CONNECTIONIST TEMPORAL CLASSIFICATION

UCS749: SPEECH PROCESSING AND SYNTHESIS

Raghav B. Venkataramaiyer

CSED TIET Patiala India.

September 16, 2024

Outline

OUTLINE

SUMMARY

AUTHOR Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J.

URL [ACM], [PDF from toronto.edu], [PDF from tum.de], [Pytorch] [Tensorflow] [Julia]

DATE 2006

BOOKTITLE Proceedings of the 23rd International Conference on Machine Learning

SUMMARY

- type of neural network output and associated scoring function (for RNN's);
- to tackle sequence problems where the timing is variable;
- CTC refers to the outputs and scoring, and is independent of the underlying neural network structure.

SUMMARY

> *For example, in speech audio there can be multiple time slices which correspond to a single phoneme. Since we don't know the alignment of the observed sequence with the target labels we predict a probability distribution at each time step. — Wikipedia*

See also: https://distill.pub/2017/ctc/

OUTLINE

KEY OBSERVATION

An input waveform for the word HELLO may vary in the following ways,

- A quick and slow speaker may stretch it at varying lengths, *e.g.* HELLLOO vs HEELLLOOOO; and the same may be extend to syllable stresses and intonations when speaking in further detail;
- The start points and blanks may vary, *e.g.* ---HEE-LLOO- vs -HELLLOO-

We call this as an alignment problem, where given an alphabet, say $\{H, E, L, O\}$, and a sequence $[x_1, \ldots, x_T]$, find the correspondence.

KEY OBSERVATION

- Inherent to this formulation, there's no way to distinguish between HELO vs HELLO.
- This problem is like mode collapse.
- To this end, the author introduces a special character called CTC blank ϵ, that suppresses mode collapse.
- *E.g.* $\mathtt{LLLL} \rightarrow \mathtt{L}$ in post-processing, but $\mathtt{LL}\text{-}\mathtt{LL} \rightarrow \mathtt{LL}$.
- Hence, the alphabet now becomes $\{H, E, L, O, \epsilon\}$.
- This problem grows polynomially in alphabet size and exponentially in sequence size.

OUTLINE

Minimise transcription mistakes from speech to text or handwriting to text, where the natural measure is a *label error rate* LER of a temporal classifier h, defined as follows.

$$\mathrm{LER}(h, \mathbb{S}') = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{z})\sim\mathbb{S}'} \left[\frac{\mathrm{ED}(h(\mathbf{x}), \mathbf{z})}{|\mathbf{z}|} \right]$$

where,

- $\mathbb{S}' \subset \mathcal{D}_{\mathcal{X}\times\mathcal{Z}}$ is the test sample;
- ED is the edit distance.

OUTLINE

The problem may be seen as,

$$Y_* = \arg\max_Y P(Y|X)$$

$$P(Y|X) = \sum_{A \in \mathcal{A}(X,Y)} \prod_{t=1}^{T} P_t(\mathbf{a}_t|X)$$

where,

- $X \equiv [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be the input sequence;
- $Y \equiv [\mathbf{y}_1, \ldots, \mathbf{y}_m]$ be the output sequence;
- $A \equiv [\mathbf{a}_1, \ldots, \mathbf{a}_T]$ be an alignment between $\mathbf{x}$ and $\mathbf{y}$ — also the network output; and
- $\mathcal{A}(X, Y)$ be such an alignment space;

KEY CONTRIBUTION (CONTD...)

INPUTS $X \in \mathbb{R}^{(\cdot) \times n}$ is a spectrogram-like audio input, like MFCC, providing n time step sequence of input.

NETWORK OUTPUT (or RNN output) is the alignment $A \in \mathbb{R}^{|L'| \times T}$. Here $L' \equiv L \cup \epsilon$ is the alphabet augmented with CTC blank.

ALPHABET $Y \in \mathbb{R}^m$ (also known as Y_{mask} in some implementations) is the output sequence augmented by blanks, *e.g.* -HEL-LO- or HEL-LO for HELLO, as a sequence of indices; or seldom tokens dependent upon implementation detail.

BACK-PROPAGATION WITH FORWARD-BACKWARD ALGORITHM

$\mathcal{A}(X, Y)$ is grows exponentially in the length of sequence, *i.e.* $\mathcal{O}(m^T)$

But similar to the HMM, a recursive definition enables us to compute the loss efficiently in $\mathcal{O}(m^2 T)$.

OUTLINE

IN FRAMEWORKS

[Pytorch]
[Tensorflow]