Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

Receipt-AVQA-2023 Challenge

Artur Begaev
Budapest, Hungary
artur.begaev@aol.com

Evgeny Orlov
Budapest, Hungary
eugene.a.orlov@gmail.com

Abstract

In this work, we introduce a new challenging Document VQA dataset, named Receipt AVQA, and present the results of the associated RECEIPT-AVQA-2023 shared task. Receipt AVQA is comprised of 21835 questions in English over 1957 receipt images. The receipts contain a lot of numbers, which means discrete reasoning capability is required to answer the questions. The associated shared task has attracted 4 teams that have managed to beat an extractive VQA baseline in the final phase of the competition. We hope that the published dataset and promising results of the contestants will inspire further research on understanding documents in scenarios that require discrete reasoning.

Keywords: VQA, computer vision, multimodal dataset
DOI: 10.28995/2075-7182-2023-22-1-11

Соревнование Receipt-AVQA-2023

Артур Бегаев
Будапешт, Венгрия
artur.begaev@aol.com

Евгений Орлов
Будапешт, Венгрия
eugene.a.orlov@gmail.com

Аннотация

В данной работе мы представляем новый датасет для задачи VQA, названный Receipt AVQA, и результаты проведенного соревнования RECEIPT-AVQA-2023. Датасет Receipt AVQA состоит из 21835 вопросов на английском языке к 1957 изображениям товарных чеков. Товарные чеки содержат большое количество числовой информации, что требует опредленной степени дискретного мышления для ответа на вопросы. Сопровождающее датасет соревнование привлекло 4 команды, которые смогли улучшить результаты по сравнению с базовой экстрактивной VQA моделью. Мы надеемся, что опубликованный датасет и многообещающие результаты участников соревнования вдохновят дальнейшие исследования в области автоматического понимания изображений документов в сценариях, где требуется дискретное мышление.

Ключевые слова: VQA, компьютерное зрение, мультимодальный датасет

1 Introduction

Receipt understanding is an important problem, which has to be solved in many applications. For example, customers want to analyze the prices of positions and total paid money, extract information about quantities of products, and how individuals could adjust their budget as so to purchase the needed amount of goods. In addition, most people would like to ask questions about these properties without the usage of any APIs or complicated computational programs. There are no such existing datasets, which cover the issue of answering the natural language questions over a receipt. In this paper, we propose such a dataset based on SROIE (Huang et al., 2019) and CORD (Park et al., 2019) receipt datasets, which is expected to cover and reveal problems with existing solutions to the question answering tasks.

VQA (Antol et al., 2015) is quite a novel task in the machine learning domain. In order to resolve such an issue a solution should combine approaches from both computer vision (process an image) and

natural language processing domains (process a question). However, most of the existing datasets focus on real-life photos; documents in general, such as invoices, industry documents, food and nutrition-related collections; and screenshots (Patadia et al., 2021). These datasets provide a markup of layouts for images or bounds of the objects and relations between the objects in an image. Meanwhile, the questions in these datasets are mostly formulated in an extractive manner, meaning that required answers are already presented on an image.

There is a special subset of such datasets – TextVQA (Singh et al., 2019). The markup for this subset also contains the recognized text (OCR Tokens) from the scene on an image. TextVQA datasets require some reasoning about the placing and the nature of OCR tokens. However, the extraction of such answers doesn't involve complex mathematical reasoning or calculations.

In this challenge, we present a new dataset: Receipt AVQA Dataset, comprising 21837 questions over 1957 images. By introducing this dataset we want to stress an important problem in the TextVQA task: extracting and calculating answers from the data presented on the image of a receipt. This task is not so trivial, because most of the current state-of-the-art methods for the TextVQA problem mostly focus on the extraction of the answers without any calculations using extracted tokens.

VQA tasks commonly contain questions about the objects in an image or the relations between them. In the TextVQA datasets questions about text tokens are offered. Solutions are required to extract the requested tokens from an image following the rules defined in a question. Our dataset also offers a new challenge – a solving model has to make calculations and aggregations over the extracted text tokens when requested in a question. We used the receipts from two datasets, in which the scales of numbers are different. Expected solutions should take into account this variation in the scaling and produce a required numerical answer.

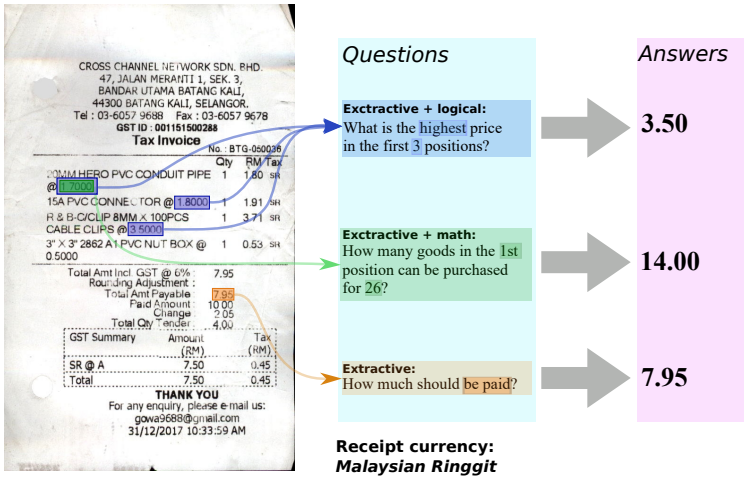

Figure 1: An example of a receipt with questions from the Receipt AVQA dataset

Moreover, our dataset introduces several types of questions like mathematical, logical, and finding the ratio between some values. The complete scheme for one receipt and some questions for that is represented in Figure 1.

The main contributions of this work can be summarized as follows:

- We introduce Receipt AVQA, a dataset of 1957 receipt images, over which we have defined 21837 questions and answers (§3) [1];
- We introduce a baseline solution for the shared task (§5);
- We conduct an analysis of the received submissions for both sub-tasks (§6) and discuss potential research directions (§7);
- We set up the shared task environment, which remains open for community submissions to facilitate future research in the area (§4.2)[2].

[1] https://github.com/dialogue-evaluation/Receipt-AVQA-2023
[2] https://codalab.lisn.upsaclay.fr/competitions/11087

2 Related Work

ST-VQA (Biten et al., 2019) and TextVQA (Singh et al., 2019) datasets extend VQA over natural images to a new direction where understanding scene text on the images is necessary to answer the questions.

OCR-VQA (Mishra et al., 2019) introduces a task similar to ST-VQA and TextVQA, but instead of natural images, images of book covers are used. Template questions are generated from book metadata such as author name, title, and other information.

Related to OCR-VQA, DVQA (Kafle et al., 2018), FigureQA (Kahou et al., 2017), and LEAF-QA (Chaudhry et al., 2019) are VQA datasets that operate on various chart images, InfographicVQA (Mathew et al., 2021) introduces VQA dataset of infographics images.

DocVQA (Mathew et al., 2020) dataset shifts the image domain to documents completely. DocVQA is a VQA dataset that is comprised of the document images of industry/business documents, and questions requiring understanding document elements such as text passages, forms, and tables. Similarly to most aforementioned VQA datasets, DocVQA is focused on *extractive questions*, where answers can always be extracted verbatim from a text on the images.

To our best knowledge, there are few VQA datasets operating on documents that are focused on the *abstract questions*, where answers cannot be directly extracted from text in the images or questions.

FiqureQA is comprised of abstract questions, but an answer to any question in the dataset is limited to yes/no.

InfographicVQA contains some questions that require certain discrete operations resulting in numerical non-extractive answers. These discrete operations are limited to counting and sorting and are employed only in 20% of questions in the dataset.

VisualMRC (Tanaka et al., 2021) dataset is built for the abstract question answering. The dataset employs the screenshots of web pages and questions that don't involve operating on numericals for answers.

Most similar to Receipt-AVQA dataset is TAT-DQA (Zhu et al., 2021) dataset. TAT-DQA comprises the high-quality images of financial reports. In order to answer the questions in the dataset a wide range of operations on numericals is required. The distinct feature of Receipt-AVQA stems from the document type employed. Receipt images on average contain fewer text tokens than financial reports but a higher quantity and relative share of numerical tokens on the image.

In Table 1 we present a high-level summary of Document VQA datasets related to ours.

Dataset	Images	Synthetic Images	Template Questions	# Images	# Questions	Answer type
DocVQA	Industry documents	No	No	12K	50K	Ex
FigureQA	Charts	Yes	Yes	120K	1.5M	Y/N
InfographicVQA	Infographics	No	No	5.4K	30K	Ex, Nm
VisualMRC	Webpage screenshots	No	No	10K	30K	Ab
TAT-DQA	Financial reports	No	No	3K	16.5K	Nm, Ex
Receipt-AVQA	Receipts	No	Yes	2K	21.8K	Nm, Ex

Table 1: Summary of Document VQA datasets. Answer type abbreviations are: Extractive: Ex, Abstractive: Ab, Yes/No: Y/N, and Numerical (the answer is numerical and not extracted from image or question; but derived): Nm.

3 Dataset and Shared Task

In this section, we present the definition of the RECEIPT-AVQA-2023 task, the construction of the Receipt AVQA dataset, and the statistical analysis of the dataset.

3.1 Task Definition

The challenge has a focus on question answering for receipts. Two tracks are offered:

1. VQA Track - question answering over the images of receipts.

2. QA Track - in this track in addition to receipt images the solutions can use the ground-truth text tokens with their corresponding coordinates extracted from the images (essentially, error-less OCR output).

The competition is formulated as a regression task. Unlike the vast majority of VQA tasks where the answer is a string, each answer is a float number here, and we use a metric operating on numbers to score submitted solutions.

To come up with a correct answer, the VQA model needs not only to recognize and extract tokens from the receipt image but to apply a number of operations (e.g. sorting, counting, arithmetic operations) over it. As a result, discrete reasoning capability is required from a potential solution.

All the questions are in English and can be divided into several types. Answering each question can be done independently for each type, however, we expected from participants to build a solution to answer questions in an end-to-end manner.

In our challenge we propose the following types of questions:

- Amount – finding and extracting required values, sometimes with some aggregation: "How much should be paid?", "What is the average price of a position?"
- Count – counting or extracting the number of elements of some type (eg. positions, a changed amount of goods): "How many positions were bought?", "How many goods are in the 1st position?"
- Ratio – finding a ratio between required values: "What is net total and total amount ratio?", "What share of cash was returned as change?"

To make our task easier, we offer a list of operations for each question. Explicit formulas are not provided. Types of proposed operations are the following: division, sorting, subtraction, summation, counting, and multiplication. A question can contain zero or several operations.

The values of prices are scaled differently because we made our dataset based on SROIE and CORD. These datasets contain receipts from different countries. We introduced the currencies of receipts as follows: "Malaysian ringgit", "Indonesian rupiah".

We split our dataset into train, validation, and test subsets in the following manner:

- Train: 16611 questions over 1537 images
- Validation: 2302 questions over 210 images
- Test: 2924 questions over 210 images

The splits in Receipt AVQA are consistent with the splits in SROIE and CORD datasets.

3.2 Dataset construction and verification

Original data from SROIE and CORD is not labeled for the VQA task: it doesn't contain any class labels for the fields on the receipt. We developed a method that allowed us to introduce the necessary labels in a semi-automatic way.

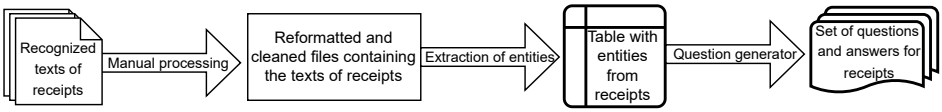

Figure 2: The scheme of data processing and question generation

We used recognized texts with bounding boxes made by authors of SROIE and CORD. In the first step all unnecessary data, like names of the shops, their addresses, and telephone numbers, was cleaned from the files. Then these files were used to extract the needed entities for our dataset: positions with prices and amounts, key-value pairs, containing, information about paid amounts, discount amounts and etc.

We used special heuristics for automatic data processing. After that, the information extracted for each receipt was reviewed and cleaned in order to contain valid data. SROIE and CORD were processed independently. So, a special table format was developed to merge the entities from both datasets. This table is intermediate and is not available for challengers.

The compilation of such a table made us available to generate the questions for receipts. We made up 13 types of questions for the positions in receipts and 9 types of questions for a total section of each

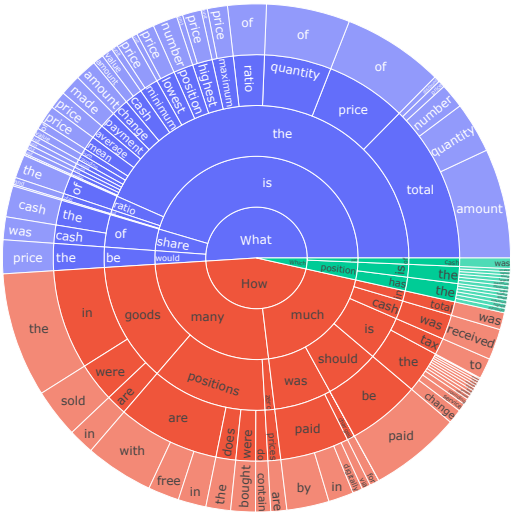

Figure 3: The word map of questions for the first five words.

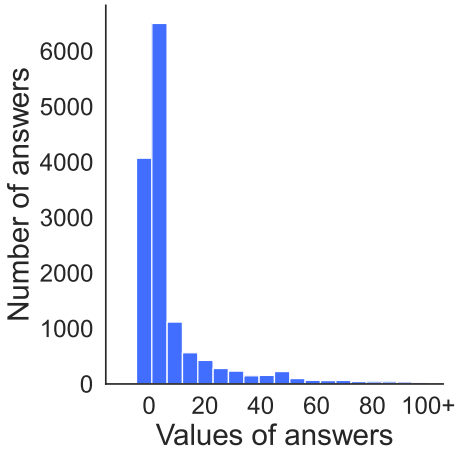

Figure 4: The distribution of answers lower than 100

receipt. For each kind of questions, several templates were prepared (from 3-5, depending on the type of question). Each template could be varied by several parameters: the numbers inside a question, the type of aggregation, and a type of value that is indented to be extracted. Some questions require calculations to get an answer. We obtained the answers explicitly by defining the needed formulas. However, applied formulas and functions are not provided in our dataset so the difficulty of provided tasks was increased.

Questions and answers have to be validated. Markup, entity extraction, and heuristics could introduce mistakes in the reasoning of getting the answers. Unfortunately, most part of the generated questions cannot be validated automatically. Still, the extractive questions for the total sections can be validated by calculating the prices of positions and summing them up. Other questions were validated by a human in several iterations.

Firstly, we prepared a large subset of our questions for the detection of mistakes in the extraction of answers. Answers and questions were validated manually by looking at a sample and checking the answers provided by the generative algorithm. This step revealed some issues with entity extraction and question generation. Some templates were reworked due to inappropriate or invalid formulations. After fixing the errors smaller subset was extracted. The same approach was applied to the new subset. We continued such iterations until there were no revealed mistakes in the dataset.

3.3 Dataset statistics and analysis

Hereby we provide some statistics over generated questions and answers. Challengers were awaited to do this by themselves in order to reveal hidden problems, which their models were intended to deal with.

In Figure 3 it can be seen that most of the questions start with the words "What" and "How" and the smallest part starts with the word "Which". There are a lot of questions that require getting some value by its extraction or calculation.

A hidden issue in our dataset was introduced – unbalanced data. Figure 4 shows the distribution of the values of answers lower than 100. Answers over 100 were omitted for more reasonable visualization. This distribution has a heavy tail – there are a lot of questions with an answer lower than 10 and many of them have an answer between 0 and 1. Extractive models should take into account this problem and resolve it properly.

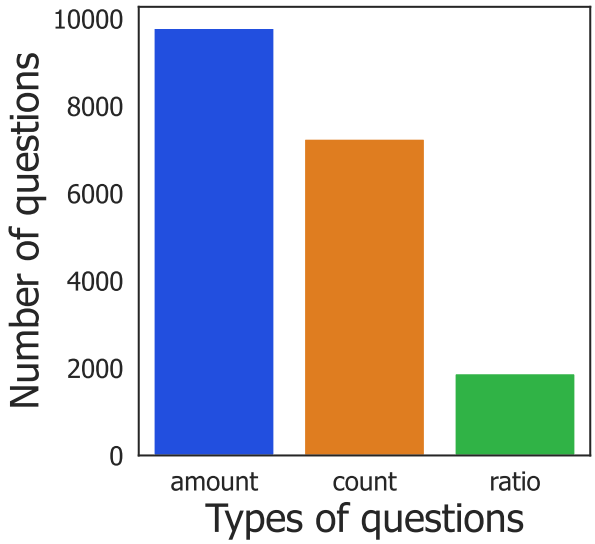

Figure 5: The distribution of questions by types

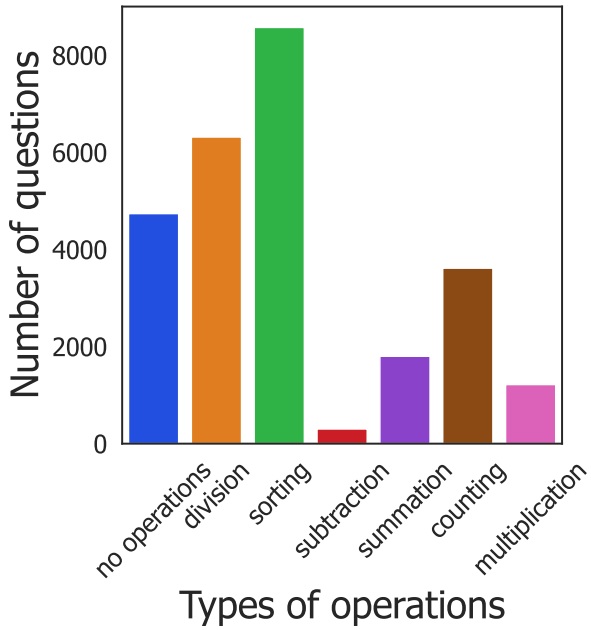

Figure 6: The distribution of questions by operations

Figure 5 provides an intuition on the number of questions from every proposed category. Most of the questions require answering counting questions and questions that ask for a specified amount of some kind.

4 Evaluation

In this section, we discuss our primary evaluation metric for the shared task and organization of the evaluation process.

4.1 Evaluation metric

Unlike traditional VQA tasks, where answers to questions are usually text, all answers in the Receipt AVQA dataset are real numbers. Consequently, both tracks in the RECEIPT-AVQA-2023 shared task are treated as a regression problems.

Mean absolute scaled error (MASE) (Hyndman and Athanasopoulos, 2013) metric is adopted as the primary metric for the shared task.

$$MASE = \frac{\frac{1}{N}\sum_{n=1}^{N}|QT_n - PR_n|}{\frac{1}{N}\sum_{n=1}^{N}|QT_n - \overline{QT_{train}}|}, \quad (1)$$

where N is a number of questions, QT_n is a real answer for the n-th question, PR_n is a predicted answer for the n-th question, and $\overline{QT_{train}}$ is an average of real answers on the train split of the dataset.

MASE metric was selected due to scale invariance, symmetry, and predictable behavior near 0.

One possible interpretation of the metric would be the ratio of the mean absolute error of a model to the mean absolute error of the baseline model, which is merely an average of the answers for the training dataset.

To get a better baseline model evaluation, questions are divided into 6 groups based on the receipt currency ("Indonesian rupiah", "Malaysian ringgit") and question type ("amount", "count", "ratio").

The 6 MASE values are then averaged, so the weights of all question types were equal.

In addition to MASE metric, we also use a more traditional accuracy metric, but with 10% leeway to account for possible rounding errors ($Acc_{\pm 10\%}$ metric).

4.2 Evaluation platform

We use the CodaLab (Pavao et al., 2022) competition platform to run the shared task.

Participation is allowed on either individual or team basis in both sub-tasks. The shared task consists of two stages: public and private testing. The first stage provides access to the public validation set and the leaderboard, allowing the participants to develop and improve their submissions during the competition. The second stage defines the final leaderboard ranking on the private test set, scoring up to twelve submissions selected by the participants. Participants are allowed to use any additional materials and pre-trained models, except for direct markup of the test set and looking for answers on the Internet.

5 Baseline

In this section, we describe the baselines we evaluated on the Receipt AVQA. These include heuristic baselines and upper bounds, and a document understanding model that was adopted as a baseline for the shared task.

5.1 Heuristics and Upper bounds

Heuristic baselines and upper bounds we evaluate are similar to the ones evaluated in other VQA benchmarks like DocVQA, and InfographicVQA.

Heuristic Baselines. The following heuristics were evaluated.

- *Random OCR number* measures performance when a random number from OCR results for the receipt image is picked as the answer;
- *Majority answer* measures performance when the most frequent answer in the train split is considered as the answer.

Upper Bounds. We also compute the following upper bounds:

- *Vocab UB* measures the upper bound on performance if the answer is predicted correctly, provided it is in the vocabulary of most common answers (> 1) of the train split;
- *OCR UB* measures the upper bound on performance if the answer is predicted correctly, provided it is one of the text tokens present on the corresponding receipt;
- *Vocab + OCR UB* measures the upper bound on performance if the answer satisfies either *Vocab UB* or *OCR UB*.

In the calculation of upper bounds, if the correct answer is not found within the defined scope, the mean of the corresponding question type answers on the train set is used instead.

The results of heuristic baselines and upper bounds are shown in Table 2.

Baseline	$MASE$ **val**	$MASE$ **test**	$Acc_{\pm 10\%}$ **val**	$Acc_{\pm 10\%}$ **test**
Random OCR number	1e11	3e11	6.17%	5.81%
Majority answer	0.8576	0.9662	18.64%	17.44%
Vocab UB	0.4420	0.5277	82.54%	81.67%
OCR UB	0.8216	0.8175	51.56%	47.02%
Vocab + OCR UB	0.4184	0.4812	88.44%	87.10%

Table 2: Results of heuristics and upper bounds.

5.2 VQA Baseline

It is not obvious whether SOTA methods for the TextVQA task, such as LayoutLMv3 (Huang et al., 2022), would be enough for solving the proposed tasks without any heavy modifications. We expect a model that can do some reasoning and calculations on extracted tokens. In this work, we want to present LayoutLMv3 as the baseline for the challengers to beat by introducing novel modules and methods for finding the relations between questions and texts from a receipt. We intentionally focus on the extraction questions in order to find out competitors who had beaten our solution.

LayoutLMv3 is a visual transformer that accepts bounding boxes, an image, and a tokenized text. We use pre-trained LayoutLMv3 from Hugging Face (Hug,) with LayoutLMv3TokenizerFast as a tokenizer. Still, it is not suitable to use for our dataset without some modifications.

Data preprocessing. LayoutLMv3 requires the recognized text from an image, we use PaddleOCR (Pad,) for the text extraction with default settings. Our task was defined as regressive, so it is expected to extract numbers, not strings, both from images and questions. Still, most of the TextVQA models treat answers as strings. Solutions for the TextVQA datasets extract the answers by using OCR tokens. We made a simple lookup algorithm for finding an answer in OCR tokens. This algorithm matches the decimal, the whole and fractional parts of a number. We treat an OCR token as matching to the answer when the absolute error between the token converted to a float and the answer itself equals zero. Perhaps, the OCR algorithm is not perfect and introduces some errors which our algorithms fail to resolve. Unresolved answers are encoded by a special token, which should be extracted by the model if the answer wasn't found.

Model training. A token prediction head was added on top of the LayoutLMv3. This head consists of 2 fully-connected layers with LeakyReLU between and Dropout before and after the first layer. The first layer outputs 768 features, and the second outputs 2 features - predicting is a token the answer or not. For the optimization, we used Adam with $learning\_rate = 0.00001$. We trained our model for 50 epochs, but the best result (by loss) was achieved on the 9th epoch. In order to train the model to select an appropriate token we use cross-entropy loss.

Answer extraction. The answer decoding is not as straightforward as could be expected. The tokenizer splits a string by punctuation marks and spaces, although we try to extract only one OCR token without finding spans. These factors introduce some complications to our algorithm for the extraction of answers. So, the output from our model represents a sequence from 2-dimensional vectors. Because we trained our model for the binary classification, we can extract the needed tokens after the tokenization by finding all positions in the sequence where the number in the last dimension is greater than 0.5. We concatenate these tokens to get a complete token representing the extracted answer. However, when the model fails to get the answer our post-processing outputs a special token. In order to get an answer to such questions we had pre-calculated means for each type of question and currency. When this unwanted situation occurs we put the mean defined by the type of question as the answer.

Results. Such a simple approach doesn't provide good results. Especially for the questions requiring calculations or aggregations.

$MASE$ **Total**	$MASE$ **Amount**	$MASE$ **Count**	$MASE$ **Ratio**	$Acc_{\pm 10\%}$
0.8786	0.8068	0.8291	1.0000	14.60%

Table 3: The values of metrics produced by our LayoutLMv3 model.

As it can be seen in Table 3 our model completely fails in the Ratio questions and doesn't provide precise results for the Amount and Count questions. This leads to the conclusion that a good model, which can properly solve our task, should inherit some architectural and methodical properties from the state-of-the-art TextVQA models. Thus, further modifications, such as computational trees over the extracted tokens or specific rules for token processing, are required.

6 Submitted solutions

The final phase of the Receipt-AVQA-2023 shared task attracted 5 participants. We provide brief descriptions of the 3 solutions, which outperformed the baseline for at least one track. We denote each team by their CodaLab user names. In case of multiple submissions from one team, we report only the best result. The scores of the teams are shown in Table 4.

surkov_evseev The solution of the team is based on two core pipelines, one for extraction and structuring of the information from receipt images and another for translating each question into a mathematical expression. To extract the textual information from an image fine-tuned PP-OCRv3 (Li et al., 2022) /

TrOCR (Li et al., 2021) pipeline is employed. The extracted text data is then tagged via finetuned BERT (Devlin et al., 2019) models to establish a standardized structure for the receipt. A finetuned T5 (Raffel et al., 2019) model is employed for parsing question text into a mathematical expression required for the answer calculation. The final answer is then derived based on the structured receipt data and the mathematical expression for the question. The team had to build auxiliary markup using custom rules to train their BERT and T5 models.

s231644 The team manually converted all questions types into mathematical expressions, and built a multimodal UDOP (Tang et al., 2022) model to predict sequence of operands and operands types for these mathematical expressions. If the output of the UDOP model is inconsistent with the mathematical expression for the question, a fallback T5 model is employed instead, which tries to answer the question directly without parsing it into a mathematical expression. The sequences in models operate on character level. The models require text tokens from the image as an input; these text tokens are extracted via custom OCR pipeline. The team had to annotate the dataset to create custom labels for the UDOP model.

daniyallaiev The participant created a separate solution for each question type in the dataset. For the *ratio* question type a finetuned multimodal LayoutLMv3 model outputs the numerator and the denominator tokens of the answer. For the *amount* question type a multimodal LayoutLMv3 model is trained to output the best suitable token for the answer. LayoutLMv3 models work in extractive fashion by trying to select the most appropriate tokens for the answer among all OCR'ed tokens available to the models as input. The pipeline for the *count* question type is different. Firstly, a BERT model is used to extract key tokens from the question text that are required to answer the question. If no token is found, a fallback ViT (Dosovitskiy et al., 2020) model is used to predict the answer directly. If the tokens are found, they are used as an extra input to another LayoutLMv3 model, which tries to output operand tokens required to answer the question. The participant had to use custom rules to annotate labels to finetune the BERT model which processes question text and to extract operand tokens for each question with the *count* type.

7 Results and discussion

Track	Solution	$MASE$	$Acc_{\pm 10\%}$
QA	surkov_evseev	**0.1164**	**91.45%**
VQA	s231644	**0.2165**	**90.08%**
VQA	surkov_evseev	0.2331	81.91%
VQA	fireee80	0.2652	86.15%
VQA	daniyallaiev	0.7874	25.31%
VQA	LayoutLMv3 baseline	0.8786	14.60%
VQA	poddiving	6892	10.53%

Table 4: The shared task results on the test dataset. The best results for each track are in **bold**.

We report the shared task results for both tracks in Table 4.

Our observations from the results table are the following.

- Only a single team hasn't managed to beat an extractive VQA baseline;
- The best models according to $MASE$ also perform the best according to the accuracy based metric $Acc_{\pm 10\%}$;
- The need to apply OCR to extract textual information from images puts significant pressure on the quality of question answering system as highlighted by the drop in metrics of the *surkov_evseev* solution going from the QA track to the VQA track.

The submitted solutions share certain commonalities:

- The performance of the solutions adapted to the task significantly exceeds the performance of a generic extractive VQA model, which hopefully indicates potential for further research in the area;
- The participants have to rely on a preliminary OCR step to explicitly extract text data; building an end-to-end OCR-free solution remains unattainable in practical setting;

- As expected, the participants turn to pretrained VQA models to improve solution performance; building a custom network architecture given the limited size of the dataset remains challenging;
- For practical reasons all the participants instead of doing arithmetic operations on numbers within neural network computations decided on predicting operands involved in calculation / mathematical expression of the calculation and doing the required operations in the post processing step; the ways on how operations on real numbers can be incorporated into the compute within neural networks remains under-explored.

One potential extension of the Receipt AVQA dataset would be the addition of explicit calculation steps for obtaining answers to questions. This should significantly decrease the need for custom annotation efforts when building the models using the dataset and facilitate creation of end-to-end models that require less post processing.

Another interesting direction of future work is expanding the number of templates used for QA generation in order to encourage building solutions that try to estimate required calculation steps automatically rather than rely on predefined rule-based formulas.

8 Conclusion

In this work, we introduce a new challenging Document VQA dataset, named Receipt AVQA, and present the results of the associated RECEIPT-AVQA-2023 shared task.

Receipt AVQA is comprised of 21835 questions over 1957 receipt images. The questions in the dataset are formulated in a way, that to answer them, VQA solution needs not only to recognize and extract tokens from the receipt image, but to apply a number of operations (e.g. sorting, counting, arithmetic operations) over it, thereby testing discrete reasoning capability of the solution.

The associated shared task has attracted 4 teams that have managed to beat an extractive VQA baseline in the final phase of the competition.

We hope that the dataset and promising results of the contestants will inspire further research on understanding documents in scenarios that require discrete reasoning.

Acknowledgements

The authors gratefully thank Vasily Loginov for leading contribution in the Receipt AVQA dataset annotation and verification effort and valuable input in related discussions.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. *CoRR*, abs/1505.00468.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. *// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2019. Leaf-qa: Locate, encode & attend for figure question answering. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, P 3501–3510.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. *// 2019 International Conference on Document Analysis and Recognition (ICDAR)*, P 1516–1520.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*.

Hugging face – the ai community building the future. https://huggingface.co. Accessed: 2023-04-06.

Rob J Hyndman and George Athanasopoulos. 2013. Forecasting: principles and practice.

Kushal Kafle, Scott D. Cohen, Brian L. Price, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, P 5648–5656.

Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *ArXiv*, abs/1710.07300.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *ArXiv*, abs/2109.10282.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoyue Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. 2022. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *ArXiv*, abs/2206.03001.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, P 2199–2208.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. 2021. Infographicvqa. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, P 2582–2591.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. *// 2019 International Conference on Document Analysis and Recognition (ICDAR)*, P 947–952.

Paddleocr: Awesome multilingual ocr toolkits based on paddlepaddle. https://github.com/PaddlePaddle/PaddleOCR. Accessed: 2023-04-06.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: A consolidated receipt dataset for post-ocr parsing.

Devika Patadia, Shivam Kejriwal, Richa Shah, and Neha Katre. 2021. Review of vqa : Datasets and approaches. *// 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)*, P 1–6.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. *CoRR*, abs/1904.08920.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. *ArXiv*, abs/2101.11272.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Chao-Yue Zhang, and Mohit Bansal. 2022. Unifying vision, text, and layout for universal document processing. *ArXiv*, abs/2212.02623.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *ArXiv*, abs/2105.07624.