφρασης βρίσκεται στη θέση r_j του γερμανικού λεξιλογίου.)

$$L = -\sum_j \log o_{j,r_j}$$

(Ελαχιστοποιώντας το L, μεγιστοποιούμε την πιθανότητα που δίνει το μοντέλο στις σωστές λέξεις, σε όλες τις θέσεις της μετάφρασης.)