视觉与语言相结合的跨媒体智能分析及应用

2020年9月20日 15:30-18:00

王鹏，西北工业大学计算机学院

报告题目：Richer and Deeper: Vision and Language Understanding with Richer Visual Content and Deeper Non-visual Knowledge

报告摘要：In this talk, I will introduce two recent works on vision and language understanding. The first one is a question-conditioned graph attention network for TextVQA, which is capable of reasoning over a heterogenous graph with text and object nodes. The second one is a dataset and pipeline that performs referring expression understanding using external commonsense knowledge. By incorporating more visual and non-visual information, we see an increasingly comprehensive visual reasoning ability.

刘偲，北航计算机学院

报告题目：Referring Expression Comprehension

报告摘要：指代表达理解（Referring Expression Comprehension）是视觉与语言交叉领域中的热门研究课题，包含 localization 和 segmentation 两个分支，对于智能机器人和交互式图像编辑等实际应用有重要意义。Localization 分支的主流方法采用两阶段式架构，模型复杂且速度受限，而 segmentation 分支的主流方法聚焦于多模态特征融合，缺乏利用语言信息进行上下文建模和推理的能力。在本次讨论中，会尝试对上述问题提出针对性的解决方法，提升模型对于 referring expression 的理解能力。

吴琦，澳大利亚阿德莱德大学

报告题目：视觉-语言问题中的深层推理研究

报告摘要：视觉-语言（Vision-and-Language）问题是近年来非常热门的一个研究课题，这个领域内比较主流的问题有 Image Captioning， Visual Question Answering 以及 Referring Expression. 目前解决这些问题的主流方法基本是基于深度学习，依靠观察大量数据“记忆”出一个从输入到输出的对应关系。而我们认为这些问题的价值在于如何让机器懂得“推理”，这个报告中，我会通过介绍我们近期的几个工作， 来阐述如何通过改变任务目标，模型架构，测试标准等方法，来真正体现视觉-语言问题中的深层推理问题与挑战。

黄岩，中科院自动化所

报告题目：Few-Shot Image and Sentence Matching via Aligned Cross-Modal Memory

报告摘要：The task of image and sentence matching has attracted much attention recently, and many effective methods have been proposed to deal with it. But its intrinsic few-shot problem, i.e., uncommonly appeared instances and words in images and sentences cannot be well associated, is usually ignored and seldom studied, which has become a bottleneck for further performance improvement in real applications. This talk will introduce our recent work on the few-shot image and sentence matching, by proposing an Aligned Cross-Modal Memory (ACMM) model to handle it.

潘滢炜，京东AI研究院

报告题目：图像描述生成：从自洽、交互到共生

报告摘要：Vision and language are two fundamental capabilities of human intelligence. Humans routinely perform tasks through the interactions between vision and language, supporting the uniquely human capacity to talk about what they see or hallucinate a picture on a natural-language description. Image captioning, as one of the hottest task in such type of research, is to automatically produce a natural-language sentence that describes the image content. The talk will briefly review existing innovations on this topic, covering three bases of visual perception via encoder, language modeling through decoder, and the multi-modal interaction in between. Moreover, we will also discuss the reflection on what is likely to be the next big leap in captioning.

论坛讲者

王鹏

西北工业大学

刘偲

北京航空航天大学

吴琦

阿德莱德大学

黄岩

中科院自动化所

潘滢炜

京东AI研究院

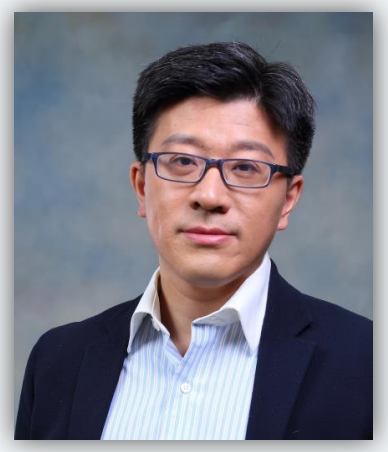
论坛组织者

梅涛

京东AI研究院

于静

中科院信息工程研究所



秦曾昌

北京航空航天大学