DocVQA Challenge - Task 1

January 14, 2021

Name	Guilherme Moraes Rosa
College	Faculty of Electrical and Computer Engineering
University	State University of Campinas - UNICAMP
Email	G264437@dac.unicamp.br

1 Abstract

Systems that use natural language processing architectures, such as BERT, have proven to be very effective in dealing with textual information, such as in the Question Answering task. However, the Document Visual Question Answering (DocVQA) challenge focuses on the Visual Question Answering task, in which a visual understanding of the information in a document image is also necessary to provide the correct answer. Therefore, in order to accurately recognize the text fields of interest and obtain a good performance in the task, it is inevitable to take advantage of the multi-modal nature of the documents, where the textual, visual and layout information must be modeled together and learned end to end in a single structure. This article explores the effectiveness of the T5, LayoutLM and EffficientNet models applied to the context of visual question answering and also explores the feasibility of architectures formed from the combination of these models in task 1 of the DocVQA challenge. The experiments showed that T5 only and T5+EfficientNet have a reasonable performance for the proposed task, with T5 reaching a F1 Score up to 60 points, while T5+LayoutLM presented learning difficulties during training. Code is available at https://github.com/guilhermemr04/IA376J-Final-Project.

2 Introduction

Understanding documents is a very challenging task for neural networks due to the diversity of layouts and formats, as well as the complexity of their structure. Developing models capable of seeing, reading, understanding and analyzing documents automatically is a relatively new and very promising research topic. These documents can be of digital origin, written, digitized or printed on paper. Some common examples include financial reports, purchase orders, business emails, letters, invoices, sales agreements, supplier contracts, receipts, resumes and more. The format of the documents varies widely, but the information is usually presented in natural language and is organized in different ways, from simple text to different layouts or a wide variety of tables, forms and figures.

Distinct from conventional information extraction tasks, the Visual Question Answering task is more complex and goes far beyond understanding the textual content of a document, but also visual and layout information including fonts, colors, as well as elements such as marks, checkboxes, separators, diagrams and other information like page structure, forms or tables, which is vital for understanding the documents. The challenge is composed of three different tasks. In task 1, questions are defined in a single document and the answer needs to be given by interpreting the document's image. The Task 2 goal is to identify and recover all documents in a collection that are relevant to answering this question, as well to provide one. Finally, Task 3 is a VQA task in which questions are also defined in single images, but this time the images are infographics on different topics where visual information is more relevant

to answer the questions asked.

To this end, I propose three approaches to deal with task 1 of the DocVQA challenge [1], where each approach uses different architectures and input data. The first experiment consisted of using a T5 Base model and the textual information extracted by OCR as an input, which includes the question and the document's context. For the second experiment I used LayoutLM [2] and T5 combined with the textual features used in the previous step and the OCR 2D position information provided by the dataset as the inputs. For the third experiment, two models were also applied. This time EfficientNet [3] and T5[4], with the visual features provided by the scanned document and the textual features from the OCR. The experiments illustrate the difficulties imposed by the challenge. Among the three architectures implemented, the combination of T5 and LayoutLM had a hard time learning the task, while the other two had better results and outperformed one of the BERT-Large models provided by the authors of the paper DocVQA.

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

• The evaluation of three different architectures on the DocVQA dataset
• The code used to perform the task

This work is organized as follows. Sections 3, 4, 5, 6 and 7 present the theoretical framework necessary to understand the present paper. Section 8 presents the methodology used to develop this proposal. Section 9 presents the experimental results obtained. Finally, section 10 presents the conclusion of the work together with some perspectives for future work.

3 Related Work

Recently, pre-trained language models, such as GPT [5], BERT [6], XLNet [7], RoBERTa [8], ALBERT [9] and T5 [4] have pushed great advances on NLP tasks. These models have been used in several applications, such as summarization, sentiment analysis, and questions-answering. More recently, there has been a growing interest in supervised learning for multi-modal tasks. In this type of task information is received by the model in more than one mode, for example a VQA model, which is challenged to deal with both visual and textual information. For example, VideoBERT [10] and CBT [11] applied BERT to learn a distribution about video frame features and linguistic tokens from video text pairs. LXMERT [12] introduced the two-stream architecture, where two Transformers are applied to images and text independently and are merged by a third Transformer at a later stage and ViLBERT [13] were proposed for learning task-agnostic representations of image content and natural language by extending the popular BERT [6] architecture to a multi-modal two-stream model. On the other hand, VisualBERT [14] and VL-BERT [15] proposed a single flow architecture, that adopt the BERT model as the backbone, and extend it to take both visual and linguistic embedded features as input. VLP [16] applied pre-trained models for both image captioning and VQA.

The Visual Question Answering (VQA)task aims to provide an accurate answer in natural language given an image and a question. This field of study has attracted an intense research effort in recent years. Some examples of datasets include ST-VQA [17] and TextVQA [18], which were introduced in 2019 and were quickly widely used by researchers and developers. The ST-VQA dataset has more than 31,000 questions in more than 23,000 images collected from different public data sets. The TextVQA dataset has more than 45,000 questions about 28,000 images sampled from specific categories in the OpenImages dataset. Another well-known dataset is the OCR-VQA, which has more than 1 million pairs of questions and answers in more than 207,000 book cover images.

Compared to these VQA datasets that have been cited, DocVQA comprises a great diversity of images. The data set covers a multitude of different types of documents that include elements such as tables, figures, forms and graphics.

4 Datasets

In this section I explain the datasets used to train the models.

4.1 DocVQA

DocVQA [1] comprises 50,000 questions in 12,767 images. Data is randomly divided by an 80/10/10 ratio for training, validation and testing. The test dataset has 39,463 questions and 10,194 images, while the validation dataset has 5,349

questions and 1,286 images and the test contains 5,188 questions and 1,287 images.
The images provided in the dataset come from documents hosted at the Industry Documents Library, maintained by UCSF. The set of images was extracted from more than 6,000 documents used in the industry, most of these, from a period between 1960-2000. There are documents from all five main industries for which the library hosts documents, which are tobacco, food, drugs, fossil fuels and chemicals. The documents contain a mixture of printed, typewritten and handwritten content. A wide variety of document types are used for this task, including letters, memos, notes, reports, etc. The answers to the questions are short excerpts from the text. This means that the responses comprise a set of contiguous text tokens present in the document. There may be more than one valid answer per question. In this case, a list of possible correct answers is provided in the dataset.

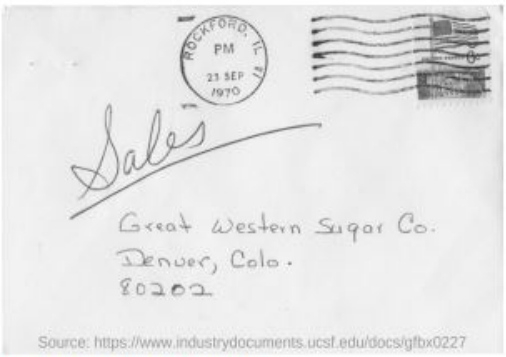

Figure 1: DocVQA example

4.2 Natural Questions

Natural Questions (NQ) is a challenging dataset for open domain question-answering [19]. It was created to help boost research advances in question-answering systems. It consists of 307,373 training, 7,830 development, and 7,842 test examples, with each example including a question next to a Wikipedia article that may or may not contain the answer. According to Kwiatkowsk et al., NQ is the first dataset to use naturally occurring queries and to focus on finding answers by reading an entire page, rather than extracting from a short paragraph. In order to create the dataset, aggregated queries, real and anonymous, that users made to the Google search engine were cataloged. After collection, human annotators were asked to find answers by reading an entire Wikipedia page. They sought long answers that covered all the information needed to infer an answer and also short answers that answered the question succinctly.

5 Task

Current Document Analysis and Recognition research focuses on generic information extraction tasks, such as character recognition, table extraction and word localization, largely disconnected from the final purpose for which the extracted information is used. The DocVQA challenge seeks to inspire a new point of view in Document Analysis and Recognition research, where the document's content is extracted and used to perform high-level tasks defined in natural language by humans. DocVQA is modeled as a Visual Question Answering (VQA) problem in which the task is to answer a question in natural language made on a single document image. This goes beyond passing a document image through OCR and involves understanding all types of information transmitted by a document. Textual content (handwritten or typed), non-text elements (marks, checkboxes, separators, diagrams), layout (page structure, forms, tables) and style (font, colors, highlights), are just some of the information that can potentially be necessary to answer the questions posed by the challenge. For this reason, the DocVQA dataset constitutes a new type of problem, where high-level semantic tasks dynamically lead information extraction algorithms to conditionally interpret images from documents.

6 Models

For evaluating performance of existing models on DocVQA [1] I choose three different models. In this section, I briefly review each of them. All architectures were implemented in Pytorch with the HuggingFace and Pytorch Lightning libraries.

6.1 T5

In 2019, Raffel et al. [4] introduced a new model called "Text-to-Text Transfer Transformer" or T5. This model presents an unified structure that converts all language tasks into a text-to-text framework, in which inputs and outputs are texts. This format provides a simple way to train and run inference on a single model on a wide variety of tasks, such as machine translation, document summarization, question-answering, and classification, using the same loss function and decoding procedure. The implementation of the T5 closely follows the Transformer model originally proposed by Vaswani et al. [20]. Composed of encoder and decoder, each of which consists of two sub-components: a self-attention layer, followed by a feed forward network. The model is pre-trained on the Colossal Clean Crawled Corpus (C4), which is a pre-processed version of a publicly available texts extracted from the web. The T5 model was provided in 5 different sizes (Base, Small, Large, 3B and 11B). The framework demonstrated state-of-the-art performance in several natural language processing tasks.

6.2 LayoutLM

Using BERT [6] Base as a backbone and inspired by its architecture, the authors of LayoutLM [2] added two types of input embeddings, a 2-D positional embedding and an image embedding. These extra embeddings were incorporated into the architecture because the first one can capture the relationship between tokens within a document, while the second one captures some appearance features, such as fonts, types, colors and represents these image attributes in a language representation.
LayoutLM was pre-trained on the IIT-CDIP Test Collection 1.0, which contains more than 6 million documents with 11 million document images in various categories. The authors used two tasks for pre-training:

Masked Visual-Language Model: consists of, during pre-training, randomly masking some of the input tokens, but maintaining the corresponding 2-D position embeddings and then training the model to predict the masked tokens. In this way, the model, in addition to understanding the language context, also uses the corresponding 2-D position information.

Multi-label Document Classification: Given a set of digitized documents, the authors used the document tags to supervise the pre-training process so that the model can group knowledge from different domains and generate a better representation at the document level.

For the first time, the textual and image layout information of scanned documents were pre-trained in a single structure. This made the model obtain state-of-the-art performance in tasks of understanding images in documents and significantly surpassed several other text only pre-trained models.

6.3 EfficientNet

EfficientNet [3] is a model launched in 2019 by Google. The authors used Neural Architecture Search to build an efficient network architecture. The main building block of EfficientNet consists of MBConv, to which is added the compression and excitation optimization that form a shortcut connection between the beginning and the end of a convolutional block. Input activation maps are first expanded using 1x1 convolutions to increase the depth of resource maps. This is followed by 3x3 Depth-wise convolutions and Point-wise convolutions that reduce the number of channels on the output resource map. Shortcut connections connect the narrow layers, while the wider layers are present between the jump connections. This structure helps to decrease the overall number of operations required, as well as the size of the model. In order to improve accuracy and efficiency, the authors proposed a simple and effective scaling technique, which uses a composite coefficient to uniformly dimension the width, depth and resolution of the network in a principled manner. In addition, all layers or stages in scale models use the same convolution operations as the baseline network called EfficientNet-b0. This technique allowed the authors to produce models that provided greater precision than existing convolutional networks, in addition to a reduction in the number of FLOPS and the size of the model.

7 Evaluation Metrics

In this section I explain evaluation metrics. Two different metrics were used to evaluate the models accuracy, and both ignore punctuations and articles [21].

Exact match: This metric measures the percentage of predictions that match any one of the ground truth answers exactly.
F1 score: This metric measures the average overlap between the prediction and ground truth answer. The prediction and ground truth are treated as bags of tokens, and then their F1 is computed.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

TP = True positive
TN = True negative
FP = False positive
FN = False negative

Figure 2: Evaluation metrics

8 Experiments

The models were trained on a single Tesla V100 GPU at Google Colab with a constant Learning Rate of 5e-5, over ten epochs, with Adam Optimizer and batch size of 32. A maximum of 512 input tokens and 32 output tokens were selected. I used the T5-Base and LayoutLM-Base in the experiments and did not try the T5-Large and LayoutLM-Large due to their computational cost. I also used EfficientNet-b4, since its improved speed of convergence in comparison with smaller variations.

Table 1: Models and features description

Model	Description
T5	Question and context from DocVQA
T5-NQ	Question and context from DocVQA
LayoutLM+T5	Question, context and OCR 2D positions
EfficientNet+T5	Question, context and image

8.1 T5 training

The first experiment consisted of training two T5 models, a T5 model from Transformers library and a T5-Base pretrained on Natural Questions dataset. Both models received question and document's context as input.

Figure 3: T5 data flow

8.2 LayoutLM+T5 training

For the second VQA experiment, I used an architecture composed of the LayoutLM and T5 models. The architecture receives as input the same textual information as the models of the previous experiment, together with the OCR position data provided by the dataset. All of these features are passed to LayoutLM, which creates an embedding representation, before moving on to T5, which is the model responsible for providing an answer to the question.

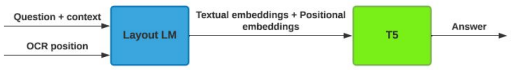

Figure 4: LayoutLM + T5 data flow

8.3 EfficientNet+T5 training

At the third VQA experiment, I used an architecture composed of the EfficientNet and T5 models. The architecture receives as input the same textual information as the previous models, but this time it also receives the image of the scanned documents. In this implementation, visual embeddings go through CNN, while textual embeddings go through the T5 encoder. After that, the two representations are concatenated and passed to the T5 decoder to provide an answer.

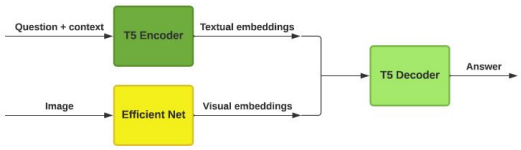

Figure 5: EfficientNet + T5 data flow

9 Results

Now, I discuss the main experimental results, evaluating the proposed system in the unseen data sets. Table 2 shows the performance of each model. For comparison purposes, the BERT models trained by the authors of the DocVQA paper were included in the table.

Table 2: Results

Model	Exact Match	F1 Score
Bert-base	45.6	
Bert-large	49.28	
Bert-squad-large	54.48	
T5-base	50.49	58.87
T5-NQ-base	53	61.8
EfficientNet+T5-base	47.3	57.2

As can be seen in Table 2, the T5 pretrained on Natural Questions obtained a relatively better result than T5-Base with the addition of a faster convergence, which show us that it is a good ideia to pretrain the model on a question answering dataset before fine-tuning on DocVQA. Besides that, both models can outperform a simple BERT-Large on the task and a T5-Base pretrained on NQ gets a performance close to a Bert-Large pretrained on SQUaD.

During the tests, the architecture formed by LayoutLM + T5 had some difficulty in the proposed task, taking a long time to converge and achieving a poor performance. So it has no results in the table. Some variations were tested, such as passing the textual information through the T5 encoder and right after concatenating with the LayoutLM output, but no attempt has achieved good results.

At the third experiment, two choices were important for the architecture to converge. An EfficientNet from size b4 and a T5-base pre-trained on Natural Questions and textual features of the DocVQA. The EfficientNet-b4 was selected due to its small size and the same convergence speed in relation to the larger variations of the same model. Other combinations between T5 and EfficientNet models were tested but not resulted in good performance, with slower convergence speed and significantly high initial loss. Despite using visual information, the architecture composed by EfficientNet + T5 obtained a result below the experiment in which only the T5 model was used, in addition to having a considerably lower convergence speed in comparison. It is possible that the features generated by EfficientNet were not of good quality and that employing other CNN's in this type of task will bring better results.

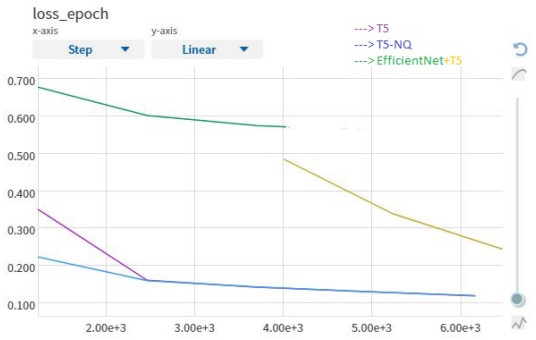

Figure 6: T5 and EfficientNet+T5 loss for first epochs

10 Conclusion

In this paper, three question and answer systems were developed to extract information from digitized documents using T5, LayoutLM and EfficientNet models trained on the DocVQA dataset. When examining the full set of results in this paper, these main conclusions can be drawn:

• The dataset motivate simultaneous use of visual and textual information for answering questions asked on document images.

• The model T5 trained with Natural Questions has been shown to be a good starting point for building question and answer systems.

• Given the difficulty of the DocVQA dataset, the architectures based on the T5 and EfficientNet models did a satisfactory job in the task of understanding text in digitized documents.

From this study, it appears that the T5-base is able to perform well in this type of task, outperforming a simple Bert-large, which gives us confidence about the use of this type of system in future applications. There are some limitations and opportunities for further work that arise from this study. For future research, I will further explore the network architectures that receive visual embeddings as input as well as some pre-training strategies, hoping that it can push the results higher.

References

[1] C.V. Jawahar Minesh Mathew, Dimosthenis Karatzas. Docvqa: A dataset for vqa on document images. *arXiv preprint arXiv:2007.00398*, 2021.

[2] Lei Cui Shaohan Huang Furu Wei Ming Zhou Yiheng Xu, Minghao Li. Layoutlm: Pre-training of text and layout for document image understanding. *https://arxiv.org/abs/1912.13318*, 2019.

[3] Quoc V. Le Mingxing Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. *https://arxiv.org/abs/1905.11946*, 2019.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Yiming Yang Jaime Carbonell Ruslan Salakhutdinov Quoc V. Le Zhilin Yang, Zihang Dai. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[8] Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer Veselin Stoyanov Yinhan Liu, Myle Ott. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[9] Sebastian Goodman Kevin Gimpel Piyush Sharma Radu Soricut Zhenzhong Lan, Mingda Chen. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[10] Myers A. Vondrick C. Murphy K. Schmid C Sun, C. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1901.08634*, 2019.

[11] Baradel F. Murphy K. Schmid C Sun, C. Contrastive bidirectional transformernfor temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019.

[12] Bansal M Tan, H. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1906.05743*, 2019.

[13] Batra D. Parikh D. Lee S.: Vilbert Lu, J. Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1906.05743*, 2019.

[14] Yatskar M. Yin D. Hsieh C.J. Chang K.W Li, L.H. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[15] Zhu X. Cao Y. Li B. Lu L. Wei F. Dai J Su, W. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1906.05743*, 2020.

[16] Palangi H. Zhang L. Hu H. Corso J.J. Gao J Zhou, L. Unified visionlanguage pre-training for image captioning and vqa. *arXiv preprint arXiv:1906.05743*, 2020.

[17] Andres Mafla Lluis Gomez Marcal Rusinol Ernest Valveny C.V. Jawahar Ali Furkan Biten, Ruben Tito and Dimosthenis Karatzas. Scene text visual question answering. *arXiv preprint arXiv:1901.08634*, 2019.

[18] Meet Shah Yu Jiang Xinlei Chen Devi Parikh Amanpreet Singh, Vivek Natarjan and Marcus Rohrbach. Towards vqa models that can read. *arXiv preprint arXiv:1901.08634*, 2019.

[19] Kwiatkowsk, Palomaki, and Redfield. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 2019.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[21] Konstantin Lopyrev Percy Liang Pranav Rajpurkar, Jian Zhang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.