Intermediate Loss Regularization for CTC-based Speech Recognition

Jaesong Lee[NAVER], Shinji Watanabe[JHU]

ICASSP 2021 #5115

Introduction

- Improve CTC-based ASR model using Intermediate CTC, a simple auxiliary loss
- Combine Intermediate CTC and stochastic depth for further improvement
- With Conformer, it reaches CER 5.2% on AISHELL-1 with greedy decoding and without any language models, nearly reaching SOTA of autoregressive+LM
- Also improves Mask CTC, CTC-based non-autoregressive model

Architecture

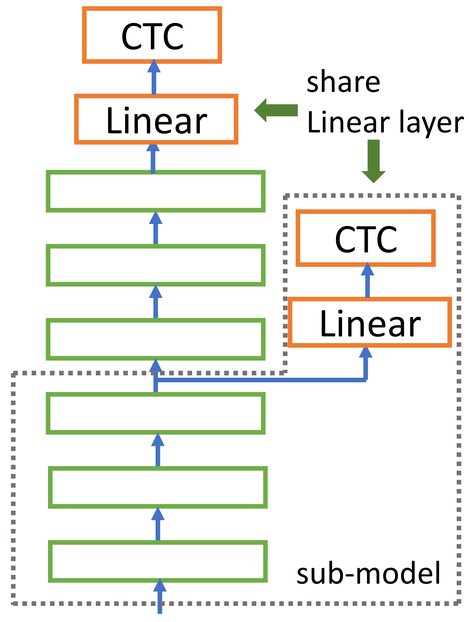
- Transformer/Conformer encoder + CTC
- Transformer: self-attention + residual network
- Conformer: augment Transformer with convolution layers
- Greedy decoding for CTC, no beam search or external language models!

Stochastic Depth

- Regularization method for residual networks (e.g. Transformer and Conformer)
- During training, each layer may be skipped with some probability
- Improves deep models (≥ 24 layers), but does not improve shallow models

Intermediate CTC

- StochDepth skips each layer independently
- Instead, consider a sub-model by skipping upper layers as a whole
- When the full model is forwarded, the sub-model can be computed with very small overhead!
- Loss = (1-w) * full model CTC + w * sub-model CTC
- During testing, only the full model CTC is used
- Intermediate CTC consistently improves CTC performance, from 6 layers to 48 layers
- Can be combined with Stochastic Depth for further improvement



Experiments

- Intermediate CTC always improves all model of any depth
- Stochastic Depth may not improve shallow network
- Using both gives better result, implying two methods are complementary to each other
- With conformer, it achieves WSJ 9.9% and AISHELL-1 5.2%, nearly reaching autoregressive state-of-the-art results (WSJ 9.3% and AISHELL-1 5.1%)

Transformer	WSJ (WER) dev93	WSJ (WER) eval92	TED-LIUM2 (WER) dev	TED-LIUM2 (WER) test	AISHELL-1 (CER) dev	AISHELL-1 (CER) test
12-layer	20.1	16.5	14.8	14.0	5.8	6.3
+ InterCTC	17.5	13.6	13.3	12.3	5.7	6.2
+ StochDepth	19.8	16.2	13.8	13.1	5.9	6.4
+ both	16.8	13.7	13.2	12.1	5.7	6.1
24-layer	17.8	13.9	12.6	12.2	5.4	5.9
+ InterCTC	15.3	12.4	11.5	10.6	5.1	5.6
+ StochDepth	16.3	12.7	11.9	11.2	5.2	5.7
+ both	14.9	**11.8**	10.9	10.2	5.2	5.5
48-layer	16.6	13.8	11.6	10.9	5.1	5.7
+ InterCTC	14.9	12.6	10.7	10.3	5.1	5.5
+ StochDepth	15.6	12.9	11.0	10.2	5.0	5.4
+ both	**14.2**	**11.8**	**10.3**	**9.9**	**4.9**	**5.3**

Conformer	WSJ (WER) dev93	WSJ (WER) eval92	TED-LIUM2 (WER) dev	TED-LIUM2 (WER) test	AISHELL-1 (CER) dev	AISHELL-1 (CER) test
12-layer	15.2	12.4	10.5	9.8	5.4	6.0
+ InterCTC	13.4	10.8	**9.7**	**9.1**	5.1	5.6
+ StochDepth	13.1	10.8	11.1	10.7	5.2	5.8
+ both	**12.0**	**9.9**	10.8	9.9	**4.7**	**5.2**

Mask CTC	threshold	dev93	eval92
12enc-6dec	0.0	16.5	13.5
	0.999	15.7	12.9
+ InterCTC	0.0	14.4	11.6
	0.999	**14.1**	**11.3**
Mask CTC [13]	0.999	15.4	12.1
Align-Refine [37]	-	13.7	11.4