


源方 (/users/299443) 2020-04-16 19:41:32 (最初创作于: 2020-04-16 19:11:49) 发表于: 人工智能实验室-认知实验室 (/teams/511) 595 阅读

知识体系: 自然语言处理 (/articles/?kid=939) 计算机视觉 (/articles/?kid=935)  修改知识体系

文章标签: 自然语言处理 (/search?q=自然语言处理&type=ARTICLE) 算法 (/search?q=算法&type=ARTICLE) 计算机视觉 (/search?q=计算机视觉&type=ARTICLE) 多模态 (/search?q=多模态&type=ARTICLE)  修改标签  标签历史 (/articles/170108/tags/history)

附加属性: 内部资料请勿外传 作者原创

多模态融合技术浅析（一）：多模态技术的分类

导读：本文作为多模态融合技术交流系列第一弹，将介绍笔者对多模态技术的一些理解，其与各模态领域知识的关系，着重介绍多模态技术的分类。本文受众主要是相关领域算法从业者。本文为作者原创，水平有限，不妥之处还望指正。

一、多模态融合的目的

1. 感知（逆问题）

感知主要解决如何将真实世界数字化，并将获得的数据转化为有价值信息的问题。多模态感知是指利用多种感官获取原始数据，将这些数据进行联合处理并转化为有价值信息的技术，比如：利用视觉和听觉的信息进行联合决策，弥补单一感官信息不足的问题。

它的意义在于，不让机器变成聋子或者瞎子，让它逐步变成一个健全的“人”。

2. 呈现（正问题）

呈现主要解决如何将数字信息进行合理转换，变为人类善于理解和接收的信息的问题。多模态呈现主要弥补单一媒体在展现力方面的不足，达到将数据转化为多种人类可接收信号进行更完整信息表达的目的。例如：我们可将文本信息，利用机器全自动展现为虚拟形象的视觉展示和音频展示。这样的多模态呈现，将给用户更加立体的体验。

二、领域知识

每个模态都有各自的领域知识。我们以天猫精灵遇到的3个模态，简要说明领域知识是什么。

	模态	原始数据	数据维度	领域（感知）	领域（呈现）
1	视觉	图像、视频	2D、3D	计算机视觉	图像处理、图形学
2	听觉	音频	1D	音频识别	音频生成
3	自然语言	文本	1D	自然语言理解	语言生成

表1：视觉、听觉、自然语言三个模态的领域知识

如表1所示，各模态均有自己的领域知识，那这些领域知识本质上在做些什么呢？

我们将从信息融合的角度进行介绍，首先明确单一模态和跨模态的概念。



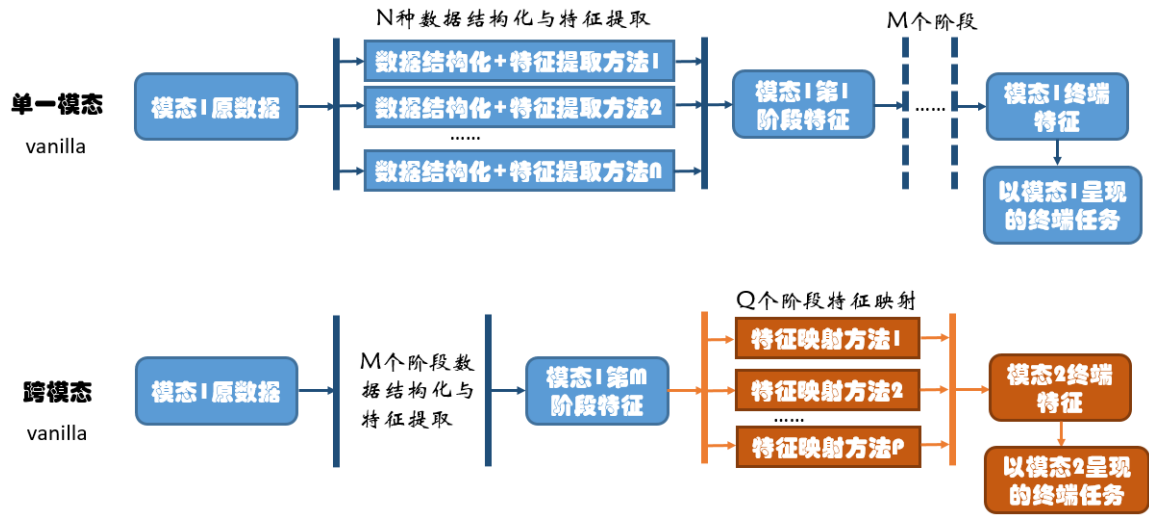


图1：单一模态与跨模态

如图1所示，在单一模态中，领域知识的作用是将各自领域的原数据进行结构化并提取特征，经过若干阶段，转化为与最终呈现形态相关的特征，最终完成各领域的end-task。比如：在图像领域常用的边缘检测（low-level特征），音频领域常用的梅尔频谱（频域特征），自然语言理解常用的分词和句法分析等。

所谓跨模态的概念也很简单，只是我们经常忽略它的存在而已。它是指首先通过模态1领域知识，经过一系列数据结构化和特征提取操作，转化为模态1下的最终特征，继而通过特征映射，将模态1特征映射到模态2的特征空间下，继而完成模态2所需展示的end-task。比如：在图像分类任务中，领域知识促使我们利用CNN进行图像特征提取，进而将图像转变为特征向量。此后，经过若干FC层进行特征映射，将特征空间由图像变为自然语言，进而判断语义类别。

综上，领域知识主要完成针对各自领域的数据结构化和特征提取工作。

三、多模态融合的分类

多模态融合按照融合阶段的不同大致可分为四类：1) 逻辑融合（Logic-fusion）2) 后融合（Late-fusion）3) 前融合（Early-fusion）4) 中期融合（Mid-fusion）。

1. 逻辑融合（Logic Fusion）

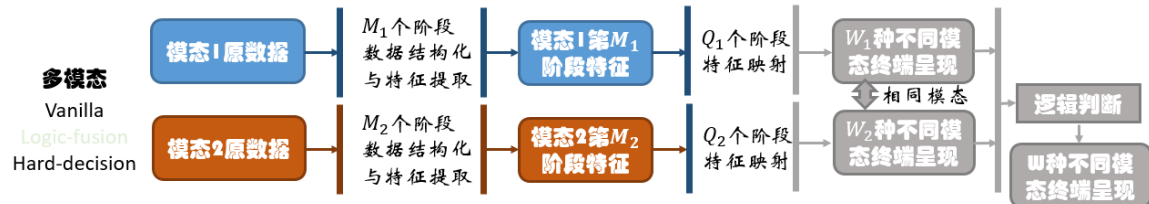


图2：多模态融合：逻辑融合

图2展示了逻辑融合的概念，简言之就是利用各自模态产生的标签，通过逻辑判断（if-else），完成最终决策。这种方式是最常见的，能够将各领域知识解耦并转化为统一有物理意义的标签，以完成最终决策。

逻辑融合至少拥有两个明显缺点：

1) 无法解决复杂逻辑问题：逻辑融合可以看作手工特征提取方法，即：发现标签出现内在的逻辑规律设计规则（新特征），继而完成最终决策。它对简单问题很有效，但面对复杂问题时，逻辑手工特征（hand-crafted feature）相比数据驱动下学习的特征（learned feature）有一定局限性。

2) 标签作为特征，维度过低，会引发特征维度不足导致的表达能力限制。即：决策过硬（hard-decision）。

因此，当面对复杂场景，逻辑融合无法满足要求时，需要做更加soft的decision，以达到目标。

2. 后融合（Late-fusion）





图3：多模态融合：后融合

图3所示为后融合的基本概念。在后融合中，模态1和模态2不产生各自的end-task标签，而是产生特征向量，继而通过可能的特征融合方法形成联合特征（joint feature）。以联合特征为基础，经过特征映射，继而将联合特征转化至输出模态的特征空间，完成end-task。

后融合相比逻辑融合，因为其对于高维特征的融合，可以给出更加soft和flexible的决策，可解决绝大部分复杂场景问题。特征融合的阶段和方法是多模态融合技术的核心。

值得指出的是，关于融合方法，不只concatenate这么简单粗暴，目前很多融合技术还利用了如bilinear pooling（参考link (<https://zhuanlan.zhihu.com/p/62532887>））等技术以应对复杂场景。关于从联合特征，具体映射到哪个模态下的特征空间进行最终信息表达，也关联了不同需求和技术方向。

3. 前融合（Early-fusion）与中期融合（Mid-fusion）

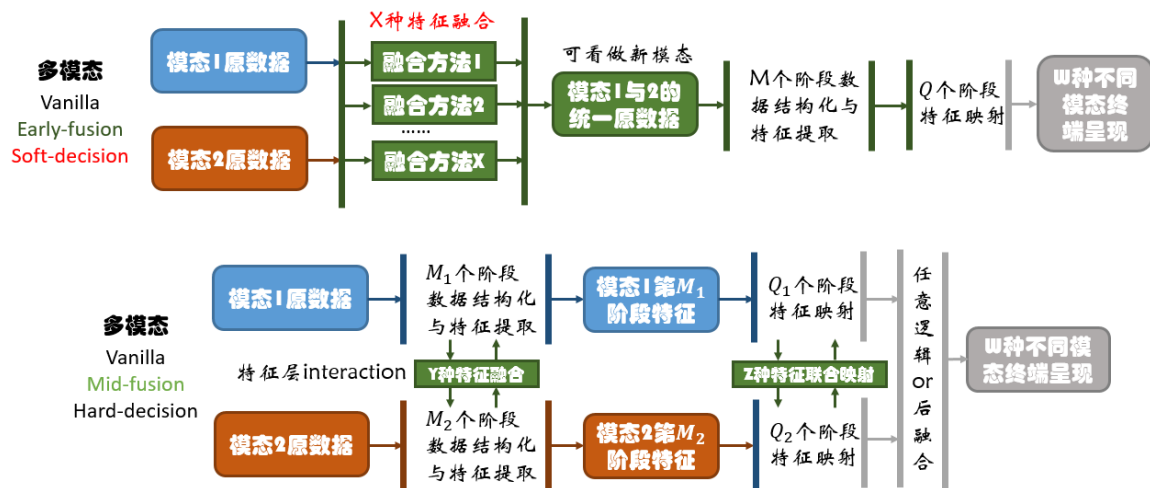


图4：多模态融合：前融合与中期融合

如图4，前融合通常是指在信息特征提取前，将不同模态的原始数据进行统一预处理，形成一种新模态原始数据，进而针对这种新模态设计后续结构化和特征提取方法，以完成end-task。这种做法简单粗暴，但是数据门槛很高，因为自定义新模态的原始数据完全依赖自有数据集，无法借力各模态领域内大规模公开数据集的预训练模型完成transfer learning。

中期融合则是在特征提取过程中，在各模态中间特征层发生信息交流，以达到更flexible的融合效果。和前融合一样，它的数据门槛相对很高，但在特定场合下非常必要。如：希望机器利用人的嘴型和音频的对应，判断声音是否来自某人。如果图像和音频在时间断面上进行特征融合，而不是时间段上的late-fusion，可能产生更好效果。

由上面叙述可知，多模态融合的难点是数据。

4. 多模态融合方式的选取建议

1. 对大多数简单场景，利用逻辑融合即可达到目标的，没有必要利用feature层融合。如：在家庭场景下，人脸识别和声纹识别的协同使用，当人脸可见时就用人脸，当人脸不可见时只能用声纹，依靠逻辑判断即可达成目标。
2. 当必须采用feature层融合，则优先使用后融合而不是前融合和中期融合。因为后融合可以借力各领域内经典问题的数据集，解决数据瓶颈，也方便把整个链路标准化和模块化。
3. 在不得不采用前融合和中期融合以达到更好效果时，尽量使用逻辑或后融合先上线，继而利用回流数据解决数据量不足的问题以达到满意效果。
4. 在end-to-end训练时，有条件下多supervise各模态特征，引入各模态标签，Multi-task进行训练。

四、小结

1. 多模态融合的目的：1) 感知：弥补单一感官信息不足；2) 呈现：弥补单一媒体在展现力方面的不足。
2. 领域知识是什么：领域知识主要完成针对各自领域的数据结构化和特征提取工作。



- 3. 多模态算法研发重点：特征融合的阶段和方法是多模态融合技术的核心。
- 4. 多模态融合的难点：多模态下的数据

五、下期展望

下期将介绍什么是狭义多模态融合，什么是广义多模态融合。将举一些例子说明多模态融合的普遍性与特殊性，敬请期待。

六、感谢



感谢与@应知@墨鲤@帆月@坎特以及组内其他小伙伴关于多模态融合技术的深入讨论。
感谢@涵毅@实一关于多模态研究与应用的技术交流分享，获益良多。

评论文章 (1) 16 (/articles/170108/voteup) 2 25 收藏 (/articles/170108/mark/)

他们赞过该文章

灵成 (/users/861088) 见水 (/users/114740) 刀锋 (/users/582152) 赢武 (/users/794077) 齐桓 (/users/32857) 爱桐 (/users/112747)
帆月 (/users/593819) 淮林 (/users/636729) 奇傲 (/users/536506) 朔元 (/users/572014) 实一 (/users/457133) 景澄 (/users/688280)
元融 (/users/430939) 坎特 (/users/277391) 应知 (/users/566850) 铂涛 (/users/404847)

上一篇：【独家干货】ICCV2019收录论文分类汇总（数据...

1F 奇傲 (/users/536506) 2020-04-17 23:38:10
期待直播分享
大迈 赞同
1 (/comments/286957/voteup) | 0

写下你的评论...

评论

