Technical Report on DocVQA

Pavan Kumar K
CVIT, IIIT, Hyderabad, India
Reproduced the paper results

***Abstract*—The DocVQA dataset, a novel contribution to the field of Visual Question Answering on document images, was thoroughly examined in this reproduction study. This dataset comprises over 50,000 questions posed on more than 12,000 document images. A comparative analysis with similar datasets in VQA and reading comprehension was conducted. The study involved applying established VQA and reading comprehension models to set baselines. While these models showed decent performance on certain question types, a significant gap in accuracy (94.36%) compared to human performance was observed, particularly in questions demanding an understanding of the document's structure. Access the DocVQA dataset, its code, and leaderboard at DocVQA**

I. Introduction

I explored Document Visual Question Answering (DocVQA) as a significant advancement in Document Analysis and Recognition (DAR). Unlike traditional DAR tasks focused on specific information extraction like character recognition or table extraction, DocVQA emphasizes a holistic understanding of document images, driven by natural language questions. This approach requires an intelligent reading system to interpret both textual and visual elements, such as layout, style, and non-textual cues, within documents. My analysis included a review of the large-scale DocVQA dataset, consisting of over 12,000 diverse document images and 50,000 questions. The study also encompassed the evaluation of various baseline methods, ranging from heuristic approaches to advanced Scene Text VQA and NLP models, highlighting the open-ended nature of answers in document VQA compared to traditional VQA tasks. The implementation of BERT and M4C models was crucial in addressing the unique challenges presented by the open-ended nature of answers in document VQA.

II. Method

This Section discusses the importance of the models in the VQA tasks.

1. BERT: Bidirectional Encoder Representations from Transformers is highly effective for Visual Question Answering (VQA) tasks, especially when combined with image processing techniques. In VQA, BERT can be utilized to understand and process the textual content of questions and possible answers. When integrated with convolutional neural networks (CNNs) for image analysis, BERT helps in accurately interpreting the context and details within images, thereby enhancing the model's ability to generate relevant and precise answers to visually grounded questions. This synergy between BERT's language understanding and visual data processing makes it a powerful tool for VQA tasks.

2. M4C: The Multimodal Multi-Copy Mesh model, designed specifically for Visual Question Answering (VQA) tasks, is notable for its ability to integrate textual and visual data. It leverages a multimodal transformer to fuse inputs from different modalities, including image features and text. M4C is particularly adept at answering questions that require understanding text in images, such as those found in document-based VQA tasks. Its architecture allows it to reference multiple parts of an image and text simultaneously, making it a robust choice for complex VQA challenges.

III. Experiments

This section discusses the evaluation metrics used, describes the experimental setup for the reproduction of the experiments, and summarizes the results obtained. Reproduces Code and more details can be found at DocVQA-Reproduced

Evaluation Metrics: In my replication of the DocVQA paper, I adopted two evaluation metrics, Average Normalized Levenshtein Similarity (ANLS) and Accuracy, as outlined in the original paper. ANLS, a metric primarily used in ST-VQA evaluations, compensates for minor discrepancies often caused by OCR inaccuracies. It's particularly useful as it doesn't severely penalize slight mismatches in answers. On the other hand, the Accuracy metric quantifies the percentage of questions where the predicted answer precisely matches the target answer, assigning zero scores for even near-accurate predictions. ANLS was chosen as the primary metric in my analysis to allow for a more nuanced evaluation of the OCR-driven results.

Experimental Setup: The DocVQA dataset comprises 50,000 questions based on 12,767 images. This dataset was divided into training, validation, and testing sets in an 80-10-10 split. Specifically, the training set included 39,463 questions across 10,194 images, the validation set comprised 5,349 questions and 1,286 images, and the test set consisted of 5,188 questions with 1,287 images.

Packages utilised: To execute the code, you'll need: Python, Git, simple transformers, PyTorch, Transformers (from Hugging Face), CUDA (for Nvidia GPUs), Nvidia GPUs (for BERT large models), Logging (Python standard library)

In my study, I utilized three pre-trained BERT models from the Transformers library for Question Answering tasks on the DocVQA dataset. These models included bert-base-uncased, bert-large-uncased-whole-word-masking, and bert-large-uncased-whole-word-masking-finetuned-squad, referred to as bert-base, bert-large, and bert-large-squad, respectively. The bert-large-squad model additionally fine-tuned on SQuAD 1.1, was adapted to handle the unique context of DocVQA, where document images replace text passages. I arranged the OCR tokens from the documents into serialized strings to fine-tune these BERT models.

The bert-base model was fine-tuned using 2 Nvidia GeForce 1080 Ti GPUs for 2 epochs, a batch size of 32, and a learning rate of 5e-05. In contrast, the bert-large and bert-large-squad models were fine-tuned on 4 GPUs for 6 epochs with a batch size of 8, eval batch size of 64, max answer length of 50 and a learning rate of 2e-05.

I utilized the official M4C model implementation from the MMF framework. Also for image feature extraction, I have followed this MMF Documentation on Image Feature Extraction. My approach mirrored the training settings and hyperparameters outlined in the original M4C model study. For the M4C vocabulary, Also as mentioned by the authors I adopted the fixed vocabulary strategy, using the 5,000 most frequent words derived from the answers in the training set, ensuring consistency with the established methodology.

Quantitative Results:

Pretrained model	**val**		**test**	
	ANLS	**Acc.**	**ANLS**	**Acc.**
bert-base	0.56	45.8	0.575	47.4
bert-large-squad (not Finetuned on DocVQA)	0.596	49.30	0.61	51.1
bert-large-squad	**0.655**	**54.50**	**0.665**	**55.79**
M4C	0.354	23.90	0.372	24.28

TABLE I
THE BEST PERFORMANCE IS EXHIBITED BY A BERTLARGE-SQUAD MODEL THAT HAS UNDERGONE FINE-TUNING ON BOTH THE SQUAD DATASET AND DOCVQA.

CONCLUSION

In conclusion, the replication of the DocVQA paper highlighted the potential and challenges of VQA models in interpreting complex document images. The BERT and M4C models demonstrated promising capabilities, with the fine-tuned BERTLARGE-SQuAD model achieving the best results. This study underscores the importance of specialized training and the need for further research to bridge the gap between model performance and human accuracy, particularly in understanding the nuanced interplay of textual and visual elements in documents.