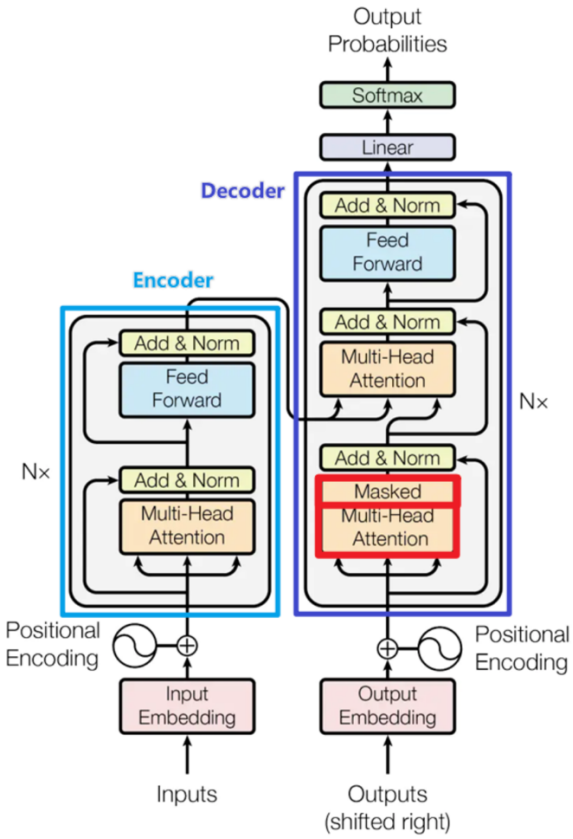
Mask



1. 概念

- Mask 机制经常被用于 NLP 任务中，按照作用总体来说可以分成两类
  a) 用于处理非定长序列的 padding mask（非官方命名）
  b) 用于防止标签泄露的 sequence mask(非官方命名）
- Transformer 中同时用到了这两种 Mask 机制

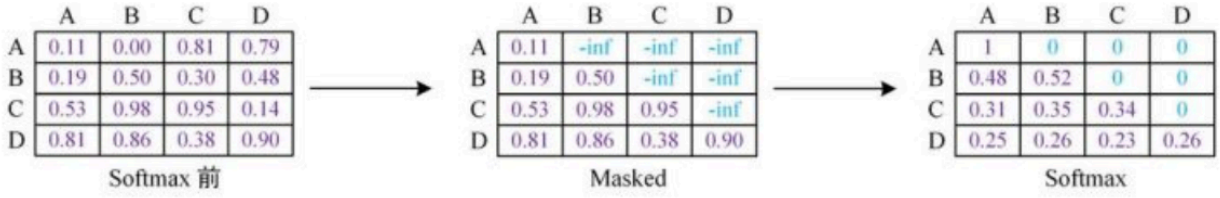
2. padding mask

- 在 NLP 任务中，文本通常是不定长的，所以在输入一个样本长短不一的 batch 到网络前，要对 batch 中的样本进行截断（truncating）/补齐（padding）操作，以便能形成一个张量的形式输入网络，如下图所示。对于一个 batch 中过长的样本，进行截断操作，而对于一个长度不足的样本，往往采用特殊字符"<PAD>"进行 padding（也可以是其他特殊字符，但是 pad 的字符要统一）。Mask 矩阵中可以用 1 表示有效字，0 代表无效字（也可以用 True/False）。

- padding mask 的生成和使用:
  a) padding（补齐）操作在 batch 输入网络前完成，同步生成 padding mask 矩阵。
  b) padding mask 矩阵常常用在最终结果输出、损失函数计算等等，一切受样本实际长度影响的计算中，或者说不需要无用的 padding 参与计算时候。

3. sequence mask

- sequence mask 有各种各样的形式和设计，最常见的应用场景是在需要一个词预测下一个词的时候，如果用 self-attention 或者是其他同时使用上下文信息的机制，会导致模型"提前看到"待预测的内容，这显然不行，所以为了不泄露要预测的标签信息，就需要 mask 来"遮盖"如下图所示，这也是 Transformer 中 Decoder 的 Masked Multi-Head self-attention 使用的 Mask 机制。



- 除了在 decoder 部分加入 mask 防止标签泄露以外，还有模型利用这种填空机制帮助模型学的更好，比如说 BERT 和 ERNIE 模型中利用到的 Masked LM (MLM)。(注意: BERT 模型只有 Transformer 的 Encoder 层，是可以学习上下文信息的)

4. BERT 中的 Mask

- Masked LM 随机掩盖部分输入词，然后对那些被掩盖的词进行预测。在训练的过程中，BERT 随机地掩盖每个序列中 15%的 token，并不是像 word2vec 中的 cbow 那样去对每一个词都进行预测。MLM 从输入中随机地掩盖一些词，其目标是基于其上下文来预测被掩盖单词的原始词汇。
- 而 ERNIE 不是在 token 级进行掩码，而是在短语级进行掩码。