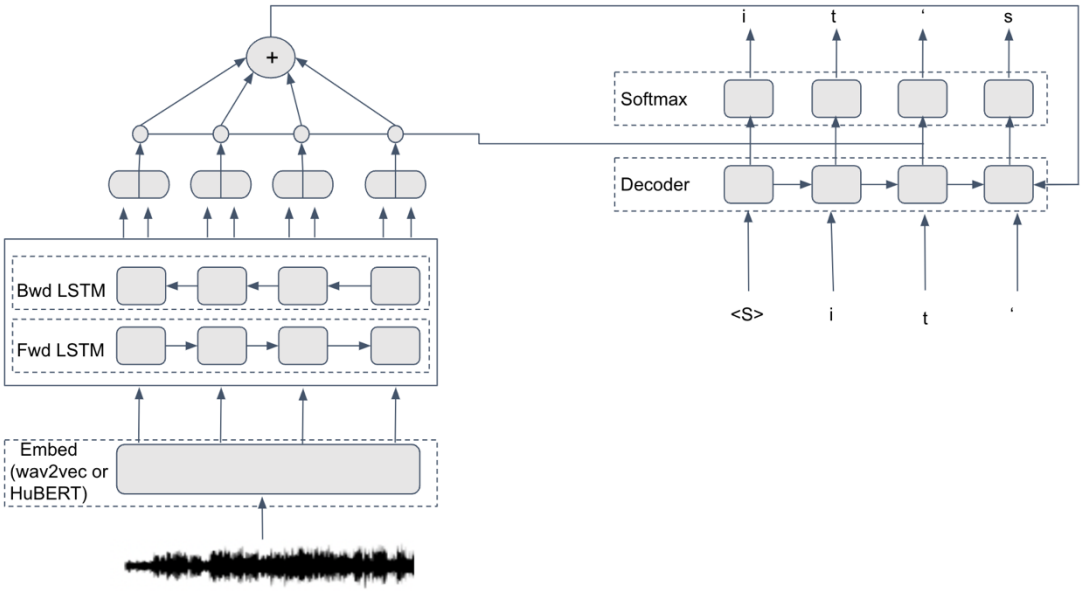
Exercises on Speech Recognition with Deep Learning

Ion Androutsopoulos, 2023–24

There are no exercises to be handed in as assignments in this part of the course.

1. Modify the formulae of exercise 5 of Part 4 (machine translation with BiLSTM encoder and LSTM encoder) for the case where the input to the encoder is a sequence of vectors representing speech frames (each vector represents a speech frame, e.g., vectors produced by wav2vec) instead of a sequence of word embeddings, and the decoder outputs at each time-step a single letter (as in slides 18–20).

Απάντηση:[1]
Το διάγραμμα της άσκησης 5 της ενότητας 4, που αφορούσε μηχανική μετάφραση, θα μετατραπεί όπως φαίνεται παρακάτω, για αναγνώριση ομιλίας:



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]

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

και (ii) μια ακολουθία one-hot διανυσμάτων:

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

[1] Ο διδάσκων ευχαριστεί την κ. Σοφία Ελευθερίου για την προετοιμασία των απαντήσεων των ασκήσεων 1 και 2.
[2] Στην πράξη χρησιμοποιούμε και άλλα ειδικά tokens, όπως είναι το <s>, τα οποία παραλείπονται εδώ.
[3] Στην πράξη συχνά υπο-δειγματοληπτούμε την ακολουθία διανυσμάτων που έχει δημιουργήσει ένα μοντέλο σαν το wav2vec ή το HuBERT.

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 $[\ldots;\ldots]$ παριστάνει συνένωση (concatenation). Τα f και g παριστάνουν συναρτήσεις ενεργοποίησης.

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

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

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

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

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

$t_j = E\, y_j \in \mathbb{R}^{d^{(a)}}$ (Το embedding του σωστού γράμματος εξόδου στη θέση *j*.)

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

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

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

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

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

$o_{j,k} = \frac{\exp(\tilde{o}_{j,k})}{\sum_{k'=1}^{|A|} \exp(\tilde{o}_{j,k'})}$ (Πόσο πιθανό θεωρεί το μοντέλο το *k*-στό γράμμα του αλφαβήτου να είναι το σωστό για τη *j*-στή θέση της ακολουθίας εξόδου.)

$r_j = \mathrm{argmax}_l\, y_{j,l}$ (Σύμφωνα με το 1-hot y_j, το σωστό γράμμα στην *j*-στή θέση της εξόδου βρίσκεται στη θέση r_j του αλφάβητου.)

$L = -\sum_j \log o_{j,r_j}$ (Ελαχιστοποιώντας το L, μεγιστοποιούμε την πιθανότητα που δίνει το μοντέλο στα σωστά γράμματα, σε όλες τις θέσεις της ακολουθίας χαρακτήρων της εξόδου.)

2. Modify the formulae of exercise 4 of Part 5 (machine translation with CNN encoder and LSTM decoder) for the case where the input to the encoder is a sequence of real (or integer) numbers obtained by sampling an input speech signal (as in slide 12) instead of a sequence of word embeddings, and the decoder outputs at each time-step a single letter. Hint: At the first convolutional layer, the filters are now applied to a window $[x_{i-k}, \ldots, x_i, \ldots, x_{i+k}]$ of $2k+1$ real numbers (or integers), where $2k+1$ is the size of the filters, instead of a window of $3 \cdot d^{(e)}$ real numbers obtained by concatenating three (for 3-gram filters) word embeddings of $d^{(e)}$ dimensions each.

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):

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

και (ii) μια ακολουθία one-hot διανυσμάτων:

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

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 $[\ldots;\ldots]$ παριστάνει συνένωση (concatenation). Τα f και g παριστάνουν συναρτήσεις ενεργοποίησης.

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

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

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

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

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

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

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

$t_j = E\, y_j \ \in \mathbb{R}^{d^{(a)}}$ (Το embedding του σωστού γράμματος εξόδου στη θέση j.)

[4] Στην πράξη χρησιμοποιούμε και άλλα ειδικά tokens, όπως είναι το <s>, τα οποία παραλείπονται εδώ.

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

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

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

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

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

$o_{j,k} = \dfrac{\exp(\tilde{o}_{j,k})}{\sum_{k'=1}^{|A|} \exp(\tilde{o}_{j,k'})}$ (Πόσο πιθανό θεωρεί το μοντέλο το *k*-στό γράμμα του αλφαβήτου να είναι το σωστό για την *j*-στή θέση της ακολουθίας εξόδου.)

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