NPU-ASLP-LiAuto团队 CNVSRC 2023视觉语音识别系统描述

成员：**王贺**[1]、郭鹏程[1]、陈伟[2]、周盼[2]、谢磊[1]

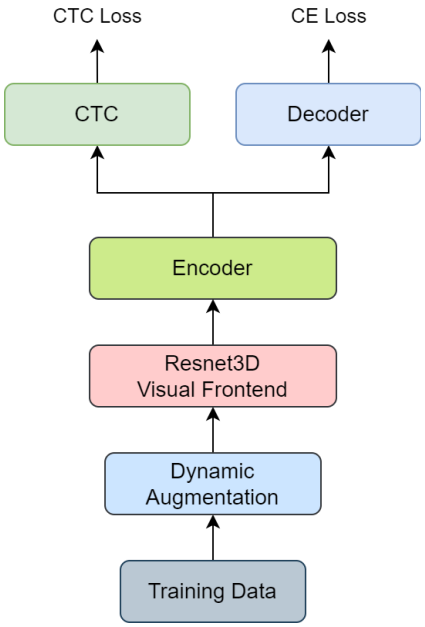
单位1：西北工业大学-音频语音与语言处理研究组（NPU-ASLP）

单位2：理想汽车（Li Auto）

目录 | CONTENT

概述

- **视觉语音识别定义**
  - 将说话人的连续图像信息（视频）自动识别为相应文本的技术



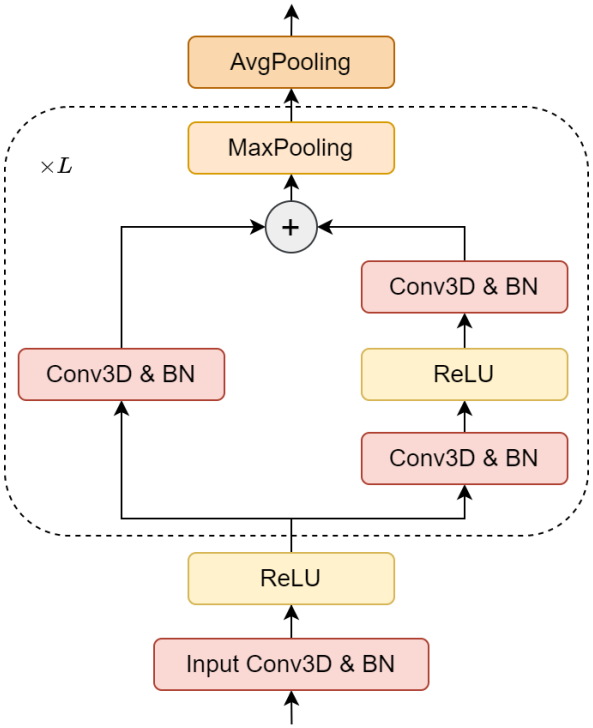
- **视觉语音识别系统**
  - 联合CTC/attention端到端架构
  - 主要由数据增广、视频前端、编码器和解码器组成

- **竞赛成绩（CER）**
  - 单人VSR任务
    - 验证集34.47%，测试集34.76%，fixed和open赛道第一名
  - 多人VSR任务
    - 验证集41.39%，测试集41.06%，open赛道第一名

目录 | CONTENT

视觉前端

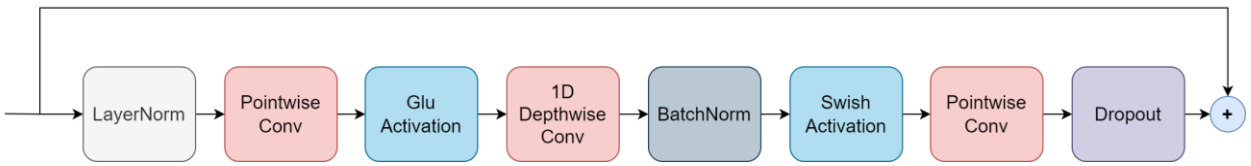
- **ResNet3D 视觉前端**
  - 灵感来源于经典网络ResNet并进行简化
  - 主要由三维卷积、最大池化和平均池化构成
  - 输入三维卷积将视频通道数投射到更高维度
  - 核心部分为5层ResNet Block堆叠
    - 每层分为两支独立的卷积模块
    - 每层输出通道数分别为32,64,64,128,256
    - 每层后MaxPooling对宽高维度进行2倍下采样
  - AvgPooling对宽高维度进行平均



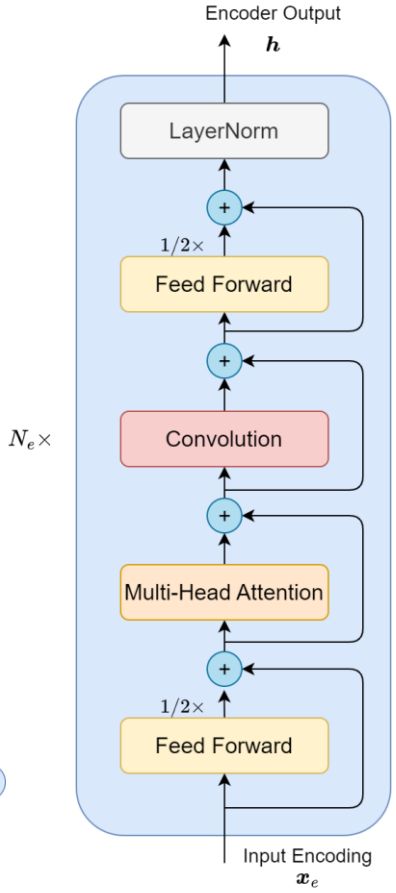
目录 | CONTENT

多种编码器探索

- **Conformer[1] 模型**
  - Multi-Head Self-Attention 模块用于建模全局信息，Convolution 模块用于学习局部信息
  - Feed Forward一分为二并引入LayerNorm，同时使用Swish激活函数替代Transformer中的Relu



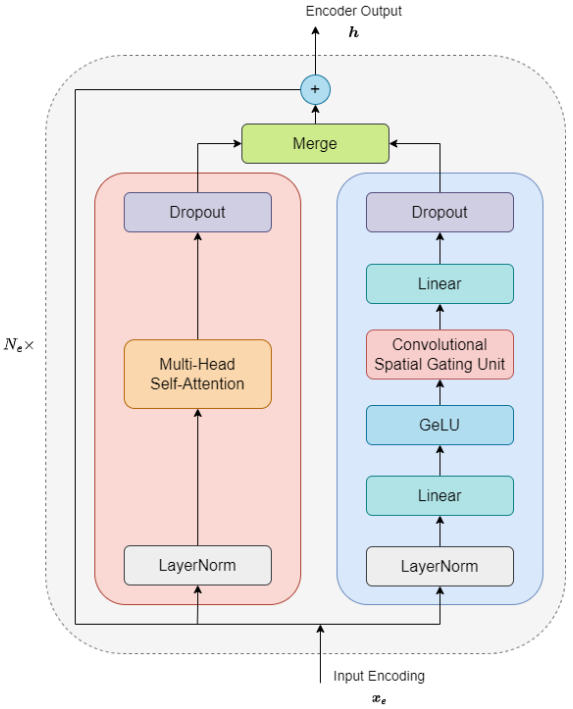
Convolution Module



[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmer, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, “Conformer: Convolution-augmented Transformer for speech recognition,” Interspeech, 2020

多种编码器探索

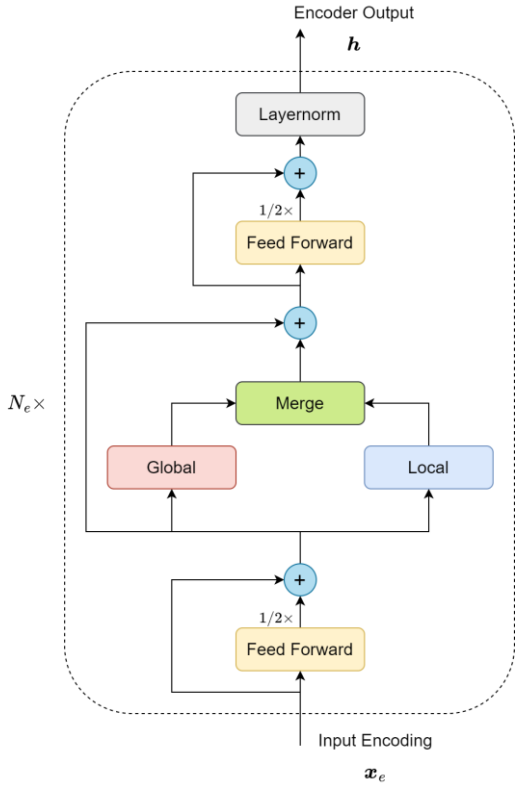
- **Branchformer[2] 模型**
  - 将全局信息和局部信息分为两个分支进行建模
  - 全局信息采用Multi-Head Self-Attention进行建模
  - 局部信息则采用使用线性映射和卷积模块进行建模
  - 最终两分支的结果会经过Merge模块完成合并
    - 两分支结果拼接后过线性映射



[2] Yifan Peng, Siddharth Dalmia, Ian Lane, et al., “Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding,” in Proc. ICML. PMLR, 2022, pp. 17627–17643.

多种编码器探索

- **E-Branchformer[3] 模型**
  - 整体结构与Branchformer一致，分两支进行全局和局部信息建模
  - 采用和Conformer类似的1/2 Feed Forward模块
  - 更有效的Merge模块
    - 两分支结果拼接后，先经过深度卷积，再与拼接结果残差连接，最后经过线性映射

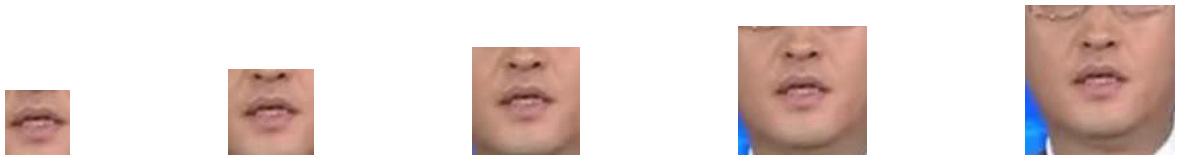


[3] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J Han, and Shinji Watanabe, “E-branchformer: Branchformer with enhanced merging for speech recognition,” in Proc. SLT. IEEE, 2023, pp. 84–91.

目录 | CONTENT

多尺寸数据建模与数据增广

- **多尺寸视频数据建模**
  - 使用baseline[1]提供的视频预处理工具生成48, 64, 80, 96, 112，5个尺度下的唇动数据，分别使用E-Branchformer模型进行训练。



size 48　size 64　size 80　size 96　size 112

CN-CVS[4]
sample n001_00001_001

[4] Chen Chen, Dong Wang, and Thomas Fang Zheng, “Cn-cvs: A mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis,” in Proc. ICASSP.IEEE, 2023, pp. 1–5.
1 https://github.com/MKT-Dataoceanai/CNVSRC2023Baseline

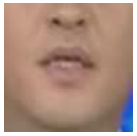
多尺寸数据建模与数据增广

- **数据增广**
  - 静态增广：利用moivepy[2]工具对视频数据进行0.9,1.0,1.1三倍速度扰动
  - 动态增广：训练过程中利用kornia[3]工具完成对每一个batch的数据进行随机水平翻转、旋转以及颜色变换（亮度、对比度、饱和度、色调、灰度处理）

speed 0.9

normal

speed 1.1



h-flip
color jiggle

rotation
gray scale

h-flip
rotation

h-flip
rotation
color jiggle

2 https://pypi.org/project/moviepy
3 https://github.com/kornia/kornia

目录 | CONTENT

数据集与实验设置

- **数据集**
  - CN-CVS+CNVSRC-Single.Dev / CNVSRC-Multi.Dev
- **实验设置**
  - 所有系统均采用ESPnet[4]工具包进行构建

Module	Block Layers	Attention Dim	Attention Heads	Feed Forward
Conformer	12	256	4	1024
Branchformer	24	256	4	1024
E-Branchformer	12	256	4	1024
Transformer (Decoder)	6	256	4	2048

VSR系统编码器和解码器主要参数设置
Block Layers: 基本模块堆叠次数；Attention Dim: 注意力维度；
Attention Heads：注意力头数；Feed Forward：前馈网络映射维度

4 https://github.com/espnet/espnet

不同编码器实验

- **不同编码器系统结果(CER %)**

ID	Encoder	Crop	SP	Single.Dev	Single.Eval	Multi.Dev	Multi.Eval
M1	Conformer	96	√	39.43	39.99	46.08	45.73
M2	Branchformer	96	√	39.00	39.36	46.63	46.37
M3	E-Branchformer	96	√	**38.59**	**38.61**	**46.26**	**45.80**

Crop: 训练所用视频尺寸；SP: 训练数据是否进行三倍变速；
Single.Dev/Eval：单人VSR任务的验证集/测试集；Multi.Dev/Eval：多人VSR任务的验证集/测试集；

- **分析**

  训练数据相同的情况下，三种编码器的性能呈现出如下趋势：
  E-Branchformer > Branchformer > Conformer

不同训练数据尺寸实验

- **不同训练数据尺寸系统结果(CER %)**

ID	Encoder	Crop	SP	Single.Dev	Single.Eval	Multi.Dev	Multi.Eval
M4	E-Branchformer	48	×	46.88	45.81	55.58	55.51
M5	E-Branchformer	64	×	44.40	43.59	53.64	52.98
M6	E-Branchformer	80	×	42.95	42.26	50.77	50.38
M7	E-Branchformer	96	×	40.56	40.42	47.16	46.53
M8	E-Branchformer	112	×	**38.46**	**38.95**	**45.17**	**44.87**

Crop: 训练所用视频尺寸；SP: 训练数据是否进行三倍变速；
Single.Dev/Eval：单人VSR任务的验证集/测试集；Multi.Dev/Eval：多人VSR任务的验证集/测试集；

- **分析**

  训练数据尺寸在48到112范围内，尺寸越大，VSR系统性能越好

多系统融合实验

- **Recognizer Output Voting Error Reduction[5](ROVER)**
  - 利用不同识别结果的出现频次来进行判决。

ID	Encoder	Crop	SP	Single.Dev	Single.Eval	Multi.Dev	Multi.Eval
M1	Conformer	96	√	39.43	39.99	46.08	45.73
M2	Branchformer	96	√	39.00	39.36	46.63	46.37
M3	E-Branchformer	96	√	38.59	38.61	46.26	45.80
M4	E-Branchformer	48	×	46.88	45.81	55.58	55.51
M5	E-Branchformer	64	×	44.40	43.59	53.64	52.98
M6	E-Branchformer	80	×	42.95	42.26	50.77	50.38
M7	E-Branchformer	96	×	40.56	40.42	47.16	46.53
M8	E-Branchformer	112	×	38.46	38.95	45.17	44.87
ROVER	-	-	-	**34.47**	**34.76**	**41.39**	**41.06**

[5] J. G. Fiscus, et al., “A post-processing system to yield reduced word error rate: Recognizer output voting error reduction (ROVER),” ASRU, 1997

总结

- **多编码器探索**
  - 使用目前主流的三种编码器进行VSR系统的构建
  - 单系统结果上 E-Branchformer > Branchformer > Conformer
- **多尺寸数据建模与数据增广**
  - 使用48,64,80,96,112，5个尺寸下的视频数据进行VSR系统构建
  - 在48 ~ 112范围内，尺寸越大，VSR系统效果越好
  - 数据增广包括静态增广（速度扰动）以及动态增广（水平翻转、旋转、颜色变换）
- **竞赛成绩**
  - 最终通过ROVER技术完成对8个系统结果的融合
    - 在单人VSR任务测试集上CER 34.76%，位列fixed和open赛道第一名
    - 在多人VSR任务测试集上CER 41.06%，位列open赛道第一名

谢谢聆听

Thank You

王贺

西北工业大学 音频语音与语言处理研究组

网址：www.npu-aslp.org

微信搜一搜

西工大音频语音与语言处理研究组