

Ibrahim Souleiman Mahamoud ⁺⁺, Mickaël Coustaty ^{*}, Aurélie Joseph ⁺, Vincent Poulain d'Andecy ⁺, Jean-Marc Ogier ^{*}
^{*} Email : firstname.lastname@univ-lr.fr , ⁺ Email: firstname.lastname@getyooz.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. predefine 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

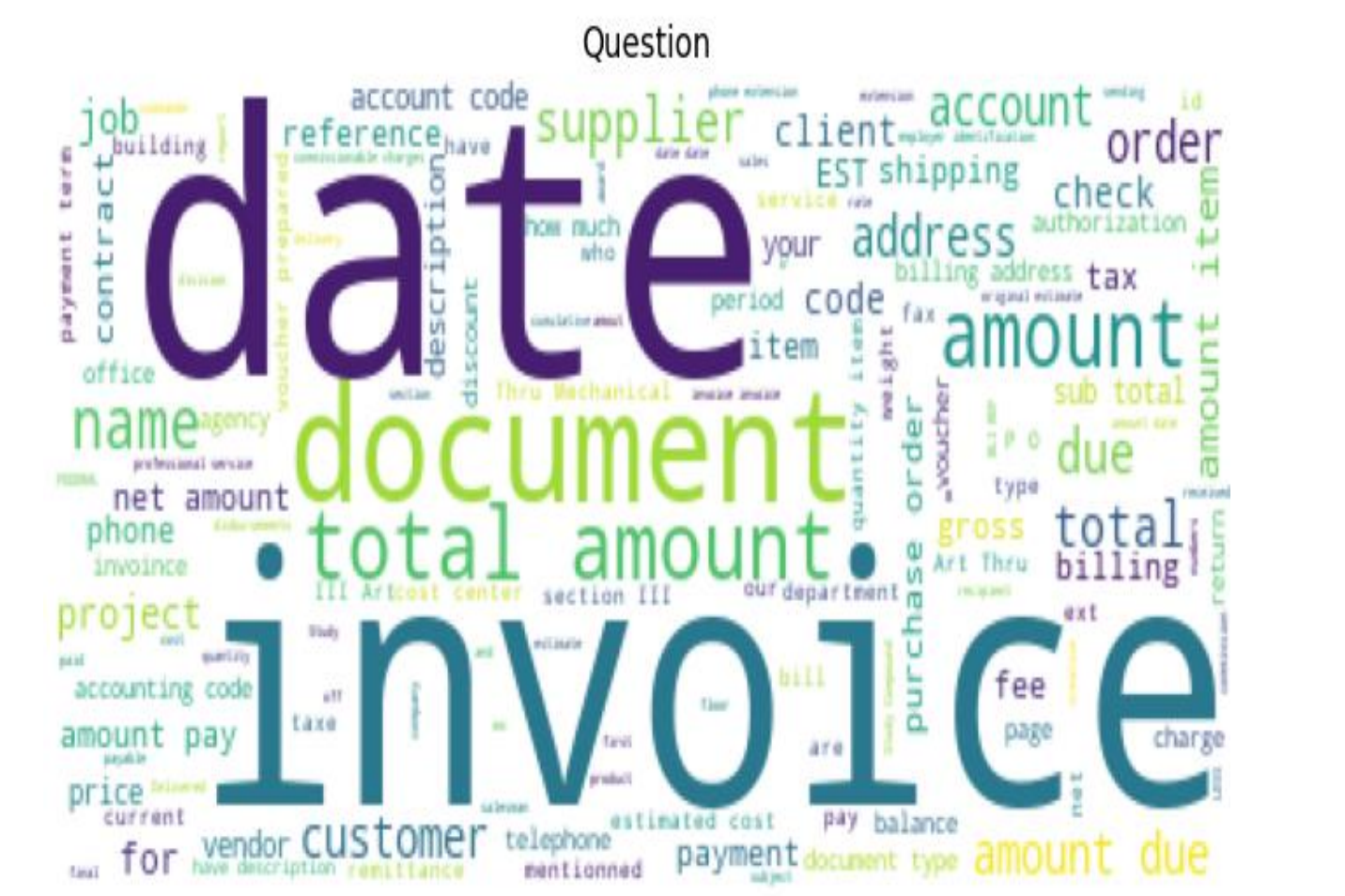


Fig 1. Distribution of questions on corporate documents (the new dataset VQA-CD)



QR Code 1. Sunburst chart with animations on the train corpus (VQA-CD)



QR Code 2. Sunburst chart with animations on the test corpus (VQA-CD)

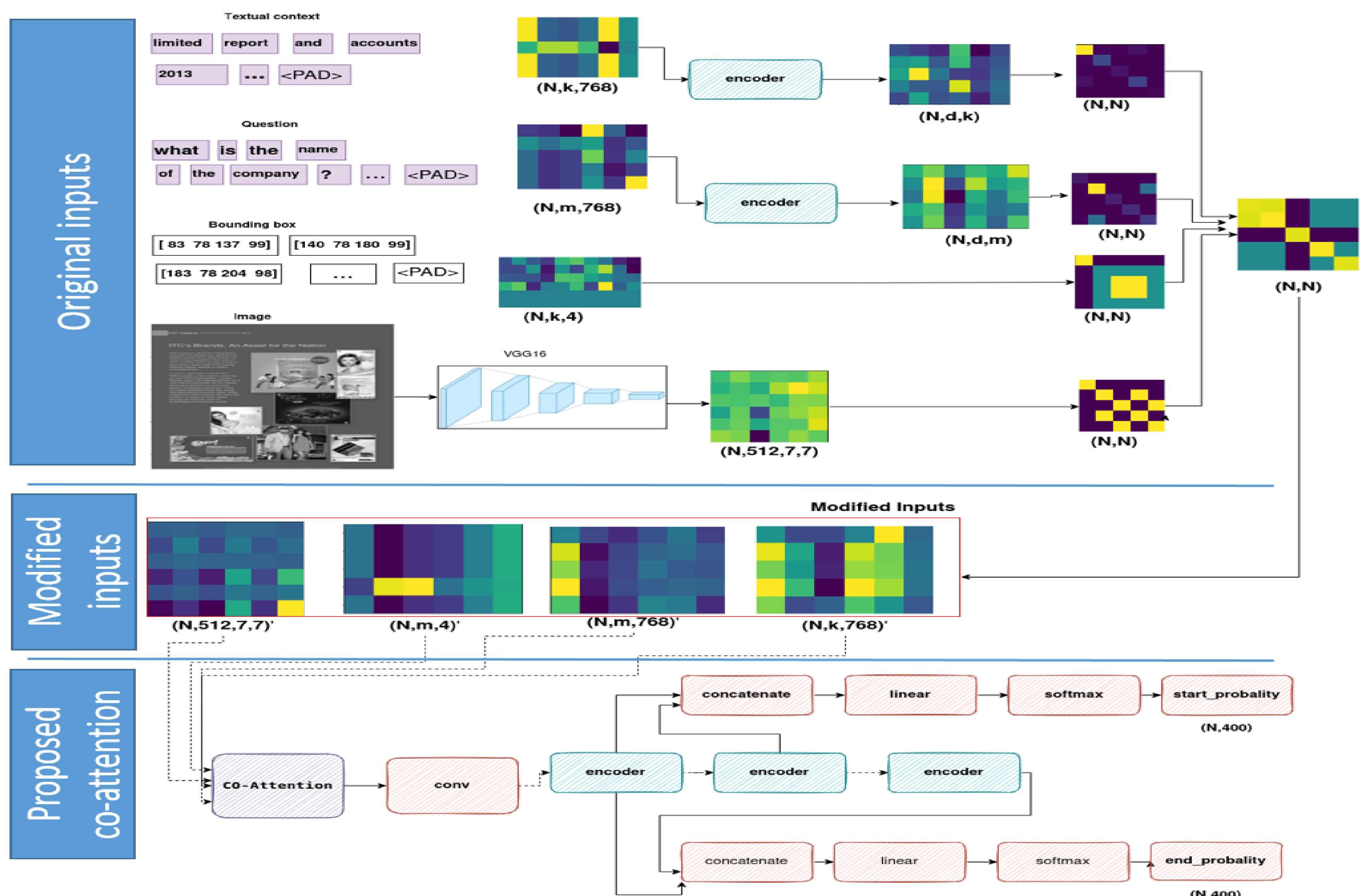


Fig 2. Schematic description of QALayout method

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

- Yu, A.W., Dohan, D., Luong, M.T., Zhao, R., Chen, K., Norouzi, M., Le, Q.V.: Qanet: Combining local convolution with global self-attention for reading comprehension (2018)
- Cheng, H., Zhou, J.T., Tay, W.P., Wen, B.: Attentive graph neural networks for few-shot learning (2020)
- Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension (2018)

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

Modality	Method	param	SQUAD	DOCVQA
			F1-score	ANLS
Text	Bert	~ 110 M	74.43%	45.57%
	QALayout (only_Text)	~ 1M	82.19%	48.63%
Text + Image + Layout	LayoutLM	~ 160 M	-	68.93%
	QALayout	~ 8M	-	77.65%

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