Team: CrimsonNoir

1. **JithendraKatta-G34424060**
2. **NikhilReddyKodumuru-G27606239**

Literature Review

Our project InterAct utilizes an open vocabulary object detection model to identify objects based on user-provided text input, which can be in various formats such as questions or expressions.

Open vocabulary object detection refers to the ability to detect and recognize objects beyond a predefined set of categories, unlike traditional object detection models that are constrained to a fixed vocabulary of classes. The model recognizes the objects mentioned in the text and displays relevant visual representations by drawing bounding boxes around the detected objects. These open vocabulary models differ from conventional object detection models like Region-based Convolutional Neural Networks (RCNN, an early two-stage object detection model that first generates region proposals and then classifies and refines those proposals) and You Only Look Once (YOLO, a family of one-stage real-time object detection models that use a single neural network to predict bounding boxes and class probabilities directly from the input image) versions 1 to 9.

Several open vocabulary object detection models exist, such as ZSD-YOLO [1] (Zero-Shot Detected YOLO, an open-vocabulary extension of the YOLO model that aligns the detector with a language model for zero-shot detection), GLIP [2] (Grounded Language-Image Pre-training, a model pre-trained on image-text pairs for open-vocabulary object detection and phrase grounding), and Meta's DINO [4] (Detecting Instances by matching Normalized Overlays, a transformer-based object detection model trained with a normalized set loss). However, these models tend to be computationally expensive and resource-intensive, making them less suitable for real-time object detection on typical devices with limited computational resources.

The latest development in this domain is the YOLO-World model [5], a lightweight open vocabulary detector capable of real-time performance. YOLO-World achieves this by combining a traditional YOLO backbone with the Contrastive Language-Image Pre-training (CLIP, a neural network trained on a large dataset of image-text pairs to learn visual-linguistic representations, enabling zero-shot transfer to various vision-language tasks) model and employing techniques like re-parameterizable Vision-Language Parallel Attention Networks (a technique used in YOLO-World to connect vision and language features by re-parameterizing the attention maps to align the two modalities) to bridge the gap between vision and language features. Additionally, it incorporates an open vocabulary region-text contrastive pre-training scheme, which allows the model to learn from large-scale image-text data and generalize to unseen object categories.

Interestingly, some of the data used to train the YOLO-World model is generated by the GLIPv2 model, reducing the need for manual annotation. This approach is known as model distillation in the field of machine learning, where knowledge from a larger, more complex model (in this case, GLIPv2) is transferred to a smaller, more efficient model (YOLO-World) through the generation of synthetic data or intermediate representations.

While models like BLIP [3](Bidirectional Language-Image Pre-training, a model that can understand and generate natural language descriptions of images) and Instruct BLIP (an extension of BLIP that can follow natural language instructions for image understanding and generation tasks) are designed for image understanding and answering text-based questions, they lack the capability to draw precise bounding boxes around objects, a task at which YOLO-World excels. YOLO-World was trained on the COCO (Common Objects in Context, a large dataset for object detection and segmentation tasks, containing over 300,000 images with annotations for 80 object categories) and GQA [6] (Grounded Question Answering, a dataset containing over 20 million images and questions designed to evaluate visual reasoning and grounding abilities) datasets, equipping it with knowledge for image understanding and drawing accurate bounding boxes based on annotations.

[1] Xie, Johnathan, and Shuai Zheng. "Zero-shot object detection through vision-language embedding alignment." *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, Nov. 2022, https://doi.org/10.1109/icdmw58026.2022.00121.

[2] Zhang, H., Zhang, P., Hu, X., Chen, Y., Li, L. H., Dai, X., Wang, L., Yuan, L., Hwang, J., & Gao, J. (2022). GLIPv2: Unifying Localization and Vision-Language Understanding. *ArXiv*. /abs/2206.05836

[3] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *ArXiv*. /abs/2201.12086

[4] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., . . . Bojanowski, P. (2023). DINOv2: Learning Robust Visual Features without Supervision. *ArXiv*. /abs/2304.07193

[5] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., & Shan, Y. (2024). YOLO-World: Real-Time Open-Vocabulary Object Detection. *ArXiv*. /abs/2401.17270

[6] Hudson, D. A., & Manning, C. D. (2019). GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. *ArXiv*. /abs/1902.09506