Document Visual Question Answering with CIVQA

Czech Invoice Visual Question Answering Dataset

Šárka Ščavnická, Michal Štefánik, and Petr Sojka

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

Abstract. Applications of document processing become increasingly popular across multiple industries, resulting in a growing amount of research on the applications of artificial intelligence in document processing (Document AI). This paper focuses on a subtask of Document AI, Document Visual Question Answering (DVQA), recently getting well-deserved attention thanks to its universality. However, the limited availability of data sources for languages outside English restrains the applicability of DVQA in non-English languages.
For this reason, we created the CIVQA (Czech Inovice Visual Question Answering) dataset covering 15 entities of financial documents, consisting of more than 6,000 invoices in the Czech language.
We used the CIVQA dataset to create the first-of-its-kind DVQA models specifically tailored for applications to Czech documents. Striving to create DVQA models able to generalize, we specifically evaluate our models on the entities not covered in the training mix and find that multilingual LayoutLM models are able to respond to questions about previously unseen entities substantially more accurately than other models.
The CIVQA dataset and experiment observations offer new opportunities for Document AI in the Czech Republic, with potential applications in research and commercial fields.

Keywords: Question Answering, Visual Question Answering, Document Visual Question Answering, Czech Invoice Visual Question Answering Dataset

1 Introduction

Document AI is transforming how businesses and organizations process, store, and analyze vast amounts of data. [26] have stated that The Big Four accounting firms (Deloitte, Ernst & Young (EY), PricewaterhouseCoopers (PwC), and Klynveld Peat Marwick Goerdeler (KPMG)) have launched their own Document AI systems. These systems are able to automatically recognize data from Visually Rich Documents, enter invoices into the systems, and generate financial reports.

Document Visual Question Answering (DVQA) is a subgroup of Document AI. The task of question answering is usually combined with the use of the

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 23–34, 2023.

Large Language Models (LLM). Demand after these systems is most visible in the domain of office documents like invoices [19].

For creating suitable models for DVQA, it is essential to have appropriate learning data [11,21]. While several English datasets exist, no DVQA datasets exist for the Czech language. To address this gap, we have created the first dataset in the Czech language, focused on invoices. We conducted experiments to evaluate the quality and robustness of the CIVQA dataset and DVQA models trained on CIVQA. Our results demonstrate that CIVQA-trained models can to a certain extent generalize to unseen question types, and the robustness of resulting DVQA models can be further supported by using multilingual base models and mixing of CIVQA datasets with existing datasets in English.

2 Background

This section gives a brief theoretical introduction to the Visually Rich Documents and Document Visual Question Answering.

2.1 Visually Rich Documents

Visually Rich Documents (VRD) contain documents whose semantic structure is not determined only by the text but also by the layout and visual elements of the documents. These visual elements are, for example, typesetting formats, tables, and figures. Invoice is an example of VRD; its layout pieces of information are crucial for the overall understanding because they are usually split into several sectors. [9]

2.2 Document Visual Question Answering

Document AI can be divided into four groups: Document Layout Analysis, Visual Information Extraction, Document Visual Question Answering, and Document Image Classification [9]. In this paper, we will be focusing on the third part: Document Visual Question Answering.

The Question Answering (QA) systems are tools for retrieving specific information that some users have requested. One of the most significant features is that the Document Visual Question Answering systems can retrieve these pieces of data from the VRDs. [27] The usual QA systems have two inputs: the first is the question, and the second is a document or collection of multiple documents, where we search for the answer.

The question-answering systems have evolved over time. First, there were purely text-based systems; for example, these systems worked with Wikipedia articles and looked for factual answers. The BERTmodel [2], fine-tuned on the SQuAD dataset [18], is an example of a textual QA model.

Subsequently, Visual question-answering (VQA) models began to emerge. Antol et al. [1] define VQA as an artificial intelligence technology that enables a computer system to answer questions about an image. VQA combines natural

language processing, object recognition, and computer vision to interpret the content of an image and provide answers based on its understanding.

Document Visual Question Answering [10] seeks to obtain knowledge from documents through answering 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 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. Due to their ability to work with VRDs, Document VQA popularity constantly increases across different fields. For example, they can help process invoices and other documents in the financial sector.

Only a small amount of Document VQA datasets have been created recently, primarily in English. These datasets typically feature web pages, scanned documents, born-digital documents, as well as pages sourced from textbooks or posters.

Nowadays, the most popular dataset for Document VQA is DocVQA [17]. This dataset comprises several documents from the UCSF Industry Documents Library [23], which also includes invoices. The documents are either born-digital, scanned, handwritten, or typewritten from 1960 to 2000.

2.3 Models

The models from the LayoutLM family play an essential role in Document AI, mainly because during pre-training, they combine both the visual part of the document and its textual part. [9] Hence, they are improving the performance of Document AI models.

In this paper, we have focused on five different versions of the LayoutLM model family: LayoutXLM (Layout Cross-Lingual Language Model) [25], LayoutLMv2 [24], LayoutLMv3 [14], Impira QA (Impira model for Visual Question Answering) [16], and Impira Invoice (Impira for Invoices) [15]. Impira models are the finetuned versions of the LayoutLM models.

3 CIVQA Dataset

Presently, coverage of non-English models for Document Visual Question Answering is lacking. For this reason, we have created the first Czech dataset for document question-answering, called the *CIVQA dataset*.

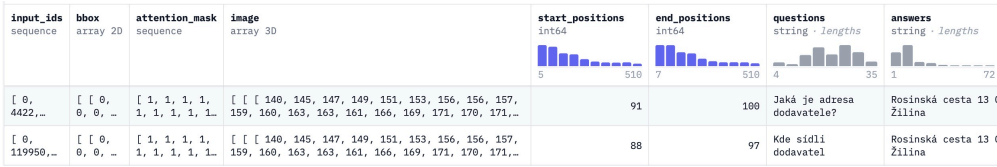
The CIVQA dataset consists of 6,849 invoices, which were obtained from public sources. Over these invoices, we focused on 15 different entities, which are crucial for processing the invoices. We included each entity in at least one of these four groups: numeric, textual, pattern, and shape. The difference between the pattern and shape is that patterns are entities like QR codes, which do not contain words or numbers but have some *visual* pattern. The shape group is for

Table 1. Entities' categories

Entity	Numeric	Textual	Pattern	Shape
Invoice number	X			
Variable symbol	X			
Specific symbol	X			
Constant symbol	X			
Bank code	X			X
Account number	X			X
ICO	X			X
Total amount	X			
Invoice date	X			X
Due date	X			X
Name of supplier		X		
IBAN	X	X		X
DIC	X	X		X
QR code			X	
Supplier's address		X		

id string · *lengths*	words sequence	answers string · *lengths*	bboxes sequence	answers_bboxes sequence	questions string · *lengths*	image string · *lengths*
420000000	["12008626", "FAKTÚRA", " (FV)", "Strana", "1",…	Rosinská cesta 13 010 08 Žilina	[[78.69774919614147,…	[[15.434083601286174…	Jaká je adresa dodavatele?	f8f55985f6a82596baa43a72543fa4e4
420000001	["12008626", "FAKTÚRA", " (FV)", "Strana", "1",…	Rosinská cesta 13 010 08 Žilina	[[78.69774919614147,…	[[15.434083601286174…	Kde sídli dodavatel	f8f55985f6a82596baa43a72543fa4e4

Fig. 1. CIVQA pre-encoded dataset



input_ids sequence	bbox array 2D	attention_mask sequence	image array 3D	start_positions int64	end_positions int64	questions string · *lengths*	answers string · *lengths*
[0, 4422,…	[[0, 0, 0, …	[1, 1, 1, 1, 1, 1, 1…	[[[140, 145, 147, 149, 151, 153, 156, 156, 157, 159, 160, 163, 163, 161, 166, 169, 171, 170, 171,…	91	100	Jaká je adresa dodavatele?	Rosinská cesta 13 0… Žilina
[0, 119950,…	[[0, 0, 0, …	[1, 1, 1, 1, 1, 1, 1…	[[[140, 145, 147, 149, 151, 153, 156, 156, 157, 159, 160, 163, 163, 161, 166, 169, 171, 170, 171,…	88	97	Kde sídli dodavatel	Rosinská cesta 13 0… Žilina

Fig. 2. CIVQA encoded dataset

entities that have some given rules and constraints like IBAN or DIC. All entities and their groups can be seen in Table 1.

The first step needed for creating the datasets was obtaining annotated invoices. The annotation was provided by a third party in the Intelligent Backoffice (IBO) project [20,13]. The annotations and OCR results were received in a JSON file, which includes several items for each invoice. These items consist of the unique identifier of the image, image path, bounding boxes of all words on the invoice obtained by OCR from the invoice, a list of words corresponding to the respective bounding boxes, and a list of labels corresponding to the fifteen entities, as shown in Table 1. We have also added a label marked "O" to the list

of labels. This label belongs to those words that, during annotation, were not assigned any entity.

This JSON is further processed to add questions and gather answers. We have created at least three questions per entity. Each entity corresponds to an answer that we are looking for. With the help of these questions, we covered an extensive range of possibilities a user can ask about an entity. Examples of questions created for the invoice number entity are: Jaké je číslo faktury? (What is the invoice number?), Pod jakým číslem je vedena faktura? (Under which number is the invoice kept?), Číslo faktury? (Invoice number?), Jaké je označení faktury? (What is the label on the invoice?).

There are two types of CIVQA datasets. The first was created using Tesseract OCR [22], and the second was created with EasyOCR [12]. At the same time, both datasets have two versions. The first one is suitable for further use and subsequent adaptation of this dataset for other models. The part of the dataset can be seen in Figure 2 on the facing page. It contains words, respective bounding boxes, image names, questions, and answers. There are no labels, as these labels may vary for different models for document visual question answering, though this dataset is ready to be encoded for future models.

Figure 2 displays the second dataset, which is ready for training on selected models. The images in this dataset are resized to *224×224*, and it is established that they have the correct order of color channels. Due to the resizing of the images, it was also necessary to recalculate all the bounding boxes. The words and bounding boxes are transformed into token-level parameters like input_ids, attention_mask, token_type_ids, and bbox. The processor adds special tokens ([CLS] and [SEP]) to separate questions from word tokens. For the chosen LayoutLM model, we also need labels; for these models, the label consists of starts and end positions, indicating which token is at the start and which token is at the end of the answer. Based on this, we can find the correct answer to our questions. The dataset also contains the textual answer, which can be used for verification if the model is predicting the correct answer. It is important to note that these datasets contain errors created by OCR retrieval of incorrect letters and mistakes made by annotators.

CIVQA datasets are available as CIVQA TesseractOCR Dataset [7], CIVQA TesseractOCR LayoutLM Dataset [8], CIVQA EasyOCR Train Dataset [5], CIVQA EasyOCR Validation Dataset [6], CIVQA EasyOCR LayoutLM Validation Dataset [4], CIVQA EasyOCR LayoutLM Train Dataset [3].

4 Experiments

In this section, we introduce experiments done with CIVQA datasets. We fine-tuned the models from Section 2.3 on CIVQA datasets in order to find out which OCR method is the best for Czech Document VQA. In the second part of the experiments, we evaluated the robustness of the resulting models on unseen types of questions.

4.1 The OCRs and the CIVQA Dataset

The objective of this experiment was to determine the impact of using different OCRs on the quality of document visual question-answering systems and find the best OCR for future experiments. The model's quality is determined by its ability to return a prompt identical to the answer from the dataset. The better the model, the better the precision it achieves. For this experiment, we focused on all 15 entities throughout all invoices. Table 2 shows the measured results for both OCR frameworks. We see that the best results were obtained by fine-tuned models on the dataset used by Tesseract OCR, which can recognize more languages than EasyOCR. Furthermore, LayoutXLM achieves the best results for both types of datasets.

Table 2. CIVQA results: comparison of Tesseract and EasyOCR frameworks by Precision, Recall, and F1 score.

	Tesseract			EasyOCR		
Model	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**
LayoutXLM	**0.7422**	**0.7117**	**0.7079**	0.6636	0.6633	0.6455
LayoutLMv2	0.6917	0.6750	0.6634	0.6323	0.6129	0.6011
LayoutLMv3	0.6989	0.6382	0.6410	0.6370	0.6164	0.6065
Impira QA	0.6773	0.6291	0.6313	0.6373	0.6015	0.5984
Impira Invoice	0.6948	0.6440	0.6434	0.6345	0.6019	0.5962

As stated, the LayoutXLM is the overall best model for the Czech document visual question-answering task. This model is unique because it was trained using a multilingual dataset consisting of these languages: Chinese, Japanese, Spanish, French, Italian, German, and Portuguese [25]. Even though it was not trained on the Czech dataset, the languages with diacritic (Spanish, French, German, Portuguese, Italian) may have helped the LayoutXLM to work better than other models (trained only on English data).

We also delved into exploring the performance of individual questions. In Figure 3 on the next page, it can be seen how many percentages of questions were answered correctly for the best model from Table 2. Individual questions are separated by color and assigned to individual entities.

The success of individual entities has been observed to be influenced by the entity groups that were defined in Table 1, as we can see in the percentage results for individual questions in Figure 3 on the next page. Entities with a specific or numerical shape were more successful than purely textual entities without a specific shape. The supplier's name and the supplier's address are the entities with one of the lowest success rates, and these are also the only purely textual entities. Below them is the QR code entity, which is neither textual nor number and is spread on multiple lines, which is a problem when predicting starting and ending positions.

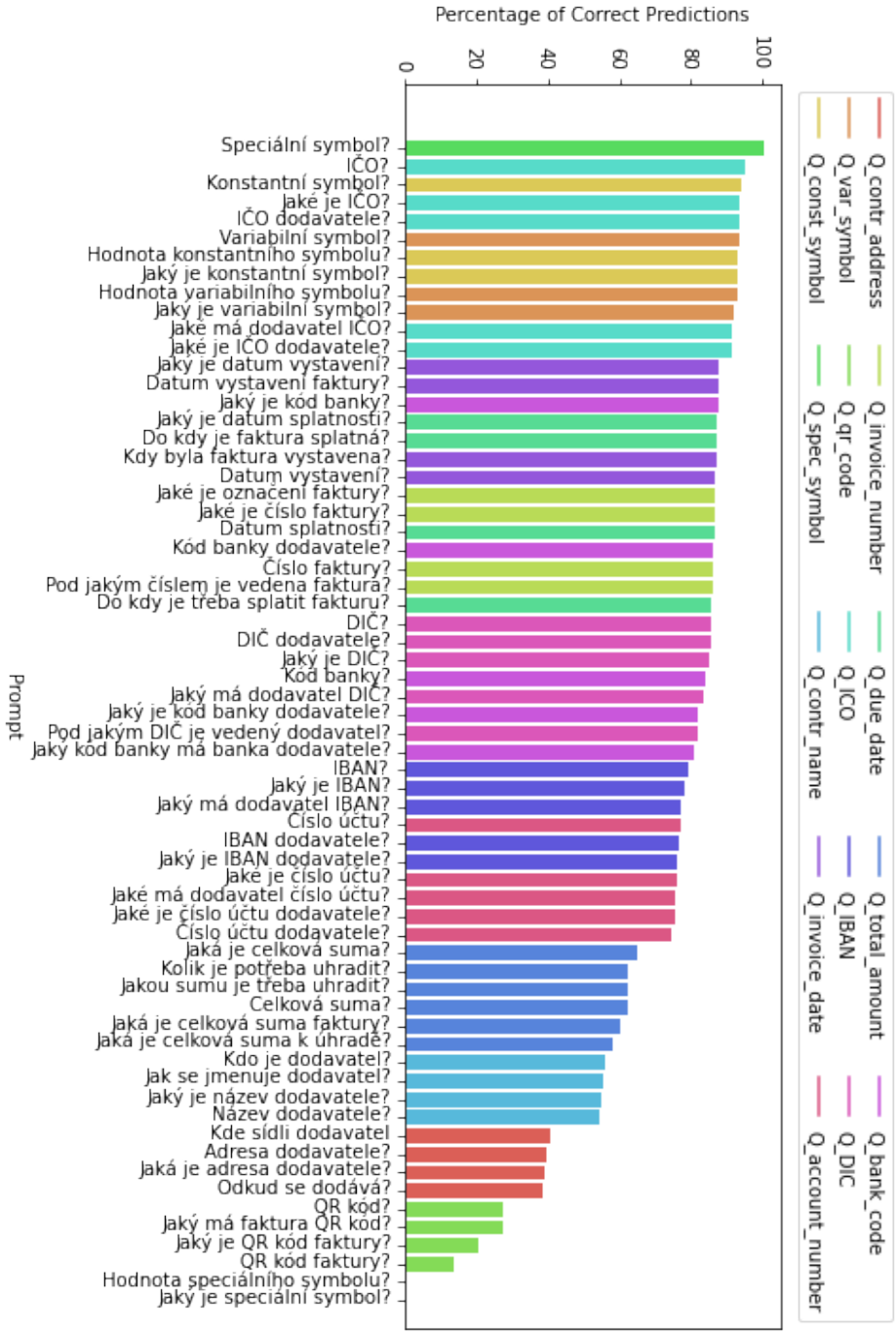


Fig. 3. Validation dataset of CIVQA Tesseract OCR: LayoutXLM model success rate by individual question percentage.

The entity with the identifier *ICO*, which is a numeric entity with a given shape, performed the best in the evaluation. Although in the first place, with 100% success, we have a question focused on the *Specific symbol* a numeric entity. However, this happened for one out of three types of questions for this entity. All other questions related to this entity achieved a 0% success rate. Based on this, we can not say this is the best-performing entity when *ICO* had good results for every question. It should be also noted that the *Specific symbol* entity was presented less than the other entities in the obtained invoices because this is not that much used on the invoices.

4.2 CIVQA and unseen types of questions

In this set of experiments, our focus was on developing a practical and robust solution for unseen entities. We would like to create a model that could be used on new entities, and in that case, it could be more beneficial for users. For this task, we have separated the CIVQA_TesseractOCR dataset into two datasets. One is for unseen entities with five entities, and the other is for known entities. We choose these five entities (invoice number, ICO, supplier's address, IBAN, due data), in the way we would cover the most different types of entities, based on the Table 1 on page 26.

In the following subsections, we will present various experiments where we observed how the models behave on unknown entities. Initially, we trained individual models from Section 2.3 on a new dataset of ten known entities and verified their success on unknown entities. Subsequently, we tried to improve their success with various attempts.

Table 3. CIVQA results: comparison of models when handling unknown entities

	Baseline			Known data			DocVQA + Known data		
Model	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**
LayoutXLM	0	0	0	0.1920	0.0413	0.0582	0.3731	0.2163	0.2465
LayoutLMv2	0	0	0	0.0343	0.0270	0.0261	0.0665	0.0334	0.0279
LayoutLMv3	0	0	0	0.1022	0.0341	0.0456	0.1504	0.0455	0.0611
Impira QA	0	0	0	0.1512	0.0455	0.0652	0.2326	0.0895	0.1148
Impira Invoice	0	0	0	0.1360	0.0530	0.0724	0.2226	0.0807	0.1063

In Table 3, we have presented the results of our experiments on unknown entities. The first column (baseline) contains the results for each model without training them on known values. Neither model was successful and failed to correctly predict any entity, resulting in a precision, recall, and F1-score of 0.

The second column (known data) represents the results of the models, which were finetuned with the dataset of ten known entities. This is the first introduction of the Czech language to these models, and we can see that the models were now more successful with predicting some entities.

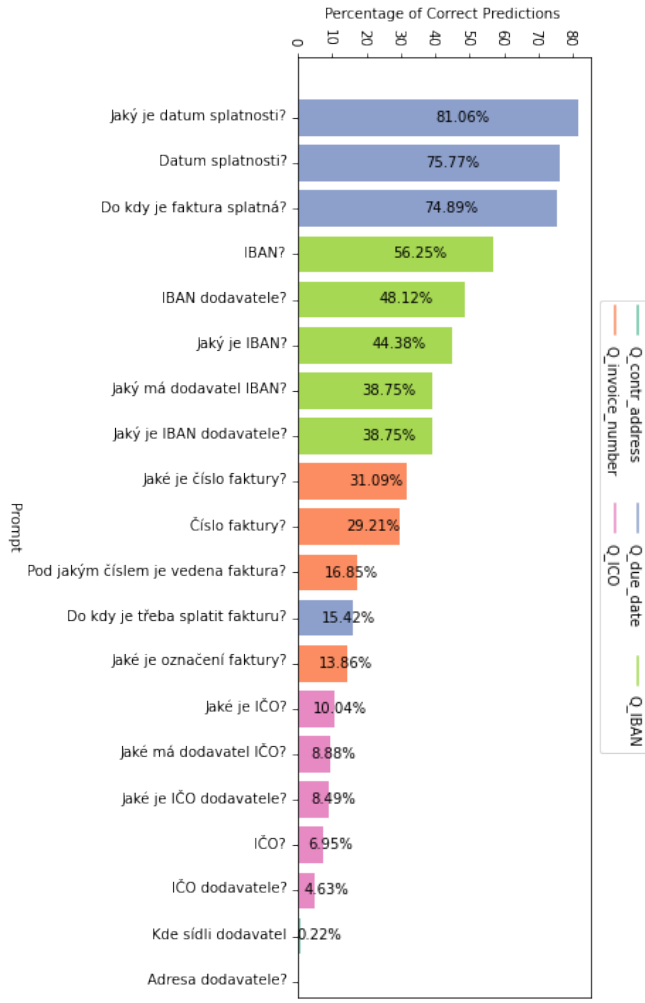


Fig. 4. Validation dataset of CIVQA unknown entities: LayoutXLM model success rate by individual question percentage fine-tuned on DocVQA plus CIVQA known dataset.

The third column (DocVQA + known data) shows the outcomes obtained from fine-tuning models on a dataset of known entities concatenated with the DocVQA dataset, which is currently one of the best datasets for DVQA. We can see that even though the DocVQA only consists of invoices in English, it has improved the score on unknown entities. The LayoutXLM model consistently achieved the best results.

For the best model, we have created a graph of the success of individual questions, where we will take a closer look at the success of each entity. This graph can be seen in Figure 4. In this case, entities with some clear structure and

shape are also more successful. The entity known as 'due date' (a numeric entity with a specific format) has achieved first place with an outstanding success rate. This entity is also similar to the invoice date entity, which is in the known entities dataset. Based on this, we can claim that if there is a similar type of entity to the unknown in the dataset of known values, the model will correctly predict this unknown but similar entity.

4.3 Effect of introduction of a small number of unknown entities

In this experiment, we have tried introducing a small amount of unknown data to the trained models. We choose 5% and compare these results with the results obtained without the introduction of unknown data in order to see how it affects the models.

In Table 4, we compare results obtained from various models. The first column results were not exposed to any unknown data, though they were fine-tuned on known data. We then fine-tuned these models with 5% of unknown data and compared the results, which can be seen in the third column. Lastly, we will compare the models from the first column, but it was further fine-tuned on the known dataset concatenated with 5% of unknown data. This experiment aimed to evaluate how different amounts of data, even those already known to the models, could improve the overall results.

According to Table 4, we can see that no other model has not surpassed LayoutXLM. Fine-tuning with a small part of unknown entities showed noticeable improvement in all models. However, after using the concatenated dataset of known entities with 5% of unknown entities, LayoutXLM did not obtain as much of an improvement as other models.

Table 4. CIVQA results: Comparing results on baseline models, then models trained on a 5% subset of unknown entities and then models fine-tuned on the concatenation of known dataset with a subset of 5% unknown entities.

	Known data			**5% of unknown**			**Known + 5% unknown**		
Model	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**
LayoutXLM	0.1920	0.0413	0.0582	0.7002	0.6594	0.6617	**0.7069**	**0.6693**	**0.6700**
LayoutLMv2	0.0343	0.0270	0.0261	0.5944	0.5154	0.5192	0.6223	0.5726	0.5755
LayoutLMv3	0.1022	0.0341	0.0456	0.5793	0.5125	0.5254	0.6344	0.5528	0.5631
Impira QA	0.1512	0.0455	0.0652	0.6186	0.5356	0.5466	0.6318	0.5487	0.5670
Impira Invoice	0.1360	0.0530	0.0724	0.5999	0.5255	0.5369	0.6353	0.5577	0.5681

5 Conclusion

This paper introduces the CIVQA datasets, which are opening new doors in the field of Document VQA in the Czech language. We have discovered that

numeric answers obtained better results than purely textual ones. Furthermore, we have shown that combining the CIVQA dataset with another DVQA dataset can improve the robustness of DVQA on unseen entities.

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: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2425–2433 (2015). https://doi.org/10.1109/ICCV.2015.279
2. Chan, B., Möller, T., Pietsch, M., Soni, T.: Hugging Face (2022), https://huggingface.co/deepset/roberta-base-squad2
3. CIVQA EasyOCR LayoutLM Train Dataset, https://huggingface.co/datasets/fimu-docproc-research/CIVQA_EasyOCR_LayoutLM_Train, accessed 2023-11-21
4. CIVQA EasyOCR LayoutLM Validation Dataset, https://huggingface.co/datasets/fimu-docproc-research/CIVQA_EasyOCR_LayoutLM_Validation, accessed 2023-11-21
5. CIVQA EasyOCR Train Dataset, https://huggingface.co/datasets/fimu-docproc-research/CIVQA_EasyOCR_Train, accessed 2023-11-21
6. CIVQA EasyOCR Validation Dataset, https://huggingface.co/datasets/fimu-docproc-research/CIVQA_EasyOCR_Validation, accessed 2023-11-21
7. CIVQA TesseractOCR Dataset, https://huggingface.co/datasets/fimu-docproc-research/CIVQA-TesseractOCR, accessed 2023-11-21
8. CIVQA TesseractOCR LayoutLM Dataset, https://huggingface.co/datasets/fimu-docproc-research/CIVQA-TesseractOCR-LayoutLM, accessed 2023-11-21
9. Cui, L., Xu, Y., Lv, T., Wei, F.: Document AI: Benchmarks, Models and Applications (2021). https://doi.org/10.48550/arXiv.2111.08609
10. 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. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.02083
11. DocILE 2023: ICDAR 2023 Competition and CLEF 2023 Lab on Document Information Localization and Extraction (2023), https://docile.rossum.ai/, accessed 2023-10-31
12. EasyOCR, https://github.com/JaidedAI/EasyOCR, accessed 2023-08-10
13. Geletka, M., Bankovič, M., Meluš, D., Ščavnická, Š., Štefánik, M., Sojka, P.: Information Extraction from Business Documents: A Case Study. In: Horák, A., Rychlý, P., Rambousek, A. (eds.) Proceedings of the 16th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2022, Karlova Studánka, Czech Republic, December 9–11, 2022. pp. 35–46. Tribun EU (2022), https://nlp.fi.muni.cz/raslan/2022/paper18.pdf
14. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: LayoutLMv3: Pre-training for Document AI with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4083–4091 (2022)
15. LayoutLM for Invoices, https://huggingface.co/impira/layoutlm-invoices, accessed 2023-10-25

16. LayoutLM for Visual Question Answering, https://huggingface.co/impira/layoutlm-document-qa, accessed 2023-10-25
17. 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 (Jan 2021)
18. Rajpurkar, P., Jia, R., Liang, P.: Know What You Don't Know: Unanswerable Questions for SQuAD. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the ACL (Volume 2: Short Papers). pp. 784–789. ACL, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-2124
19. Rossum raises record $100 million Series A from General Catalyst to reinvent B2B document communication (2021), https://rossum.ai/blog/rossum-raises-record-100-million-series-a-from-general-catalyst-to-reinvent-b2b-document-communication/, accessed 2023-10-31
20. Ščavnická, Š., Štefánik, M., Kadlčík, M., Geletka, M., Sojka, P.: Towards General Document Understanding through Question Answering. In: Horák, A., Rychlý, P., Rambousek, A. (eds.) Proceedings of the 16th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2022, Karlova Studánka, Czech Republic, December 9-11, 2022. pp. 181–188. Tribun EU (2022), https://nlp.fi.muni.cz/raslan/2022/paper17.pdf
21. Šimsa, Š., Šulc, M., Uřičář, M., Patel, Y., Hamdi, A., Kocián, M., Skalický, M., Matas, J., Doucet, A., Coustaty, M., Karatzas, D.: DocILE Benchmark for Document Information Localization and Extraction (2023), https://arxiv.org/abs/2302.05658
22. Tesseract Open Source OCR Engine (main repository), https://github.com/tesseract-ocr/tesseract, accessed 2023-10-08
23. UCSF Industry Documents Library, https://www.industrydocuments.ucsf.edu/
24. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on NLP (Volume 1: Long Papers). pp. 2579–2591. ACL, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.201
25. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding (2021). https://doi.org/10.48550/arXiv.2104.08836
26. Zhang, Y., Xiong, F., Xie, Y., Fan, X., Gu, H.: The Impact of Artificial Intelligence and Blockchain on the Accounting Profession. IEEE Access **8**, 110461–110477 (2020). https://doi.org/10.1109/ACCESS.2020.3000505
27. Zitouni, I.: Natural Language Processing of Semitic Languages. Springer (2014). https://doi.org/10.1007/978-3-642-45358-8