07

实战演练之预训练模型

（1）预训练介绍

（2）代码实战，掩码语言模型

（3）代码实战，因果语言模型

预训练介绍

预训练简介

- **什么是预训练**
  - 预训练（Pretrain）是指通过自监督学习从大规模数据中获得与具体任务无关的预训练模型的过程，最终产出为预训练模型（Pretrained Model）。

模型类型	常用预训练模型	适用任务
编码器模型，自编码模型	ALBERT, BERT, DistilBERT, RoBERTa	文本分类、命名实体识别、阅读理解
解码器模型，自回归模型	GPT, GPT-2, Bloom, LLaMA	文本生成
编码器解码器模型，序列到序列模型	BART, T5, Marian, mBART	文本摘要、机器翻译

预训练介绍

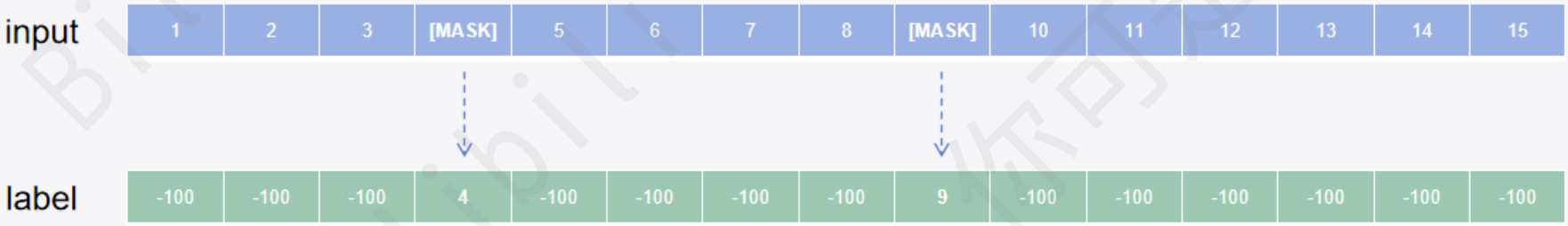
预训练简介

- **预训练任务**
  - 掩码语言模型，自编码模型
    - 将一些位置的token替换成特殊的[MASK]字符，预测这些被替换的字符
  - 因果语言模型，自回归模型
    - 将完整序列输入，基于上文的token预测当前token
  - 序列到序列模型
    - 任务较为多样化，只是采用编码器解码器的方式，预测部分放在解码器中

预训练介绍

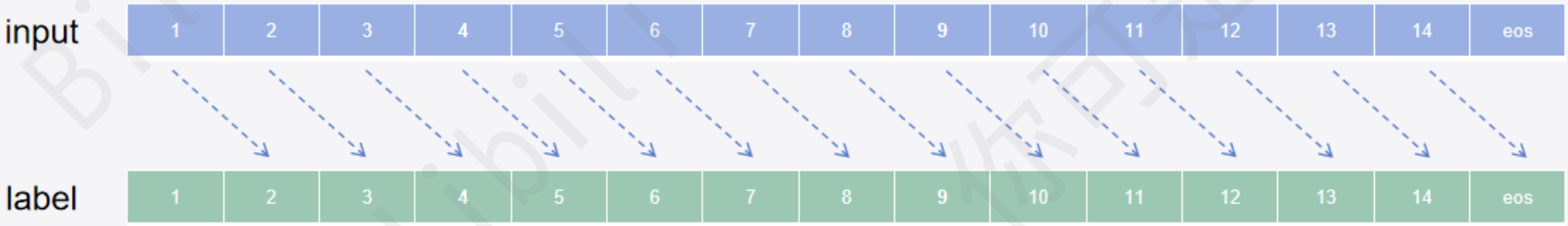
预训练简介

- **预训练任务**
  - 掩码语言模型，自编码模型
    - 将一些位置的token替换成特殊的[MASK]字符，预测这些被替换的字符
    - 只计算掩码部分的loss，其余部分不计算loss



预训练简介

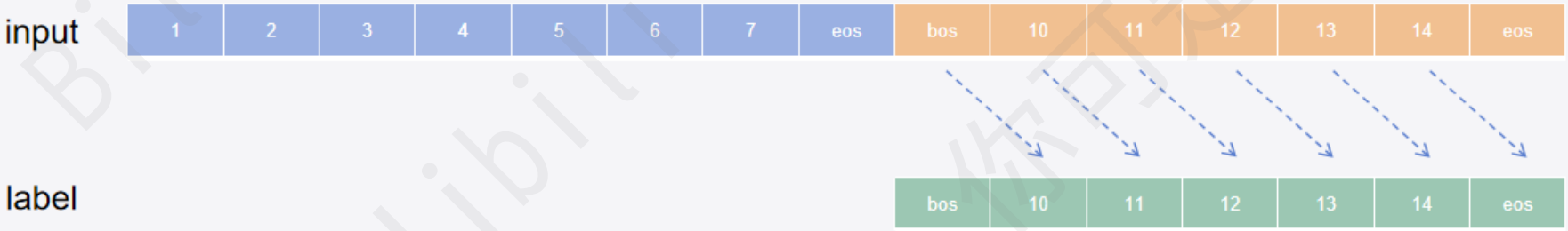
- **预训练任务**
  - 因果语言模型，自回归模型
    - 将完整序列输入，基于上文的token预测当前token
    - 结束位置要有特殊token，eos_token



预训练介绍

预训练简介

- **预训练任务**
  - 序列到序列模型，前缀语言模型（Prefix Language Model）
    - 任务较为多样化，如掩码生成、片段自回归、乱序修正等
    - 采用编码器解码器的方式进行实现，计算解码器部分的loss



代码实战演练（掩码语言模型）

- **数据集**
  - https://huggingface.co/datasets/pleisto/wikipedia-cn-20230720-filtered
  - 百科语料
- **预训练模型**
  - hfl/chinese-macbert-base

代码实战演练（因果语言模型）

- **数据集**
  - https://huggingface.co/datasets/pleisto/wikipedia-cn-20230720-filtered
  - 百科语料
- **预训练模型**
  - Langboat/bloom-389m-zh