Introduction to Deep Learning for Speech and Text Processing
Exercise Sheet 10: Tricks & Sequence-to-Sequence Models

Thang Vu

8th January 2026

Dropout

When we apply dropout on the outputs of a layer during training, we need to scale the weights at test time. This is because we need the expected value of the layer outputs in the test phase to be consistent with the expected value in the training phase, otherwise the learned parameters of the network will be useless. The expected value of a random variable is similar to the mean but in a stochastic setting. For a random variable x with $\{x_1...x_m\}$ outcomes, each with probability $\{p(x_1)...p(x_m)\}$, the expectation $E(x)$ is defined as $\sum_{i=1}^{m} p(x_i)x_i$. This corresponds to the probability-weighted average of all possible values of the random variable.

Exercise 1.

(1) Compute $E(x)$ of a neuron x with probability p to dropout during training.

(2) What scaling factor do you need at test time to obtain the same expectation of x?

(3) Can you think of a way to ensure the same expectation of x during training and testing without having to scale x at test time?

Solution 1.

(1) $E(x) = (1-p)x + p * 0 = (1-p)x$

(2) Without scaling we would have $E(x) = x$, therefore we need to multiply x by $1-p$ to obtain the same expectation as during the training.

(3) If during training we divide x each time it is not dropped out by $1-p$, we obtain $E(x) = x$ during training and we need not scale x at test time: $E(x) = \frac{1-p}{1-p}x + p * 0 = x$

Seq2seq Architecture

Exercise 2.

You are given a Seq2seq model with the following specifications: Each input in the source sequence has 100 dimensions and each output in the target sequence has 150 dimensions. The recurrent functions in the encoder and decoder are vanilla RNNs with a hidden state of 300 dimensions. In the case of attention, the output of $\tanh(W_c[c_t, h_t] + b_c)$ has 50 dimensions. Assume W_a being a square matrix for sum attention.

Compute the number of parameters for each of the following Seq2seq variants:

(1) without attention

(2) with attention using dot product as a scoring function

(3) with attention using bilinear function as a scoring function

Solution 2.
We assume a vanilla RNN of the form

$$h_t = \phi(W_{ih}x_t + b_ih + W_{hh}h_{t-1} + b_hh)\,,$$

so the number of parameters is

$$\#\text{params} = \underbrace{H \cdot D}_{W_{ih}} + \underbrace{H}_{b_ih} + \underbrace{H \cdot H}_{W_{hh}} + \underbrace{H}_{b_hh},$$

where D is the input dimension and H is the hidden dimension.
Given $D_{\text{src}} = 100$, $D_{\text{tgt}} = 150$, and $H = 300$:

Encoder RNN.

$$\#\theta_{\text{enc}} = 300 \cdot 100 + 300 + 300 \cdot 300 + 300 = 30{,}000 + 300 + 90{,}000 + 300 = 120{,}600.$$

Decoder RNN.

$$\#\theta_{\text{dec}} = 300 \cdot 150 + 300 + 300 \cdot 300 + 300 = 45{,}000 + 300 + 90{,}000 + 300 = 135{,}600.$$

Output layer (no attention). We map decoder hidden state $h_t \in R^{300}$ to the output space R^{150}:

$$\#\theta_{\text{out}} = 150 \cdot 300 + 150 = 45{,}000 + 150 = 45{,}150.$$

Attention combination layer. With attention, $c_t \in R^{300}$ and $h_t \in R^{300}$, so $[c_t; h_t] \in R^{600}$. The attentional hidden has dimension 50:

$$\#\theta_c = 50 \cdot 600 + 50 = 30{,}000 + 50 = 30{,}050.$$

Output layer (with attention). We map $\tanh(W_c[c_t; h_t] + b_c) \in R^{50}$ to R^{150}:

$$\#\theta_{\text{out}}^{\text{att}} = 150 \cdot 50 + 150 = 7{,}500 + 150 = 7{,}650.$$

(1) **Without attention:**

$$\#\theta = \#\theta_{\text{enc}} + \#\theta_{\text{dec}} + \#\theta_{\text{out}} = 120{,}600 + 135{,}600 + 45{,}150 = 301{,}350.$$

(2) **With attention (dot product scoring):**
Dot-product scoring has no parameters, hence

$$\#\theta = \#\theta_{\text{enc}} + \#\theta_{\text{dec}} + \#\theta_c + \#\theta_{\text{out}}^{\text{att}} = 120{,}600 + 135{,}600 + 30{,}050 + 7{,}650 = 293{,}900.$$

(3) **With attention (bilinear scoring):**
Bilinear scoring uses $W_a \in R^{300\times300}$ (square), so

$$\#\theta_a = 300 \cdot 300 = 90{,}000.$$

Thus

$$\#\theta = 293{,}900 + 90{,}000 = 383{,}900.$$

Machine Translation

Exercise 3.

In this exercise we will decode the output of a Seq2seq model for English → German machine translation. The Seq2seq model consists of an encoder and decoder, both with vanilla RNNs that have ReLU activations. The decoder uses a dot product attention mechanism.

The input to the encoder is the sequence 'a christmas tree ⟨EOS⟩', where ⟨EOS⟩ is a special item of the vocabulary denoting the end of the sequence. The encoder outputs are already computed and are as follows:

$\overline{h_1} = \begin{bmatrix}1\\1\end{bmatrix}, \overline{h_2} = \begin{bmatrix}1\\0.5\end{bmatrix}, \overline{h_3} = \begin{bmatrix}0.5\\0\end{bmatrix}, \overline{h_4} = \begin{bmatrix}0\\0\end{bmatrix}$

The decoder has the following weight matrices, all bias vectors are 0:

$$W_{\mathrm{i}} = \begin{bmatrix}0 & -1 & 1 & 1\\1 & 0 & 1 & 1\end{bmatrix}, W_{\mathrm{h}} = \begin{bmatrix}0 & 1\\0 & 1\end{bmatrix}, W_{\mathrm{c}} = \begin{bmatrix}0 & 1 & 1 & 2\\1 & -1 & 0.5 & 1\end{bmatrix}, W_{\mathrm{o}} = \begin{bmatrix}0 & 0\\0 & 0\\0.9 & 0\\0 & 1\end{bmatrix}.$$

The output vocabulary is V = $\{v_0 :$ ⟨SOS⟩, $v_1 :$ ⟨EOS⟩, $v_2 :$ ein, $v_3 :$ Weihnachtsbaum$\}$.

(1) Compute the first two decoder outputs. To start decoding, we give a special start of sentence token ⟨SOS⟩ as input with $y_0 = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}$. Always round to two decimals.

(2) When does the decoding process end?

Solution 3.

(1) First decoding step $t = 1$:

$$h_1 = \mathrm{ReLU}(W_i y_0 + W_h h_0) = \mathrm{ReLU}(W_i y_0 + W_h \overline{h_4}) = \mathrm{ReLU}(\begin{bmatrix}0\\1\end{bmatrix}) = \begin{bmatrix}0\\1\end{bmatrix}$$

$$\mathrm{score}(h_1, \overline{h_1}) = h_1^\top \cdot \overline{h_1} = \begin{bmatrix}0 & 1\end{bmatrix} \cdot \begin{bmatrix}1\\1\end{bmatrix} = 1, \mathrm{score}(h_1, \overline{h_2}) = 0.5, \mathrm{score}(h_1, \overline{h_3}) = 0, \mathrm{score}(h_1, \overline{h_4}) = 0$$

$$a_1 = softmax(\begin{bmatrix}1\\0.5\\0\\0\end{bmatrix}) = \begin{bmatrix}0.427\\0.259\\0.157\\0.157\end{bmatrix}$$

$$c_1 = \sum_{s=1}^{4} a_1(s) \cdot \overline{h_s} = 0.427 \cdot \begin{bmatrix}1\\1\end{bmatrix} + 0.259 \cdot \begin{bmatrix}1\\0.5\end{bmatrix} + 0.157 \cdot \begin{bmatrix}0.5\\0\end{bmatrix} + 0.157 \cdot \begin{bmatrix}0\\0\end{bmatrix} = \begin{bmatrix}0.765\\0.557\end{bmatrix}$$

$$W_{\mathrm{c}}[c_1, h_1] = W_{\mathrm{c}} \begin{bmatrix}0.765\\0.557\\0\\1\end{bmatrix} = \begin{bmatrix}2.557\\1.208\end{bmatrix}$$

$$\tanh(\begin{bmatrix}2.557\\1.208\end{bmatrix}) = \begin{bmatrix}0.988\\0.836\end{bmatrix}$$

$$W_{\mathrm{o}}(\begin{bmatrix}0.988\\0.836\end{bmatrix}) = \begin{bmatrix}0\\0\\0.898\\0.836\end{bmatrix}$$

$$p_t = \mathrm{softmax}(\begin{bmatrix}0\\0\\0.898\\0.836\end{bmatrix}) = \begin{bmatrix}0.148\\0.148\\0.363\\0.341\end{bmatrix}$$

Therefore we have $y_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$. The output of the first decoding step is $v_2 =$ 'ein '.

Second decoding step $t = 2$:

$$
\begin{aligned}
h_2 &= \text{ReLU}(W_i y_1 + W_h h_1) == \text{ReLU}(\begin{bmatrix} 2 \\ 2 \end{bmatrix}) = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \\
\text{score}(h_2, \overline{h_1}) &= \begin{bmatrix} 2 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 4, \text{score}(h_2, \overline{h_2}) = 3, \text{score}(h_2, \overline{h_3}) = 1, \text{score}(h_2, \overline{h_4}) = 0 \\
a_2 &= softmax(\begin{bmatrix} 4 \\ 3 \\ 1 \\ 0 \end{bmatrix}) = \begin{bmatrix} 0.696 \\ 0.256 \\ 0.035 \\ 0.013 \end{bmatrix} \\
c_2 &= \sum_{s=1}^{4} a_2(s) \cdot \overline{h_s} = 0.696 \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0.256 \cdot \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} + 0.035 \cdot \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} + 0.013 \cdot \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.97 \\ 0.824 \end{bmatrix} \\
W_\text{c}[c_2, h_2] &= W_\text{c} \begin{bmatrix} 0.97 \\ 0.824 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 6.824 \\ 3.146 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\tanh(\begin{bmatrix} 6.824 \\ 3.146 \end{bmatrix}) &= \begin{bmatrix} 1 \\ 0.996 \end{bmatrix} \\
W_\text{o}(\begin{bmatrix} 1 \\ 0.996 \end{bmatrix}) &= \begin{bmatrix} 0 \\ 0 \\ 0.9 \\ 0.996 \end{bmatrix}
\end{aligned}
$$

$$
p_t = \text{softmax}(\begin{bmatrix} 0 \\ 0 \\ 0.9 \\ 0.996 \end{bmatrix}) = \begin{bmatrix} 0.14 \\ 0.14 \\ 0.343 \\ 0.378 \end{bmatrix}
$$

Therefore we have $y_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$. The output of the second decoding step is $v_3 =$ 'Weihnachtsbaum '.

(2) The decoding process does not end by itself, it goes on infinitely. Therefore you have to declare stopping criteria for the decoding process. Usually, you want to stop decoding as soon as the decoder has produced the ⟨EOS⟩ symbol to indicate that the sequence has ended. Furthermore, you should declare a maximum number of decoding steps to avoid running in an infinite loop if the decoder fails to generate the ⟨EOS⟩ symbol.