ARIA
APPLIED RESEARCH IN ACTION

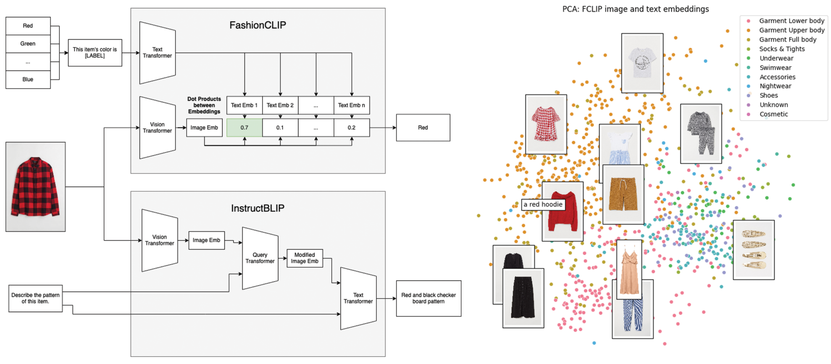
Image Captions Are Worth a Thousand Words: Enhancing Search in Conversational

Improving Conversational Agents through Image-Informed Text Generation.

Jason Tang

Eldan Cohen
ACADEMIC SUPERVISOR

Garrin McGoldrick and Marie Al Ghossein
INDUSTRY SUPERVISORS



PROJECT SUMMARY

In the realm of eCommerce, conversational agents powered by Large Language Models (LLMs) offer a natural means for users to convey shopping intents and locate desired products. As text-based models, these LLMs output textual search queries based on user conversations, which are subsequently used for retrieving relevant items from store catalogs. This limits the adoption of conversational agents in many eCommerce settings, particularly for smaller websites that struggle with creating and maintaining meaningful textual descriptions for items. However, nearly every eCommerce website is compelled to maintain high-quality product images that cater to the fundamentally visual nature of online shopping. To leverage this, our project investigates strategies such as pre-trained image captioning models to leverage these catalog images in enhancing text-based retrieval within conversational recommendation systems.

For evaluation, we use Amazon's publicly available ESCI dataset, consisting of customer search queries and labeled relevance scores for a subset of Amazon products. We employ the customer queries as search queries and contrast the relative retrieval capabilities between our image-generated text and the original text data. This investigation has the potential to significantly enhance user experiences in product discovery and shopping, and can also prove valuable in the separate application of automatic data labeling.

REFERENCES

[1] W. Dai et al., InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. 2023.
[2] P. J. Chia et al., Contrastive language and vision learning of general fashion concepts. 2023.
[3] J. Li, D. Li, S. Savarese, and S. Hoi, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023.
[4] "Introducing chatgpt," Introducing ChatGPT, https://openai.com/blog/chatgpt (accessed Aug. 23, 2023).
[5] A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision. 2021.
[6] A. Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021.
[7] A. Vaswani et al., "Attention is All You Need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010.
[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019.

Computer Science
UNIVERSITY OF TORONTO
Master of Science in Applied Computing