Paper Reading 11

沙之洲 2020012408

**UnlimiFormer**

1 现有 LLM 主要关注的能力是生成高质量文本内容的能力。但是在 UnlimiFormer 这里，对于 scale up LM，encode chunked input 是一个更重要的能力，因为将 chunked input 编码成更适合 cross attention 的形式似乎会对模型生成效果产生影响。我们是否应该花费一些精力去研究如何让模型 encode 更好？

2 结合上一节课提到的，模型权重中可能保存有知识图谱。是否可以将 UnlimiFormer 里边的 encoded vector 看作是模型的知识图谱？换句话说，是否可以认为 encode chunked input 这个过程，可以实现用自然语言表达的人类知识，向 vector 表达的 LM 能理解的知识的转换？

3 本文中只进行了一层的 encode chunk input，这个过程是否可以做成 hierarchy 的？

4 会有比 knn 更好的 retrieval 方法吗？

Knn 使用的是 kq 得到的权重，但是权重本身并不能反应重要性，要考虑到 V 的模长与 权重的乘积 作为 knn 选择的依据。

**MegaByte**

1 本文提出的观点是将所有模态的数据都看做是 Byte 来进行 encode。但是不同模态的数据的 distribution 并不相同，是否会出现能够很好 encode language 的 model ，encode image 的能力很差的情况出现？不同模态的 encode 的模型知识之间是否会出现冲突？

2 megabyte 和 UnlimiFormer 的区别在于，前者是以 patch 的方式 generation 的，而后者是 token generation。这两种 generation 各有什么优劣？

这种 multi scale 的结构，local model 经常会成为瓶颈。因为在不知道和远的过去和未来的情况下，很难知道应该提取什么样的信息。

Token or not：现在的 token 使用 bpe 的启发式算法做了一个预压缩。但是现在的 architecture 并没有为更好地压缩设计结构。虽然提供更高的无损压缩这个想法是好的，但是现有的实现是在每个 Byte 上分配同样的 attention 这件事情是不合理的。