Catastrophic Forgetting

Ali Ahmadi Esfidi

July 14, 2025

1 Introduction

This report investigates two versions of sequential learning implementation for the MNIST digit classification problem. The first version is trained without any protective mechanism, while the second version utilizes the **Orthogonal Gradient Descent (OGD)** method to prevent catastrophic forgetting.

2 Version 1: Simple Training

2.1 Data and Preprocessing

- The MNIST dataset was divided into two subsets:
  - **Task A:** Digits 0 to 4
  - **Task B:** Digits 5 to 9
- **Transformations:** Converted to tensors and normalized with a mean of 0.1307 and a standard deviation of 0.3081.

2.2 Model Architecture

```
model = nn.Sequential(
    nn.Flatten(),
    nn.Linear(28*28, 100), nn.ReLU(),
    nn.Linear(100, 100),   nn.ReLU(),
    nn.Linear(100, 10)
)
```

2.3 Training Settings

- **Optimizer:** SGD with a learning rate of $1e^{-3}$
- **Loss Function:** CrossEntropyLoss
- **Number of Epochs per Phase:** 3

2.4 Results

- **Before Task B (Task A phase):** Task A test accuracy was 97.55%.
- **After Task B:** Task A test accuracy dramatically dropped to 0%.

2.5 Analysis

This phenomenon, where the model loses all knowledge of a previous task after learning a new one, is called **catastrophic forgetting**. This implementation lacks any mechanism to preserve prior knowledge.

3 Version 2: Using OGD

3.1 Orthogonal Gradient Descent

1. Calculate gradients ∇L on Task A samples.
2. Apply the Gram-Schmidt process to extract orthogonal, normalized vectors.
3. During Task B training, project the overall model gradient onto the subspace orthogonal to these vectors.

3.2 Results

- **After Task A:** Task A test accuracy was 93.75%.
- **Number of Stored Directions:** 32
- **After Task B:** Task A test accuracy experienced only a slight decrease, reaching 93.44%.

4 Comparison and Final Analysis

Metric	**Version 1**	**Version 2**
Task A Accuracy Before Task B	97.55%	92.90%
Task A Accuracy After Task B	0.00%	92.16%
Severity of Catastrophic Forgetting	Severe	Almost Resolved

Version 2, using OGD, successfully maintained a functional memory of Task A throughout Task B learning.

5 References

- *PyTorch: An Imperative Style, High-Performance Deep Learning Library*
- *Orthogonal Gradient Descent for Continual Learning (Farajtabar, et al.)*