Explainable Document Visual Question Answering

Abstract

Document visual question answering (DocVQA) is an important tool to perform high-level reasoning and interpret document images. Despite the advancements in DocVQA models, the absence of explanations for the provided answers remains a notable limitation. This project endeavors to bridge this gap by developing a novel DocVQA model that not only generates accurate responses to questions but also offers comprehensive explanations to support the provided answers.

Keywords: Visual Question Answering, DocVQA, Explainable AI.

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 [2–4].

DocVQA models include usually a vision encoder, a text encoder and a text decoder with a language model as illustrated in fig 1. This complex architectures of the DocVQA models do not allow the probing and the analyzing of the internal reasoning process that led to the answer. However, in many real-world cases, documents carry sensitive information that is used to inform subsequent decisions, further increasing the transparency requirements for DocVQA methods.

Motivated by this, we plan in this project to explore the explainable AI methods [5], especially the ones using vision transformers [6]. Then, to develop a transparent DocVQA model that allows the explanation of its predictions.

Work required by the student

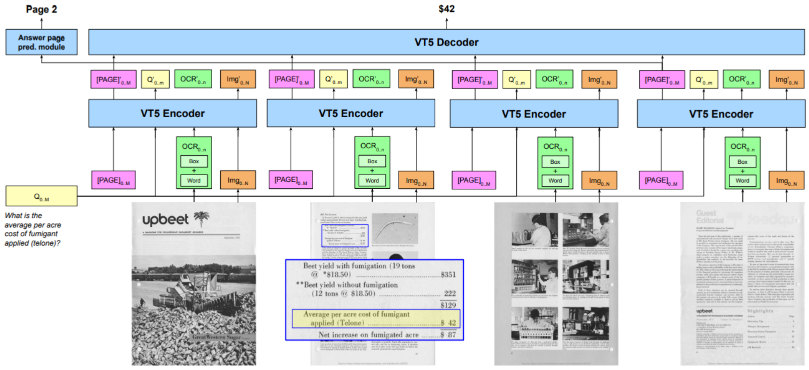
Page 2
$42
Answer page pred. module
VT5 Decoder
VT5 Encoder
[PAGE]
Q'
OCR'
Img'
OCR
Box
Word
Img
Q
What is the average per acre cost of fumigant applied (telone)?
upbeet
Beet yield with fumigation (19 tons @ $18.50) $351
**Beet yield without fumigation (12 tons @ $18.50) 222
$129
Average per acre cost of fumigant applied (Telone) $ 42
Net increase on fumigated acre $ 87

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

- Exploring the state-of-the-art DocVQA systems and reproducing their architectures and results.
- Integrating a new module that allows the explainability of the DocVQA model predictions.
- Train all the components in an end-to-end fashion.

Excellent students might be given the possibility to continue for a PhD at the Computer Vision Center following the successful completion of this project.

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] Lee, K., Joshi, M., Turc, I.R., Hu, H., Liu, F., Eisenschlos, J.M., Khandelwal, U., Shaw, P., Chang, M.-W., Toutanova, K.: Pix2struct: Screenshot parsing as pretraining for visual language understanding. In: International Conference on

Machine Learning, pp. 18893–18912 (2023). PMLR

[5] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., *et al.*: Explainable ai (xai): Core ideas, techniques, and solutions. ACM Computing Surveys **55**(9), 1–33 (2023)

[6] Böhle, M., Fritz, M., Schiele, B.: Holistically explainable vision transformers. arXiv preprint arXiv:2301.08669 (2023)