GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences

Source

The article addresses the challenge of handling extensive nucleic acid sequences in biological tasks using transformer models. Specifically, it focuses on improving the input size capacity of existing models to effectively analyze and predict biological signals in DNA sequences.

Related works in the field include DNABERT, BigBird, and Nucleotide Transformer models, which allow advanced genomic sequence analysis in downstream tasks. Despite these advancements, there is a need for new solutions like GENA-LM due to limitations in input size capacity of existing models. The proposed GENA-LM model addresses this limitation by incorporating techniques like Byte-Pair Encoding (BPE) and Recurrent Memory Transformer (RMT) to handle longer sequences and improve performance on biological tasks.

RMT is a relatively new architecture that enhances the input capacity of transformer models by incorporating recurrent memory mechanism. It divides input sequences into segments, processing them sequentially while utilizing memory tokens to pass information between consecutive segments. This design allows the model to effectively handle extended input sequences and maintain constant memory consumption. By integrating RMT, transformer models can function as a single recurrent unit, addressing one segment at a time and improving performance on tasks requiring analysis of long genomic sequences.

The input data for GENA-LM consists of genomic sequences processed into 'sentences' and 'documents' for training, with data augmentation techniques applied. Performance metrics include evaluating the model's accuracy on specific tasks like promoter prediction using human sequences from the EPDnew database. Specialized tokenization and fine-tuning strategies are employed to optimize the model's performance on various other biological tasks.

The major results of the proposed GENA-LM solution include outperforming existing models like DNABERT and BigBird, especially in handling longer input sequences and improving performance on biological tasks. However, limitations may include the computational resources required for training specialized models and the need for further optimization to achieve superior quality across all biological tasks and datasets.