信息网络专题研究之应用层

一、文献信息

1.论文题目：Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

2.作者：Colin Raffel,Noam Shazeer,Adam Roberts,Katherine Lee,Sharan Narang ,Michael Matena ,Yanqi Zhou ,Wei Li ,Peter J. Liu

3.发表时间：2019

二、问题意义

1. 研究背景及意义

近些年，自然语言处理技术迅速发展，迁移学习是一种强大的自然语言处理技术，它首先对模型进行数据丰富任务的预训练，然后再对后续任务进行微调。迁移学习的有效性带来了方法、方法和实践的多样性。而技术的快速进步和多样性，使得比较不同的算法、梳理新贡献的效果和理解现有的转移学习方法变得困难。出于便于理解的需要，本文提出了一种统一的迁移学习方法，使我们能够系统地研究不同的方法，并突破该领域的当前限制。

2. 主要研究问题

1）提出一个通用框架 T5，建立模型

2）进行大量的实验，包括预训练的试验，对比结果，给出推荐参数

3）公开 C4（The Colossal Clean Crawled Corpus）数据集

三、思路方法

本文工作的基础设想是每一个 NLP 问题当作一个“text-to-text”的问题，即以文本为输入，以生成新文本为输出。本文的“text-to-text”的框架允许直接将相同的模型、目标、训练过程和解码过程应用于每一项任务。利用这种灵活性，可以评估各种基于英语的 NLP 问题的性能，包括问题回答、文档摘要和情绪分类等。同时，可以比较不同迁移学习目标、未标记的数据集和其他因素的有效性，通过扩展模型和数据集来探索 NLP 迁移学习的极限。

然后，系统地研究比较了培训前的目标、预训练目标、体系结构、未标记的数据集、迁移方法和其他因素对十个自然语言理解任务的影响，此外，还引入一个新的数据集：Colossal

Clean Crawled Corpus，名为 C4。该数据集引入的初衷是为了探索尺度规模(包括模型规模和数据规模)在 NLP 中的影响

本文其实并没有引入新的模型或者新的方法，而是将现有的方法和技术做一次集大成，进行统一。

由于内容较多，下面主要对 T5 模型与实验进行总结

1）Text-to-Text Transfer Transformer（T5）模型：

(1)输入序列 被映射为一序列的嵌入表示

(2)将嵌入表示输入到编码器。编码器由一堆块组成，每个块由自注意力层和前馈神经网络层组成。每个块中的子层都加了层归一化和残差连接。解码器与编码器很类似，所不同的是解码器在每个自注意力层后面紧接一层标准的注意力层。这层标准注意力层将编码器的输出注入到解码器中。解码器中的自注意力使用的是自回归方式或者因果自注意力。这意味着，在解码器中模型只能看到历史过往的输出信息，不可窥探未来。

(3)最后解码器中的输出输入到一个全连接层(最简单的线性变换)，再接一个 softmax 函数。

2）实验

本文实验真详细对比分析了各种相关因素：预训练目标、模型框架、无标注数据等。

文中选用的 baseline 模型与 BERT_base 很接近，模型采用标准的 transforme。

对不同的框架（Encoder-decoder、Language model 和 Prefix LM）进行了探讨对比，结果是结果是 Encoder-Decoder 结构效果最好。Encoder-Decoder 型，即 Seq2Seq 常用模型，分成编码器和解码器两部分，对于编码器部分，输入可以看到全体，之后结果输给 编码器，而解码器因为输出方式只能看到之前的。Language model 相当于 Encoder-Decoder 的解码器部分，当前时间步只能看到之前时间步信息，典型代表是 GPT2。Prefix LM 型，可看作是上面编码器和解码器的融合体，一部分如编码器一样能看到全体信息，一部分如解码器一样只能看到过去信息。以上三种结构都是有 Transformer 构成，主要的区别是注意力机制的不同。

在无监督目标函数方面，对 prefix language modeling、masked language modeling (MLM) 和 deshuffling objective 这三种函数进行了对比，实验发现 MLM 较为出色，然后对 MLM 进行了进一步分析。

微调策略上，对比了两种。第一种，dapter layers，在 Transformer 每个 block 中的前馈神经网络后添加 dense-ReLU-dense blocks。新的前馈网络使得输出可以与输入维度匹配。微

调阶段，只有 adapter layer 和 layer normalization 的参数被更新。这种方法的超参数是内部前馈网络的维度。第二种，gradual unfreezing。这种方式是更新模型参数的范围随着随着时间扩大。初始微调时，只有最后一层的参数被更新，训练一段时间后，倒数第 2 层及其之后层的参数被更新，直至整个网络的参数都被更新。应用到本文的框架，这种方式是有所改动的。

多任务学习，本文这里是简化版的多任务学习，并不热衷于多任务之间的参数共享，而是更关注于用同一个时间训练多个任务。对于每个任务需要用多少数据进行训练这个问题，本文探索了三种方案：Examples-proportional mixing、Temperature-scaled mixing 和 Equal mixing，实验发现，多任务训练一般是无法于预训练-微调方法相媲美的。后续还进一步研究了如何缩小多任务训练和预训练-微调的差距，研究了三种方案。最后探讨了规模的影响。

在做了上述对比和分析后，文章中综合一起设计出最优模型。

四、实验结论

本文研究结果提供了一些高层次的视角，提出的 text-to-text 框架将 NLP 进行统一，并详尽地分析了架构、无监督目标函数、数据集、训练方法和规模等因素的影响。而大模型表现虽然好，但不是长久之计，distillation、parameter sharing 和 conditional computation 或许是一条新出路。我们需要一个更有效的方法来学到通用的知识，BERT-style loss 的效率或许不高，并且需要一个衡量预训练和下游任务相似性的方法。

五、启发思考

文章中虽然没有提出一个新的模型与方法，而是提供一个全面的视野，使我们能够系统地研究不同的方法，并突破该领域的当前限制。这也是一种贡献，而且是相当大的贡献。

文中提出了 T5 框架，这个统一的框架不同于当前的实践，其优点在于简单性和强大的性能。在这其中，另一个起到决定性作用的是实验部分。文中从各个角度，系统地对比了各种相关因素，梳理各种因素的对结果性能的贡献，提供了代码、数据集和预训练模型，并且公开了 C4 数据集，正是因为这些可以复现的材料，这片论文才会引起如此大的反响。

论文阅读过程中，庞大而严谨的实验部分给我留下了深刻的印象，大量的图表辅助说明使得读者更容易理解，思路更加清晰，结果更加一目了然。