机器学习作业

[redacted] [redacted]

2021/12/14

> 👉如何用机器学习方法处理多源模态信息数据（表格、文本、语音、图像、视频等）？
>
> 请大家对这个问题进行自己的思考，课后查看相关文章和理论并进行总结，提交内容word/pdf 1-2页(可附图）。

1. 什么是多元模态数据?

多模态数据涉及不同的感知通道如视觉、听觉、触觉、嗅觉所接收到的信息；在数据层面理解，多模态数据则可被看作多种数据类型的组合，如图片、数值、文本、 符号、音频、时间序列，或者集合、树、图等不同数据结构所组成的复合数据形式，乃至来自不同数据库、不同知识库的各种信息资源的组合。

图 1 “下雪”场景的多模态数据（图像、音频与文本）
Fig. 1 Multimodal data for a “snow” scene (images, sound and text)

2. 多模态表示学习

单模态的表示学习负责将信息表示为计算机可以处理的数值向量或者进一步抽象为更高层的特征向量，而多模态表示学习是指通过利用多模态之间的互补性，剔除模态间的冗余性，从而学习到更好的特征表示。主要包括两大研究方向：联合表示（Joint Representations）和协同表示（Coordinated Representations）。

- 联合表示将多个模态的信息一起映射到一个统一的多模态向量空间；

- 协同表示负责将多模态中的每个模态分别映射到各自的表示空间，但映射后的向量之间满足一定的相关性约束（例如线性相关）。

2.1 转化 Translation / 映射 Mapping

转化也称为映射，负责将一个模态的信息转换为另一个模态的信息。常见的应用包括：

机器翻译（Machine Translation）：将输入的语言A（即时）翻译为另一种语言B。类似的还有唇读（Lip Reading）和语音翻译 （Speech Translation），分别将唇部视觉和语音信息转换为文本信息。

图片描述（Image captioning）或者视频描述（Video captioning）： 对给定的图片/视频形成一段文字描述，以表达图片/视频的内容。

语音合成（Speech Synthesis）：根据输入的文本信息，自动合成一段语音信号。

2.2 对齐 Alignment

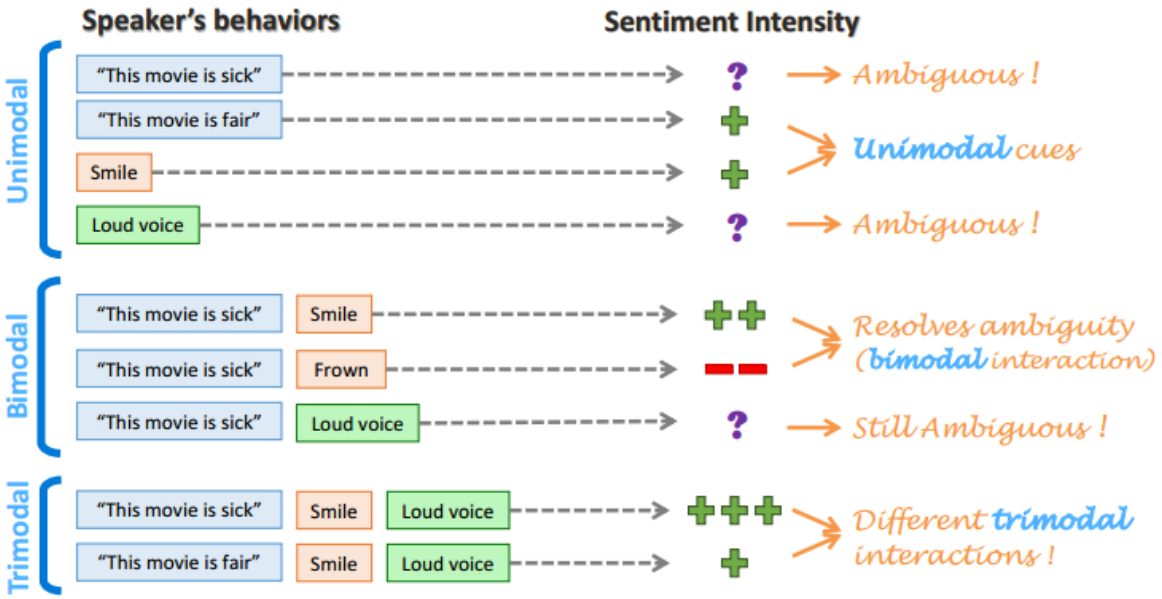
多模态的对齐负责对来自同一个实例的不同模态信息的子分支/元素寻找**对应关系**。这个对应关系可以是时间维度的，又可以是空间维度的，比如**图片语义分割 （Image Semantic Segmentation）**：尝试将图片的每个像素对应到某一种类型标签，实现视觉-词汇对齐。

2.3 多模态融合 Multimodal Fusion

常见的机器学习方法都可以应用于多模态融合，下面列举几个比较热门的研究方向。

视觉-音频识别（Visual-Audio Recognition）： 综合源自同一个实例的视频信息和音频信息，进行识别工作。

多模态情感分析（Multimodal sentiment analysis）： 综合利用多个模态的数据（例如下图中的文字、面部表情、声音），通过互补，消除歧义和不确定性，得到更加准确的情感类型判断结果。



手机身份认证（Mobile Identity Authentication）： 综合利用手机的多传感器信息，认证手机使用者是否是注册用户。

2.4 协同学习 Co-learning

协同学习是指使用一个资源丰富的模态信息来辅助另一个资源相对贫瘠的模态进行学习。