Lecture AI3611:智能感知认知实践

05 实践项目

Mengyue Wu & Xie Chen

Cross Media Language Intelligence Lab (X-LANCE)
Department of Computer Science & Engineering
Shanghai Jiao Tong University

2024

感知认知课程实践

- 单模态理解
  - 基于Vall-E模型的语音合成
  - 声音事件检测
  - 语言模型
  - 图像生成
- 多模态及跨模态交互
  - 图片摘要生成
  - 音视频场景识别

感知认知课程实践

6选4进入计分，可以自由选择

- 单模态理解
  - 基于Vall-E模型的语音合成
  - 声音事件检测
  - 语言模型
  - 图像生成
- 多模态及跨模态交互
  - 图片摘要生成
  - 音视频场景识别
- 实践项目单个21%，6选4，共84%
  - 按兴趣选择
  - 按自我最优提交

实践考核

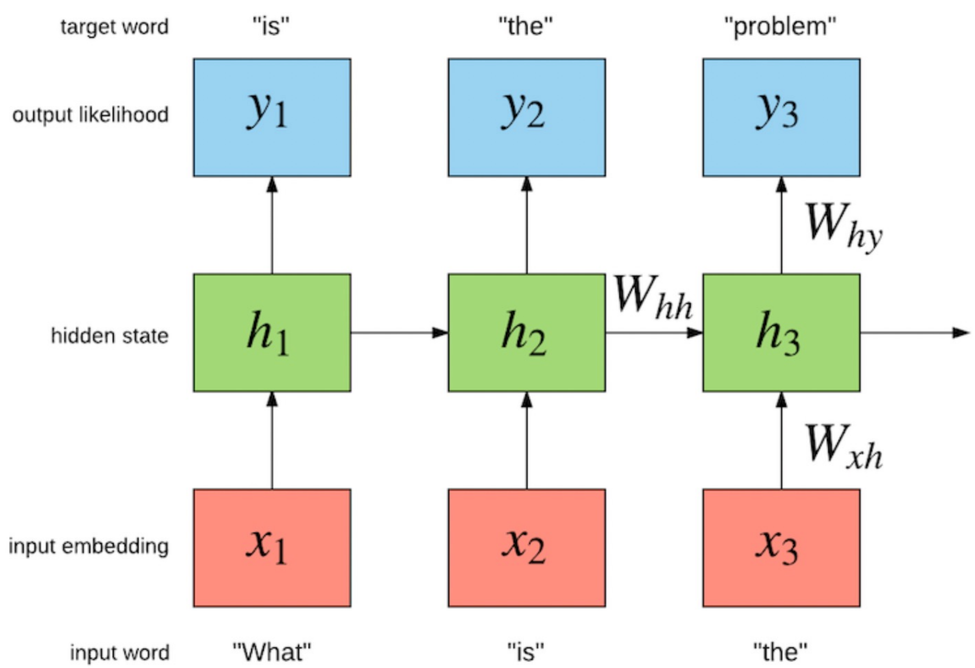
- 实践考核
  - 1. 项目报告 - 40%　清楚阐述实验过程，实验结果分析
  - 2. 模型性能 - 40%　模型创新程度，测试集性能
  - 3. 代码可读性 - 20%　逻辑清晰，易复现，注释等

- **语言模型**
- 语音合成
- 图像生成
- 声音事件检测
- 图片摘要生成
- 音视频场景识别

语言模型

- Perplexity (PPL)
  - the lower the better
  - average divergence of each prediction

$$PPL = \exp(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|h))$$



语言模型

- 基本要求
  - 使用不同模型结构，训练基于神经网络的语言模型（至少尝试Feedforward，RNN，GRU，LSTM，Transformer等结构中的三种不同神经网络模型结构）
  - 讨论和尝试不同超参数，对语言模型性能的影响
  - 模型总参数量不得超过60M，最优化测试集上的PPL

- 实践报告要求
  1. 回答基本要求中的各个要点
  2. 写出详细的实验过程和实验分析
  3. 提交代码，给出重现最优结果的脚本和配置

- 加分项目
  - 使用tensorboard画出训练阶段每个epoch的train, valid和test loss(或PPL）趋势
  - 解释和解决Transformer LM性能不如LSTM LM的问题
  - 可以使用相关文献中或Github中开放的更先进的语言模型，对本课程提供的数据持续提升性能

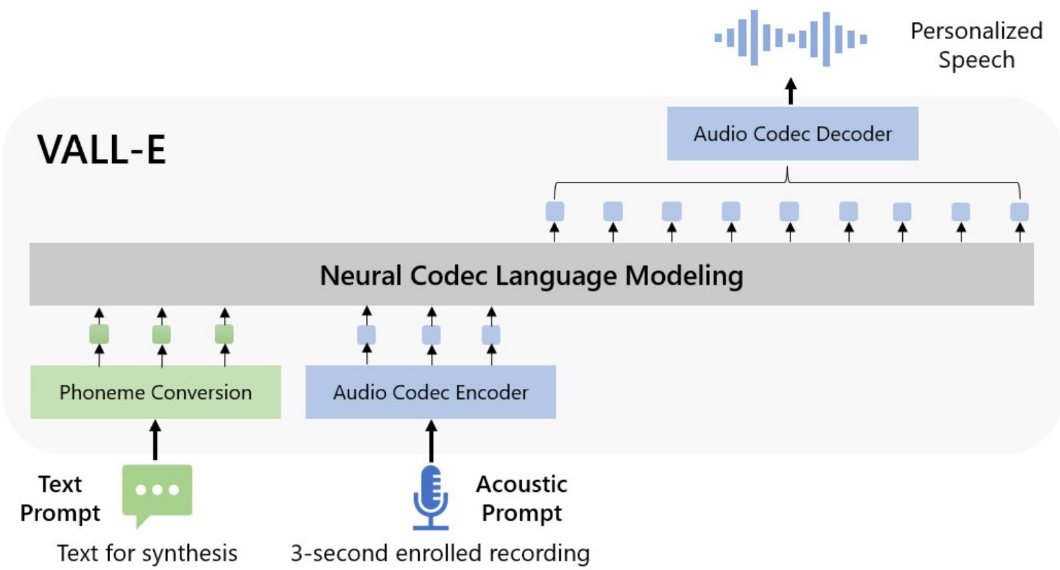
语言模型

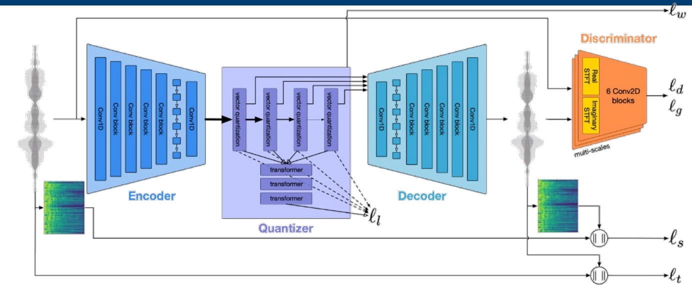
- 参考文献

  - Mikolov, Tomas, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. "Recurrent neural network based language model." In *Interspeech*, vol. 2, no. 3, pp. 1045-1048. 2010.
  - Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. "LSTM neural networks for language modeling." In *Thirteenth annual conference of the international speech communication association*. 2012.
  - Irie, Kazuki, Albert Zeyer, Ralf Schlüter, and Hermann Ney. "Language modeling with deep transformers." *arXiv preprint arXiv:1905.04226* (2019).

- 语言模型
- **语音合成**
- 图像生成
- 声音事件检测
- 图片摘要生成
- 音视频场景识别

语音合成 – Vall-E模型





- **This outward mutability** indicated, and did not more than fairly express, the various properties of her inner life.

Prompt VALL-E

- **This she said was true hospitality,** and I am not sure that I did not agree with her.

Prompt VALL-E

Wang, Chengyi, et al. "Neural codec language models are zero-shot text to speech synthesizers." *arXiv preprint arXiv:2301.02111* (2023).
Défossez, Alexandre, Jade Copet, Gabriel Synnaeve, and Yossi Adi. "High fidelity neural audio compression." *arXiv preprint arXiv:2210.13438* (2022).

语音合成

- 基本要求
  1. 基本理解代码逻辑，熟悉基于Vall-E语音合成过程
  2. 得到不同训练数据规模下（100，360和960小时）的语音合成性能对比
  3. 各模型参数总数量不超过500M（现有例子模型参数已超过60M）
  4. 评价指标：最小化词错误率 （WER）；说话人相似性
     1. 基于已有的语音识别系统评测识别性能
     2. 基于已有的说话人识别模型评价说话人相似性
- 实践报告要求
  1. 回答基本要求中的各个要点
  2. 写出详细的实验过程和实验分析
  3. 提交代码，给出重现最优结果的脚本和配置
- 加分项
  - 尝试使用已有的中文语音数据，类似思路搭建一个中文语音合成系统
  - 尝试不同的neural codec模型，提高合成语音质量
  - 尝试不同模型结构或算法，提高模型稳定性或语音质量

注：该题目对算力有一定要求，使用学校集群，可能会导致超过最长时长，需要断点

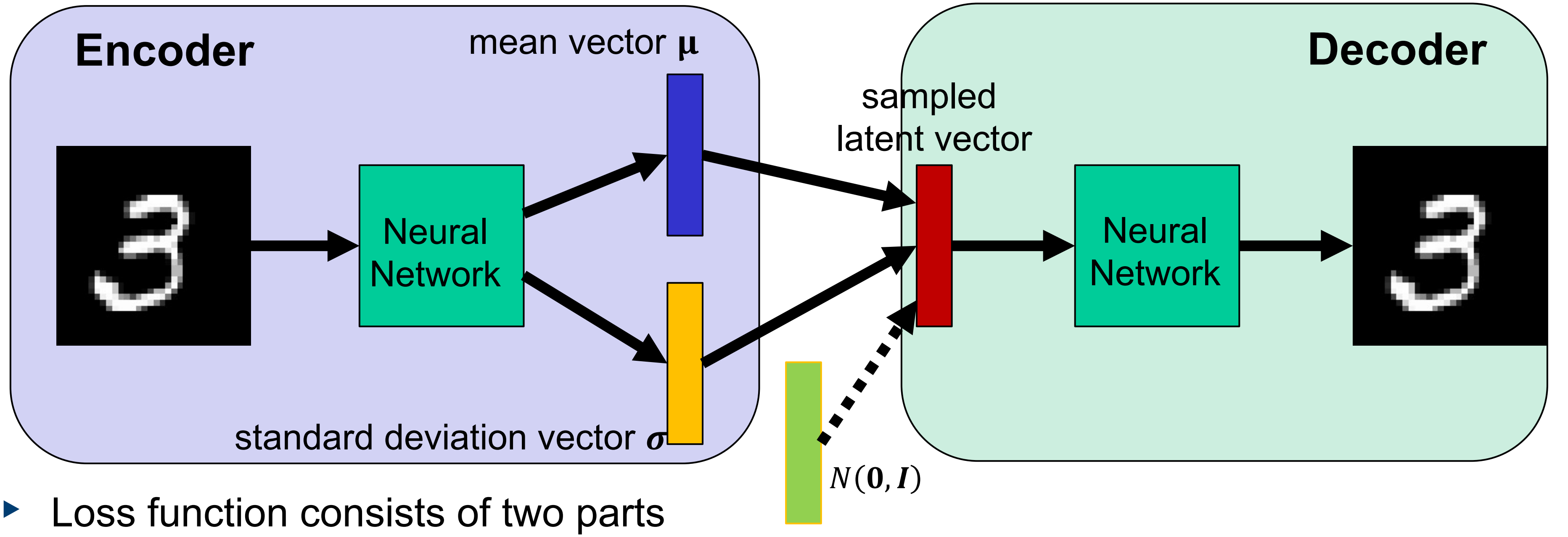
语音合成

- 参考文献

  - Défossez, Alexandre, Jade Copet, Gabriel Synnaeve, and Yossi Adi. "High fidelity neural audio compression." *arXiv preprint arXiv:2210.13438* (2022).
  - Wang, Chengyi, et al. "Neural codec language models are zero-shot text to speech synthesizers." *arXiv preprint arXiv:2301.02111* (2023).
  - Kumar, Rithesh, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. "High-fidelity audio compression with improved rvqgan." Advances in Neural Information Processing Systems 36 (2024).

- 语言模型
- 语音合成
- **图像生成**
- 声音事件检测
- 图片摘要生成
- 音视频场景识别

图像生成 – Variational AutoEncoders (VAE)



- Loss function consists of two parts
  - regularization loss: KL-distance between $N(\boldsymbol{\mu}, \boldsymbol{\sigma})$ *and* $N(\mathbf{0}, \boldsymbol{I})$
  - reconstruction loss: reconstruct the image in the output layer of decoder
- Notes and tips:
  - KL distance between two Gaussian distributions: $KLD(p,q) = log\,\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2+(u_1-u_2)^2}{2\sigma_2^2}$
  - The error (or gradient) not back propagation from sampling (dotted line)
  - The sampled latent vector used the $\boldsymbol{\mu}, \boldsymbol{\sigma}$ and samples from $N(\mathbf{0}, \boldsymbol{I})$ to generate the input for decoder. only $\boldsymbol{\mu}, \boldsymbol{\sigma}$ connection used for error backpropagation
  - value in MNIST used here are binary value 0 or 1, consider using BCELoss in Pytorch

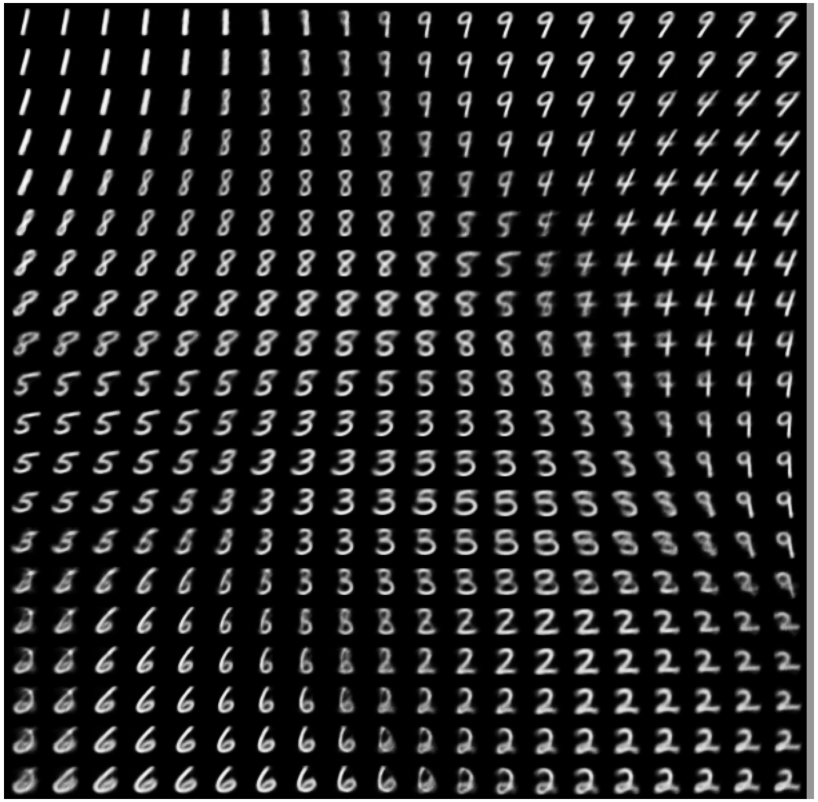
图像生成

- 自由发挥题
  - 提供MNIST数据集，和data loader代码，不提供模型训练样本代码
  - 不限模型结构，模型大小
  - 可以借鉴Github代码，但不要照搬，发现N份雷同代码(N>1), 每份雷同代码作业扣N分

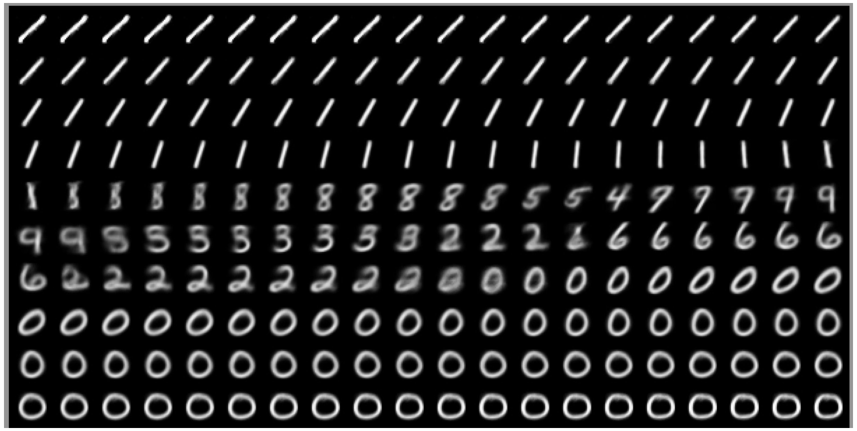
- 要求和目标
  - 给定MNIST数据集（train和valid），自定义encoder和decoder，实现VAE训练
  - 将隐层向量z维度设置为1，比较VAE训练完成后不同的z值对应的生成图片效果
  - 将隐层向量z维度设置为2，找出隐层向量的两个维度[-5, 5]值区间内对应的图片生成效果
  - 最优化valid数据集的重构误差

- 报告要求
  - 根据自己的理解， 描述VAE的原理，可用数学公式，或图片，或文字
  - 写出详细的实验过程和实验分析
  - 提交代码（需包含代码注释，方便阅读），给出重现最优结果的脚本和配置
  - 鼓励加上你觉得这个模型有意思的观察或思考！

2-dim gaussian



1-dim gaussian



- 参考文献
  - Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).
  - Doersch, Carl. "Tutorial on variational autoencoders." *arXiv preprint arXiv:1606.05908* (2016).

课程作业问题联系人

授课教师： 陈谐 邮箱： chenxie95@sjtu.edu.cn
助教： 马子阳 邮箱： zym.22@sjtu.edu.cn

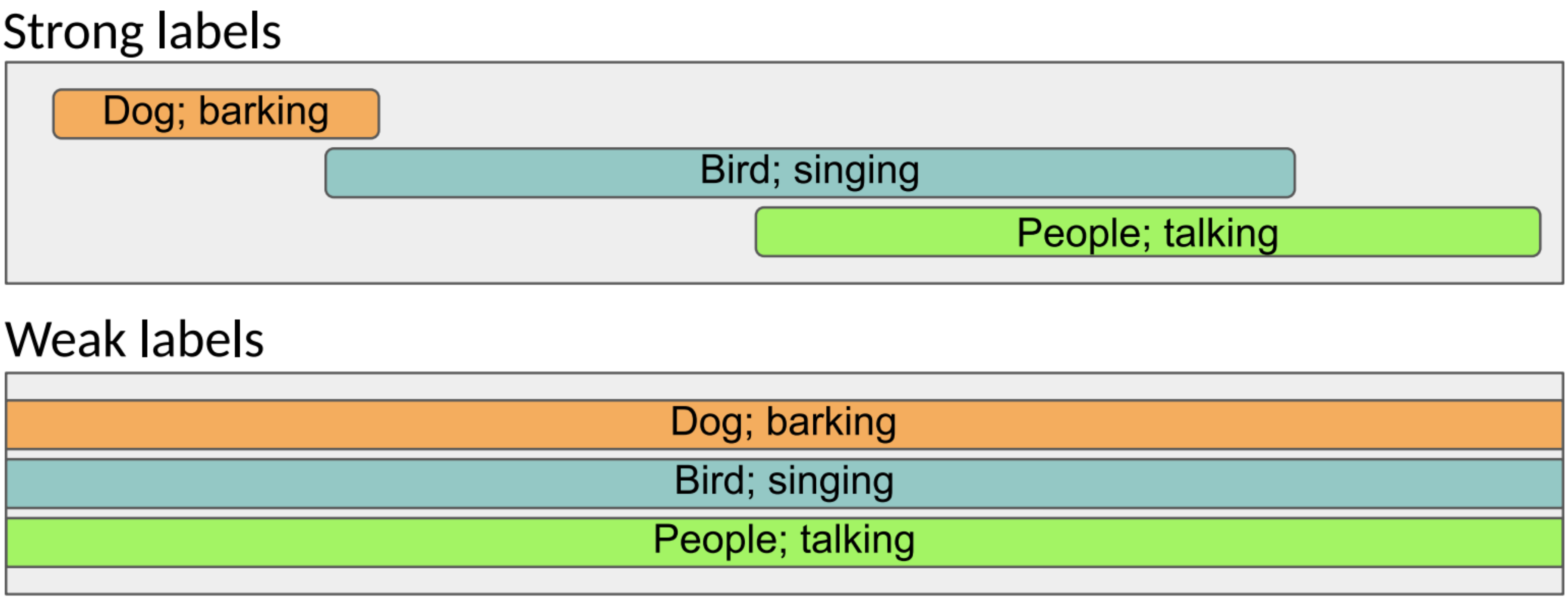
微信可从课程微信群里添加

如有相关问题，建议请随时与我们联系！

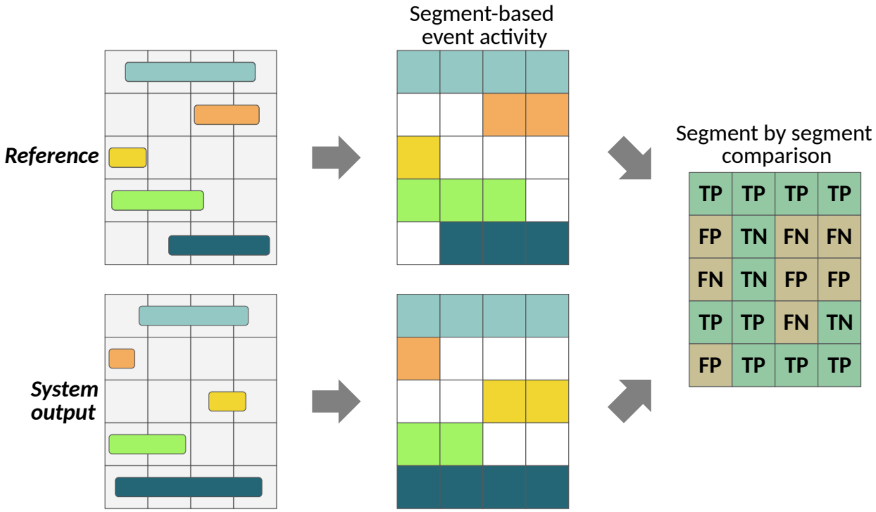
- 语言模型
- 语音合成
- 图像生成
- **声音事件检测**
- 图片摘要生成
- 音视频场景识别

声音事件检测

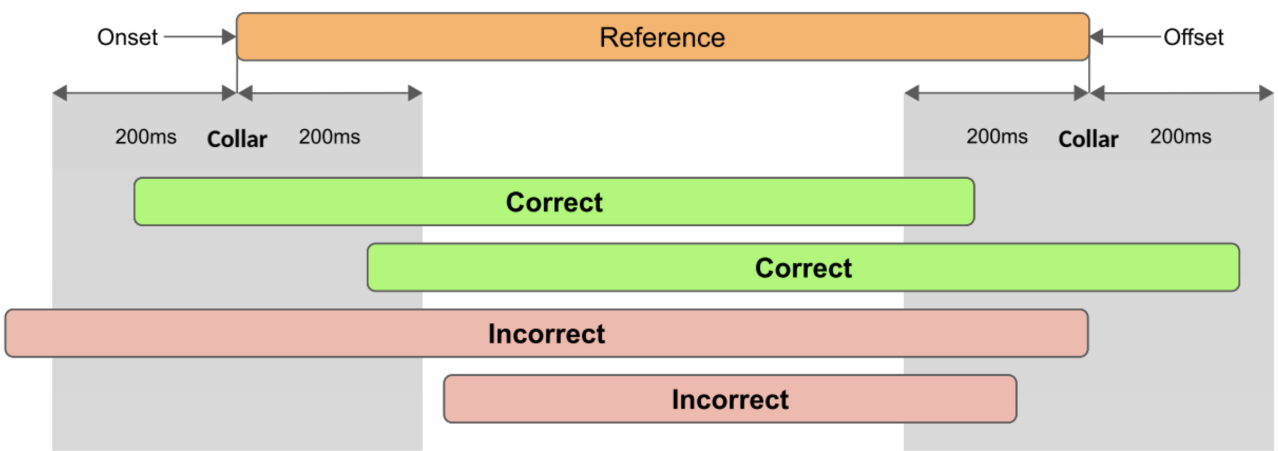
- Weakly-labelled sound event detection
- Output: tagging (classification) and boundary detection



- Metrics: F1 score，Tagging evaluated by segment- and **event- F1** $F = \frac{2PR}{P+R}$



Segment-level F1

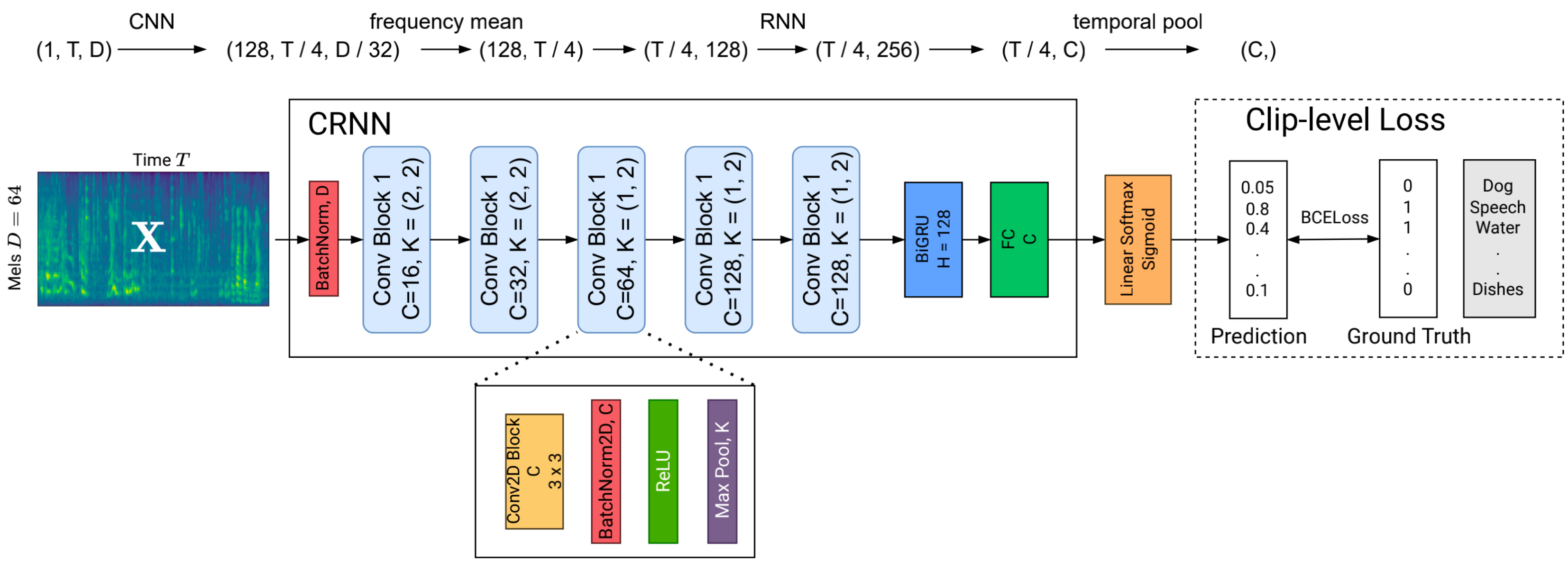


Event-level F1

声音事件检测

- Data: DCASE18
- Baseline: CRNN
- Feature: LMS 64d
- Loss: Binary cross entropy (BCE):

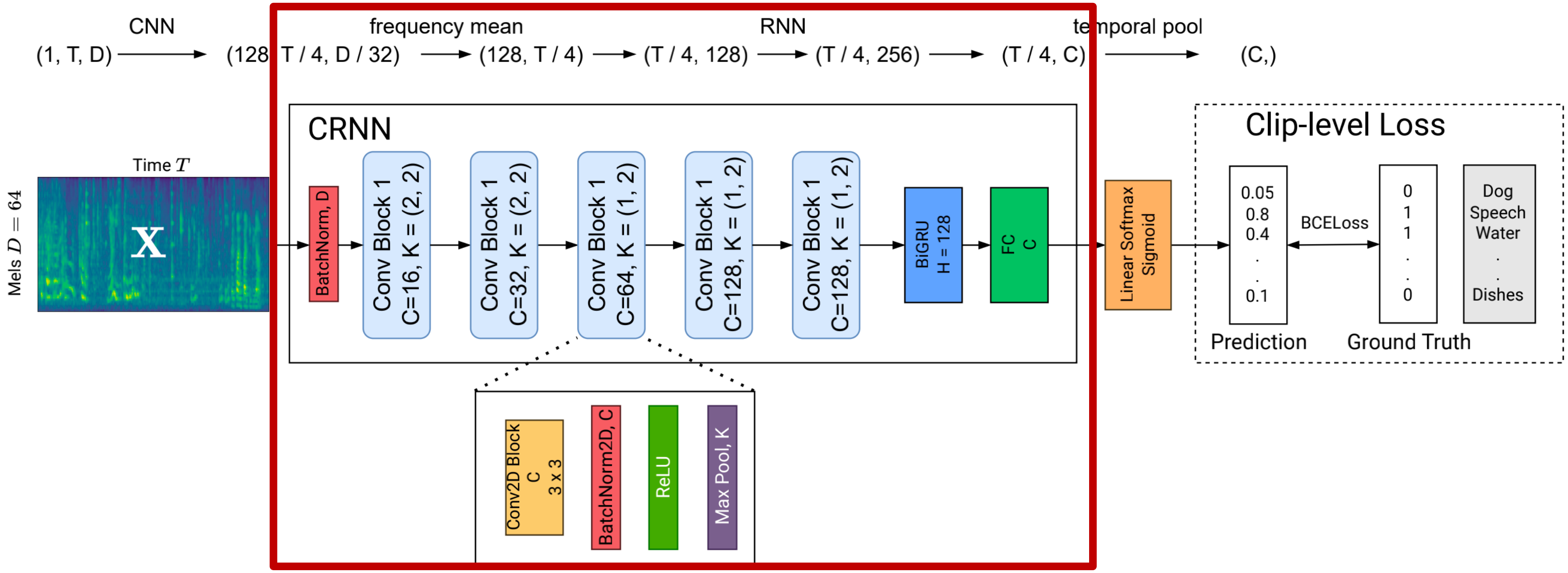
$$\mathcal{L}(y, \hat{y}) = -\hat{y}\log(y) + (1-\hat{y})\log(1-y)$$



声音事件检测

- Data: DCASE18
- Baseline: CRNN
- Feature: LMS 64d
- Loss: Binary cross entropy (BCE):

$$\mathcal{L}(y,\hat{y}) = -\hat{y}\log(y) + (1-\hat{y})\log(1-y)$$



声音事件检测

基本要求

1. 理解代码逻辑，熟悉声音事件检测模型
2. 了解弱监督情况下进行时间轴预测的难点，以及基线模型设计的原理
3. 按模型框架实现CRNN模型

高阶要求

1. 可以修改模型深度、参数以及超参数，优化模型结果
2. 调研学习音频中的数据增强方法以应用

实践报告要求

1. 回答基本要求中的各个要点，若有额外工作，写明白修改的地方和额外的方法
2. 写出详细的实验过程和实验分析
3. 提交代码，给出重现最优结果的脚本和配置

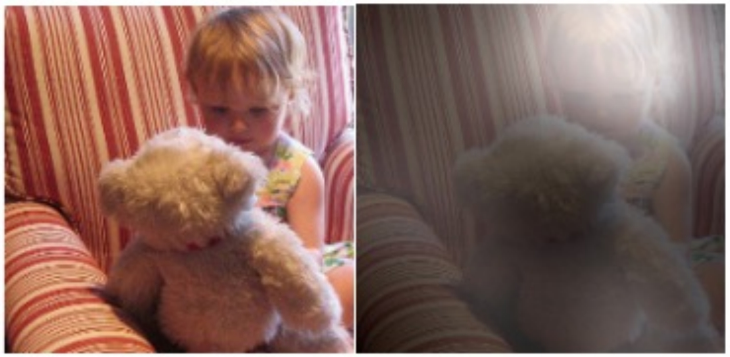
- 参考文献

  - Cao, Yin, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, and Mark D. Plumbley. "Polyphonic sound event detection and localization using a two-stage strategy." arXiv preprint arXiv:1905.00268 (2019).
  - Mesaros, Annamaria, Toni Heittola, Tuomas Virtanen, and Mark D. Plumbley. "Sound event detection: A tutorial." IEEE Signal Processing Magazine 38, no. 5 (2021): 67-83.
  - Dinkel, Heinrich, Mengyue Wu, and Kai Yu. "Towards duration robust weakly supervised sound event detection." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 887-900.

- 语言模型
- 语音合成
- 图像生成
- 声音事件检测
- **图片摘要生成**
- 音视频场景识别

图片摘要生成

- A modality translation task
- Input: images; Output: natural language to describe the image
- Supervision signal: human-written captions, 5 captions/image



A little girl sitting on a bed with a teddy bear.

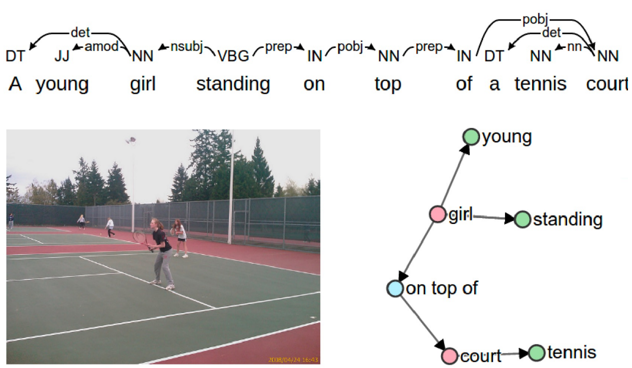
A group of people sitting on a boat in the water.

图片摘要生成

Evaluation: language generation metrics and image-language metrics

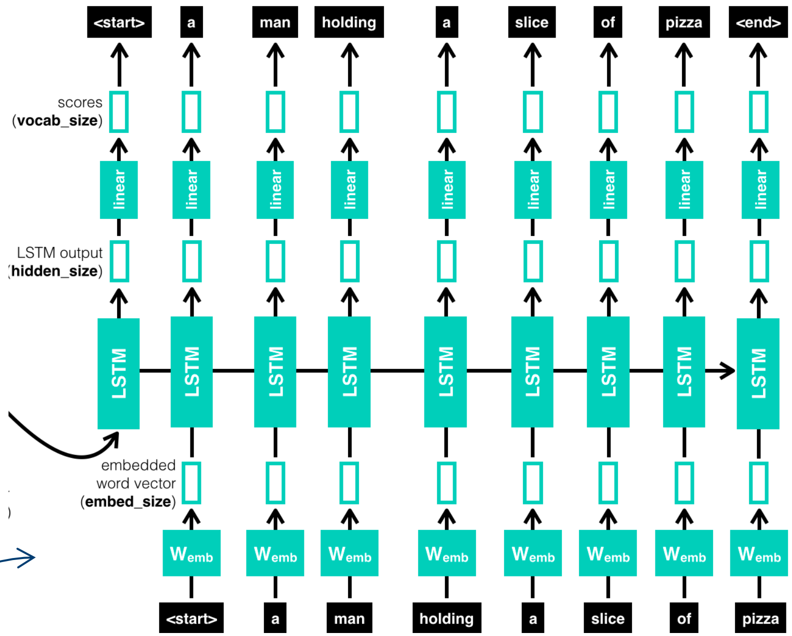
- BLEU4, 4-gram overlap precision
- METEOR: 衡量生成词语的 F1
  - 对词语做了lemmatization (speaking -> speak) 并用 wordnet 映射到对应的概念
- CIDEr: 把句子表示成各个词语的 TF-IDF 组成的向量，计算 reference 和 candidate 向量的 cosine similarity
- SPICE: 把句子转化成从 scene graph 中提取的 tuple set，计算 reference 和 candidate tuple set 的 F1

- **SPIDEr:** (SPICE+CIDEr)/2

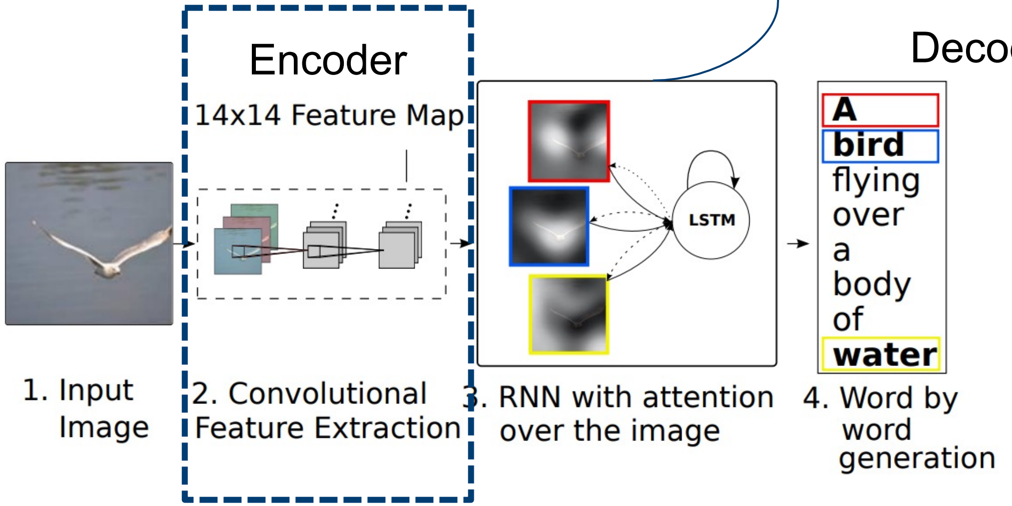


$$\{\ (\text{girl}),\ (\text{court}),\ (\text{girl, young}),\ (\text{girl, standing}) \\ (\text{court, tennis}),\ (\text{girl, on-top-of, court})\ \}$$

- Dataset: Flicker8k
- Baseline: CNN-LSTM with attention
- Encoder output: 14×14×512 feature map of the 4th convolutional layer before maxpool



Decoder Mechanism



图片摘要生成

- 基本要求
  1. 理解跨模态模型对于不同模态数据处理的方式，编码器-解码器模型的整体框架
  2. 实现scheduled sampling (知道在哪改，怎么改）*
  3. 了解对于跨模态生成任务不同指标衡量的目的，打印生成最好和最坏的指标样例比较客观&主观评测

- 高阶要求
  1. 通过修改代码和超参数进行模型调优，改进模型性能
  2. 可以使用额外数据以及预训练模型

- 实践报告要求
  1. 回答基本要求中的各个要点，若有额外工作增加详细做法说明
  2. 写出详细的实验过程和实验分析，分析不同指标下的生成样例差异，与个人主观评测进行对比
  3. 提交代码，给出重现最优结果的脚本和配置

*Scheduled sampling: Bengio, Samy, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. "Scheduled sampling for sequence prediction with recurrent neural networks." Advances in neural information processing systems 28 (2015).

图片摘要生成

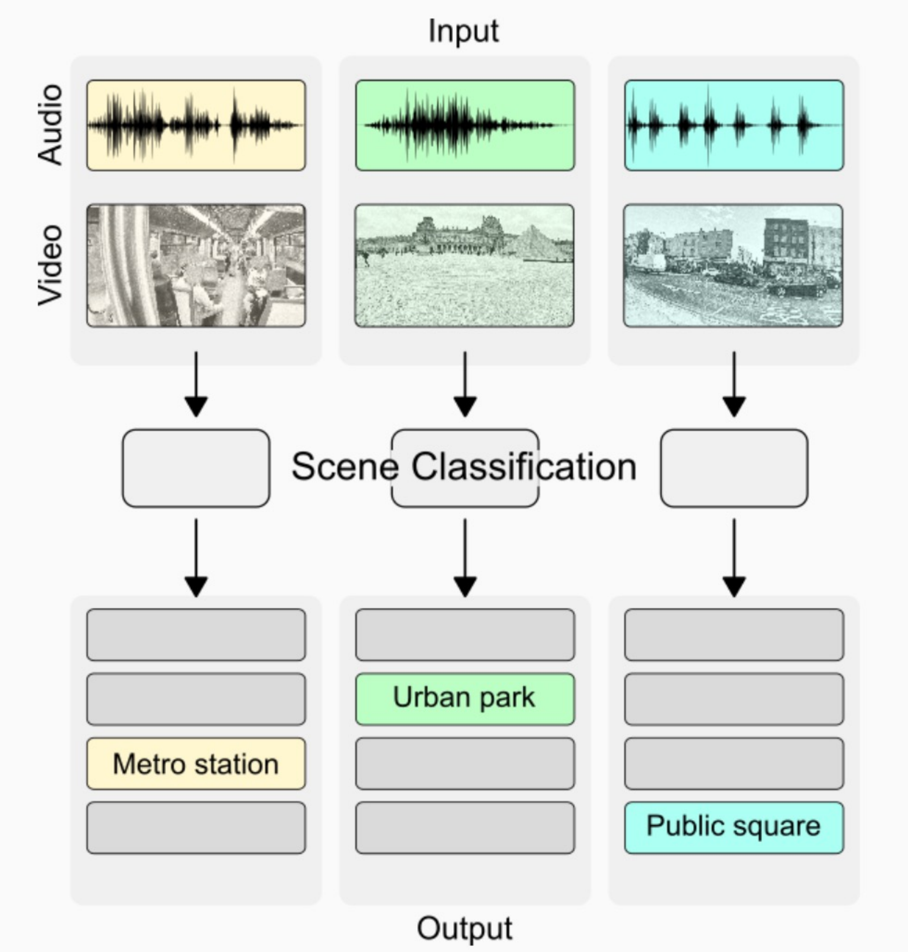
- 参考文献

  - Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In International conference on machine learning, pp. 2048-2057. PMLR, 2015.
  - Wang, Haoran, Yue Zhang, and Xiaosheng Yu. "An overview of image caption generation methods." Computational intelligence and neuroscience 2020.
  - You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image captioning with semantic attention." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4651-4659. 2016.

- 语言模型
- 语音合成
- 图像生成
- 声音事件检测
- 图片摘要生成
- **音视频场景识别**

音视频场景识别

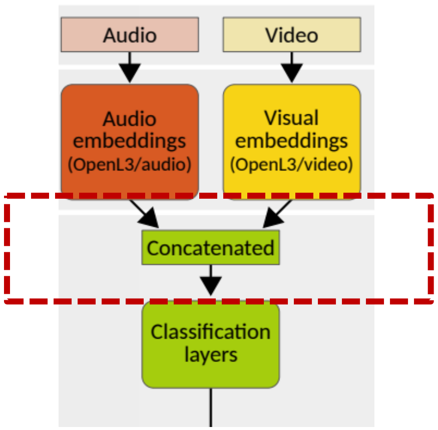
- Multi-modal representation learning
- Input: video and audio
- Ouput: a uni-label scene classification result

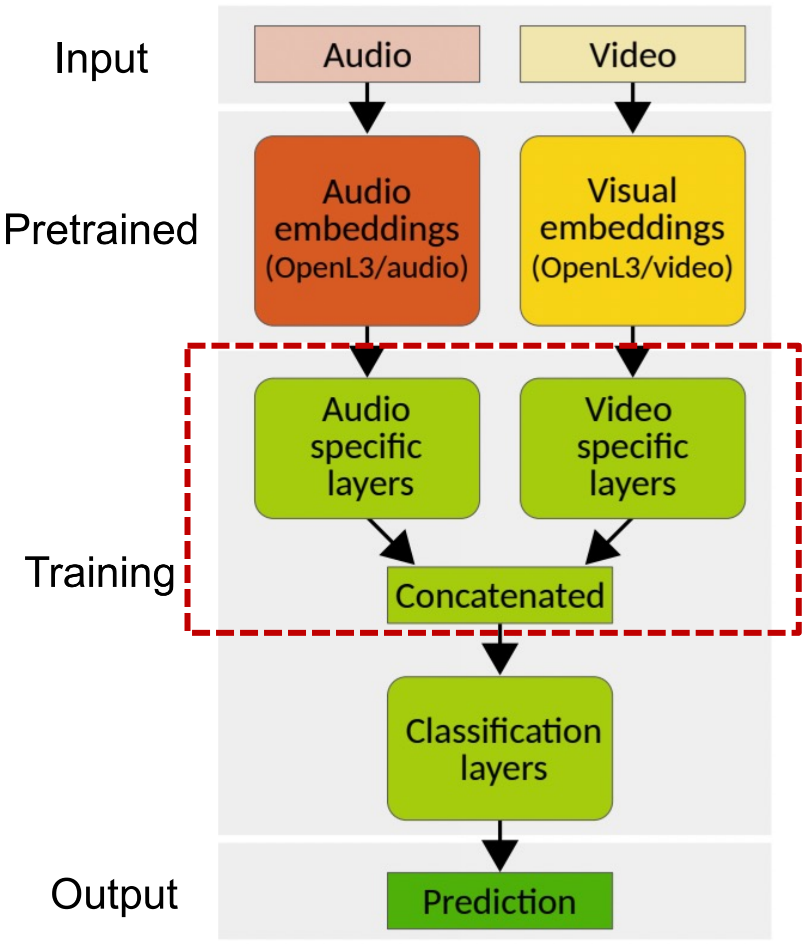


1. Airport
2. Indoor shopping mall
3. Metro station
4. Pedestrian street
5. Public square
6. Street with medium level of traffic
7. Travelling by a tram
8. Travelling by a bus
9. Travelling by an underground metro
10. Urban park

音视频场景识别

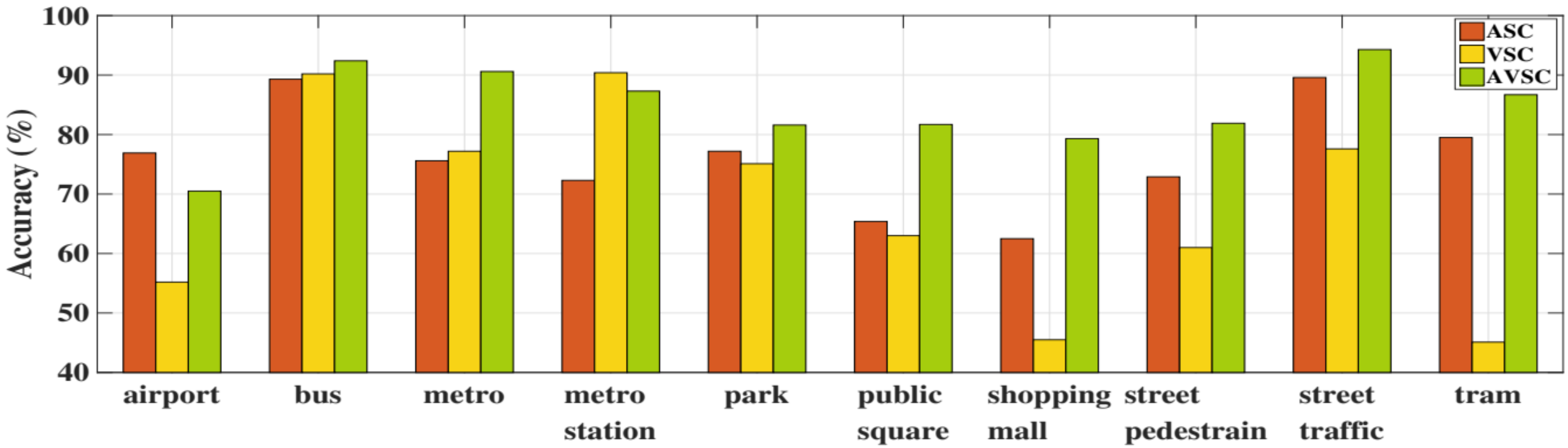
- Pretraining can be a useful tool for multi-modal representation: baseline utilized pretrained audio and video embeddings
- Feature concat is the most basic modality fusion method, however we can also involve separate audio and video networks first to learn task-specific representations





- Evaluation metrics:
  - Audio-only, Video-only, A/V combined
  - Class-wise Results
  - LogLoss

$$L_{\log}(y,p) = -(y\log(p) + (1-y)\log(1-p))$$



音视频场景识别

- 基本要求
  1. 理解多模态融合工作不同表征获取的方式、模态融合的方式
  2. 分析有冲突的模态结果，i.e. 在何种类别上多模态融合更有效，为什么
  3. 替换为early特征融合和late决策融合的两类融合方式

- 高阶要求
  1. 通过替换特征、修改模型、和超参数进行模型调优，改进模型性能

- 实践报告要求
  1. 回答基本要求中的各个要点，若有额外工作提供详细说明
  2. 写出详细的实验过程和实验分析
  3. 提交代码，给出重现最优结果的脚本和配置

- 参考文献

  - https://dcase.community/challenge2021/task-acoustic-scene-classification#subtask-b
  - Wang, Shanshan, Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. "Audio-visual scene classification: analysis of DCASE 2021 Challenge submissions." arXiv preprint arXiv:2105.13675 (2021).
  - Wang, Shanshan, Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. "A curated dataset of urban scenes for audio-visual scene analysis." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 626-630. IEEE, 2021.

课程作业问题联系人

授课教师： 吴梦玥 邮箱： mengyuewu@sjtu.edu.cn

- 助教： 谢泽宇 邮箱： zeyu_xie@sjtu.edu.cn

Github QA page:

https://github.com/chenxie95/deeplearning_course_sjtu/issues

按疑问需求设置线上/线下统一答疑时间

数据和代码

- 语言模型、语音识别、VAE图片生成
  - 代码：https://github.com/chenxie95/deeplearning_course_sjtu
  - 数据：/lustre/home/acct-stu/stu281/deeplearning_course_sjtu
- 音视频场景识别、声音事件检测、图片摘要生成：
  - 代码：/lustre/home/acct-stu/stu282/Project/{av_scene_classify，sound_event_detection，image_captioning}
  - 数据：/lustre/home/acct-stu/stu282/Data/{av_scene_classify，sound_event_detection，image_captioning}
  - 环境："conda env create -f /lustre/home/acct-stu/stu282/Project/env.yaml"

“交我算”平台使用

- 登录节点
  - 使用 SSH 登录:
    [local] $ ssh [username]@pilogin.hpc.sjtu.edu.cn
  - 写入 ~/.ssh/config

```
Host pi
    HostName pilogin.hpc.sjtu.edu.cn
    User [username]
    ServerAliveInterval 240
```

    登录命令:
    [local] $ ssh pi
  - 免密登录 (Linux / Mac)
    - 生成私钥: [local] $ ssh-keygen
    - 发送私钥到登录节点: [local] $ ssh-copy-id pi

“交我算”平台使用

- 数据传输
  - 基本命令: $ scp [source] [destination]
  - 例: 从超算上拷贝数据到本地:
    [local] scp pi:~/data.txt ./
- 终端复用：Tmux
  - 建立新的 session: [pi] tmux new -s [session_name]
  - 暂离session: Ctrl+b d (先同时按Ctrl和b，再按d，下同)
  - 重新进入session: [pi] tmux a -t [session_name]
  - 切换session: Ctrl+b s

"交我算" 平台使用

- 用 Slurm 系统提交作业
  - 交互式作业 srun / salloc:
    - srun: [pi] $ srun -N 1 -p small --pty /bin/bash
    - salloc: [pi] $ salloc -N 1 -p small
    - 随后登录分配到的节点
- 提交作业脚本 sbatch:
  - 写好脚本，如 run.sh:

```
#!/bin/bash
#SBATCH --job-name hostname
#SBATCH -p small
#SBATCH -N 1
/bin/hostname
```

  - 提交任务: [pi] $ sbatch run.sh
  - 该作业与 [pi] $ srun --job-name hostname -p small -N 1 /bin/hostname 一致
- 提交 gpu 任务:
  - 指定队列: -p dgx2
  - 申请一块GPU: --gres gpu:1

“交我算”平台使用

使用注意事项：

- 为避免计算资源浪费，教学账号限制作业运行数量1个、核数10个、GPU卡数1卡、最长运行时间12小时。请不要在登录节点运行作业，否则将会被封禁！！！教学支撑gpu队列为dgx2（单卡拥有32G显存）。目前集群GPU资源紧张，可能会出现排队的现象，请妥善安排作业提交时间，并且不要运行与课程无关的任务。
- 集群状态查询：https://status.hpc.sjtu.edu.cn/

相关文档：

- 登录：https://docs.hpc.sjtu.edu.cn/login/index.html
- 作业提交：https://docs.hpc.sjtu.edu.cn/job/index.html

如有任何问题，请联系助教。

课程论坛：https://github.com/chenxie95/deeplearning_course_sjtu/issues