EMLO

ELMO采用了典型的两阶段过程，第一个阶段是利用语言模型进行预训练；第二个阶段是在做下游任务时，从预训练网络中提取对应单词的网络各层的Word Embedding作为新特征补充到下游任务中。上图展示的是其预训练过程，它的网络结构采用了双层双向LSTM，目前语言模型训练的任务目标是根据单词 W_i 的上下文去正确预测单词 W_i ， W_i 之前的单词序列Context-before称为上文，之后的单词序列Context-after称为下文。图中左端的前向双层LSTM代表正方向编码器，输入的是从左到右顺序的除了预测单词外 W_i 的上文Context-before；右端的逆向双层LSTM代表反方向编码器，输入的是从右到左的逆序的句子下文Context-after；每个编码器的深度都是两层LSTM叠加。这个网络结构其实在NLP中是很常用的。使用这个网络结构利用大量语料做语言模型任务就能预先训练好这个网络，如果训练好这个网络后，输入一个新句子 Snew ，句子中每个单词都能得到对应的三个Embedding:最底层是单词的Word Embedding，往上走是第一层双向LSTM中对应单词位置的Embedding，这层编码单词的句法信息更多一些；再往上走是第二层LSTM中对应单词位置的Embedding，这层编码单词的语义信息更多一些。也就是说，ELMO的预训练过程不仅仅学会单词的Word Embedding，还学会了一个双层双向的LSTM网络结构，而这两者后面都有用。

GPT

从词向量到句子向量

- 无监督句子表示：将句子表示成定长向量
- 基线模型：word2vec
- 现有模型：AE，LM，Skip-Thoughts
  - 本身的信息
  - 上下文的信息
  - 任务的信息

BERT

BERT原理与GPT有相似之处，不过它利用了双向的信息，因而其全称是Bidirectional Encoder Representations from Transformers。

BERT做无监督的pre-training时有两个目标：

一个是将输入的文本中 k%的单词遮住，然后预测被遮住的是什么单词。 另一个是预测一个句子是否会紧挨着另一个句子出现。 预训练时在大量文本上对这两个目标进行优化，然后再对特定任务进行fine-tuning。