RESEARCH PROPOSAL

Sai Rajeswar *Research scientist*

1 Research Proposal: Extracting Document Structure from Text-Intensive Images

1.1 Research Abstract

This project aims to utilize multimodal models to convert images, especially those rich in text and structured content, into structured code (E.g. Latex, Markup, Docx) format. The ability to accurately translate complex diagrams, tables, and mathematical equations from images into editable text could revolutionize the creation and editing process of scientific and other text-intensive documents. 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

Large Language Models (LLMs) have predominantly focused on processing textual information, leaving the realm of visual information largely unexplored. However, recent advancements in Large Multimodal Models (LMMs) have begun to bridge this gap [2]. LMMs integrate visual and textual data within a single Transformer-based model, enabling the model to learn and generate content based on both modalities [1, 5]. These models have shown potential in diverse applications, including understanding natural images and text images [3, 6, 1]. While the current focus of LMMs has been on natural images with lower resolutions, the exploration of text images, especially high-resolution and noisy document images, remains an area ripe for further research [8, 4, 9]. Leveraging large-scale multimodal pre-training for high-resolution text images represents a significant direction for future LMM research.

1.4 Research Objectives and Sub-Projects

The primary objective of this research is to develop and refine LMMs capable of accurately extracting document structure from text-intensive images. This overarching goal can be broken down into several sub-objectives or sub-projects:

1. **Image to LaTeX Conversion**: Develop an LMM capable of converting text-intensive images into LaTeX code, focusing on complex diagrams, tables, and mathematical equations.
2. **High-Resolution Text Image Understanding**: Investigate the application of LMMs to high-resolution text images, an area currently under-explored in the field.
3. **Large-Scale Multimodal Pre-Training**: Explore the benefits of large-scale multimodal pre-training 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.

2 Datasets

1. **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.
2. **ArXiv Documents:** ArXiv, a platform for open-access sharing of research, serves as another prominent data source, contributing approximately 20.9 million pages. We will scrape a significant chunk of data, which comprises PDF and LATEX source files, right from the official ArXiv repository.
3. **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.
4. **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.
5. **Web Screenshots:** Further, a subset of the mC4 web pages would be used and rendered into screenshots, capturing nearly 100 million or more pages
6. **DOCX files** widely used in research such as TableBank [7] and ReadingBank, The original DOCX files are converted into PDF files, with each page aligned to the corresponding markdown content span based on a heuristic method.
7. **LATEX documents** from arXiv are used to generate PDF files for texts with bounding boxes. We also convert the LATEX content into markdown texts, similar to the Nougat approach. We use LaTeXML to convert the LATEX code into the HTML sequence, which is then transformed into the markdown format.

3 Evaluation

1. **Text Recognition**:
   - **Character Accuracy (CER)**: Measures the percentage of incorrect characters in recognized text compared to ground truth.
     - Example: Ground truth: "Hello," Recognized text: "Helo," CER = 20%.
   - **Word Accuracy (WER)**: Measures the percentage of incorrectly recognized words in the text.
     - Example: Ground truth: "The cat is fast," Recognized text: "The dog is slow," WER = 66.67%.
2. **Image to Markdown Generation**:
   - **BLEU Score**: Measures similarity between generated and reference Markdown.

        - Example: Reference: " *Italic* and **bold**," Generated: " *Italic* and **bold**," BLEU = 100%.
    - **Semantic Evaluation**: Experts rate generated Markdown for correctness, coherence, and relevance.
        - Example: Experts rate generated Markdown as 4/5 for correctness, coherence, and relevance.

3. **Table Extraction**:
    - **Precision**: Ratio of correctly identified tables to total tables in the algorithm's output.
        - Example: Algorithm identifies 10 tables, of which 8 are correct, Precision = 80%.
    - **Recall**: Ratio of correctly identified tables to total actual tables in the ground truth dataset.
        - Example: There are 12 actual tables, and the algorithm finds 8 of them, Recall = 66.67%.

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] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[3] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. ICML'23. JMLR.org, 2023.

[4] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[5] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. MM '22. Association for Computing Machinery, 2022.

[6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[7] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: A benchmark dataset for table detection and recognition, 2019.

[8] Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *AAAI Conference on Artificial Intelligence*, 2021.

[9] 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.