Week5

Transformer 原理老师已经讲的非常清楚，此笔记对 Transformer 的关键点及优缺点进行简要概览。

Transformer 是 Google 的研究者于 2017 年在《Attention Is All You Need》一文中提出的一种用于 seq2seq 任务的模型，它是首个完全抛弃 RNN 的 recurrence，CNN 的 convolution，仅用 attention 来做特征抽取的模型。原理图见附件中‘Transformer Block 概览图。

Transformer 整个网络结构仅由 Self-attention 和 Feed Forward Neural Network 组成。一个基于 Transformer 的可训练的神经网络可以通过堆叠 Transformer 的形式进行搭建，论文中的实验是通过搭建编码器和解码器各 6 层，总共 12 层的 Encoder-Decoder，并在机器翻译中取得了 BLEU 值得新高。

1. 采用 Attention 机制

采用 Attention 机制是考虑到 RNN（或者 LSTM，GRU 等）的计算限制为是顺序的，也就是说 RNN 相关算法只能从左向右依次计算或者从右向左依次计算，这种机制带来了两个问题：

a.时间片 t 的计算依赖 t-1 时刻的计算结果，这样限制了模型的并行能力；

b.顺序计算的过程中信息会丢失，尽管 LSTM 等门机制的结构一定程度上缓解了长期依赖的问题，但是对于特别长期的依赖现象,LSTM 依旧无能为力。

Transformer 的提出解决了上面两个问题，它使用了 Attention 机制，将序列中的任意两个位置之间的距离是缩小为一个常量；其次它不是类似 RNN 的顺序结构，因此具有更好的并行性，符合现有的 GPU 框架。

2. 使用多头注意力

若使用单头注意力的平均注意力加权，会降低有效的分辨率，即它不能充分体现来自不同表示子空间的信息。而使用多头注意力机制有点类似于 CNN 中同一卷积层内使用多个卷积核的思想。可以增强模型对于文本在不同子空间中体现出的不同的特性，避免了平均池化对这种特性的抑制。

3. 引入位置编码

Transformer 模型并没有捕捉顺序序列的能力，为了解决这个问题，论文中在编码词向量时引入了位置编码（Position Embedding）的特征，此特征为模型捕捉单词之间的相对位置关系提供了非常大的便利。

Transformer 的优势很明显，但也有其缺点。

1. Transformer 更多的关注全局的相关性，局部信息的获取不如 RNN 和 CNN 强。

2. 位置信息编码存在问题

在使用词向量的过程中，会做如下假设：对词向量做线性变换，其语义可以在很大程度上得以保留，也就是说词向量保存了词语的语言学信息（词性、语义）。然而，位置编码在语义空间中并不具有这种可变换性，它相当于人为设计的一种索引。那么，将这种位置编码与词向量相加，就是不合理的，所以不能很好地表征位置信息。

3. 顶层梯度消失

Transformer 模型实际上是由一些残差模块与层归一化模块组合而成。目前最常见的 Transformer 模型都使用了 LN，即层归一化模块位于两个残差模块之间。因此，最终的输出层与之前的 Transformer 层都没有直连通路，梯度流会被层归一化模块阻断，从而导致顶层梯度消失。

4. 不能处理所有问题

例如，当我们需要输出直接复制输入时，Transformer 并不能很好地学习到这个操作。

5. 不适合处理超长序列

当针对文章处理时，序列的长度很容易就超过 512。而如果选择不断增大模型的维度，训练时计算资源的需求会平方级增大，难以承受。因此一般选择将文本直接进行截断，而不考虑其自然文本的分割（例如标点符号等），使得文本的长距离依赖建模质量下降。

6. 计算资源分配对于不同的单词都是相同的

在 Encoder 的过程中，所有的输入 token 都具有相同的计算量。但是在句子中，有些单词相对会更重要一些，而有些单词并没有太多意义。为这些单词都赋予相同的计算资源显然是一种浪费。

参考文献：

1. 详解 Transformer （Attention Is All You Need），大师兄，知乎，
   https://zhuanlan.zhihu.com/p/48508221
2. Transformer 及其变种，蒋润宇，2020， 哈工大 SCIR，
   https://www.jiqizhixin.com/articles/2020-06-28-8
3. Transformer 优缺点分析，人工智能，知乎，
   https://zhuanlan.zhihu.com/p/330483336

附件：

Transformer Block 概览图