ECE 537 Fundamentals of Speech Processing
Problem Set 8

UNIVERSITY OF ILLINOIS
Department of Electrical and Computer Engineering

Assigned: Sunday, 10/30/2022; Due: Friday, 11/4/2022
Reading: Graves, Fernández, Gomez & Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," 2006

1. The CTC article defines as the joint probability of current and future states, $\beta_t(s) = p(\mathbf{l}'_{s:(2U+1)}|\mathbf{x}_{t:T})$ (Eq. (9) in the article), as a consequence of which the posterior state probability has the awkward form shown in the summand of Eq. (14), $p(\pi_t = l'_s|\mathbf{l}, \mathbf{x}) = \frac{1}{y^t_{l'_s}}\alpha_t(s)\beta_t(s)$. Note that there is an inconsistency in the article: Equations (5) and (9) define s as an index into the length-$|\mathbf{l}|$ sequence $\mathbf{l}$, but Equations (6)-(8) and (10)-(16) define s as an index into the length-$(2|\mathbf{l}|+1)$ sequence $\mathbf{l}'$. We will assume the latter definition, and will use the symbol U to mean $|\mathbf{l}|$.

   In this problem, we will consider a definition of $\beta_t(s)$ that gives slightly cleaner equations. Consider the following definition:

$$\beta_t(s) \equiv p(\mathbf{l}'_{s:(2U+1)}|\mathbf{x}_{(t+1):T}, \pi_t = l'_s) \tag{1}$$

   Note that Eq. (1) specifies that the label sequence starting from time t is $\mathbf{l}_{s:(2U+1)}$, but that, rather than depending on $\mathbf{x}_{t:T}$, this probability is dependent on $\mathbf{x}_{(t+1):T}$ and $\pi_t = l'_s$.

   (a) (1 point) Given the definition of $\beta_t(s)$ in Eq. (1), and the definition of $\alpha_t(s)$ in the article's Eq. (5), what is $p(\pi_t = l'_s|\mathbf{l}, \mathbf{x})$ as a function of $\alpha_t(s)$ and $\beta_t(s)$?

   Solution:

$$p(\pi_t = l'_s|\mathbf{l}, \mathbf{x}) = \alpha_t(s)\beta_t(s)$$

   (b) (1 point) Given the definition of $\beta_t(s)$ in Eq. (1), what is the "initialize" step of the backward algorithm? In other words, find a formula for $\beta_T(s)$ in terms of any of the neural net outputs. If you wish, you can use U to mean the length of $\mathbf{l}$, and $2U+1$ to mean the length of $\mathbf{l}'$.

   Solution:

$$\begin{aligned}\beta_T(s) &\equiv p(\mathbf{l}'_{s:(2U+1)}|\mathbf{x}_{(T+1):T}, \pi_T = l'_s)\\ &= \begin{cases} 1 & s \in \{2U, 2U+1\}\\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

   (c) (1 point) Given the definition of $\beta_t(s)$ in Eq. (1), what is the "iterate" step of the backward algorithm? In other words, find a formula for $\beta_t(s)$ in terms of $\beta_{t+1}(s')$, and in terms of any of the neural net outputs. Be sure to take into account the fact that the character at time $t+1$ may be s, $s+1$, or, if $l'_s \neq b$ and $l'_s \neq l'_{s+2}$, $s+2$. Note that your answer will be a little different from Eq. (10) in the article, because our definition of $\beta_t(s)$ is a little different.

Solution:

$$\beta_t(s) \equiv p(\mathbf{l}'_{s:(2U+1)}|\mathbf{x}_{(t+1):T}, \pi_t = l'_s)$$
$$= \begin{cases} y^{t+1}_{l'_s}\beta_{t+1}(s) + y^{t+1}_{l'_{s+1}}\beta_{t+1}(s+1) & l'_s = b \text{ or } l'_s = l'_{s+2} \\ y^{t+1}_{l'_s}\beta_{t+1}(s) + y^{t+1}_{l'_{s+1}}\beta_{t+1}(s+1) + y^{t+1}_{l'_{s+2}}\beta_{t+1}(s+2) & \text{otherwise} \end{cases}$$

(d) (1 point) The new definition of $\beta_t(s)$ requires a revision of Equations (14) and (15) in the article. How should these two equations read if one is using the new definition of $\beta_t(s)$?

Solution: Equation (14) should read:

$$p(\mathbf{l}|\mathbf{x}) = \sum_{s=1}^{|\mathbf{l}'|} \alpha_t(s)\beta_t(s)$$

Equation (15) should read:

$$\frac{\partial p(\mathbf{l}|\mathbf{x})}{\partial y^t_k} = \frac{1}{y^t_k} \sum_{s \in \mathbf{lab}(\mathbf{l},k)} \alpha_t(s)\beta_t(s)$$

2. Using the un-numbered equations preceding Eq. (15) in the article, it's possible to re-write Eq. (15) as

$$\frac{\partial p(\mathbf{l}|\mathbf{x})}{\partial y^t_k} = \frac{1}{y^t_k} p(\mathbf{l}, \pi_t = k|\mathbf{x}),$$

where $p(\mathbf{l}, \pi_t = k|\mathbf{x})$ is the probability that the label sequence is $\mathbf{l}$, and that the character generated at time t is k. Differentiating the loss function $\mathcal{L} = -\ln p(\mathbf{l}|\mathbf{x})$ therefore gives us

$$\frac{\partial \mathbf{L}}{\partial y^t_k} = -\frac{1}{y^t_k}\frac{p(\mathbf{l}, \pi_t = k|\mathbf{x})}{p(\mathbf{l}|\mathbf{x})} = -\frac{1}{y^t_k} p(\pi_t = k|\mathbf{l}, \mathbf{x})$$

Defining $\gamma_t(k) = p(\pi_t = k|\mathbf{l}, \mathbf{x})$ gives the equation reported in lecture:

$$\frac{\partial \mathbf{L}}{\partial y^t_k} == -\frac{\gamma_t(k)}{y^t_k}$$

Suppose we know that the softmax outputs, y^t_k, are defined in terms of the softmax logits, u^t_i, as

$$y^t_k = \frac{e^{u^t_k}}{\sum_j e^{u^t_j}}$$

You may recall, from homework 7, that the derivative of the softmax can be written in this way:

$$\frac{\partial y^t_k}{\partial u^t_i} = \begin{cases} y^t_k(1 - y_k) & i = k \\ -y^t_k y^t_i & \text{otherwise} \end{cases} \tag{2}$$

Eq. (2) is sometimes written as:

$$\frac{\partial y^t_k}{\partial u^t_i} = y^t_k(\delta_{ik} - y^t_i), \tag{3}$$

where δ_{ik} is an indicator function, defined as

$$\delta_{ik} = \begin{cases} 1 & i = k \\ 0 & \text{otherwise} \end{cases}$$

Remember that the chain rule is

$$\frac{\partial \mathbf{L}}{\partial u_i^t} = \sum_k \frac{\partial \mathbf{L}}{\partial y_k^t} \frac{\partial y_k^t}{\partial u_i^t}$$

Use the chain rule to prove Eq. (16) in the article, i.e., to show that $\frac{\partial \mathbf{L}}{\partial u_i^t} = y_i^t - \gamma_t(i)$. Hint: what is $\sum_k \gamma_t(k)$?

Solution:

$$\begin{aligned} \frac{\partial \mathbf{L}}{\partial u_i^t} &= \sum_k \frac{\partial \mathbf{L}}{\partial y_k^t} \frac{\partial y_k^t}{\partial u_i^t} \\ &= -\sum_k \frac{\gamma_t(k)}{y_k^t} y_k^t (\delta_{ik} - y_i^t) \\ &= -\sum_k \gamma_t(k)\delta_{ik} + \sum_k \gamma_t(k) y_i^t \\ &= -\gamma_t(i) + y_i^t \end{aligned}$$