PhenoGPT: Towards Language Interaction with Vision Models for Plant Phenotyping

Anonymous ECCV 2024 Submission

Paper ID #17

1 Introduction

Plant phenotyping greatly benefits from deep learning and computer vision, which enable precise quantitative measurement of plant traits [6,11,14]. However, interacting with vision models requires substantial computer science knowledge (*e.g.* understanding coding and data processing), which increases the entry barrier for scientists lacking such knowledge. The advancement of large language models (LLMs) [2, 12] has captured the attention of the natural science community due to their strong capability in natural language understanding and reasoning [8, 9, 13]. Nevertheless, the ability of LLMs (incl. multi-modal LLMs) to directly solve vision tasks (*e.g.* image classification and instance segmentation) remains questionable.

To facilitate simple interaction with computational models while maintaining accurate measurements for plant phenotyping, we present the prototype of PhenoGPT. PhenoGPT leverages an LLM to invoke the most appropriate pretrained vision models to address plant tasks specified by free text.

2 PhenoGPT

We define our **application scenario** as follows: the user sends a free text containing the specification of a computer vision task and the location of an image to the LLM. The LLM will then identify the correct vision model(s) to solve the task and return the results to the user.

Prompt-engineered LLM. We select GPT-3.5 Turbo as the base LLM. The key to enabling model calling is to set a system prompt that provides descriptions of the available models to the LLM and instructs it to output a JSON file based on the user prompt. The JSON file should contain the model name and the corresponding arguments needed to correctly invoke the model.

Vision Models. In this prototype, we provide the LLM access to two different vision models. The first is a Vision Transformer (ViT) [7] trained for cassava disease classification [1, 3], and the second is a Mask2Former [5] trained for leaf instance segmentation [4, 10].

Use Cases. In Fig. 1 (a) and (b), we show that PhenoGPT correctly performs leaf instance segmentation and disease classification. To illustrate the importance of using task specific vision models, we provided similar prompts directly to GPT-4o. We observed in Fig. 1 (c) and (d) that GPT-4o failed to solve the leaf instance segmentation task correctly, and produced ambiguous results for leaf disease classification (multiple possible diseases suggested).

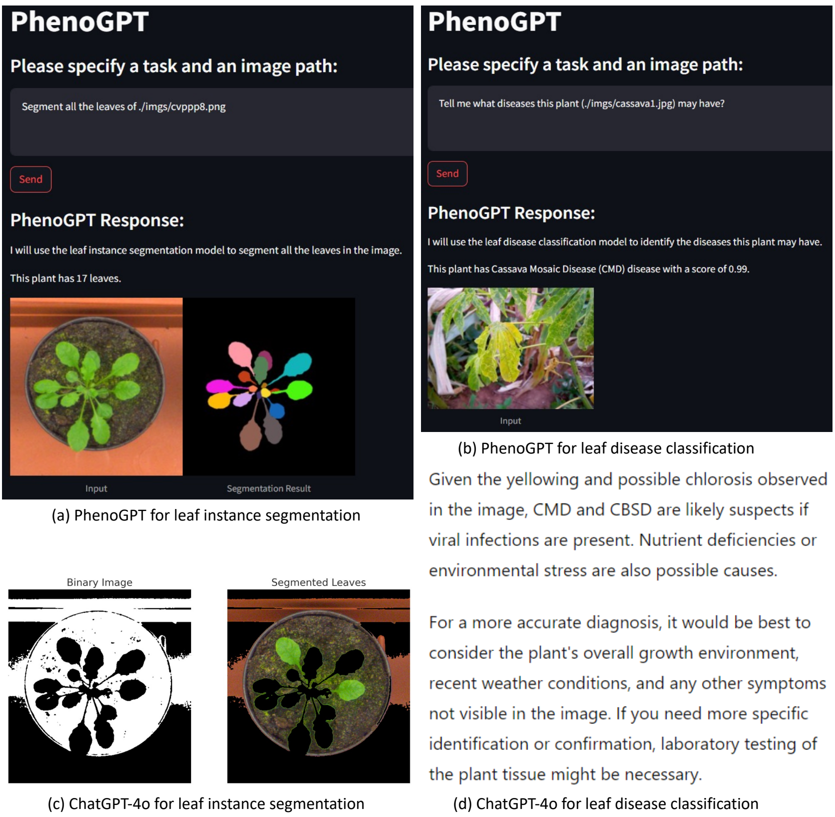

Fig. 1: Response comparison between PhenoGPT and ChatGPT-4o. (a)(c) Leaf instance segmentation. (b)(d) Leaf disease classification.

Conclusion. The prototype of PhenoGPT demonstrates the potential of combining LLMs with vision models to create a convenient natural language interaction interface while maintaining high accuracy in plant trait measurement. Our future work will focus on expanding the range of accessible vision models to enhance usability and accuracy in various plant phenotyping tasks.

References

1. Bell, J., Dee, H.M.: Aberystwyth leaf evaluation dataset [data set] (2016), http://doi.org/10.5281/zenodo.168158 1
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A.,

Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020), https://arxiv.org/abs/2005.14165 1

3. Chen, F., Giuffrida, M.V., Tsaftaris, S.A.: Adapting vision foundation models for plant phenotyping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 604–613 (2023) 1
4. Chen, F., Tsaftaris, S.A., Giuffrida, M.V.: Gmt: Guided mask transformer for leaf instance segmentation. arXiv preprint arXiv:2406.17109 (2024) 1
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022) 1
6. Dobrescu, A., Valerio Giuffrida, M., Tsaftaris, S.A.: Leveraging multiple datasets for deep leaf counting. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 2072–2079 (2017) 1
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 1
8. Luo, X., Rechardt, A., Sun, G., Nejad, K.K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A.O., Borghesani, V., Pashkov, A., Marinazzo, D., Nicholas, J., Salatiello, A., Sucholutsky, I., Minervini, P., Razavi, S., Rocca, R., Yusifov, E., Okalova, T., Gu, N., Ferianc, M., Khona, M., Patil, K.R., Lee, P.S., Mata, R., Myers, N.E., Bizley, J.K., Musslick, S., Bilgin, I.P., Niso, G., Ales, J.M., Gaebler, M., Murty, N.A.R., Loued-Khenissi, L., Behler, A., Hall, C.M., Dafflon, J., Bao, S.D., Love, B.C.: Large language models surpass human experts in predicting neuroscience results (2024), https://arxiv.org/abs/2403.03230 1
9. M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A.D., Schwaller, P.: Augmenting large language models with chemistry tools. Nature Machine Intelligence pp. 1–11 (2024) 1
10. Minervini, M., Fischbach, A., Scharr, H., Tsaftaris, S.A.: Finely-grained annotated datasets for image-based plant phenotyping. Pattern recognition letters **81**, 80–89 (2016) 1
11. Qi, C., Sandroni, M., Westergaard, J.C., Sundmark, E.H.R., Bagge, M., Alexandersson, E., Gao, J.: In-field classification of the asymptomatic biotrophic phase of potato late blight based on deep learning and proximal hyperspectral imaging. Computers and Electronics in Agriculture **205**, 107585 (2023) 1
12. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023), https://arxiv.org/abs/2302.13971 1
13. Yang, X., Gao, J., Xue, W., Alexandersson, E.: Pllama: An open-source large language model for plant science (2024), https://arxiv.org/abs/2401.01600 1
14. Yasrab, R., Atkinson, J.A., Wells, D.M., French, A.P., Pridmore, T.P., Pound, M.P.: Rootnav 2.0: Deep learning for automatic navigation of complex plant root architectures. GigaScience **8**(11), giz123 (2019) 1