**BERT**：通过大规模无监督预训练学习文本的深层双向表示，再通过微调适应各类下游任务。多层双向的transformer的编码器（通过masked模型实现），能实现句子、词元层面的任务，更适合做文本分类、命名实体识别、问答系统，不太适合做文本翻译、文本摘要

**预训练**（pre-training）：使用大量没有标号的数据训练

预训练任务：

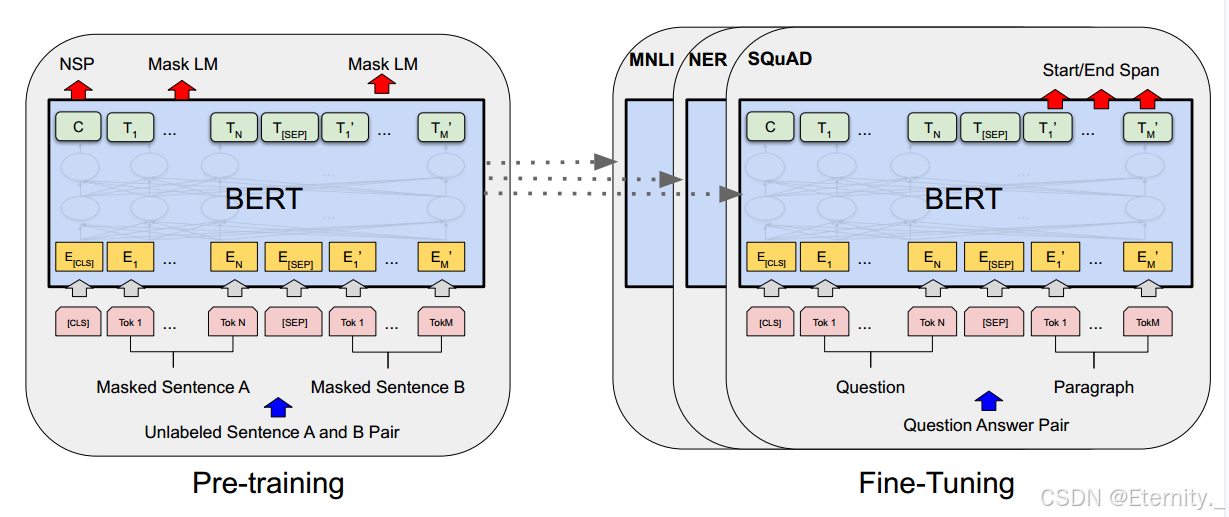
(1) 掩码语言模型（Masked Language Model, MLM）

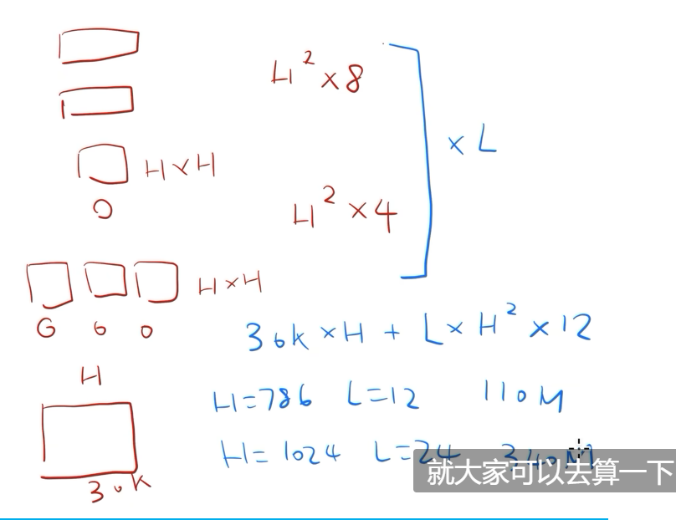
* 方法：随机掩盖输入中15%的词（如用[MASK]替换），模型预测被掩盖的词。
  + 掩盖策略：80%用[MASK]，10%随机替换，10%保留原词。
* 目的：强制模型通过双向上下文推断被掩盖词，增强语义理解能力。

(2) 下一句预测（Next Sentence Prediction, NSP）

* 方法：输入两个句子A和B，预测B是否为A的下一句。
  + 正样本：实际相邻的句子对（50%）。
  + 负样本：随机选取的无关句子对（50%）。
* 目的：学习句子间关系，提升段落级任务（如问答、推理）表现。
* 示例：
  + 输入："[CLS] A: How are you? [SEP] B: I am fine. [SEP]" → 标签：IsNext。
  + 输入："[CLS] A: How are you? [SEP] B: The sky is blue. [SEP]" → 标签：NotNext。

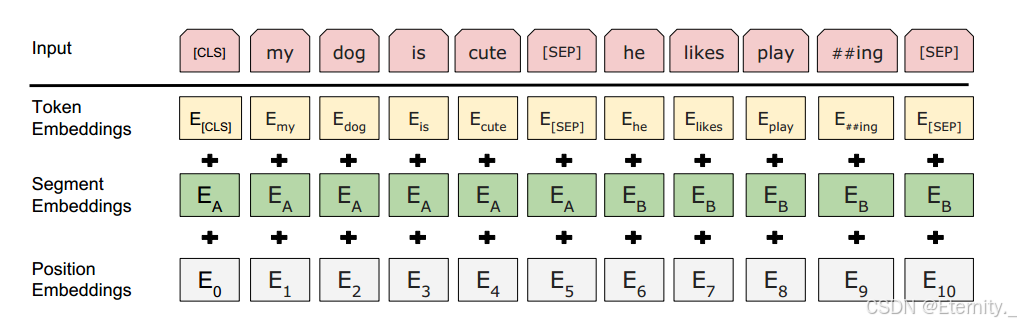
**微调**（fine-tuning）：权重被初始化为预训练时得到的权重，使用有标号的数据训练权重。微调时将输入变换为之前训练的模型对应的输入格式，得到输出，然后加一个输出层



**可学习参数**：L是transformer层个数，H是隐藏层维度，A是注意力头数

分词：先用空格分词，然后对分词后的结果划分出常用的子序列，从而用较小的词典表示较大的文本

输入文本添加**[CLS]**标记，取其对应输出作为整体表示

输入两个句子时将其合并为一个序列，如何区分这两个句子：1.第二个句子前加上【SEP】；2.学一个嵌入层来表示到底是第一个句子还是第二个  
Token Embeddings：词嵌入（WordPiece分词）。

Segment Embeddings：区分句子对（如句A和句B）。

Position Embeddings：位置编码，标记词序。