Visual Language - Let the Product Say What You Want

Jiaying Wang, Shuailing Hao, Jing Shan, Xiaoxu Song

Shenyang University of Technology
No.111, Shenliao West Road, Economic and Technological Development Zone
Shenyang, Liaoning 110870 P.R.CHINA
{jiaying, shanjing}@sut.edu.cn

Abstract

Visual Language is a multitasking on-line system focusing on e-commerce, which involves in generating accurate product descriptions for sellers and providing convenient product retrieval service for customers. To achieve this goal, the system adopts image description technology and multi-modal retrieval technology. By utilizing cross-modal generation technique, we could help sellers on rapid uploading products and customers on rapid retrieval, which could improve the experience of both sellers and customers.

System Framework

Visual Language system utilizes current state-of-the-art technologies and modules to extract fine-grained features from user-supplied product images, videos, and text. By combining multimodal technology with a knowledge base, feature similarity computation and conversion between different modalities are realized, thus enabling product description and multimodal graphic retrieval.

System Mechanism

Product label generation In the e-commerce task, products only exist in a specific region in the image. However, considering the image as a whole for cross-modal alignment with the text will inevitably bring in a large number of noisy background objects. In order to solve this problem, we need to label products in the image to extract the product information. First, we used the pre-trained model CLIP (Radford et al. 2021) as an image feature extractor to obtain image features. Second, we use ViT (Dosovitskiy et al. 2021), high quality phrases (Wang et al. 2021) and image library to build a knowledge base that contains a large number of fine-grained semantic labels, such as "Nordic style" and "soft decoration living room". Finally, we calculate similarity between image features and semantic labels in the knowledge base to get product labels. We use product labels and image features to generate fine-grained descriptions of products.

Fusion of multidimensional features It is a non-trivial task to ensure the full fusion of multidimensional features. To ensure that the fused features and the corresponding text features can be perfectly connected and smoothly integrated into the language model GPT-2 (Radford et al. 2019), we utilized a spatial mapping approach (Mokady, Hertz, and Bermano 2021) to accomplish the fusion of multidimensional features of products and modal transformations to provide more informative and contextualized inputs for subsequent tasks. Specifically, we utilize the pre-trained BERT (Devlin et al. 2019), and use image coding features with matched product label features as inputs. The powerful fusion capability of the model not only fully fuses the representation between multi-dimensional product labels and image features, but also ensures that the fused features and the subsequently generated text features are in the same mapping space. Thus, we get fine-grained product descriptions.

System Function Description

Our target is to develop a simple, easy-to-use, multimodal, multitasking and unified system to meet the needs of sellers and customers. Fig. 2 presents functional demonstration of Visual Language system.

Image and Video Description Generation This system provides two interfaces to support generating descriptions of product images and videos. For images uploaded by users, the system is able to obtain their fine-grained representations; for videos, the system first splits the frames according to the time sequence and then extracts the product features in the images. The fused multidimensional feature representation of the product obtained through the encoder is integrated into the language model as a visual prerequisite for generating content, which continuously influences the model generation process and makes the generated content more focused on the encoded characteristics of the product. This is designed to improve the accuracy and relevance of the generated content, making it more consistent with the real characteristics of the product.

Multimodal Retrieval In addition to image and video description generation, the system can also perform multimodal image retrieval, which extracts the features from the product image or text described by a customer, and matches them with image features in the knowledge base, it can achieve the fast retrieval of image and text. Our approach is inspired by the retrieval task in (Yang et al. 2022). Based on the knowledge base of image features constructed in the previous module, we use a multimodal encoder to encode the

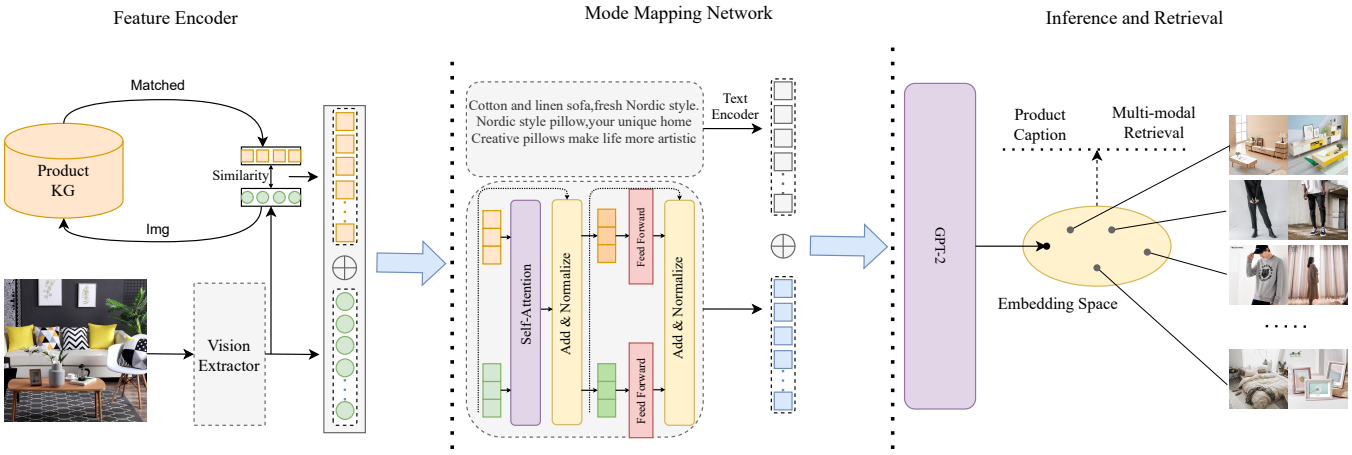

Figure 1: Framework of the Visual Language system. The framework is divided into three modules. Feature Encoder: The fine-grained product knowledge base is introduced into the feature encoding stage to enrich the image features. Mode Mapping Network: The product label attribute features and image features are integrated so that the acquired features contain more elements of the product. Inference and Retrieval: It handles downstream tasks, including the generation of product descriptions and the retrieval of products.

image and text respectively. The system quickly matches the best image features using a retrieval algorithm (e.g. KNN) by calculating the similarity between query encoded features and image features in the knowledge base.

Custom Settings: In this module, we have built in a variety of custom setting options, including the selection of image or video descriptions, the number of descriptions generated, diversity control, and support for different languages. These customized features are designed to meet the individual needs of users. Through these functions, users can gain greater flexibility and creative idea in the process of product display and promotion, and further enhance the user experience.

Related Works

Multimodal image descriptions have gained popularity in e-commerce for their ability to automatically generate product descriptions, accelerate the speed of product uploading, and optimize platform efficiency. Although previous methods have addressed some challenges in cross-modal alignment and information integration, there is still room for improvement.

To address fine-grained cross-modal alignment, FashionViL (Han et al. 2022) introduced a framework for learning fashion-centric visual and linguistic representations. However, it is lacking in fine-grained cross-modal alignment. Another approach, K3M (Zhu et al. 2021), proposed a structural aggregation module to integrate information from images, text, and knowledge modalities. While it focused on entities in textual modalities and lacked sufficient cross-modal interaction.

In terms of image description techniques, approaches like object detection (Li et al. 2020) and scene graph parsing (Cui et al. 2021) have limited effectiveness in e-commerce tasks. In the product retrieval task, we discovered a more effective

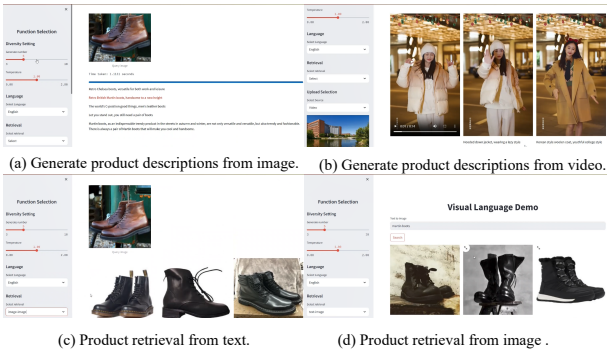

Figure 2: Functional demonstration of Visual Language system. Functions include generating product descriptions based on images or videos, and retrieving products based on images or text.

solution by incorporating more detailed product introduction labels. Inspired by the construction of multimodal concept-level knowledge graphs (Wang et al. 2023), we were able to address the issue of inadequate granularity in product descriptions.

Conclusion and Future Work

Visual Language System is an approach that combines a pre-trained model and a knowledge base. It generates accurate product descriptions for sellers and provides convenient product retrieval for customers. In future work, we plan to further explore the application of the fine-grained labeling task for products in image feature extraction, especially for the accurate identification of unique geographic indications for different products and generate corresponding descriptions.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China(Nos.61702346, 61702345 and 61802268) and Basic Scientific Research Foundation of Liaoning Provincial Department of Education(No. JYTMS20231226).

References

Cui, Y.; Yu, Z.; Wang, C.; Zhao, Z.; Zhang, J.; Wang, M.; and Yu, J. 2021. ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross- and Intra-modal Knowledge Integration. In Shen, H. T.; Zhuang, Y.; Smith, J. R.; Yang, Y.; César, P.; Metze, F.; and Prabhakaran, B., eds., *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 797–806. ACM.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Han, X.; Yu, L.; Zhu, X.; Zhang, L.; Song, Y.; and Xiang, T. 2022. FashionViL: Fashion-Focused Vision-and-Language Representation Learning. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV*, volume 13695 of *Lecture Notes in Computer Science*, 634–651. Springer.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, 121–137. Springer.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. ClipCap: CLIP Prefix for Image Captioning. *CoRR*, abs/2111.09734.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Wang, J.; Shan, J.; Santos, O. E.; and Bao, J. 2021. High quality error-tolerant phrase mining on text corpus. *Expert Syst. Appl.*, 171: 114557.

Wang, X.; Wang, C.; Li, L.; Li, Z.; Chen, B.; Jin, L.; Huang, J.; Xiao, Y.; and Gao, M. 2023. FashionKLIP: Enhancing E-Commerce Image-Text Retrieval with Fashion Multi-Modal Conceptual Knowledge Graph. In Sitaram, S.; Klebanov, B. B.; and Williams, J. D., eds., *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL 2023, Toronto, Canada, July 9-14, 2023*, 149–158. Association for Computational Linguistics.

Yang, A.; Pan, J.; Lin, J.; Men, R.; Zhang, Y.; Zhou, J.; and Zhou, C. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *CoRR*, abs/2211.01335.

Zhu, Y.; Zhao, H.; Zhang, W.; Ye, G.; Chen, H.; Zhang, N.; and Chen, H. 2021. Knowledge Perceived Multi-modal Pretraining in E-commerce. In Shen, H. T.; Zhuang, Y.; Smith, J. R.; Yang, Y.; César, P.; Metze, F.; and Prabhakaran, B., eds., *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 2744–2752. ACM.