nlp中的预训练语言模型，主要包括3大方面、涉及到的模型有：

单向特征表示的自回归预训练语言模型，统称为单向模型：

ELMO/ULMFiT/SiATL/GPT1.0/GPT2.0；

双向特征表示的自编码预训练语言模型，统称为BERT系列模型：

(BERT/MASS/UNILM/ERNIE1.0/ERNIE(THU)/MTDNN/ERNIE2.0/SpanBERT/RoBERTa)

双向特征表示的自回归预训练语言模型：XLNet；

预训练思想的本质是模型参数不再是随机初始化，而是通过一些任务（如语言模型）进行预训练；预训练属于迁移学习的范畴，本文的【预训练语言模型】主要指无监督预训练任务（有时也称自学习或自监督），迁移的范式主要为特征集成和模型精调（finetune）。

语言模型表示序列文本的联合概率分布，为降低对长文本的概率估算难度，通常使用一个简化的n-gram模型[5]。为缓解n元语言模型概率估计时遇到的数据稀疏问题，提出了神经网络语言模型NNLM，第一层参数可用作词向量表示。词向量可看作是NNLM的一个副产品，而word2vec通过一些优化技巧专注于词向量的产生，后来的glove词向量是通过共现语料矩阵进行高效分解产生的，glove也可看作是更换了目标函数和权重函数的全局word2vec。由于word2vec、glove等静态词向量未考虑一词多义、无法理解复杂语境，可通过预训练语言模型产生上下文相关的特征表示（动态词向量）。

BERT存在哪些优缺点？

优点：能够获取上下文相关的双向特征表示；

缺点：

生成任务表现不佳：预训练过程和生成过程的不一致，导致在生成任务上效果不佳；

采取独立性假设：没有考虑预测[MASK]之间的相关性，是对语言模型联合概率的有偏估计（不是密度估计）；

输入噪声[MASK]，造成预训练-精调两阶段之间的差异；

无法文档级别的NLP任务，只适合于句子和段落级别的任务；

BERT擅长处理哪些下游NLP任务？

适合句子和段落级别的任务，不适用于文档级别的任务（如长文本分类）；

适合处理文本语言本身就能处理好的任务（如QA/机器阅读理解），不依赖于额外特征（如推荐搜索场景）；

适合处理高层语义信息提取的任务，对浅层语义信息提取的任务的提升效果不大（如文本分类/NER，文本分类关注于“关键词”这种浅层语义的提取）；

适合处理句子/段落的匹配任务，因为BERT在预训练任务中引入NSP；因此，在一些任务中可以构造辅助句（类似匹配任务）实现效果提升（如关系抽象/情感挖掘等任务）；

不适合处理NLG任务，因为BERT在生成任务上效果不佳；

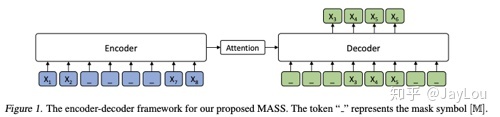
BERT基于“字输入”还是“词输入”好？（对于中文任务）

如果基于“词输入”，会出现OOV问题，会增大标签空间，需要利用更多语料去学习标签分布来拟合模型。

随着Transfomer特征抽取能力，分词不再成为必要，词级别的特征学习可以纳入为内部特征进行表示学习。

BERT系列模型进展介绍

1. MASS(微软)

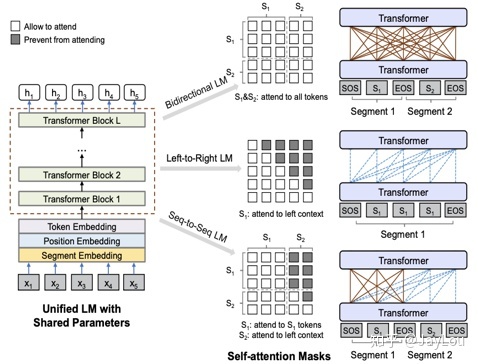


统一预训练框架:通过类似的Seq2Seq框架，在预训练阶段统一了BERT和LM模型；

Encoder中理解unmasked tokens；Decoder中需要预测连续的[mask]tokens，获取更多的语言信息；Decoder从Encoder中抽取更多信息；

当k=1或者n时，MASS的概率形式分别和BERT中的MLM以及GPT中标准的LM一致（k为mask的连续片段长度）

1. UNILM (微软)



统一预训练框架:和直接从mask矩阵的角度统一BERT和LM；

3个Attention Mask矩阵：LM、MLM、Seq2Seq LM；

注意：UNILM中的LM并不是传统的LM模型，仍然是通过引入[MASK]实现的；