Meeting Note

Author: Henry Lin
Date: 2024/07/08 – 2024/07/12

Summary
CLIP. [1]
Dense Retrieval / Learned Sparse Retrieval. [2]
Multi-Label CLIP. [3, 4, 5]
Difference between contrastive learning and dense retrieval.

Plan for Next Week
A text-to-text contrastively pretrained model for language translation. (Different text encoders)
A text-to-text (topic-to-content / content-to-content) contrastively pretrained retrieval model.

Problems
We want to design an architecture for multi-modal / single-modal dense retrieval task. What are the existing methods and researches? [2]

Details
Alex et al. [1] introduced the Contrastive Language-Image Pre-training (CLIP) model which can learn the multimodal representation for each text and image data. The simplified training process is as follow: first, prepare a pair of text and image; second, obtain two vectors of text and image from encoders; third, calculate the contrastive loss of the two vectors.
Basically, the encoder plays two roles in CLIP, one is the feature extractor, the other is projector (projecting feature map onto same latent space). What if we separate the projector from the encoders? The research in the next section shows the similar method.

Thong et al. [2] introduced a multi-modal sparse projector called Dense2Spare (D2S), this model takes the outputs of text and image encoder as the input, and sparse vectors as output. In the training step, the projector will learn how to project data from their feature map to a shared latent space.

This architecture may address the problem of semantic deviation, but they didn't consider what if the output of image encoder is similar to the output of the text encoder although they have the different labels, this may let the projector "confuse". In this study, they used the pretrained image and text encoder using different architecture and even trained on different dataset, so there's no promise that the feature map obtained from the encoders will be distinguishable.

Conclusions

Contrastive learning is basically a process to project a pair of data onto a high-dimensional space, and pull the similar data closer, push the different data further. And dense retrieval is a process to convert a data into a dense vector, then use the vector to find another similar data.

There are many similarities between contrastive learning and dense retrieval, for example, we can use the contrastive loss to train a single-modal shared encoder dense text retrieval modal (specifically replacing the image encoder with text encoder in the original CLIP modal and use text-text pairs to train it). This model may not only be useful for topic-to-content retrieval, but may also be useful for language translation (the model can learn the contrastiveness from the article pairs of two different languages).

References

[1] RADFORD, Alec, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, 2021. p. 8748-8763.

[2] NGUYEN, Thong, et al. Multimodal Learned Sparse Retrieval with Probabilistic Expansion Control. In: European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2024. p. 448-464.

[3] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, Song Wang: "CDUL: CLIP-Driven Unsupervised Learning for Multi-Label Image Classification", 2023; arXiv:2307.16634.

[4] Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, Deng Cai: "TagCLIP: A Local-to-Global Framework to Enhance Open-Vocabulary Multi-Label Classification of CLIP Without Training", 2023; arXiv:2312.12828.

[5] Yanming Guo: "Multimodal Multilabel Classification by CLIP", 2024; arXiv:2406.16141.