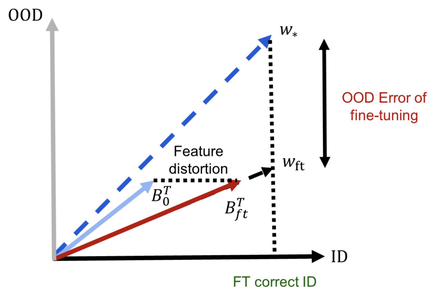
Effects of Fine-Tuning on In-Distribution and Out-of-Distribution Performance

Kunal Kapoor, Suhas Kotha *(Carnegie Mellon University)*

Introduction

- Fine-Tuning refers to the process of taking a pre-trained model and training it further on some dataset
- Fine-Tuning has impacts on performance of model In-Distribution (I.D.) and Ou-of-Distribution (O.O.D.)
- In-Distribution ⟶ distribution of training data
- Out of Distribution ⟶ distribution of non-training data



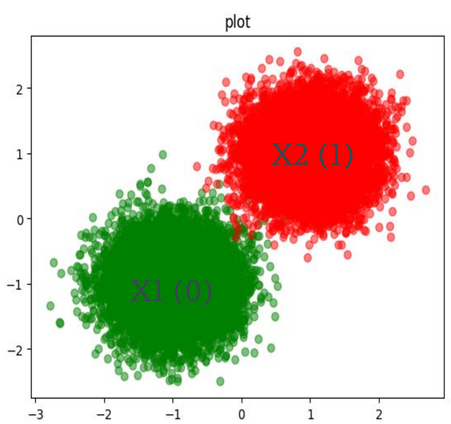
(b) Toy example (fine-tuning)

Methodology

- Goal: Develop a theoretical understanding of phenomena such as catastrophic forgetting and data replay during fine-tuning
- Created a synthetic setup to study effects of fine-tuning on simple neural networks
  - Tested impact of scale on forgetting
  - Implemented data replay
- Pre-trained over 2 gaussians as a regression problem; fine-tuned by shifting label over X1

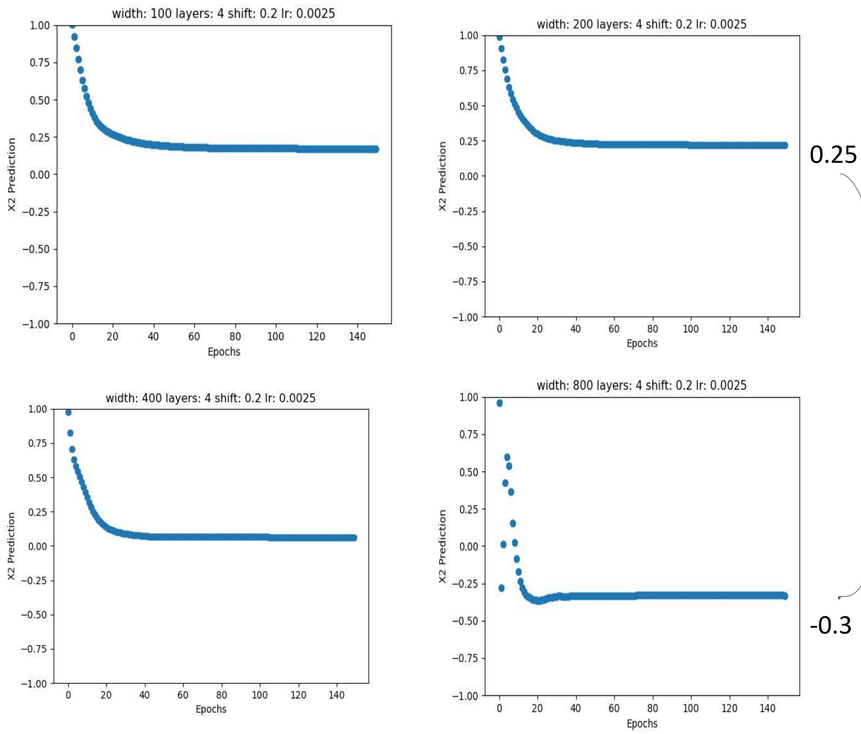
<u>Pre-training</u>
X1 ~ N(-1,0.2)
Label: 0
X2 ~ N(1, 0.2)
Label: 1

<u>Fine-Tuning</u>
X1 ~ N(-1,0.2)
Label: **-0.2**
X2 ~ N(1, 0.2)
Label: 1



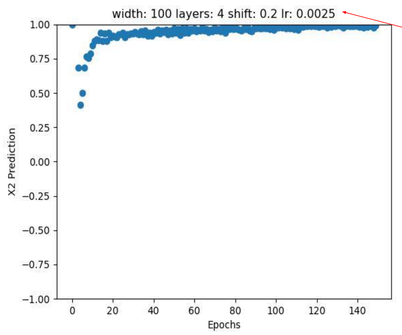
Results

- Width Increasing ⟶ More Forgetting



Decrease in X2 prediction represents forgetting

- Data Replay at 5-10% significantly reduces forgetting



X2 Prediction stays near 1; almost no decrease

5-10% of data during fine-tuning is drawn from pre-training distribution

- Found empirical examples of catastrophic forgetting in Llama-2-7B
  - Fine-tuned on Alpaca dataset (instruction training)
  - Evaluated on MNLI (Multi-Genre Natural Language Inference Corpus)
  - Accuracy decreased from **42%** to **33%**, demonstrating forgetting

Discussion

<u>Key Takeaways</u>

- Scale solving forgetting via capacity is likely an incorrect conclusion
- It's more likely that **optimization differences** explains why scale solves forgetting
- Data replay boosts O.O.D. performance even at a **low** percentage of fine-tuning data

<u>Analysis of I.D. and O.O.D. performance via NTK</u>

- Neural Tangent Kernel (NTK)
  - Measures sensitivity of function value at x to prediction errors at x'

$$k_\theta(x, x') = \left\langle \frac{df_\theta(x)}{d\theta}, \frac{df_\theta(x')}{d\theta} \right\rangle$$

- NTK can model how predictions change as the model performs updates over training data
- Idea: Write a function that determines how model's predictions change I.D. and O.O.D. as function of time
- NTK for explaining data replay
  - Calculate changes I.D. and O.O.D. over fine-tuning and data replay distribution
  - Determine optimal replay rate and replay curriculum

Future Work

- Formalize the NTK's explanation for data replay
- Investigate more complicated synthetic setups for determining if scale solves forgetting
- Find more instances of catastrophic forgetting in LLMs
- Test out scale hypothesis among LLMs

Contact
kunalkapoor@cmu.edu