DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome

Jiayou Zhang

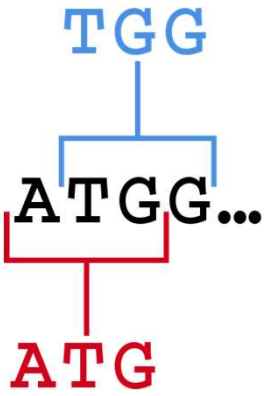
Aug 24

TLDR

- Proposed a new BPE-based tokenization for genome data
- Introduced a DNABERT-2, a refined genome foundation model
- Proposed a multi-species genome dataset for training and a multi-species genome classification dataset for evaluation

Part 1: Tokenization

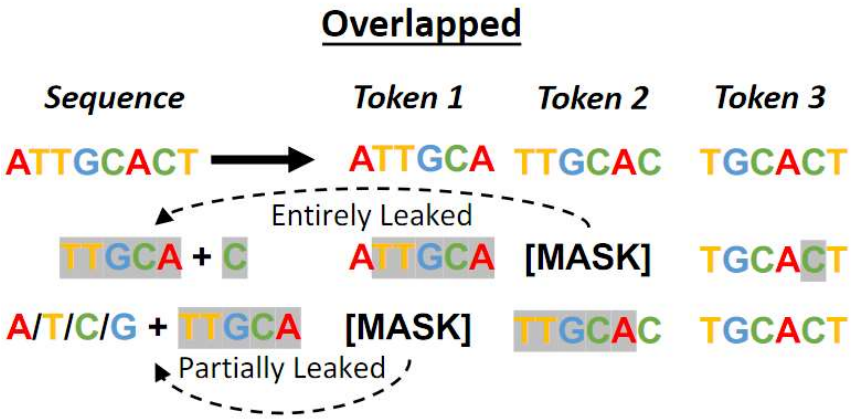
- k-mer tokenization:
  - ***k*-mers** are substrings of length *k* contained within a biological sequence.
  - In NLP, this is also known as n-gram.

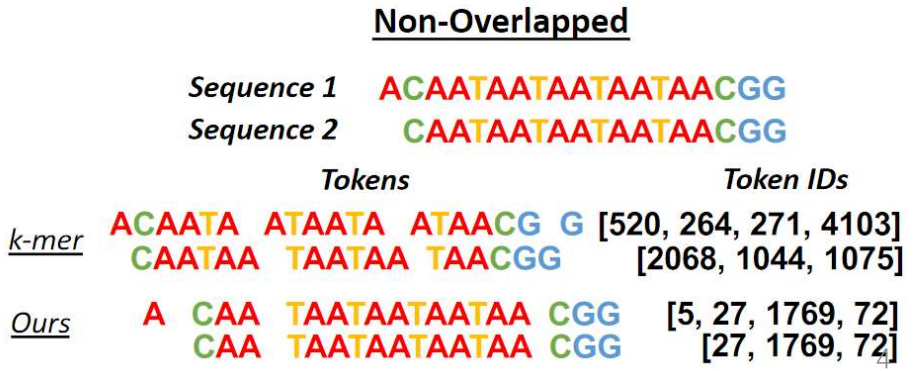


An example of 3-mers.

Drawbacks of k-mers

- Although k-mer tokenization is widely used in existing works, it suffers from
  - Left fig: For overlapped k-mer tokenization, the information of a masked token is leaked by its neighbors
  - Right fig: For non-overlapped k-mer tokenization, the tokenization of a sequence can be drastically changed due to insertion/deletion





BPE-based tokenization

- The construction of Byte Pair Encoding (BPE) [Sennrich et al., 2016]
  1. Start with a vocabulary with each character as a token
  2. Tokenize the corpus with this vocabulary
  3. Merge the most frequent token pairs and add it to the vocabulary
  4. Repeat 2 until the vocabulary reaches the desired size

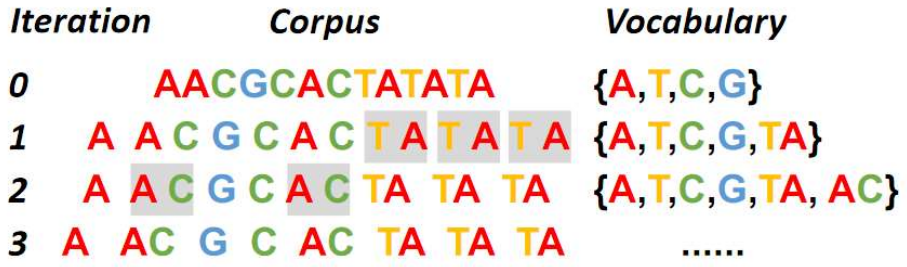
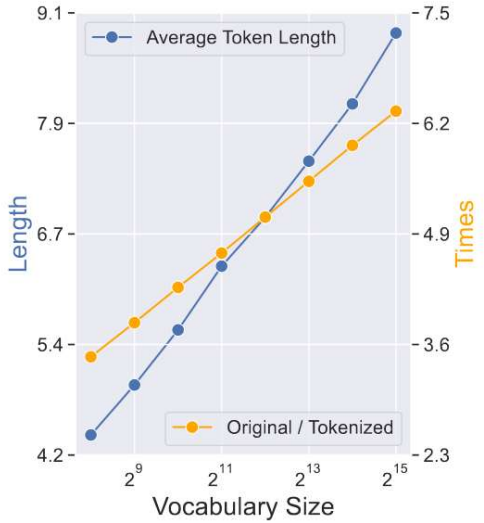


Figure 2: Illustration of the BPE vocabulary constructions.

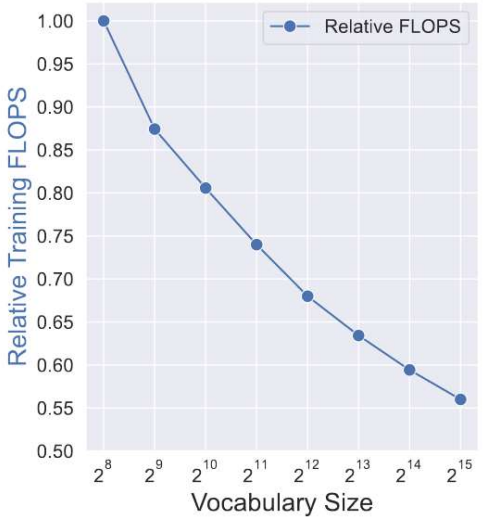
- The encoding process greedily encodes the most lengthy tokens. If two tokens share the same length, the more frequent one will be encoded first.

The choice of desired vocabulary size

- The larger vocabulary size, the more lengthy tokens
  - The shorter tokenized sequence 👍
  - The lower computational cost 👍
  - The less common each token is ☹



(a) Average token length and the length ratio of original sequence v.s. tokenized sequence.



(b) Training FLOPs on 500-length sequences compared to model with 2^8 vocabulary.

(c) Model performance averaged over each tasks (macro) and individual dataset (micro).

Best trade-off

Part 2: DNABERT-2

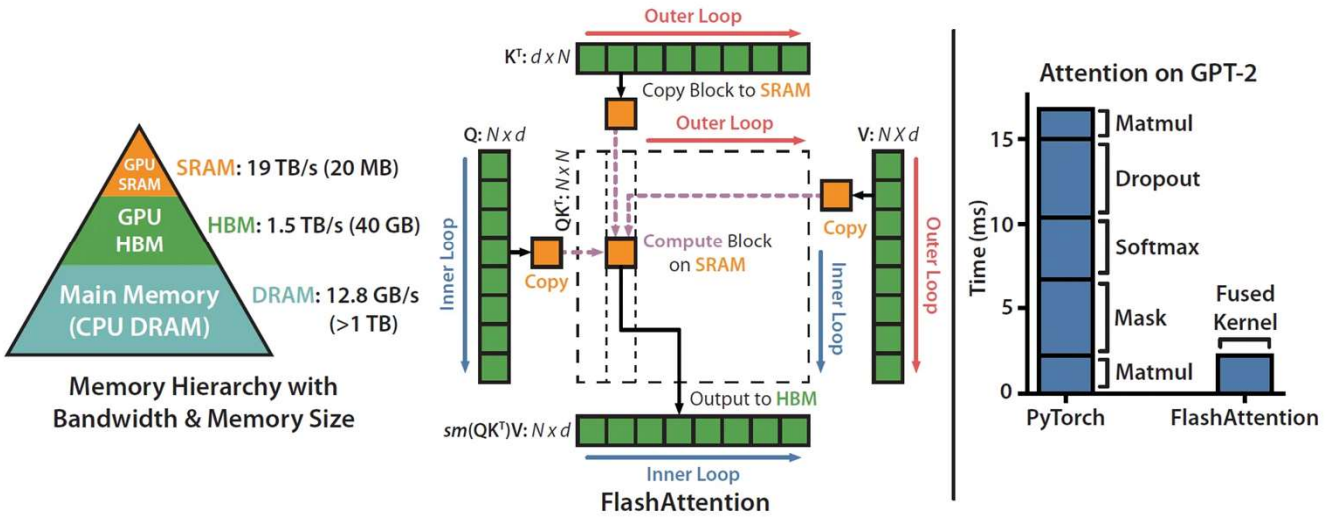
- Incorporate a series of recent advances to BERT architecture
  - Attention with Linear Biases (ALiBi)
    - Instead of using positional embedding, the position is encoded as a distance penalty to the attention score

$$softmax(q_i K - md_i)$$

where d_i is the distance between the i-th query and the keys and m is a head-specific constant. ALiBi uses $m = \frac{1}{2^1}, \frac{1}{2^2}, \frac{1}{2^3}, \ldots$ for different heads.

Part 2: DNABERT-2

- Incorporate a series of recent advances to BERT architecture
  - Flash attention
    - The attention is slow due to the O(N^2) computational and memory cost
    - Flash attention avoids IO on HBM but try to use the faster SRAM
    - To achieve this, it splits the K/Q/V matrices into blocks and incrementally performs softmax over the entire input.
    - To fit in the SRAM, it does not explicitly store the attention scores but computes it on the fly during forward/backward, trading extra computation for fewer IO with HBM.



Part 2: DNABERT-2

- Incorporate a series of recent advances to BERT architecture
  - Low-Rank Adaptation (LoRA)
  - After pre-training, we obtain weight matrices, each denoted as W_0
  - During finetuning, W_0 is kept frozen and a low rank matrix ΔW is added to W_0

$$W_1 = W_0 + \Delta W, \qquad \text{where } \Delta W = BA$$

The learnable part is low-rank matrices B and A

Part 2: DNABERT-2

- Implementation
  - Mask ratio = 15%. Tokens are masked independently.
  - Batch size = 4096
  - Max sequence length = 128
  - 500k training steps with AdamW optimizer
  - 5e-4 learning rate with 30k-step warmup and linear decay
  - 14 days using eight Nvidia RTX 2080Ti GPUs

Part3: Datasets

- Pre-train:
  - The human genome dataset from [Ji et al., 2021] with 2.75B nucleotide bases
  - The multi-species dataset from 135 species, spread across 7 categories, with 32.49B nucleotide bases in total.

Part3: Datasets

- Benchmark: Genome Understanding Evaluation (GUE)
  - Starting with various biologically important genome analysis datasets,
  - Datasets where the majority of models yielded moderate (e.g., F1-scores between 0.3 and 0.8) and distinguishable performance scores were retained
  - Datasets that did not meet these criteria underwent a restructuring process involving various strategies such as class balancing, adversarial sample inclusion, and reduction of training sample volume, among others

Species	Task	Num. Datasets	Num. Classes	Sequence Length
Human	Core Promoter Detection	3	2	70
	Transcription Factor Prediction	5	2	100
	Promoter Detection	3	2	300
	Splice Site Detection	1	3	400
Mouse	Transcription Factor Prediction	5	2	100
Yeast	Epigenetic Marks Prediction	10	2	500
Virus	Covid Variant Classification	1	9	1000

Table 1: Summarization of the Genome Understanding Evaluation (GUE) benchmark.

Experiments

Model	Num. Params. ↓	FLOPs ↓	Trn. Tokens	Num. Top-2 ↑	Ave. Scores ↑
DNABERT (3-mer)	86M	3.27	122B	2 || 0	61.62
DNABERT (4-mer)	86M	3.26	122B	0 || 1	61.14
DNABERT (5-mer)	87M	3.26	122B	0 || 1	60.05
DNABERT (6-mer)	89M	3.25	122B	0 || 1	60.51
NT-500M-human	480M	3.19	50B	0 || 0	55.43
NT-500M-1000g	480M	3.19	50B	0 || 1	58.23
NT-2500M-1000g	2537M	19.44	300B	0 || 1	61.41
NT-2500M-multi	2537M	19.44	300B	7 || 9	66.93
DNABERT-2	117M	1.00	262B	8 || 4	66.80
DNABERT-2◆	117M	1.00	263B	**11** || **10**	**67.77**

Table 2: The statistics and performance of each model. The five columns represent the number of model parameters, relative FLOPs compared to DNABERT-2, the number of tokens used in pre-training, and the number of being top-2 among all the models (`1st || 2nd`) and the average evaluation scores on the 28 datasets of the GUE benchmark. ◆: perform further masked language modeling pre-training on the training sets of the GUE benchmark.

Experiments

	Yeast	Mouse	Virus	Human			
	EMP	**TF-M**	**CVC**	**TF-H**	**PD**	**CPD**	**SSP**
DNABERT (3-mer)	49.54	57.73	62.23	64.43	84.63	**72.96**	84.14
DNABERT (4-mer)	48.59	59.58	59.87	64.41	82.99	71.10	84.05
DNABERT (5-mer)	48.62	54.85	63.64	50.46	84.04	72.03	84.02
DNABERT (6-mer)	49.10	56.43	55.50	64.17	81.70	71.81	84.07
NT-500M-human	45.35	45.24	57.13	50.82	85.51	66.54	79.71
NT-500M-1000g	47.68	49.31	52.06	58.92	86.58	69.13	80.97
NT-2500M-1000g	50.86	56.82	66.73	61.99	86.61	68.17	85.78
NT-2500M-multi	58.06	67.01	**73.04**	63.32	**88.14**	71.62	**89.36**
DNABERT-2	55.98	67.99	71.02	**70.10**	84.21	70.52	84.99
DNABERT-2◆	**58.83**	**71.21**	68.49	66.84	83.81	71.07	85.93

Table 3: The models' averaged performance on the 8 tasks in the GUE benchmark, including Epigenetic Marks Prediction (EMP), Transcription Factor Prediction on the Human genome and the Mouse genome (TF-H and TF-M), Covid Variants Classification (CVC), Promoter Detection (PD), Core Promoter Detection (CPD), and Splice Site Prediction (SSP).