(520|600).666 Information Extraction

Homework # 6

Due Friday, April 26, 2024.

For this homework, please carefully review notes and the google colab posted on the course page.

Connectionist Temporal Classification

Consider the task of recognizing an M length sequence of tokens, y_1^M, from a T length input $\mathbf{x}_1^T$. The CTC objective function is one objective for such sequence transduction tasks which works by converting y_1^M into an alignment sequence s_1^T of the same length as $\mathbf{x}_1^T$ by using an additional symbol $\oslash$ to fill the extra space. Note that it assumes $T \geq M$. $\beta^{-1}\left(y_1^M\right)$ represents the set of all possible T-length alignments, s_1^T, for the sequence y_1^M.

$$\log p\left(y_1^M | \mathbf{x}_1^T\right) = \log \sum_{s_1^T \in \beta^{-1}\left(y_1^M\right)} \prod_{t=1}^{T} p\left(s_t | \mathbf{x}_1^T\right) \tag{1}$$

We will consider that $p\left(s_t | \mathbf{x}_1^T\right)$ is obtained by normalizing the outputs of a neural network, ϕ. Element, $\phi_{s_t}^t \propto p\left(s_t | \mathbf{x}_1^T\right)$, of the matrix, ϕ, can be used to compute the CTC objective where

$$\phi = \begin{bmatrix} \phi_1^1 & \phi_1^2 & \dots & \phi_1^T \\ \phi_2^1 & \phi_2^2 & \dots & \phi_2^T \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{|S|}^1 & \dots & & \phi_{|S|}^T \end{bmatrix}, \tag{2}$$

and $|S|$ is the number of output units in the neural network.

1. **WFST Representation of the CTC Objective**

   Consider the task of modeling the word “`all`” using characters as the tokens. In this case $M = 3$. For this question, we will model words using both the Hybrid DNN-HMM, and CTC models and examine the different unweighted (i.e., the weights on each arc are 1.0) WFST topologies used for each. Remember that we can represent an HMM as a WFST. Remember that final states are indicated by using a double circle instead of a single circle around the state label. Mark all the input and output labels.

(a) Draw the WFST for the word "`all`" corresponding to using a 1-state **HMM** for each letter and uniform state transitions. Assume the output labels could be fed into a pronunciation lexicon to recover the word, i.e., `a - l - l` → `all`. In this case each state roughly represents a letter. The arcs accept symbols that the neural network could predict and produce the letters such that for any alignment of symbols `a, l, l`, i.e., `a a a l l l l l`, the letters, `a-l-l`, corresponding to the word "all" will be produced.

(b) Draw the WFST corresponding to the **CTC** topology. Recall that it should be able to accept strings starting or ending with $\oslash$.

(c) Enumerate all possible length-5 alignments accepted by the "`all`" **HMM** topology, e.g., `a a l l` would be a length-4 sequence.

(d) Enumerate all possible length-5 alignments of the word "`all`" accepted by the **CTC** topology

(e) Draw the "**HMM**" trellis, often called a lattice, corresponding to all possible length 5 sequences, but as a WFST. (It should still look very similar to a normal trellis). Ignore the weights on the arcs. Do not draw any unnecessary (unused) nodes. Please include input and output labels.

(f) Repeat (e) but for the **CTC** trellis, corresponding to all possible length 5 sequences as a WFST.

2. **WFST Composition**

The outputs of a neural network, ϕ, can themselves be represented as a WFST. Imagine the neural network has 4 outputs. In other words, it outputs vectors $\phi^t \in \mathbb{R}^{1x4}$. The first output corresponds to the $\oslash$ symbol, the second symbol corresponds to `a`, the third symbol to `b` and the fourth to `l`.

For this problem consider the matrix of scores, ϕ, where each column corresponds to a time index, and each row corresponds to one of the network outputs. These are like to observation probabilities in HMMs. We can draw this matrix as a WFST with 6 states corresponding to the length of the neural network output ($5+1$, where $+1$ is for a start state). Between each node there are as many arcs as there are output symbols (i.e., rows in the matrix). The input and output labels for this WFST will be the same (i.e., it is a WFSA), and the weights on the arcs correspond to the scores of those symbols at that position in time. Use the following matrix.

$$\begin{bmatrix} -0.002 & -0.556 & -6.605 & -5.620 & -0.109 \\ -8.502 & -0.856 & -0.005 & -1.620 & -7.309 \\ -6.402 & -6.956 & -7.705 & -2.620 & -5.609 \\ -8.802 & -7.756 & -5.605 & -0.320 & -2.309 \end{bmatrix} \tag{3}$$

(a) Draw this WFST. Feel free to use some shorthand notation provided it is logical if this process seems tedious. We will call it Φ.

(b) Use the WFST, Φ, from part (a) and compose it, with the CTC WFST, which we will call T, from problem (1.) using the log-semiring, i.e., $\Phi \circ T$. Assume unweighted arcs all had a weight of **1**. Recall that in the log-semiring, $a \bigoplus b = -\log e^{-a} + e^{-b}$, $a \bigotimes b = a + b$, $\mathbf{1} = 0$, and $\mathbf{0} = \infty$. Show each step of the composition by denoting states with the pairs of labels on states used in Φ and T, like (Φ_i, T_j). Feel free to use any computational aid, i.e. a program or calculator, to help. Remember to remove any branches that don't end in final states. A final state occurs when the both states in the pair of WFSTs are final.

(c) Compare the resulting WFST with the CTC trellis computed in problem (1.f) and comment.

(d) Bonus: Use the forward algorithm on the result of part (b) to compute the CTC objective for this utterance. Again feel free to use any computational aid.