Final Project Progress Report 1

Team name: ***MapCap***
TA name: ***Srinath Sridhar***

Note: when submitting this document to Gradescope, make sure to add all other team members to the submission. This can be done on the submission page after uploading.

Progress Report

Restate the goal of the project succinctly. The goal of our project is to augment attribution methods for CLIP zero-shot image classification using CLIPCap. This is so that we are able to understand why a model classifies something the way it does by making attribution maps more interpretable through the use language models. The language model will be used to describe the regions of images that an attribution method highlights across an entire class. We will then summarize all of the descriptions such that a human will be able to quickly understand how CLIP is making its decisions. Afterward, we will analyze trends between classification performance and properties of the patch caption summaries.

What has the team collectively accomplished? So far, our team has collectively made significant progress towards our goal. We were able to get access to our required data by taking a partition of TinyImageNet (5 classes, 10 images per class) so that we could run our pre-trained model on a smaller sample size. Additionally, we ran inference over CLIP for zero-shot image classification over this partition (See Table). We experimented with gradient-based saliency maps and found that ViT feature extraction renders them difficult to interpret (See Figure 1). From there, we explored more advanced attribution methods, settling on an attention-based attribution method presented in *Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers*. We then ran this method on two images that were misclassified in the zero shot setting (see Figure 2). Finally, we ran CLIPCap on these same two images, as well as a third image that was correctly classified (See Figure 3) Both the CLIPCap and the Attention explainability method papers provide Google Colaboratory notebooks, which we leverage.

What individual tasks have been accomplished? In terms of individual tasks that have been accomplished, one team member created a script that will partition Tiny Imagenet while another member ran inference on the data using the CLIP model. Finally, the last member created and visualized the saliency map for each image and generated initial data using CLIPCap and the attention-based attribution method.

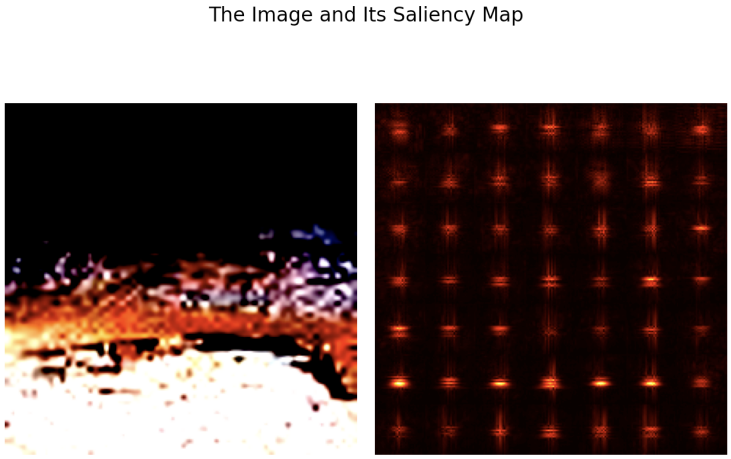

Figure 1: TinyImageNet image of a slug, and it's corresponding gradient based saliency map. Note how the center of every ViT patch is considered salient.

Class	Accuracy
Slug	90%
Lawn Mower	100%
Pill Bottle	90%
Goldfish	100%
Birdhouse	100%

What are the current tasks? Our current task is to implement segmentation over the heatmap generated by the attention-based attribution method, generate image representations for each segment, and feed them into CLIPCap.

What tasks remain undefined? We currently do not have a good way to quantify our natural-language explainability method's usefulness.

What are the next steps? After the current task is completed, we must implement unsupervised extractive summarization to create a simple description of the features used by the CLIP. Finally, we must increase our dataset size, and rewrite the CLIPCap and Attention Attribution method Colaboratory notebooks to run over our sequences of images.

Are you missing resources? Data, compute, skills? We have all resources that we need.

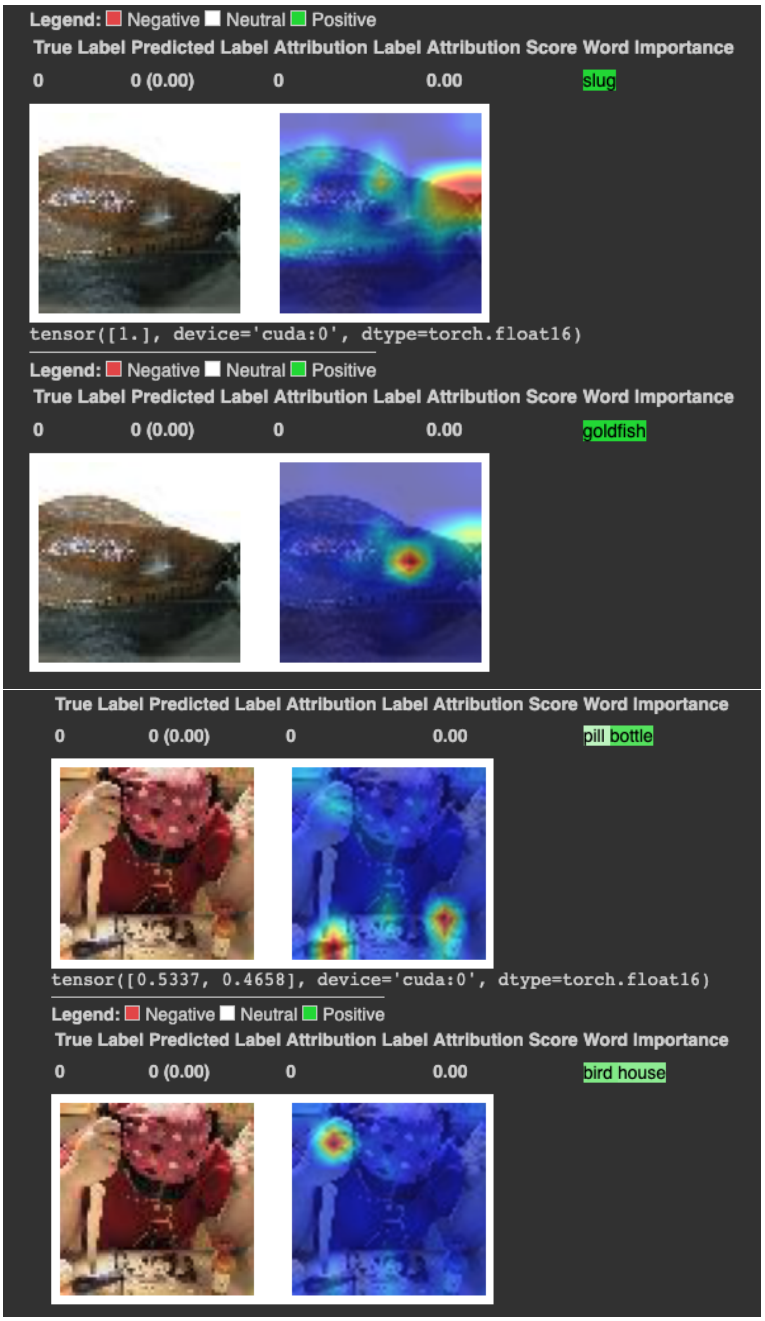

Figure 2: This attention-based attribution method computes an attribution map based on an image and the corresponding text. (Top) Image of a slug that was misclassified as a goldfish. Attribution map generated for both the correct class and the incorrect class. (Bottom) Image corresponding to the label “Pill Bottle” that was misclassified as a bird house.

Figure 3: CLIPCap captions for three images. (Top) Slug, misclassified as goldfish. (Middle) Pill Bottle, misclassified as birdhouse. (Bottom) Birdhouse, correctly classified.