Creating a Visual Explainable AI Interface with Grounded-Segment-Anything

An Assignment Report for the evaluation of CMPSC 291I

Master of Science

by

Alisetti Sai Vamsi
Perm Number - A1U0061

Department of Computer Science
Santa Barbara, CA - 93106
October 2024

Contents

1 Evaluation of Interface

The interface is designed to be simple and light. The interface is divided into three panels. First panel lets you upload the image, and input the prompt. It additionally provides the strength of segmentation and text on the output to tweak and play with. Once the run segmentation button is hit, the second panel displays the segmented output, and the corresponding segmentation masks. Upon hitting the explain segmentation button, the third panel displays the activation maps from the attention layers of the SAM model to see what the model is focusing on while making its decisions by overlaying these activation maps on the original image.

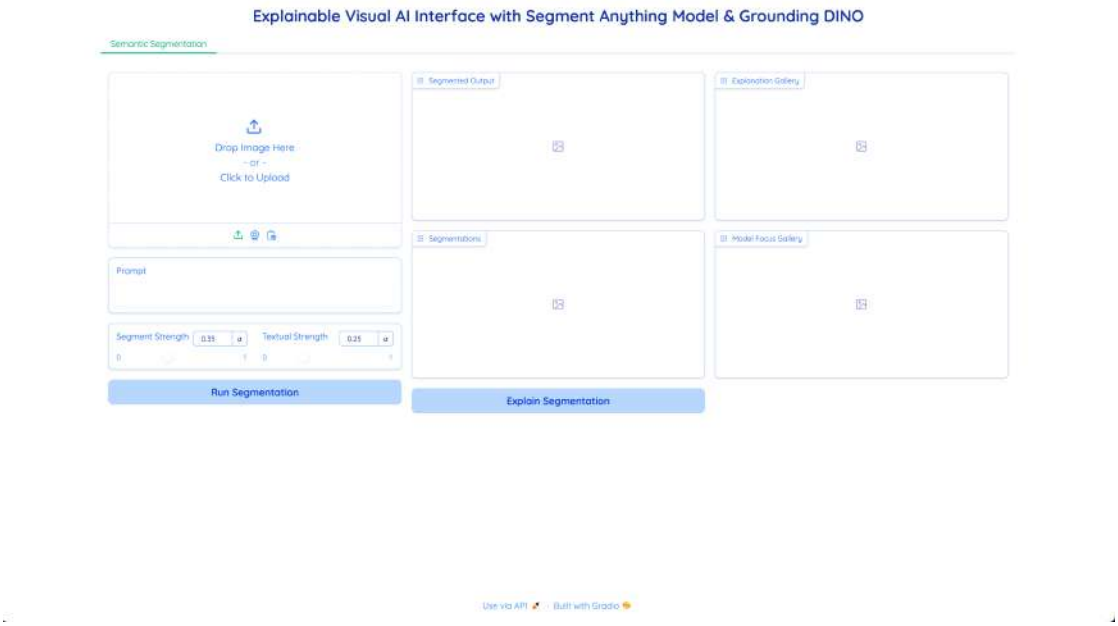
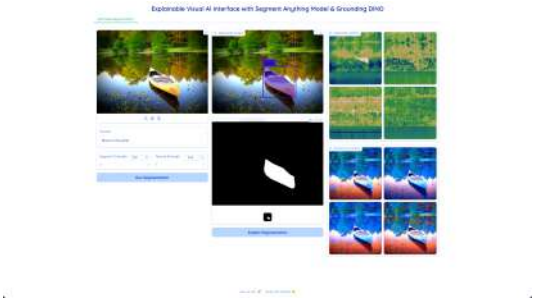


Figure 1: Visual Explainable Interface

2 Strengths and Limitations of Visual Explainable AI

Visual Explainable AI (XAI) provides significant advantages in fields where transparency is essential. For instance, in healthcare, heatmaps can assist radiologists by highlighting regions in an MRI scan that an AI model identifies as indicative of a tumor, making the AI's decision more transparent and verifiable. This fosters trust, as experts can corroborate the model's findings. However, visual XAI also has limitations. In autonomous driving, a model may indicate focus on specific road areas but may not clarify why certain road features influenced its decisions, introducing ambiguity and potential safety concerns. Additionally, creating visual explanations can be computationally demanding, which may restrict their usability in real-time applications like traffic monitoring or emergency response. Thus, while visual XAI enhances interpretability, relying solely on these



(a) Query: Boat on the water (Able to detect things)

(b) Query: Two lions laying down (Able to count objects)

(c) Query: Black Horses (Unable to identify colors)

(d) Query: Female adults running (Able to detect gender)

Figure 2: Examples

methods can sometimes lead to oversimplified or ambiguous conclusions, underscoring the need for a holistic approach to model interpretation.

The use of attention maps in Visual Explainable AI offers an in-depth look at how models process and interpret visual data. In medical applications, segmentation can isolate specific regions in a scan—such as tumors or lesions, giving clinicians clear, interpretable insights into areas of interest. Attention maps further complement this by showing where the model's focus lies during decision-making. This layered approach proves especially useful in fields like autonomous driving, where attention maps reveal which environmental elements, such as pedestrians, road signs, or obstacles, the model prioritized, ensuring that critical areas receive appropriate attention for safety.

Nevertheless, these methods also have drawbacks. Attention maps may highlight regions without fully explaining why they are significant, which can lead to interpretive ambiguity. To ensure visual explanations are both accurate and meaningful, these techniques demand rigorous training and validation, along with diverse, high-quality data and careful model tuning, to avoid potential misinterpretations.

(a) Query: Kids at the dinner table (Able to detect age)

(b) Query: Person next to the adult wearing pink shirt (Able to detect spatial relations)

(c) Query: Walking (Able to detect actions)

(d) Query: Sun and the clouds (Able to detect multiple diverse segments)

Figure 3: Examples

3 Potential Applications and Ethical Considerations

Visual XAI has broad potential in fields like finance, healthcare, and criminal justice. For instance, in financial services, visual explanations can illustrate which features of a client's profile influenced a loan approval or rejection, helping clients and regulators understand the basis for decisions. In criminal justice, AI models used for risk assessment can display factors influencing parole decisions, potentially reducing biases. However, ethical concerns arise when visual XAI unintentionally highlights biased or irrelevant factors, such as demographics, perpetuating inequality. Misinterpretation of visual outputs also poses ethical risks; for example, over-reliance on AI in healthcare or hiring decisions could reinforce systemic biases. Responsible deployment of visual XAI requires careful monitoring, transparency, and fairness assessments to mitigate these ethical risks and promote accountable AI applications.

In particular this interface could be used in medical imaging. For instance, this tool could assist radiologists by segmenting specific regions-of-interest, like tumors or inflamed areas based on descriptive prompts and highlighting relevant regions through attention maps. This capability would support diagnostic accuracy by clarifying how

(a) Query: Elephants in the background (Able to detect objects in blurry pictures)

(b) Query: Green leafy vegetation (Unable to detect depth of field)

and why the model identifies certain features. Additionally, this could be utilized in wildlife conservation, where the interface could be used to analyze drone or satellite images by isolating specific animal species or environmental features, aiding researchers in tracking biodiversity and monitoring ecosystems. Moreover, in industrial safety, the interface could help detect and segment hazardous equipment or unsafe conditions within factory images, allowing for real-time safety assessments and risk management.

However, ethical considerations must be addressed. In healthcare, for example, overreliance on automated explanations without human validation could lead to incorrect interpretations and potentially harm patients. Similarly, in wildlife and environmental monitoring, improper segmentation or explanation biases could misrepresent ecological data, influencing misguided conservation policies. Transparency around how attention maps are generated and interpreted will help users rely on this tool responsibly, promoting its ethical application across fields.

4 Suggestions for Improving the Model & Interface

The biggest improvement would be to provide textual explanations of the attention maps, which I could not do because of the lack of an LLM that could generate textual

explanations given the attention maps and the original image. This could have been a massive improvement on the interface because as of now, it is just showing the attention maps and leaving the end-user pondered about what exactly they signify. Additionally there could be more UI improvements, to provide more control to the user. Some more hyper parameters from the Grounding DINO model and the SAM model could have been included to provide a better contextualized control to the user.

Another improvement could be to select the segmentation masks individually and get an explanation about each segmentation mask with respect to the original image and its attention maps. Furthermore I have tried other variants of explainability including LIME, & GradCAM which were significantly harder to integrate because of the large number of parameters these models hold, and because of the transformer nature of these model architecture. Especially GradCAM is more suited for CNN-like models, it was much harder to adapt it to SAM and Grounding DINO, which is super interesting to make a custom-implementation out of, but in the interest of time, I had to find other alternatives.

Additionally, for this assignment, I could not finish the part of generating textual description of the output because of the pricing of the ChatGPT API. I have written the code however to prompt the ChatGPT API to compare the attention maps and provide a detailed analysis on the contextual features. This is commented in the colab notebook.