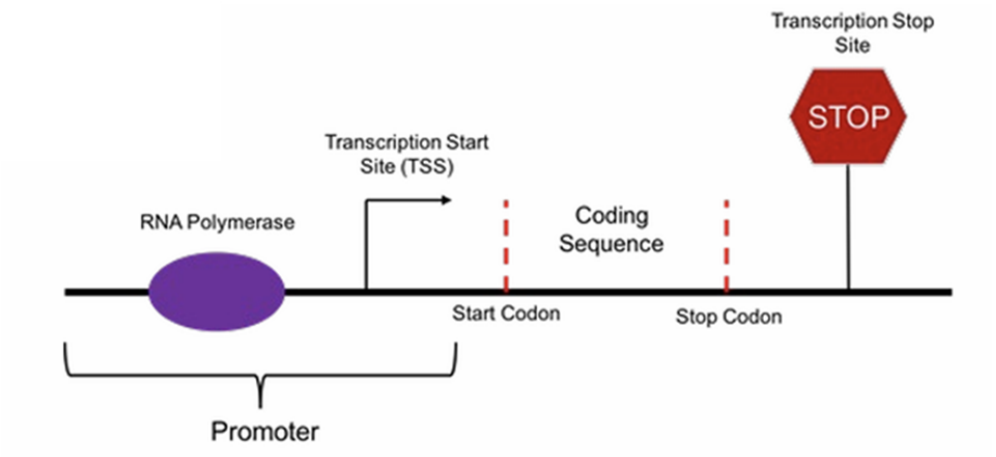
Promoter detection

Author	資工四甲 U0924028 符絮棻
Tags	BERT self-attention
程式碼來源	https://github.com/MChatzakis/promoter-detection/blob/main/promoter-detection.ipynb

What is a promoter?

- 啟動子是基因上游的 DNA 區域，相關蛋白質（例如 RNA 聚合酶和轉錄因子）在此結合以啟動該基因的轉錄。啟動子檢測是為了鑑定給定的DNA序列中是否存在啟動子區域。
- 指示轉錄起始位置的 DNA 序列
- 指示「細胞機器」（聚合酶）將在何處啟動轉錄



位置編碼

由於句子中的每個單詞同時流經 Transformer 的編碼器/解碼器堆棧，因此模型本身對每個單詞沒有任何位置/順序感。因此，仍然需要一種方法來將單詞的順序合併到我們的模型中。

給模型一些秩序感的一個可能的解決方案是給每個單詞添加一條關於它在句子中的位置的資訊。我們稱之為「資訊片段」即位置編碼。

理想情況下，應滿足以下條件

- 它應該為每個時間步長（單詞在句子中的位置）輸出唯一的編碼
- 在具有不同長度的句子中，任何兩個時間步長之間的距離應該一致。
- 我們的模型應該不費吹灰之力地推廣到更長的句子。其值應是有界的。
- 它必須是確定性的。

此程式碼使用正弦位置編碼，公式如下：

Let t be the desired position in an input sentence, $\overrightarrow{p_t} \in \mathbb{R}^d$ be its corresponding encoding, and d be the encoding dimension (where $d \equiv_2 0$) Then $f : \mathbb{N} \to \mathbb{R}^d$ will be the function that produces the output vector $\overrightarrow{p_t}$ and it is defined as follows:

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k . t), & \text{if } i = 2k \\ \cos(\omega_k . t), & \text{if } i = 2k+1 \end{cases}$$

where

$$\omega_k = \frac{1}{10000^{2k/d}}$$

使用資料

train.csv

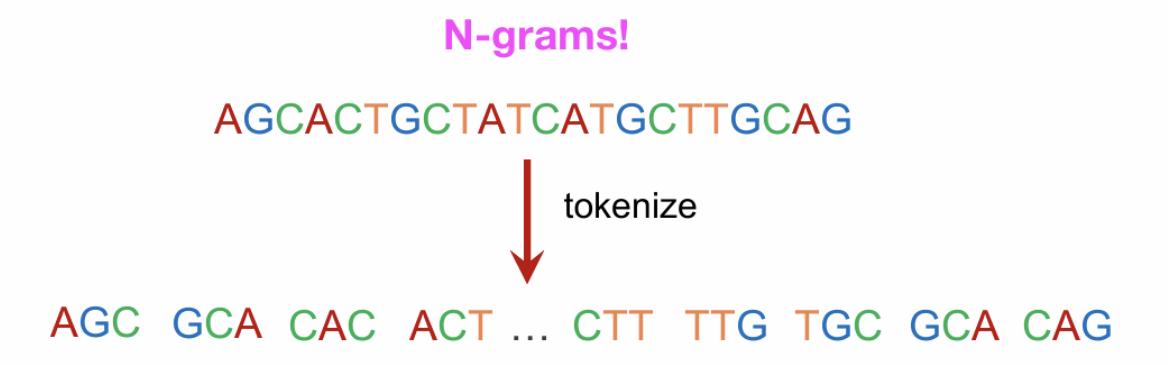
	sequence	label
1	TATAATAATAACGAAGATGAGACGACAGTCGACAAGAAAAGCACCAGCTGTCCCC……CCTGCGCCGA	0
2	AAAGCCCGAGCGGCGGCCACGCCTCGGTGGCGATTTTATTAGCGCTTGGTTGGG……GGCGGCTCCA	1
3	AGTCCGCGATATTCTGAGGGGACTTTCGACACAAAAAAGTTGACACGGGTCGTTT……CCCAGCTGCG	0
……		
47356	CGGGGCCGCAGCGCTAGCCGCAACTATTGAGTGCTCTCCATACTATGCGATTTCG……GCTCGCCGCC	0

test.csv

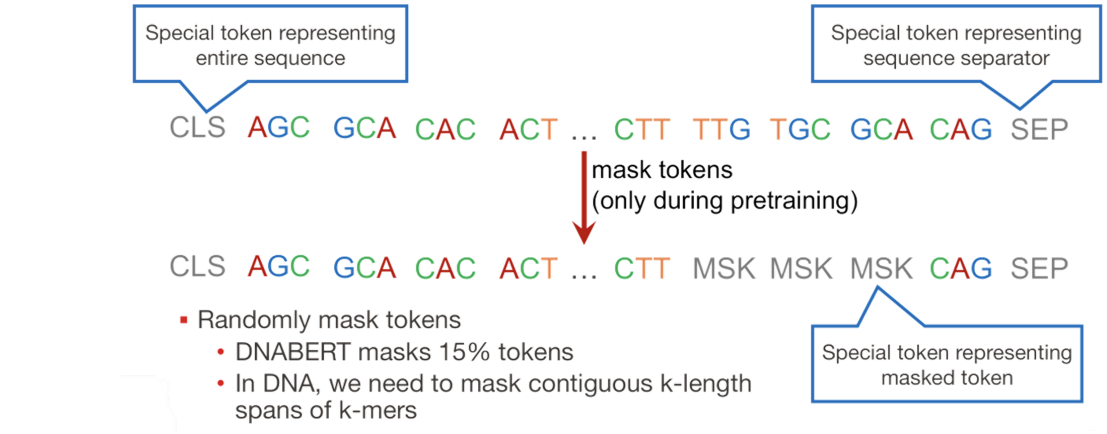
	sequence	label
1	CTAAATATTAACTGGTCTTGTGAGATGTCTTCTTGGCTGGAGCCTGACCACCTAACTT……CCTTAGCTCTT	1
2	CAGCCTCTCGACCGCCGGCCACCCGCCAGCCACGCGCGCCCAGACAGGAGGG……TCCACGGAGG	1
3	GTGGGATCCCCACGGACCTGGAAATTCTCGCCTGTCTTCCCTTCACCCAGAGCAA……GAGCGCGGCG	0
……		
5920	CTACACCAAAATGTAAGCGTCCAGGGATAACCCATGTAAAAGCCCATGGGGGAAG……CACTTTGGGG	0

DNABERT 架構

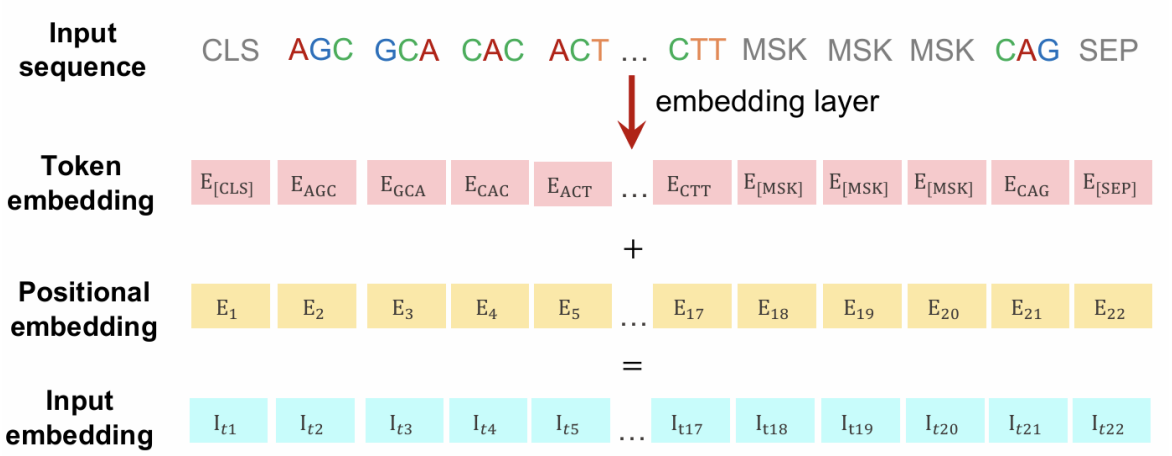
1. 文字和 DNA 序列之間的主要區別在於如何標記序列 ⇒ 需要為 DNA 序列實作一個分詞器。



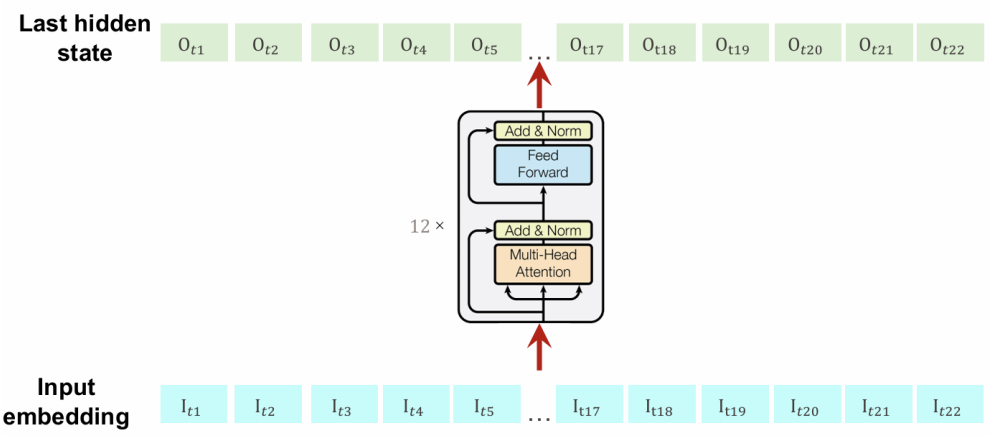
2. 對於每個序列，隨機遮罩構成序列 15% 的 k 個連續標記的區域。



3. 通過結合每個核苷酸的身份和位置，將原始DNA序列轉換成適合機器學習模型的格式。

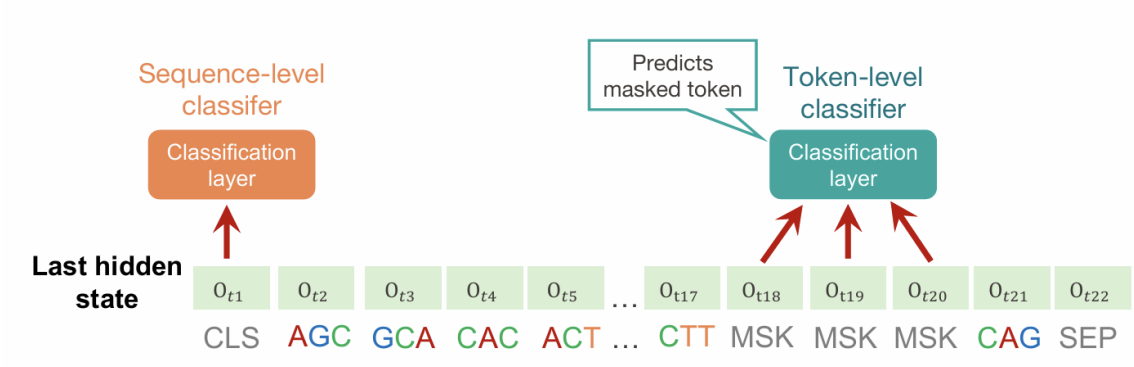


4. Transformer encoder

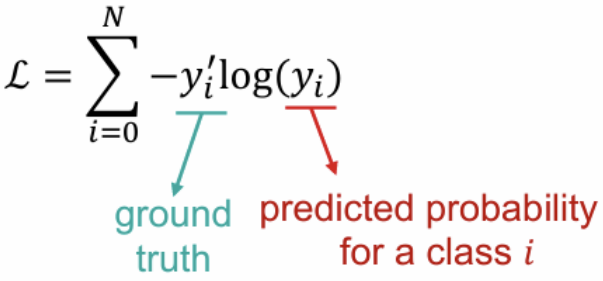


5. Classifier

   DNABERT根據剩餘tokens預測屏蔽序列



Cross-entropy loss function



用於微調 DNABERT 的具體數據

- 陽性樣本：啟動子序列（取自Promoter 資料庫)
- 陰性樣本：啟動子區域外的隨機序列
  - 隨機序列是不夠的
  - 使用包含相似基序的隨機序列 (TATA)
- 選擇「困難」的負序列有助於 DNABERT 學習區分陽性/陰性樣本的不太明顯的特徵