Milestone 3 Writeup

Our first extension implements a watermark detection system using a Vision Transformer (ViT) classifier to classify images as watermarked or non-watermarked. It uses a StegaStamp encoder to initially embed watermarks into images and is designed to work with the COCO Captions dataset. The ViT then takes the dataset of both watermarked and unwatermarked images and classifies them. Some of the practical applications of our extension include protecting IP with the proper use of AI. One specific application would be to watermark all images that are published by an author. If someone were to use these images, even for training in a diffusion model (https://arxiv.org/abs/2303.10137), it is likely that it could be detected using a watermark classifier.

The watermark detection model had strong performance using a Vision Transformer (ViT) classifier trained over 10 epochs using Cross-Entropy Loss. The evaluation metrics for the binary classification on 250 validation images is as follows: precision of 91.04%, a recall of 97.60%, an F1-score of 94.21%, and an overall accuracy of 93.99%. We used an AdamW optimizer and a learning rate of 10^-4 during training to train on a 1000 images.

The model specifically consisted of 2 convolutional layers. Conventionally, ViTs process non-overlapping patches of the input image with a single convolution and then pass the output to a transformer. However, we noticed that this doesn’t capture the fine detail of the embedded watermark. Therefore, we substituted a single convolution for 2 convolutions that produces 64 output channels (much more than the original number of output channels). This allowed our model to approximately converge in 10 epochs.

Some next steps are to actually decode a signature from an image using a transformer. For this, we would likely need to have both an encoder and decoder ViT that tries to embed and extract a signature from the image. We would also like to start introducing some noise to the images in the hopes that the model/watermark is resistant to image perturbations. If this works, we hope to answer additional questions, such as: Can a ViT learn to do both image classification/captioning and watermark classification/decoding. This is similar to how modern language models can be instruction tuned on a variety of tasks.