T5 Summary

沙之洲 2020012408

T5(Text-to-Text Transfer Transformer) 这篇文章提出了一个简单统一的基于 Transformer的 Text to Text 框架，将 NLP 的大多数任务转换成了基于不同 prompt 的 text-to-text 的任务。同时， T5 还提供了新数据集 C4，还提供了对 NLP 领域现有方法党的全面总结和实验。

T5 这篇文章的优点在于，它将 NLP 中的很多任务统一成 text to text 任务，例如，问答，句子分类，情感分析，句子因果分析。这意味着 T5 的模型具有更强的泛化性能，也一定程度上证明了模型学到了真实的知识。除此之外，T5 还做了过于模型结构，unsupervised方法，fine-tuning 技巧上非常细致的研究。不但可以帮助初学者了解每个方法的原理，还能够提供一个全面的视角。

对于T5 这篇文章的缺点，我觉得在比较模型架构的时候，T5 只做了计算复杂度层面上的分析，也缺少具体的实验来比较不同模型结构在具体任务上的表现。除此之外。 T5缺少对于 Fully visible, causal, casual with prefix 的 attention mask 的 intuitive 的理解。同样，对于多任务训练，对于 temperature scale mixing，只是通过实验找到了最优的 T，缺少逻辑上的解释。

对于 T5 的一些可能改进的方向，我认为在 scaling up 一节的讨论中，随着模型大小的增长，在各个任务的性能上有所提升。但是，对于这些具体的小任务而言，过大的模型可能会带来更多的计算开销，更高的成本。是否有可能先 scaling up 大模型提升性能，再通过 teacher student 的模式训练出若干针对不同任务的小模型。这样，大模型学到的关键知识可以被小模型蒸馏得到，同时小模型也意味着能够在更少的卡上运行，以及更快的推理速度，更平易近人的价格。

最后是一些对于课上讨论思路的总结。对于 encoder-decoder 和 decoder 结构的对比，encoder 相当于强行在输入序列和输出序列中间加了一层 latent space，这就导致模型不得不学习 text 到 latent space 和 latent space 到 text 两个映射；而 decoder 结构省去了中间强行设定的 latent space 这个步骤，所以目前来看效果会比 encoder-decoder 更好。除此之外，另一种解释也很有意思，encoder-decoder 相当于完整看完了用户的问题之后，再给出答案，而 decoder 是需要一边读用户的问题，一边预测用户的问题，最后要 auto-regressive 回答用户的问题。类比到现实世界，decoder 不但能够学到和 encoder-decoder 一样的问题回答的能力，还能学到对于用户问题的“未卜先知”的能力，也就是“情商”，自然 decoder 就能有更好的效果了。除此之外，我对于课上提到的语言模型为什么没有推理能力这个问题印象深刻。因为语言模型的训练 loss 只是保证了语言模型输出的句子是通顺的，但是无法保证语言模型的回答是符合逻辑的。例如，对语言模型提问“我有5个苹果，他有5个苹果，我们一共有\_\_个苹果”，语言模型可能认为这里只要填一个数字，这句话都是通顺的。所以解决语言模型推理能力的一个可能办法是通过 loss 对语言模型的推理能力进行矫正。