[BIOINFO 2023]

Multi-task learning with modeling gene interaction using transformer for predicting patient outcomes

Bonil Koo[1,2], Dohoon Lee[3,4], and Sun Kim[1,2,5,6*]

[1] *Interdisciplinary Program in Bioinformatics, Seoul National University*
[2] *AIGENDRUG Co., Ltd.*
[3] *Bioinformatics Institute, Seoul National University*
[4] *BK21 FOUR Intelligence Computing, Seoul National University*
[5] *Department of Computer Science and Engineering, Seoul National University*
[6] *MOGAM Institute of Biomedical Research, Seoul National University*
** Corresponding author: sunkim.bioinfo@snu.ac.kr*

Multi-task learning aims to jointly learn multiple related tasks for generalized representation of the data useful in multiple contexts. It is widely studied in various fields such as computer vision, natural language processing (NLP), and drug discovery. In healthcare, although the omics data underlies a patient's multiple conditions and information about these conditions can be learned jointly, the prediction task has mainly been performed as a single task. Transformer contributes greatly to interpretability as well as performance improvement in the various fields. It is widely used not only in the field of NLP where it was developed, but also in computer vision, DNA/protein sequences, and chemical compounds. The core module of transformer is multi-head attention layer, which is expected to be effective in modeling gene functions in multiple contexts since genes can have multiple interaction modes with other genes. In this study, we propose a multi-task learning framework that combines the advantages of the transformer architecture with the unique challenges of using omics data as input to model the multi-functional roles of genes effectively. Transformer is designed to work with sequences of tokens, but omics data usually comes in the form of numerical values (e.g. gene expression quantity). We devised a novel positional encoding strategy tailored for omics data, and demonstrated that our framework outperforms single task learning and is effective in various prediction tasks related to patients' multiple outcomes or diagnoses using omics data. Our results reveal that our model captures meaningful biological patterns and dependencies among genes, thus providing valuable insights into gene interactions in driving different patient outcomes. This work paves the way for more effective and interpretable multi-task learning approaches in the field of bioinformatics, with potential applications in healthcare and personalized medicine.