УДК 004.89

ENHANCING BRAIN TUMOR CLASSIFICATION VIA PROMPT OPTIMIZATION IN VISION-LANGUAGE MODEL

Yang Jiafeng (ИТМО)
Scientific supervisor – Doctor, Associate professor Zhukova Natalia Alexandrovna (ИТМО)

Introduction. Modern multimodal models such as CLIP (Contrastive Language-Image Pre-training) [1] open new possibilities for medical diagnostics through their ability to combine visual and textual information. A key advantage of such models is their capability to utilize natural language descriptions to improve image classification. However, the effectiveness of pretrained models in specialized medical tasks is often limited due to the semantic gap between general and medical data. This study investigates the application of the Context Optimization (CoOp) [2] method to overcome this limitation in the classification of brain tumors using MRI images.

Main Content. This study utilized CLIP, a vision-language model that learns visual concepts from natural language supervision. CLIP employs a dual-encoder architecture where an image encoder and a text encoder are trained to maximize the similarity between matching image-text pairs while minimizing it for non-matching pairs. Experiments were conducted on a dataset of 7,023 brain MRI images comprising four classes: glioma, meningioma, pituitary, and no tumor [3]. When applied directly to brain tumor MRI classification with standard text prompts, the pretrained CLIP model (ViT-L/14@336px) achieved only 26.3% accuracy, highlighting the challenge of adapting general-domain visual-language models to specialized medical tasks. To address this limitation, we implemented the Context Optimization (CoOp) method, which learns a set of continuous context vectors that act as optimizable prompt tokens. Unlike traditional fixed prompts, CoOp allows the model to adapt its text representations specifically to medical imaging features while leveraging CLIP's pretrained visual-semantic knowledge. By employing CoOp to optimize the prompting strategy, the classification accuracy significantly improved to 94.8%. This substantial improvement demonstrates that appropriate prompt tuning can effectively bridge the domain gap between general visual-language pretraining and specific medical image analysis tasks.

Conclusions. This study demonstrates the critical role of text prompt optimization in improving the effectiveness of multimodal classification of medical images. The significant improvement in accuracy (from 26.3% to 94.8%) confirms that targeted tuning of textual descriptions through prompt tuning can overcome the limitations of pretrained models when working with specialized medical data.

List of references::

1. Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.
2. Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." International Journal of Computer Vision 130.9 (2022): 2337-2348.
3. Brain Tumor MRI Dataset. – URL: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.