RESEARCH PROPOSAL

Sai Rajeswar *Research scientist*

1 Research Proposal: Enhancing User-AI Interaction: A Study on Multimodal Models for Document-Based Conversations (VQA)

1.1 Research Abstract

This project is designed to investigate the role of multimodal models in enabling document-based dialogues. This exploration will delve into how these models comprehend the context of a document and generate pertinent, coherent responses, thereby augmenting the communication between users and AI systems. The focus will be on interpreting a document (either an image or a representation in Word, latex, or PowerPoint format), extracting relevant and useful information, and answering queries related to it. Examples of such questions could include inquiries about the total dollar amount of a bill or the specific products purchased as indicated on the bill. The research combines natural language processing capabilities along with visual understanding [3, 1, 4]. This opens up possibilities for applications such as generating relevant captions for images, answering questions related to visual content, and contextual understanding of images. We would leverage existing multimodal architectures and pre-trained models (LLaVa, CogVLM, or Fuyu-8B) to propose novel ideas for improving the state-of-the-art.

1.2 Internship Details:

The project is managed by a team of researchers and academic professors at ServiceNow Research. Initially, the internship will be conducted remotely. However, based on successful performance during this period, the student will be invited to join us for a paid, in-person internship with the ServiceNow Research team in Montreal.

1.3 Background and Literature Review

Document VQA methodologies can generally be classified into three categories. The first, utilize a single modality [7] and, despite good performance using transformer-based models, they fall short when multimodal understanding is required.

The second category incorporates multimodal architectures to simultaneously process visual and textual content [5, 6]. However, these models often require retraining with each new type of input due to varying document layouts. For instance, a model trained on a dataset where addresses are typically at the top of documents may struggle to locate them in different positions in new documents or datasets. This necessitates retraining on new datasets with their respective annotations.

To address this, recent models like LayoutLM [2] have been developed to incorporate a third modality representing the layout information of documents. This category includes notable models like LayoutLM, LamBERT, and ViBERTgrid. LamBERT, for instance, is based on the Transformer encoder architecture RoBERTa.

1.4 Research Objectives and Sub-Projects

In this project, we propose a method capable of processing various types of documents and extracting diverse information as per the end user's requirements for a question-answering model. The advantage of employing visual question answering on text-intensive documents, as opposed to predefined extraction, lies in its ability to extract more general information. This information can then be adapted to a new corpus of data, enhancing the model's flexibility and applicability. Here are some specific details:

1. **Bottom up fashion: Recognize and answer**: Approach first produces structured text output, capturing content and styles. Later it can be fine-tuned as a single model for question answering. The model itself combines a ViT-based vision encoder and a Transformer-based language decoder linked by a association module for sample efficiency. And is pretrained on a large corpus of text-intensive images.

2. **Prompt based VQA**: Here we aim to develop a captioning model that enhances the connection between images and black-box language models. This model will differ from generic captions by using a natural-language prompt to guide the description of visual entities in the produced caption. The prompt will include a question that the caption should assist in answering. The idea here is to leverage more powerful LLMs like GPT-3.5 or GPT-4. The output from our model can serve as a context for better prompt engineering.

3. **Large-Scale Multimodal Pre-Training**: Explore the benefits of large-scale multimodal pretraining for improving the performance of LMMs on text-intensive images.

4. **Real-World Applications**: Apply the developed LMMs to real-world scenarios, such as the creation and editing of scientific documents, to evaluate their effectiveness and identify areas for further improvement.

5. **Document conditional chat models for VQA:** In this effort, our goal is to devise strategies where generation of an existing language model able to carry out conversations can be conditioned on a document. A simple initial approach could be as simple as carrying out OCR with an external model and then feed the OCR results as part of the prompt to query a language model of choice. As a second step, our focus will be to then feed pixels directly into our language model without the use of an external OCR system, which could be achieved via extra cross-attention layers added to the language model of interest and fine-tuned to enable this new capability.

2 Datasets

1. **Corporate Documents**: These documents are varied in both content and form (i.e. invoices, order forms, resumes, pay slips etc). VQA-CD is a new public dataset containing 3000 questions extracted around 693 documents from RVL-CDIP.

2. **DocVQA dataset**: could be the most complete dataset in both content and number of samples. The dataset consists of 50,000 questions defined on 12,000 document images.

3. **IIT-CDIP Dataset:** This dataset is a comprehensive public repository of scanned document images on a large scale. Roughly 27.6 million pages from this dataset will be employed to train our model.

4. **PowerPoint Documents:** We plan to gather a corpus of at least a million pages from various web pages presenting PowerPoint documents, thereby substantially augmenting our training data's diversity.

5. **Universal PDF:** On top of this, we plan tp conduct web crawling for diverse, open-domain digital PDF files. This resulted in the accumulation of a vast corpus.

6. **Web Screenshots:** Further, a subset of the mC4 web pages would be used and rendered into screenshots, capturing nearly 100 million or more pages

3 Evaluation

1. **Average normalized Levenshtein (ANLS)**

2. **F1-Score**

3. **Content Extraction**

4. **Human preference evaluation**

5. **Automated evaluation using state-of-the-art VLMs.**

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[2] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022.

[3] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

[4] Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. Kosmos-2.5: A multimodal literate model. 2023.

[5] Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, and Jean-Marc Ogier. Qalayout: Question answering layout based on multimodal attention for visual question answering on corporate document. In *Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings*, page 659–673, 2022.

[6] Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Ryan A. Rossi, and Franck Dernoncourt. Pdftriage: Question answering over long, structured documents, 2023.

[7] Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*, 2018.