speech and language processing 第十章报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. P6页“For neural generation, where we are trying to generate novel outputs, we can simply sample from the softmax distribution. However, for applications like MT where we’re looking for a specifific output sequence, random sampling isn’t appropriate and would likely lead to some strange output。”这里为什么MT会有一些不好的结果。？

讨论后的理解：对于neural generation，其目标是产生一个合理的句子输出，也就是说句子符合语法规则并且符合人们的使用习惯即可，而对于MT来说，句子不仅要符合人们的使用习惯，也受到上下文翻译内容以及翻译源文本结构的制约，有着更为严格的形式上的限制，直接使用softmax distribution去直接产生备选项会让翻译很奇怪，所以要考虑其他因素，比如上下文。

1. Beam search上一段中最后一句话说这个方法"invariant doesn't hold"具体指什么意思吗？

讨论后的理解：这个应该是说Seq2Seq模型并不符合动态规划的一些条件，比如其输出不一定是定长的序列，而动态规划往往是一个整齐的矩阵，而且其规模很难运用动态规划来快速的运算。

1. attention机制是否就是指通过给予编码器和解码器一定权重从而使得机器可以更有针对性的学习？

讨论后的理解：attetion我认为其本质其实和我们人类进行翻译的过程是比较类似的，因为两种文本一般有一种对齐的关系，某些部分会对应某些部分，如果使用的是之前那种上下文向量的话，其实是不符合人的思考习惯的。attetion的意义在于其进行了一种对齐操作，比如说根据我现在翻译的信息，我可以通过向量积的形式找到对于接下来一步翻译最重要的原文部分，这样就为每一步翻译有效的提供了其需要的上下文，而且信息缺失也很少

1. 别人提出的问题的理解
2. Attentio机制与直接上下文之间的比较，Attention是否存在缺点。

自己的理解：Attetion机制的优势是考虑了每一个部分对于翻译的影响，给出了具体的上下文向量。但是由于是并行对每一个序列元素进行了相关的运算，这也就意味着输入序列顺序的缺失。一般在NLP和机器翻译中，序列顺序是占有很大的比重的，其往往会对结果造成明显的影响。Attetion这里也仅仅知识考虑了向量相似性这种简单的指标，其得出来的上下文向量虽然更加合理，但是还有这优化的空间。

1. Self-Attention and Transformer Networks这两个概念是指什么？是否是Attention的延申。

自己的理解：Self-Attention和Transformer Networks可以理解为Attention机制的延申，Self-Attention可以理解为是attetion机制的一种，相当于对一个句子进行对自己的attention，比如观察某个词语和句子其他部分之间的联系性，用来分析句子的结构。

Transformer Networks是Seq2Seq的变体，其基本来说就是对于输入序列进行编码，对输出序列进行解码，而编码和解码都使用了attention机制，这样会使得输入序列之间的元素联系的十分紧凑，有着更好的效果。

1. 读书计划

1、本周完成的内容章节：阅读完第10章。

2、下周计划：阅读完speech and processing第11章部分内容。

四、读书摘要总结：

1. Attention:

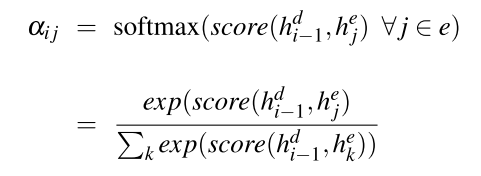
尽管Seq2Seq的结构非常适合处理序列，但是其存在一个显著的问题，上下文向量在绝大多数情况下只能保存大部分后半部分的内容。尽管双向RNN在某种程度上可以缓解这个问题，而且存在一些处理方法，但是attention机制提供了一种新颖的方法来处理这种问题。

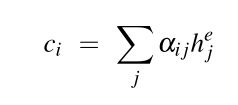


其目标在于为每一步翻译都分配其需要的上下文部分，通过向量积或者引入矩阵乘积的方式，来为每一步的decode的状态计算context向量，而在原始版本中，这个context是一个固定的向量。

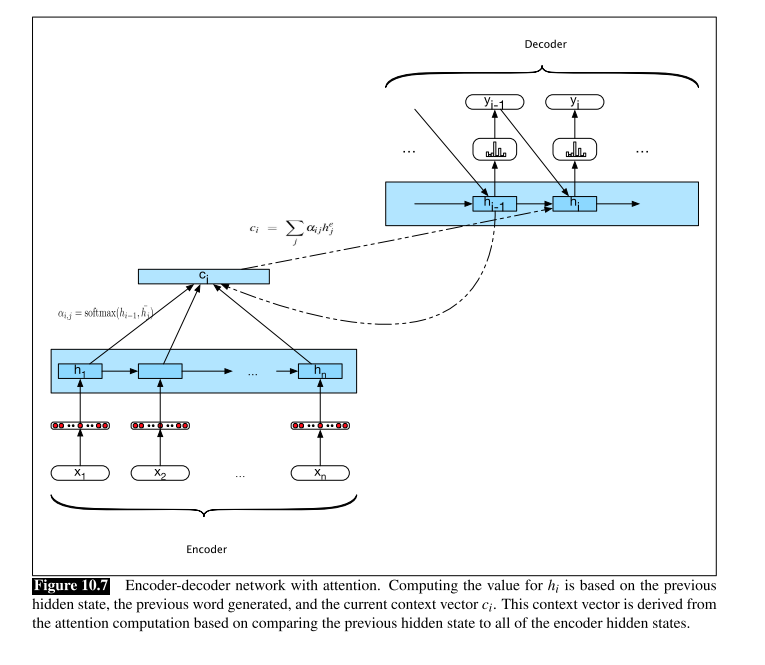


其通过上面的公式比较输入部分对于现在的输出部分是否重要，并且通过加权的方式得到新的context向量。





attention的具体结构可以理解为下图：



除此之外，self-attention以及transformer网络也有着广泛的应用，比如应用广泛的BERT模型就是一种transformer网络。

思考：机器翻译或者说对序列的处理在NLP中一直是重要的工作，并且非常具有挑战性，因为其要求机器理解语言的语法，词法结构以及一些复杂特性。现在的机器翻译比起以前已经改进了很多，但是仍然称不上完美，说明在分析语言的过程中，仍然有着现在研究没有涵盖到的内容，比如引入相关的背景知识，比如寻找更好的上下文之间的联系以及更全面的考虑序列的顺序信息以及词汇的类型信息等等，我认为都会产生新的突破。