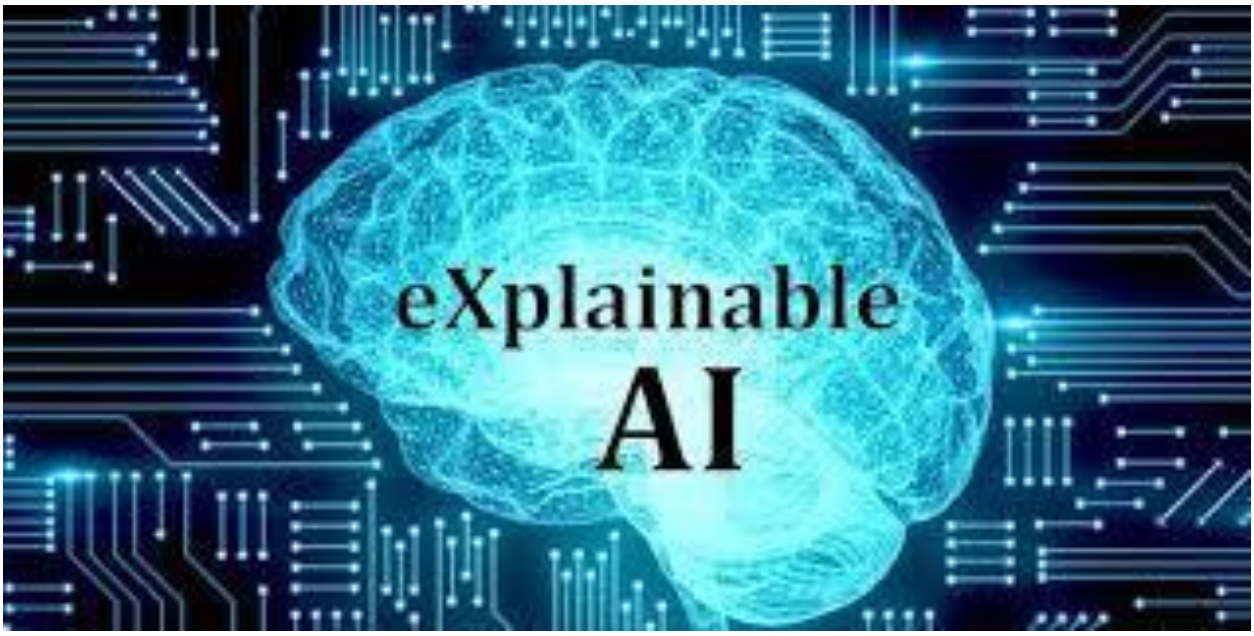
DAI COURSE PROJECT

EXPLAINABILITY FOR INTERPRETING CLIP MODELS



OBJECTIVE:

The aim of this project is to provide explainability for the CLIP (Contrastive Language-Image Pre-training) model developed by OpenAI and understand how the CLIP model makes predictions by generating visualizations and saliency maps that highlight the important regions and features in the input data and deployment of the project at Hugging face.

INTRODUCTION & MOTIVATION:

The CLIP is a powerful deep learning model that learns joint representations of images and text.

As machine learning models become more sophisticated, there is an increasing need for explainability to understand how they make predictions. Explainability in CLIP models refers to the ability to interpret and understand the model's decision-making process, shedding light on the important regions and features in the input data that contribute to its predictions. By providing explainability, CLIP models can enhance transparency, trust, and enable users to gain insights into the inner workings of the model, facilitating better decision-making and model understanding.

PROPOSED METHODOLOGY:

The proposed methodology for achieving explainability in the CLIP model is inspired by the paper "Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers" by Hila Chefer, Shir Gur, and Lior Wolf.

Here is the briefs of methodology described in paper.

- The method uses attention layers of a Transformer-based architecture to generate relevancy maps for interactions between input modalities.
- Relevancy maps are constructed for each interaction, including self-attention interactions for text and image tokens, and multi-modal attention interactions.
- The relevancy maps are updated by a forward pass on the attention layers using update rules that modify the maps impacted by the mixture of token embeddings.
- The method calculates the relevancy maps by accumulating each layer's contribution to the aggregated relevancies.
- For self-attention layers, the update rules add the attention map to the relevant aggregated relevancy scores, taking into account previous contextualization.
- The method uses gradients to average across attention heads and removes negative contributions before averaging.
- The final attention map is defined as the mean across the heads dimension.
- The method can be generalized to address more than two modalities and can be applied to any Transformer-based architecture.

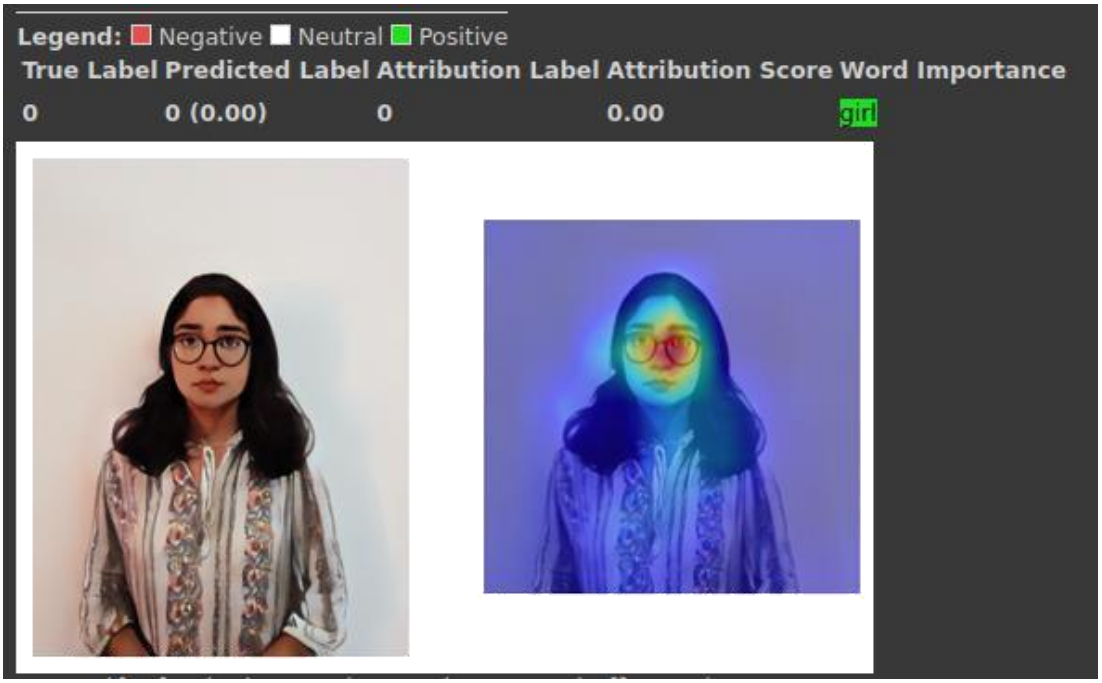
WORK DONE:

- The explainability technique for the CLIP model was implemented based on the research paper "Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers."

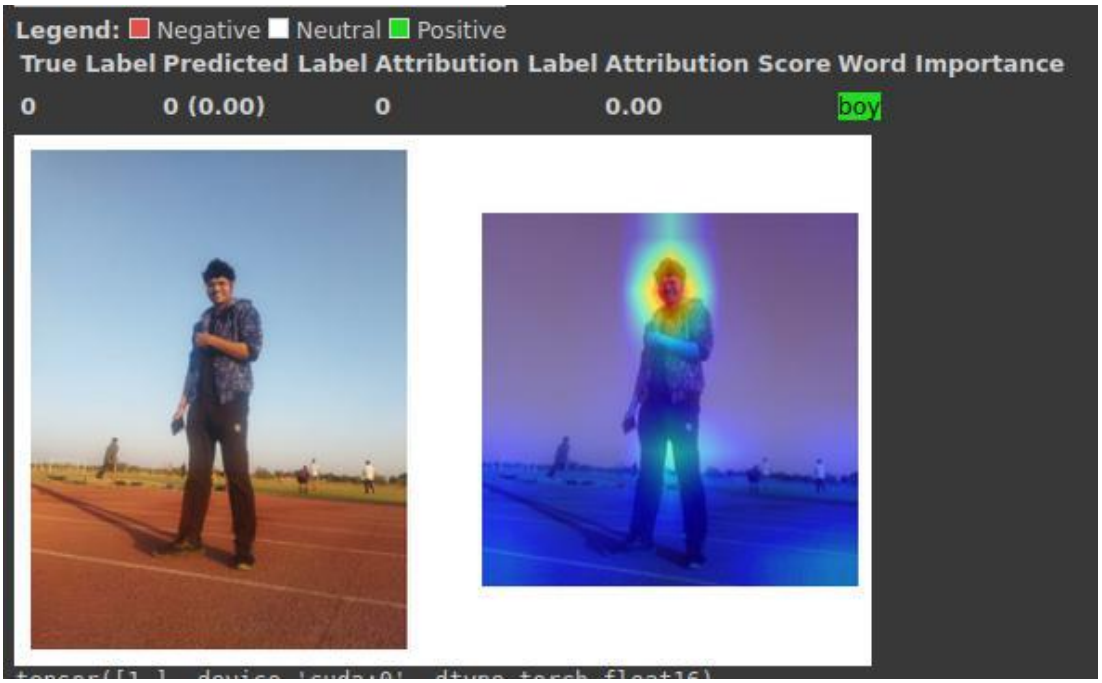
- The methodology and insights provided in the paper were studied to understand how attention-based explainability can be applied to CLIP.
- The official GitHub code provided by the authors was also examined to understand the codeflow and implementation details.
- The CLIP model was chosen for explainability due to its attention mechanism, which allows it to focus on relevant regions and features in images and text.
- The attention mechanism in CLIP makes it suitable for generating attention visualizations and saliency maps, providing interpretability and insights into the model's decision-making process.
- The implementation of the explainability technique for CLIP is available on a GitHub repository at **https://github.com/himanshi-2602/CLIP-Explainability**.
- The repository includes the main code for the implementation, and the results can be observed by running the provided Jupyter notebook.
- The project is also deployed on Hugging Face at CLIPGroundingExplainabilityDemo - a Hugging Face Space by harsh001, making it easily accessible for users to utilize the explainability method.

RESULTS:

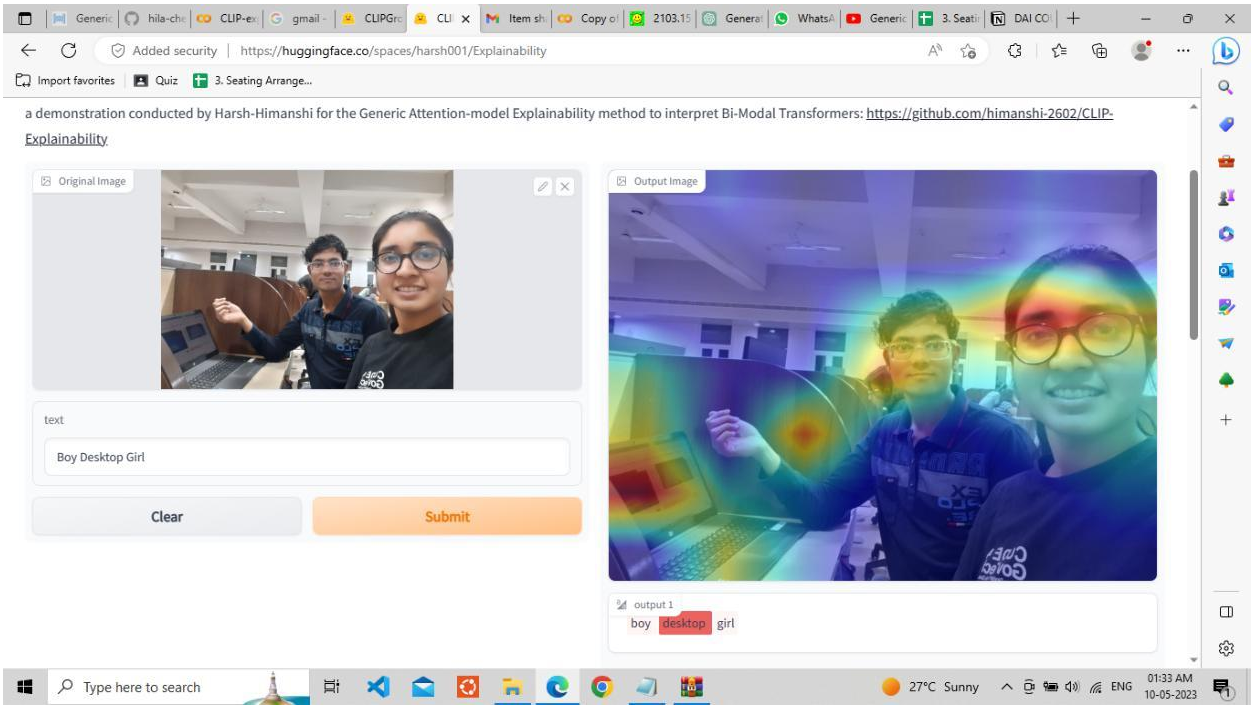
Here you can see the when we provided the model a picture of Himanshi and “girl” as text input. So here you can see that the model gave the output that highlights those areas of the image that is responsible for classifying this image as girl. Hence serving the purpose of explainability.



Here you can see the when we provided the model a picture of Harsh and “boy” as text input. So here you can see that the model gave the output that highlights those areas of the image that is responsible for classifying this image as boy. Hence serving the purpose of explainability.



Deployed Model



Here you can see we have deployed the model. Here we have uploaded our image of doing DAI Project in the left side and given a text prompt in the text field and submitted and here you can see we got the result at the right side which we can see the model has highlighted those features corresponding to boy, girl and desktop.

PROJECT :

https://github.com/himanshi-2602/CLIP-Explainability

CLIPGroundingExplainabilityDemo - a Hugging Face Space by harsh001

Discover amazing ML apps made by the community

https://huggingface.co/spaces/harsh001/Explainability

To verify the results download and run the below collab**:**

https://github.com/himanshi-2602/CLIP-Explainability/blob/main/CLIP_explainability.ipynb

---

RFERENCES:

https://github.com/hila-chefer/Transformer-MM-Explainability

https://arxiv.org/pdf/2103.15679v1.pdf