Multimodal Language Modeling for Science and Financial Documents

Jaemin Cho

UNC Chapel Hill (Advisor: Mohit Bansal)

Abstract

Recent large language model-based AI agents have shown promising text understanding and generation capabilities and have been used in many applications. However, they are limited to only handling text data, while human experts deal with non-text data, such as reports, news, papers, social media, and blogs, where text, charts, diagrams, and tables are mixed. To address this, I propose developing a multimodal science/financial language model to browse and understand different non-text modalities, such as diagrams, charts, and tables. My research would allow existing language model-based systems, such as BloombergGPT, to handle various types of modalities and help human experts save time and make more intelligent decisions.

Collaborators: Mohit Bansal (UNC Chapel Hill)

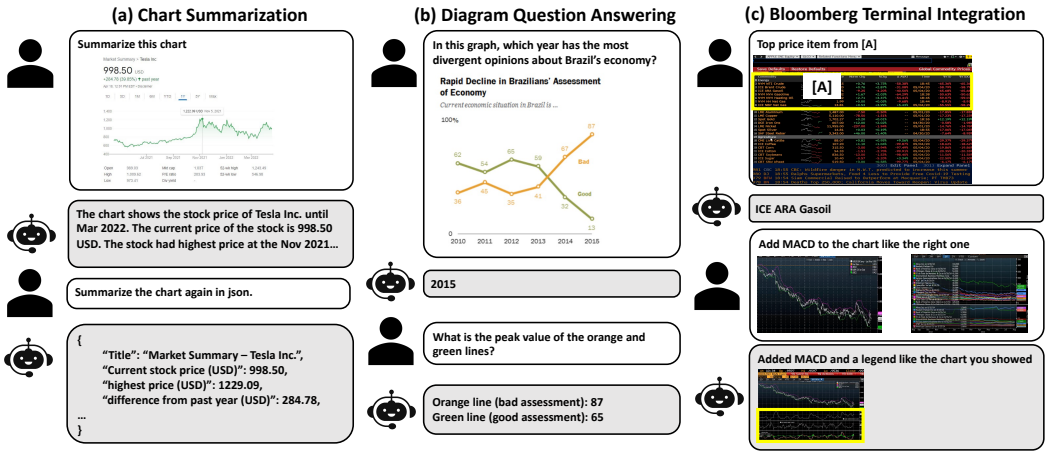

Figure 1: Example use cases of the proposed Multimodal Sci-Fi (Science/Financial) Language Model: **(a)** Summarize charts in open-ended text or program; **(b)** Answering open-ended questions about visual input; **(c)** Integration in domain-specific software, such as Bloomberg Terminal.

1 Introduction

Large language models (LLMs), such as ChatGPT and GPT4, have shown remarkable performance in understanding and generating open-ended text. As most LLMs were trained on general web-crawled corpora, the recently released BloombergGPT [Wu et al., 2023] demonstrates that training LLMs on a wide range of domain-specific datasets can yield significant improvements in target domain over general-purpose baselines. Although these domain-specific pretraining of LLMs show promising results, they are limited to understanding text data. However, in the science and finance domain, human experts deal with different sources of information beyond text, such as charts, diagrams, tables, etc.

A multimodal AI agent that browses and understands different non-text data types would help human experts save significant time. In addition, a multimodal agent that provides visual explanations that complement textual answers would help users understand the responses and share the results with co-workers. In Fig. 1, we illustrate several use cases, including chart summarization, diagram question answering, and integration into domain-specific software, such as Bloomberg Terminal.

In this proposal, we introduce two research projects to develop **a multimodal AI agent in Sci-Fi (science/financial) domain**: **(1) Multimodal Sci-Fi Dataset** to teach machine learning models science/financial tasks involved with heterogenous input/output, connected to my previous works on hierarchical multimodal retrieval and summarization dataset [Zala* et al., 2023]; **(2) Multimodal Sci-Fi Language Model** that can understand different data types and provides more informed and useful responses to human experts, following my previous works on unified multimodal language models [**Jaemin Cho** et al., 2021], parameter-efficient multimodal finetuning [Sung et al., 2022a,b], and reinforcement learning with multimodal rewards [**Jaemin Cho** et al., 2022]. These two projects complement each other to achieve the final goal of developing a multimodal science/financial AI agent.

2 Proposed Work

2.1 Multimodal Sci-Fi Dataset

We need a multimodal Sci-Fi corpus to train a multimodal language model where different data types complement each other. Existing LLMs are usually pre-

trained on combinations of general-domain (e.g., The Pile [Gao et al., 2020], C4 [Raffel et al., 2019], BookCorpus [Zhu et al., 2015]), and multimodal LMs are developed by being finetuned from these language models with image-text pairs or image-text interleaved corpus (e.g., M3W [Alayrac et al., 2022], MMC4 [Zhu et al., 2023]). While the multimodal LMs show interesting understanding performance on images with common objects, they cannot understand scientific/financial contents, such as charts and tables. Recently, Wu et al. [2023] introduced FinPile, a large-scale financial corpus that only consists of text data.

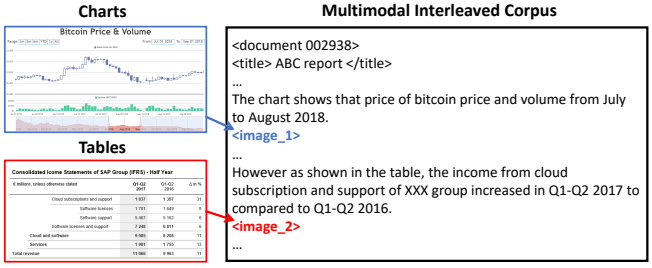

Figure 2: Multimodal Sci-Fi Dataset. Text and non-text data (e.g., charts, tables, diagrams) are interleaved.

Motivated by this, we propose to collect a novel multimodal Sci-Fi dataset. As shown in Fig. 2, different non-text contents, such as charts and tables, are interleaved within a text corpus. We use special tokens (`<image_1>`, `<image_2>` in Fig. 2) next to the sentences describing the content to point the original contents. A multimodal language model trained on this new corpus will have contextualized understanding of the text and different types of images and opens up new capabilities, including question answering, captioning, and summarization using non-text contents. To evaluate such models, we can set benchmarks for multimodal Sci-Fi tasks, including content retrieval, captioning, summarization, and hierarchical combination of these tasks, which is well connected to my previous work on hierarchical multimodal retrieval and summarization dataset Zala* et al. [2023].

2.2 Multimodal Sci-Fi LM

Although many previous AI applications have used architectures specialized in specific tasks and modalities, designing and training a new model for each new task often takes a long time. A general framework that can handle different modalities and tasks would enable quick prototype experiments with new ideas. To address this, we propose a multimodal Sci-Fi language model that can flexibly handle different tasks within a unified generative framework, following my previous work [**Jaemin Cho** et al., 2021]. This unified multimodal framework has been widely used by recent works in many research groups such as SimVLM/CoCa/Flamingo (Google Deepmind) [Wang et al., 2022, Yu et al., 2022, Alayrac et al., 2022], BLIP (Salesforce) [Li et al., 2022], and UDOP (Our lab and Microsoft) [Tang et al., 2023]. In Fig. 3, we illustrate the overall architecture of the proposed model.

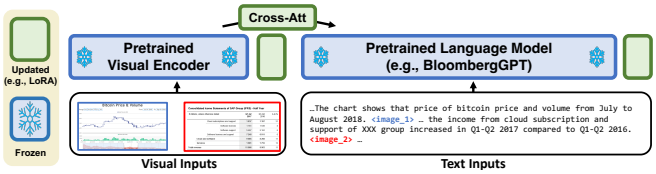

Figure 3: Multimodal Sci-Fi LM Architecture. We can bootstrap the training process with parameter-efficient finetuning of a small set of newly inserted parameters (colored in green) on top of existing pretrained visual encoder and language models (colored in blue).

Since training large visual encoders and language models from scratch is very expensive, we propose using parameter-efficient finetuning of pretrained visual encoders (e.g., CLIP [Radford et al., 2021]) and language models (e.g., BloombergGPT) by only tuning a small portion of newly inserted Adapter [Houlsby et al., 2019], LoRA [Hu et al., 2022], and cross-attention parameters [Alayrac et al., 2022]. This parameter-efficient training for multimodal language models has been initially introduced in my work [Sung et al., 2022a] and applied in recent work by other groups, such as BLIP-2 [Li et al., 2023] and Flamingo [Alayrac et al., 2022].

To further improve human alignment of responses, after first training the model using maximum likelihood estimation on the multimodal interleaved image-text dataset, we propose to further train the model with instruction tuning with multimodal rewards in reinforcement learning with human feedback (RLHF) [Ziegler et al., 2019]. While the existing RLHF methods for LMs typically use a reward based only on text inputs, we propose to experiment with using a multimodal reward conditioned on both visual and textual inputs for the tasks that involve visual editing/generation, extending the idea of my previous work that introduces the use of CLIP as a multimodal reward model for improving image captioning [**Jaemin Cho** et al., 2022].

3 Expected Results & Impact

Expected Results. The research outcomes would include: (1) A **dataset** to train multimodal science/financial machine learning models, where texts and non-text contents are interleaved. (2) A **multimodal science/financial language model framework**, based on a unified generative framework, parameter-efficient tuning, and reinforcement learning based on multimodal rewards. (3) A comprehensive **evaluation** of the capabilities of the new model.

Expected Impact. We believe that multimodal understanding and generation capabilities enhance the existing AI agents for the science/financial domain by opening up many new applications; researchers and analysts could summarize reports written in different formats without explicit text-parsing; software such as Bloomberg Terminals could be equipped with new functionalities taking visual input as well as text inputs.

Data, Software, and Ethics Policy. We plan to publish papers at top-tier conferences and open-source the public version of our models and code.

References

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. URL http://arxiv.org/abs/2204.14198.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. In *ICML*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, pages 1–26, 2022. URL http://arxiv.org/abs/2106.09685.

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. URL http://arxiv.org/abs/2201.12086.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023. URL http://arxiv.org/abs/2301.12597.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Jong Wook, Kim Chris, Hallacy Aditya, Ramesh Gabriel, Goh Sandhini, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. URL http://arxiv.org/abs/2103.00020.

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. 21:1–67, 2019. URL http://arxiv.org/abs/1910.10683.

Yi-Lin Sung, **Jaemin Cho**, and Mohit Bansal. VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks. In *CVPR*, 2022a.

Yi-Lin Sung, **Jaemin Cho**, and Mohit Bansal. LST: Ladder Side-Tuning for Parameter and Memory Efficient Transfer Learning. In *NeurIPS*, 2022b.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying Vision, Text, and Layout for Universal Document Processing. In *CVPR*, 2023. URL http://arxiv.org/abs/2212.02623.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying Vision-and-Language Tasks via Text Generation. In *ICML*, 2021.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained Image Captioning with CLIP Reward. In *Findings of NAACL (short)*, 2022.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *ICLR*, 2022. URL http://arxiv.org/abs/2108.10904.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A Large Language Model for Finance. 2023. URL http://arxiv.org/abs/2303.17564.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research*, 2022. URL http://arxiv.org/abs/2205.01917.

Abhay Zala*, **Jaemin Cho***, Satwik Kottur, Xilun Chen, Barlas Oğuz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In ***CVPR***, 2023.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved With Text. 2023. URL http://arxiv.org/abs/2304.06939.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences. 2019. URL `http://arxiv.org/abs/1909.08593`.