专访 | 英特尔刘茵茵：持续优化NLP服务，助推人工智能创新和落地

Original 2018-04-27 Synced 机器之心

机器之心原创

作者：邱陆陆

> 去年六月，英特尔人工智能产品事业部（AIPG）数据科学主任、首席工程师刘茵茵在机器之心主办的第一届全球机器智能峰会（GMIS 2017）上发表了《演变中的人工智能，与模型俱进》主题演讲，探讨了深度学习如何用同一种模型为不同行业提供解决方案，以及如何让各个行业的专家建议推动整个人工智能生态系统的发展。会后，刘茵茵也接受了机器之心的专访，分享了英特尔在 AI 领域的整体规划，以及 AIPG 部门如何计划通过构建相应的框架、资源库等实现这一目标。

日前，机器之心受邀参加了由英特尔与 O'Reilly 联合主办的中国人工智能大会，并再次与刘茵茵进行了深入的对话，我们以英特尔在自然语言处理方面的工作为切入点，聊了聊英特尔是如何构建自然语言基础模块能力，为企业用户提供人工智能服务的，以下是对话实录。

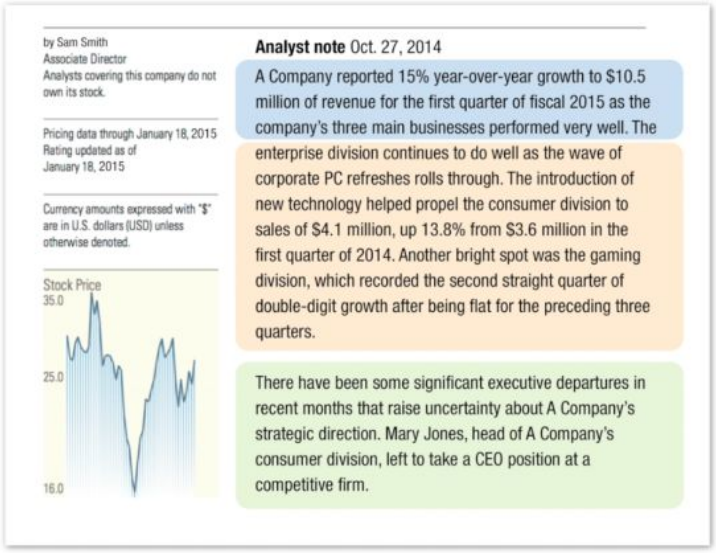
自然语言处理用例：主题分析、趋势分析与情绪分析

机器之心：很高兴再次见到您！如今一年时间过去了，AIPG 也完成了很多工作，尤其是在自然语言方面。您在演讲中提到了三个客户案例，分别是主题分析、趋势分析和情绪分析。首先，能否从这三个案例出发，为我们介绍一下 AIPG 定义问题与解决问题的流程呢?

刘茵茵：首先是主题分析。主题分析的主要目标是为需要处理大量专业领域文档的客户进行以段落为单位的主题连接，让客户能够集中阅读自己感兴趣的、与自身工作相关的内容。深度学习网络接收文档句子/ 作为输入，然后将其映射到数十个主题上，输出给用户。

我们的团队先和客户进行沟通，了解其应用场景并确定主题：在实际工作中，他们需要处理的文档数据是什么样子的？有哪些资源可以辅助数据标注过程？此外，还要了解实际应用过程中的数据流程（pipeline）、延迟要求、存储要求，最终根据所有的需求从英特尔的整套工具中选择模块，通过一些设计转化成算法，再转化成一套整体方案。

Document Understanding & Knowledge



机器之心：在获取训练数据方面，主题分析并不是一个天然存在有标注数据的任务，如何在初期收集有标注的训练集呢?

刘茵茵：首先我们鼓励客户在初期可以做一些数据方面的投资，这样可以更有效的利用英特尔多样的深度学习产品。充足的数据相当于一个能够尝试多种算法的环境，我们可以使用多种算法进行试验，找到效果最好的方法。其次，在标注数据不足的情况下，也可以利用无监督学习方法进行预训练。尤其是在自然语言领域里，语言的连贯性特点使我们可以根据其上下文关系进行无需额外标注的特征提取和特征学习等无监督预训练。英特尔也确保在框架里支持各类不同的训练方式。

机器之心：趋势分析解决哪些问题呢?

刘茵茵：趋势分析的分析对象是文本库，目标是从文本库中提取关键的名词短语（noun phrase），然后通过衡量每个短语的相关性和重要性并进行加权打分，来比较不同文本库之间的趋势与变化。我们已经将算法用于学术期刊的趋势分析，旨在让初学者，尤其是刚刚开始研究深度学习的数据科学工作者能够看到领域里一些概念在学术期刊汇总的热度与趋势。算法也可以用于其他领域，例如产品分析、市场分析、热门话题分析，都是理想的应用场景。

机器之心：情绪分析的应用场景有哪些?

刘茵茵：影视作品评论是一个比较直观的例子。也有很多合作伙伴其希望在商业角度进行产品评估或者是竞争对手分析。因此模型将用户评价作为输入，在进行语法结构分析（parsing）的基础上，进行命名实体识别（name entity recognition），然后通过名词和形容词连接，给出细粒度的（fine-grid）、多方面的评价分析，进而为合作伙伴提供明确的反馈以优化自己的产品设计。

机器之心：在使用深度学习模型完成这些具体用例的过程中，有哪些不一样的体会呢?

刘茵茵：一个是深度学习模块的可共享性。实际上，主题分析、趋势分析和情绪分析模型都是由我们的自然语言堆栈中的众多底层模块组成的。例如，趋势分析的第一个步骤是「名词短语提取」（noun phrase extraction），其当前最优（state-of-the-art）的模型结构是「词嵌入（word embedding）+ 深层 LSTM」，而这也是主题抽取任务中最常用的模型结构，更是情绪分析中语法结构分析的当前最优模型。因此，虽然目标不同、功能不同，但模型中的非常多模块是能够共享的。模块的可共享性让我们在每做一个客户案例的同时都为建立 NLP 能力堆栈积累了非常可观的结构经验，而作为企业用户，如果团队对模块的应用有基本的了解，也会很快利用同样的模块来搭建新的方案。

另外一个则是领域的专门性。例如在情绪分析中，数据科学人员观察到的一个非常有趣的现象是，在不同的领域中，同样的形容词可能表达截然不同的意义。可能一个形容词在形容影视作品时是褒义词，然而如果用来形容产品可能就变成了贬义词。因此，设计出有效的机制，能够引入领域内的专家来对模型进行领域专门的调整，也是非常重要的。

AIPG 的角色：完成数据科学与技术服务，最终提供开源组件与能力

机器之心：英特尔的数据科学专家和领域内的专家在用户案例的设计与搭建过程中分别承担着怎样的角色?

刘茵茵：人工智能仍然处在起步阶段，算法能够触及的领域在不断扩展，而很多领域专家也刚刚开始逐渐理解如何利用 AI 帮助他们解决实际问题。因此，在进行方案设计前，我们要和领域专家进行多次沟通，理解他们的问题，确定 AI 是否可以帮助解决这个特定问题；如果不可以，是否可以将问题转换一下，变成一个当前的 AI 可以帮助解决的问题。

领域专家在这个过程中贡献出自己对业务的理解：希望从哪种角度收集数据，希望看到算法给出何种结果。有时候，他们需要的并不是情感分析或者趋势分析这种已经有成熟定义和解决方案的模型，而是结合不同的深度学习模块，组合成一个他们需要的全新的东西。英特尔在此基础上进行数据科学工作和技术服务，在了解了问题之后，判断何种模型可以帮助他们，再提供算法设计，并将整个算法连接到英特尔的深度学习框架乃至硬件上面。

机器之心：自然语言问题对框架乃至硬件层面提出了哪些独特的需求呢?

刘茵茵：自然语言处理是一个很有挑战、很有发展空间的领域。大部分自然语言任务需要用递归神经网络（RNN）处理时间序列（temporal sequence）、进行循环展开，这是一个很难并行的过程，因此在硬件方面，对从内存中快速提取数据的能力、内存能够支持的模型容量等都有较高要求；在框架方面，也有与可并行模型截然不同的优化需求。所以英特尔在高层直接优化（HLO，提供多核架构优化）和 nGaph（提供框架和底层硬件连接优化）层面都会对众多 NLP 模型进行持续的优化和基准衡量（benchmarking），确保其在硬件以及框架层面获得最好的支持。

机器之心：您在演讲以及刚才的采访中多次提到了「堆栈」的概念，能具体解释一下「堆栈」是什么吗?

刘茵茵：「堆栈」与其说是一个模块集合，不如说是一种看待 NLP 问题的观点和认识。英特尔数据科学团队和研究团队自成立以来，以 AI Lab 的形式解决了许多方面的问题，在计算机视觉、NLP、机器人学习乃至时间序列学习方面都积累了许多的能力。

在 NLP 方面，我们希望在积累了大量经验，有了自己的理解后，能够把不同组件组合在一起，可以通过英特尔的直接优化或者 nGraph，以开源库的形式返回给公众。无论是机器翻译、命名实体识别还是主题分析，都能够通过开源的框架，以平台的形式将做法示范给大家。

英特尔的战略目标：充分利用软硬件联合优化优势

机器之心：目前，AI Lab 有哪些主要目标，又有哪些典型用户？

刘茵茵：AI Lab 主要致力于开发具有创新性的算法，进行创新性的研究。它的目标有以下几层。首先我们希望能够自行进行新算法的研究，数据科学人员在应用最新的、最好的算法的同时，也会产生众多的关于如何改进这些算法的想法，并且希望把它们变成现实。下一个目标是，将算法推荐给

合适的用户，用以解决一些之前无法解决的事记问题。现在有一些合作伙伴来自英特尔内部，例如之前我们帮英特尔的制造部门，对晶片图像进行分类和分割，用以检测晶片内部是否有缺陷。基于深度学习的方法能够同时提高传统方法的速度和准确率。在今后我们也会将用例以论文的形式分享出来。

机器之心：英特尔 AI Lab 在中国进行了哪些实践？AI Lab 期望未来在中国获得何种发展?

刘茵茵：英特尔在中国非常活跃地参与了众多讨论。在中国，有很多研究所、研究院以及大学，通过各种方式了解到英特尔正在进行的应用研究。英特尔为他们提供了在软件框架和最新的优化算法方面的一些支持，帮助学者了解如何在原型的基础上扩展模型解决实际问题，而他们也为英特尔提供了一些特别的数据与用例。

如今，众多英特尔硬件产品被广泛应用在各行各业中，如果我们能够充分地理解这些硬件的长处、短处，适用之处，然后设计出能够根据其特性有效地实施和部署的方案，就可以高效地把一些早期的好的想法变成最终可以解决问题的方案。

机器之心：现在一个普世的观点是，数据、算法和计算力是 AI 实践上的三个关键节点，这三个方面重要性相当，且很难用一方面的长处弥补另一方面的短处。英特尔在这几方面的有哪些优势?

刘茵茵：这三个元素都是非常重要的，也是需要紧密结合的。英特尔人工智能产品事业部的数据科学家，不单单是在算法方面有丰富的经验，也能够将算法与算力紧密契合，找到最适合特定应用场景的组合。

数据则永远是一个非常关键，也非常棘手的部分。很多时候我们要想办法如何能够不局限于监督学习，充分利用无监督学习，例如在数据使用方面，可以努力找寻一些隐藏的数据来源和数据关系来加强无监督学习、配合监督学习。

前瞻：从学界到业界，以及英特尔未来一年规划

机器之心：过去的一年里，学术界有哪些新的方法或者趋势让你觉得会对自然语言处理的实践应用产生新的影响?

刘茵茵：一个是名为「稀疏」的做法。很多时候人们发现密集型的深度学习网络能够被更大、更稀疏的模型所取代。这些大而稀疏的模型，在各种软件和硬件良好配合的前提下，能够极大提升最终的准确度。这样的模型虽然稀疏，但是需要的内存并不会因此减小，尤其是大模型通常倾向于与维度更高的数据配合，这要求大型的存储密集型的硬件对模型进行支持。

例如在英特尔和浙江大学合作的医疗影像案例中，如果内存方面受到较多限制，就必须把 CT 影像切割成小块，在看不到全局的情况下完成分类、分割算法。然而当采用英特尔至强处理器来做，就可以对 2D 全影像乃至 3D 影像行处理，大型的数据加上诸如 U-net 这类大型的深层神经网络，久而久之，就会大幅度提升精准度。

另外，自然语言相比于计算机视觉还有更大的上升空间，诸多基于深度学习的视觉算法都可以转而应用到自然语言处理上。比如说计算机视觉中常见的「风格迁移」任务，也可以在离散的、不连续的自然语言数据上进行。

自然语言处理方面和增强学习方面还是有很大的发展空间，最近经常看到一些多模态数据，比如图像的文本描述，就能利用增强学习训练一个行动器（agent），逐渐了解如何认知图像中的一些概念并且能够描述出来。我认为这是非常基础而有用的研究，因为它不再将图像和语言作为单独的问题处理，而是将图像、语言等通过各种传感器集合在一起进行输入。

机器之心：AI Lab 在新的一年有哪些计划呢?

刘茵茵：在研究层面，我们希望把一些研究成果通过发表论文、分享白皮书或者开源案例的形式分享给其他的研究员或者从业者。在 NLP 方面，也有很多正在进行中的研究，其中主要致力于搭建一个较为全面的堆栈，为使用英特尔软件与硬件的用户提供一个能力层。希望未来一年能够更多把成果分享给大家。 SYNCED