# 机器之心技术周报

**知识工程 知识图谱**

2019年10月19日（总第3期）第3期

本期相关主题： 注意力图卷积网络，知识图谱嵌入，BERT神经网络，生产对话系统，图数据库

本周知识图谱领域发表了多篇论文，为知识图谱中应用中出现的挑战提供了发人深醒的思路和行之有效的解决方案。例如在问答系统任务上，引入基于本体的语义成分增强正则化方法，大大提高了准确率；颠覆传统费时的静态知识图谱嵌入模式，用动态知识图谱嵌入实现高效，实时地嵌入；解决知识图谱图查询由结构不一致导致的匹配错误，系统响应时间过长的问题；成功使得面向非目标的生成对话系统能根据问题提供有充分依据的答案等。在理论方面，一种基于范畴论的新形式框架也被提出，为RDF和SPARQL的主要基本特征提供了清晰简洁的形式定义。在学习资源方面，本周推荐了一份基于python的spaCy库，手把手实现维基百科知识图谱创建的讲义。而开发资源方面，也介绍了几个有效的小工具，比如集成网络和动态推理组装器，知识图谱的可视化和知识模式建模等。更详细内容，请参照本周更新。

# 本周概览

### 热点事件 | 同盾科技云图为智能反欺诈、信贷风控和智能决策提供了强有力的科技支撑。

### 论文更新 | Efficiently Embedding Dynamic Knowledge Graphs

### 论文更新 | Bridging the Knowledge Gap: Enhancing Question Answering with World and Domain Knowledge

### 论文更新 | Semantic Guided and Response Times Bounded Top-k Similarity Search over Knowledge Graphs

### 论文更新 |On foundational aspects of RDF and SPARQL

### 学习资源 | Knowledge Graph – A Powerful Data Science Technique to Mine Information from Text

### 开发资源 | AmpliGraph (knowledge graph library)

### 开发资源 | Grakn Workbase – An Integrated Environment

### 开发资源 | Integrated Network and Dynamical Reasoning Assembler

### 近期活动 | Graphorum – 知识图谱专题演讲（Oct.14，芝加哥·美国）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 类别 | | 标题 | | | 时间 | |
| SOTA | | [AnnaParser: Semantic Parsing for Tabular Data Analysis](https://arxiv.org/abs/1910.10363) | | | **2019.10.23** | |
| 论文提出了一种将给定表的自然语言问题转换为SQL查询语句的新颖方法，该方法满足数据分析应用程序的三个要求：跨域，多语言和快速启用。方法包括：（1）在解析器之前添加表不可知性分析的新颖数据抽象步骤； （2）一组语义规则，用于将抽象的数据分析问题解析为树形派生的中间逻辑形式，以减少搜索空间； （3）将神经网络模型作为在跨度语义解析器上的局部评分函数，以进行结构化优化和高效推理。实验表明，论文方法优于大型开放基准数据集WikiSQL上的榜单最先进算法。此外，该方法还可以在小型数据集上以英语和中文进行更复杂的查询，并获得较好的结果，这也证明了论文方法的语言扩展能力和快速启动能力。  *文本转SQL，语义解析* | | | | | | |
| SOTA | | [**ChineseGLUE：为中文NLP模型定制的自然语言理解基准**](https://www.jiqizhixin.com/articles/2019-10-23-8) | | | **2019.10.24** | |
| GLUE 是一个用于评估通用 NLP 模型的基准，其排行榜可以在一定程度上反映 NLP 模型性能的高低。然而，现有的 GLUE 基准针对的是英文任务，无法评价 NLP 模型处理中文的能力。为了填补这一空白，国内关注 NLP 的热心人士发布了一个中文语言理解测评基准——ChineseGLUE。ChineseGLUE 目前拥有八个数据集的整体测评及其基线模型，并有 20多位来自各个顶尖机构的自愿者加入并成为了创始会员。  *中文自然语言处理，GLUE，Benchmark* | | | | | | |
| SOTA | | [**谷歌T5预训练模型刷新GLUE榜单，110亿参数量，17项NLP任务新SOTA**](https://www.jiqizhixin.com/articles/2019-10-25-8) | | | **2019.10.25** | |
| 2018 年，谷歌发布基于双向 Transformer 的大规模预训练语言模型 BERT，为 NLP 领域带来了极大的惊喜。而后一系列基于 BERT 的研究工作如春笋般涌现，用于解决多类 NLP 任务。预训练模型也成为了业内解决 NLP 问题的标配。今日，谷歌又为大家带来了新的惊喜，新模型在 17 个 NLP 任务中都达到了 SOTA 性能。  *Transformer，迁移学习，text-to-text* | | | | | | |
| 热点事件 | | **麻省理工新研究：**[**点云深度学习**](http://news.mit.edu/2019/deep-learning-point-clouds-1021) | | | **2019.10.21** | |
| 麻省理工学院的一项新的点云深度学习的研究希望能让自动驾驶汽车、机器人和其他应用程序更容易理解 3D 世界。已有的大多数方法并未特别成功地从数据中捕获模式，而这些数据是从一堆 3D 空间点中获取有意义的信息所必需的。在这一项新研究中，研究人员证明了他们的「EdgeConv」方法能使用一种称为动态图卷积神经网络的神经网络来分析点云，从而使他们能够对单个对象进行分类和分段。除了开发 EdgeConv，该团队还探索了点云处理的其他特定方面。例如，一个挑战是大多数传感器在 3D 世界中移动时都会改变视角—每次我们对同一对象进行新的扫描时，其位置都可能与上次看到的位置不同，要将多个点云合并到一个单一的详细视图中，研究员提出了一个被称为「注册」的过程中，通过注册从而对齐多个 3D 点，将来自不同来源的 3D 数据集成到一个共同的坐标系中。  *点云，动态图卷积神经网络，3D* | | | | | | |
| 热点事件 | | [**CCL 2019 各大奖项揭晓，复旦大学、哈工大、北语等团队获最佳论文奖**](https://www.jiqizhixin.com/articles/2019-10-21-11) | | | **2019.10.21** | |
| CCL 会议创办于 1991 年，由中国中文信息学会计算语言学专业委员会主办，是国内自然语言处理领域权威性最高、规模和影响最大的学术会议之一，会议内容主要聚焦于中国境内各类语言的智能计算和信息处理。此次最佳论文奖共有四篇获奖论文。包括复旦大学邱锡鹏等关于 BERT 微调用于文本分类的任务，福州大学等机构研究者提出的重建选项的观点型阅读理解任务模型，哈尔滨工业大学关于结合规则蒸馏的情感原因发现方法以及北京语言大学关于汉语复合名词短语语义关系知识库构建与自动识别研究。  *文本分类，情感发现，关系知识库* | | | | | | |
| 热点事件 | | [**Facebook 人工智能实验室：通过压缩搜索结果来改善长形式的问题回答**](https://ai.facebook.com/blog/research-in-brief-training-ai-to-answer-questions-using-compressed-search-results/) | | | **2019.10.22** | |
| 近日，Facebook 人工智能实验室更新了一篇名为「通过压缩搜索结果来改善长形式的问答系统」的博客，介绍了一种可通过使长格式问答（QA）系统更有效地搜索相关文本来提高其性能的新方法。该方法基于该团队自然语言处理任务 long-form QA 基础之上。在该任务中，模型必须使用前 100 个网络搜索结果来回答自然语言的问题，例如「爱因斯坦为什么出名？」。尽管答案通常会出现在这些结果中，序列对序列（seq2seq）模型在分析如此大量的数据时会遇到困难，因为需要处理数十万个单词。而通过将文本压缩为知识图并引入更细粒度的注意力机制，FB 的新技术实现了让模型使用整个网络搜索结果来解释相关信息。  *长格式问答系统，序列到序列* | | | | | | |
| 热点事件 | | [**NVIDIA实验室：全景分割帮助自动驾驶汽车更准确理解周围环境**](https://blogs.nvidia.com/blog/2019/10/23/drive-labs-panoptic-segmentation/) | | | **2019.10.23** | |
| NVIDIA DRIVE Labs 系列的最新文章以工程为重点研究了自动驾驶汽车的挑战以及 NVIDIA DRIVE 的解决方案。自动驾驶汽车「认知」世界通常是通过边界框—这些边界框将汽车，行人和停车标志等等都整齐地包裹在红色和绿色矩形中。但是，在现实世界中，并非所有内容都适合放在盒子里。对于高度复杂的驾驶场景，例如以交通锥标为标志的建筑区域，公路中间的沙发椅或其他道路碎屑，或从货车后方卸下货品的行人，给予车辆的感知软件更多细节有助于其了解周围复杂的环境。通过以像素级精度对图像内容进行分割，又称为全景分割，可以获得这种细粒度的结果。通过全景分割，可以针对语义内容（哪个像素代表汽车与行人相对于可驾驶空间）以及实例内容（哪个像素代表同一汽车与不同的汽车对象）准确地解析图像。NVIDIA 的方法使用单个多任务学习型深度神经网络实现了相机图像的像素级语义和实例分割—这种方法能够训练一个全景分割的深度神经网络，进而不需分段直接整体理解场景。  *自动驾驶技术，全景分割，语义分割，实例分割* | | | | | | |
| 热点事件 | | [**谷歌量子计算登上Nature封面，首次实现量子优越性，里程碑式突破**](https://www.jiqizhixin.com/articles/2019-10-24) | | | **2019.10.23** | |
| 谷歌在被称为「量子优越性」方向上的重大突破研究,开发了一种名为「Sycamore」的新型 54 比特处理器，该处理器由快速，高保真量子逻辑门组成，以执行基准测试。谷歌声称：该机器在 200 秒内完成了目标计算，而世界上最快的超级计算机要想产生类似的输出也需要花费一万年。  *量子优越性，量子计算，机器学习* | | | | | | |
| 热点事件 | | [**谷歌更博：学习气味：使用深度学习预测分子的嗅觉特性**](https://ai.googleblog.com/2019/10/learning-to-smell-using-deep-learning.html) | | | **2019.10.24** | |
| 谷歌今日发表了一篇新博客分享了其团队在「机器学习中的气味：学习小分子的通用感知表示」研究中的发现。研究人员利用图神经网络（GNN）（一种旨在将图作为输入进行操作的深度神经网络），直接预测单个分子的气味描述符，而无需使用任何手工制定的规则。研究结果证明，与目前的最新技术相比，该方法在气味预测方面具有显着改善的性能，并且是未来研究的有希望的方向。  *图神经网络，通用感知表示* | | | | | | |
| 论文更新 | | [Cross-Lingual Machine Reading Comprehension](https://www.jiqizhixin.com/articles/2019-10-21-11) | | | **2019.10.21** | |
| 近年来，机器阅读理解研究得到了飞速发展，然而多数工作面向的是英文数据，忽略了机器阅读理解在其他语言上的表现，而其根本原因在于大规模训练数据的缺失。这篇论文提出跨语言机器阅读理解任务来解决非英文下的机器阅读理解。首先给出了若干传统方法来实现跨语言机器阅读理解，其中包括一些零样本（zero-shot）学习方法。紧接研究人员创新地提出了Dual BERT模型，对<篇章，问题>在双语环境中建模，同时利用富资源语言来帮助低资源语言下的机器阅读理解。通过在两个中文机器阅读理解数据集上验证得知，该方法能够显著提升低资源下的机器阅读理解效果，为未来低资源下的机器阅读理解提供了一种解决方案。  *机器阅读理解，零样本学习* | | | | | | |
| 论文更新 | | [**GPU 加速的 VITERBI 精确网格解码器，用于分批在线和离线语音识别**](https://arxiv.org/pdf/1910.10032v1.pdf) | | | **2019.10.22** | |
| 近日，英伟达研究人员提出一种优化的加权有限状态传感器（WFST）解码器，能够实现在线流式传输和脱机使用图形处理单元对音频进行分批批处理。研究称，解码器的内存利用率很高，输入 / 输出带宽，并使用旨在最大化并行度的新颖 Viterbi 实现。节省内存使解码器能够处理比以前更大的图形同时支持更大数量的手机连续流处理。总体而言，所提出的改进使单次加速达到 240 倍核心 CPU 解码。该架构还可以部署大型硬件上的生产级模型，实现数据中心服务器对低功耗边缘设备的实用性。  *加权有限状态传感器，在线流式传输，语音识别* | | | | | | |
| 论文更新 | | [**斯坦福:通用可编程光子架构，用于量子信息处理**](https://arxiv.org/pdf/1910.10141.pdf) | | | **2019.10.22** | |
| 研究人员提出了一种用于量子可编程门阵列（QPGA）的光子集成电路结构，该结构能够准备任意量子态和算子。该架构由一个相位调制的 Mach-Zehnder 干涉仪的格子组成，该格子在路径编码的光子量子位上执行旋转；嵌入式量子发射器使用两个光子散射过程在相邻量子位之间实现确定性的受控 zz 操作。通过在晶格内适当设置相移，可以对器件进行编程以实现任何量子电路，而无需进行硬件修改。研究团队提供了用于在设备上精确准备任意量子状态和算符的算法，并证明了基于梯度的优化可以训练模拟的 QPGA，以具有逼近度的保真度自动实现对重要量子电路的高度紧凑的近似。  *量子可编程门阵列，梯度优化* | | | | | | |
| 论文更新 | | [**获顶会最佳论文，天津大学等用强化学习寻找游戏bug**](https://www.jiqizhixin.com/articles/2019-10-23-8) | | | **2019.10.22** | |
| 该论文主要融合了进化算法与深度强化学习算法，从多目标优化的角度，旨在解决大规模商业游戏的自动化智能测试问题。首先论文针对四款网易游戏产品中的 1349 个真实 bug 进行深入分析，并针对性的提出了四个用于 bug 检测的 oracle。其次，提出了实时游戏测试框架 Wuji，通过融合了进化算法，DRL 算法和多目标优化机制，实现了智能的自动化游戏测试。Wuji 在赢得游戏和探索游戏空间之间取得了较好的平衡。其中，赢得游戏可以使得智能体在游戏中取得进展；而空间探索则可以增加发现错误的可能性。最后，论文使用一个仿真游戏和两个大型商业游戏对 Wuji 算法的效果进行了大规模评估，结果证明了 Wuji 在探索游戏状态空间方面以及检测 bug 方面的有效性。此外，Wuji 算法还检测到了游戏中先前从未被发现过的漏洞，进一步论证了算法的有效性。  *进化算法，深度强化学习，多目标优化* | | | | | | |
| 论文更新 | | [**DeepMind 新研究：终身语言学习中的情节记忆**](https://arxiv.org/abs/1910.09840) | | | **2019.10.23** | |
| DeepMind 研究团队引入了终生的语言学习设置，其中模型只需从文本示例流中学习，而无需任何数据集标识符。他们提出了一种情景记忆模型，该模型通过执行稀疏的体验重播和局部适应，以减轻此设置中的灾难性遗忘。文本分类和问题解答的实验证明了稀疏体验重播和局部适应的互补优势，即模型可以从新数据集中不断学习。研究结果表明，通过随机选择要在内存中以最小的性能下降来存储的示例，可以显着降低情景存储模块的空间复杂度（约 50-90％）。研究团队认为情节记忆成分是一般语言智能的重要组成部分，而论文提出的模型朝着这个方向迈出的第一步。  *情景记忆，文本分类* | | | | | | |
| 论文更新 | | [**2019 ICCV收录论文：基于弱监督学习的病理影像分析框架**](https://www.jiqizhixin.com/articles/2019-10-23-3) | | | **2019.10.23** | |
| 人工智能病理诊断成为当今学术研究和工程落地的热点。数字病理切片的体积通常都在500MB至2GB，像素数超过百亿，有监督的病理诊断模型需要进行繁琐的像素级切片标注，对这一领域的快速发展带来了挑战。相比之下，弱监督学习仅需要图像级别的粗粒度标签，能够大幅降低标注的工作量。由于单个数字病理切片尺寸过大，仅通过切片级诊断很难获得高准确率的结果（建立弱监督学习模型通常需要超过1万张数字切片 [1]），研究人员通常将数字病理切片划分为若干的图像，对每一个图像打上标签，进而获得弱监督学习模型。但是，由于对监督信息利用率不足，在之前发表的研究工作中，弱监督学习算法的准确率远低于有监督学习 。而这篇论文提出弱监督学习框架CAMEL，通过多实例学习（multiple instance learning, MIL），CAMEL能够通过建模自动生成细粒度（像素级）的标注信息，从而可以使用有监督的深度学习算法完成图像分割模型的建立。通过在CAMELYON16和解放军总医院肠腺瘤数据集上的验证，CAMEL能够取得接近完全有监督模型的准确率。  *弱监督学习，多实例学习，自动标注* | | | | | | |
| 论文更新 | | [**深度学习识别脑出血，准确度高达 99%**](https://blogs.nvidia.com/blog/2019/10/23/deep-learning-brain-hemorrhages/) | | | **2019.10.23** | |
| 加州大学伯克利分校和加州大学旧金山分校医学院的研究人员正在研究一种深度学习模型，以减少在 CT 扫描中诊断颅内出血（颅骨出血）所需的时间。仅在美国，每年就进行超过 8000 万次 CT 扫描，因此人工智能在大量图像中可以提高放射科医生的效率。研究人员通过 Amazon Web Services 使用英伟达 V100 Tensor Core GPU 来训练和推断其人工智能模型，该模型可以分割出血区域，并以 99％的准确度识别出脑出血。  *深度学习，图像识别* | | | | | | |
| 学习资源 | | [**14周无监督学习课程，UC伯克利出品，含课件、视频**](https://www.jiqizhixin.com/articles/2019-10-22-10) | | | **2019.10.22** | |
| 这份来自 UC 伯克利最新的无监督深度学习 CS294-158 课程总共 14 周，资源中包括 YouTube 高清视频和课件，绝对是不容错过的学习课程。其中，课程涵盖了两个无需标签数据的深度学习领域：深度生成模型（Deep Generative Models）和自监督学习（Self-supervised Learning）。生成模型的最新进展使得可以对高维原始数据进行逼真的建模，例如自然图像，音频波形和文本语料库；自我监督学习的进步已经开始缩小监督表征学习与无监督表征学习之间的差距，以微调未见任务。  *无监督学习，生成模型，自监督学习* | | | | | | |
| 开发资源 | | [**14周无监督学习课程，UC伯克利出品，含课件、视频**](https://www.jiqizhixin.com/articles/2019-10-22-10) | | | **2019.10.22** | |
| 这份来自 UC 伯克利最新的无监督深度学习 CS294-158 课程总共 14 周，资源中包括 YouTube 高清视频和课件，绝对是不容错过的学习课程。其中，课程涵盖了两个无需标签数据的深度学习领域：深度生成模型（Deep Generative Models）和自监督学习（Self-supervised Learning）。生成模型的最新进展使得可以对高维原始数据进行逼真的建模，例如自然图像，音频波形和文本语料库；自我监督学习的进步已经开始缩小监督表征学习与无监督表征学习之间的差距，以微调未见任务。  *无监督学习，生成模型，自监督学习* | | | | | | |
| 开发资源 | |  | |  | **2019.10.24** | |
| 基于知识图谱进行套餐推荐、套餐信息查询、个人信息查询等任务的中移动智能对话系统。系统使用了三个识别器：用户动作识别器、属性识别器和值识别器，模型主要使用了 词袋模型(BOW)和富对话状态跟踪模型 (EDST)。  *知识工程，知识模式，知识图谱可视化* | | | | | | |
| 近期活动 | | [**国际顶级学术会议CIKM2019北京开幕在即，重量级嘉宾带你窥探人工智能前沿**](https://www.jiqizhixin.com/articles/2019-10-23-12) | | | **2019.11.03-11.07** | |
| 第28届国际顶级“信息提取，知识管理和数据库” 年会（CIKM - The ACM International Conference on Information and Knowledge Management）将于今年11月3日-7日在中国北京召开。自1992年成立以来，CIKM致力于将这三个相关领域的研究员和开发者聚集在一起，实现跨领域的深入合作与交流。CIKM今年以“未来生活的人工智能”为主题，聚焦大数据和人工智能未来愿景的技术和见解。此次落户北京，CIKM为国内研究和工程人员提供了一个绝佳的机会和平台去参与发表、介绍和讨论高质量的理论和应用研究成果，发现行业所面临的重要挑战，提出高效的解决方案，并共塑未来。  *知识管理，数据科学，信息检索* | | | | | | |