Lecture13 Contextual Word Representations and Pretraining

Reflections onword representations

词的表示：word2vec、fastText，GloVe等
无监督的预训练方法
有了非线性、正则化之后，用于有监督学习中，在神经网络中能更好的工作

在词向量中unk的处理，可以加入字符向量（上节讲到）
Tip： 在测试时遇到新单词，可能是无监督的单词，因为预先训练的word embedding有更大的词汇量，在测试时遇到新词时，而这个词出预训练中出现，就可以直接使用这个词的向量。
Tip： 遇到unk时当场分配一个随机word vector

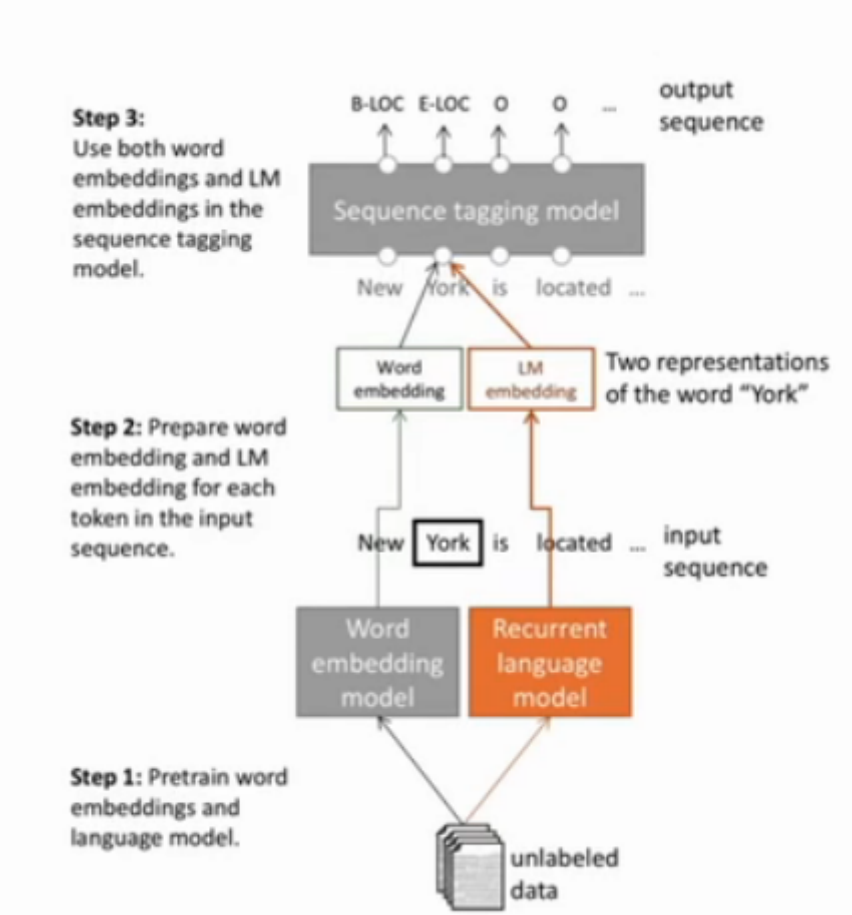
但是词向量有两个问题：

1. 一个词有多种含义。一种解决方法是区分它们的含义并为他们分配不同的单词向量；或许将向量看作是多个的混合，让模型将它们分开。另一种：它们的含义之间可能会存在一些相似，我们想知道在特定环境中单词的含义。
2. 一个词有不同的维度。句法、词性、语法等。单词也有注册和内涵的区别

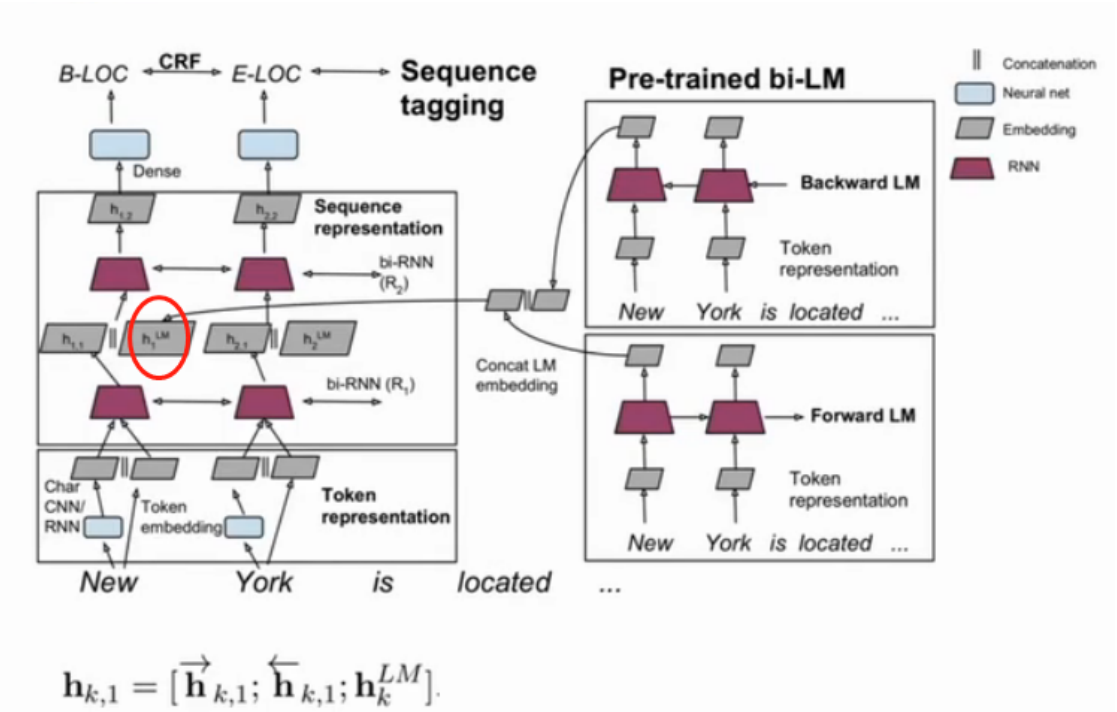
基于以上问题，就需要提出新的上下文词嵌入

Pre-ELMo and ELMO

Tag LM



将未标记的数据分别进行word2vec和RNN，将两个输出表示同时输入到序列标记模型中，使它更好地工作



$$\mathbf{h}_{k,1} = [\overrightarrow{\mathbf{h}}_{k,1}; \overleftarrow{\mathbf{h}}_{k,1}; \mathbf{h}_k^{LM}]$$

上图右侧是预训练的双向的语言模型，得到的每个位置的上下文单词表示（作为参数，无法参与反向传播）与左侧RNN隐藏层得到的表示一起输入到下一层

提升较小

ELMo

- 使用CNN来构建单词表示，减少了存储的参数数量
- 层之间使用残差连接，做了参数绑定
  tagLM仅使用LSTM栈的顶层，而ELMo使用biLSTM的所有层

$$R_k = \left\{ \mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} | j = 1, \ldots, L \right\}$$
$$= \left\{ \mathbf{h}_{k,j}^{LM} | j = 0, \ldots, L \right\}$$

$$\mathbf{ELMo}_k^{task} = E\left(R_k; \Theta^{task}\right) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}$$

全局比例因子gamma：允许控制某些任务中语境词嵌入是否有用
计算每个位置的ELMo表示作为加权平均值，将他连接到隐藏状态并生成输出
可以用于各种任务中（NER、情感分析等）

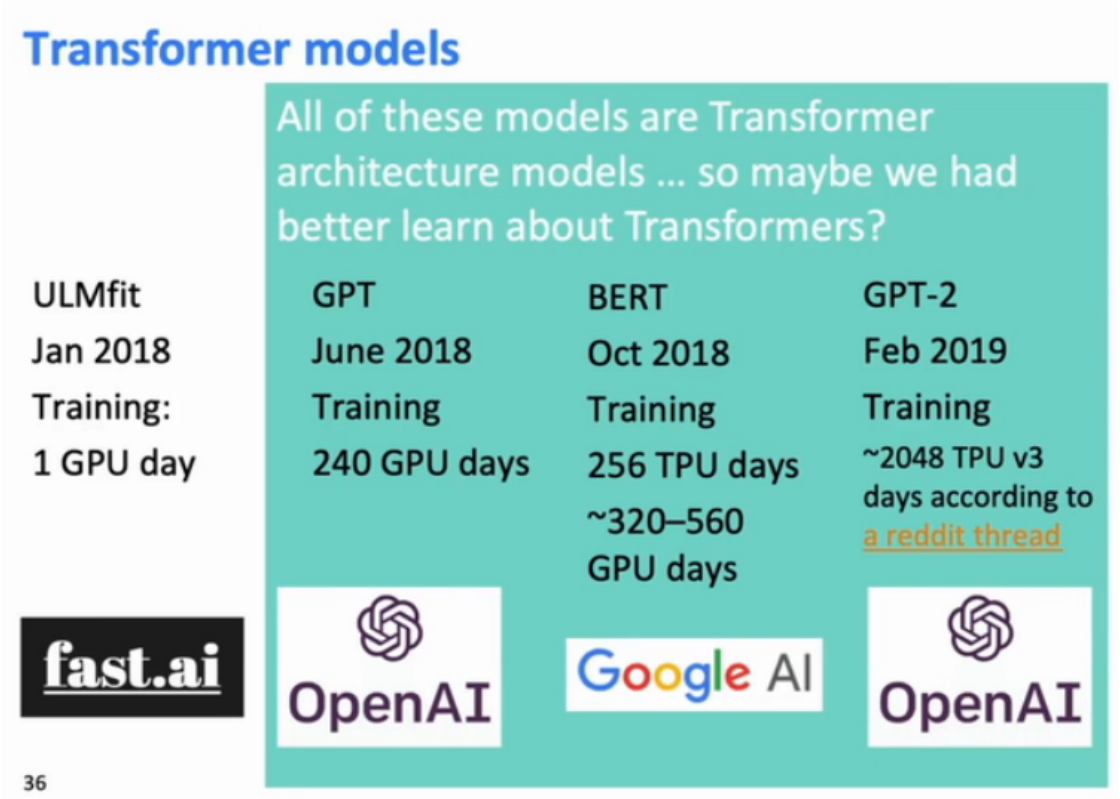
ULMfit anf onward

在一个很大的无监督语料库中训练神经语言模型，对模型进行微调来应用到自己感兴趣的领域，比如用作文本分类器，使用相同的模型，但在最顶层是完全不同的来适应特定的任务。

ULMfit transfer learning

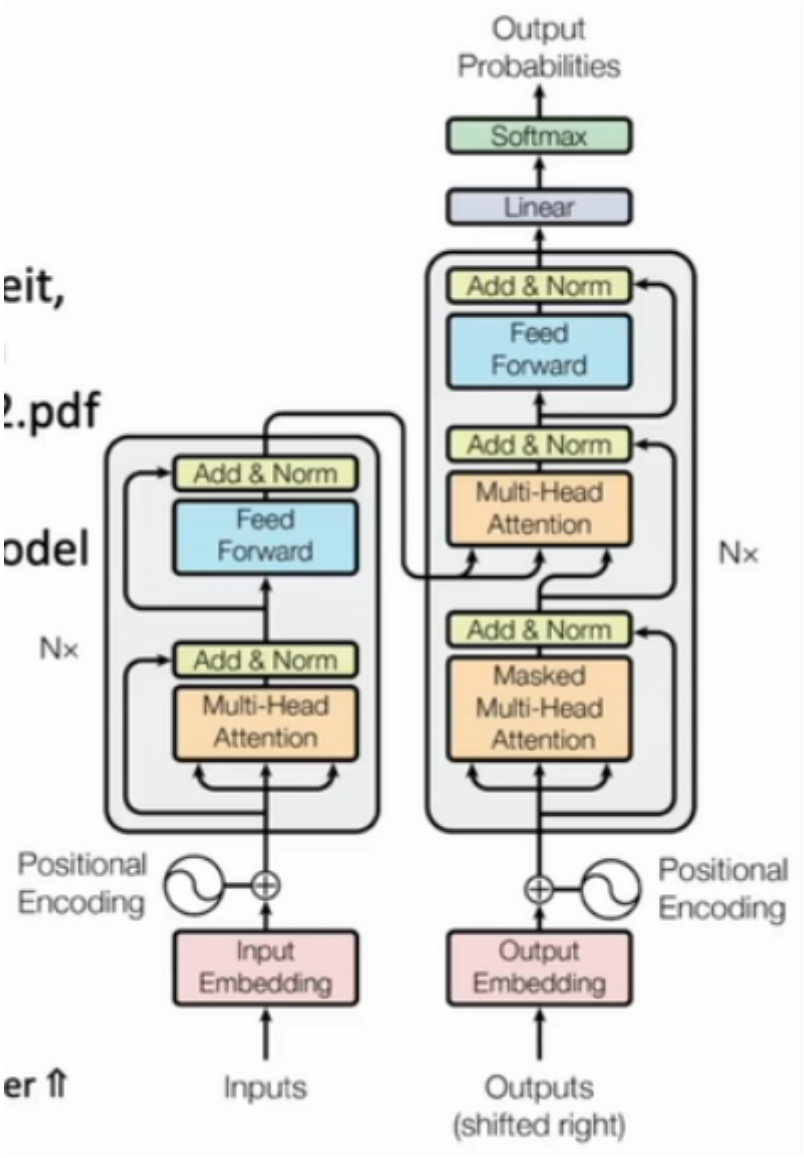
在大量数据上训练这种神经语言模型，在有监督的任务上就恩能够做的更好，即便训练数据很少

Transformer architectures



GRUs、RNNs在长序列上遇到的问题可以通过增加注意力来改进

- Attention is all you need. 2017. Aswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin https://arxiv.org/pdf/1706.03762.pdf

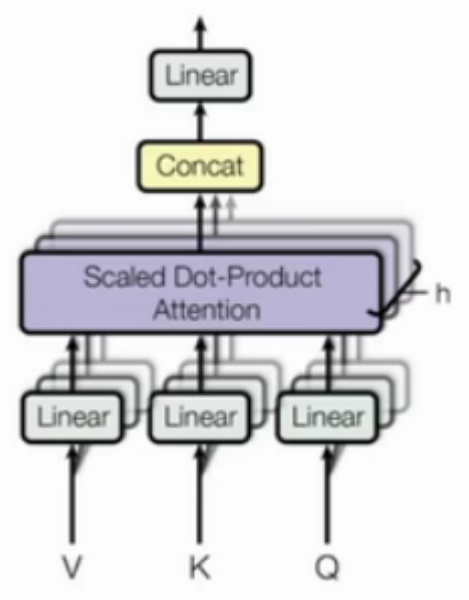


非循环的seq-to-seq编码解码模型用于神经机器翻译
Q,K,V通过超隐藏状态维度的大小进行归一化（细节留待下课）

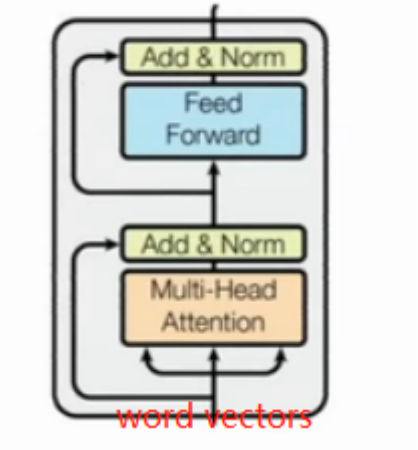
$$A(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Multi-head attention

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

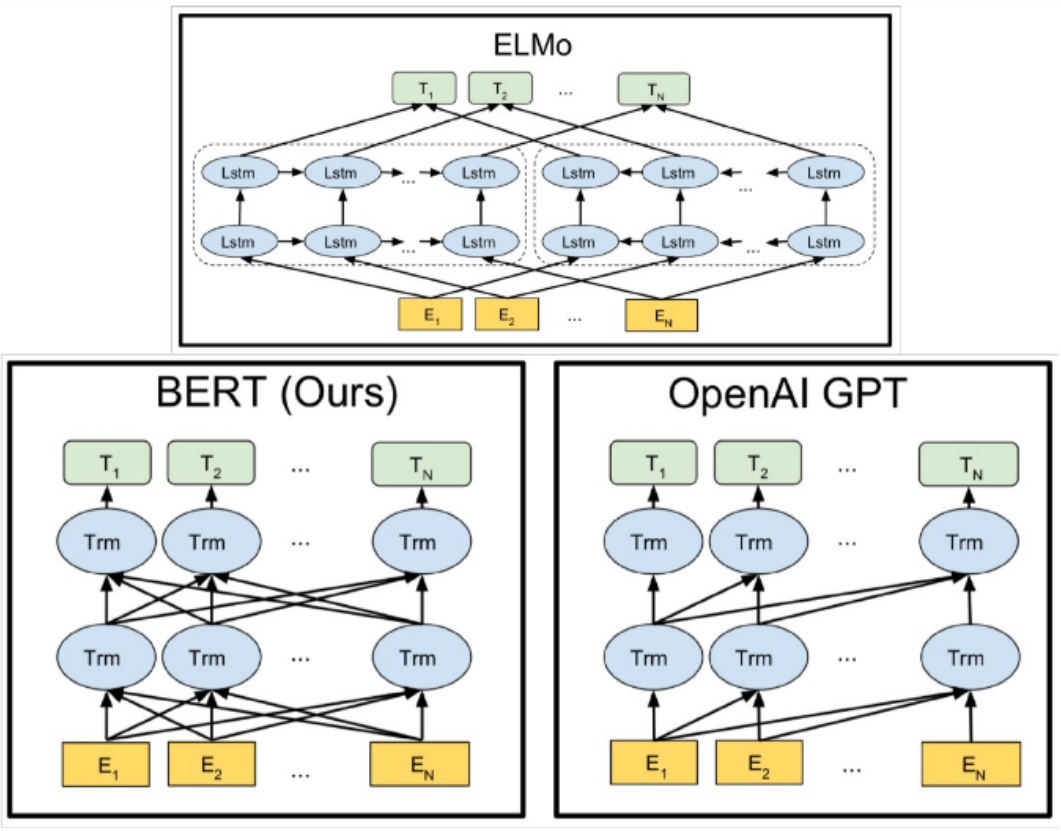


每个block 都有两个“子层”，每个block使用和上一层相同的QKV，重复六次，就可以开始在序列上逐步推送信息来计算感兴趣的值
多头attention有两层的前馈神经网络，使用 ReLU



BERT

Bidirectional encoder representation from transformers
使用transformer网络的编码器，多头注意力机制来计算句子的表示
掩盖一些词，并预测他们的概率，训练模型来预测mask掉的词，掩盖太少，训练花费大；掩盖太多丢失上下文，常取15%



- GPT 是经典的单项的语言模型
- ELMo虽然又从左到右和从右到左，但是独立训练的，只是将他们的表现联系在一起，因此即便建立了上下文词表示，并没有使用双方的上下文。
- BERT 使用 mask 的方式进行整个上下文的预测，使用了双向的上下文信息

BERT可以用来学习句子间的联系，来预测下一句。为特定任务微调