An Improved Single Step Non-autoregressive Transformer for Automatic Speech Recognition

Ruchao Fan[1], Wei Chu[2], Peng Chang[2], Jing Xiao[2], Abeer Alwan[1]

fanruchao@g.ucla.edu

[1]Dept. of Electrical and Computer Engineering, UCLA, USA

[2]PAII Inc., USA

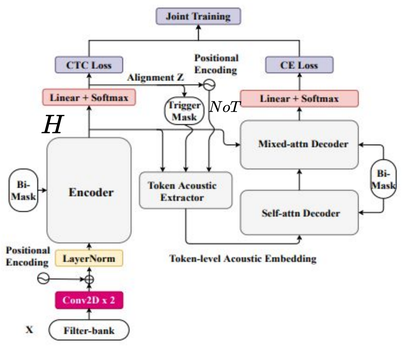
Introduction

- In end-to-end ASR, the autoregressive mechanism in a transformer decoder slows down the inference speed.
- Our previous work (CASS-NAT) [13] has shown a significant real time factor (RTF) improvement over autoregressive transformers (AT).
- However, the word error rate (WER) performance of CASS-NAT is still far behind the AT.
- **In this work, we propose an improved CASS-NAT** by 1): applying convolution augmented self-attention blocks to the transformer decoder; 2) expanding the time boundary of each token to increase the robustness of CTC alignment; and 3) using iterated loss to enhance the gradient updates of low-layer parameters.
- The improved CASS-NAT achieves an impressive WER gain (9.2%->7.2% using Librispeech "test other" dataset without an external LM), and is close to the WER of AT (7.0%).
- **The visualization of the attention distribution and token-level acoustic embedding** may explain why the improved CASS-NAT performs similarly to AT.

System Overview

1. Framework

Figure 1. A review of the CASS-NAT architecture.



- **Encoder:** extract high level representation H
- **CTC:** optimize CTC alignment to offer auxiliary information for token-level acoustic embedding extraction.
  - Time boundary for each token (trigger mask)
  - Number of tokens for decoder input (NoT)
- **Token-acoustic extractor:**
  - 1 self-attention block, where K and V are equal to H
  - Q: sinusoidal positional embedding with NoT
  - Mask: trigger mask from CTC alignment (Viterbi in training, error-based sampling in inference [13])
- **Decoder:**
  - self-att block (not considering H)
  - mix-att block (considering H)
- **CE:** cross entropy loss to optimize the final WER.

2. Proposed training strategies

- Conv-augmented transformer decoder:

$$\hat{s}_i = s_i + \frac{1}{2}\mathrm{FFN}(s_i) \quad (3)$$

$$s_i^{'} = \hat{s}_i + \mathrm{LN}(\mathrm{Attn}(\hat{s}_i, \hat{s}_i, \hat{s}_i, \mathrm{BiMask})) \quad (4)$$

$$s_i^{''} = s_i^{'} + \mathrm{Conv}(s_i^{'}) \quad (5)$$

$$s_i^{'''} = s_i^{''} + \mathrm{LN}(\mathrm{Attn}(s_i^{''}, H, H, \mathrm{BiMask})) \quad (6)$$

$$o_i = \mathrm{LN}(s_i^{'''} + \frac{1}{2}\mathrm{FFN}(s_i^{'''})) \quad (7)$$

  - FFN is decomposed into two sub-layers to be placed at the beginning and the end of the block.
  - A convolution layer similar to that in [2] is inserted between the self-attention and cross- attention layer.
  - LN is the layer normalization and Attn is:

$$Attn(Q, K, V, M) = Softmax(\frac{QK^T}{\sqrt{d_k}}) \otimes M \cdot V \quad (2)$$

- Expansion of trigger mask:
  - Time boundary [t1, t2] from Viterbi alignment may not be accurate.
  - Extending the boundary to [t1-τ, t2+τ] to compensate for the inaccuracy of CTC alignment.
- Iterated functions in encoder and decoder:
  - Deep models suffer from gradient vanishing, especially for parameters that are distant from the output layers.
  - Iterated loss is proposed to add additional loss functions after each layer to boost the gradient update.
  - Iterated CTC loss is added in the middle encoder layer
  - Iterated CE loss is added in the decoder as follows:

$$L_{\text{joint}} = \lambda_{CE} L_{dec}^{final} + (1-\lambda_{CE}) L_{dec}^{middle} + \lambda_{CTC} L_{CTC}^{final} + (1-\lambda_{CTC}) L_{CTC}^{middle} \quad (8)$$

Experiment - Librispeech

1. Experimental Setup

- Input and output:
  - 80-dim log-mel filter bank features
  - Every 3 frames are concatenated to be a 240-dim input.
  - Output: 5k word-pieces obtained by SentencePiece [34].
- Model
  - 2 CNNs: 64 filter, kernel size 3, stride 2
  - AT baseline: $N_e = 10, N_d = 5, d_{FF} = 2048, H = 8,$
  - CASS-NAT: $d_{MHA} = 512$
    - 1-layer token-acoustic extractor
    - Decoder: 3 self-att blocks and 4 mix-attn blocks
  - SpecAug, Label smoothing, **Encoder initialization**
- Proposed methods (empirically chosen):
  - Maximum relative position k is 20 in the encoder and 8 in the decoder
  - Trigger mask expansion τ=1
  - λ_{CE} : 0.9, λ_{CTC} : 0.5
- Decoding:
  - RTF was conducted using an NVIDIA Tesla V100 GPU with batch size of one.

2. Results - Improvement of the proposed methods

Table 1. WERs of the proposed methods for improving CASS-NAT on Librispeech. No external language model is used. SpecAug is used in all configurations. WERR is the incremental relative WER

Model w/o LM	dev-clean	dev-other	test-clean	test-other	*WERR*
Conformer AT	2.7	7.2	3.0	7.0	
CASS-NAT	3.7	9.2	3.8	9.1	-
+ Conv-aug Enc.	3.1	7.9	3.3	7.9	13.2%
+ Conv-aug Dec.	3.0	7.8	3.1	7.6	3.8%
+ Tri. Mask Exp.	3.0	7.6	3.1	7.5	1.3%
+ Iterated CTC	2.8	7.3	3.1	7.3	2.7%
+ Iterated CE	2.8	7.3	3.1	7.2	1.4%

- Each proposed method can incrementally improve the WER performance of the CASS-NAT.
- The final model has a very close performance to the AT baseline, which is promising.

3. Results - comparisons with previous work

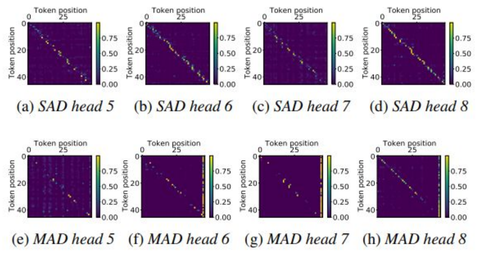
Table 2. RTFs of previous work are missing because the authors did not report them, or the machines used to test RTFs are different. †: use SpecAug. *: use speed perturbation.

Librispeech (WER)	LM	test clean	test other	RTF test-other
Previous work (NAT)				
A-FMLM [7] †	w/o	6.6	12.2	-
Imputer [8]	w/o	4.0	11.1	-
Align-refine [14] †	w/o	3.6	9.0	-
CASS-NAT [13] †	w/o	3.8	9.1	0.010
Conformer AT†	w/o	3.0	7.0	0.499
	w/	2.3	5.2	0.568
Improved CASS-NAT †	w/o	**3.1**	**7.2**	0.014
	w/	2.8	6.5	0.188

- Compared to CASS-NAT [13], the WER has a significant improvement with little RTF degradation.
- The shallow fusion for CASS-NAT does not work well as shown in [13].

4. Visualization Analysis

Figure 1. Attention weight distributions of the last four heads in multi-head self-attention of the last block in the self-attention decoder (SAD) and mixed-attention decoder (MAD) for the first utterance in the Lirispeech train-clean-100 subset.



(a) *SAD head 5* (b) *SAD head 6* (c) *SAD head 7* (d) *SAD head 8*

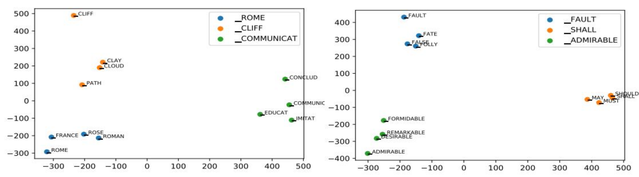
(e) *MAD head 5* (f) *MAD head 6* (g) *MAD head 7* (h) *MAD head 8*

- Most of the heads learn a monotonic alignment indicating that each token relies more on adjacent tokens.
- This is similar to the idea of word embedding using continuous bag of words (CBOW) and skip-gram [37].

Acknowledgement

This work was supported in part by NSF and PAII.

Figure 2. Visualization of token-level acoustic embedding for two examples, each having three word pieces using the first two PCA dims.



- The first 3000 utterances in the train-clean-100 subset was used for acoustic embedding computation by averaging the embeddings for each token.
- Three randomly chosen word pieces and their four closest embeddings are reduced to a 2-dimensional space using principal component analyses (PCA) and are then plotted.
- The behaviour of acoustic embedding is similar to that of word embedding in AT.

Experiment - Aishell1 and child corpus

1. Results - Aishell1

Table 3. A comparison of WERs on Aishell1 with previous studies.

Aishell1 (CER)	LM	dev	test	RTF test
Previous work (NAT)				
ST-NAT [10] †	w/o	6.9	7.7	-
A-FMLM [7] ⋆	w/o	6.2	6.7	-
Insertion-NAT [12] †	w/o	6.1	6.7	-
Enhanced-NAT [15] †⋆	w/o	5.3	5.9	-
BERT-LASO [18] †	w/o	5.2	5.8	-
CASS-NAT [13] †⋆	w/o	5.3	5.8	0.011
Conformer AT †⋆	w/o	4.8	5.2	0.200
Improved CASS-NAT †⋆	w/o	**4.9**	**5.4**	0.023

- Similar experimental setup as Librispeech, please refer to the paper for model details.
- The same behaviour is observed for the Aishell1 Mandarian corpus, where the improved CASS-NAT has similar WERs compared to AT.

2. Results - OGI kids corpus

Table 4. WERs for the development and test sets and RTF for the test set using the scripted part of the OGI corpus. Both experiments used SpecAug.

	dev	test	RTF on test
Conformer AT	1.8	2.5	0.081
Improved CASS-NAT	2.2	2.6	0.018

- The RTF improvement may be suitable for child ASR-based educational applications.

Conclusion

- This paper presented three methods to improve the WER performance of CTC alignment-based single step NAT (CASS-NAT), followed by performance analyses.
- We achieved a 7%~21% relative WER/CER improvement over the original CASS-NAT on Librispeech and Aishell1 datasets with little RTF degradation.
- Embedding visualization might explain why the improved CASS-NAT performed similarly to AT.

References

Reference numbers are the same as those in the paper.