Streaming Automatic Speech Recognition with the Transformer Model

Niko Moritz, Takaaki Hori, Jonathan Le Roux

ICASSP
May, 2020

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com
Motivation

• End-to-end automatic speech recognition (ASR) has greatly simplified the pipeline for building and applying ASR systems.

• Offline end-to-end ASR systems have shown to surpass the performance of traditional hybrid DNN-HMM solutions.

• Streaming end-to-end architectures are still lacking behind this success.

• Encoder-decoder based architectures have demonstrated to achieve the best end-to-end ASR results but are difficult to apply in a streaming fashion.

This work

• Our proposed triggered attention (TA) concept is used to overcome these difficulties.

• The TA concept is applied to the transformer architecture, achieving SOTA streaming end-to-end ASR results.
Outline

• Encoder-Decoder Neural Networks
  – Attention
  – Transformer
  – Self-attention
  – Time-Restricted Self-Attention
  – Streaming Encoder-Decoder Attention (prior work)

• Triggered Attention
  – Architecture
  – Frame-Synchronous Decoding Algorithm

• LibriSpeech Results
Attention

Input sequence:

Query vector:

Embedding/Feature vector:

Output vector:

Weight distribution
Encoder-Decoder Attention

Hello World <eos>

state

Decoder

Encoder States

Encoder

Acoustic Features

Feature Extraction

Audio waveform
Transformer Architecture

Encoder

- $E \times \text{Feature Extraction}$
  - Positional Encoding
  - Add & Norm
  - Multi-Head Attention
  - Audio Input

Decoder

- $D \times \text{Encoder-decoder attention}$
  - Output Embeddings
  - Previous Output
  - Add & Norm
  - Multi-Head Attention
  - Positional Encoding
  - Linear
  - Softmax

- $E=12$
- $D=6$
Self-Attention

Output sequence:

Input sequence:

current frame

Query vector:
Input sequence:

- past frames
- current frame
- future frames

Self-Attention
Time-Restricted Self-Attention

Output sequence: 

Input sequence: 

Algorithmic delay: \( \# \text{layers} \cdot \varepsilon^{\text{enc}} = 4 \) frames
Adaptive Chunking based on Selection Probability

Example:
- Monotonic Chunkwise Attention (MoChA) [1]

Problems:
- Backpropagation with discrete decisions is not possible.
- No frame-synchronous decoding algorithm.
- Detecting word or word-piece positions is a good part of the ASR job that defines insertion and deletion errors.

[1] C. Chiu and C. Raffel, “Monotonic chunkwise attention,” in Proc. ICLR, Apr. 2018.
Triggered Attention (TA) Architecture

Decoding output: \( Y = (y_1, \ldots, y_l) \rightarrow \)

Encoder output: \( X_E = (x_1^E, \ldots, x_N^E) \rightarrow \)

Transformer model
- #encoder layers \( E = 12 \)
- #decoder layers \( D = 6 \)
- attention dimension = 512
- #attention heads = 8

Acoustic features: \( X = (x_1, \ldots, x_T) \rightarrow \)

N. Moritz, T. Hori, and J. Le Roux, “Triggered attention for end-to-end speech recognition,” in Proc. ICASSP, May 2019, pp. 5666–5670.

© MERL
Frame-Synchronous Decoding

Frame-synchronous CTC prefix beam search [1]:

\[ \ell_1 = (\langle \text{sos} \rangle, \text{Hello}) \]
\[ \ell_7 = (\langle \text{sos} \rangle, \text{Hey}, \text{World}, \text{World}) \]
\[ \ell_8 = (\langle \text{sos} \rangle, \text{Hello}, \text{World}) \]

Set of prefix sequences after pruning:

\[ \Omega = \{ \ell_1, \ell_2, \ell_8 \} \]

\[ \log p(\ell|X_{1:n}^E) = \log p_{\text{prfx}}(\ell|X_{1:n}^E) + \alpha \log p_{\text{LM}}(\ell) + \beta |\ell| \]

- \( \ell \): prefix sequence
- \( X_{1:n}^E \): Encoder state sequence for frame (1, ..., n)
- \( p_{\text{prfx}} \): CTC prefix probability
- \( p_{\text{LM}} \): Language model (LM) probability
- \( \alpha \): LM weight
- \( \beta \): insertion bonus weight
- \( |\ell| \): prefix sequence length

[1] A. L. Maas, A. Y. Hannun, D. Jurafsky, and A. Y. Ng, “Firstpass large vocabulary continuous speech recognition using bidirectional recurrent DNNs,” arXiv preprint arXiv:1408.2873, 2014.
Frame-Synchronous Decoding

Frame-synchronous one-pass TA decoding [1]:

$$\log p_{\text{joint}}(\ell|X^E_{1:n}) = \lambda \log p_{\text{prfx}}(\ell|X^E_{1:n}) + (1 - \lambda) \log p_{\text{ta}}(\ell|X^E_{1:n}) + \alpha \log p_{\text{LM}}(\ell) + \beta |\ell|$$

$p_{\text{ta}}$: Triggered attention probability
$\lambda$: CTC weight
$\nu = n' + \varepsilon_{\text{dec}}$
$n'$: trigger frame
$\varepsilon_{\text{dec}}$: decoder look-ahead

[1] N. Moritz, T. Hori, and J. Le Roux, “Streaming end-to-end speech recognition with joint CTC-attention based models,” in Proc. ASRU, Dec. 2019, pp. 936–943.
| Encoder | Full-sequence CTC-attention decoding [1,2] |
|---------|------------------------------------------|
|         | Clean Dev | Test | Other Dev | Test |
| Full-sequence | 2.4      | 2.7   | 6.0      | 6.1   |

| Time-restricted encoder | Frame-synchronous CTC prefix beam search | TA: $\varepsilon_{\text{dec}} = 18$, delay: $\varepsilon_{\text{dec}} \cdot 40 \text{ ms} = 720 \text{ ms} |
|-------------------------|------------------------------------------|---------------------------------|
| $\varepsilon_{\text{enc}} / \text{ delay}$ | Clean Dev | Test | Other Dev | Test | Clean Dev | Test | Other Dev | Test |
| 0 / 30 ms               | 3.3      | 3.7   | 9.4      | 9.4   | 2.9      | 3.2   | 8.1      | 8.0   |
| 1 / 510 ms             | 3.0      | 3.3   | 8.4      | 8.6   | 2.8      | 3.0   | 7.5      | 7.8   |
| 2 / 990 ms             | 2.9      | 3.1   | 8.0      | 8.2   | 2.7      | 2.9   | 7.3      | 7.4   |
| 3 / 1470 ms            | 2.8      | 2.9   | 7.8      | 8.1   | 2.7      | 2.8   | 7.1      | 7.2   |
| Full-sequence          | 2.5      | 2.8   | 6.9      | 7.0   | 2.4      | 2.6   | 6.1      | 6.3   |

* Algorithmic encoder delay: $E \cdot \varepsilon_{\text{enc}} \cdot \text{frame-rate} + \text{CNN-delay}

$E = 12$, frame-rate = 40 ms, CNN-delay = 30 ms

[1] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” J. Sel. Topics Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.
[2] S. Karita, N. Yalta, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, “Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration,” in Proc. ISCA Interspeech, Sep. 2019, pp. 1408–1412.
Conclusions

• The triggered attention (TA) concept enables frame-synchronous decoding with an encoder-decoder based model for the first time.
• The TA concept enables joint scoring of an CTC and attention-based decoder model in a streaming fashion.
• The proposed system achieves state-of-the-art results for streaming end-to-end ASR on the LibriSpeech corpus.
MITSUBISHI ELECTRIC

Changes for the Better