050 | 如何将“深度强化学习”应用到视觉问答系统?

2018-01-26 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 08:07 大小 3.72M

本周我们一起来剖析 ICCV 2017 的论文，周一和周三分别讲了最佳研究论文和最佳学生论文。今天，我们来分享一篇完全不同的文章，题目是《使用深度强化学习研究协作性视觉对话机器人》（Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning），讲的是如何通过“深度强化学习”来解决视觉问答系统。

作者群信息介绍

第一作者阿布谢克·达斯（Abhishek Das）是一名来自佐治亚理工大学的在读博士生。他于 2017 年和 2018 年在 Facebook 人工智能研究院实习，已经获得了 Adobe 的研究奖学金和 Snapchat 的研究奖学金，可以说是一名非常卓越的博士生。之前在智能系统，特别是在利用强化学习研究智能机器人会话系统的领域已经发表了多篇论文。

共同第一作者萨特维克·库托儿（Satwik Kottur）来自卡内基梅隆大学，博士第四年，研究领域为计算机视觉、自然语言和机器学习。2016 年暑假他在 Snapchat 的研究团队实习，研究对话系统中的个性化问题。2017 年暑假在 Facebook 研究院实习，研究视觉对话系统。近两年，萨特维克已在多个国际顶级会议如 ICCV 2017、ICML 2017、IJCAI 2017、CVPR 2017、NIPS 2017 以及 EMNLP 2017 发表了多篇高质量研究论文，可以说是一颗冉冉升起的学术新星。

第三作者何塞·毛拉（José M. F. Moura）是萨特维克在卡内基梅隆大学的导师。何塞是美国工程院院士和 IEEE 院士，长期从事信号处理以及大数据、数据科学的研究工作。他当选 2018 年 IEEE 总裁，负责 IEEE 下一个阶段的发展。

第四作者斯特凡·李（Stefan Lee）是来自乔治亚理工大学的研究科学家，之前在弗吉尼亚理工大学任职，长期从事计算机视觉、自然语言处理等多方面的研究。斯特凡 2016 年博士毕业于印第安纳大学计算机系。

第五作者德鲁·巴塔（Dhruv Batra）目前是 Facebook 研究院的科学家，也是乔治亚理工大学的助理教授。德鲁 2010 年博士毕业于卡内基梅隆大学；2010 年到 2012 年在位于芝加哥的丰田理工大学担任研究助理教授；2013 年到 2016 年在弗吉尼亚大学任教。德鲁长期从事人工智能特别是视觉系统以及人机交互系统的研究工作。文章的第四作者斯特凡是德鲁长期的研究合作者，他们一起已经发表了包括本文在内的多篇高质量论文。

论文的主要贡献

我们首先来看一下这篇文章的主要贡献，理解这篇文章主要解决了什么场景下的问题。

这篇论文是建立在这么一个虚拟“游戏”（Game）的基础上的。

首先，我们有两个“机器人”（Agent），一个叫“Q 机器人”（Q-Bot），一个叫“A 机器人”（A-Bot）。这个游戏的规则是这样的。一开始，A 机器人得到一张图片 I，Q 机器人一开始得到 I 的一个文字描述 c，而并不知道图片本身。然后，Q 机器人开始问 A 机器人关于图片的各种问题，A 机器人听到问题之后进行作答，帮助 Q 机器人更进一步理解图片。Q 机器人最终的目的是能够把这个图片“猜到”，也就是说能够把图片从一个数据库中“提取”（Retrieve）出来。当然在实际的操作中，这一步可以是去衡量 Q 机器人对于图像的理解，也就是“描述图像的向量”和“真实图像的描述向量”的差距，差距越小说明越成功。

那么，你可以看到，这其实是一个很难的问题。Q 机器人必须从 A 机器人提供的图像文字描述中寻找线索，并且能够提出有意义的问题。而 A 机器人必须了解 Q 机器人到目前为止究竟理解什么信息，才能帮助 Q 机器人成功。

整个游戏，或者叫任务，常常被称作是“协作性的视觉对话系统”（Cooperative Visual Dialog System）。**这篇文章的主要贡献就是第一个利用深度加强学习来对这样一个系统进行建模，并且，与之前的非加强学习模型相比，提出的解决方案极大地提高了准确度。**

论文的核心方法

那么，既然要把整个问题使用深度强化学习来建模，我们肯定就需要定义强化学习的一些构件。

第一，我们来看看模型的**“动作”（Action）**。两个机器人的动作空间就是自然语言的词汇表。因为，在这个游戏或者说在强化学习的每一轮中，两个机器人都是需要根据现在的状态，来进行下一步的动作，也就是问问题的语句。这是一个离散的动作空间。除此以外，Q 机器人还需要在每一轮之后对自己理解的图像向量进行更新。那么，这是一个连续的动作空间。

第二，我们来看看模型的**“状态”（State）**。对于 Q 机器人来说，每一轮的状态，是一个这些信息的集合，包括最初的 A 机器人提供的图像的描述，以及到目前为止所有轮问答的每一句话。而 A 机器人的状态空间，则包括最初的图像本身，图像的描述，以及到目前为止所有轮的对话。

第三，我们来看看模型的**“策略”（Policy）**。对 A 机器人和 Q 机器人来说，都是要根据现在的状态，来评估下面的语句的可能性。这里，评估的机制其实分别用两个神经网络来学习 A 机器人和 Q 机器人的策略。同时，Q 机器人还需要有一个神经网络来根据现有的 A 机器人的回答，来更新对图像的一个认识。

第四，我们来看一看模型的**“环境”（Environment）和“回馈”（Reward）**。在这个游戏里，两个机器人都会得到一样的回馈，而这个回馈的根据是 Q 机器人对图像的认识所表达的向量和图像的真实表达向量的一个距离，或者更加准确地说是距离的变化量。

以上就是整个模型的设置。

那么，我们来看两个模型策略神经网络的一些细节。首先，对于 Q 机器人来说，有这么四个重要的部件。第一，Q 机器人把当前轮自己问的问题和 A 给的回答，当做一个组合，用 LSTM 进行编码产生一个中间变量 F。第二，当前步骤的 F 和以前的所有 F 都结合起来，再经过一个 LSTM，产生一个中间变量 S。然后第三步，我们根据这个 S 来产生下一步的语句，以及当前对图像的一个重新的认识。也就是说，**F 其实就是一个对历史所有状态的描述，而 S 则是一个压缩了的当前描述信息，并且我们使用 S 来作为下一步的一个跳板。**A 机器人的策略神经网络的架构非常类似，这里就不赘述了，区别在于不需要去产生图像的理解。

整个模型采用了目前深度强化学习流行的**REINFORCE 算法**来对模型的参数进行估计。

这篇文章其实有不少技术细节，我们在今天的分享里只能从比较高的维度帮助你进行总结，如果有兴趣一定要去阅读原文。

方法的实验效果

作者们在一个叫 VisDial 的数据集上做了实验。这个数据集有 6 万 8 千幅图像，是从我们之前提到过的 COCO 数据集里抽取出来的，并且提供了超过 68 万对问答。可以说这个数据集还是比较大型的。

文章比较了利用普通的监督学习以及“课程学习”（Curriculum Learning）的方法。从效果来看，强化学习的效果还是很明显的。**最直接的效果是，强化学习能够产生和真实对话相近的对话效果**，而其他的办法，比如监督学习，则基本上只能产生“死循环”的对话，效果不理想。不过从图像提取的角度来讲，强化学习虽然比监督学习的效果好，但是差距并不是特别明显，基本上可以认为目前的差距依然是在误差范围内的。

小结

今天我为你讲了 ICCV 2017 的一篇有意思的文章。这篇文章介绍了如何利用深度强化学习来搭建一个模型去理解两个机器人的对话并能够理解图像信息。

一起来回顾下要点：第一，我们简要介绍了这篇文章的作者群信息。第二，我们详细介绍了这篇文章要解决的问题以及贡献 。第三，我们重点介绍了的文章提出方法核心内容 。

最后，给你留一个思考题，你认为把强化学习用在这样的对话场景中，难点是什么？

欢迎你给我留言，和我一起讨论。

上一篇　049 | 精读2017年ICCV最佳学生论文

下一篇　内参特刊 | 和你聊聊每个人都关心的人工智能热点话题

精选留言 (1)

写留言

林彦
2018-01-28

强化学习里累积奖赏的状态-动作值函数如何获得。对话后对状态的改变和后续动作的选择造成图像与推测的差距缩小或放大，差距的改变，特别是改变的值范围很大时，如何转换成合适数值的奖赏，期望有相应的理论支持。

展开