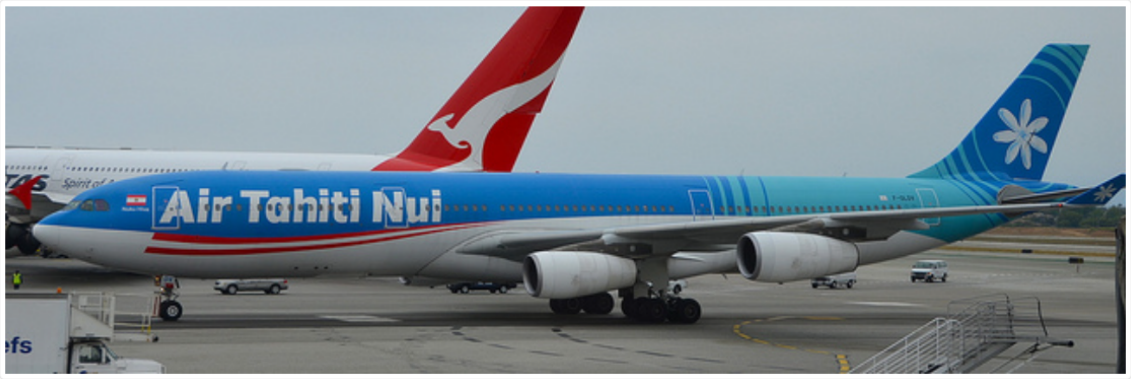
Incorporating Vision Encoders into Retrieval Augmented Visual Question Answering

Question : What country is named here?



Answer: tahiti
Answer Occurence: 5 / 5
Category: Vehicles and Transportation

Figure 1: Knowledge Based Visual Question Answering: an example drawn from the OK-VQA dataset [5]. Answering the question requires both image understanding and real-world knowledge – e.g. 'Tahiti is a country.' – which can be found in external knowledge sources.

External Knowledge Visual Question Answering can be considered a relatively simple task in machine learning that requires answering a direct question based on information in an accompanying image. The OK-VQA (Outside Knowledge) data set [5] is a more interesting alternative. VQA becomes more challenging when the answer to the question is not readily apparent in the image. In such cases the VQA system must retrieve information from external knowledge sources to generate a complete and accurate answer (see Figure 1). OK-VQA provides such a task.

Vision Encoders A straightforward approach to VQA is to convert images directly into text using image captioning models. Transformer-based large language models can then achieve excellent VQA performance [4, 2, 8] if the captioning model generates an accurate text-based description of the image. However this approach risks losing visual information by transforming images independently of the question to be answered. Vision-and-Language Transformers (Figure 2) offer a possible approach to avoiding information loss through joint embedding of the image and query, although such models can be costly to implement in terms of computational and data requirements.

RA-VQA Retrieval Augmented Visual Question Answering (RA-VQA) [3] is a VQA framework that first retrieves documents (as passages and snippets) from an external knowledge base and then generates an answer to question from the image and the retrieved documents. RA-VQA achieved state-of-the-art performance in 2022. However it relies on an image-to-text mapping which risks losing information, as described above.

Project Overview The project will investigate modelling approaches to enhance the vision understanding component of RA-VQA with the aim of improving performance on the OK-VQA dataset. Work will begin with a literature review of recent of developments, data sets, metrics, and modelling techniques that combine vision models and language Transformers. A plan of work will then be decided upon based on the student's interests and the resources available. Possible lines of work are:

- Introducing a *mapping network* into the RA-VQA framework to connect the vision model with a language model, as inspired by ClipCap [6].

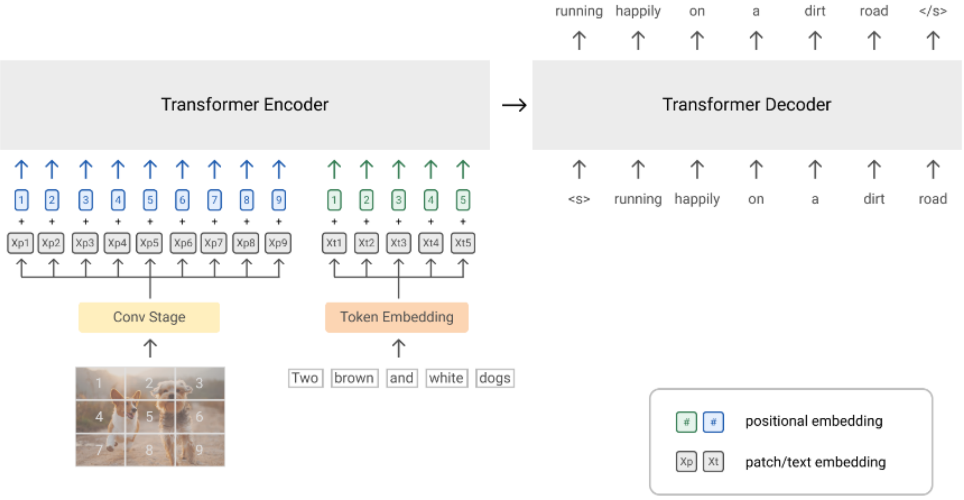


Figure 2: **SimVLM** [7] an end-to-end Vision-and-Language Transformer.

- Fine-tuning CLIP text/image encoders and BART encoders for multimodal generation [1].

Available Resources

- Pre-trained Vision-and-Language Transformers in Hugging Face: VisualBERT and ViLT.
- RA-VQA codebase. The framework can be easily migrated to other VQA datasets, if needed. https://github.com/LinWeizheDragon/Retrieval-Augmented-Visual-Question-Answering.

References

[1] W. Dai, L. Hou, L. Shang, X. Jiang, Q. Liu, and P. Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022.

[2] F. Gao, Q. Ping, G. Thattai, A. Reganti, Y. N. Wu, and P. Natarajan. A thousand words are worth more than a picture: Natural language-centric outside-knowledge visual question answering. *arXiv preprint arXiv:2201.05299*, 2022.

[3] W. Lin and B. Byrne. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.772.

[4] M. Luo, Y. Zeng, P. Banerjee, and C. Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.517.

[5] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019.

[6] R. Mokady, A. Hertz, and A. H. Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[7] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[8] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. *arXiv preprint arXiv:2109.05014*, 2021.