Towards General Document Understanding through Question Answering

Šárka Ščavnická, Michal Štefánik, Marek Kadlčík, Martin Geletka, and Petr Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
527352@mail.muni.cz

Abstract. Document Visual Question Answering is a relatively new extension of Visual Question Answering. The aim is to understand the documents and to be able to obtain information that corresponds to the question that was asked. This proposition aims to approach the problem of the lack of datasets and a model for Slavic languages. Therefore we would like to create a model and dataset for Document VQA suitable for the non-English language. This paper overviews the field of Question Answering and also describes the first Czech Document VQA dataset and model.

Keywords: Question Answering, Visual Question Answering, Document Visual Question Answering

1 Introduction

Document processing and analysis is a rapidly growing field. It applies not only to analysts who try to obtain and analyze critical information. It is also used in economic sectors, where they speed up the processing of legal documents and invoices. Several new works approach the document information extraction task through Visual Question Answering, due to which Document Visual Question Answering is created nowadays. This domain is very young; therefore, most of the models and datasets are only in English. This paper proposes a plan to create a system that can improve non-English Document understanding, specifically for the Czech language. First, we are trying to form the first Czech dataset for Document VQA. Furthermore, we plan to develop the first model for DVQA, which will process Czech invoices. Last but not least, we propose to address three research questions extending the applicability of general document extraction technology to non-English languages.

2 Background

This section discusses the current situation in Document processing and Question answering. The first part focuses on a general overview of Intelligent Document Understanding and Visual Question Answering. Subsequently, we overview available datasets and models in these fields.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 183–190, 2022. © Tribun EU 2022

2.1 Intelligent Document Understanding

With Intelligent Document Understanding (IDU), machines are able to comprehend and analyze unstructured data. Earlier works for document understanding [8,12] stood on a predefined set of rules. Therefore they required an exact definition of a template for every type of document that they were processing. [17]

IDU combines Natural Language Processing (NLP), Computer Vision, Machine Learning, and Deep Learning. Transformers are best suited for these tasks; for example, simple transformers are practical for NLP and Computer Vision tasks. On the other hand, Layout-aware Transformers are used for IDU because they can comprehend the layout information for the given document. As a result, LayoutLM [19] combines text, document layout, and visual information to extract practical knowledge from a document.

2.2 Visual Question Answering

Question Answering (QA) [3] models work with text and retrieve answers to the given question [13] based on the information they got from the text. This process is a combination of natural language processing and information retrieval fields. On the other hand, Visual Question Answering (VQA) focuses on understanding the visual data. Even if the images contain some text, this text is not considered when answering the given questions. However, there is also a combination of QA and VQA, which incorporates text from the scene of the images.

Document Visual Question Answering [7] seeks to obtain knowledge from documents to answer questions. The asked questions may relate to different parts of the examined document, not only the text part; for example, they may refer to inserted images, tables, and forms, but they may also refer to the overall arrangement of the text. Therefore, for Document VQA, we also need to incorporate the detection of scene objects and an understanding of the document's layout and the relations between different parts of the layout.

Presently, there is lacking coverage of non-English models for Document Visual Question Answering. For this reason, we would like to expand the coverage of the models on Document VQA by the Czech language model.

Figure 1 presents an overview of our proposed system for document processing based on VQA. The model will be trained on our Czech dataset for Document VQA. The system will be able to answer the entered questions by writing the required answer and highlighting this answer in the document. Highlighting in the document is essential, mainly because the text of the correct answer may be in several places in the document, but the correct answer is only one of them. On the other hand, it is also possible that the correct answer will be found more than once in the text, and it is crucial to highlight all the right answers.

2.3 Datasets

This section describes datasets used for Document Question Answering or Named Entity Recognition.

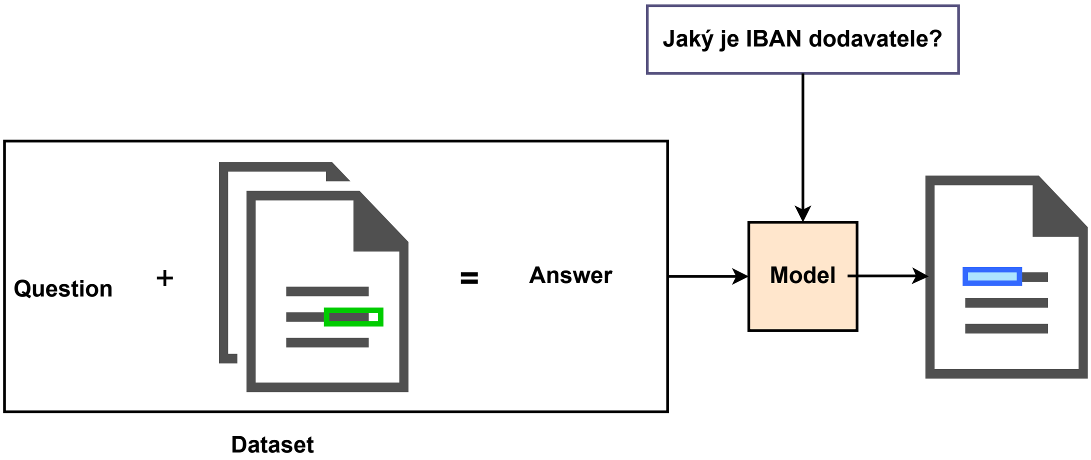

Fig. 1: Top-level approach for Document Visual Question Answering

Question Answering (QA) Nowadays, there are Question Answering datasets available in multiple languages. The most significant datasets are in English. These datasets differ in the size of question-answer pairs as well as the origin of these datasets. A large part uses texts from Wikipedia or newspaper articles, which are used to find answers to professional questions. Next, some datasets focus on technical areas, for example, on mathematical questions are also worth mentioning.

The first frequently used English dataset is the Stanford Question Answering Dataset (SQuAD) [13]. The original version of this dataset consists of over 100 000 questions posed on a set of more than 500 Wikipedia articles. The second version of the dataset has an additional 50 000 unanswerable questions. Conversational Question Answering (CoQA) [14] is another dataset in English. As is mentioned in the name of this dataset, it consists of more than 8 000 conversations and contains 127 000 questions, which also include evidence for answers. Answers can be free-form text.

The French Question Answering Dataset (FQuAD) [5] is an example of a dataset that is not in the English language. This french dataset is similar to the SQuAD dataset; both use Wikipedia articles. There are also two versions of FQuAD. The first one contains over 25,000 samples, and the second one is bigger, with more than 60,000 samples.

Although these datasets work with text, they cannot be used directly on Document VQA, given that they do not contain a visual stand. In Document VQA, it is crucial to look at the document from its visual standpoint; different parts of documents have different meanings. With the text-only QA datasets, this information is lost.

Visual Question Answering (VQA) In the VQA datasets, we can use the same images for different languages. This is possible because we focus on the objects in the picture, not the text it contains. One of the English datasets is VQA dataset [1], which includes 204,721 images from the MS COCO dataset [10] with 614,163 questions and 7,984,119 answers. This dataset also possesses 50,000 abstract scenes with 150,000 questions and 1,950,000. The Image-Set Visual Question Answering (ISVQA) [2] dataset went one step further and focused on a multi-image setting. The dataset consists of 141,096 questions about objects and relationships in one or more images. In total ISVQA has 60,884 image sets.

The Indic Visual Question Answering dataset [4] is an example of a non-English VQA. The exciting thing about this dataset is that it focuses on three languages: Hindi, Kannada, and Tamil. This Indic dataset consists of 3.7 million image-question pairs for each of these three languages on the same set of images.

Named Entity Recognition (NER) Named Entity Recognition aims to locate and identify entities in the given text. One of the datasets for this task is the CoNLL-2003 dataset [15], which consists of two languages: English and German, and four types of named entities. The English data originate from Reuters Corpus, composed of 22,137 sentences. The German part consists of 18,933 from the ECI Multilingual Text Corpus, precisely from the German newspaper Frankfurter Rundschau. Another dataset for NER is Few-NERD [6]; this dataset is more extensive than CoNLL-2003 and consists of 188,238 sentences from Wikipedia, eight coarse-grained types, and 66 fine-grained types.

There are also datasets for legal documents, which are closer to our task. An example of one of these datasets, which is also non-English, is a Dataset of German Legal Documents for NER [9]. This dataset contains 66,723 sentences and two versions of annotations. The first version consists of 19 fine-grained semantic classes, and the second has seven coarse-grained classes.

Document Visual Question Answering datasets Only a few Document VQA datasets have been created recently, primarily in English. These datasets consist of web pages, scanned documents, or born-digital documents, and also various pages are from textbooks or posters.

Currently, the best dataset on Document VQA is DocVQA [11]. This dataset consists of several documents from the UCSF Industry Documents Library [18]. The important thing is that it also contains invoices, and all answers can be retrieved directly from the document's text. This is similar to the task of our future model; processing and analyzing invoices and developing an extractive model. As for the quality of the documents used, the dataset contains born-digital documents, scanned documents, and handwritten or typewritten documents. The reason why there are handwritten or typewritten documents is that the documents are from the period between 1960 and 2000. Overall the dataset consists of 50,000 questions over 12,767 document images, which are extracted from 6071 scanned documents. [11]Of the 50,000 questions, 36,170 of them are unique questions.

Visual Machine Reading Comprehension (VisualMRC) [16] is an example of a dataset consisting of screenshots of web pages. Another difference between this dataset and DocVQA is that it has images from different sources, which increases its diversity and usability. Altogether the VisualMRC [16] consists of 30,562 questions, where 29,419 are unique.

2.4 Models

This section will discuss some applicable models for our Document Question Answering model and the models we are using for the baseline.

Question Answering models Question Answering models can retrieve the response to a question from a text. There are two different types of models based on their answers. The first type is Extractive QA; in this type, answers are directly written in the text. The second type is Generative QA, where the answer is a free text based on the context of the text. An example of a model for QA is the BERT Base model [3], which was fine-tuned on the SQuAD dataset.

LayoutLM family models The LayoutLM family consists of three generations of multimodal Transformer models, which were pre-trained on the IIT-CDIP Test Collection containing English scanned documents. Additionally, the LayoutLM family also offers one multilingual model trained in 53 languages, including Czech and Slovak.

3 Czech Document Visual Question Answering Dataset

This section will introduce our Document Visual Question Answering dataset in the Czech language. We will focus on collecting documents, creating questions, and mapping them to desired answers.

3.1 Data Collection

Our dataset contains 6,849 documents, the vast majority of which are invoices. The documents are primarily in the Czech and Slovak languages, but it also has Polish and Slovene languages. However, there are also non-Slavic languages like English, German, and Hungarian.

For each invoice, we asked questions about 15 entities it contains. For each entity, we created several different types of questions, covering multiple variants. This also helps the model to learn to answer more than one type of question for each entity.

In Figure 2, we can see how we created our dataset. At first, we have an invoice with several entities. We have chosen the 15 most important from them, for example, IBAN, account number, total sum to be paid, or invoice number. From annotators, we have obtained the exact positions of these entities on our invoices. Next, we created the questions for these entities, which were then mapped to the bounding boxes with the correct one or multiple answers.

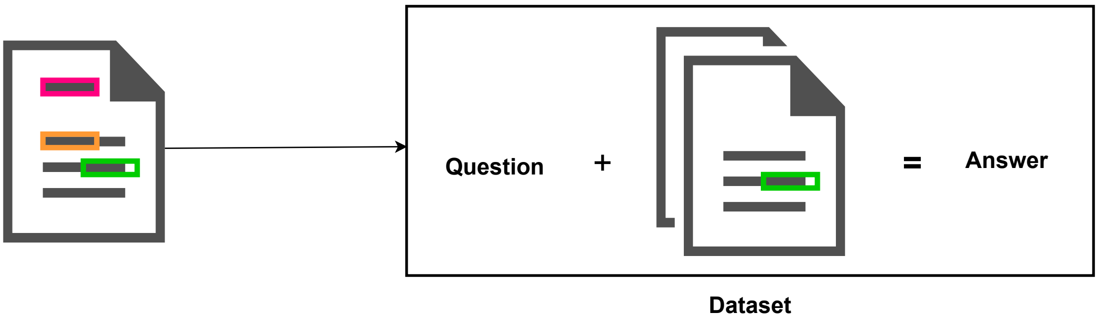

Fig. 2: Visual question answering dataset design: the green rectangle is the box that contains the information crucial to answer the question.

4 Research Proposal

This section introduces three research questions that we propose to study in our future work.

4.1 How can Textual QA and Visual QA datasets improve Document Understanding?

Our model will be trained on the Czech Document VQA dataset, which contains a large number of relevant, but *narrow* questions regarding the document's content. We hypothesize that if we first introduce our model to much more general and comprehensive QA datasets, the resulting model might be more robust in the unseen evaluation scenarios and hence, perform better as a result. Given that the textual QA datasets contain no layout, we propose to experiment with (i) a synthetic text layout and (ii) where applicable, the retrieval of the QA contexts from their original sources over the internet websites. Analogically, we would like to use the VQA dataset to see how it can improve the understanding of the document as an image. Finally, we would like to try all the relevant datasets and examine the results.

We will compare the performance of our Document VQA models to the baselines of (i) a competitive textual QA model, such as XLM-RoBERTa-Large, and (ii) the Visual Named Entity Recognition model, such as LayoutLM, trained solely on our Document VQA dataset. We will experiment with three models from the LayoutLM family: LayoutLMv2 Base, LayoutLMv2 Large, and Layout-XLM model.

4.2 How well can Document VQA models generalize beyond training entity types?

Conventional Named Entity Recognition models, for example, token classification, can identify only a closed set of entity types present in the training set and must be retrained when a new entity needs to be recognized. QA models can theoretically circumvent this limitation by having the question as a part of the input. However, our model will be trained using a closed set of questions – one or a few for each entity type.

Therefore, it remains unclear whether such a QA model will be able to answer questions that are beyond the scope of its training set. We measure how well our model can generalize to unseen entities by selecting entity types as either training or evaluation and splitting the dataset accordingly.

To provide a reference to the results, we will compare our model with a competitive text-only model for QA and a model for the Visual Named Entity Recognition model from the LayoutLM family trained on the evaluation entities.

4.3 How much can Document VQA in non-English languages benefit from English datasets?

Lastly, we will quantify the benefit of utilizing English Document VQA datasets for document understanding in other languages. In this set of experiments, we will compare the model performance on a *target language*, i.e., a language of the final application, trained using (i) English data only, (ii) using a mixture of English and the target-language data, and (iii) using solely the target-language data. Details of the experimental setup can be found in Section 4.1.

Notably, the evaluation of the approach will assess the applicability of our models to *unseen languages*, which is necessary for relevancy in a vast majority of the world languages.

Our evaluation setup will include target languages where *any* document-understanding datasets are currently available. Thanks to our dataset, these will include Czech. Also, small-scale datasets for Document NER or QA are available in French and German.

5 Conclusion

This paper offers a basic overview of systems and datasets for Document Visual Question Answering. We created the first Czech dataset for Document VQA. Last but not least, we pose three research questions outlining future work: (i) improvement of Document Understanding, (ii) generalizing Document VQA beyond training entity types, (iii) benefit of English datasets for non-English Document VQA.

Acknowledgements We acknowledge the support of grant Intelligent Back Office, project number CZ.01.1.02/0.0/0.0/21_374/0026711.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2425–2433 (Dec 2015). https://doi.org/10.1109/ICCV.2015.279

2. Bansal, A., Zhang, Y., Chellappa, R.: Visual Question Answering on Image Sets. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 51–67. Springer International Publishing, Cham (2020)
3. Chan, B., Möller, T., Pietsch, M., Soni, T.: Hugging face, https://huggingface.co/deepset/roberta-base-squad2
4. Chandrasekar, A., Shimpi, A., Naik, D.: Indic Visual Question Answering. In: 2022 IEEE International Conference on Signal Processing and Communications (SPCOM). pp. 1–5 (2022). https://doi.org/10.1109/SPCOM55316.2022.9840835
5. d'Hoffschmidt, M., Belblidia, W., Brendlé, T., Heinrich, Q., Vidal, M.: FQuAD: French Question Answering Dataset (2020). https://doi.org/10.48550/ARXIV.2002.06071
6. Ding, N., Xu, G., Chen, Y., Wang, X., Han, X., Xie, P., Zheng, H.T., Liu, Z.: Few-NERD: A Few-Shot Named Entity Recognition Dataset (2021). https://doi.org/10.48550/ARXIV.2105.07464
7. Ding, Y., Huang, Z., Wang, R., Zhang, Y., Chen, X., Ma, Y., Chung, H., Han, S.C.: V-Doc: Visual questions answers with Documents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21492–21498 (2022)
8. Klink, S., Dengel, A., Kieninger, T.: Document structure analysis based on layout and textual features. In: Proc. of International Workshop on Document Analysis Systems, DAS 2000. pp. 99–111 (2000)
9. Leitner, E., Rehm, G., Schneider, J.M.: A Dataset of German Legal Documents for Named Entity Recognition. CoRR **abs/2003.13016** (2020), https://arxiv.org/abs/2003.13016
10. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft COCO: Common Objects in Context (2014). https://doi.org/10.48550/ARXIV.1405.0312, https://arxiv.org/abs/1405.0312
11. 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 (WACV). pp. 2200–2209 (January 2021)
12. Niyogi, D., Srihari, S.N.: A Rule-Based System for Document Understanding. In: Proceedings of AAAI-86. pp. 789–793 (1986)
13. Rajpurkar, P., Jia, R., Liang, P.: Know What You Don't Know: Unanswerable Questions for SQuAD (2018). https://doi.org/10.48550/ARXIV.1806.03822
14. Reddy, S., Chen, D., Manning, C.D.: CoQA: A Conversational Question Answering Challenge (2018). https://doi.org/10.48550/ARXIV.1808.07042
15. Sang, E.F.T.K., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. CoRR (2003). https://doi.org/10.48550/ARXIV.CS/0306050
16. Tanaka, R., Nishida, K., Yoshida, S.: VisualMRC: Machine Reading Comprehension on Document Images. Proceedings of the AAAI Conference on Artificial Intelligence **35**(15), 13878–13888 (May 2021). https://doi.org/10.1609/aaai.v35i15.17635
17. Tito, R., Karatzas, D., Valveny, E.: Document collection visual question answering. In: International Conference on Document Analysis and Recognition. pp. 778–792. Springer (2021)
18. Industry documents library, https://www.industrydocuments.ucsf.edu/
19. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-Training of Text and Layout for Document Image Understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192—1200. KDD '20, ACM, New York, NY, USA (2020). https://doi.org/10.1145/3394486.3403172