用于声学信号内容理解的机器学习算法前瞻

李圣辰
西交利物浦大学
SHENGCHEN.LI@XJTLU.EDU.CN

Xi'an Jiaotong-Liverpool University
西交利物浦大学

音频内容理解

- 所谓音频内容理解系统，即利用机器学习的方法，通过分析音频的波形，对音频中含有的信息进行分析，完成指定任务。
- 常见的音频内容理解任务包括：
  - 特定声音检测：判断特定声音的出现（检测）
  - 声学场景分类：判断声音采录时所处的环境（识别）
  - 音频自动描述：根据声音内容生成相应的文字叙述（描述）
  - 自动报警：根据采集到的声音内容进行实时分析并在必要时报警（回应）

音频内容处理的优势与难点

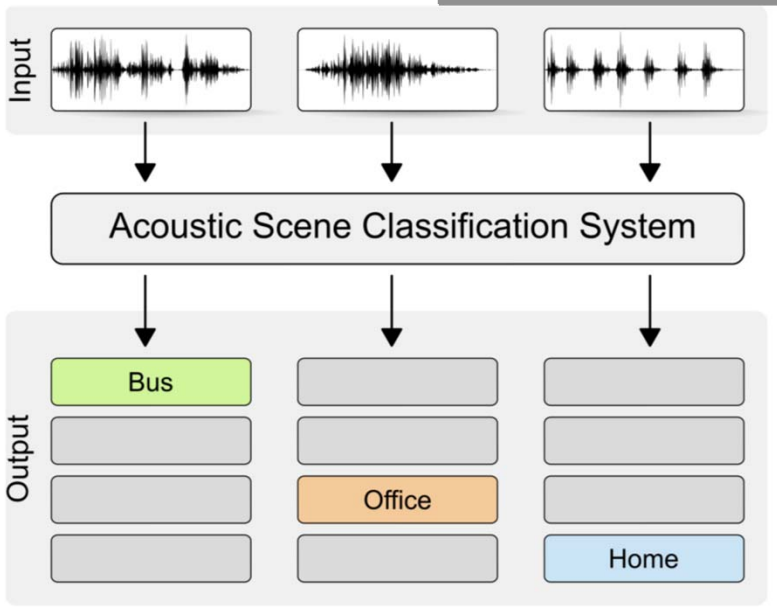
- 优势
  - 隐私性相对较好
  - 低功耗潜力高
  - 对客观条件要求较低
- 难点
  - 人类听觉感知相对敏感
  - 音频信息量相对较大
  - 音频缺乏驻在性，难以准确描述

DCASE 数据挑战赛

- 2013年在英国玛丽女王大学（Queen Mary University of London）数字音乐研究中心（Centre for Digital Music, C4DM）时任主任Mark D. Plumbley教授的倡导下举行。
- 2016年举办第二届，此后每年举办一届。
- 常见任务
  - 声学场景分类
  - 声学事件检测（标记）
  - 异常声音检测
  - 音频内容描述

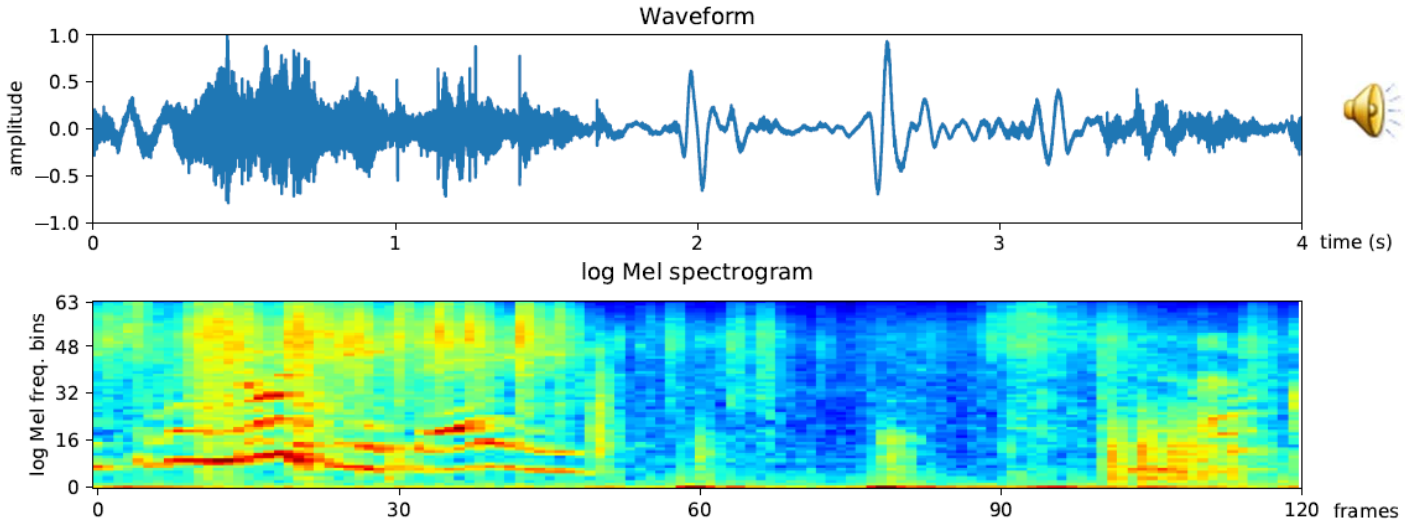
声学场景分类

- 自2013年第一次DCASE比赛开始，该项目即存在。
- 在众多的DCASE任务中，声学场景分类的正确率较高。
- 由于简单声学场景分类系统的性能较好，声学场景分类的研究主要集中在算法适应性的扩展上
  - 开放数据集问题
  - 多设备学习问题
  - 声学信号采集地点问题
  - 模型简化问题
  - 低计算复杂度问题
  - 多模态融合问题



声学事件检测

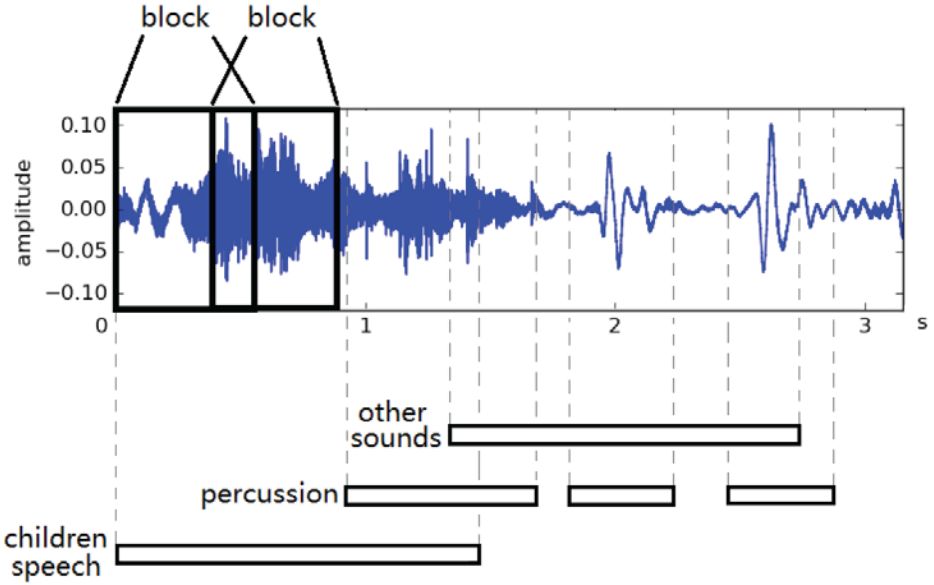
- 自第一届DCASE比赛开始，一直有所涉及。



Audio tagging: children (c), percussion (p), other sounds (o).

Sound event detection:

声学事件检测

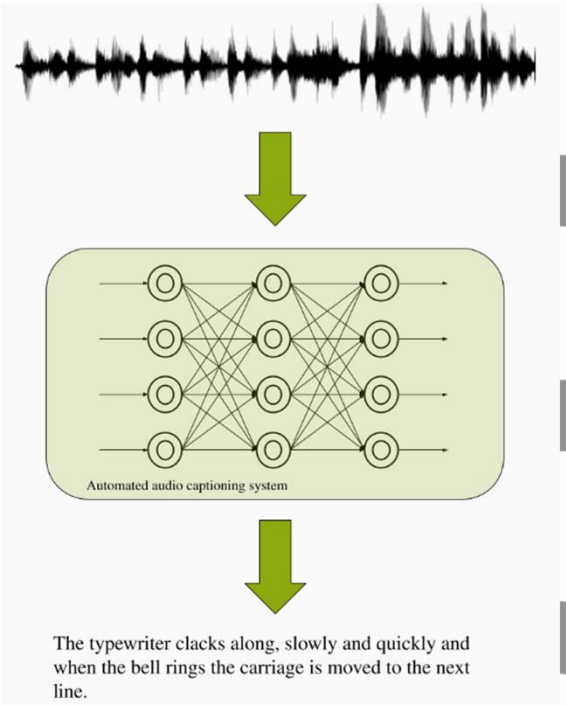


- 声学事件检测项目的发展历程
  - 合成声音检测
  - 声音事件标记
  - 实际声音检测
  - 半监督学习声音事件检测
  - 小样本动物叫声检测

- 声学事件检测较声学场景分类复杂，主要原因有三：
  - 标签标记复杂耗时
    - 强标签→（序列标签）→弱标签
  - 声音事件混叠
  - 声音事件定义仍欠完备

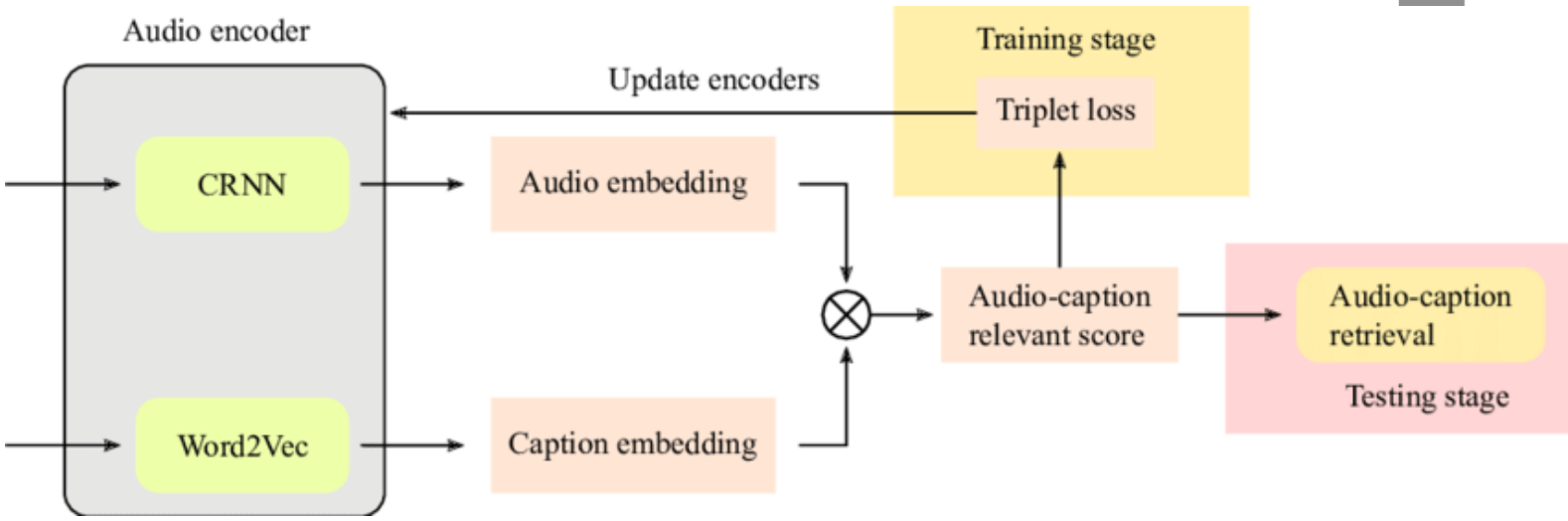
声音内容描述

- 音频内容描述是指系统根据声音，生成一段描述声音内容的文字
- 多模态信息
- 编码器——解码器结构
- 特殊问题：
  - 复杂深度学习系统下的预加载模型效率问题
  - 复杂音频信号中的语义分割、组合与理解



音频的自然语言检索

- 根据描述，检索相关音频
- 可以被看做音频内容描述的逆向应用



Xie, Huang & Lipping, Samuel & Virtanen, Tuomas. (2022).
DCASE 2022 Challenge Task 6B: Language-Based Audio Retrieval. 10.48550/arXiv.2206.06108.

未来发展方向

- 声音事件标签体系构建
- 领域自适应与领域泛化
- 半监督学习与自监督学习
- 小样本学习
- 多尺度信息分析
- 多模态信息融合
- 深度学习模型简化

声音事件标签体系构建

- Google AudioSet (2017)
  - 210万条10秒音频
  - 632种声音事件
  - 平均每段2.7个标签
  - 最大层级：6级
  - 物体声音→车辆→机动车→特种车辆→警笛→救护车警笛

Human sounds
- *Human voice*
- *Whistling*
- *Respiratory sounds*
- *Human locomotion*
- *Digestive*
- *Hands*
- *Heart sounds, heartbeat*
- *Otoacoustic emission*
- *Human group actions*

Animal sounds
- *Domestic animals, pets*
- *Livestock, farm animals, working animals*
- *Wild animals*

Natural sounds
- *Wind*
- *Thunderstorm*
- *Water*
- *Fire*

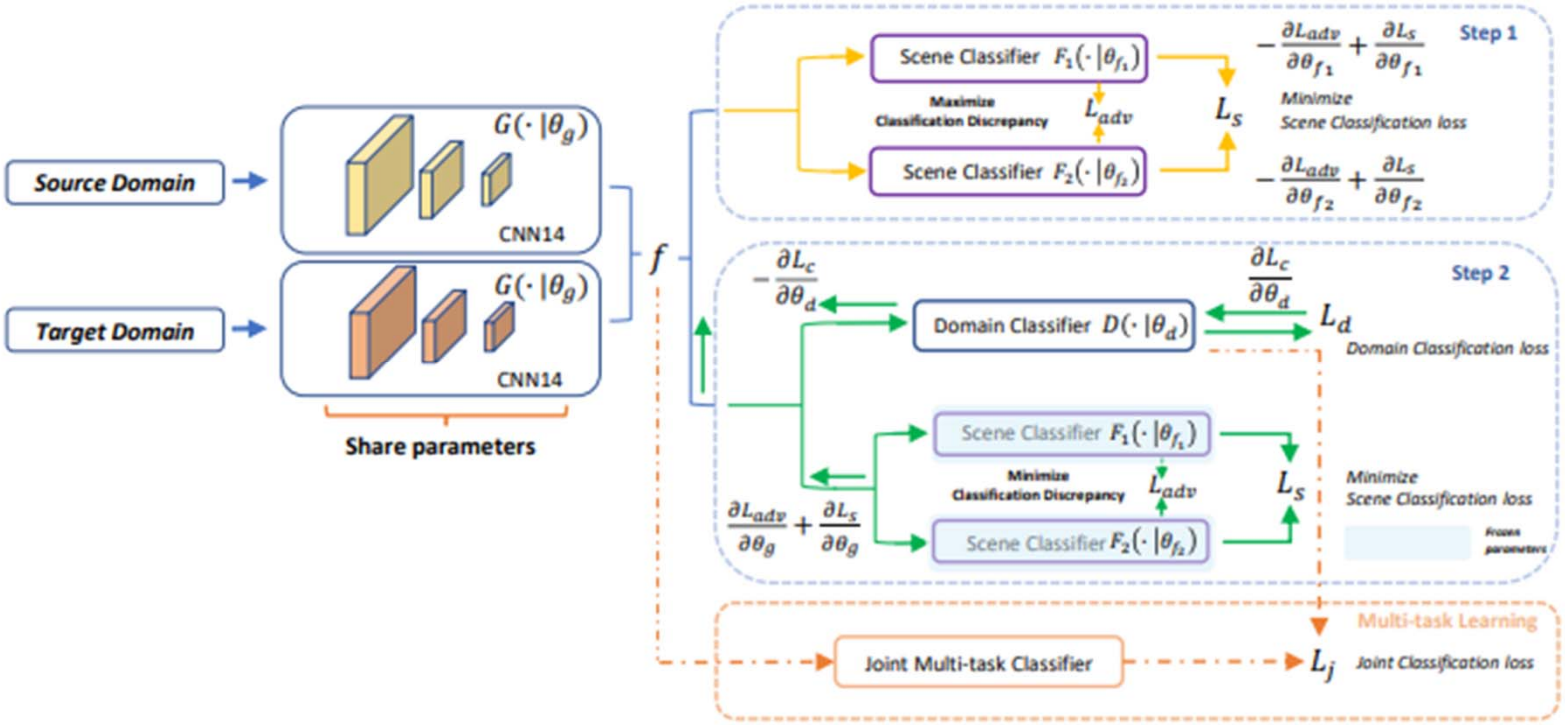
Music
- *Musical instrument*
- *Music genre*
- *Musical concepts*
- *Music role*
- *Music mood*

Sounds of things
- *Vehicle*
- *Engine*
- *Domestic sounds, home sounds*
- *Bell*
- *Alarm*
- *Mechanisms*
- *Tools*
- *Explosion*
- *Wood*
- *Glass*
- *Liquid*
- *Miscellaneous sources*
- *Specific impact sounds*

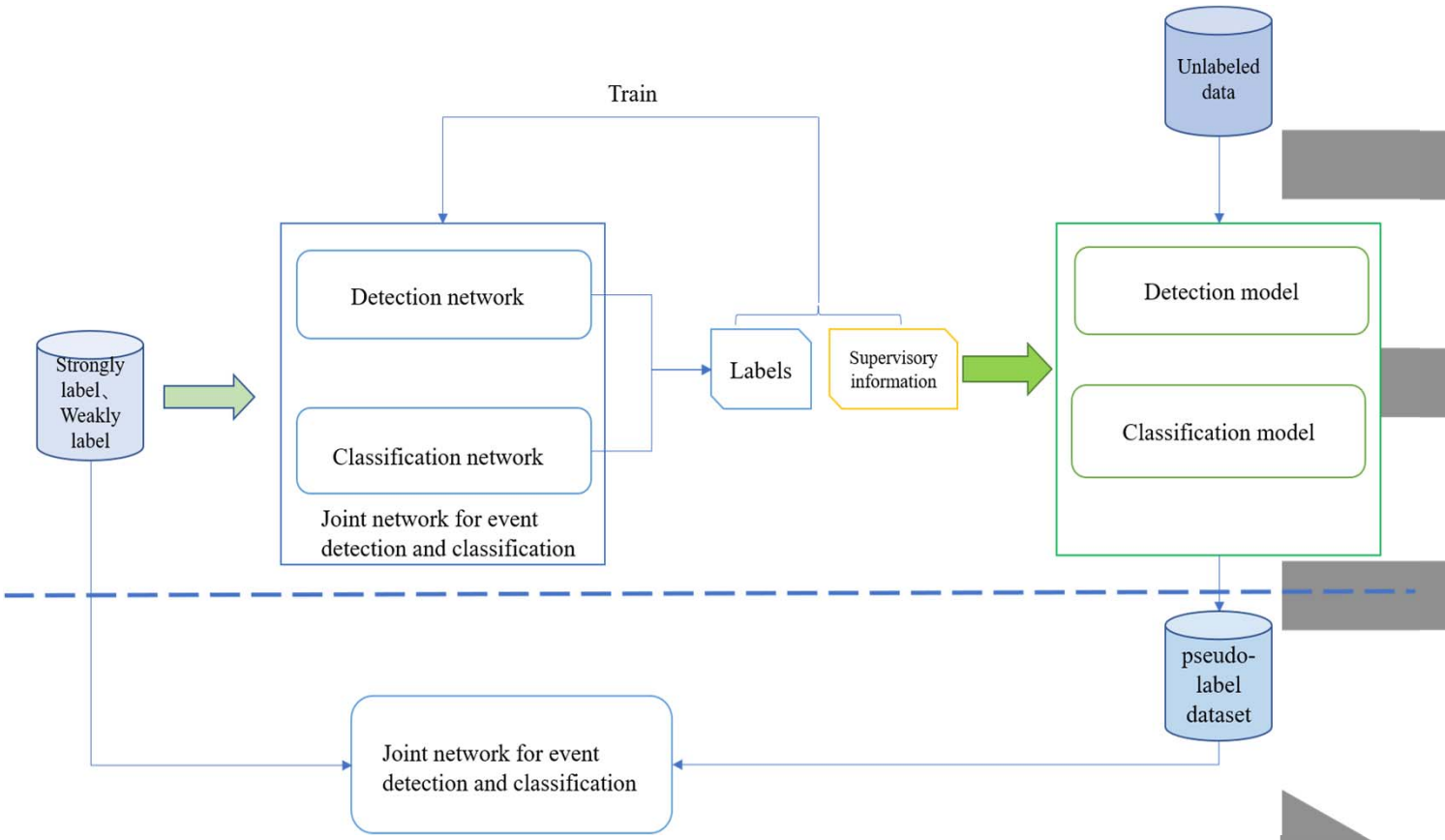
Source-ambiguous sounds
- *Generic impact sounds*
- *Surface contact*
- *Deformable shell*
- *Onomatopoeia*
- *Silence*
- *Other sourceless*

Channel, environment and background
- *Acoustic environment*
- *Noise*
- *Sound reproduction*

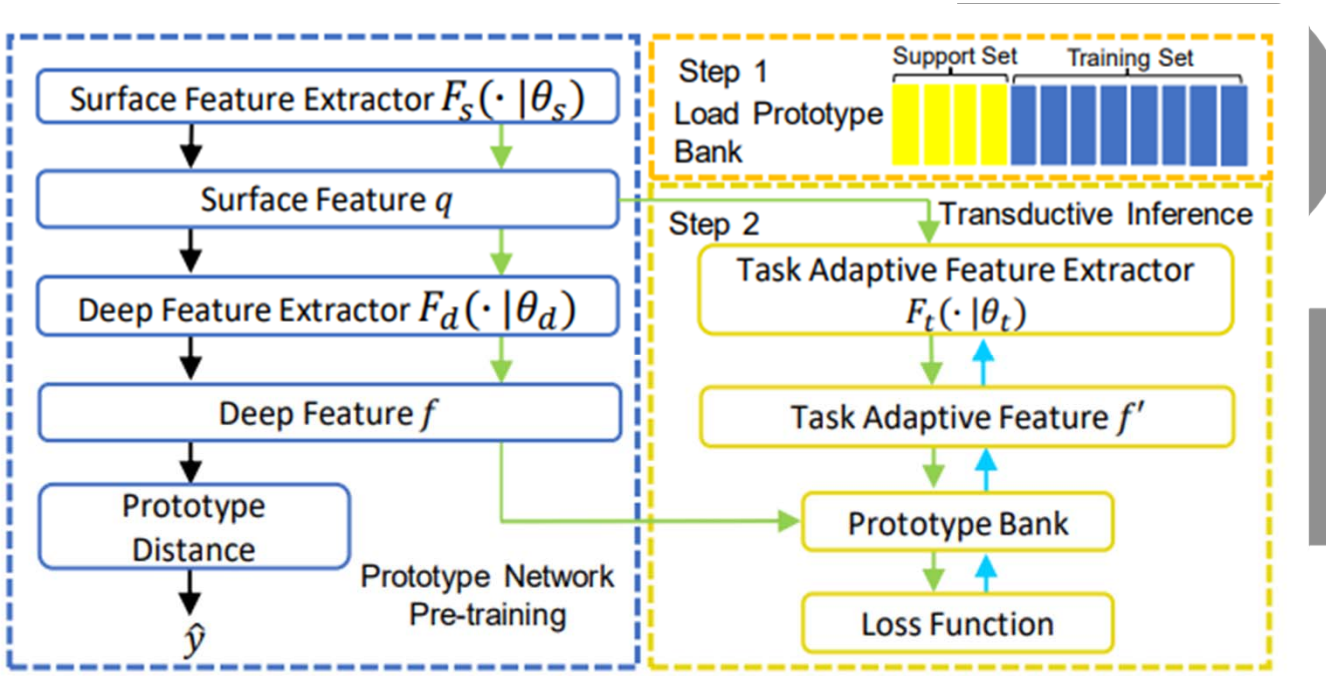
领域自适应与领域泛化

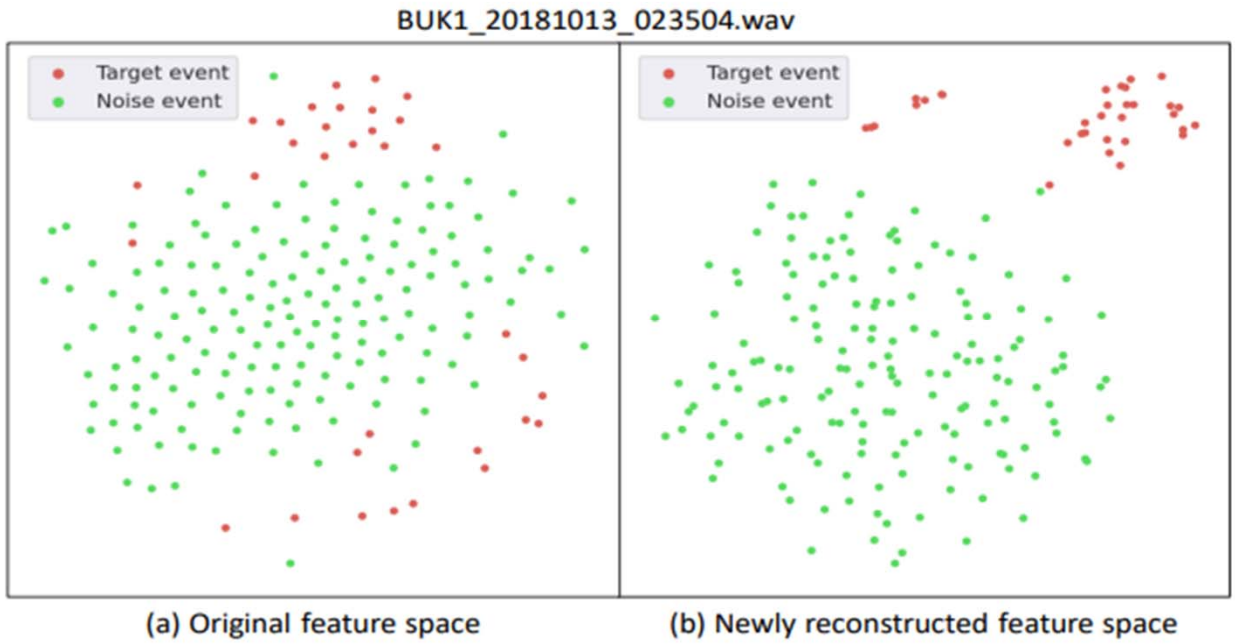

Source Domain
Target Domain
$G(\cdot|\theta_g)$
CNN14
$G(\cdot|\theta_g)$
CNN14
Share parameters
f
Scene Classifier $F_1(\cdot|\theta_{f_1})$
Maximize Classification Discrepancy
L_{adv}
Scene Classifier $F_2(\cdot|\theta_{f_2})$
L_s
$-\frac{\partial L_{adv}}{\partial \theta_{f_1}} + \frac{\partial L_s}{\partial \theta_{f_1}}$
Step 1
Minimize Scene Classification loss
$-\frac{\partial L_{adv}}{\partial \theta_{f_2}} + \frac{\partial L_s}{\partial \theta_{f_2}}$
$-\frac{\partial L_c}{\partial \theta_d}$
Domain Classifier $D(\cdot|\theta_d)$
$\frac{\partial L_c}{\partial \theta_d}$
L_d
Step 2
Domain Classification loss
Scene Classifier $F_1(\cdot|\theta_{f_1})$
Minimize Classification Discrepancy
L_{adv}
Scene Classifier $F_2(\cdot|\theta_{f_2})$
L_s
$\frac{\partial L_{adv}}{\partial \theta_g} + \frac{\partial L_s}{\partial \theta_g}$
Minimize Scene Classification loss
Multi-task Learning
Joint Multi-task Classifier
L_j
Joint Classification loss

半监督学习与自监督学习



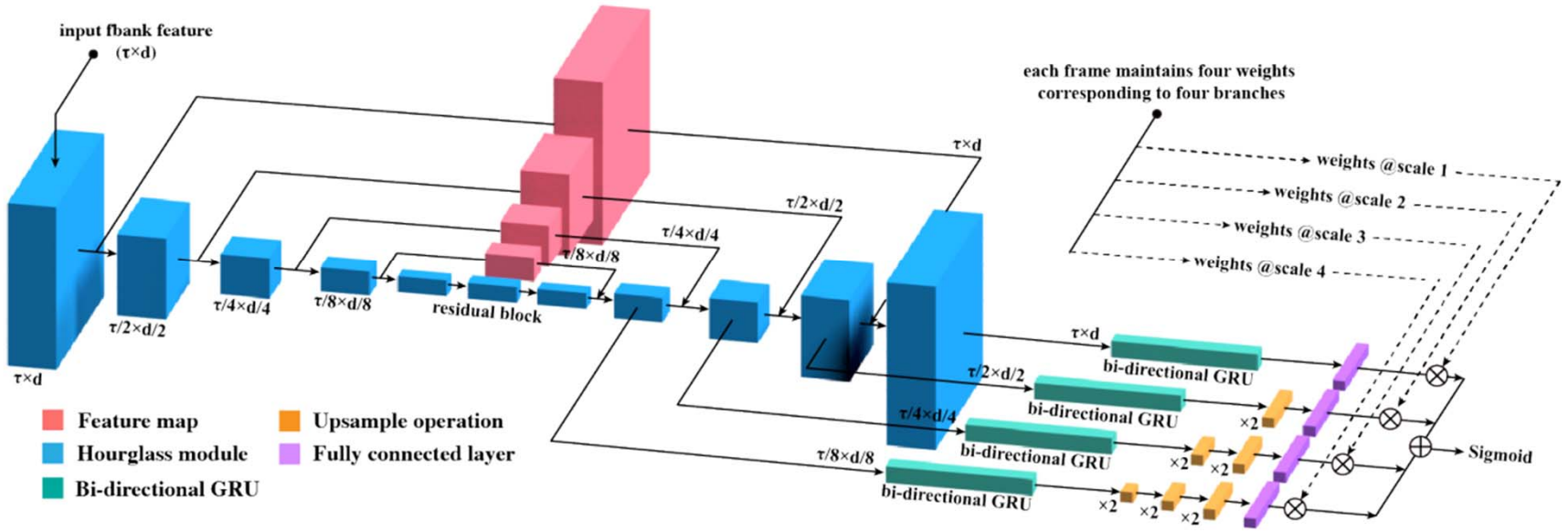
小样本学习





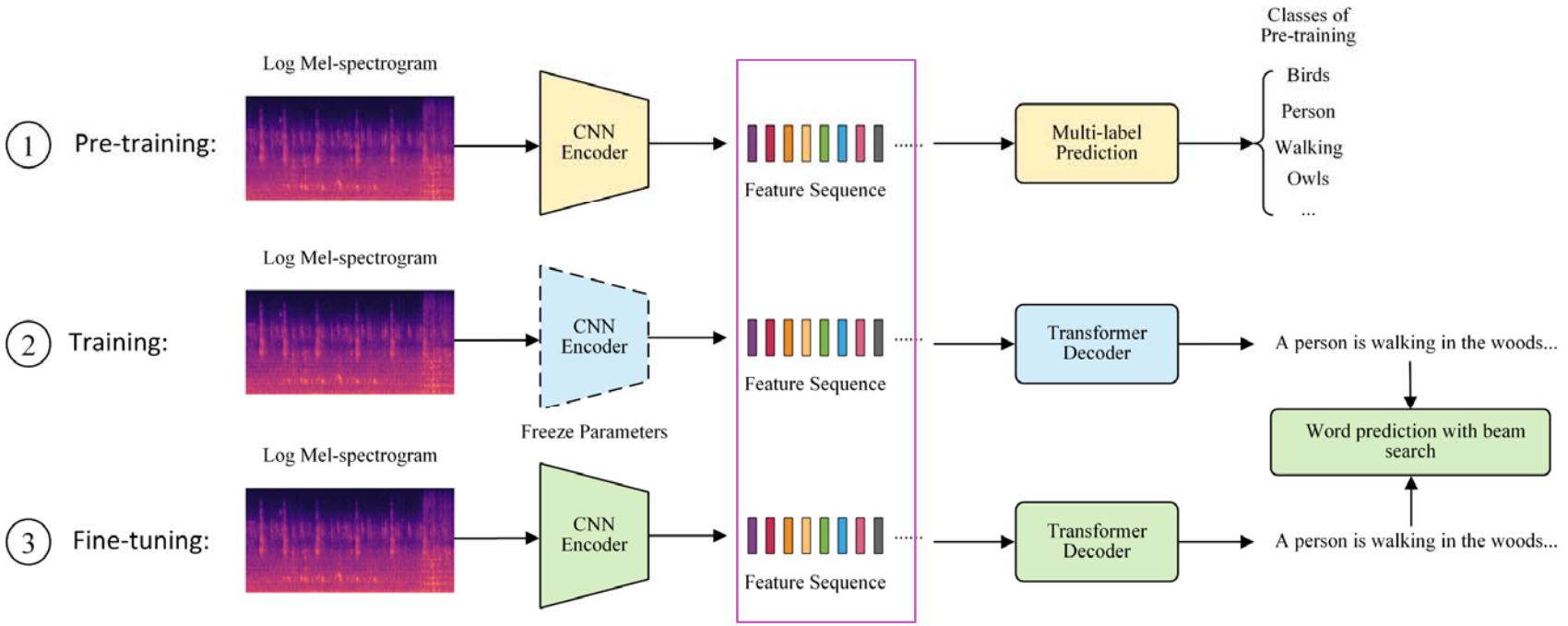
(a) Original feature space (b) Newly reconstructed feature space

多尺度信息分析



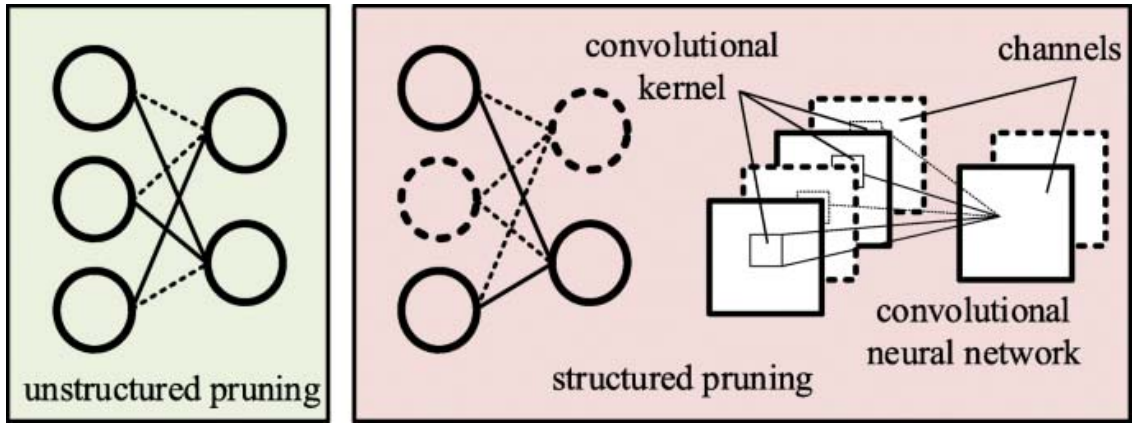
Ding, W., & He, L. (2019). Adaptive multi-scale detection of acoustic events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 294-306

多模态信息融合

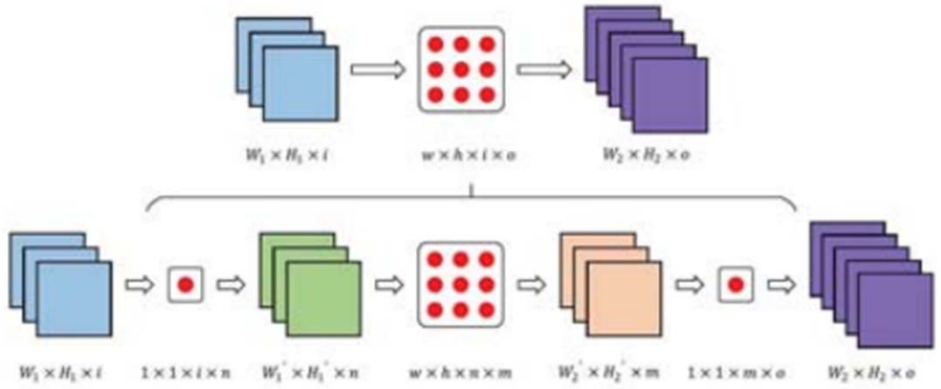


Shared Feature Space?

深度学习模型简化



Qi, C., Shen, S., Li, R. *et al.* An efficient pruning scheme of deep neural networks for Internet of Things applications. *EURASIP J. Adv. Signal Process.* **2021**, 31 (2021).



Wang, Jun, Shengchen Li, and Wenwu Wang. "Svd-based channel pruning for convolutional neural network in acoustic scene classification model." *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2019.

谢谢

访问西浦

WWW.XJTLU.EDU.CN

关注西浦

@西交利物浦大学

Xi'an Jiaotong-Liverpool University

西交利物浦大学