



Ασκήσεις μελέτης B10

Lab 11

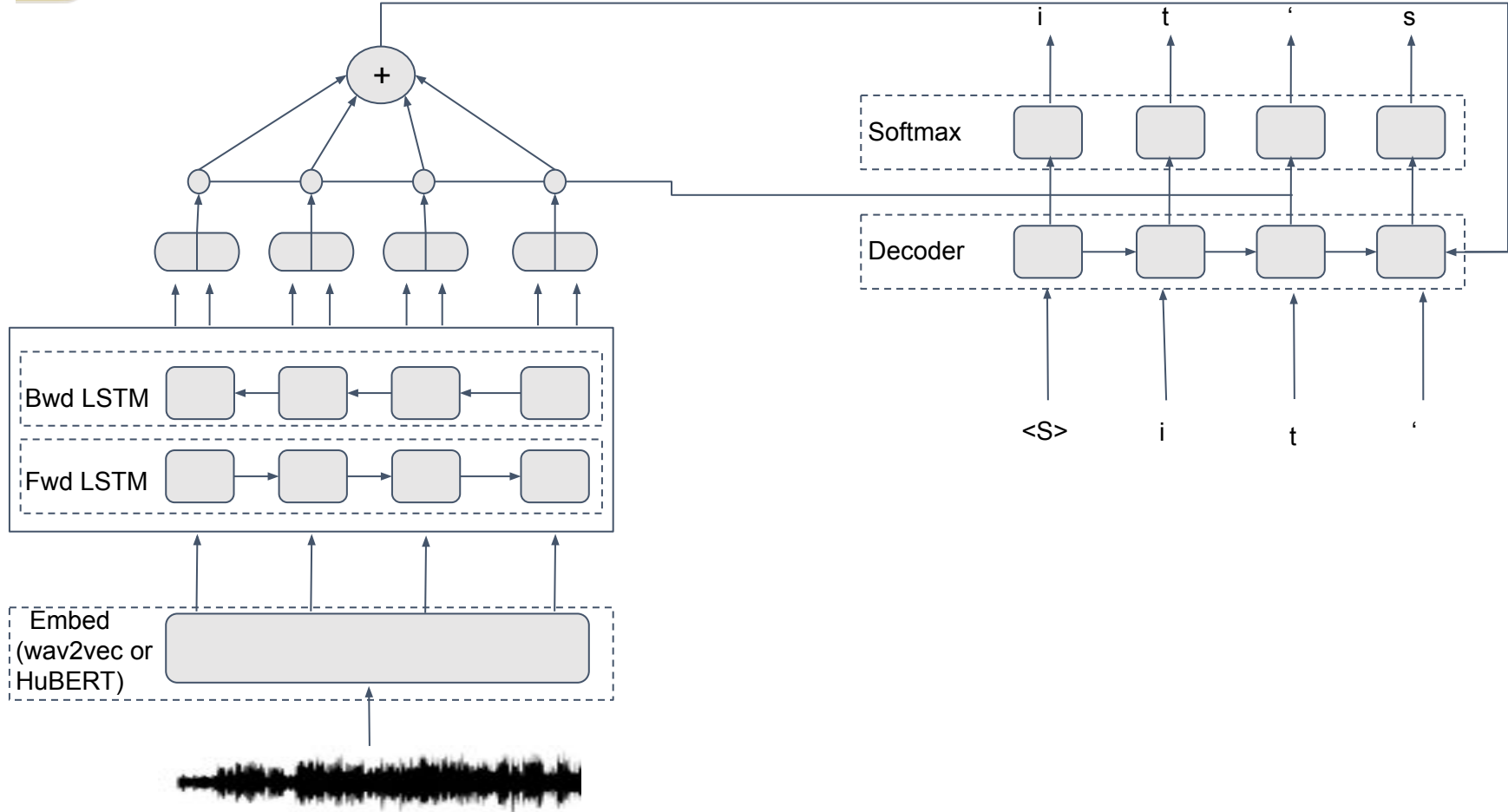
Human-Computer Interaction, AUEB
Εαρινό εξάμηνο 2022-2023

Lab Assistant: Sofia Eleftheriou



Άσκηση B10.1

Προσαρμόστε τους τύπους της άσκησης 4 της ενότητας B6 (μηχανική μετάφραση με κωδικοποιητή BiLSTM και αποκωδικοποιητή LSTM) για την περίπτωση όπου στην είσοδο του κωδικοποιητή έχουμε διανύσματα που έχουν δημιουργηθεί με wav2vec ή HuBERT, ένα για κάθε τμήμα (frame) ήχου (ή λιγότερα, αν κάνουμε υπο-δειγματοληψία), αντί για ενθέσεις λέξεων (word embeddings) και σε κάθε βήμα του αποκωδικοποιητή παράγεται ένα γράμμα, όπως στις διαφάνειες 18–20.





Έστω A το αλφάβητο της γλώσσας προς αναγνώριση (π.χ. Αγγλικά).² Κάθε παράδειγμα εκπαίδευσης είναι ένα ζεύγος αποτελούμενο από (i) μια ακολουθία διανυσμάτων που έχει δημιουργήσει το wav2vec ή το HuBERT (ένα διάνυσμα $d^{(h)}$ διαστάσεων ανά τμήμα ήχου) από κάποιο σήμα ομιλίας (π.χ. από την εκφώνηση μιας πρότασης):³

$$e_1, e_2, e_3, \dots, e_n \in \mathbb{R}^{d^{(h)}}$$

και (ii) μια ακολουθία one-hot διανυσμάτων:

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

η οποία δείχνει τη σωστή ακολουθία χαρακτήρων που πρέπει να παραχθεί για το σήμα ομιλίας (κάθε διάνυσμα δείχνει σε ποια θέση του αλφάβητου A βρίσκεται το αντίστοιχο σωστό γράμμα).

Έστω $E \in \mathbb{R}^{d^{(a)} \times |A|}$ ο πίνακας με τις ενθέσεις χαρακτήρων (character embeddings) του αλφαβήτου A . Κάθε ένθεση χαρακτήρα είναι διάνυσμα $d^{(a)}$ διαστάσεων.

Οι παρακάτω τύποι περιγράφουν αναλυτικά τη λειτουργία του μοντέλου και τον υπολογισμό του σφάλματος (L) για ένα παράδειγμα εκπαίδευσης. Ο συμβολισμός $[\dots; \dots]$ παριστάνει συνένωση (concatenation). Τα f και g παριστάνουν συναρτήσεις ενεργοποίησης.



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

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

$$\vec{h}_0 \in \mathbb{R}^{d^{(h)}}$$

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

$$\overleftarrow{h}_{n+1} \in \mathbb{R}^{d^{(h)}}$$

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

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

$$t_j = E y_j \in \mathbb{R}^{d^{(a)}}$$

(Το embedding του σωστού γράμματος εξόδου στη θέση j .)

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

$$z_0 \in \mathbb{R}^{d^{(z)}}$$

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

$$W^{(a)} \in \mathbb{R}^{d^{(z)} \times 2 \cdot d^{(h)}}$$

$$U^{(a)} \in \mathbb{R}^{d^{(z)} \times d^{(z)}}$$

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



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

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

$$W^{(c)} \in \mathbb{R}^{d^{(a)} \times 2 \cdot d^{(h)}} \\ b^{(c)} \in \mathbb{R}^{2 \cdot d^{(h)}}$$

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

$$W^{(o)} \in \mathbb{R}^{|A| \times d^{(z)}} \\ b^{(o)} \in \mathbb{R}^{|A|}$$

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

(Πόσο πιθανό θεωρεί το μοντέλο το k -στό γράμμα του αλφαβήτου να είναι το σωστό για τη j -στή θέση της ακολουθίας εξόδου.)

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

(Σύμφωνα με το 1-hot y_j , το σωστό γράμμα στην j -στή θέση της εξόδου βρίσκεται στη θέση r_j του αλφαβήτου.)

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

(Ελαχιστοποιώντας το L , μεγιστοποιούμε την πιθανότητα που δίνει το μοντέλο στα σωστά γράμματα, σε όλες τις θέσεις της ακολουθίας χαρακτήρων της εξόδου.)



Άσκηση B10.2

Προσαρμόστε τους τύπους της άσκησης 2 της ενότητας B8 (επεξεργασία φυσικής γλώσσας με συνελικτικά νευρωνικά δίκτυα και Transformers) για την περίπτωση όπου στην είσοδο του κωδικοποιητή CNN έχουμε μια ακολουθία πραγματικών (ή ακεραίων) αριθμών που έχει προέλθει από δειγματοληψία του σήματος ήχου (διαφάνεια 12), αντί για μια ακολουθία από ενθέσεις λέξεων, ενώ ο αποκωδικοποιητής LSTM παράγει σε κάθε βήμα ένα γράμμα.

Υπόδειξη: Στο πρώτο συνελικτικό επίπεδο του κωδικοποιητή, τα φίλτρα εφαρμόζονται τώρα σε ένα παράθυρο $[x_{i-k}, \dots, x_i, \dots, x_{i+k}]$ με $2k + 1$ πραγματικούς αριθμούς (ή ακεραίους), όπου $2k + 1$ είναι το μέγεθος των φίλτρων, αντί για ένα παράθυρο με $3 \cdot d^{(e)}$ πραγματικούς αριθμούς που παραγόταν προηγουμένως συνενώνοντας τρία (αν χρησιμοποιούσαμε φίλτρα τριγραμμάτων λέξεων) ενθέσεις λέξεων με $d^{(e)}$ διαστάσεις η κάθε μία.



Απάντηση:

Έστω A το αλφάβητο της γλώσσας προς αναγνώριση (π.χ. Αγγλικά).⁴ Κάθε παράδειγμα εκπαίδευσης είναι ένα ζεύγος αποτελούμενο από (i) ένα διάνυσμα διακριτών τιμών ήχου, το οποίο δημιούργησε η δειγματοληψία κάποιου σήματος ομιλίας (π.χ. της εκφώνησης μιας πρότασης):

$$x_1, x_2, x_3, \dots, x_n \in \mathbb{R}$$

και (ii) μια ακολουθία one-hot διανυσμάτων:

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

η οποία δείχνει τη σωστή ακολουθία χαρακτήρων που πρέπει να παραχθεί για το σήμα ομιλίας (κάθε διάνυσμα δείχνει σε ποια θέση του αλφάβητου A βρίσκεται το αντίστοιχο σωστό γράμμα).

Έστω $E \in \mathbb{R}^{d^{(a)} \times |A|}$ ο πίνακας με τις ενθέσεις χαρακτήρων (character embeddings) του αλφαβήτου A . Κάθε ένθεση χαρακτήρα είναι διάνυσμα $d^{(a)}$ διαστάσεων.

Οι παρακάτω τύποι περιγράφουν αναλυτικά τη λειτουργία του μοντέλου και τον υπολογισμό του σφάλματος (L) για ένα παράδειγμα εκπαίδευσης. Ο συμβολισμός $[\dots; \dots]$ παριστάνει συνένωση (concatenation). Τα f και g παριστάνουν συναρτήσεις ενεργοποίησης.



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

Για $i' < 1$ και $i' > n$, θεωρούμε ότι $x_{i'} = 0$ (padding). Έστω ότι χρησιμοποιούμε $d(h)$ φίλτρα μεγέθους $2k + 1$ σε κάθε συνελκτικό επίπεδο. Τότε:

$$h_i^{(1)} = \text{ReLU}(W^{(1)}[x_{i-k}, \dots, x_i, \dots, x_{i+k}] + b^{(1)}) + x_i \in \mathbb{R}^{d^{(h)}}$$

$$W^{(1)} \in \mathbb{R}^{d^{(h)} \times (2k+1)}$$

$$b^{(1)} \in \mathbb{R}^{d^{(h)}}$$

$$h_i^{(l)} = \text{ReLU}(W^{(j)}[h_{i-k}^{(l-1)}, \dots, h_i^{(l-1)}, \dots, h_{i+k}^{(l-1)}] + b^{(j)}) + h_i^{(l-1)} \in \mathbb{R}^{d^{(h)}}$$


$$W^{(j)} \in \mathbb{R}^{d^{(h)} \times (2k+1)}$$

$$b^{(j)} \in \mathbb{R}^{d^{(h)}}$$

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

$$t_j = E y_j \in \mathbb{R}^{d^{(a)}}$$

(To embedding του σωστού γράμματος εξόδου στη θέση j .)


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

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

$$W^{(a)} \in \mathbb{R}^{d^{(z)} \times d^{(h)}}$$

$$U^{(a)} \in \mathbb{R}^{d^{(z)} \times d^{(z)}}$$

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

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

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

$$W^{(c)} \in \mathbb{R}^{d^{(a)} \times d^{(h)}}$$

$$b^{(c)} \in \mathbb{R}^{d^{(h)}}$$

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

$$W^{(o)} \in \mathbb{R}^{|A| \times d^{(z)}}$$

$$b^{(o)} \in \mathbb{R}^{|A|}$$

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

(Πόσο πιθανό θεωρεί το μοντέλο το k -στό γράμμα του αλφαβήτου να είναι το σωστό για την j -στή θέση της ακολουθίας εξόδου.)

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

(Σύμφωνα με το 1-hot y_j , το σωστό γράμμα στην j -στή θέση της ακολουθίας εξόδου βρίσκεται στη θέση r_j του αλφαβήτου.)

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

(Ελαχιστοποιώντας το L , μεγιστοποιούμε την πιθανότητα που δίνει το μοντέλο στα σωστά γράμματα, σε όλες τις θέσεις της ακολουθίας χαρακτήρων της εξόδου.)