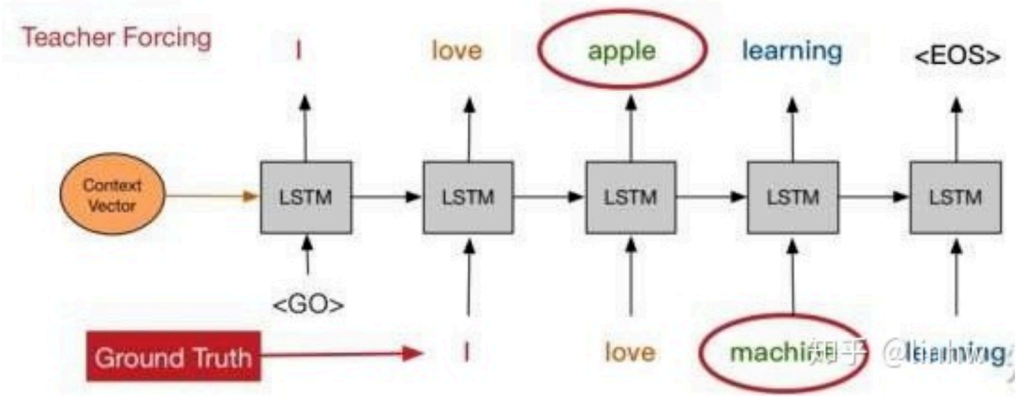
这是我学习CS224N的一些笔记和问题

以下是一些我做的笔记：

1. seq2seq在训练时与测试时有什么不一样？

   seq2seq训练时，decoder的输入是ground truth，而测试时的输入是前一个decoder的输出。

   Teacher Forcing可能会导致exposure bias。



teacher forcing

在机器翻译的模型中通常使用的是seq2seq模型。在seq2seq模型的解码端中，当前词是根据前一个词来生成的，但是在训练时使用的是Teacher Forcing，前一个词是从ground truth中得到的，而在测试时，前一个词是模型自己生成的，这就使得在训练和测试时预测出的单词实际上从不同的分布中得到的，这就是exposure bias。由于exposure bias的存在，在测试时，如果某一步出现错误，那么错误就会一直累积（因为训练时前一个单词总是正确的），最终导致生成不正确的文本。

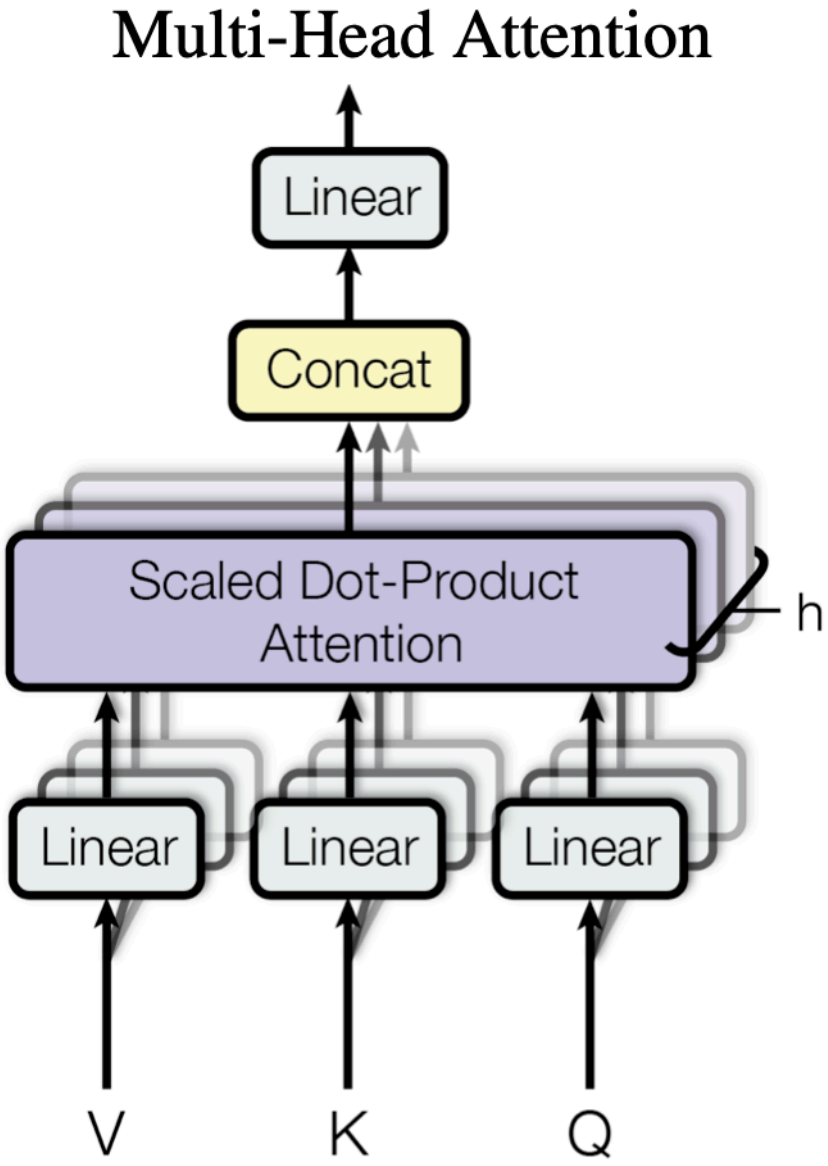
除此之外作者认为在机器翻译的过程中还有另一个问题：overcorrection。即在机器翻译中，损失函数通常是交叉熵，交叉熵函数会严格匹配预测的输出和ground truth是否一致，如果预测的词和ground truth中的词不同，尽管这个翻译是合理的，但也会被交叉熵纠正，这降低了翻译结果的多样性。

2. transformer中encoder和decoder的区别？

   2.1 encoder中有两层sub-layer（multi-head self-attention和fully connected feed-forward），decoder中有三层sub-layer（Masked Multi-head Attention、Encoder-Decoder Attention和Feed Forward）。每个sub-layer都加了Add&Norm（residual connection和layer Normalization）

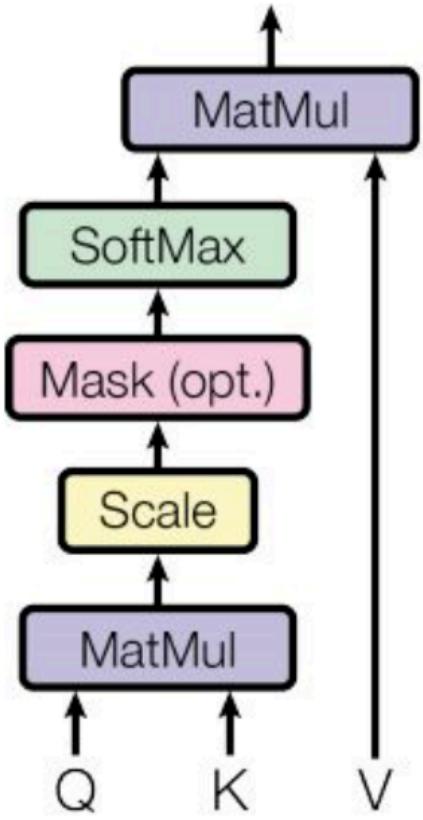
2.2 encoder中是并行计算的，而decoder也得按顺序得出一个一个的结果。

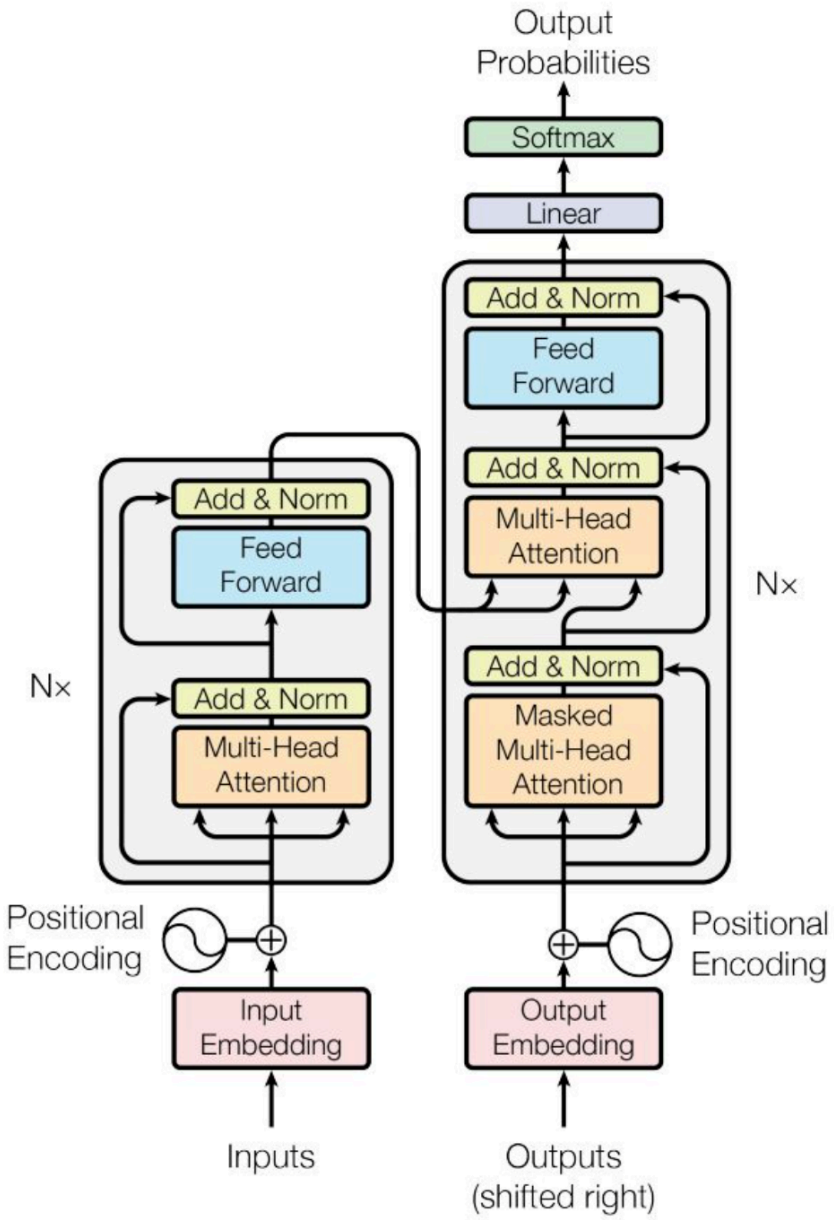
2.3 encoder-decoder attention的输入：encoder的输出和前一个decoder的输出


Multi-Head Attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q

.4

Scaled Dot-Product Attention





3. beam search的思想?

   在Beam Search中只有一个参数B，叫做beam width(集束宽)，用来表示在每一次挑选top B的结果。

4. word embedding如何解决一词多义的问题?

   可以赋予文章中的同一个word不同位置的词一个权重，最终输出的word embedding是一个weighted sum

5. 讲一讲word2vec的negative sampling和hierarchical softmax?

本质是word2vec高效训练的计算框架。

5.1 负采样：选取k个负样本（错误分类的样本），通过k个负样本，将模型复杂度由O(V)降至O(K)

5.2 层序softamx：讲单词以Huffman树的形式组织起来，频率最高的所对应的路径最短，模型复杂度降至log倍数

6. LayerNormalization和BatchNormalization的区别?

6.1 BatchNorm是对同一个batch里的不同data的同样的dimension做normalization，我们希望整个batch里面同一dim的mean=0，variance=1。

6.2 LayerNorm是不用考虑Batch的，给一笔data，我们希望各个不同dim的mean=0，variance=1。LayerNorm一般会搭配RNN一起使用，

7. 做Attention的方式除了Inner Porduct还有别的吗，这些有什么不同?

7.1 Soft Attention：

传统的Attention Mechanism就是Soft Attention,即通过确定性的得分计算来得到attended之后的编码隐状态。Soft Attention是参数化的（Parameterization），因此可导，可以被嵌入到模型中去，直接训练。梯度可以经过Attention Mechanism模块，反向传播到模型其他部分。

也有称作TOP-down Attention。

7.2 Hard Attention：

相反，Hard Attention是一个随机的过程。Hard Attention不会选择整个encoder的隐层输出做为其输入，Hard Attention会依概率Si来采样输入端的隐

状态一部分来进行计算，而不是整个encoder的隐状态。为了实现梯度的反向传播，需要采用蒙特卡洛采样的方法来估计模块的梯度。

两种Attention Mechanism都有各自的优势，但目前更多的研究和应用还是更倾向于使用Soft Attention，因为其可以直接求导，进行梯度反向传播。

8. BERT里面训练LM的随机替换为什么是结果更好?

基于 interpolation 的 LM 平滑

9. 为什么需要position embedding?

因为每一个input都要做一次self-attention，而self-attention的softmax 输出考虑了全部的input，即考虑了全部的信息，那这样对于所有的attention输出，并没有考虑到具体的输入顺序，可以说第一个和最合一个最起来是一样的，所以需要PE来提取输入句子中单词的顺序信息。

10. self-attention的流程?

每一个input embedding分别乘以Q、K、V三个Matrix得到qi、ki、vi，然后拿每个query qi对所有的key ki做attention(一般为inner product) 得到一个attention score，然后通过sofxmax层，得到score-hat，最后每一个与对应的value vi进行矩阵相乘，得到b1，即self-attention该层的输出。注意的是b1-b4可以parallel地计算出来。

self-attention解决了并行计算、长距离依赖的问题

11. 什么是迁移学习

12. 什么是ELMo?

ELMo：Embeddings from Language Models。

在EMLo中，他们使用的是一个双向的LSTM语言模型，由一个前向和一个后向语言模型构成，目标函数就是取这两个方向语言模型的最大似然。

在预训练好这个语言模型之后，ELMo就是根据下面的公式来用作词表示，其实就是把这个双向语言模型的每一中间层进行一个求和。最简单的也可以使用最高层的表示来作为ELMo。

总结一下，不像传统的词向量，每一个词只对应一个词向量，ELMo利用预训练好的双向语言模型，然后根据具体输入从该语言模型中可以得到上下文依赖的当前词表示（对于不同上下文的同一个词的表示是不一样的），再当成特征加入到具体的NLP有监督模型里。

13. 什么是GPT?

目标是学习一个通用的表示，能够在大量任务上进行应用。

然后再具体NLP任务有监督微调时，与ELMo当成特征的做法不同，OpenAI GPT不需要再重新对任务构建新的模型结构，而是直接在transformer这个语言模型上的最后一层接上softmax作为任务输出层，然后再对这整个模型进行微调。他们额外发现，如果使用语言模型作为辅助任务，能够提升有监督模型的泛化能力，并且能够加速收敛。

14. 讲一讲BERT?

BERT这篇论文把预训练语言表示方法分为了基于特征的方法（代表ELMo）和基于微调的方法（代表OpenAI GPT）。而目前这两种方法在预训练时都是使用单向的语言模型来学习语言表示。

BERT最大输入长度为512。

15. fine-tune怎么做?

我们假设在Resnet101后面加上一个全连接层，然后我们锁住前面Resnet的参数，不参加梯度更新，然后只更新最后一个全连接层的参数。当全连接层的

loss足够小的时候，再释放所有的参数一起训练。这样Resnet的参数也会微微调整，这就是finetune；

16. SubWord Model是什么？

    BPE：Byte Pair Encoding，即寻找经常出现在一起的Byte对，合并成新的Byte Pair加入到词汇库。

    课程在这里介绍了介于word-level和char-leval之间的Sub-word models，主要一般有两种结构，一种是仍采用和word-level相同的结构，只不过采用更小的单元word pieces来代替单词；另一种是hybrid architectures, 主要部分依然是基于word, 但是其他的一些部分用characters。

    16.1 Hybrid architecture的核心思想：大部分时候都使用word-level的模型来做translate，只有在遇到rare or unseen的words的时候才会使用character-level的模型协助。这种做法产生了非常好的效果。

    混合模型即两种方式并存的模型，在正常处理时采用word-level的模型，当出现奇怪的词的后，使用char-level级的模型。

17. 什么是high-way Network？与Residua Network的区别？

18. sad