ReVL: Midterm Report

Authors
Amaad Martin
Wadih Pazos

1 Introduction

GUI agents aim to automate tasks on digital devices purely through natural language. Previous attempts at GUI agents varied from utilizing HTML content for web interactions to relying solely on images for actions. The main bottleneck in these attempts has been GUI grounding which is the task of locating screen elements from natural language. The goal of this research is to explore new formulations of the GUI grounding task to achieve state-of-the-art performance. For our data we will be using a set of screenshot, task, bounding box tuples gathered from web, and mobile data. To evaluate our results we plan to use a recently created evaluation benchmark, ScreenSpot, which was made specifically for the GUI Grounding task.

2 Background/Literature

For the GUI Agent problem, there has been progress with the rise of LLMs (Kim et al., 2023; Deng et al., 2023). There have also been attempts at only using images (Shaw et al., 2023). Now we see Visual Language Models that are being used for GUI agent tasks as well (Bai et al., 2023; Yan et al., 2023; Hong et al., 2023; Zhang et al., 2024). In addition, Recent publications have found some success in the area of GUI grounding (Cheng et al., 2024), and more work is being done in creating evaluation benchmarks for this specific task (Cheng et al., 2024). To improve on what has been done we are learning from the insight that Gui grounding is the bottleneck (Cheng et al., 2024) and will be trying to achieve state-of-the-art performance on the task to improve the ultimate problem of creating a GUI agent.

3 Methods/Model

As a baseline model, we designed a model that uses ResNet-50 to extract features from the input image and then uses BertTransformer to encode the natural language task. We then concatenate both embeddings and pass them through a linear layer to predict the partition of the image, the task resides in. The input to the model is simply the text instruction along with the image, and the output is a label from 1-10000, which represents a partition of the image, after we split it up into 100x100 partitions.

For training, we used a Cross Entropy Loss objective, and we used a subset of the training data that was used for See-Click. In addition, we evaluated the baseline model on the ScreenSpot benchmark.

4 Preliminary Results

5 Evaluation of preliminary work

6 Future Work

When thinking about how we as humans interact with computers we look at and focus on wherever we are clicking before we do. This project plans to introduce new formulations of the GUI grounding

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

problem involving focusing on specific regions of the input image to mirror this human behavior in the hopes of seeing improved performance. We will first try splitting the image up into several patches which will be upscaled to the input resolution of the VLM. Then we will try recursively splitting the image up using the model to choose which partition to look into. We will evaluate our final method using ScreenSpot and MiniWob.

7 Teammates and Work Division

March 11: Implement fine-tuning infrastructure
March 18th: Finish fine-tuning QwenVL and evaluating on ScreenSpot
March 25th: Finish formulation of Mixture of Images Model
April 1st: Finish implementation of Mixture of Images Model, train, and evaluate
April 8th: Finish formulation of Recursive Visual Language Model
April 15th: Finish implementation of Recursive Visual Language Model, train, and evaluate
April 22: Document all findings, write up final report

References

[1] Kim, G., Baldi, P., & McAleer, S. (2023). Language models can solve computer tasks. arXiv. https://arxiv.org/abs/2303.17491

[2] Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., & Su, Y. (2023). Mind2Web: Towards a generalist agent for the web. arXiv. https://arxiv.org/abs/2306.06070

[3] Shaw, P., Joshi, M., Cohan, J., Berant, J., Pasupat, P., Hu, H., Khandelwal, U., Lee, K., & Toutanova, K. (2023). From pixels to UI actions: Learning to follow instructions via graphical user interfaces. arXiv. https://arxiv.org/abs/2306.00245

[4] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv. https://arxiv.org/abs/2308.12966

[5] Yan, A., Yang, Z., Zhu, W., Lin, K., Li, L., Wang, J., Yang, J., Zhong, Y., McAuley, J., Gao, J., Liu, Z., & Wang, L. (2023). GPT-4V in Wonderland: Large multimodal models for zero-shot smartphone GUI navigation. arXiv. https://arxiv.org/abs/2311.07562

[6] Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Zhang, Y., Li, J., Xu, B., Dong, Y., Ding, M., & Tang, J. (2023). CogAgent: A visual language model for GUI agents. arXiv. https://arxiv.org/abs/2312.08914

[7] Cheng, K., Sun, Q., Chu, Y., Xu, F., Li, Y., Zhang, J., & Wu, Z. (2024). SeeClick: Harnessing GUI grounding for advanced visual GUI agents. arXiv. https://arxiv.org/abs/2401.10935

[8] Zhang, C., Li, L., He, S., Zhang, X., Qiao, B., Qin, S., Ma, M., Kang, Y., Lin, Q., Rajmohan, S., Zhang, D., & Zhang, Q. (2024). UFO: A UI-focused agent for Windows OS interaction. arXiv. https://arxiv.org/abs/2402.07939

[9] OpenAI. Various publications on LLMs and VLMs for digital interaction.

[10] Rabbit, Startup. "Hardware Solutions for Enhanced VLM Interaction." Internal Report, 2023.

[11] Imbue, Company. "Advancements in Natural Language Processing for GUI Navigation." Tech White Paper, 2023.

[12] Adept, Company. "Integrating VLMs for Desktop Environment Control." Research Findings, 2023.