Large Language Models for Document Visual Question Answering

Abstract

Document visual question answering is an important tool to perform high-level reasoning and interpret document images. Nowadays, Large language models are becoming popular in question answering tasks. In this project, we aim to incorporate the large language models in a machine learning model that answers user questions and queries about a document image in a multi-modal fashion.

Keywords: Visual Question Answering, DocVQA, Large Language Models

Description:

Document intelligence is the field of research at the meeting point between Computer Vision (CV) and Natural Language Processing (NLP), focusing on techniques and methods for extracting, interpreting and inferring information from documents. It is a research field in rapid expansion with applications in many different sectors where the processing of large volumes of documents is critical, such as finance, insurance, public administration, businesses or personal document management. One of the subfields of document intelligence is Document Visual Question Answering (DocVQA) [1]. DocVQA is considered as a tool to perform high-level reasoning on document images and conditionally interpret the document information. The process consists of the following: inputting the document image with a requested question about it to a machine learning model, then, the model should output the answer. Several approaches have recently appeared following this process [2, 3]. DocVQA models include usually a vision encoder, a text encoder and a text decoder with a language model as illustrated in fig 1.

Large Language Models (LLM) are now popular in NLP tasks, mainly after the appearance of GPT variants, LLaMA [4], PaLM, etc. LLMs showed a great performance on question answering and context understanding, especially with the tool ChatGPT. However, GPT models are big and require huge resources to fine-tune them on desired tasks like DocVQA. Recently, some works appeared and show the possibility to fine-tune LLaMA with reasonable resources and get similar results to ChatGPT on many tasks [5, 6]. Another work presented in [7] even fine-tune these models to multi-modal tasks to simulate GPT 4 with less resources.

Motivated by this, we plan in this project to use the LLMs in a multi-modal DocVQA task. The intuition is that the large knowledge possessed by those models

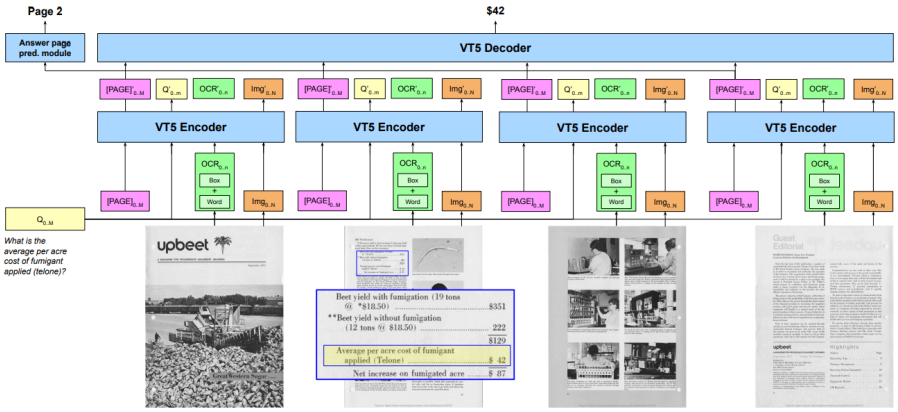

Fig. 1 An example of a DocVQA model. This figure is taken from [2].

shall benefit the outputted answer quality and precision for our task.

Work required by the student:

- Exploring the state-of-the-art DocVQA systems and reproducing their architectures and results.
- Including the LLM part to the DocVQA model and fine-tuning all the components in an end-to-end fashion.

Contact:

- **Mohamed Ali Souibgui**
  email: msouibgui@cvc.uab.es
- **Dimosthenis Karatzas**
  email: dimos@cvc.uab.es

References

[1] Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2200–2209 (2021)

[2] Tito, R., Karatzas, D., Valveny, E.: Hierarchical multimodal transformers for multi-page docvqa. arXiv preprint arXiv:2212.05935 (2022)

[3] Wu, X., Zheng, D., Wang, R., Sun, J., Hu, M., Feng, F., Wang, X., Jiang, H., Yang, F.: A region-based document vqa. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4909–4920 (2022)

[4] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[5] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford Alpaca: An Instruction-following LLaMA model. GitHub (2023)

[6] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90(2023). https://vicuna.lmsys.org

[7] Zhu, D., Chen, J., Shen, X., Li, Elhoseiny, M.: MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models (2023)