BERT的模型变体学习笔记

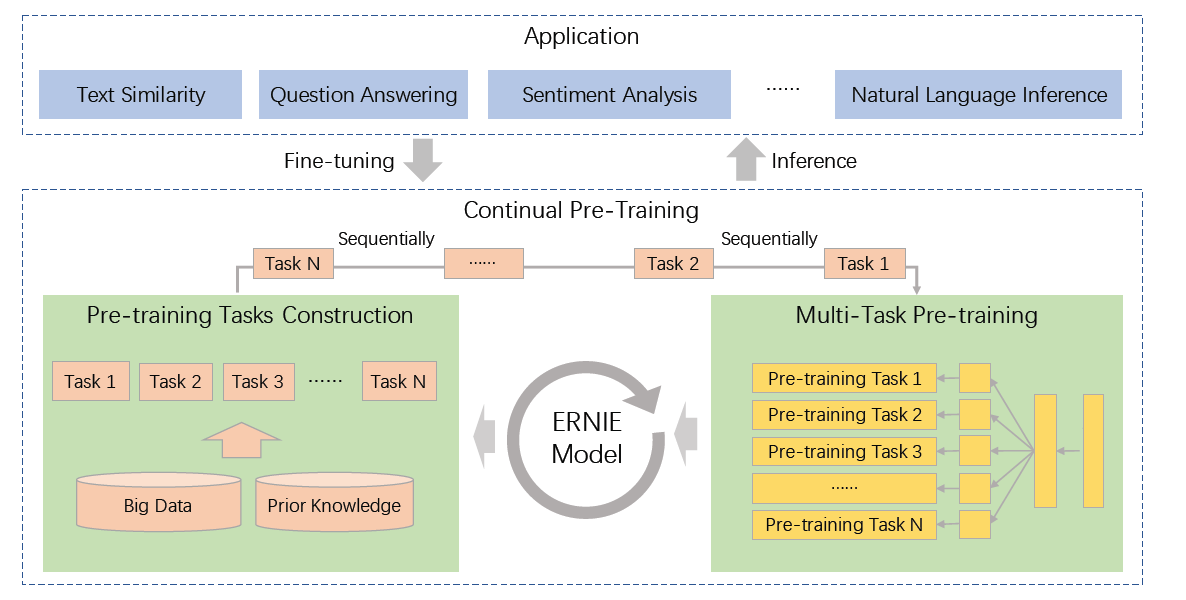
这次课程中，BERT的模型变体主要介绍了ERNIE，Transformer-XL以及XLNet这三种模型变体框架。

ERNIE是Enhanced Representation through Knowledge Integration，其主要的贡献在于通过实体和短语mask学习到了语法和句法信息，并且将中文维基百科、百度百科、百度新闻以及百度贴吧加入了训练数据集，是的中文自然语言处理任务方面的性能达到了世界领先水平。其中采用的方法仍然基于BERT，其中层数为12，隐藏单元为768，采用12个多头注意力机制。ERNIE的特点是同时实现了单字级别、实体级别以及短语级别的MASK。

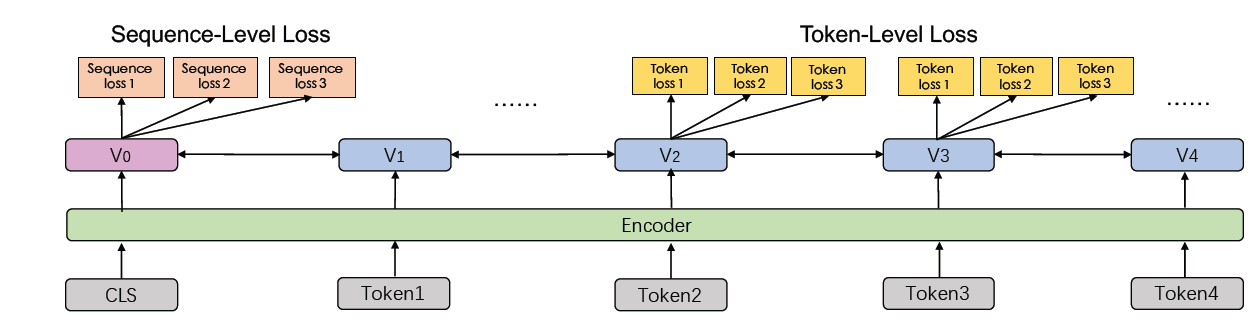
在ERNIE1.0阶段主要是通过提供更多的语料以及实体级别的连续MASK整合进了大量的先验知识，从而提升了预测语言的表示效果，其在自然语言推理、语义相似性、命名实体识别NER、情感分析、问答抽取等任务方面进行了相关的测试，在准确度上要强于BERT超过1%.

在ERNIE2.0方面采用了多任务持续学习的预训练框架，构建了三种无监督任务训练模型，即词法级别的预训练任务、语言结构级别预训练任务以及基于语法级别预训练任务进行预训练。其中，词法级别的预训练任务学习的是token存在段落A中是否也会在文档的段落B中出现；语言结构级别预训练主要是关注文档中的句子被打乱时识别正确的句子顺序；词法级别的预训练任务时计算两句之间的语义与修辞关系从而获取短文本信息检索关系。

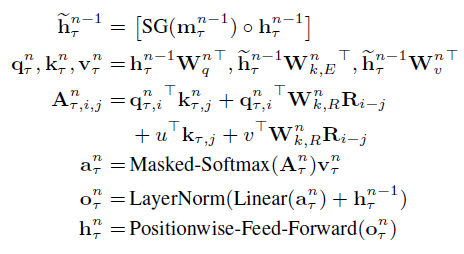
ERNIE2.0的模型框架如下，其集成的数据除了ERNIE1.0的百度百科等数据之外，还集成了Reddit搜素数据等。



ERNIE2.0的这种连续多任务学习的效果是不遗忘之前的训练结果，进行多任务高效的训练，按照序列使用上一任务的参数并且将新旧任务放在一起进行训练，每个任务分成多次迭代，框架完成不同迭代的训练并且自动分配。其中句子级别的损失函数和词级别的损失函数在不同的任务里互相独立，句子级别的任务和词级别的任务可以一起进行训练，如下图所示，



通过回顾Transformer的位置编码、多头注意力以及编码解码过程，比较Transformer和RNN的区别，可以看出Transformer的优点在于并行和获取长距离语义的信息，缺点在于位置嵌入所导致的限制，将Transformer的优点结合RNN的优点，则构成了Transformer-XL的BERT变体架构。Transformer-XL来源于谷歌的论文“Attentive Language Models Beyond a Fixed-Length Context”，主要是解决长文本的建模并且捕获超长距离的依赖，其贡献主要在于transformer建模机制和一种相对位置编码的方式而不是Transformer的绝对位置编码方式。Transformer-XL的传递方式是下一层接受上一层的信息，并且下一层接受上一时刻的上层信息，不同片段的信息内容是通过拼接的方式实现的，其使用相对位置编码的原因主要集中在描述不同token之间的相对关系，避免绝对位置编码造成的性能损失，其最终形态如下图所示：

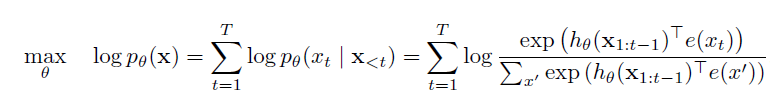


ERNIE、 Transformer等新的变体相比于BERT的优点在于以下4点：

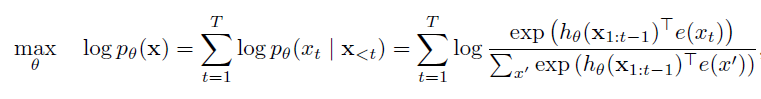
1. 在几种不同的数据集（大/小，字符、单词级别）均实现了最先进的语言建模结果
2. 结合深度学习的两个重要概念——循环机制和注意力机制，解决了长距离依赖的问题
3. 学习了超长的依赖关系
4. 推理阶段较快。

同时，其不足在于这些模型并么有在具体的情感分析或者问答方面有所应用，没有明确给出相对于BERT的明显优势，在训练模型过程中需要消耗大量的计算资源。

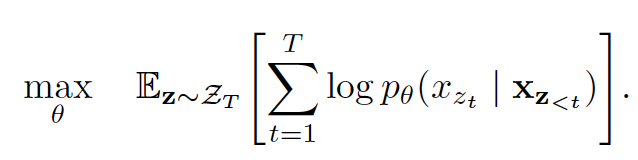
XLNet也是一个类BERT模型，出自谷歌的论文Generalized Autoregressive Pretraining for Language Understanding。XLNet高度融合了以GPT为代表的自回归模型和以BERT为代表的自编码模型二者的优点，属于双向语言模型，并且可以根据上下文单词预测被MASK掉的噪音，完成生成类NLP任务。AR的目标是在前向自回归因子分解方法如下图所示：



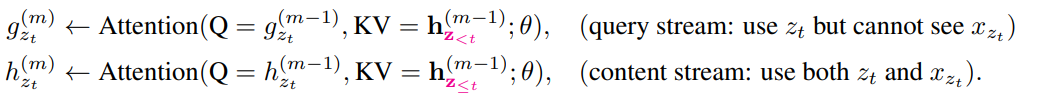
AE的目标是通过随机选取15%的tokens进行遮挡，训练目标是从被遮挡处理之后的版本 生成 ，其生成方法如下：



为了更好的利用好AR模型的优点并且双向捕捉到上下文信息，XLNet采用了Permutation Language Modeling这种基于排列组合的输入方法，不会改变原始词的顺序，实现通过Attention的MASK来对应的不同分解方法，其建模方法的目标是最大化以下的目标函数：



为了学习更多有用的表示方式，将目标的位置变化重新参数化，在目标位置 上从前文获取注意力，采用了两套表征方式content stream或者query stream进行双流自注意力的表征，



训练方法是采用50%的概率是连续的句子，50%的概率是不连续的句子，随机采样排列组合数据，与bert比较不同的地方在于xl-net采用了相对片段编码（relative segment encodings），双流自注意力的架构如下图所示：

