1. ELMo 的优点

ELMo 实现了两个转变：

1）实现从单纯的词嵌入（Word Embedding）到情景词嵌入（Contextualized Word Embedding）的转变；

2）实现预训练模型从静态到动态的转变。

2. ELMo 的缺点

ELMo 预训练模型的特征提取器使用了双向循环神经网络（如 Bi-LSTM），循环神经网络的训练需要按序列从左到右或从右到左，严格限制了并发处理能力。此外，ELMo 的每一层会拼接两个方向的向量，所以这种操作实际仍然属于单向学习，无法做到同时向两个方向学习。

7.2　可视化 BERT 原理

前面提到，ELMo 是预训练模型由静态转为动态的重要转折点，不过它基于循环网络 LSTM，这就严重限制了其并发能力，在面对巨大的训练语料库，这是非常致命的。不过，接下来将介绍的 BERT 和 GPT 预训练模型就很好地解决了这个问题，它们不再基于 LSTM，而是基于可平行处理的 Transformer。

7.2.1　BERT 的整体架构

BERT 的整体架构如图 7-1 所示，它采用了 Transformer 中的 Encoder 部分。

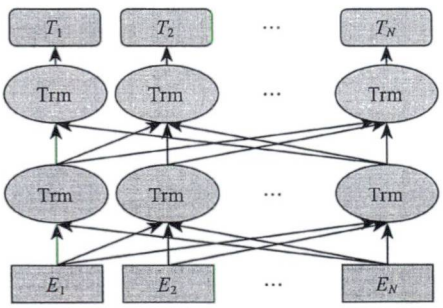

图 7-1　BERT 模型的整体架构

图 7-1 中的 Trm 指 Transformer 的 Encoder 模块，如图 7-2 所示。

BERT 提供了简单和复杂两个模型，对应的超参数分别如下。

- $\text{BERT}_{\text{BASE}}$：$L$=12，$H$=768，$A$=12，参数总量 110MB；
- $\text{BERT}_{\text{LARGE}}$：$L$=24，$H$=1024，$A$=16，参数总量 340MB。