Summary of atacFormer

GE Muyang

Feb 24, 2025

1 Current Framework

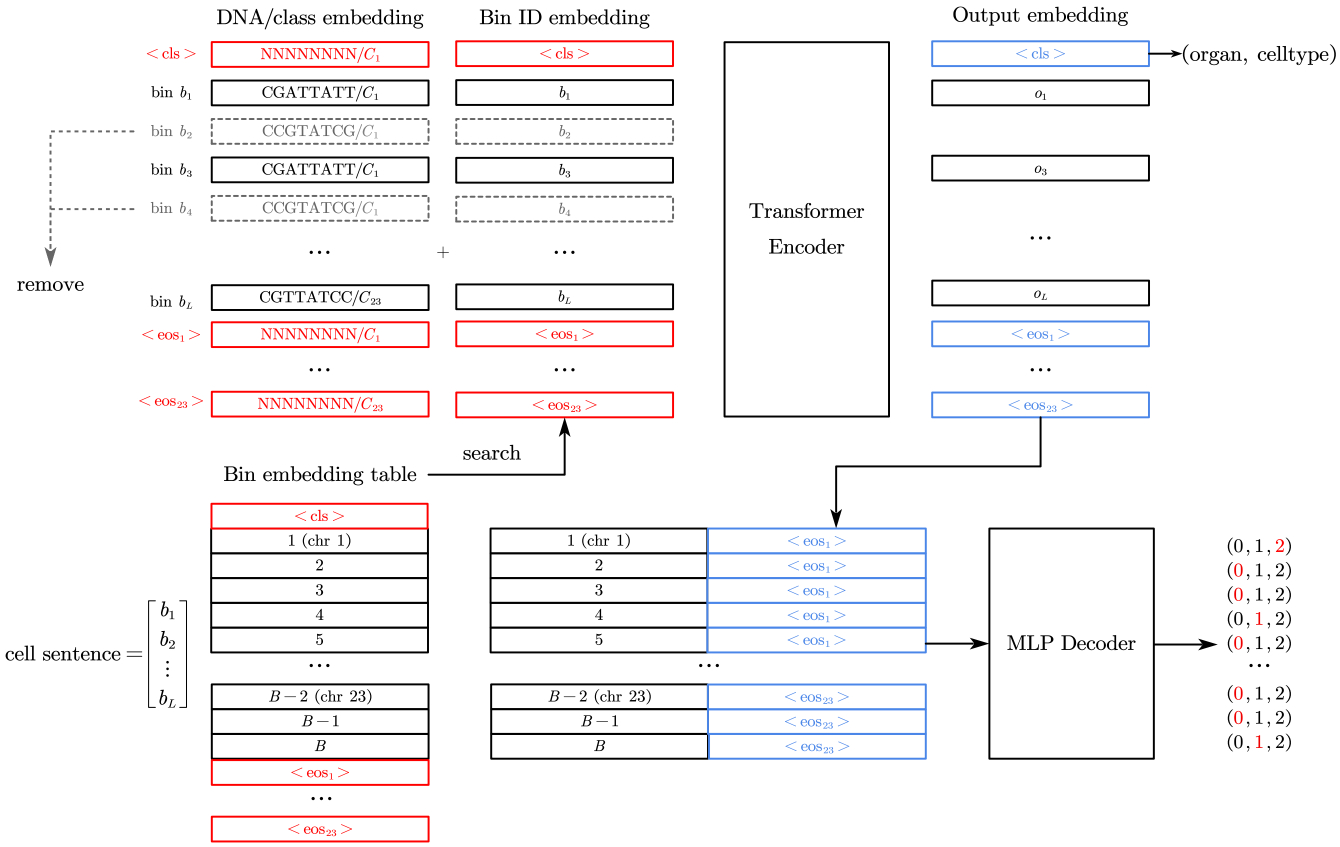

Fig. 1 | Current framework of atacFormer.

Suppose that we have B different ATAC bins in total. Each bin corresponds to a DNA sequence with length 5000, as in SnapATAC. In each cell, suppose that ATAC bins $b_1, b_2, \cdots, b_L$ are open. The cell sentence is formulated as $(b_1, b_2, \cdots, b_L)$. As the input to the model, some open bins will be removed from the sentence (as masked tokens in BERT).

Bin ID Embedding

Since the vocabulary size is very large, to reduce the number of parameters, we use **embedding factorization**: The embedding matrix E (of size $V \times d$, where V is the vocabulary size and d is the embedding dimension) is factorized into:

- E_1: A smaller embedding matrix of size $V \times k$ (where $k \ll d$);

- E_2: A projection matrix of size $k \times d$, which maps the reduced k-dimensional embeddings to the final d-dimensional space.

In our case, we take $V = B + 25$, where 25 is for the embeddings of $< cls >$, $< pad >$ and $< eos_1 >, \cdots, < eos_{23} >$ for 23 chromosomes. We take $k = 64$ and $d = 512$.

Note: If this does not work, we can use full size bin ID embedding (512 dimensions).

Transformer Encoder

12 transformer encoder layers built by Flash Attention. The maximum sequence length (of cell sentence) is fixed to 6800 (8000 × 0.15) by default.

Decoder and MLM Loss

The output embeddings of $< eos_1 >, \cdots, < eos_{23} >$ (512 dimensional) are first compressed to 64 dimensional, and then concatenated with bin ID embeddings on the corresponding chromosome (e.g. $< eos_1 >$ is concatenated with all bins on chromosome 1) into 128-dimensional feature vectors. These feature vectors are feed into an MLP decoder to output the probability of 3 states: 0 (not in the cell sentence), 1 (in the cell sentence) and 2 (in the cell sentence but be removed).

Note: If this does not work, we can concatenate full size embedding (512 dimensions + 512 dimensions) as the input to the decoder, but it will increase the computational burden.

DNA Embedding (Optional)

We plan to use **Nucleotide Transformer** to build DNA embedding table for all 5000-bp DNA sequence of each bin. For special tokens, the DNA embedding can be fixed to zero vector.

Supervised Loss (Optional)

The output embedding corresponding to $< cls >$ is used to predict the organ and the celltype of the cell using MLP. For multi-omics data, we achieve the celltype label for scATAC-seq data by running **sCimilarity** for corresponding scRNA-seq data.

2 Future Perspective

Decoder by Genomic Regions

The $< eos >$ can be designed for different genomic regions instead of chromosomes. The following are possible annotated regions for ATAC bins.

1. **Gene-related Regions**: Promoter, Exon, Intron, 5' UTR, 3' UTR, TTS.
2. **Non-gene-related Regions**: Intergenic, Non-coding RNA (ncRNA).
3. **Regulatory Elements**: Enhancer, Insulator, CTCF Binding Site.
4. **Other Features**: Repeat, Conserved.

We can use **HOMER** to annotate ATAC bins. For each class c of bins, we use one $< eos_c >$ to predict the open/closed state of ATAC bins in this class.

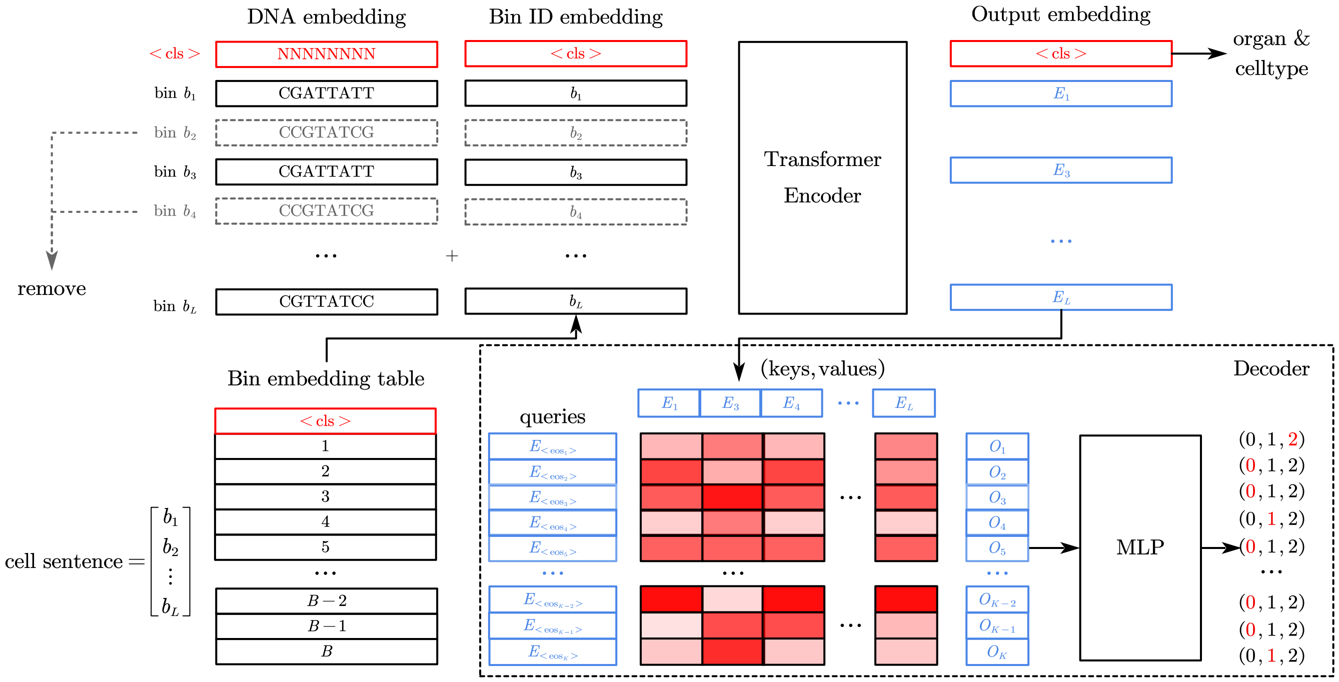

Fig. 2 | Future perspective of using decoder by cross attention.

Decoder by Cross Attention

Use independent $< eos >$ embedding vectors (not in the cell sentence).

- Use a Transformer encoder to encode the cell sentence, obtaining embeddings $E_{\text{encoder}} \in \mathbb{R}^{L\times d}$.
- Define $< eos_k >$ tokens corresponding to different regions. Each $< eos_k >$ token has an embedding $E_{<eos_k>} \in \mathbb{R}^d$.
- For each $< eos_k >$ token, compute cross-attention: $\text{Output}^k = \text{Attention}(E_{<eos_k>}, E_{\text{encoder}}, E_{\text{encoder}})$. Each output Output^k represents the interaction between the $< eos_k >$ token and the cell sentence.
- For each Output^i, predict the added/dropped bins in the corresponding functional region: $\text{Prediction}^i = \text{MLP}(\text{Output}^i)$

The advantages of this strategy are:

- Compared with self-attention, each $< eos >$ token can focus on a specific functional area and dynamically select bins related to it through cross-attention.
- Each $< eos >$ token can interact with the encoder output independently, which can avoid $< eos >$ tokens interacting with each other in the self-attention.

3 Pretraining Plan

Pretraining Steps

1. Pretraining for **masked ratio=0**, across datasets from HuBMAP.
    - Sampling closed bins v.s. no sampling (EpiAgent);

    - Class embedding v.s. no class embedding;
    - Bin ID embedding: full size v.s. small size;
    - Decoder input: full size v.s. small size;
    - MLP v.s. decoder by cross attention (optional);
    - $< eos >$ assigned according to chromosomes v.s. genomic regions (optional).

2. Pretraining for **masked ratio=0.15**, across datasets from HuBMAP.
    - DNA embedding v.s. no DNA embedding;
    - MLP v.s. decoder by cross attention;
    - $< eos >$ assigned according to chromosomes v.s. genomic regions.
3. Pretraining with **supervised learning**, across datasets from HuBMAP.
4. Pretraining across all available datasets.

Additional Tasks

1. Annotate the genomic regions for ATAC bins by **HOMER**.
2. Obtain DNA embeddings for the DNA sequences of ATAC bins by **Nucleotide Transformer**.
3. Obtain or align the celltype labels by **sCimilarity**.
4. Collect datasets from other data sources, and transform them into **cell by bin** format.