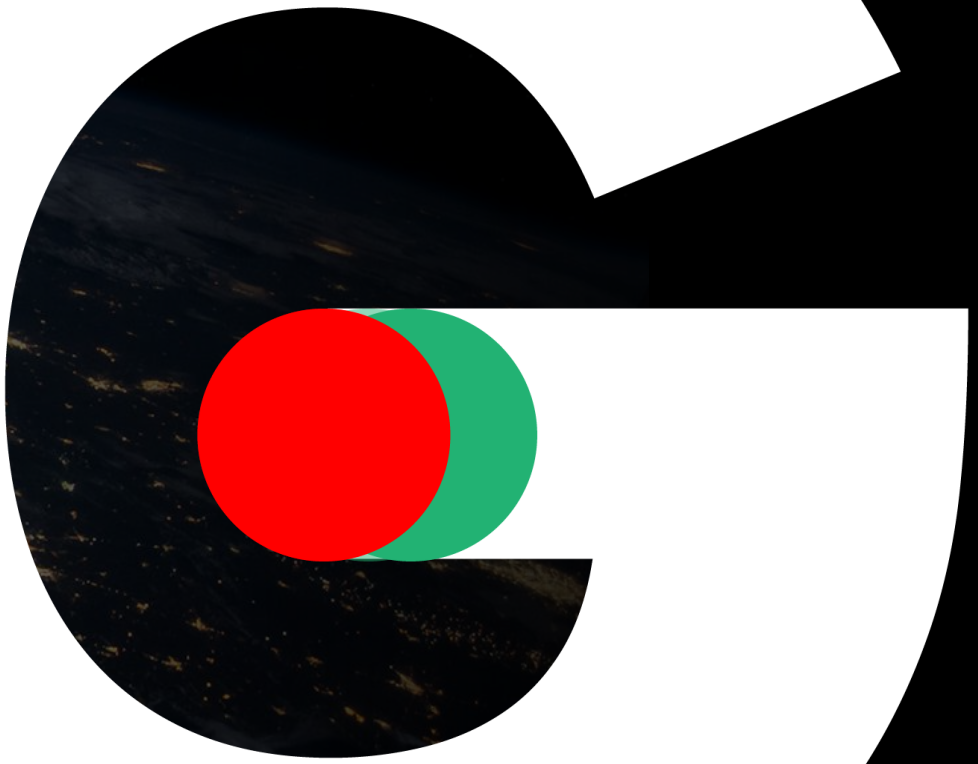
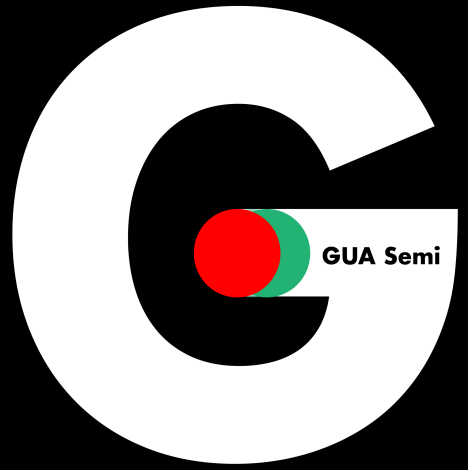
CNVSRC2023 分享

Task 1 Single-speaker VSR fixed track

T266 GUA Speech
雷超、李盛强、马宝忠

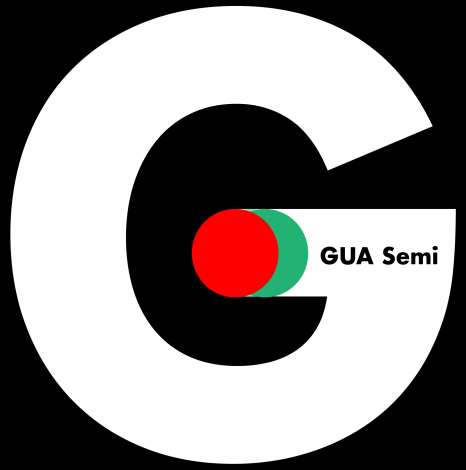
2023.12





目录

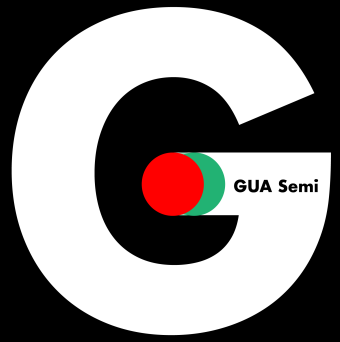
CONTENTS



01

数据

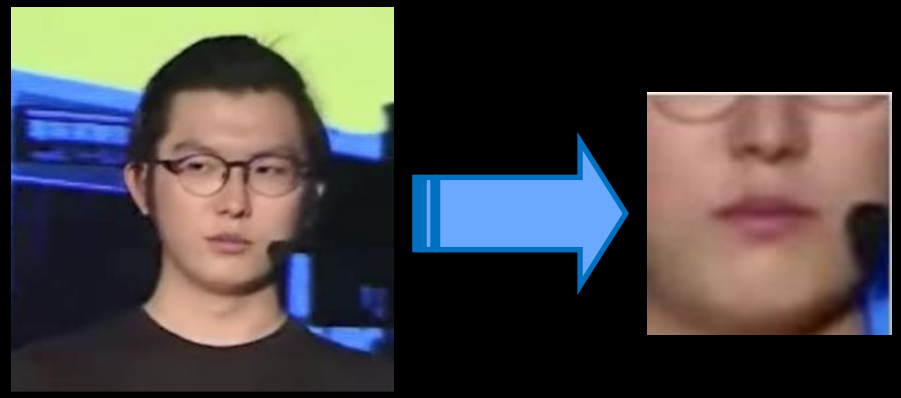
数据预处理；训练集、验证集划分；



GUA

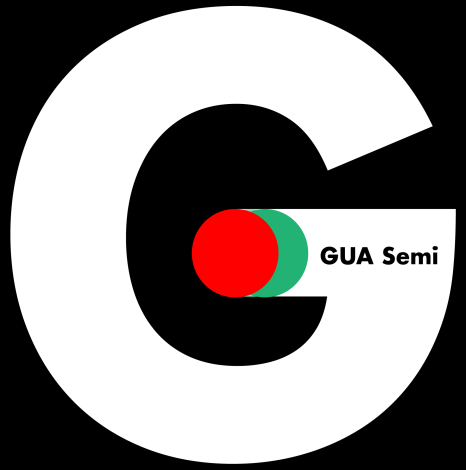
数据预处理

将原始MP4，使用人脸检测器识别关键点，最终以96 × 96 的边界框裁剪嘴唇区域作为训练数据



数据分布

	CN-CVS	CNVSRC-Single Dev
train	175058 (252h)	25038(90.6h)
valid	913(1.4h)	568(2h)
test	903(1.4h)	



02 模型

Visual Front-end

Back-end Networks

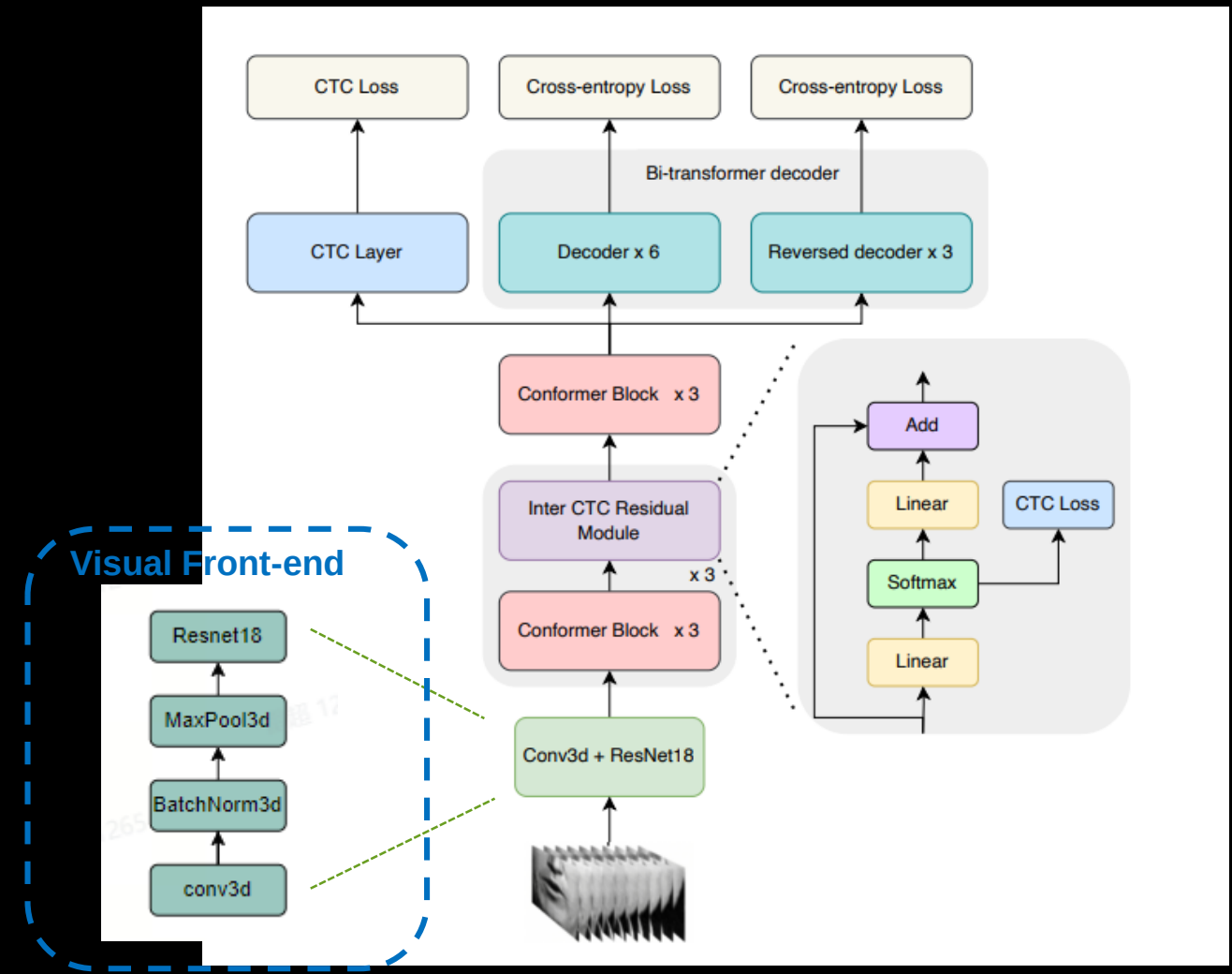
Objective Function

模型

Visual Front-end

Conv3d+ResNet18

Conv3d： (5, 7, 7), (1, 2, 2), (2, 3, 3)



模型

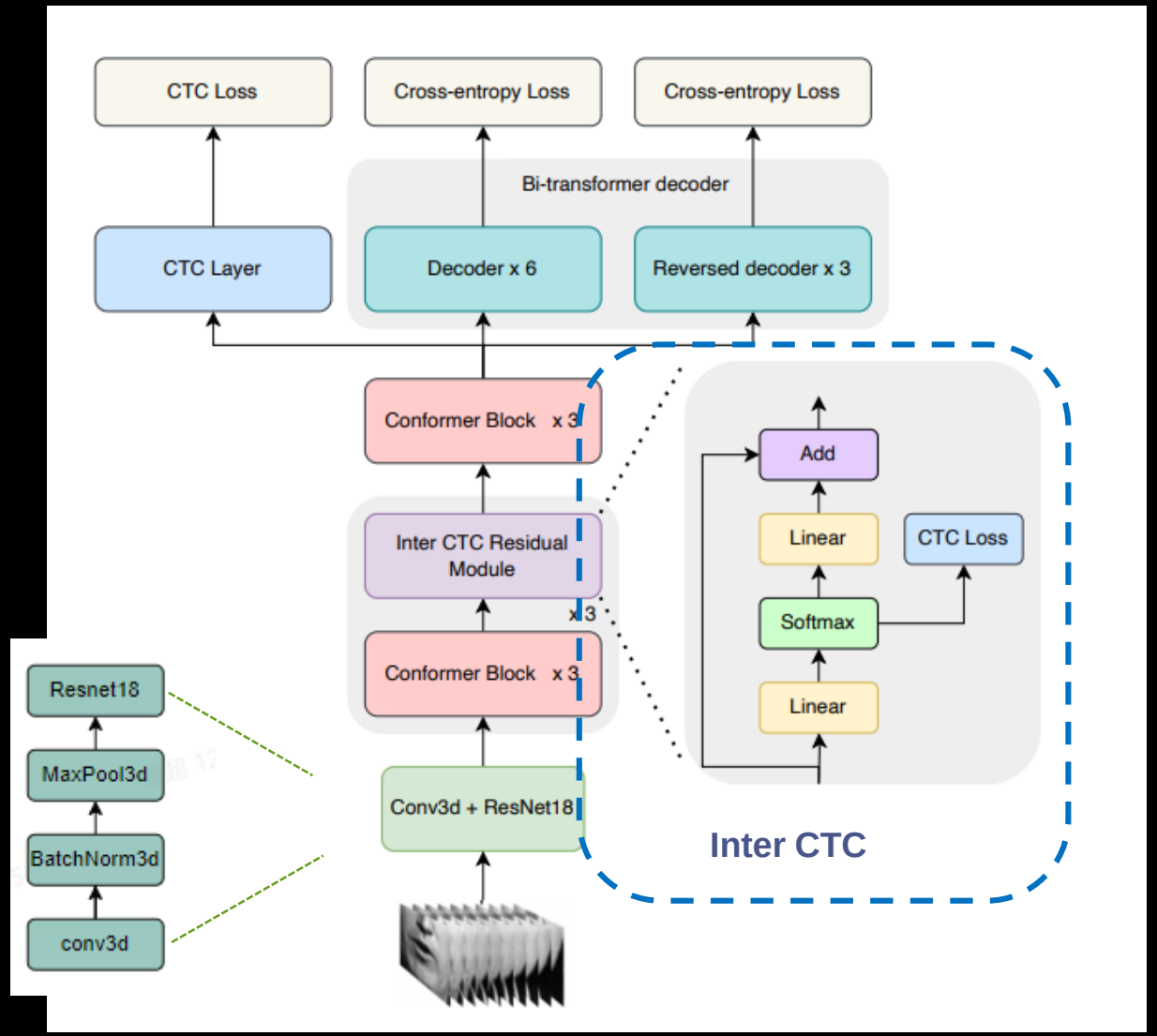
Back-end Networks

1. Conformer
2. **Inter CTC residual modules**
3. Bi-transformer decoder

$$\boldsymbol{Z}_l = \text{Softmax}\left(\text{Linear}\left(\boldsymbol{X}_{\text{out}}^l\right)\right)$$
$$\boldsymbol{X}_{\text{in}}^{l+1} = \boldsymbol{X}_{\text{out}}^l + \text{Linear}\left(\boldsymbol{Z}_l\right)$$

Inter-CTC loss:

$$\mathcal{L}_{\text{inter}} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{inter}}^k$$



Back-end Networks

1. Conformer
2. Inter CTC residual modules
3. **Bi-transformer decoder**

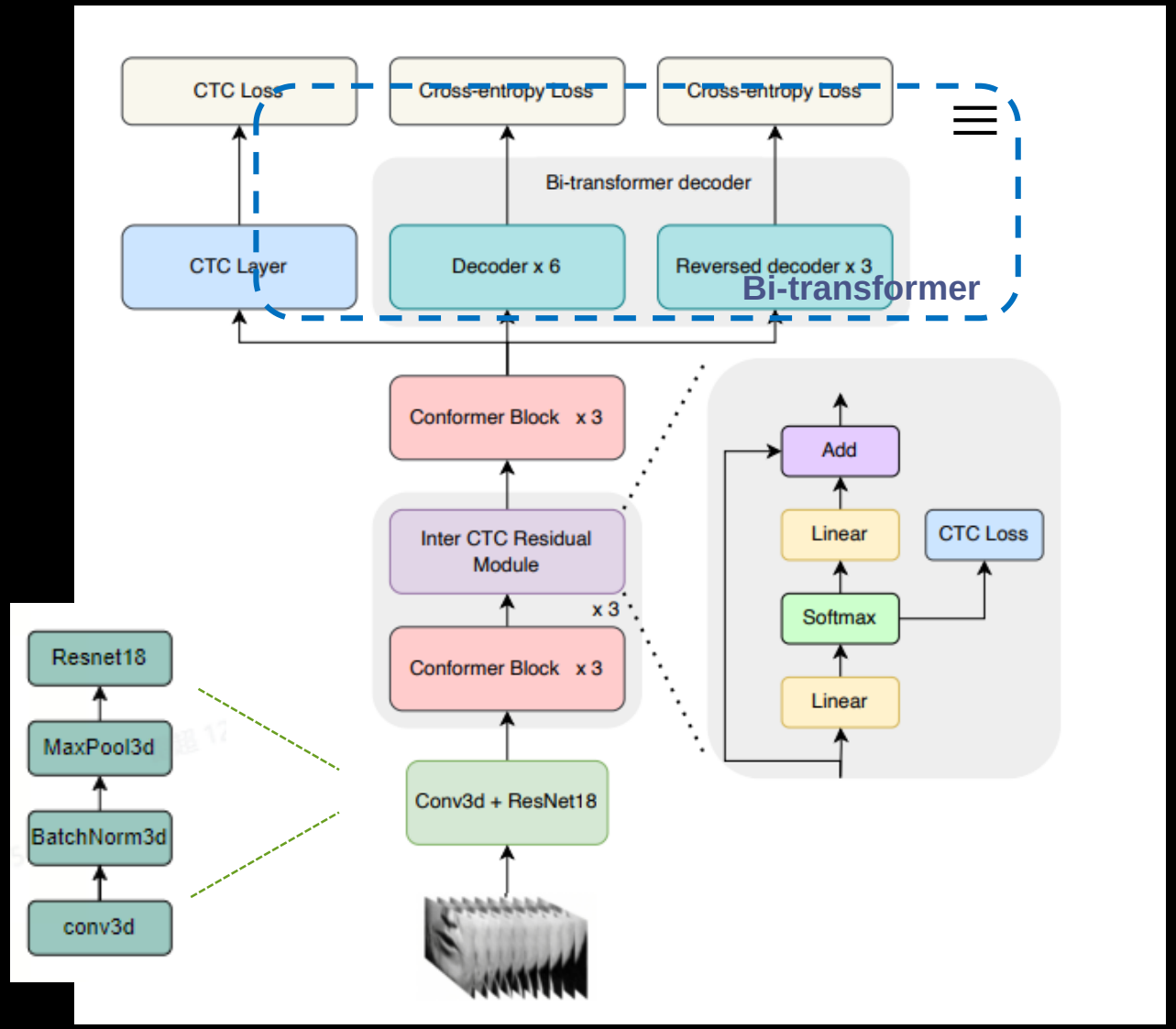
decoder loss:

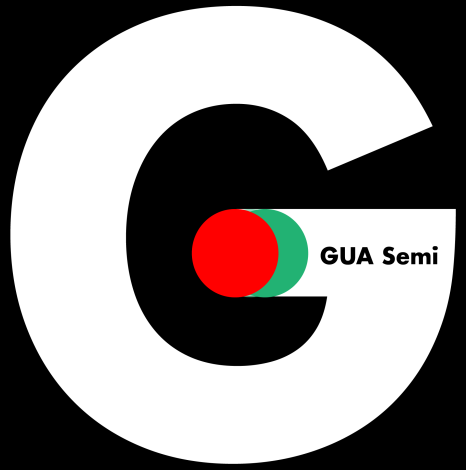
$$\mathcal{L}_{left} = -\log\left(\prod_{l=1}^{L} p\left(y_l \mid \boldsymbol{y}_{1:l-1}, \boldsymbol{X}_e\right)\right)$$

$$\mathcal{L}_{right} = -\log\left(\prod_{l=L}^{1} p\left(y_l \mid \boldsymbol{y}_{L:l-1}, \boldsymbol{X}_e\right)\right)$$

$$\mathcal{L}_{attn} = (1-\alpha)\,\mathcal{L}_{left} + \alpha\mathcal{L}_{right}$$

$$\mathcal{L} = \lambda\left(\gamma\mathcal{L}_{inter} + (1-\gamma)\,\mathcal{L}_{ctc}\right) + (1-\lambda)\,\mathcal{L}_{attn}$$





03 实验

Pre-processing & Experimental settings

Training & Inference

Results

Pre-processing & Experimental settings

Pre-processing

Lip regions：96 × 96

Augment：random cropping and adaptive time masking

Units：使用汉字作为建模单元，词表由 CN-CVS 和 CNVSRC-Single.Dev 训练集的文本生成的 4466 个汉字和 < blank >，< unk >，< eos/sos >组成。

Experimental settings

Encoder：

conformer layers：12

attention dimension：768，attention heads：12

cnn module kernel size：31，feed-forward network dimension：3072

Decoder：

6 transformer decoders and 3 reversed transformer decoders

attention dimension：768，attention heads：12

cnn module kernel size：31，feed-forward network dimension：3072

RNN-LM：

layers：2，hidden size：650

实验

Training & Inference

Training

Stage1：使用CN-CVS 中视频时长不超过 4 秒数据来训练 vsr 模型，model average： epoch 14-23

Stage2：加载stage1 Pre-trained Model，使用CN-CVS 的全量数据集继续训练， model average：epoch 65-74

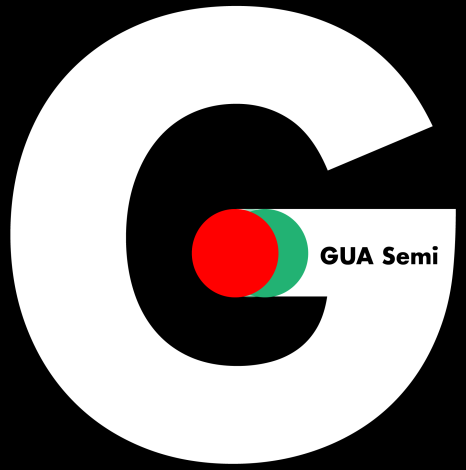
Stage3：加载stage2 Pre-trained Model，使用CNVSRC-single.dev数据训练，model average：epoch 23-28

RNN-LM：使用CN-CVS和CNVSRC-Single.Dev的train data训练60epoch

Inference

解码方式：使用 joint CTC/attention one-pass decoding，并使用RNN语言模型进行浅层融合。

解码时CTC权重为0.3，lm权重为0.1，beam size为40



04 结论

Results

结论

Results

分析

Tabel1：测试集cer

Tabel2：

trian set1： 22767 utterances of CNVSRC-Single.Dev

valid set1：2839 utterances of CNVSRC-Single.Dev

trian set2： 25038 utterances of CNVSRC-Single.Dev

valid set2：568 utterances of CNVSRC-Single.Dev

总结

1. Inter CTC 残差模块和Bi-transformer 带来了显著的效果
2. 以汉字为建模单元也提高了识别能力
3. RNN语言模型提升了系统的性能

System	Model	CER (%)
B1	Official baseline	48.60
M1	Proposed system	38.09

Tabel1：测试集cer

System	Model	Valid set 1	Valid set 2
M1D2	Proposed system	-	36.46
M1D1	Proposed system	40.46	40.37
M2D1	M1D1 - RNNLM	40.62	40.51
M3D1	M2D1 - char unit	42.36	42.38
M4D1	M3D1 - Bi-transformer decoder	43.19	43.15
M5D1	M4D1 - Inter CTC residual module	48.57	48.34

Tabel2：验证集cer

感谢各位专家批评指正

T266： 雷超，李盛强，马宝忠