Task 3: Training Considerations

ML Apprentice Take-Home Exercise

Introduction

Fine-tuning a multi-task Transformer demands careful choices about which parameters to update and which to hold fixed. These decisions affect convergence speed, overfitting risk, and how well the model can adapt to each downstream objective. Below, we examine three freezing scenarios and then outline a progressive transfer-learning workflow.

1 Entire Network Frozen

All parameters—both the shared Transformer encoder and the task-specific heads—are held fixed at their pretrained values.

Advantages

- **Zero fine-tuning cost**: No gradient computations, enabling fast inference and feature extraction.
- **No catastrophic forgetting**: Preserves all pretrained knowledge intact.

Limitations

- **No task adaptation**: Cannot specialize representations to new tasks, often leading to suboptimal performance.

When to Use

- Rapid prototyping or very low-resource settings where any fine-tuning risks overfitting.

2 Freeze Transformer Backbone Only

The Transformer encoder layers are frozen; only the task-specific heads are trainable.

Advantages

- **Fast convergence**: Trains only a small number of head parameters.
- **Regularization**: Retains general-purpose pretrained representations.

Limitations

- **Limited representational flexibility**: Deeper features cannot adapt to new task-specific patterns.

When to Use

- Moderate dataset sizes where head adaptation suffices while minimizing overfitting.

3 Freeze One Task-Specific Head Only

One of the two task heads (classification or NER) is frozen, while the backbone and the other head remain trainable.

Advantages

- **Selective stability**: Maintains performance on the frozen task.
- **Targeted capacity**: Focuses model capacity on improving the underperforming task.

Limitations

- **Asymmetric adaptation**: The frozen head cannot benefit from updated shared representations.

When to Use

- Imbalanced tasks where one has abundant data and the other is low-resource or noisy.

4 Progressive Transfer Learning Workflow

To leverage pretrained knowledge effectively while adapting to multi-task objectives, we recommend staged unfreezing:

Stage	**Frozen Layers**	**Trainable Layers**
Head-Only Tuning	All encoder layers	Both task heads
Partial Unfreeze	Bottom N transformer blocks	Top transformer blocks + heads
Full Fine-Tuning	None	Entire model

5 Rationale Behind Choices

- **Progressive Unfreezing:** Gradually expose more layers to training to prevent catastrophic forgetting.
- **Differential Learning Rates:** Apply lower learning rates to encoder layers (e.g., 1e-5) and higher rates to task heads (e.g., 5e-5).
- **Validation-Guided Strategy:** Use separate validation metrics for each task to inform unfreezing and learning rate adjustments.