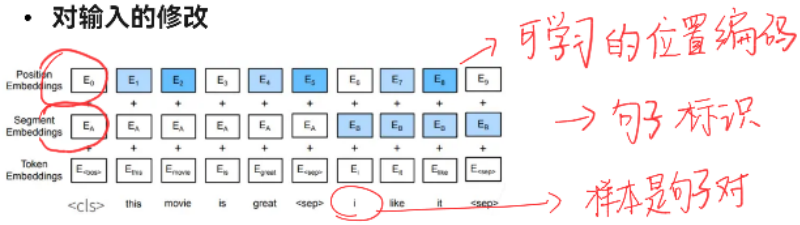
BERT

- **NLP里的迁移学习**
  - 使用预训练好的模型来抽取词句特征（例如word2vec和语言模型）
  - 不更新预训练好的模型
  - 需要构建新的网络来抓取新任务需要的信息（如word2vec忽略时序信息）
- **BERT的动机**
  - 基于微调的NLP模型，权重复用
  - 预训练的模型抽取了足够多的信息
  - 新任务只需要增加一个简单的输出层
- **BERT的架构**
  - 只有编码器的Transformer
  - 两个版本：base(12 Transformer block, 768 hidden size, 12 head, parameters 110M); large(24 Transformer block, 1024 hidden size, 16 head, parameters 110M)
  - 在大规模数据训练 > 3B词
- **对输入的修改**



- **预训练任务1：掩码语言模型**
  - Transformer 编码器是双向的，标准语言模型要求是单向的
  - 带掩码的语言模型每次随机（15%概率）将一些词元换成<mask>
  - 因为微调任务中不出现<mask>
    - 80%概率下将选中的词变<mask>
    - 10%概率下换成一个随机词元
    - 10%概率下保持原有词
- **预训练任务2：下一句子预测**
  - 预测一个句子对中两个句子是不是相邻
  - 训练样本中：50%概率选择相邻句子对，50%概率选择随机句子对
  - 将<cls>对应输出放入一个全连接层来预测
- **总结**
  - BERT针对微调来设计
  - 基于Transformer的编码器做了如下修改
    - 模型更大，训练数据更多
    - 输入句子对，片段嵌入，可学习的位置编码
    - 训练时的两个新任务