DYNAMIC ALIGNMENT MASK CTC: IMPROVED MASK CTC WITH ALIGNED CROSS ENTROPY

平安科技 PING AN TECHNOLOGY

Xulong Zhang[1,#], Haobin Tang[1,2,#], Jianzong Wang[1,*], Ning Cheng[1], Jian Luo[1], Jing Xiao[1]

[1]Ping An Technology (Shenzhen) Co., Ltd. & [2]University of Science and Technology of China

[#] Indicates equal contribution ⋆ Indicates corresponding author

Introduction

Problem in previous researches:

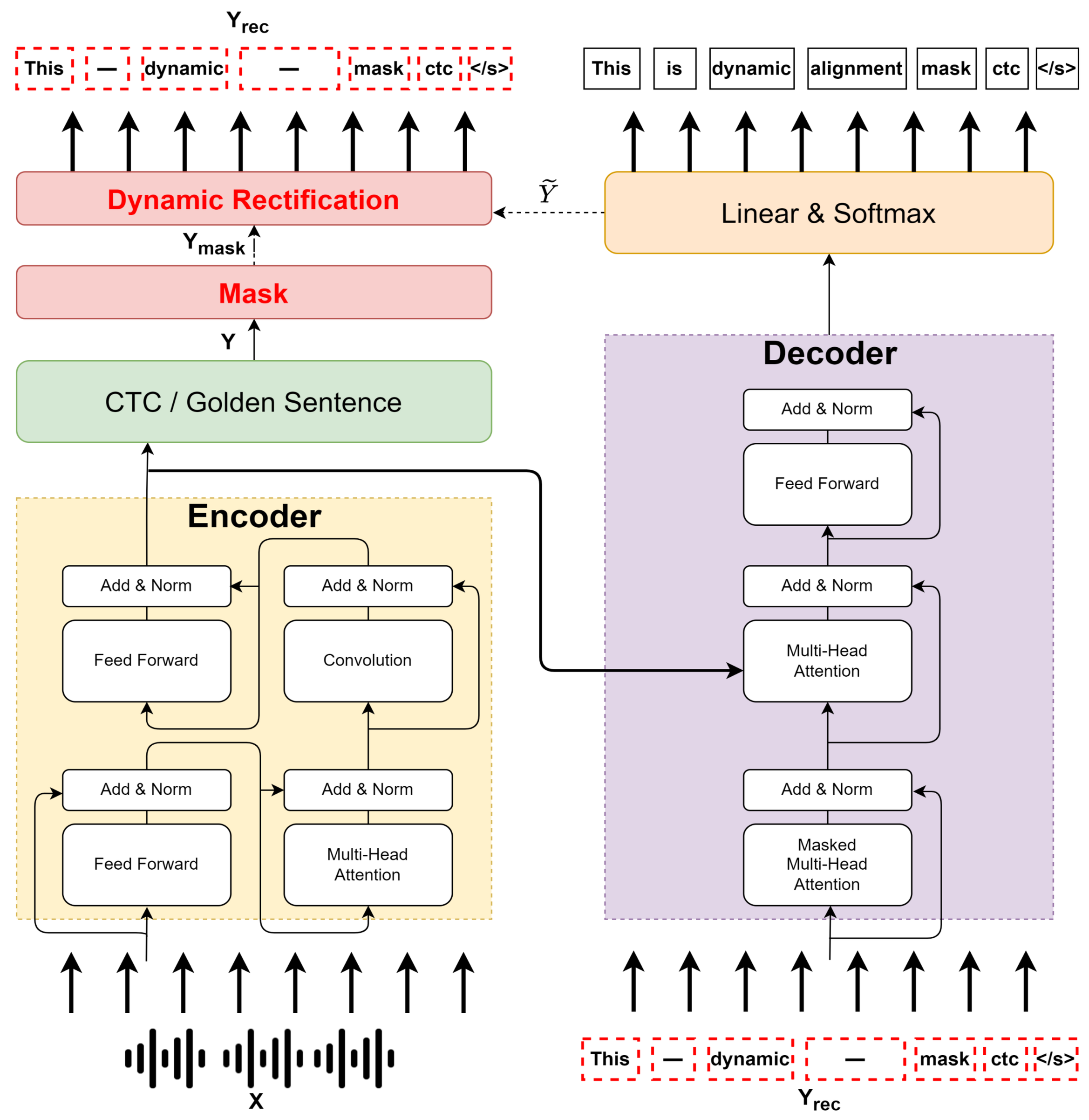
- decoder network of NAR model is usually trained on cross entropy (CE) loss, small shift in predicted tokens will result in large loss penalty, even if the content of tokens matches very well.
- The decoder input of Mask CTC is the greedy CTC search at the inference stage, while the ground truth sentence is inputted to the decoder at the training stage. This causes a mismatch between the training and inference.

Our contributions:

- We introduce the AXE loss as a relaxed loss instead of CE loss to the decoder training of Mask CTC making the model focus on the tokens matching instead of tokens ordering.
- we propose a dynamic rectification method to alleviate mismatch problem between the training and inference.

Proposed Method

Dynamic Rectification



- We started with the ground truth sentence Y, and randomly mask some tokens to get Ymask.
- After that, Ymask is inputted into dynamic rectification algorithm. We used current best model to predict Ỹ based on Ymask, and masked the tokens again. Therefore, the output sentence Yrec may has wrong predicted tokens (like "task" in Yrec instead of "mask" in Y).
- Finally, the sentence Yrec, Y, and audio features, compose the new training sample.

Y	This	is	dynamic	alignment	**mask**	CTC	non	autoregressive	speech	recognition
Y_{mask}	This	⟨*mask*⟩	dynamic	⟨*mask*⟩	⟨*mask*⟩	CTC	non	⟨*mask*⟩	⟨*mask*⟩	recognition
$\tilde{Y}$	This	is	dynamic	alignment	**task**	CTC	non	autoregressive	speech	recognition
Y_{rec}	This	is	⟨*mask*⟩	⟨*mask*⟩	**task**	CTC	⟨*mask*⟩	autoregressive	speech	recognition

AXE Loss

Dynamic programming is used by AXE to determine the optimal alignment between the current prediction Yj and ground truth token Yi.

- **align**, aligning the current prediction Yj and ground truth token Yi with probability P(Yj |Yi, X),
- **skip prediction**, skipping the current prediction Yj and inserting a special token ε to the ground truth token Yi
- **skip target**, skipping the current ground truth token Yi without incrementing the prediction j.

Experiments

Dataset: WSJ

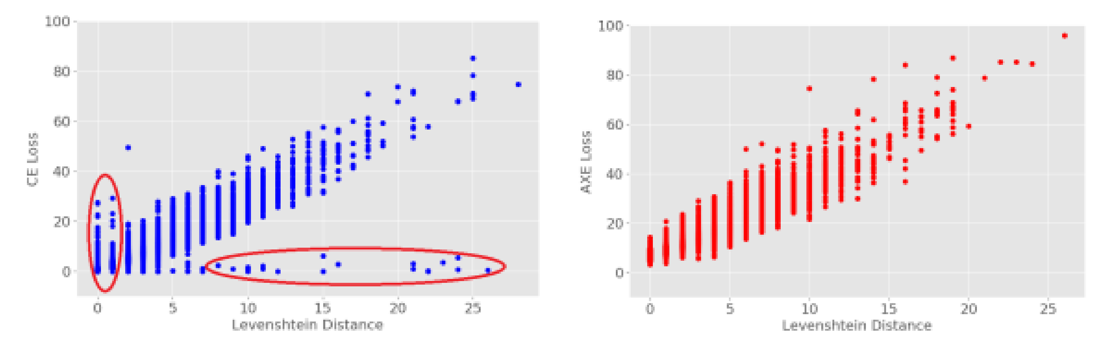
- All experiments are conducted on the ESPNET2 toolkit with **WSJ1** and **WSJ0.**
- si284, dev93, and eval92 for training, validation, and testing respectively.

Different Training Methods & Decoding Iterations,

Training Method	Iter	dev93	eval92	RTF
Transformer				
Mask + CE	1	16.8	14.3	0.04
Mask + AXE	1	15.8	12.5	0.04
Mask + Rec + AXE	1	15.3	11.8	**0.04**
Mask + CE	10	16.5	13.9	0.07
Mask + AXE	10	15.7	12.2	0.07
Mask + Rec + AXE	10	**15.2**	**11.6**	0.07
Conformer				
Mask + CE	1	14.6	12.1	0.04
Mask + AXE	1	13.9	11.4	0.04
Mask + Rec + AXE	1	13.7	11.3	**0.04**
Mask + CE	10	14.1	11.7	0.07
Mask + AXE	10	13.7	11.2	0.07
Mask + Rec + AXE	10	**13.6**	**11.1**	0.07

Compared with Other Non-Autoregressive and Autoregressive Methods

Model	Iter	dev93	eval92	RTF
Autoregressive				
Transformer				
CTC-Attention	S	13.5	10.9	4.62
Conformer				
CTC-Attention	S	11.1	8.5	5.09
Non-Autoregressive Previous Work				
Transformer				
CTC	1	19.4	15.5	0.03
Mask CTC*	10	16.5	13.9	0.06
Mask CTC + DLP	10	13.8	10.8	0.07
Imputer (IM)	8	-	16.5	-
Imputer (DP)	8	-	12.7	-
Align-Refine	10	13.7	11.4	0.06
Conformer				
CTC	1	13.0	10.8	0.03
Mask CTC*	10	14.1	11.7	0.06
Mask CTC + DLP	10	11.3	9.1	0.08
Our Work				
Transformer				
Proposed	10	15.2	11.6	0.07
Conformer				
Proposed	10	13.6	11.1	0.07



(a) CE Loss (b) AXE Loss

- The first outliers have large CE loss but have small Levenshtein distance, mainly caused by token order mismatch
- The second outliers have small CE loss, but their predictions are quite different from ground truth sentence. (e.g. "form or" and "for more")

Conclusion

- Our proposed method use AXE loss which makes the model focus on the tokens matching and relaxes the restriction of tokens order.
- The dynamic rectification could reduce the mismatch of decoder input between training and inference, simulating the high confidence but possible wrong tokens of greedy CTC output.