Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής
Μάθημα: Αλληλεπίδραση Ανθρώπου – Υπολογιστή, 2023–24
Διδάσκων: Ι. Ανδρουτσόπουλος

Ασκήσεις μελέτης της ενότητας Β6 (επεξεργασία φυσικής γλώσσας με ανατροφοδοτούμενα νευρωνικά δίκτυα)

1. Θέλουμε να χρησιμοποιήσουμε το ανατροφοδοτούμενο νευρωνικό δίκτυο (RNN) της διαφάνειας 7, για να αναγνωρίζουμε ονόματα προσώπων, οργανισμών και τοποθεσιών. Χρησιμοποιούμε ετικέτες (κατηγορίες) B-I-O, π.χ. B-Person για την πρώτη λέξη ονόματος προσώπου, I-Person για τις άλλες λέξεις ονόματος προσώπου, B-Org, I-Org, …, O για όλες τις υπόλοιπες λέξεις, άρα 7 κατηγορίες. Το μέγεθος του λεξιλογίου είναι $|V| = 100.000$. Κάθε ενσωμάτωση λέξης (word embedding) είναι ένα διάνυσμα 300 διαστάσεων. Το κρυφό επίπεδο (η κατάσταση του RNN) αποτελείται από 500 νευρώνες, δηλαδή το $\vec{h}_i$ είναι διάνυσμα 500×1. Ποιες είναι οι διαστάσεις των $E, \vec{e}_i, W^{(h)}, W^{(e)}, W^{(o)}, \vec{o}_i$; Αιτιολογήστε τις απαντήσεις σας.

Απάντηση: Ο πίνακας E περιέχει (ως στήλες) τις ενσωματώσεις των 100.000 λέξεων του λεξιλογίου. Κάθε ενσωμάτωση είναι διάνυσμα (στήλη) 300 διαστάσεων. Άρα ο E έχει διαστάσεις 300×100.000.

Το διάνυσμα $\vec{e}_i$ είναι η ενσωμάτωση (embedding) της i-στής λέξης της εισόδου (π.χ. μιας πρότασης), άρα είναι διαστάσεων 300×1. Το ίδιο συμπέρασμα προκύπτει και από την παρατήρηση ότι ο πολλαπλασιασμός $E\vec{x}_i$ επιστρέφει την i-στή στήλη του πίνακα E.

Ο πίνακας $W^{(h)}$ έχει διαστάσεις 500×500, ενώ ο πίνακας $W^{(e)}$ έχει διαστάσεις 500×300, ώστε τα $W^{(h)}\vec{h}_{i-1}$ και $W^{(e)}\vec{e}_i$ να έχουν τις ίδιες διαστάσεις (500×1), να μπορούν να προστεθούν $(W^{(h)}\vec{h}_{i-1} + W^{(e)}\vec{e}_i)$ και η νέα κατάσταση $\vec{h}_i = \tanh\left(W^{(h)}\vec{h}_{i-1} + W^{(e)}\vec{e}_i\right)$ να έχει πάλι διαστάσεις 500×1, όπως η προηγούμενη κατάσταση $\vec{h}_{i-1}$. Η tanh εφαρμόζεται σε κάθε στοιχείο του διανύσματος $W^{(h)}\vec{h}_{i-1} + W^{(e)}\vec{e}_i$, χωρίς να αλλάζει τις διαστάσεις του.

Ο πίνακας $W^{(o)}$ έχει διαστάσεις 7×500, ώστε ο πολλαπλασιασμός $W^{(0)}\vec{h}_i$ να παράγει διάνυσμα 7×1 με έναν πραγματικό αριθμό για κάθε κατηγορία. Η softmax στον υπολογισμό $\vec{o}_i = \text{softmax}\left(W^{(o)}\vec{h}_i\right)$ μετατρέπει τους αριθμούς αυτούς σε κατανομή πιθανότητας (μία πιθανότητα για κάθε κατηγορία), χωρίς να αλλάζει τις διαστάσεις του $W^{(o)}\vec{h}_i$. Επομένως το $\vec{o}_i$ έχει και αυτό διαστάσεις 7×1.

2. Write down the equations for a modified version of the "RNN with deep self-attention" (slides 18–20), where the uni-directional RNN with GRU cells is replaced by a stacked bi-directional RNN with GRU cells. Use the notation $\text{GRU}(h_{t-1}, \tau_t)$ to denote the new state of a GRU cell with previous state h_{t-1} and input τ_t.

Answer: At the first layer of the GRU RNN, we have (for $t = 1, \ldots, k$):

$$\vec{h}_t^{(1)} = \text{GRU}\left(\vec{h}_{t-1}^{(1)}, x_t\right)$$

$$\overleftarrow{h}_t^{(1)} = \text{GRU}\left(\overleftarrow{h}_{t+1}^{(1)}, x_t\right)$$

$$h_t^{(1)} = [\vec{h}_t^{(1)}; \overleftarrow{h}_t^{(1)}]$$

where $\vec{h}_0^{(1)}$ is the initial state of the left-to-right GRU RNN of the first layer, $\overleftarrow{h}_{k+1}^{(1)}$ is the initial state of the right-to-left GRU RNN of the first layer, ";" denotes concatenation, and $x_1, \ldots, x_k$ are the word embeddings of the input word sequence.

Similarly, at the m-th layer of the GRU RNN:

$$\vec{h}_t^{(m)} = \text{GRU}\left(\vec{h}_{t-1}^{(m)}, h_t^{(m-1)}\right)$$

$$\overleftarrow{h}_t^{(m)} = \text{GRU}\left(\overleftarrow{h}_{t+1}^{(m)}, h_t^{(m-1)}\right)$$

$$h_t^{(m)} = [\vec{h}_t^{(m)}; \overleftarrow{h}_t^{(m)}]$$

The other equations remain as on slide 20.

3. (a) Modify the equations of the neural network of the previous exercise to support *multi-label classification*, i.e., cases where the same text (e.g., tweet) may belong in multiple classes (labels). As a twist, use a separate *label-specific self-attention-head* for each class, which will produce a different distribution of attention scores $a_{c,1}, \ldots, a_{c,k}$ (where k is again the length of the input text, counted in words) and a different $h_{sum,c}$ for each class c. Feed the $h_{sum,c}$ of each class c to a separate (different per class) dense layer with a sigmoid to produce the probability that the input text should be assigned class c.

Answer: Let C be the set of possible classes (labels). We modify the self-attention MLP of slide 20, so that $a_t^{(l)} \in \mathbb{R}^{|C|}$, i.e., $a_t^{(l)}$ is now a vector (not a scalar) containing $|C|$ attention scores $a_{1,t}, \ldots, a_{|C|,t}$ for word position t, one for each possible class. To achieve this, we modify the dimensions of $W^{(l)}$ and $b^{(l)}$ of layer l (the last layer) of the self-attention MLP, to be $|C| \times d$ and $|C|$, respectively, where d is the dimensionality of the previous layer $a_t^{(l-1)}$.

The softmax of slide 20 is now applied *label-wise*, on the attention scores of a particular class, i.e., for each possible class c:

$$a_{c,t} = \text{softmax}\left(a_{c,t}^{(l)}; a_{c,1}^{(l)}, \ldots, a_{c,k}^{(l)}\right) = \frac{\exp(a_{c,t}^{(l)})}{\sum_{t'=1}^{k} \exp(a_{c,t'}^{(l)})}$$

We form a separate weighted sum $h_{sum,c}$ for each possible class c:

$$h_{sum,c} = \sum_{t=1}^{k} a_{c,t} h_t^{(M)}$$

where M is the number of stacked GRU RNNs of the previous exercise, and we feed each $h_{sum,c}$ to a separate dense layer $W_{p,c}$ (with bias term $b_{p,c}$) per class c, to compute the probability of the corresponding class:

$$P(c|x_1, \ldots, x_k) = \sigma(W_{p,c} h_{sum,c} + b_{p,c})$$

If $h_{sum,c} \in \mathbb{R}^d$, then $W_{p,c} \in \mathbb{R}^{1 \times d}$ and $b_{p,c} \in \mathbb{R}$. The other equations of the neural network remain as in the previous exercise.

(b) Couldn't we use a single (shared) self-attention-head (and a single h_{sum}) for all the classes? What would change in that case in the equations above? What is the advantage of using a separate *label-specific* self-attention-head for each class?

4. Εκπαιδεύουμε το παρακάτω νευρωνικό μοντέλο μηχανικής μετάφρασης.

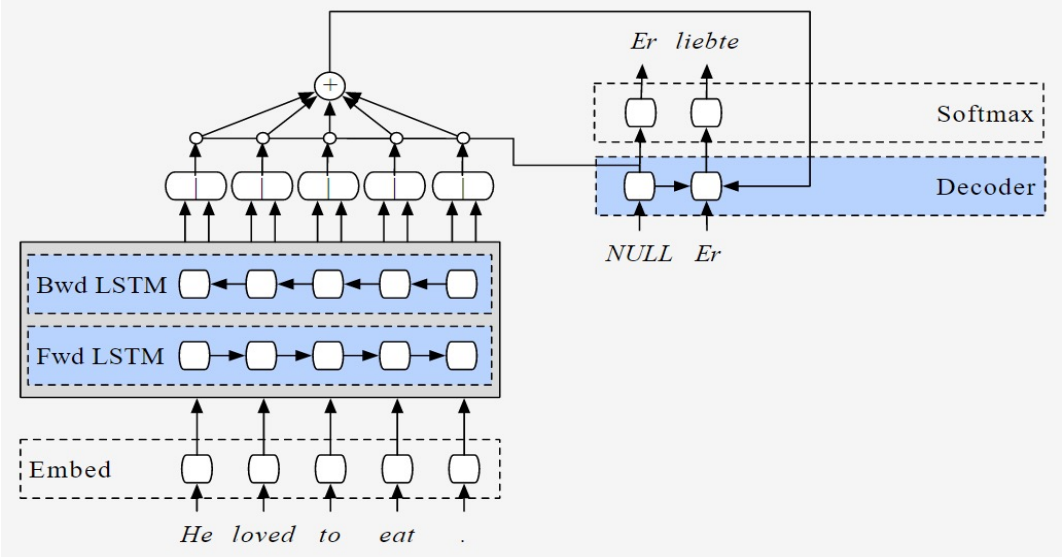

Image from Stephen Merity's http://smerity.com/articles/2016/google_nmt_arch.html

Έστω V και V' τα λεξιλόγια της γλώσσας-πηγής (Αγγλικά) και της γλώσσας-στόχου (Γερμανικά) αντίστοιχα. Κάθε παράδειγμα εκπαίδευσης είναι ένα ζεύγος αποτελούμενο από μια ακολουθία one-hot διανυσμάτων:

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

που αντιστοιχούν σε μια αγγλική πρόταση (κάθε διάνυσμα δείχνει σε ποια θέση του αγγλικού λεξιλογίου V βρίσκεται η αντίστοιχη λέξη) και μια ακολουθία one-hot διανυσμάτων:

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

που αντιστοιχούν σε μια γερμανική πρόταση που είναι η σωστή (gold) μετάφραση της αγγλικής (κάθε διάνυσμα δείχνει σε ποια θέση του γερμανικού λεξικού V' βρίσκεται η αντίστοιχη λέξη).

Έστω $E \in \mathbb{R}^{d^{(e)} \times |V|}$ και $E' \in \mathbb{R}^{d^{(e)} \times |V'|}$ οι πίνακες με τα word embeddings (το καθένα $d^{(e)}$ διαστάσεων) των δύο γλωσσών αντίστοιχα.

Οι παρακάτω τύποι περιγράφουν αναλυτικά τη λειτουργία του μοντέλου και τον υπολογισμό του σφάλματος (L) για ένα παράδειγμα εκπαίδευσης. **Συμπληρώστε τα κενά (στη λύση έχουν συμπληρωθεί με κόκκινο).** Ο συμβολισμός $[\dots;\dots]$ παριστάνει συνένωση (concatenation). Τα f και g παριστάνουν συναρτήσεις ενεργοποίησης.

Κωδικοποιητής: $(i \in \{1, 2, 3, \dots, \mathrm{n}\})$

$e_i = E\, x_i \in \mathbb{R}^{d^{(e)}}$

$\vec{h}_i = \mathrm{LSTM}(\vec{h}_{i-1}, e_i) \in \mathbb{R}^{d^{(h)}}$ $\qquad$ $\vec{h}_0 \in \mathbb{R}^{d^{(h)}}$

$\overleftarrow{h}_i = \text{LSTM}(\overleftarrow{h}_{i+1}, e_i) \in \mathbb{R}^{d^{(h)}}$ $\quad\quad$ $\overleftarrow{h}_{n+1} \in \mathbb{R}^{d^{(h)}}$

$h_i = [\vec{h}_i; \overleftarrow{h}_i] \in \mathbb{R}^{2 \cdot d^{(h)}}$

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

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

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

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

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

$c_j = W^{(c)} \cdot g(\sum_i a_{i,j} h_i + b^{(c)}) \in \mathbb{R}^{d^{(e)}}$ $\quad\quad$ $W^{(c)} \in \mathbb{R}^{d^{(e)} \times 2 \cdot d^{(h)}}$
$b^{(c)} \in \mathbb{R}^{2 \cdot d^{(h)}}$

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

$o_{j,k} = \dfrac{\exp(\tilde{o}_{j,k})}{\sum_{k'=1}^{|V'|} \exp(\tilde{o}_{j,k'})}$ $\quad\quad$ (Πόσο πιθανό θεωρεί το μοντέλο η k-στή λέξη του γερμανικού λεξιλογίου να είναι η σωστή για την j-στή θέση της μετάφρασης.)

$r_j = \text{argmax}_l\ y_{j,l}$ $\quad\quad$ (Σύμφωνα με το 1-hot y_j, η σωστή λέξη στην j-στή θέση της μετάφρασης βρίσκεται στη θέση r_j του γερμανικού λεξιλογίου.)

$L = -\sum_j \log o_{j,r_j}$ $\quad\quad$ (Ελαχιστοποιώντας το L, μεγιστοποιούμε την πιθανότητα που δίνει το μοντέλο στις σωστές λέξεις, σε όλες τις θέσεις της μετάφρασης.)