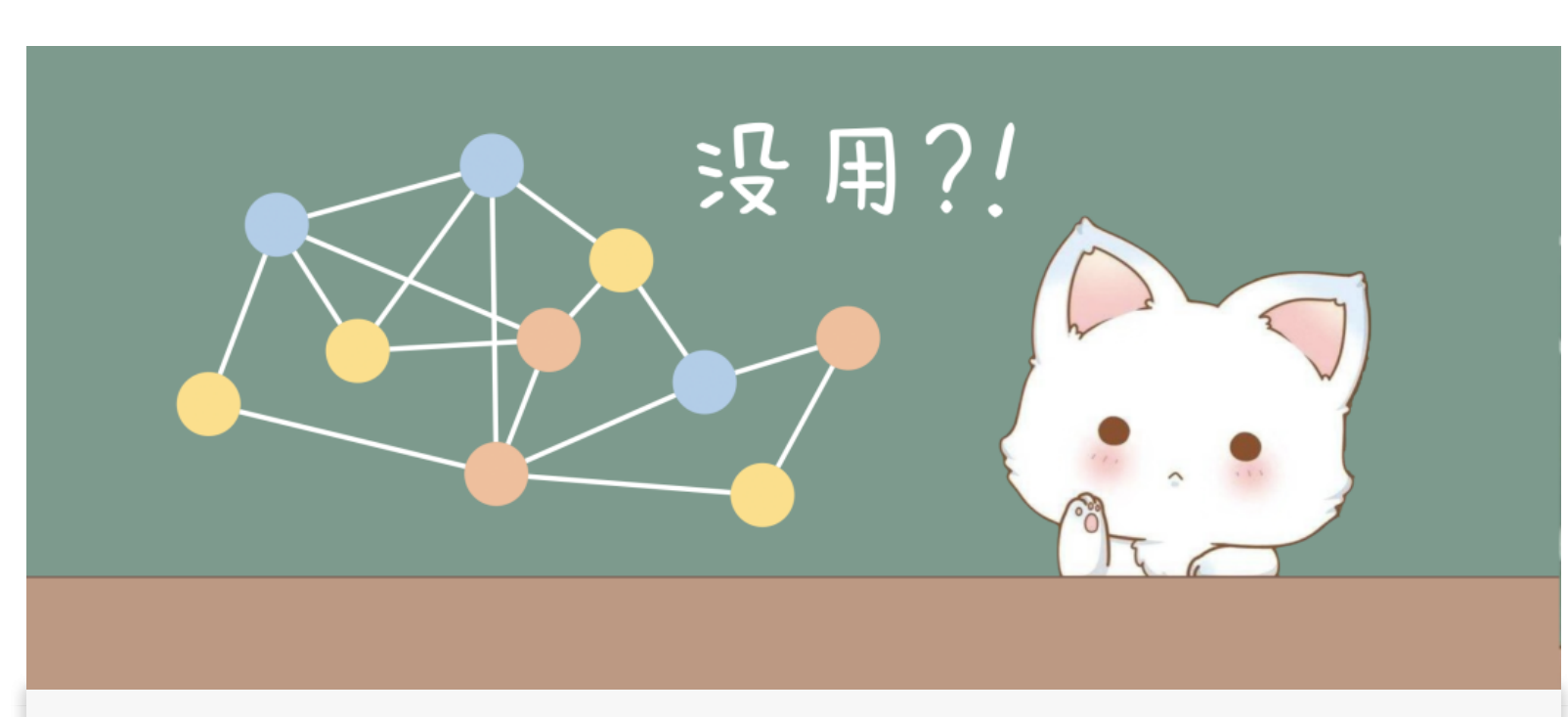
你的 GNN，可能 99% 的参数都是冗余的

原创 iven 夕小瑶的卖萌屋 2021-10-27 12:05



文 | iven

自从图卷积神经网络（GCN）面世以来，图神经网络（GNN）的热潮一瞬间席卷 NLP。似乎在一切 NLP 任务上，引入一个图结构，引入一个 GNN，就能让模型拥有推理能力。更重要的是，似乎在实验结果上，也能证明 GNN + NLP 的有效性。

具体地，GNN + NLP 可以分成以下两类任务：

在本来就需要图的任务上，比如知识图谱问答（KBQA），大家从问题和答案中抽取关键实体，从知识图谱中将这些实体，以及及所有路径提取出来，作为知识图谱针对这个问题提取出的子图，在这上使用 GNN 进行推理。

在本来没有图的任务上，比如文档级的抽取或者理解任务，大家将文档中的关键实体作为节点，并用一些简单的规则连边（比如，在同一个句子里的实体连边、指代同一个概念的实体连边，等等），得到一张文档对应的图，在上面用 GNN 推理。

看起来建图是有用的，可接下来，为啥一定要用 GNN 呢？最近的文章里，人们都说 GNN 有“推理能力”，即 GNN 在图上的信息传播过程相当于在图上找路径，这些路径可以解释答案得到的推理步骤。

在 KBQA 任务里，GNN 能在图中挑选出从问题实体到答案的推理路径。比如：

这里提问：哪里能找到有电梯的地下室呢？衣柜、教堂、办公楼，三选一，答案显然是办公楼。在这个 case 里，模型预测出了 elevator → building → office building 和 basement → building → office building 两条路径，这看起来都能解释答案的选择逻辑。

在文档级关系抽取任务里，GNN 的推理路径就表示了关系的传递。比如：

图中右边是文本，大概讲的是二战中几个国家军队的故事，下划线的是实体；左边是 World War II 这个实体，在两层 GNN 中，分别对所有实体的注意力权重。这样我们也能得到与前一个例子相似的一些推理路径：World War II → New Ireland → Japan 和 World War II → New Ireland → Imperial Japanesae Army。这也解释了为什么 Japan 与 World War II 是“参与者”的关系。

但是，问题来了！推理的必要条件其实只有一张图，我们真的需要 GNN 中复杂的节点与边的表示、复杂的信息聚集和传播，才能“推理”吗？

今天我们带来一篇杨迪一老师领衔的文章 *GNN is a Counter? Revisiting GNN for Question Answering*。这篇文章表明，GNN 被我们滥用了，推理真的不需要 GNN！简简单单的一个计数器就足够！

论文题目：
GNN is a Counter? Revisiting GNN for Question Answering

论文链接：
https://arxiv-download.xixiaoyao.cn/pdf/2110.03192.pdf

GNN 真的有用吗

在介绍这篇文章之前，我们还是先来回顾下在 KBQA 问题上，大家用 GNN 的做法。

KBQA 的主要知识来源有两个方面：预训练模型中隐含的知识、知识图谱中显式的知识。为了用上预训练模型的知识，大家用预训练模型作为 encoder，得到实体和问题的表示；为了用上知识图谱中的知识，大家从知识图谱中抽取问题相关的子图。接下来将节点表示、边的表示作为输入，过几层 GNN，得到优化的节点表示，最后送给分类器分类。

为了探究有没有必要使用 GNN，作者使用 Sparse Variational Dropout (SparseVD) 给 GNN 的网络结构解剖。SparseVD 原本是用来寻找网络结构中，哪些参数是不重要的，以此对模型进行剪枝和压缩。在这篇文章中，作者使用 SparseVD 探寻 GNN 中各层对推理过程的贡献，sparse ratio 越低，代表这些参数越没用。

作者在之前的 SOTA QA-GNN[1] 上进行剪枝，得到的结果令人震惊：

随着训练的推进，GNN 前面节点的 embedding 层越来越没用，但边的表示一直对最后的预测准确率有很大影响。

这张图表明，不仅节点 embedding 层参数没用，节点的初始化也没用。甚至作者在其他模型中也对节点初始化剪枝，发现所有方法里都没用！

在第二层 GNN 上，图注意力模块中的 key 和 query 完全没用，只有 value 比较有用。那么图注意力计算注意力权重不也就完全无效了吗？注意力模块注意不到任何元素，这和一个线性变换又有什么区别呢？

综上所述，似乎 GNN 里面很多部分都是不需要的？为了证明这一点，作者设计了一个 GNN 的简化版本 —— graph soft counter。

简单的 counter，一样有效

通过上面的实验我们可以发现，GNN 中边的表示，以及信息传递和聚合，都是很重要的；其它诸如图注意力、节点表示，都可有可无。于是作者只留下两个结构：Edge encoder 用来构建边的表示，Graph Soft Counter layer 用来做信息传递和聚合。

Edge encoder 是最简单的两层 MLP，输入边的 1-hot 表示 $[u_s, e_{st}, u_t]$。其中 u_s, u_t 表示四种节点类别，e_{st} 表示 38 种边的类别（这里的 38 种是 17 种关系类别，加上问题/答案的边，以及所有类别的反向）。MLP 最后就输出一个 [0,1] 之间的 float 数字，作为边的表示。

Graph Soft Counter layer（GSC） 完全遵照了 MPNN 信息聚合与传播的思路，并且这是无参数的！具体步骤如下图所示，一层 GSC 包含两步，即先将节点的值加到边上，再将边的值加到节点上。

对，就是这么简单的一个模型！参数还不到 GNN 的 1%！

边的表示的维度是 1，因此这个表示就可以被看做边的重要性分数；GSC 的信息聚集，因此也能被看做“数数”：数一数边两端的结点有多重要，数一数结点周围的边有多重要。

实验

作者们在 CommonsenseQA 和 OpenBookQA 两个数据集进行了实验。CommonsenseQA 需要模型对常识进行推理，而 OpenBookQA 需要对科学知识进行推理。作者们不仅在这两个数据集的 leaderboard 上进行了评测，还基于同一个预训练模型，与前人所有基于 GNN 推理的模型进行了对比。

在 CommonsenseQA 上，GSC（本方法）超过了所有基于 GNN 的方法，在 dev 和 test 上分别由 2.57% 和 1.07% 的提升。

在 CommonsenseQA 的 Leaderboard 上，GSC 排名也非常靠前。这里排在首位的 UnifiedQA，其参数量是 GSC 的 30 倍。

在 OpenBookQA 上，GSC 也有相似的惊人效果，甚至在 leaderboard 上超过了 30 倍参数的 UnifiedQA！

怎么才能证明 GSC 也有推理的能力呢？作者们采用了这样的一个假设：如果 GSC 的预测结果和基于 GNN 推理的模型预测结果比较一致，那么就说明 GSC 也有与 GNN 差不多的推理能力。实验发现，两次 GSC 的结果与 ground truth 的交集有 69% 的重合率（下图第一个），而 GSC 与前面不同 baseline 和 ground truth 也有 60% 左右的重合率，且与基于 GNN 推理的模型重合率更大。这表明 GSC 与 GNN 有差不多的推理能力。

此外，作者还举出一个例子，来演示 GSC 的推理过程。直接通过每一步的分数，我们就能得到推理路径，最终答案节点也得到一个分数，在不同的答案之间就用这个分数做出选择。

思考和总结

这篇文章表明 GNN 中很多模块对于推理都是可有可无的。但似乎这和我们之前的印象有些不同？

大家一直都说，信息在图上传播的路径就是推理路径。GAT 的 attention 权重就是传播信息的一个权重，因此大家在 case study 上看信息传播路径的时候，都是找 attention score 大的，看做信息传播的下一跳。然而本文却表明，attention 这部分参数对于结果几乎没有用？另外，在基于 counter 的模型上，case study 中依然能复现出信息传播的过程。那这是不是说，节点之间的 attention score 没有必要，节点自己的表示就足够了？那 GAT 为什么又会比 GCN 好呢？

GNN 里面到底哪些是有用的参数？推理真正需要什么模块？这些都需要更多的研究和思考。

萌屋作者：iven

在北大读研，目前做信息抽取，对低资源、图网络都非常感兴趣。希望大家在卖萌屋玩得开心ヾ(=･ω･=)o

作品推荐

1. 老板让我用少量样本 finetune 模型，我还有救吗？急急急，在线等！
2. 谷歌：CNN 击败 Transformer，有望成为预训练界新霸主！LeCun 却沉默了...
3. 中文 BERT 上分新技巧，多粒度信息来帮忙
4. 恕我直言，很多小样本学习的工作就是不切实际的

后台回复关键词【入群】
加入卖萌屋NLP/IR/Rec与求职讨论群
后台回复关键词【顶会】
获取ACL、CIKM等各大顶会论文集！

[1] Michihiro Yasunaga, et.al., "QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering", NAACL 2021, https://arxiv-download.xixiaoyao.cn/pdf/2104.06378.pdf

[2] Guoshun Nan, et.al., "Reasoning with Latent Structure Refinement for Document-Level Relation Extraction", ACL 2020, https://arxiv-download.xixiaoyao.cn/pdf/2005.06312.pdf