Multimodal Machine Learning: A Survey and Taxonomy

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency

Note

- Introduction
  - 真实世界的信息总是多模态的，AI理解世界的过程不能局限于unimodal的信息
  - 主要集中于这几种信息的模态：
    - 自然语言：现实世界的表征方式
    - 视觉信号：图像，视频
    - 语音信号：语言信息，韵律信息，语音表达
  - 多模态领域五大挑战：
    - Representation：如何将不同模态的信息表征到一个空间内？一个模态中哪些信息是redundancy的，哪些信息是互补的。比如一张狗的图片和一段狗的叫声，狗的图片中的背景信息（比如蓝天，草地）实际上是和狗的叫声的语音信息没有交互的。
    - Translation：如何将一个模态的信息转化到另一个模态的空间中？
    - Alignment：？
    - Fusion：将多种模态的信息进行融合然后用于certain tasks
    - Co-Learning：在不同的模态之间传递knowledge，用其他模态的信息来提高指定模态的performance
- Application 应用
  - Speech Recognition：语音识别 （AVSR audio-visual-speech-recognition）
    - 早期试验发现：辅助的visual information可以提高语音识别的robustness，但是并不会提高noiseless环境下的语音识别的性能，“supplementary rather than complementary”
  - 多模态的信息检索
  - 理解人类的社交中的交互行为
  - Multimedia Generation
    - Image Captioning， 产生描述图片的自然语言
    - 问题：如何准确评价生成图片/语言的质量？
      - 可以使用VQA（visual question answering）系统来进行evaluation
- Multimodal Representation
  - Joint Representations：

    所有模态的features作为输入，然后可以通过可学习的DL-model将其映射到同一个multimodal空间内：$x_m = f(x_1, x_2, \dots x_n)$, 其中$f$可以是NN，BM, RNN等
    - 使用神经网络：提取不同modal中的抽象过后的信息然后fusion
      - 自监督/无监督 VAE来减少特征提取时的数据依赖
      - 好处：能够减少对label的依赖，fine-tune
      - 缺点：不能处理**数据缺失**，难训练

        - 概率图模型：Deep Boltzmann Machine （DBM）
            - 类似神经网络的思路，but can cope with data missing naturally
            - 缺点：计算开销巨大，需要使用approximation methods
        - 序列表示模型（Sequential Representation）：
            - RNN， LSTM
            - RNN encoder-decoder 做自监督的特征提取
    - Coordinated Representations**(?)**:

      对两种模态进行分别的处理，然后map到一个coordinated multimodal space中，即 $f(x_1) \sim g(x_2)$
        - Enforce **similarities** between features from different modals
        -

- Translation
- Alignment
- Fusion
- Co-Leaning
- 感悟