目录

对于注意力机制，应当花些精力研究一下。

original idea

2014 年的这篇文章[1] 首次提出了某种注意力机制，原文如此引入：

Each time the proposed model generates a word in a translation, **it (soft-)searches for a set of positions in a source sentence where the most relevant information is concentrated**. The model then predicts a target word based on the context vectors associated with these source positions and all the previous generated target words.

The most important distinguishing feature of this approach from the basic encoder–decoder is that **it does not attempt to encode a whole input sentence into a single fixed-length vector. Instead, it encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively while decoding the translation.** This frees a neural translation model from having to squash all the information of a source sentence, regardless of its length, into a fixed-length vector. We show this allows a model to cope better with long sentences.

读者须自己读一下。

接下来直接跳到 3.1 节，可以看到，式 (5) 中 c_i 的计算实际就是上文提到的 (soft-)search，而由于权重 a_{ij} 还要依赖于解码器的上一个隐藏状态 s_{i-1}，所以这里描述的是一种动态的注意力，实际上要训练的是生成这种动态注意力的全链接层 $a()$（注意这个不是 a_{ij} 那个 a，原文的符号确实挺混乱）

其余部分都可以不读，因为与注意力机制有关的部分只有这些。

这篇 paper 和如今的文章中，对注意力的可视化都是对 a_{ij} 进行可视化，可以说这就是注意力。进一步地，可以将注意力看作一种动态生成的全链接层。

Transformer

接下来就是这篇 Attention Is All You Need[2] 了。

intro 部分提到了另一篇文章[3]，也是不依赖于 RNN 的注意力网络，只不过其工作是用于一种分类问题（自然语言推理），而不是 enc-dec 这种 seq to seq 的任务。

哦对了，接下来得请出那张经典图：

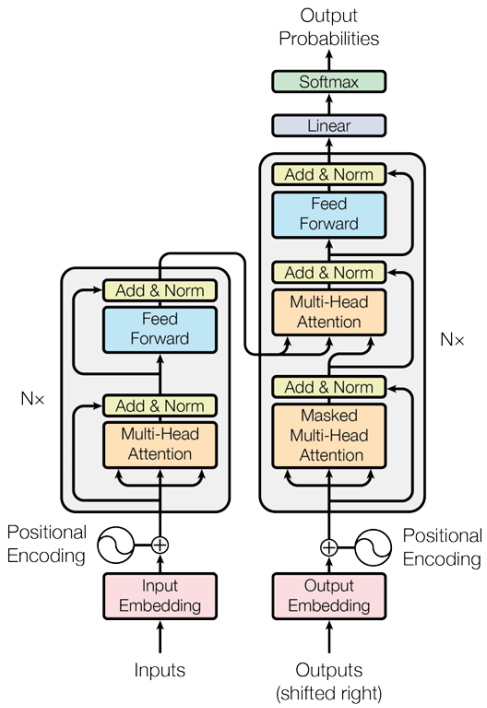

Figure 1: The Transformer - model architecture.

图 1 变形金刚 - 模型架构

batchNorm 和 layerNorm 的区别，比起翻阅原论文，用搜索引擎是更好的选择。

我们上一节谈到的就是加性注意力，而这里的点积注意力其实注意力的部分仅仅是 $\mathrm{Softmax}(QK^{\mathrm{T}})$，$\mathrm{Attention}(Q,K,V)$ 中的 V 并不是注意力本身，而是相当于加性注意力那篇论文里的 $\{h_j\}$

第 3 节中的其余部分就是一些并非细枝末节的细枝末节了，虽然很重要，但还是略过，直接跳到第 4 节。第 4 节基本是对 Table 1 的说明，值得一读。当然，各种 Attention 的变种可以留待以后讨论。

论文的其余部分不做讨论。

参考文献

[1] D. Bahdanau, K. Cho, 和 Y. Bengio, 《Neural Machine Translation by Jointly Learning to Align and Translate》. [在线]. 载于: https://arxiv.org/abs/1409.0473

[2] A. Vaswani 等, 《Attention Is All You Need》. [在线]. 载于: https://arxiv.org/abs/1706.03762

[3] A. P. Parikh, O. Täckström, D. Das, 和 J. Uszkoreit, 《A Decomposable Attention Model for Natural Language Inference》. [在线]. 载于: https://arxiv.org/abs/1606.01933