QALayout: Question Answering Layout based on multimodal Attention for visual question answering on corporate Document

Ibrahim Souleiman Mahamoud *, Mickaël Cousaty *, Aurélie Joseph +, Vincent Poulain d’Andecy +, Jean-Marc Ogier*
Email: firstname.lastname@univ-lr.fr, + Email: firstname.lastname@yooz.fr

Context
- Depending on their size, companies can process thousands of documents per day.
- Automating this process is a time and money saver for companies.
- This paper is about the work of my thesis. Yooz works with accountants and helps them to automate their processes.
- Other methods already exist and have several limitations:
  - Difficulty to extract information on a new types of documents
  - To be able to extract new information for a new client
  - Difficulty to interpret these results

Proposed Method
- QALayout is a visual question answering (VQA) method based on the state of the art of image or document processing [1].
- Our inputs are the inputs of the QANet[1] method + new features.
  - Text extracted from the document
  - The question
  - The image of the document
  - The bounding box of document
- The output of the method is the answer to the question asked.
- The encoder (convolution-layer and attention-layer) is used to have a mechanism of attention to these contexts.
- Self-attention inspired by [2] to focus our network on common features from the input (see Modified inputs in Fig 2) This will allow us to exploit the context and query correlation at the initial stage.
- Co-attention step proposed in our work is inspired by the attention flow layer from [3]. It calculates attention in several directions.

Goals and Challenges
The goals
- Extract data without business rules entered by an expert on a large and multilingual vocabulary. The models will be learned from examples of results.
- Be able to define the information to be extracted for each type of document (e.g. predfine a set of questions for each client). Then extract this information from documents using the context and link between these contexts as much as possible.
- Be able to learn continuously and extract new information on and new types of documents.

The challenges
- Simplicity and automation of learning can be achieved from naive examples, within the reach of an end-user.
- Satisfy a low processing time and respect the industrial constraints on error minimization.

New dataset VQA-CD

Table 1. This table contains the results of the proposed QALayout model and the results of the state-of-the-art method (LayoutLM,Bert)

Conclusions and future-work
- QALayout fast and accurate end-to-end method.
- This method uses several attention (attention for each input, self-attention, co-attention) for better performance and interpretation of results.
- We also contributed a new dataset VQA-CD containing 3000 questions on corporate documents.
- Some limitations exist, and we will try to provide a solution.
  - Build a graph system
  - Incremental learning
  - Add new inputs

References

Acknowledgment
This research has been funded by the LabCom IDEAS under the grand number ANR-18-LCV3-0008, by the French ANRT agency (CIFRE program) and by the Yooz company.