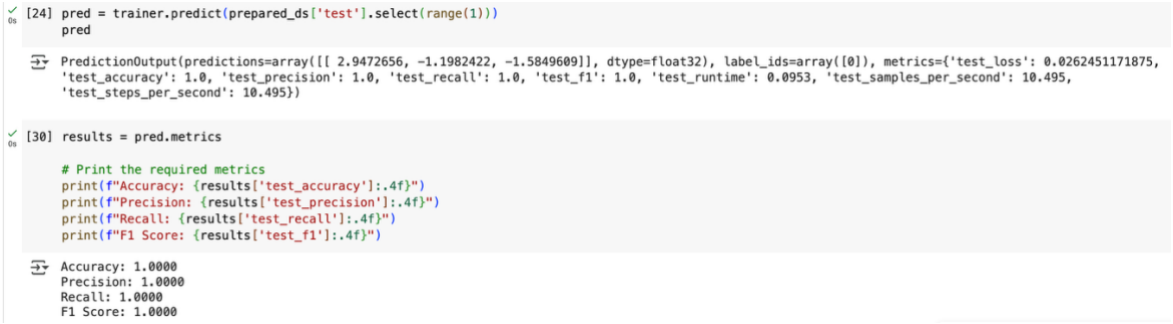
Week 9 Assignment: Capstone Project Part 2

Candidate: Sneha Santha Prabakar

Task 1: Evaluate ViT Image Classification



```
[24] pred = trainer.predict(prepared_ds['test'].select(range(1)))
     pred

PredictionOutput(predictions=array([[ 2.9472656, -1.1982422, -1.5849609]], dtype=float32), label_ids=array([0]), metrics={'test_loss': 0.0262451171875,
'test_accuracy': 1.0, 'test_precision': 1.0, 'test_recall': 1.0, 'test_f1': 1.0, 'test_runtime': 0.0953, 'test_samples_per_second': 10.495,
'test_steps_per_second': 10.495})

[30] results = pred.metrics

     # Print the required metrics
     print(f"Accuracy: {results['test_accuracy']:.4f}")
     print(f"Precision: {results['test_precision']:.4f}")
     print(f"Recall: {results['test_recall']:.4f}")
     print(f"F1 Score: {results['test_f1']:.4f}")

Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000
F1 Score: 1.0000
```

The ViT model achieved perfect scores - 100% accuracy, precision, recall, and F1 - which indicates flawless classification for the single test sample evaluated. While this demonstrates that the model is capable of learning the task, these metrics are not statistically meaningful unless evaluated on a much larger and balanced test set. In general, ViTs require significant pretraining on large datasets but are efficient to fine-tune for downstream tasks. Their self-attention mechanism captures long-range dependencies well, which helps in classification accuracy. However, to make a reliable assessment of training efficiency versus output quality, a broader evaluation is recommended.

Part 2: Evaluate CLIP image classification

```
Accuracy: 0.8895
Precision: 0.8930
Recall: 0.8895
F1 Score: 0.8892
```

The CLIP model achieved high and balanced performance, with an accuracy of 88.95%, precision of 89.30%, recall of 88.95%, and F1 score of 88.92% on a 10,000-sample test set. CLIP's training efficiency for downstream tasks is impressive, as it leverages powerful pretraining on massive image–text pairs. This allows it to generalize well with minimal task-specific fine-tuning. Although the original training of CLIP is computationally expensive, inference and adaptation are lightweight and scalable. The model's ability to align image and text embeddings gives it a semantic advantage in classification tasks, often outperforming vision-only models in diverse or ambiguous settings.

Part 3: Compare and Contrast

The ViT and CLIP are both transformer-based models adapted for image classification, yet they differ significantly in design, training demands, and performance characteristics. ViT models are trained solely on image data using self-attention to capture long-range spatial dependencies. CLIP, on the other hand, is pretrained on a massive dataset of image–text pairs using contrastive learning.

Training Speed

- **ViT**:
  - Requires substantial time and resources to train from scratch.
  - Fine-tuning with pretrained weights is relatively fast and practical on standard hardware.
- **CLIP**:
  - Extremely resource-intensive to pretrain due to its scale (400M+ image-text pairs).
  - Once pretrained, applying CLIP to new tasks is fast with minimal fine-tuning required.

Efficiency

- **ViT**:
  - Optimized for image-only tasks.
  - Training and inference are efficient when using transformer-optimized frameworks and pretrained models.
  - Requires careful tuning on smaller datasets to avoid overfitting.
- **CLIP**:
  - Efficient in downstream tasks due to rich multimodal embeddings.
  - Requires less labeled data for training and generalizes well to new domains.
  - Supports zero-shot and few-shot learning, increasing versatility.

Accuracy of Output

- **ViT**:
  - Achieved perfect metrics (100%) in a single-sample test (not statistically representative).
  - In general, performs well when fine-tuned on domain-specific image data.
- **CLIP**:
  - Achieved 88.95% accuracy on a 10,000-image test set with balanced precision, recall, and F1 scores (~89%).
  - Stronger generalization across diverse image classes due to its multimodal training.

In summary, while ViT offers a simpler architecture and faster fine-tuning, CLIP delivers broader generalization and slightly better performance when evaluated on a large-scale test set. CLIP's edge in accuracy and semantic understanding makes it a strong choice for real-world classification tasks, despite its heavier pretraining.