NUS Generative AI: Fundamentals to Advanced Techniques

Week 9: Capstone Project Part 2

Done by: A Alkaff Ahamed

A Alkaff Ahamed
10 June 2025

Learning Outcome Addressed

Learn to implement and leverage Generative AI methods, including diffusion models and multimodal systems, to develop creative and innovative AI solutions.

Overview:

In the previous week, you have understood how forward diffusion works and have begun to think about practical applications in your prototype. In Week 9, we have covered Transformers for Image Generation, focusing on models such as ViT, CLIP and DALL-E 2.

In this component of the Capstone Project, you will evaluate the use of ViT and CLIP as image classification models.

Go through the updated Jupyter Notebook for ViT, and note the accuracy, precision, recall and f1 score - `module_9_ViT_metrics_updated.ipynb`

Next, take a look at the Jupyter Notebook evaluating CLIP as an image classifier, and note the metrics given - `Image_classification_with_CLIP.ipynb`

Tasks

Task 1: Evaluate ViT Image Classification

- Display the final metrics of the model (accuracy, precision, recall, and f1 score)
- Discuss your observations regarding the efficiency of training a ViT versus the accuracy of the output (Approx. 100 words)

Task 2: Evaluate CLIP image classification

- Display the final metrics of the model (accuracy, precision, recall, and f1 score)
- Discuss your observations regarding the efficiency of training a CLIP model versus the accuracy of the output (Approx. 100 words)

Task 3: Compare and Contrast

- Having utilised both models for image classification, compare the two with respect to (Approx. 200 words):
    - Training speed and efficiency
    - Accuracy of output

Estimated time: 60-90 minutes

Submission Instructions:

- Select the **Start Assignment**button at the top right of this page.
- Upload your answers in the form of a Word or PDF file.
- Select the **Submit Assignment**button to submit your responses.

This is a graded and counts towards programme completion. You may attempt this assignment only once.

Assignment Answers

📌 Task 1: Evaluate ViT Image Classification

- Display the final metrics of the model (accuracy, precision, recall, and f1 score)
- Discuss your observations regarding the efficiency of training a ViT versus the accuracy of the output (Approx. 100 words)

📊 Model Metrics

Metric	Training	Evaluation
Accuracy	0.9746	0.835
Precision	0.9752	0.84
Recall	0.9746	0.83
F1 Score	0.9746	0.83

📝 Observations on ViT Training Efficiency vs Output Accuracy

- The model performs extremely well on the training data with metrics around **97.5%** indicating **strong learning capacity**
- Evaluation performance is **lower (~83–84%)** suggesting **some overfitting** or potential **generalization issues**
- ViT's transformer architecture leverages **global self-attention** which provides strong representational power but also increases **computational cost** and **training time**
- Compared to traditional CNNs, ViTs are generally **slower to train** and require **more GPU memory**
- Despite these limitations, the **high evaluation accuracy** demonstrates that ViT remains highly effective for image classification tasks when resources are available

📌 Task 2: Evaluate CLIP image classification

- Display the final metrics of the model (accuracy, precision, recall, and f1 score)
- Discuss your observations regarding the efficiency of training a CLIP model versus the accuracy of the output (Approx. 100 words)

📊 Model Metrics

Metric	Original Labels	"a photo of ..." Labels
Accuracy	0.692	0.727
Precision	0.7011	0.7275
Recall	0.692	0.727
F1 Score	0.692	0.727

📝 Observations on CLIP Training Efficiency vs Output Accuracy

- CLIP performed with **69.2% accuracy using raw class labels** but improved to **72.7%** when the labels were prefixed with **“a photo of...”** highlighting CLIP’s sensitivity to prompt phrasing
- This small but meaningful improvement shows that **prompt engineering** plays a critical role in maximizing CLIP's performance, especially in classification tasks
- Unlike ViT, CLIP is **not fine-tuned per task**. Instead, it operates in a **zero-shot** setting by embedding both text and image inputs and measuring similarity
- Since no training is performed, CLIP is **extremely efficient to use** requiring minimal compute and time
    - ideal for rapid experimentation
- However, the output accuracy is still lower than a trained ViT model, so there's a tradeoff between **ease-of-use** and **raw performance**.

📌 Task 3: Compare and Contrast

- Having utilised both models for image classification, compare the two with respect to (Approx. 200 words):
    - Training speed and efficiency
    - Accuracy of output

🕒 Training Speed and Efficiency

ViT

- **ViT** requires full supervised training on labeled datasets
    - Involves **multiple epochs, longer runtimes, and high GPU memory usage**
- Its transformer architecture uses **global self-attention** which enables rich feature extraction but increasing training complexity
- The ViT model used in this task achieved high training accuracy (~97.5%) but this came after extensive training and compute time

CLIP

- In contrast, **CLIP** operates in a **zero-shot setting** □ it uses **pretrained encoders** and doesn't require fine-tuning or backpropagation during classification tasks
- This makes CLIP **significantly faster to set up and run**, requiring only prompt preparation and inference-time computations
- For tasks where time, resources, or labels are limited, CLIP offers a **highly efficient alternative** with minimal infrastructure needs

🎯 Accuracy of Outputs

- **ViT** achieved strong evaluation metrics, with ~83.5% accuracy, F1 score, and other metrics, indicating **strong generalization** when properly trained
- **CLIP**, while more efficient, initially achieved **lower accuracy (~69.2%)** using unmodified labels
- However, with prompt engineering (by adding **"a photo of ..."** to class names), CLIP's accuracy improved to **~72.7%**, showing its reliance on natural language phrasing
- Despite this improvement, ViT's **task-specific training** still leads to better precision and consistency across labels

🧠 Overall Takeaway

ViT prioritizes performance at the cost of compute while **CLIP delivers speed and adaptability** with less accuracy unless carefully tuned.