System Description for T244 Team

数据

原始数据：CN-CVS，Single-dev, Multi-dev

数据增广策略：

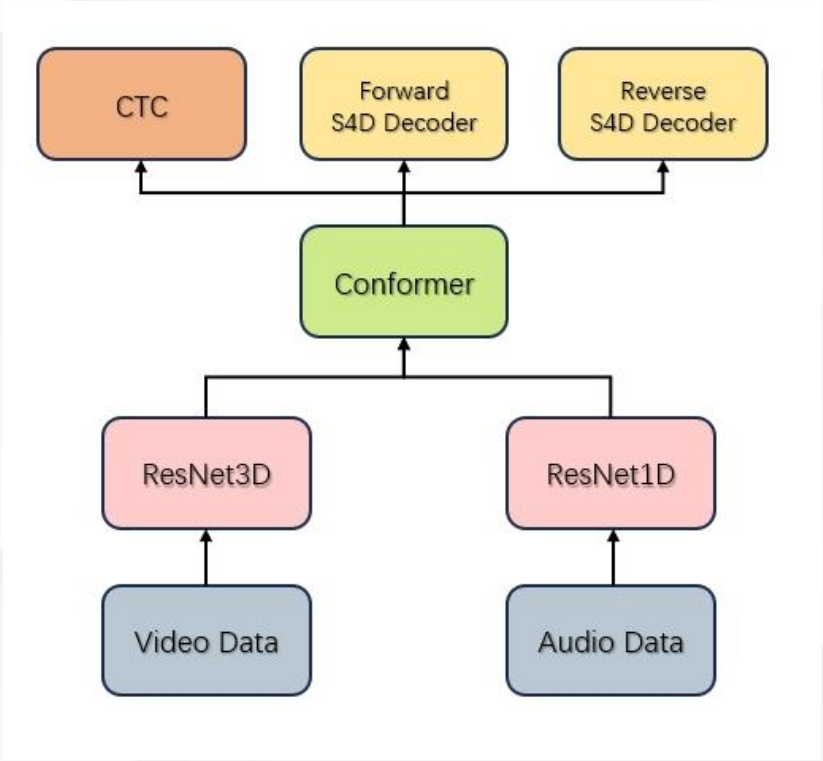
1.将视频速度调整为1.1倍和0.9倍
2.从96*96的唇部区域随机裁剪出88*88的区域
3.应用时间遮蔽

模型

与Baseline区别：

1.音视频共同输入

2.S4D Decoder



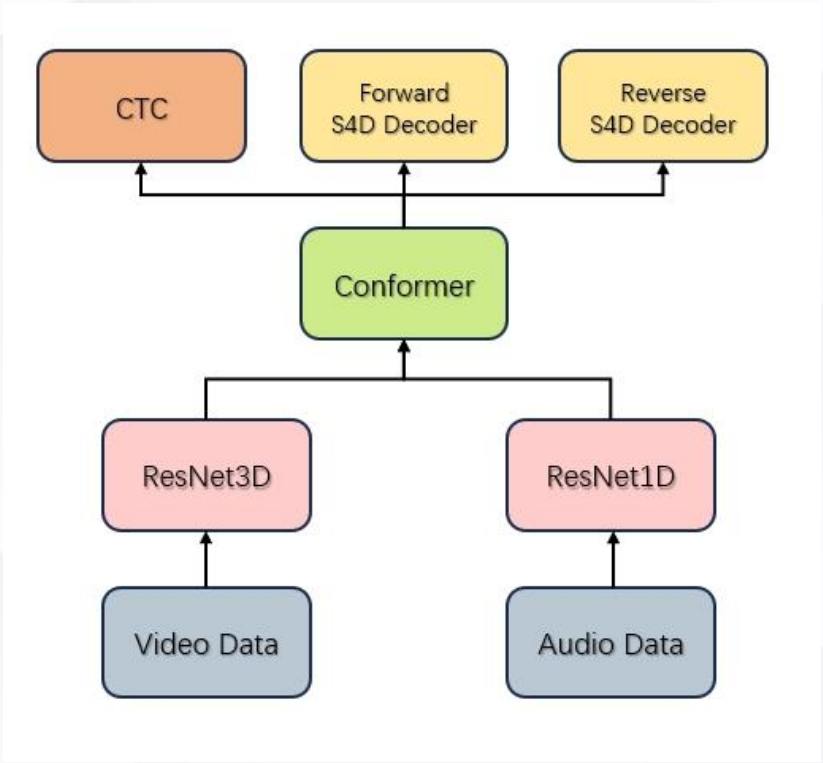
模型

音视频共同输入：

对音频和视频进行特征提取后，一同输入到Conformer模型

解决的问题：

与双向Decoder效果相近,但不再需要根据视频时长分步训练



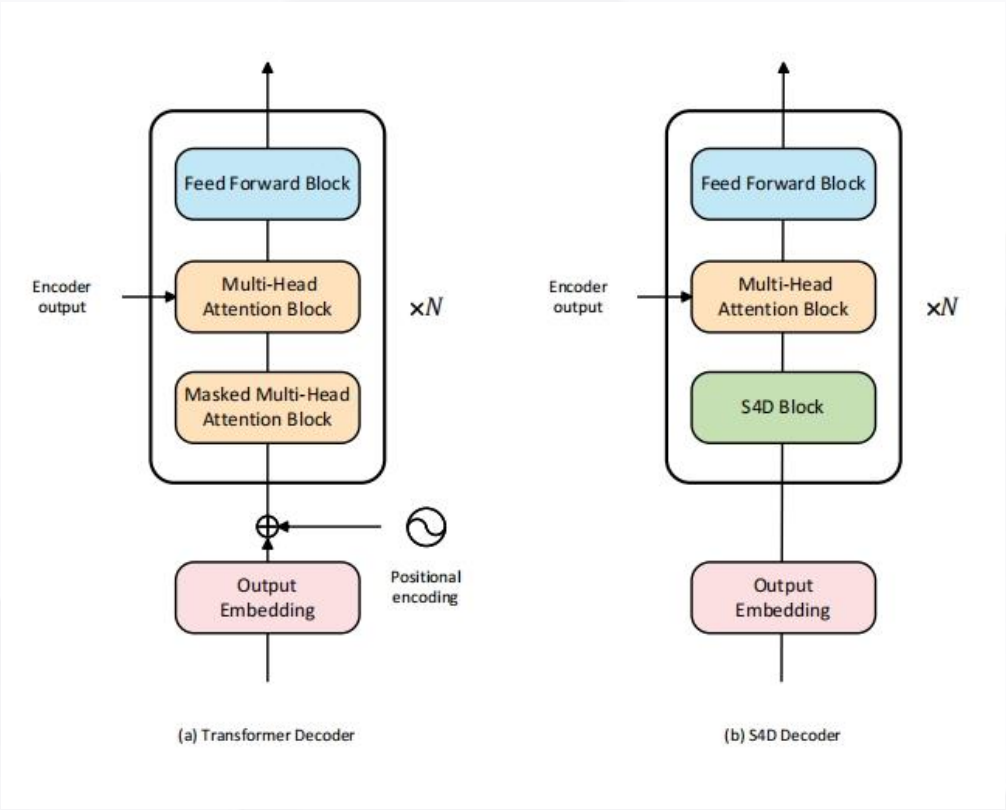
模型

S4D Decoder：

与Transformer Decoder相比去除位置编码，并用结构化状态空间模型 S4D 模型替换 Masked Multi-Head Attention Block

解决的问题：

- 掩码的设置一定程度上影响模型的训练
- S4D 模型对长文本的识别效果较好



结果

单说话人赛道（T1）：

Data	CER %
CN-CVS + Single-dev	39.31

多说话人赛道（T2）：

Data	CER %
CN-CVS + Multi-dev	47.92

内蒙古大学

INNER MONGOLIA UNIVERSITY

草原明珠 北疆风华

A SHINING PEARL ON THE GRASSLAND

THANKS