TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2019

Connectionist Temporal Classification (CTC)

Connectionist Temporal Classification (CTC)
A Successful Deep Latent Variable Model

Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Alex Graves, Santiago Fernandez, Faustino Gomez, Jurgen Schmidhuber, ICML 2006

CTC

A speech signal $x[T, J]$ is labeled with a phone sequence $y[N]$ with $N << T$.

$x[t, J]$ is a speech signal vector.

$y[n] \in \mathcal{P}$ for a set of phonemes $\mathcal{P}$.

The length N of $y[N]$ is not determined by T and the correspondence between n and t is not given.

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{\langle x, y \rangle \sim \mathrm{Train}} \; -\ln P_\Phi(y[N] \mid x[T, J]) \quad N << T$$

The CTC Model

We define a model

$$P_\Phi(z[T] \mid x[T, J])$$

$$z[t] \in \mathcal{P} \cup \{\bot\}$$

$y[N]$ is the result of removing $\bot$ from $z[T]$.

$$z[T] \;\Rightarrow\; y[N]$$

$$\bot, a_1, \bot, \bot, \bot, a_2, \bot, \bot, a_3, \bot \;\Rightarrow\; a_1, a_2, a_3$$

The CTC Model

For $p \in \mathcal{P} \cup \{\bot\}$ we have an embedding vector $e[p, I]$. The embedding is a parameter of the model.

We take the phonemes $z[t]$ to be independently distributed.

$$p_\Phi(Z[T] \mid x[T, J]) = \prod_t P_\Phi(z[t] \mid x[T, J])$$

$$h[T, \tilde{J}] = \mathrm{RNN}_\Phi(x[T, J])$$

$$P_\Phi(z[t] \mid x[T, J]) = \operatorname*{softmax}_{z[t]} \; e[z[t], I] \, W[I, \tilde{J}] \, h[t, \tilde{J}]$$

Dynamic Programming

Let $\vec{y}[t]$ to be the prefix of $y[N]$ emitted by the first t elements of z.

$$\begin{aligned} \vec{y}[t] &= z[1:t] - \perp \\ F[n,t] &= P(\vec{y}[t] = y[1:n]) \end{aligned}$$

$F[0,0] = 1$

For $n = 1, \ldots, N$ $F[n,0] = 0$

For $t = 1, \ldots, T$

  $F[0,t] = P(z[t] = \perp)F[0,t-1]$

  For $n = 1, \ldots, N$

    $F[n,t] = P(z[t] = \perp)F[n,t-1] + P(z[t] = y[n])F[n-1,t-1]$

Back-Propagation

$$\mathcal{L} = -\ln F[N, T]$$

We can now back-propagate through this computation.

END