Project Title

UNIVERSAL VISION-LANGUAGE DENSE RETRIEVAL: LEARNING A UNIFIED REPRESENTATION SPACE FOR MULTI-MODAL RETRIEVAL

Team Members

Mingjun Wen
Xuqi Zhu
Dongjing Xie

Description of the Problem

Although search engines primarily focus on textual data (Singhal et al., 2001), multi-media is necessary to satisfy user needs during retrieval. A user query can be answered by the information in variant formats, such as a text document, or a picture. The growth of multi-media content has been one of the most notable trends on the internet (Mei et al., 2014), and various studies have proved that users prefer more vivid multi-media content in search results (Datta et al., 2008).

A brief survey of what have been done and how the proposed work is different：

Current multi-media search systems often employ a divide-and-conquer approach.However, due to the modality gap, they can be only pipeline-modeled in divide-and⊴conquer, making it challenging to fuse retrieval results from different modalities.

In this paper, we explore the potential of universal multi-modal retrieval to build an end-to-end model and retrieve multi-modality documents for user queries.

More specifically, we propose a Universal Vision-Language Dense Retrieval (UniVL-DR) model to get the representations of queries, texts, and images and learn a tailored vision-language embedding space for multi-modal retrieval. UniVL-DR optimizes the vision-language embedding space using hard negatives (Xiong et al., 2021a) and balances the modalities of these negatives to alleviate the modality preference of multi-modal retrievers.

Preliminary Plan (Milestones)

10/15/2023

- Review and study all reference papers, focusing on understanding the principles and logic behind the new algorithms presented in the articles.

10/25/2023

- Reproduce the UniVL-DR algorithm demo presented in the article.

11/5/2023

- Complete the initial draft and explore potential improvements and modifications.

11/20/2023

- Validate and organize various tests, and compile the final report.

Reference

George Awad, Asad A Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, et al. 2021. Trecvid 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Ma⊴jumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective condi⊴tioned and composed image retrieval combining clip-based features. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022,pages 21434–21442. IEEE.

Patrick Cavanagh. 2021. The language of vision*. Perception, 50(3):195–215

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk.2022. Webqa: Multihop and multimodal qa. In Proceedings of CVPR.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In Proceedings of ICML, pages 1931–1942.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (Csur), (2):1–60.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186.

Sounak Dey, Anjan Dutta, Suman K. Ghosh, Ernest Valveny, Josep Lladós, and Umapada Pal. 2018. Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch. In 24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24,2018, pages 916–921. IEEE Computer Society