System Description for T267 Team System

陈顺飞 2023.12

目录

一、数据使用

- 原始数据：
  CNCVS：300hours
  Single-dev：83.7hours
  Multi-dev：18 hours

- 数据使用策略：
  加噪扩充： add salt and pepper noise；
  Gaussian blur
  Single-dev：83.7x2=166 hours
  Multi-dev： 36 hours

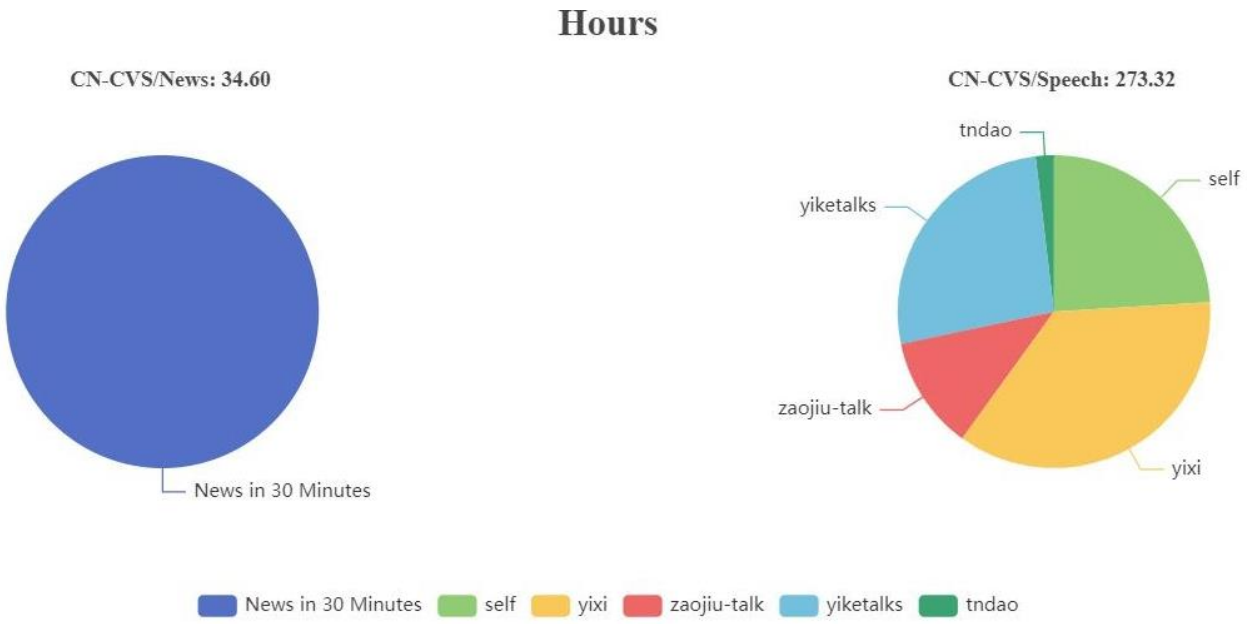


Table 1. Data profile of CNVSRC-Single and CNVSRC-Multi.

	CNVSRC-Single		CNVSRC-Multi	
DataSet	Dev	Eval	Dev	Eval
# Videos	25,947	2,881	20,450	10,269
# Hours	94.00	8.41	29.24	14.49

二、系统方案描述

同花顺

- 模型结构

  VSR Encoder： Con3dResNet

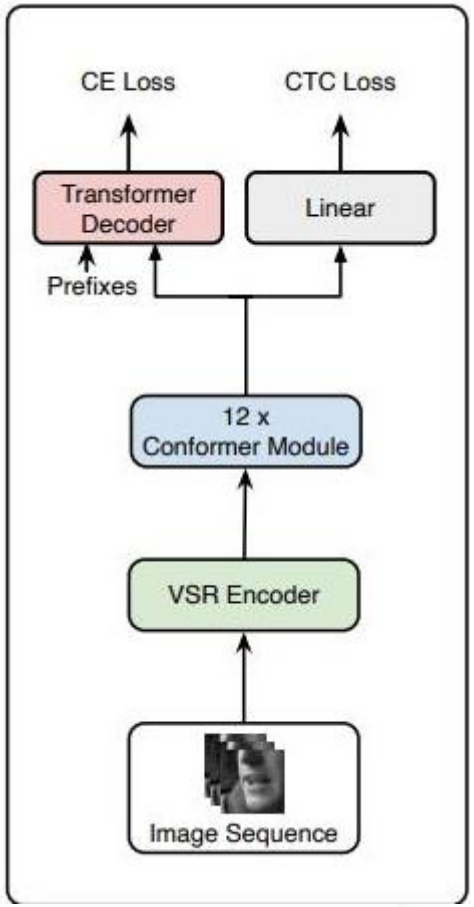
  Conformer Encoder：adim 768，ahead 12，eunits 3072，elayer 12 cnn kernel 31，

  Transformer Decoder： ddim 768，dhead 12，dunits 3072，dlayers 6

  建模单元：以汉字作为建模单元（4705个），代替bpe建模

  LSTM-LM： 6layers，1024 units

  Transformer-LM：6 layers，1024 units，embed-unit 128， att-unit：512



- 训练策略

  (1) Joint CTC/Attention training, CTC weight=0.1

  (2) 以model_avg_14_23_cncvs_4s.pth模型作为种子模型

  (3)第一阶段的训练：利用CNCVS 300hours训练模型

  (4) 第二阶段finetuning： 利用各自赛道的Dev以及扩充数据进行微调

三、模型融合与实验结果

同花顺

- 模型结果融合：
  利用ROVER方法对所有模型的解码候选文本进行融合处理

- 实验结果
  Eval： single-speaker：41.62%； multi-speaker：54.55%

Table 1: The results of our system

method	training set(hours)	finetuning set(hours)		single-speaker dev CER	multi-speakers dev CER
		T1	T2		
baseline	287	83.7	18	48.57%	58.77%
char model unit	287	83.7	18	43.59%	56.77%
+video aug	287	166	36	42.30%	54.74%
+RNN LM				42.18%	
+transformer LM				42.16%	
+model fusion				**41.50%**	

谢 谢

THANKS