УДК 004.08

Self-Sueprvised Pretraining for Visual and Language Transformer using Contextual and Weakly-annotated Youtube dataset

Ali Mohamamd (ITMO university), **Andrey Filchenkov** (ITMO university)

Введение. Cross-modality information retrieval has gained much attention in recent years. Novel approaches have benefited from advancements in deep learning methods, especially Transformer models trained with self-supervised objectives. These objectives are designed separately for each modality, and these pretrained models can then be fine-tuned on cross-modality objectives. While Masked Language Modeling (MLM) [1] is accepted as the mainstream objective for the textual modality, the visual modality lacks effective self-supervised objectives. Current approaches focus on additional offline processing using pretrained networks for region-tag identification and other semantic relationships [2], or use densely annotated datasets to learn connections between the visual and textual modalities [3]. Moreover, the datasets often used in the literature lack any contextual aspects for captions, which consist of a short description of the objects appearing in the image. This deprives the model of important semantic relationships between the different modalities. These approaches are not scalable, as they require expensive computational costs or expensive data annotation.

Основная часть. In our work, we present a new approach for Transformer's training using self-supervised objectives for both modalities. We follow the literature in using MLM as the main objective for the textual modality [4]. We add input masking and a reconstruction objective for the visual modality inspired by its success in pretraining vision Transformers [5]. As a result, the model accepts tokens from both modalities as input and is required to reproduce the input using the learned embeddings. To bridge the gap between the two modalities during training, we propose the usage of alternate modality masking, where each modality is fully masked randomly during training to incentivize cross-modality information transfer. We also present a scalable approach for collecting contextual and weakly annotated image/text pairs from video sources without the need for human annotation. In this dataset, captions are received using an off-the-shelf speech-to-text Whisper model [6], and key frames in the videos are used as images. We pretrain our architecture using our custom objectives on a sample dataset extracted using our approach and validate the results on different downstream tasks.

Выводы. Our approach has the potential to significantly improve pre-training results for cross-modality information retrieval, as it can be leveraged for large-scale pre-training on any dataset and has objectives for high and low-level understanding of both textual and visual modalities. We also proposed the use of alternate modality masking during training to encourage cross-modality information transfer. In addition, our data collection method has the potential to produce large datasets for joint textual and visual tasks, and can also be used in cross-modality generative models. Furthermore, our method can be tuned to produce data for a specific visual or textual style by simply choosing the appropriate videos. We believe that our approach can be extended to other applications as well, and has the potential to significantly advance research in cross-modality information retrieval and related areas.

Список использованных источников:

1. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep

Bidirectional Transformers for Language Understanding,” Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

2. Zhou, Mingyang, et al. "Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

3. Huang, Zhicheng, et al. "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers." arXiv preprint arXiv:2004.00849 (2020).

4. Kim, Wonjae, Bokyung Son, and Ildoo Kim. "Vilt: Vision-and-language transformer without convolution or region supervision." International Conference on Machine Learning. PMLR, 2021.Chen, Xiaokang, et al. "Context autoencoder for self-supervised representation learning." arXiv preprint arXiv:2202.03026 (2022).

5. Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." arXiv preprint arXiv:2212.04356 (2022).

Ali Mohammad (автор) Подпись

Andrey Filchenkov (научный руководитель) Подпись