11-777 Report 1: Dataset Proposal and Analysis

Hao Wu* **Jiayi Shen*** **Yanlin Feng*** **Yinghuan Zhang*** **Yuwei Wu***
{haowu3, jiayis2, yanlinf, yinghuan, yuweiwu}@andrew.cmu.edu

1 Problem Definition and Dataset Choice

TextVQA (Singh et al., 2019) is a Visual Question Answering (VQA) dataset that requires reasoning about text in images. In this section, we present a preliminary analysis of the fundamental challenges in this dataset and propose several hypotheses that can help improve performance over the current state of the art.

1.1 What phenomena or task does this dataset help address?

This dataset helps address the task of VQA that requires reading and reasoning over the text in images. The task is commonly needed by the visually impaired users.

1.2 What about this task is fundamentally multimodal?

This task involves visual and text modalities. Solving this task requires the model to leverage both visual and text content.

1.3 Hypotheses

We believe there are six places where cross-modal information can be used or improved.

1. Spatial information can be used in combination of visual and text features to help improve answer prediction.
2. The original LoRRA model architecture can be improved by introducing a text generative model, which is able to predict multi-token answers by incorporating latent encoding of both visual and text features.
3. OCR can become more robust to transformation (rotation, scaling, etc) with the help of transformation information inferred by object detection module.
4. There are several data augmentation mechanisms we can explore to enhance the model's multi-modal representation learning, including corruption of images and OCR tokens as well as spatial bias augmentation for the detection boxes.
5. To better model the interaction between visual and text features, we have also come up with two training strategies for the VQA task: one is to train with masked answer tokens in OCR, and the other is to train with relevant image patches masked.
6. In addition to prediction accuracy, evaluation can also be conducted under the Average Normalized Levenshtein Similarity (ANLS) metric to avoid a harsh zero score when the prediction is slightly different from the ground truth.

1.4 Expertise

We have the following expertise in the underlying modalities required by this task:

1. Hao Wu: Took CV in Fall 2021; experience in 3D vision.
2. Jiayi Shen: Took Intro to Machine Learning in Fall 2021, and taking Intro to Deep Learning this semester.
3. Yanlin Feng: Research experience in question answering, knowledge graph reasoning, graph neural networks and multilingual NLP.
4. Yinghuan Zhang: Took Intro to Machine Learning and Intro to Deep Learning in Fall 2021.
5. Yuwei Wu: Have conducted research in NLP areas, including question answering, dialog systems, representation learning and multi-modal machine learning.

*Everyone Contributed Equally – Alphabetical order

2 Dataset Analysis

2.1 Dataset properties

TextVQA v0.5.1 contains 45,336 questions based on 28,408 images.

1. Training set contains 34,602 questions (103 MB) based on 21,953 images (6.6 GB) from OpenImages' training set.

2. Validation set contains 5,000 questions (16 MB) based on 3,166 images from OpenImages' training set.

3. Test set contains 5,734 questions (13 MB) based on 3,289 images (926 MB) from OpenImages' test set.

2.2 Compute Requirements

1. Files: the upper bound of memory cost for the entire training set is about 6.6GB. We are able to fit it in our servers.

2. Baseline Models:

   (a) **LoRRA**: training modules include one LSTM and two attention modules, trained for 24000 iterations with a batch size of 128 on 8 GPUs (not specified in the paper);

   (b) **M4C**: training modules include BERT-base and the last layer of the Faster R-CNN, together taking 10 hours on 4 Nvidia Tesla V100 GPUs of batch size 128;

   (c) **SAM**: training modules include two BERT-Base models. It took 12 hours to train 100 epochs on 2 NVIDIA Titan XP GPUs with batch size 96;

   (d) **LayoutLM**: training modules include one BERT-large model. It requires about 12-16 GB GPU memories to train the model for 100 epochs with batch size 16.

2.3 Modality analysis

2.3.1 Text Analysis

To better understand the TextVQA dataset from the language perspective, we conduct an analysis on the questions and answers on all three data splits. TextVQA contains 45336 questions in total, of which 37912 are unique. Following the original paper, we visualize the distribution of question lengths on all three data splits in Figure 1a. Since the TextVQA validation and test split are randomly sampled, the distribution is almost identical across data splits. The average question lengths in the three data splits are 7.18/7.21/7.11 respectively. As shown in Figure 1b, more than 78% of the questions in TextVQA are what-questions. We also show the 10 most frequent questions and their frequency in Figure 1c. The distribution of the 10 most frequent answers is shown in Figure 1d. It is worth noting that more than 600 questions are labeled as unanswerable while more than 400 questions are labeled as "answering does not require reading text in the image". Other answers in TextVQA include various entity types like numbers, brands, cities, time, and people's names.

2.3.2 Vision Data Analysis

Table 1: TextVQA dataset Rosetta OCR Error Analysis

Dataset	Error Rate	Detection Error Rate	Token Error Rate
Training	29.17%	7.41%	23.50%
Validation	40.55%	6.67%	31.58%
Test	29.92%	15.35%	17.21%
Overall	32.05%	9.80%	24.67%

We conduct exploratory analysis on vision data of TextVQA dataset, i.e., Rosetta OCR data following the original paper. To measure the error rate of OCR, we randomly sample 20 images from each of training set, validation set, and test set. *Error Rate* measures the share of all incorrectly recognized tokens out of all detected boxes; *Detection Error Rate* measures the share of OCR tokens generated from non-text boxes out of all detected boxes; and *Token Error Rate* measures the recognition error rate among the detected boxes that do contain text, as shown in Table 1.

Moreover, the average numbers of objects and OCR tokens detected in each image are examined and presented in the Table 2. It is worth noting that only 2897 out of 3289 images in the testing dataset are present in OpenImages' v0.6 dataset, which should contain all the image data according to the TextVQA web page.

By glancing over the data with OCR visualization, we conclude several troublesome problems of the Rosetta OCR data:

1. OCR tokens are designed to provide the model

Table 2: Exploratory Analysis on Object & OCR Tokens Detection

Dataset	Avg objects detected	Avg OCR tokens
Training	5.04	12.45
Validation	5.31	12.89
Test	9.14[1]	9.60

[1] Only 2897 out of 3289 testing images are present in OpenImages' v0.6 dataset

with answer candidates. A high *Detection Error Rate* provides the model with more confusing and meaningless answer candidates; a high *Token Error Rate*, even when the model can "copy" the correct answer space, brings the final accuracy down due to the incorrect recognition.

2. Rosetta OCR recognizes each detection box into one token, which possibly consists of several words. The combination of the word, even each recognized correctly, is useless for the question answering model.

We present several typical errors in the model answering related to OCR data in Figure 2. Specifically, Fig 2a shows an image where correct answer can be "copied" from the OCR token, however, the OCR token is recognized incorrectly; Fig 2e detects and recognizes correctly, but multi tokens should be combined together as the answer; Fig 2c has an answer that can be "copied" from the OCR token as well, whereas the answer is not detected by OCR model; Fig 2b is not answered correctly because the answer token is partially covered and OCR model is not able to complete it; Fig 2d has partial answer token correctly detected and recognized, however, the OCR is not helpful to obtain the rest part of the answer.

2.4 Metrics

1. Accuracy: the accuracy score between predicted answers and ground truth labels.

2. Average Normalized Levenshtein Similarity (ANLS): the accuracy metric awards a zero score even when the prediction is only a little different from the target answer. Since no OCR is perfect, Mathew et al. (2021) proposed the metric called ANLS for DocVQA task. Motivated by their work, we would also like to include this metric in our evaluation.

2.5 Baselines

1. Singh et al. (2019) introduced TextVQA dataset and the Look, Read, Reason & Answer (**LoRRA**) architecture, which is based on a combination of visual detection and OCR module.

2. Hu et al. (2020) proposed model **M4C** for the TextVQA task based on a multimodal transformer architecture accompanied by a rich representation for text in images.

3. Kant et al. (2020) proposed model **SAM** with a novel spatially aware self-attention layer such that each visual entity only looks at neighboring entities defined by a spatial graph. Further, each head in the multi-head self-attention layer focuses on a different subset of relations.

4. Xu et al. (2020) proposed **LayoutLM** for document image understanding, which jointly models the interactions between text and layout information and further leverages the image features through two pretraining tasks: masked visual language model and multi-label document classification.

5. Xu et al. (2021) proposed **LayoutLM 2.0**, which is an extension of **LayoutLM** with two auxiliary pretraining tasks, namely Text-Image Alignment and Text-Image Matching. These two tasks help models to learn the correspondence between the document image and its textual content. The resulting model achieved the second place on the DocVQA dataset.

3 Team member contributions

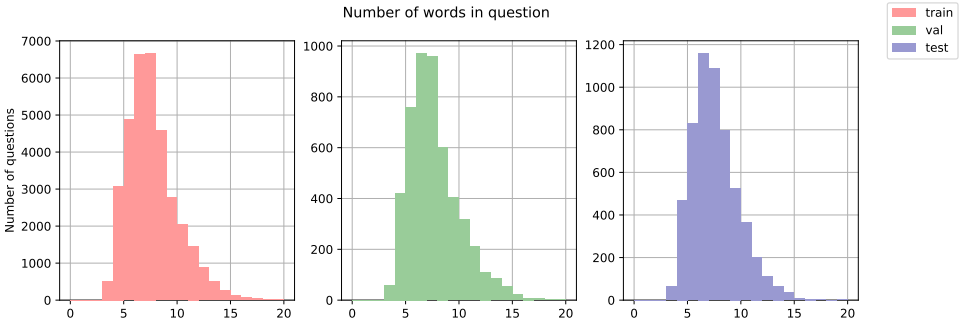
Hao Wu contributed to the hypothesis proposal, exploratory data analysis on object and OCR token detection in images, and baseline summary.

Jiayi Shen contributed to the text analysis and hypothesis ideas.

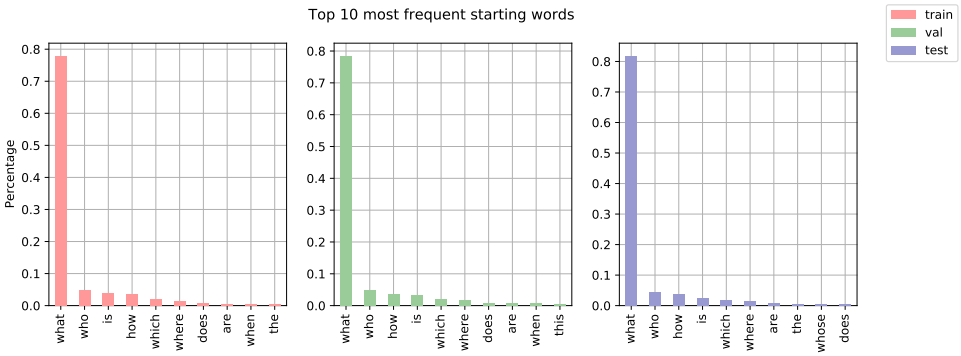
Yanlin Feng contributed to the text analysis, baseline summary, and hypothesis ideas.

Yinghuan Zhang contributed to the OCR error analysis, typical errors related to OCR data, and hypothesis ideas.

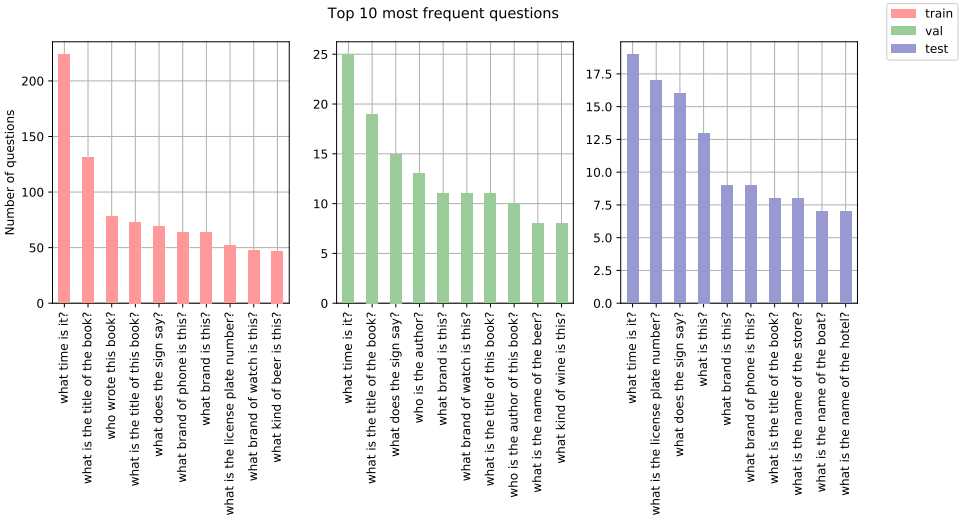
Yuwei Wu contributed to the OCR analysis, baseline summary, and hypothesis ideas.



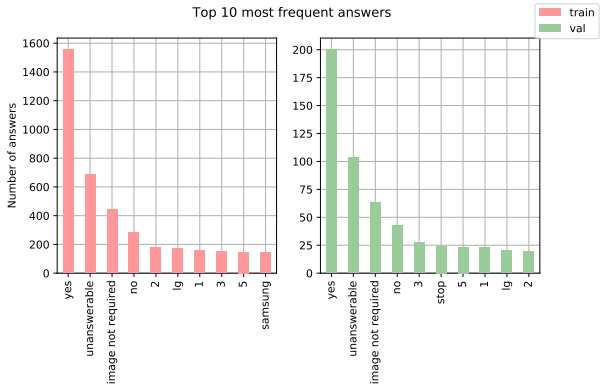
(a)



(b)



(c)



(d)

Figure 1: Text statistics of the TextVQA dataset.

(a) Incorrect Recognition

(b) Text Partially Covered

(c) Answer Not Detected

(d) OCR Not Helpful

(e) Need Multi Tokens

Figure 2: Typical errors related to OCR data.

References

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002.

Yash Kant, Dhruv Batra, Peter Anderson, Alexander Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. 2020. Spatially aware multimodal transformers for textvqa. In *European Conference on Computer Vision*, pages 715–732. Springer.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. 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*, pages 1192–1200.