多模态语义理解中的不确定性

文 / 徐童，周培伦，陈恩红

摘　要　本文介绍了目前国内外关于多模态语义理解中不确定性问题的研究进展，主要从模态信息的不确定性与模态间关系的不确定性这两个维度进行梳理和分析，并在此基础上探讨研究的未来方向。

关键词　多模态语义理解；模态缺失和高噪声；模态间语义对齐

0 引言

我们生活于其中的世界，包含着诸如文字、语音和视觉等多种感官信息，亦即多模态的世界。对于人类来说，理解多模态的信息既是一种天赋，同时也是深入、全面理解世界的基础。但对于机器而言，这个任务却由于“语义鸿沟”的存在而具有一定的困难，并由于多模态信息中的不确定性问题而更具挑战性。

具体来说，多模态语义理解中的不确定性体现在两个方面。一方面为模态信息的不确定性（inner-modality uncertainty），诸如模态信息的缺失或者信息高噪，在这种情况下，想要精确地、符合情境地、无歧义地理解模态信息显得尤为困难；另一方面，则又体现为模态间关联的不确定性（inter-modality uncertainty），这种不确定性主要来源于不同模态间各异的表达方式，而如何去对齐这些各异的表达，进而去进行模态关联上的澄清，则对于细粒度的跨模态语义交互尤为重要。因此，本文将从模态内信息和模态间关联两个维度，对于多模态语义理解中的不确定性问题进行综合性的阐述。

1 模态信息的不确定性

由于真实世界中的多模态信息在构成上具有很强的随意性，无论是信息的数量还是信息质量，在每个模态上的分布都存在一定的不均衡现象，并由此产生了模态信息的不确定性。

仅以视频为例，在富文本信息（如包含弹幕信息）的在线视频服务中，一部较为冷门的视频通常伴随着较为少量的用户评论，在这种情况下文本模态的信息就会是稀疏的甚至是缺失的。同时，用户的评论往往具有较大的随意性和主观性，在这种情况下又会为文本模态引入噪声，从而降低文本的质量。

因此，针对模态信息的不确定性问题，本文又将其细分为模态信息缺失（不足）问题和模态信息高噪声问题两个子问题，并对这两个问题分别进行阐述。

1.1 模态信息缺失下的不确定性

首先，针对模态信息缺失问题，最常用的方法是通过模态迁移（multimodal translation）技术来进行模态信息的补全，其中又以 Ngiam 和 Srivastava 等的工作为代表。2011 年，Ngiam 等设计了深度

自编码器来进行多模态信息的表征，并在含有模态缺失的音频 - 视频数据集上进行训练，试图通过跨模态的信息迁移，去重构和其他模态的表征。之后，在 2012 年，Srivastava 又提出了以深度玻尔兹曼机为代表的多模态表征方法。其中，深度玻尔兹曼机作为一个无向图模型，可接收由多模态信息构成的可见状态作为输入，并通过可见层与隐藏层的对称性交互来学习多个模态的联合表征。尤其是在某些模态信息缺失的情况下，该模型也可以通过隐藏层的状态，从条件分布中采样可见层的状态对缺失信息进行补全。此外，在具体应用层面上，Moon 等展示了如何将一个基于音频的语音识别网络中的信息进行迁移，从而增强发音时唇部动作的视觉表征；而 Arora 等则反其道而行之，利用发音时的嘴部视觉特征，基于典型相关分析方法来构建和增强相应的声学特征，同样获得了很好的效果。

类似地，在解决中文文本表征问题时，考虑到直接将复杂的中文字符进行嵌入往往会造成语义信息的遗漏。因此，我们将字词级别的部首信息独立出来作为一个新的模态（见图 1），并通过长短期记忆网络对其进行编码来构建和丰富中文文本表征，同样取得了较好的效果。

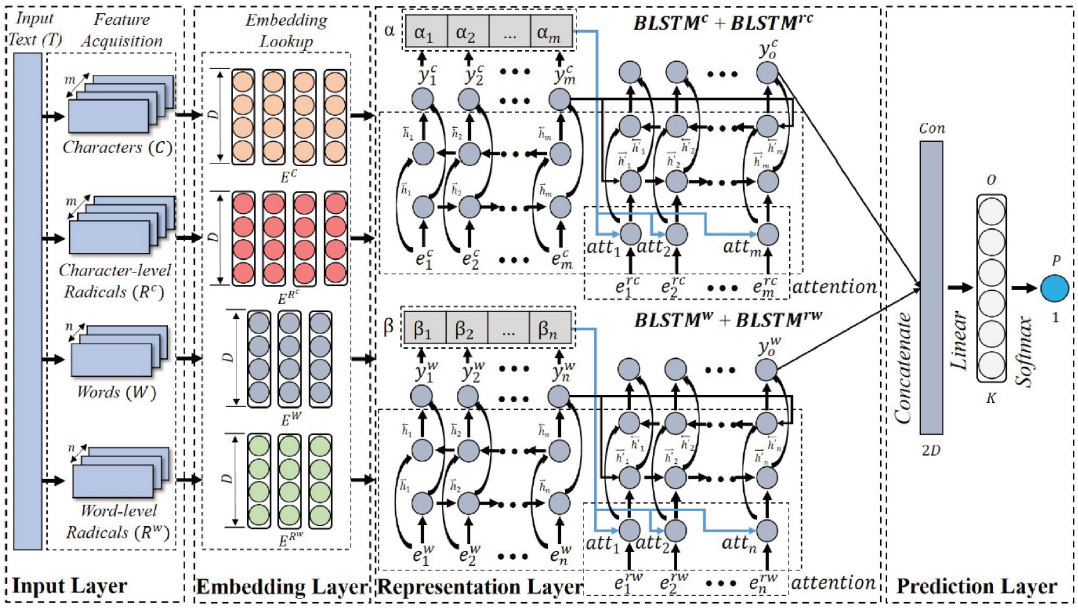

图 1　基于部首感知和注意力机制的文本表征模型

1.2 模态信息高噪声下的不确定性

针对模态信息中可能存在的高噪声问题，主流的方法往往通过多模态的协同学习（multimodal co-learning）来对信息进行整合和降噪。根据模态间数据对齐的程度，又可将这一问题进一步细分为针对平行数据或者非平行数据的降噪和过滤。

1.2.1 平行数据的降噪

在平行数据中，不同模态的信息共享同一个实例（instance）集合。例如在演讲视频中，图像和声音模态的信息均对应到同一个说话人；又如在电影中，用户的评论往往随着视频内容表达出相应的情感。在这类情况下，考虑到模态间具有较为明显且具体的相关性，所以常采用协同适应（co-adaptive）的相关方法。

其中，尤以 Christoudias 等的工作为代表。2006 年，他们设计了一种跨模态的半监督协同

训练算法，利用说话人口型和发音来进行音素和单词的识别。他们认为，在各个独立模态上进行操作的分类器之间可以分享它们各自最有信心的判断，从而可以有效规避单个模态中因为高噪声和异常数据带来的影响。基于该动机，他们设计了一种协同训练算法，在多轮迭代的过程中，不断对各个单模态分类器预测的高置信度的无标签数据进行标记，从而扩展有标签的数据集，并在重新训练中不断增强单模态分类器的性能，达到各个模态互相帮助，协同训练的效果。2008 年，Christoudias 等又在之前工作的基础上，针对弱监督的联合声音与视觉信息的视频情感极性识别问题，提出了跨模态的 Bootstrapping 算法，通过单模态分类器间的协同，在高度不平衡的数据集上获得了准确的预测性能。

面对平行数据中的噪声，我们也尝试通过使用深度学习方法，利用一个模态的高质量信息，去跨模态地辅助另一个模态上的语义理解。具体来说，我们在视频的自动标注任务中，借助于视频中的具有强主观性的文本信息，为视频添加更为细粒度的语义标签。值得一提的是，这种方法同时可以实现对于文本语义信息的准确理解，削弱了文本歧义性等因素造成的干扰，降低了文本模态的噪声。

1.2.2 非平行数据的降噪

在非平行数据中，不同的模态间仅共享某些更加抽象的概念或者模糊的关联，因而呈现出较弱的相关性。在这种情况下，通过跨模态的协同适应方法实现对共享信息之外的噪声进行过滤往往较为困难。因此，许多相关的研究工作是依照概念建构（concept grounding）的思想，通过同一个情境下多个模态信息间的融合（multimodal fusion）来增强表征能力。这一方面的研究工作最早由 Feng 等于 2010 年开始，他们依赖于多模态信息提供的情境，将概念建构实现为从多个模态信息中寻找共享的隐藏特征空间的过程。2015 年，Kiela 等将该方法用于语言学概念的表征研究，他们进一步主张将原始的感官信息，包括概念对应的图像和发音，联合语言学的文本信息，进行分阶段的模态融合，并逐步排除噪声。具体来说，他们设计了一个多模态 skip-gram 模型来建模多个模态的上下文信息，接着根据不同的模态融合方法，划分出早期、中期和后期阶段，每个阶段都在独立的数据集上进行训练，以这种方式逐步得到精确的语言学概念的表征，同年，Kiela 更是尝试着将嗅觉信号融入到某些与嗅觉相关的语言学概念表征的任务上。在此基础之上，2016 年，Shutova 等人同样基于 skip-gram 模型，以及分层的多模态融合策略，通过与单词关联的图像信息来辅助判断其是否是隐喻，并在公开数据集 MOH 和 TSV 上进行了实验，识别模型的 F1 值相比于纯语言模型有接近 4% 的提升，且召回率更是达到了 87%，能较为准确地识别出单词的隐喻特征。

为了尽可能规避非平行数据中的噪声，我们也尝试着利用融模态的方法进行更加精确的语义表征。具体来说，就是利用对抗生成网络，结合视频中的视觉信息与高噪声的、意图不明确的文本评论所提供的情景，实现融模态的视频评论的自动生成。如图 2 所示，模型在表征模块中，先是通过卷积网络获取视频的视觉特征，以及通过循环神经网络建模文本序列得到评论文本的表征；再通过变分自编码器来进行视觉与文本情境的编码；最后将融合后的情境编码输入生成器中，通过生成器与对抗器的最大最小游戏来实现评论的自动生成。实验结果表明，本文方法在 BLEU 指标上相对于传统的跨模态图像描述方法取得了显著的进步，而这也得益于我们对视觉与文本情境的融模态策略，对于有着更高自由度和复杂度的视频评论生成任务更加适用。

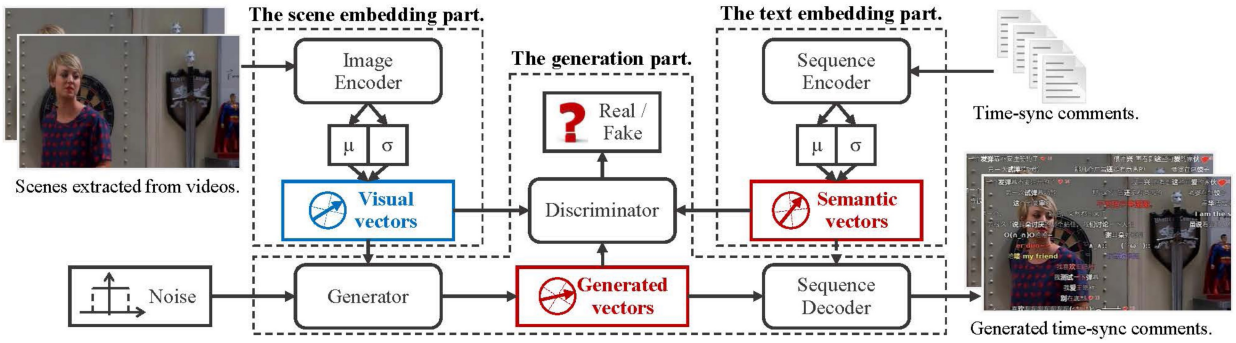

图 2 基于融模态的视频评论自动生成模型

2 模态间关联的不确定性

真实世界中的多模态信息，不仅在各个模态的分布上具有不确定性，甚至模态之间的关联也是不确定的和含糊的。举例来说，观看者对于图像的文本描述，本身可能包含多个层面的信息，诸如涉及的实体、实体间的关系或情感等。这些信息往往都只与图像的某一局部有强相关性，然而，这种关联却并非显性的、直观的，只有在对每个模态的子成分进行对齐（multimodal alignment）的基础上，才能有效地还原这种关联，从而更细粒度地、精确地理解整体的多模态信息。

而多模态对齐技术在方法论上，又可以大致分为显式的对齐和隐式的对齐。显式的对齐依赖于在不同模态的子成分间预先进行显式的相似性度量，并将之作为模型的一个独立模块为多模态的对齐提供依据。2014 年，Naim 参考隐马尔可夫模型，通过视频物体与文本名词间存在的弱共现关系，无监督地在生物实验视频与说明性文本之间进行匹配和对齐，并克服了实验顺序失误带来的影响。2015 年，Tapaswi 等尝试在电影和原著章节之间进行对齐，并以故事人物为中介，预先为每个人物显式地学习一个基于脸部特征的身份识别器，继而在每个视频场景中利用人脸追踪技术去统计不同人物出现的频数分布，将之与每一原著章节中人名的出现频数分布进行比照，判断原著章节与电影场景的相关性。

与之相对应的，隐式的对齐多数是通过注意力机制的协调，于训练过程中在模型内部自发学得的模态间的子成分的对齐。目前，隐式对齐技术已被广泛地应用于翻译相关的任务上（如语音识别和图像描述）。例如，在语音识别任务上，Chan 等于 2016 年提出了 listener-speller 的语音识别框架，先是通过由层级循环神经网络构成的 listener 对语音的滤波器组谱进行编码；而后由基于注意力机制的 speller 对语音中的每个字符进行隐式的对齐，并在此基础上识别语音内容。

在图像描述任务上，2015 年，Xu 等开创性地尝试在图像与文本描述间建立信息的细粒度对齐，从而更好地进行图像描述。具体来说，循环神经网络每生成一个单词时，都会在注意力机制的帮助下，赋予图片不同区域不等的重要性，从而使得对于该图片的描述质量可以在生成过程中不断得到提升。同年，Yao 等也将注意力机制推广到视频描述任务上，把 3D-CNN 捕捉到的视频局部信息和注意力机制捕捉到的全局信息相结合，达到了当时的领先水平。2019 年，Wu 等在 Xu 等工作的基础上，进行了更加细粒度的研究，如图 3 所示，对于每一句

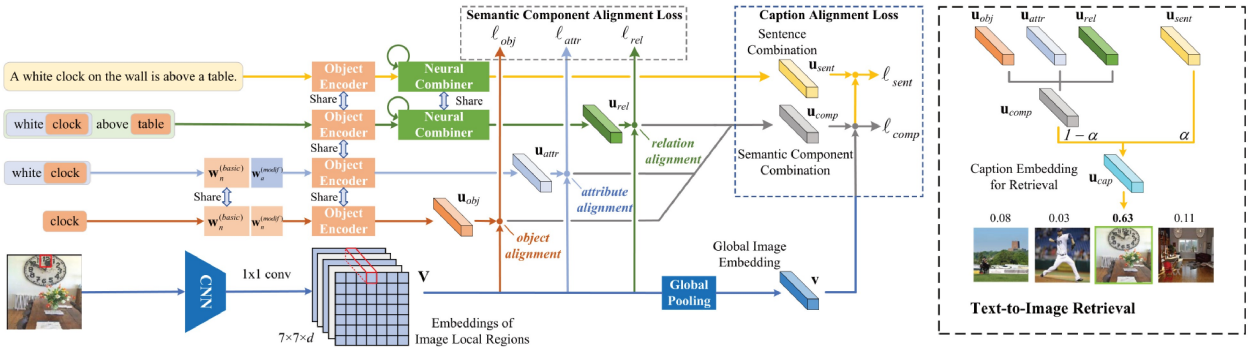

图 3 基于结构化注意力机制的图像文本对齐

图像描述，他们都在分层次的概念水平上（物体，属性，关系，情境）进行图像和文本间子成分的结构化映射，并通过对比学习的方法得到图像与文本共享的语义表征。此外，Agrawal 等更是验证了基于注意力机制的隐式对齐在视觉问答模型中发挥的积极效果，同时也细致分析了此类模型所具有的特性。

相比于 Xu 等的工作，在模态间隐式对齐的任务上，我们采取了一种反其道而行之的做法。如图 4 所示，从视频图像信息出发，通过注意力机制与时间邻域内的多条文本间建立匹配和对齐。基于这个思想，设计了一种联合图像视觉与用户评论信息的多模态人物重识别模型，并在真实数据集上进行了验证。实验结果证实了模态间的对齐是有效的，使用注意力机制可以在一定程度上识别出那些与视频人物描述更为相关的文本信息，从而有助于更精确地刻画出人物的身份特征，达到更好的人物重识别效果。

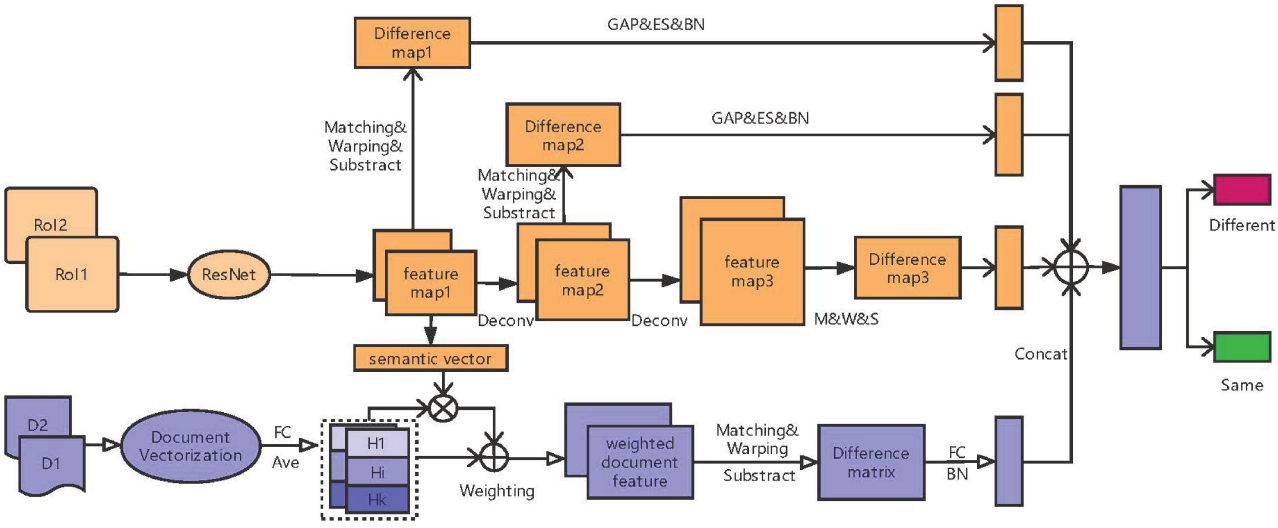

图 4 基于注意力机制的图像 – 文本隐式对齐

3 结束语

目前，对于多模态语义理解中不确定性问题的研究尚处于起步阶段，这不仅由于它包含着模态信息的不确定性与模态间关联的不确定性两方面的挑战，更为重要的是，在真实世界中，这两个难题往往又相互渗透、相互纠缠，进一步提升了多模态信息理解的难度。事实上，现有的多模态语义对齐技术往往有着较强的假设，即每个模

（下转第 56 页）

在快速发展过程中，2020 年我们已经满足了一些智能硬件边缘用户的需求，而且还在快速的增长。预计在将来的十年里，这些智能语音交互的用户体验不断提升，最终将超越智能手机现在主流用户的一些使用需求，达到所有人群对于人机交互的需求，从而成为人机交互中一个非常重要的组成部分。

刚才讲到的是智能语音交互在新一代的 AIoT 时代的重要的作用，未来还是要看两个方面，一个就是基于场景的自然语音交互智能平台，另一个是基于行业的专家系统的人工智能平台，它们会深入到某个行业里给我们带来了一个新的结果，就是每个领域、每个行业乃至每个企业都将有自己的人工智能平台，专注解决各自不同的人工智能问题，开发不同的产品和服务。

对于将来的商业生态，我们有一个什么样的理解？可以看到，每个公司都有自己的人工智能大数据和云计算，但是并不是每个公司，比如家电厂商和运营商都能够自己建立这套人工智能系统，这就需要有像科大讯飞这样的公司去帮助这些传统产业的公司，或者是没有能力建自己人工智能的公司。因为我们和他们是完全正交的，可以帮助他们建立更加紧密的合作，在新生态下知识经验数据利益的分享变得更加重要。

通过这种混合的方式，要建立新生态下的共赢，也就是会有非常多的控制节点，一起在整个商业生态系统里去共同获取我们的将来，所以这种共赢的合作是成功的关键。我坚信，将来大家会一同进入整个人工智能的时代。所以，最后的混合正交的商业生态系统，将会成为人工智能时代的主要方式。

上面的报告里我给大家回顾了人工智能，特别是我们讲到将来的从感知智能到认知智能，利用现在的大数据人工智能的方法，将能够起到什么样的作用？我相信，随着大家的努力，我们不但在中国能够用人工智能改变世界，而且在国际上也将会找到更多的合作伙伴，共同推进人工智能的视野。

（本报告根据速记整理）

（上接第 11 页）

态的信息都较为规整。然而，这一假设在现实场景中往往难以实现，因此极大地限制了现有技术的应用。

但在不远的将来，随着技术的积累，模态对齐问题势必要向高噪声数据拓展，并成为模态内部信息降噪的一种重要辅助手段。相应的，模态内部信息的降噪也能更好地支撑跨模态的对齐，最终促使上述针对多模态语义理解中的不确定性问题的技术方法能够互相促进，并自组织为一个有机的整体。总而言之，多模态语义理解问题仍具有很大的研究空间和研究价值，值得不懈地进行探索。

（参考文献略）

作者介绍

徐童

中国科学技术大学副教授。主要研究方向为社会网络与社交媒体分析。

陈恩红

中国科学技术大学教授，CAAI Fellow。主要研究方向为数据挖掘与机器学习。