Does Explainability Help in Improving ViTs?

Nagacharan Vemula
Arizona State University

I. INTRODUCTION

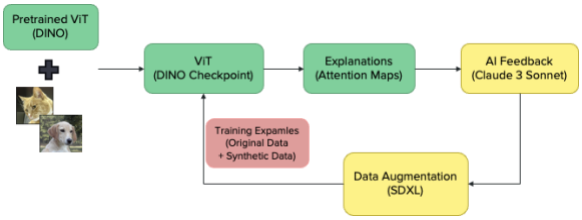
Vision transformers (ViTs) have emerged as a powerful architecture for image classification tasks, demonstrating remarkable performance gains over traditional convolutional neural networks. However, despite their success, ViTs may still encounter challenges in accurately classifying certain examples. This limitation highlights the need for strategies to enhance model robustness and improve performance further.

In this research, we explore the integration of explainability techniques, such as attention maps, with AI-driven feedback loops to identify model weaknesses and generate targeted training data. Explainability methods provide valuable insights into a model's decision-making process, enabling the identification of areas where the model struggles or exhibits biases. By leveraging these insights, we can develop a focused approach to address these weaknesses and improve model performance. Our proposed solution aims to leverage explainability data from ViTs [1,2] to guide the generation of synthetic training examples through AI-driven feedback loops. This innovative approach combines the power of explainability techniques with the ability of AI assistants to analyze and synthesize new data examples tailored to the model's needs. While previous research has explored explainability in vision transformers and data augmentation independently, our work introduces a novel contribution by integrating AI-driven feedback loops for targeted synthetic data creation.

The Project is divided into two stages:

Stage1
During the stage1, we employed a pre-trained vision transformer called ViT DINO [3]. This model was fine-tuned on the training set of the Cats-vs-Dogs dataset to establish a benchmark for performance evaluation. Then, we generated attention maps for all the misclassified samples from the validation set. These attention maps, which highlight the regions of the input images that the model focused on during prediction, will be utilized in Stage 2 to guide the process of enhancing the performance of the ViT.

Stage2
In Stage 2, we leveraged the misclassified data samples to gain insights into the model's focus areas on the input images through an AI assistant. Specifically, we employed Claude, an AI assistant created by Anthropic, to analyze the attention maps of the misclassified samples and provide feedback. Using another AI assistant, SDXL base, we synthesized new training images tailored to address the model's weaknesses identified through the attention map analysis. By augmenting the training data with these synthetic images, we produced a refined model.



II. EXPLANATION OF THE SOLUTION

A. Claude's Feedback on Attention maps

The Attention maps corresponding to misclassified data samples with the fine-tuned ViT model are generated. The attention maps generated for the misclassified samples offer valuable insights into the specific regions of the input images that the vision transformer model concentrated on during the classification process. To gain a deeper understanding of the model's behavior and decision-making patterns, these attention maps were analyzed using Claude 3 Sonnet. Claude's analysis help uncover potential patterns or deficiencies in the way the model processed and focused on different features or areas of the images, which could contribute to its misclassifications. This analysis provided crucial guidance for identifying and addressing the model's weaknesses.

The prompt shown in Figure 1 directs Claude to examine an input image along with its corresponding

attention map produced by a Vision Transformer (ViT) model trained for the cats vs dogs classification task. Additionally, Claude is informed about the specific class ({class_reflag}) that the model incorrectly predicted for the given image. The goal is for Claude to analyze the attention map and identify potential issues or limitations that may have contributed to the model's misclassification of the image as the specified class.

```
f"This image has the {class_flag}/{class_flag}s as the original image on the left and
theattention  map for that on  the right,  generated by a ViT  model for cats vs dogs
classification  task.  The model incorrectly  classified the image as {class_reflag}.
Analyze where the model is  focusing  on  and explain  what difficulties the original
image  has that made  the model misclassify this as {class_reflag}. Analyze the color
and structure of {class_flag}. Be precise about the location of attention."
```

Figure 1

Now, we used the analysis or feedback given by Claude and asked it again to generate a prompt that should be given to the SDXL [4] model to generate a synthetic image.

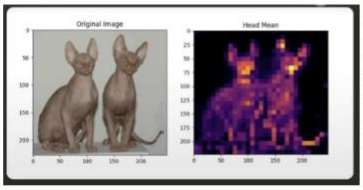
The prompt shown in Figure 2 outlines the goal of improving the ViT model's classification accuracy for the {class_flag} class, which represents the actual pet present in the image (the ground truth label). To achieve this, the plan is to further train the ViT model using additional synthetic images generated by an SDXL model. Claude's task is to formulate a prompt that can be provided to the SDXL model, instructing it to generate a single image that incorporates challenges or difficulties like those observed in the misclassified image. By generating such synthetic images that mimic the challenging scenarios, the aim is to facilitate the ViT model's learning and enhance its ability to handle complex cases effectively during the classification task.

```
f"{previous_chat}. Our objective is to improve our ViT model to classify the image as
{class_flag} by fine tuning  with more images  generated by a  SDXL model. We believe
this can  improve the ViT's focus on image. So, Can you only give a prompt(word limit
100) that  should be given to SDXL to  generate  single  image  that would improve in
finetuning the ViT.  Ensure the generated  image contains  challenges or difficulties
similar to  those present in the wrongly classified image.  The color of {class_flag}
in  original image is an important factor when genrating the prompt. This will enable
the model  to learn to classify challenging  scenarios effectively.  Only include the
prompt without any other text"
```

Figure 2

B. Generation of Synthetic Images using SDXL

The prompts generated using Claude target specific areas where the model may be struggling or misclassifying samples. SDXL, the base model, is employed to generate five synthetic images for each misclassified data sample, based on Claude's prompts. These synthetic images aim to provide additional training data to address the identified weaknesses of the model. A sample prompt generated by Claude that is given to SDXL is shown in Figure 3.



Generated Prompt:
"Hairless Sphynx cats in profile view, emphasizing distinct facial features, large ears, and wrinkled skin texture. Warm brown color matching the original cats' tone. Unique body shape and posture that highlights similarities and differences compared to dogs."

Figure 3

The synthetic images generated for the prompt shown in Figure 3 is shown in Figure 4. As you can see, the images generated by SDXL are quite close in many ways to the original image. The color, texture and body structure of the cats are same.

Figure 4

We believe these kinds of synthetic images will help improve the model's performance in its weaker areas when used in further training the model.

C. Enhancing the ViT using synthetic images

The model is further trained using the augmented dataset, consisting of the original training data and the newly generated synthetic images. The performance of both the baseline model and the data-augmented or refined model is evaluated using appropriate metrics, such as classification accuracy, precision, recall, and F1 score. This evaluation helps assess the effectiveness of leveraging explainability data and synthetic data augmentation in improving model performance.

III. RESULTS

We evaluated the effectiveness of our novel approach integrating AI-driven feedback loops with explainability data from Vision Transformers (ViTs) to enhance image classification model performance. Here, we present and discuss the experimental outcomes for various training configurations.

Table1.1

Model	Accuracy
Baseline	0.953
Fine-tuning (Generated Images Only)	0.927
Fine-tuning (Generated Images + Training Data)	**0.978**
LoRA Fine-tuning (Generated Images + Training Data)	0.934

Baseline: The baseline model, serving as the standard performance indicator of our image classification system without any synthetic image data augmentation, achieved an accuracy of **0.953**. This high baseline performance sets a robust foundation for assessing the enhancements brought by our novel methods.

Generated Images Only: Here, the model was further trained using only synthetic images generated from the explainability data. The accuracy observed was 0.927, slightly lower than the baseline. This result suggests that while the generated images provide valuable insights, they may not capture the full complexity or variability of real-world data when used in isolation.

LoRA (Generated Images + Training Data): Applying Low-Rank Adaptation (LoRA) to the model trained on both generated images and original training data resulted in an accuracy of 0.934 which is a slight decrease in overall accuracy compared to the baseline.

(Generated Images + Training Data): This experiment involved fine-tuning the model using a combination of synthetic images and the original training data, leading to a significant improvement with an accuracy of **0.978**. This indicates that the synthetic images, when utilized alongside real data, effectively enhance the model's ability to generalize, address gaps or correct misrepresentations in the training dataset.

This difference of **2.5%** in test accuracy, evaluated on a test set comprising 15,000 data samples, demonstrates the significant improvement offered by the refined model, over the baseline approach.

IV. CONTRIBUTION

In the stage 1, I contributed in the fine-tuning part of the pre-trained ViT DINO on cats and dogs image data. I also contributed to filtering out the misclassified data samples and generating attention maps.

In stage 2, I worked on testing the SDXL on images generation and I also integrated the pipeline of Claude 3 Sonnet and SDXL which helps generate five synthetic images per a misclassified sample. Finally, I involved in evaluation and comparison of both the baseline and refined model's accuracy on the test data and plotting the graphs related to test, train and validation accuracy visualizations based on the evaluation.

V. LESSONS LEARNED

During the project, I have expanded my knowledge on working with ViTs, VLMs and XAI techniques. Following are the important learnings of mine,

- Learnt how the training of a ViT (It is ViT DINO in our case) model works.
- Understood how to use models like Claude 3 Sonnet using api calls and the usage of image generation models.
- Learnt how to utilize ASU's supercomputer 'SOL' to make use of GPUs for faster training of the model.
- Learnt the process of research and approaching a research problem through this project.

Team Members

- **Nagacharan Vemula (nvemula4@asu.edu)**
- Darshan Govindaraj (dgovind5@asu.edu)
- Siqin Yang (syang240@asu.edu)
- Nivetha MK (nmurugai@asu.edu)
- Saurabh Zinjad (szinjad@asu.edu)

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008). Long Beach, CA: Curran Associates, Inc.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (pp. 1-15). Virtual: ICLR.

[3] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.

[4] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A.(2021). Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the IEEE International Conference on Computer Vision (pp. 9657-9666). Montreal, QC: IEEE.