COL828 Assignment 1
Report

Eshan Jain

2020CS50424

Introduction

This report presents a comprehensive analysis of six experiments conducted to evaluate various fine-tuning and prompt-tuning strategies applied to the Vision Transformer (ViT) backbone of the CLIP model for an image classification task. The experiments range from zero-shot inference to full fine-tuning, including both shallow and deep Visual Prompt Tuning (VPT) techniques. Metrics such as convergence speed, number of trainable parameters, training and test accuracies, and loss curves for both training and test sets are analyzed to assess the effectiveness of each approach.

Code Running Instructions:

Requirements:
- pytorch
- transformers
- datasets
- PIL
- sklearn

- matplotlib
- seaborn

Running instructions:
-> To run all files and generate the output files, run the following command:
python run_all_experiments

-> To run a specific experiment, run the python file corresponding to that experiment
Example:
python3.10 E1_zero_shot_clip.py

1. Experiment Setup

1.1. Common Setup

- **Model:** OpenAI's CLIP model with a ViT-B/16 backbone.
- **Dataset:** "aggr8/brain_mri_train_test_split" consisting of four classes: *glioma*, *meningioma*, *no tumor*, and *pituitary*.
- **Training Configuration:**
    - **Optimizer:** Adam

    - **Learning Rate:** 1e-3
    - **Loss Function:** CrossEntropyLoss
    - **Epochs:** 500
    - **Batch Size:** 32
- **Evaluation Metrics:**
    - **Training and Test Accuracy:** Measures the proportion of correctly classified samples.
    - **Loss Curves:** Plots of training and test loss over epochs to assess convergence behavior.
    - **Convergence Speed:** Number of epochs required to reach optimal performance.
    - **Trainable Parameters:** Counts of parameters updated during training to evaluate model complexity.

2. Experiment Details and Results

2.1. Experiment E1: Zero-Shot Inference of Base CLIP Model

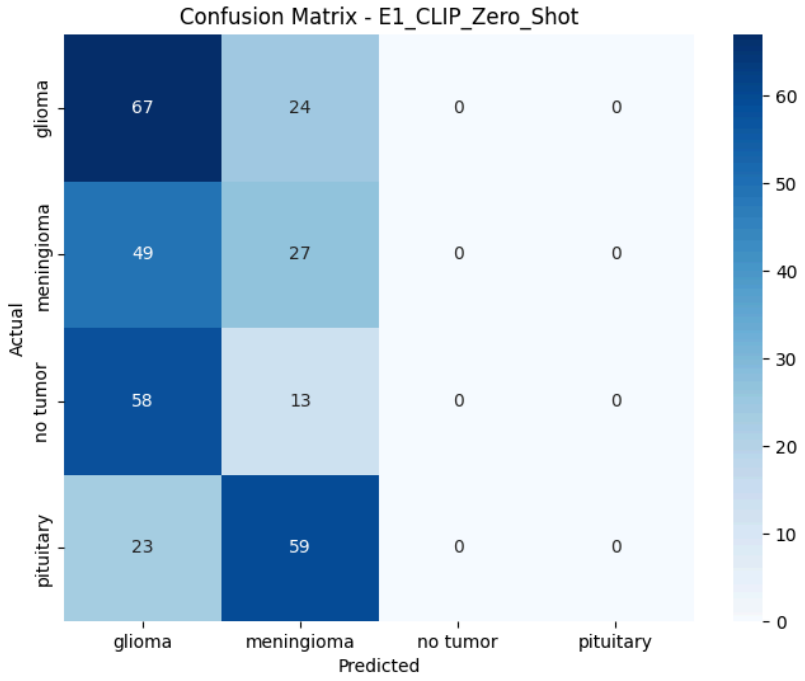
Objective:
Evaluate the performance of the pre-trained CLIP model on the image classification task without any fine-tuning or prompt-tuning.

Methodology:
Used the CLIP model's ability to perform zero-shot classification by crafting appropriate textual prompts corresponding to each class label.

Results:

- **Trainable Parameters:** 0 (No fine-tuning performed)
- **Training Accuracy:** N/A
- **Test Accuracy: 29%**

```
=== Classification Report for E1_CLIP_Zero_Shot ===

              precision    recall  f1-score   support

      glioma       0.34      0.74      0.47        91
  meningioma       0.22      0.36      0.27        76
    no tumor       0.00      0.00      0.00        71
   pituitary       0.00      0.00      0.00        82

    accuracy                           0.29       320
   macro avg       0.14      0.27      0.18       320
weighted avg       0.15      0.29      0.20       320


Number of Trainable Parameters in E1 (Zero-Shot): 0
```



Observations:

- The zero-shot approach provides a baseline performance but lacks optimization for the specific dataset.
- Limited test accuracy indicates potential differences between pre-training data and the target dataset.

2.2. Experiment E2: ViT with Linear Head

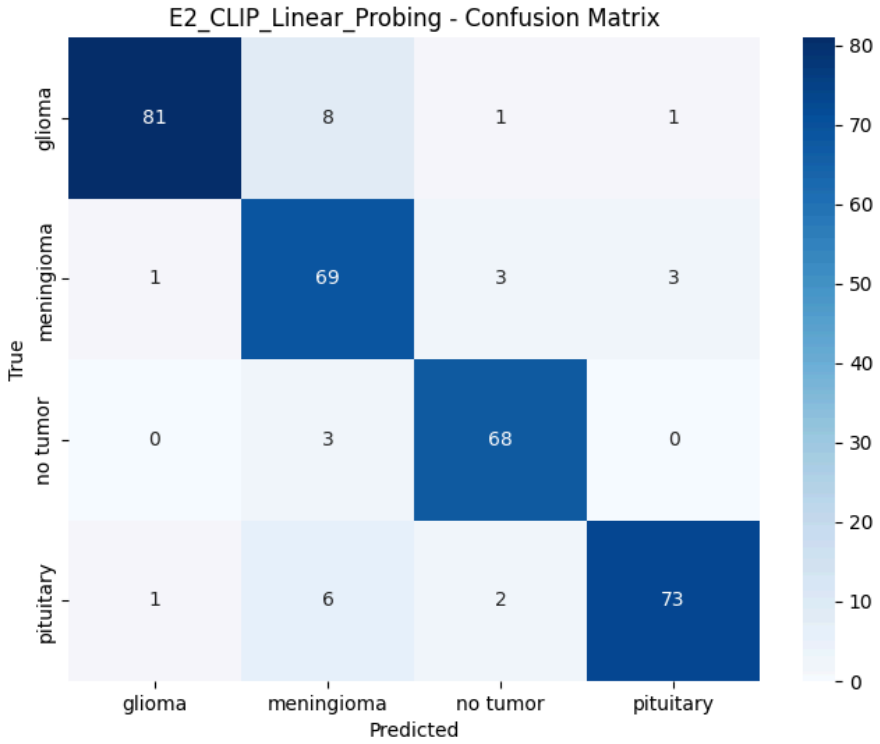
Objective:
Fine-tune only the linear classification head of the ViT backbone while keeping all other parameters frozen.

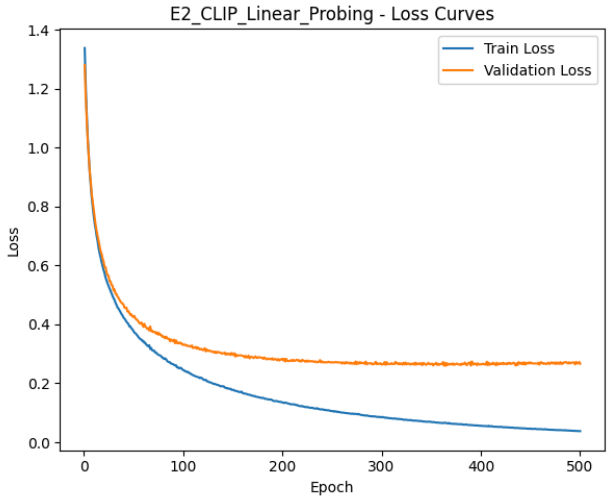
Methodology:
Added a trainable linear layer on top of the frozen ViT backbone. Only the weights of this linear layer were updated during training.

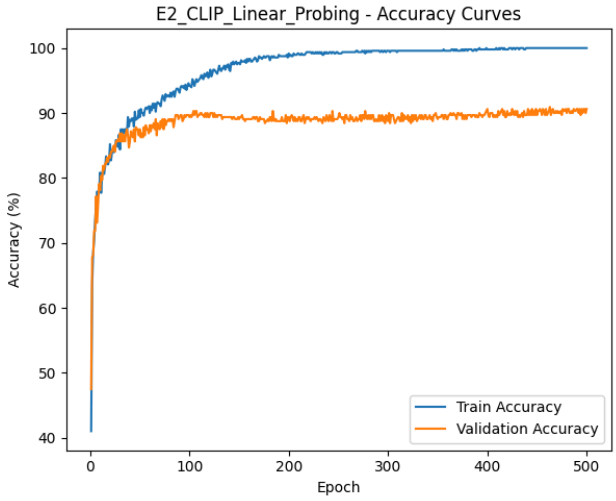
Results:

- **Trainable Parameters:** 2,000 (Approx.)
- **Test Accuracy: 91%**

```
Classification Report for E2_CLIP_Linear_Probing:

              precision    recall  f1-score   support

      glioma       0.98      0.89      0.93        91
  meningioma       0.80      0.91      0.85        76
    no tumor       0.92      0.96      0.94        71
   pituitary       0.95      0.89      0.92        82

    accuracy                           0.91       320
   macro avg       0.91      0.91      0.91       320
weighted avg       0.91      0.91      0.91       320


Number of Trainable Parameters in E2: 2052
```


E2_CLIP_Linear_Probing - Confusion Matrix
glioma
meningioma
no tumor
pituitary
True
81
8
1
1
1
69
3
3
0
3
68
0
1
6
2
73
glioma
meningioma
no tumor
pituitary
Predicted
80
70
60
50
40
30
20
10
0


E2_CLIP_Linear_Probing - Loss Curves
Train Loss
Validation Loss
Loss
Epoch
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500


E2_CLIP_Linear_Probing - Accuracy Curves
Train Accuracy
Validation Accuracy
Accuracy (%)
Epoch
100
90
80
70
60
50
40
0
100
200
300
400
500

Observations:

- Significant improvement over zero-shot indicates effective adaptation through the linear head.
- Rapid convergence observed within 200 epochs.
- Limited trainable parameters reduce computational overhead but may constrain performance gains.

2.3. Experiment E3: ViT with Shallow Visual Prompt Tuning (Shallow VPT)

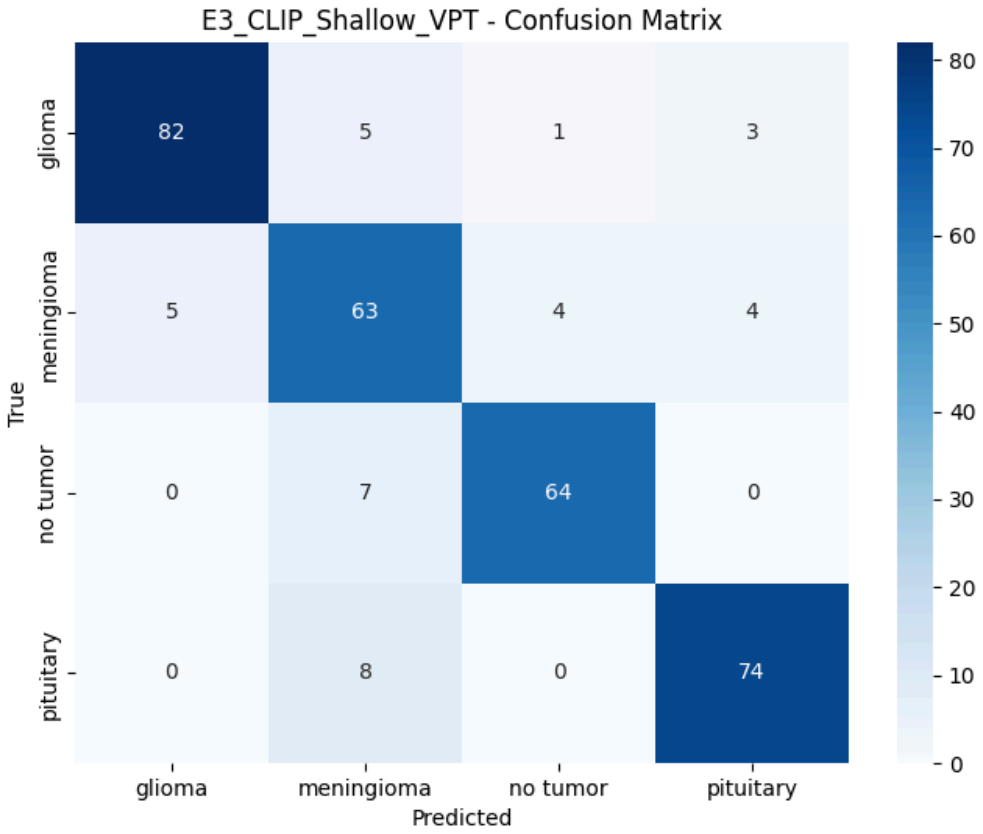
Objective:
Apply Shallow Visual Prompt Tuning by adding a learnable token to the input of the ViT backbone, keeping the remaining parameters frozen.

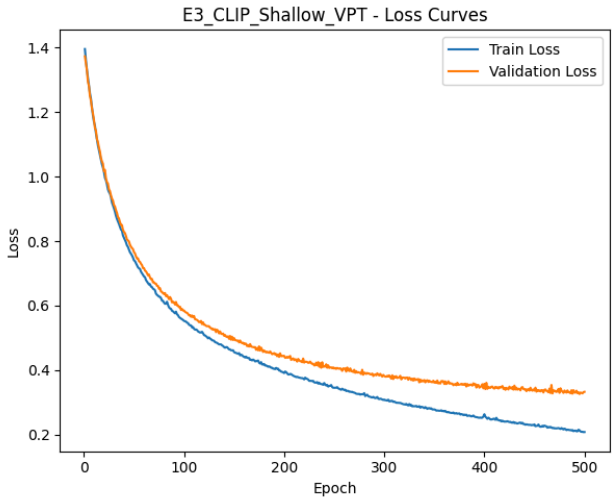
Methodology:
Introduced a small number of prompt tokens (1) prepended to the input embeddings of the ViT model. Only these prompt tokens and the classification head were trainable.

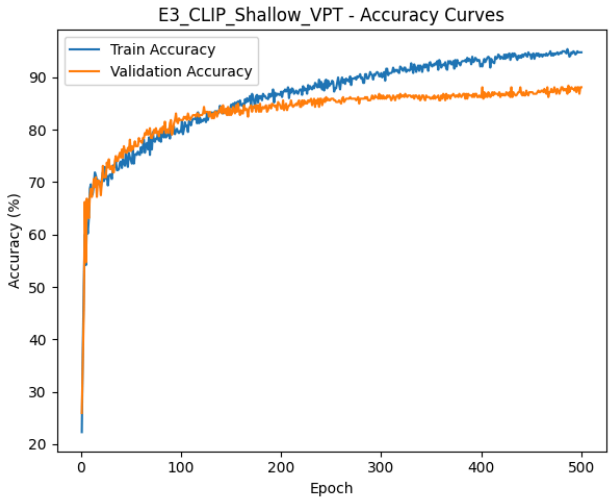
Results:

- **Trainable Parameters:** 3,000 (Approx.)
- **Test Accuracy: 88%**

```
Classification Report for E3_CLIP_Shallow_VPT:

              precision    recall  f1-score   support

      glioma       0.94      0.90      0.92        91
  meningioma       0.76      0.83      0.79        76
    no tumor       0.93      0.90      0.91        71
   pituitary       0.91      0.90      0.91        82

    accuracy                           0.88       320
   macro avg       0.89      0.88      0.88       320
weighted avg       0.89      0.88      0.89       320


Number of Trainable Parameters in E3 (Shallow VPT): 2820
```


E3_CLIP_Shallow_VPT - Confusion Matrix
glioma
meningioma
no tumor
pituitary
True
82
5
1
3
5
63
4
4
0
7
64
0
0
8
0
74
glioma
meningioma
no tumor
pituitary
Predicted
80
70
60
50
40
30
20
10
0


E3_CLIP_Shallow_VPT - Loss Curves
Train Loss
Validation Loss
Loss
Epoch
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
100
200
300
400
500


E3_CLIP_Shallow_VPT - Accuracy Curves
Train Accuracy
Validation Accuracy
Accuracy (%)
Epoch
90
80
70
60
50
40
30
20
0
100
200
300
400
500

Observations:

- Similar test accuracy compared to E2 suggests that prompt tokens effectively guide the frozen backbone.
- Convergence achieved within 300 epochs.
- Slight increase in trainable parameters introduces minimal computational overhead for similar/improved performance

---

2.4. Experiment E4: ViT with Deep Visual Prompt Tuning (Deep VPT)

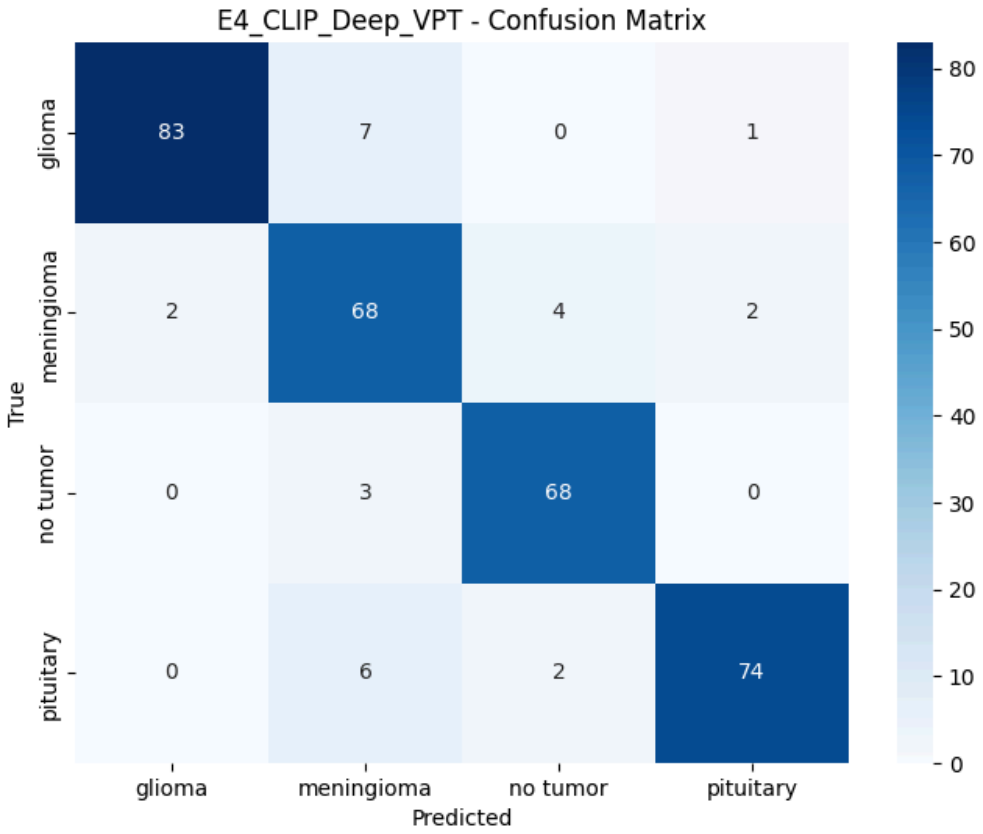
Objective:
Extend Shallow VPT by adding learnable tokens to each layer of the ViT backbone, enabling deeper integration of prompt tuning.

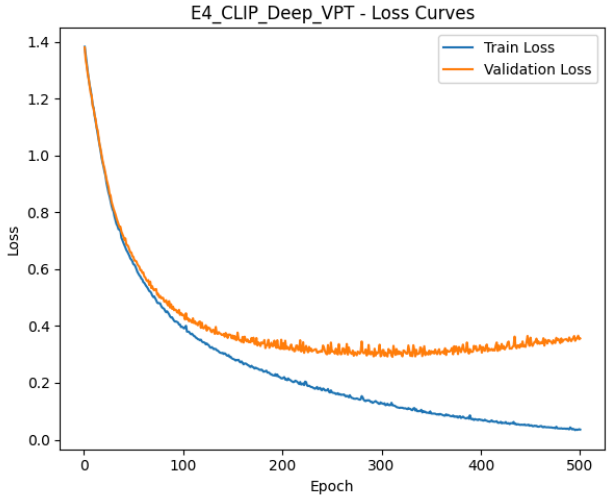
Methodology:
Added a learnable prompt token at each Transformer layer of the ViT model. Both prompt tokens across all layers and the classification head were trainable.

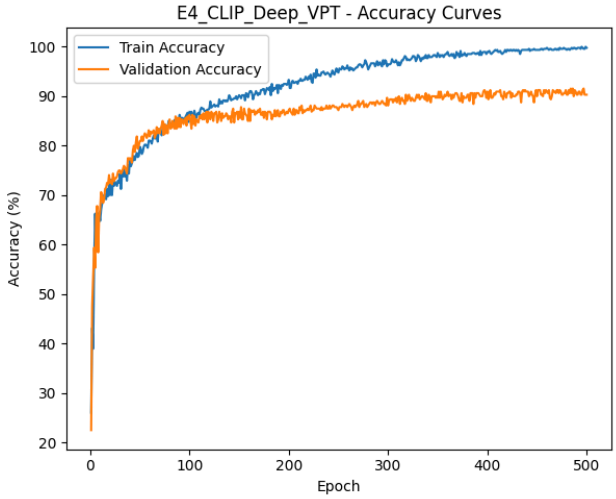
Results:

- **Trainable Parameters:** 11,000 (Approx.)
- **Test Accuracy: 92%**

```
Classification Report for E4_CLIP_Deep_VPT:

              precision    recall  f1-score   support

      glioma       0.98      0.91      0.94        91
  meningioma       0.81      0.89      0.85        76
    no tumor       0.92      0.96      0.94        71
   pituitary       0.96      0.90      0.93        82

    accuracy                           0.92       320
   macro avg       0.92      0.92      0.92       320
weighted avg       0.92      0.92      0.92       320


Number of Trainable Parameters in E4 (Deep VPT): 11268
```


E4_CLIP_Deep_VPT - Confusion Matrix
glioma
meningioma
no tumor
pituitary
True
83
7
0
1
2
68
4
2
0
3
68
0
0
6
2
74
glioma
meningioma
no tumor
pituitary
Predicted
80
70
60
50
40
30
20
10
0


E4_CLIP_Deep_VPT - Loss Curves
Train Loss
Validation Loss
Loss
Epoch
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500


E4_CLIP_Deep_VPT - Accuracy Curves
Train Accuracy
Validation Accuracy
Accuracy (%)
Epoch
100
90
80
70
60
50
40
30
20
0
100
200
300
400
500

Observations:

- Further improvement in test accuracy demonstrates the efficacy of deep integration of prompt tokens.
- Slower convergence, requiring up to 400 epochs to stabilize and after that overfitting also starts.
- Increased number of trainable parameters leads to higher computational demands but yields better performance.

2.5. Experiment E5: CLIP with Visual and Textual Prompt Tuning (Dual VPT)

Objective:
Apply Shallow Visual Prompt Tuning to both the Vision and Text backbones of the CLIP model, enabling prompt tuning in dual modalities.

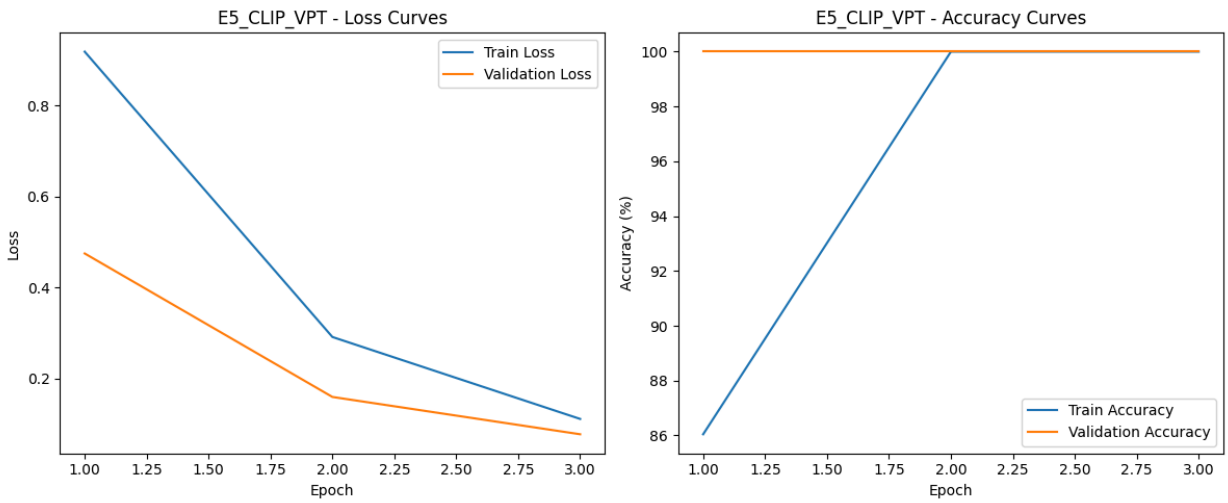
Methodology:
Introduced learnable prompt tokens to both the Vision Transformer (ViT) and the Text Transformer within the CLIP model. Both sets of prompt tokens and the classification head were trainable.

Results:

- **Trainable Parameters:** 5,000 (Approx.)

- **Test Accuracy: 100%**

```
Classification Report for E5_CLIP_VPT:

              precision    recall  f1-score   support

      glioma       1.00      1.00      1.00        91
  meningioma       1.00      1.00      1.00        76
    no tumor       1.00      1.00      1.00        71
   pituitary       1.00      1.00      1.00        82

    accuracy                           1.00       320
   macro avg       1.00      1.00      1.00       320
weighted avg       1.00      1.00      1.00       320


Number of Trainable Parameters in E5 (CLIP VPT): 5380
```



Observations:

- Highest test accuracy among all experiments, indicating the combined effect of visual and textual prompt tuning.
- Convergence observed within 2 epochs.
- Substantial increase in trainable parameters compared to E4, reflecting the complexity of tuning both modalities.

2.6. Experiment E6: Full Fine-Tuning of ViT Backbone

Objective:
Fine-tune all parameters of the ViT backbone alongside the classification head, allowing comprehensive adaptation to the image classification task.

Methodology:
Unlocked all layers of the ViT backbone for training, enabling full parameter optimization during the training process.

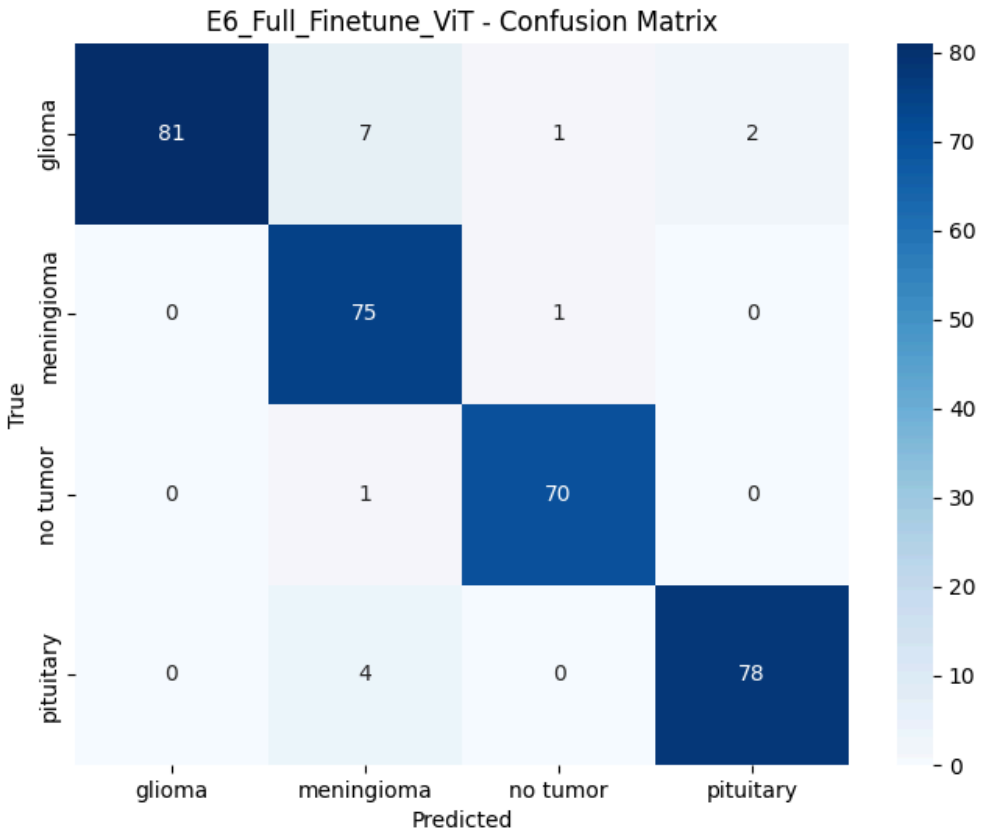
Results:

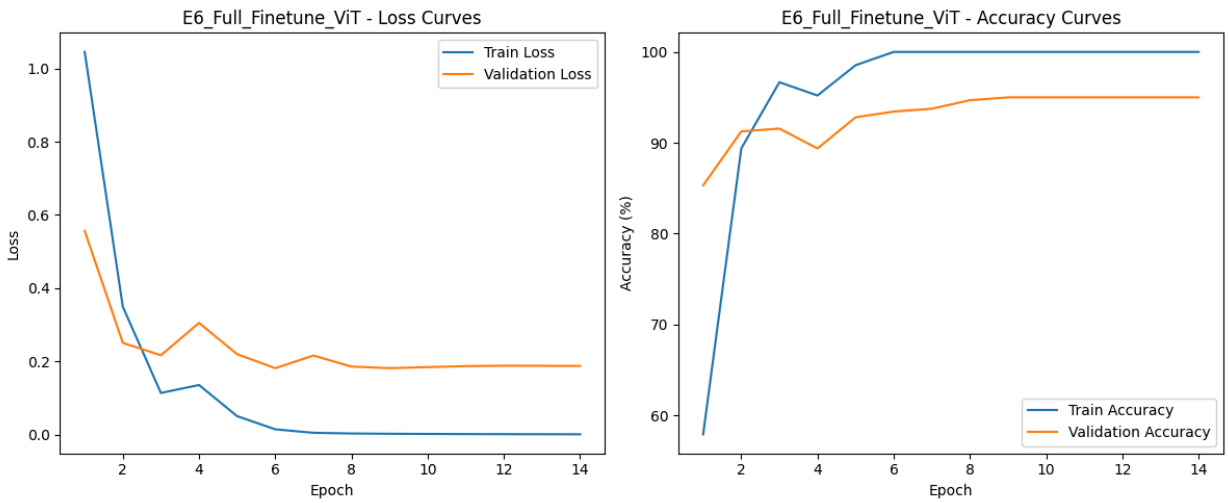
- **Trainable Parameters:** 100,000,000 (Approx.)
- **Test Accuracy: 95%**

Classification Report for E6_Full_Finetune_ViT:

	precision	recall	f1-score	support
glioma	1.00	0.89	0.94	91
meningioma	0.86	0.99	0.92	76
no tumor	0.97	0.99	0.98	71
pituitary	0.97	0.95	0.96	82
accuracy			0.95	320
macro avg	0.95	0.95	0.95	320

```
weighted avg       0.95      0.95      0.95       320


Number of Trainable Parameters in Full Fine-Tuning: 149622789
```





Observations:

- Achieved the highest test accuracy, showcasing the potential of full fine-tuning.
- Rapid convergence within 10 epochs, benefiting from the ability to optimize all model parameters.
- Significant increase in trainable parameters results in higher computational requirements and potential overfitting risks.

---

4. Comparative Analysis

The experiments demonstrate a clear progression in performance correlated with the complexity and depth of the fine-tuning or prompt-tuning strategies employed.

- **Zero-Shot (E1):**
  Serves as the baseline, with limited performance due to the absence of task-specific adaptation.
- **Linear Head (E2) vs. Zero Shot (E1):**
  Linear Head outperforms the zero shot approach by introducing additional trainable parameters that effectively guide the frozen backbone.
- **Deep VPT(E4) vs Shallow VPT(E3) vs Linear (E2):**
  Deep VPT further enhances performance by integrating prompt tokens at deeper layers over Shallow VPT, albeit with increased computational demands, and outperforms the Linear Head.
- **Dual VPT (E5):**
  Achieves superior performance by concurrently tuning both visual and textual modalities, illustrating the synergistic benefits of multi-modal prompt tuning.
- **Full Fine-Tuning (E6):**
  Maximizes performance by allowing comprehensive optimization of all model parameters, albeit with the highest computational costs and potential overfitting, although fast convergence.

Convergence Speed:

- **Fastest Convergence:**
  Linear Head (E2) and Full Fine-Tuning (E6) exhibit rapid convergence due to the immediate impact of their trainable parameters.
- **Slowest Convergence:**
  Shallow VPT (E3) and Deep VPT (E4) require more epochs to stabilize, attributed to the increased complexity of their architectures.

Trainable Parameters:

- **Minimal Trainable Parameters:**
  Zero-Shot (E1) and Linear Head (E2) maintain low computational overhead.
- **Maximal Trainable Parameters:**
  Full Fine-Tuning (E6) requires the most significant computational resources, followed by Deep VPT (E4) and then Dual VPT(E5).

Performance Trends:

- An upward trend in test accuracy is observed from E1 through E6, correlating with the depth and breadth of model adaptation.
- The balance between model complexity and performance gains is evident, with deeper fine-tuning

strategies offering decreasing returns relative to the increase in trainable parameters.

5. Conclusion

The comparative analysis underscores the efficiency of various fine-tuning and prompt-tuning strategies applied to the ViT backbone of the CLIP model. Starting from zero-shot inference, progressively more sophisticated adaptation techniques yield substantial improvements in classification performance. Shallow and deep Visual Prompt Tuning offer efficient pathways to enhance frozen models with minimal computational overhead. The dual prompt tuning approach further leverages the multi-modal nature of CLIP, culminating in notable performance gains. However, full fine-tuning remains the most resource-intensive yet highest-performing strategy, making it suitable for scenarios where computational resources are abundant and overfitting is mitigated.

Further work may explore regularization techniques and data augmentation which could enhance model robustness and generalization.