In recent years, autoregressive transformer (AT) achieves great success for automatic speech recognition. However, the autoregressive mechanism in transformer decoder slows down the inference speed. Non-autoregressive transformer (NAT) was proposed for parallel generation to accelerate the inference.

Limitations for current NAT models:
- Iterative NAT still needs multiple generation steps, which cannot fully exploit the potential of NAT.
- Single step NAT extracts incomplete acoustic representations, thus the performance is worse than AT.

Novel Contributions: 1) We propose a novel framework, CTC alignment-based single step NAT (CASS-NAT). 2) An error-based sampling alignment strategy during inference is further proposed to improve the WER performance.

The proposed CASS-NAT achieves WERs of 3.8%/9.1% on Librispeech test clean/other dataset without an external LM, and a CER of 5.8% on Aishell1 Mandarin corpus.

Compared to AT baseline, the CASS-NAT has a performance reduction on WER, but is 51.2x faster in terms of RTF.

1. Framework

Figure 1. The proposed CASS-NAT architecture.

1.1) Encoder: extract high level representation H
1.2) CTC: optimize the CASS-NAT alignment that offers auxiliary information for token-level acoustic embedding extraction.
1.3) Error-based sampling: token-level acoustic embedding extraction.

2. Training Criterion

Given $X = \{x_1, x_2, \ldots, x_T\}$ and $Y = \{y_1, y_2, \ldots, y_T\}$, the CTC alignment $Z$ is introduced, the objective function is:

$$
\log P(Y | X) = \log \sum_{Z \in q} P(Y | Z, X)
$$

where $q$ is the set of alignments which can be mapped to $Y$.

Maximum approximation is applied to reduce computation:

$$
\log P(Y | X) \geq \max_{Z \in q} \log \sum_{z \in Z} P(Y | z, X)
$$

where $f_b$ is the end boundary of token $u$.

The final objective function is:

$$
\max_{Z \in q} \log \sum_{z \in Z} P(Y | z, X) + \lambda \log \sum_{z \in Z} P(z | X)
$$

3. Analyse of the performance

Mismatch rate (MR): Deletion and insertion errors compared to the oracle alignment. Substitution errors do not affect token-level acoustic embedding extraction.

Length prediction error rate (LPER): Taking the alignment as output and removing blank and repetitions, the ratio of utterances with different length compared to ground truth.

Table 2. A comparison of different alignment generation methods in CASS-NAT decoding without LM.

For ESA, no further gains are observed when the number of sampled alignments is over 50.

Correct estimation of the decoder input length is more important for NAT.

The WER can be lowered than 2% for the utterances with correct token number estimation.

The figure shows the importance of length prediction accuracy on the encoder side.

1. Experimental Setup

The setup is almost the same as that for Librispeech except:
- 4230 Chinese characters as output from training set.
- $N_v = 6$
- Additionally use speed perturbation.

2. Results

Table 3. A comparison of WERs on Aishell1 with the existing works.

Con: beam search, slow

1. Conclusion

This work presents a novel CASS-NAT framework:
- CASS-NAT alignment is used as auxiliary information to extract token-level acoustic embedding.
- Word embedding in AT is replaced with acoustic embedding for parallel generation.
- Viterbi-alignment is used for training.
- Error-based sampling alignment is proposed for inference.
- The importance of length prediction for decoder input is shown by analyzing the relationships between different alignments and the oracle alignment.
- We decrease the gap between AT and NAT, and maintain the acceleration for NAT.

References

The number is appeared as the same in the paper.