Redes Neurais e Aprendizado Profundo

Artur Jordão
Escola Politécnica – Engenharia de Computação e Sistemas Digitais
Universidade de São Paulo

Neural Phylogeny

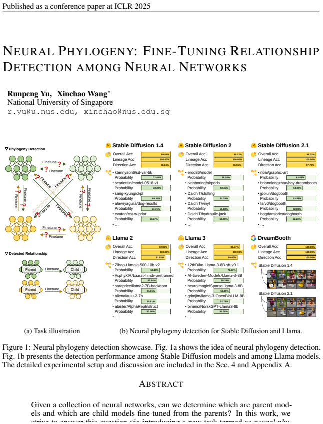
Introduction

Neural Phylogeny

- The field of deep learning has experienced transformative evolutions
    - Traditional *training from scratch* paradigm has shifted to the *pre-training-fine-tuning* approach
    - Numerous open-source model repositories have been established, and models produced by *machine learning as a service* are accumulating
- These advancements mark the emergence of a complex, interconnected network of deep learning models
- Potential applications of studying this interconnected network
    - Copyright protection
    - Understand the inheritance of knowledge and connections in the growing neural model network

Introduction

Neural Phylogeny


Published as a conference paper at ICLR 2025

NEURAL PHYLOGENY: FINE-TUNING RELATIONSHIP DETECTION AMONG NEURAL NETWORKS

Runpeng Yu, Xinchao Wang*
National University of Singapore
r.yu@u.nus.edu, xinchao@nus.edu.sg

(a) Task illustration (b) Neural phylogeny detection for Stable Diffusion and Llama.

Figure 1: Neural phylogeny detection showcase. Fig. 1a shows the idea of neural phylogeny detection. Fig. 1b presents the detection performance among Stable Diffusion models and among Llama models. The detailed experimental setup and discussion are included in the Sec. 4 and Appendix A.

ABSTRACT

Given a collection of neural networks, can we determine which are parent models and which are child models fine-tuned from the parents? In this work, we

- Yu et al.[1] investigated the novel task of neural phylogeny detection
- *Given a set of models, neural phylogeny aims to identify parent-child model pairs connected* **through fine-tuning** *behavior and determine the direction of fine-tuning*

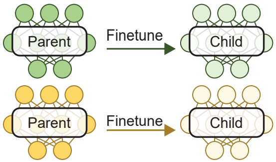
[1] Yu et al. *Neural Phylogeny: Fine-Tuning Relationship Detection among Neural Networks*. ICLR, 2025

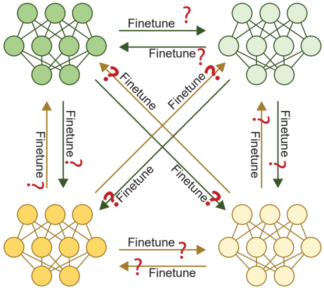
Problem Definition

Neural Phylogeny

- Consider a parent model $\mathcal{F}_p$ and its fine-tuned version, child, $\mathcal{F}_c$
  - Both $\mathcal{F}_p$ and $\mathcal{F}_c$ share the same or similar architecture

- Given a set of parent and child models, Neural Phylogeny involves:
  - Identifying all pairs of parent $\times$ child models
  - (Optional) Determining the parent and child in each pair: the fine-tuning direction





Neural Phylogeny Detection

Neural Phylogeny

- Following Yu et al.[1], neural phylogeny detection aims to solve a clustering task given a set of neural networks
  - Every model $\{\mathcal{F}\}_{i=1}^{M}$ corresponds to a point that the algorithm should cluster
  - Each cluster should contain a parent model and the child models fine-tuned from it
  - For each cluster, the algorithm should identify the parent model as the cluster centroid

- Many algorithms do not inherently include the identification of cluster centroids
  - Yu et al.[1] observed that parent models exhibit smaller parameter norms
  - From this finding, the authors assign the parent model in each cluster as the one with the smallest parameter norm: $\arg\min_i \|\theta^{(i)}\|_2$, where $\theta^{(i)}$ are the parameters of $\mathcal{F}_i$

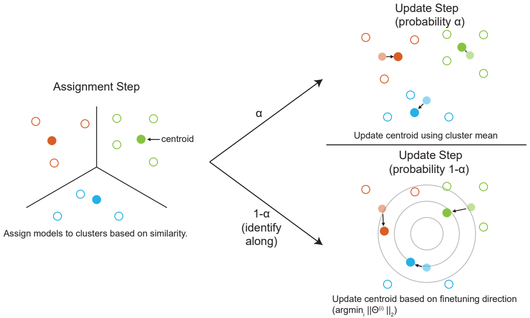
- Clustering algorithms: KMeans, MeanShift, DBSCAN, GMM

[1] Yu et al. *Neural Phylogeny: Fine-Tuning Relationship Detection among Neural Networks*. ICLR, 2025

Neural Phylogeny Detection with KMeans

Neural Phylogeny

- During the KMeans iterative process, a cluster may have no parent model. Thus, Yu et al.[1] introduce a parameter α when updating the cluster centroid
  - With a probability α, the algorithm uses the cluster's mean as the new centroid
  - With a probability $1-\alpha$, the algorithm uses the $\arg\min_i \|\theta^{(i)}\|_2$ as the new centroid



[1] Yu et al. *Neural Phylogeny: Fine-Tuning Relationship Detection among Neural Networks*. ICLR, 2025

Experimental Details

Neural Phylogeny

- A prediction counts as correct only when the parent-child pair matches and the fine-tuning direction is correct

- The authors train (yield) the child models on different datasets using various hyperparameters and training techniques
  - Including full and LoRA-based fine-tuning
  - 366 Stable Diffusion models
  - 95 Llama models from HuggingFace

Results

Neural Phylogeny

- Identify after Clustering
  - $\alpha = 1$

	KMeans		MeanShift	
Model	Average	Best	Average	Best
Stable Diffusion	14.92±7.45	18.64	18.97±1.01	19.02
Llama	0±0	0	0±0	0
DreamBooth	100.0±0	100.0	100.0±0	100.0

Results

Neural Phylogeny

- Identify along Clustering
  - $\alpha < 1$

Model	**KMeans**		**MeanShift**	
	Average	Best	Average	Best
Stable Diffusion	78.57±21.28	99.73	68.16±24.75	80.54
Llama	48.83±48.37	97.67	75.58±37.81	95.34
DreamBooth	100.0±0	100.0	100.0±0	100.0

Bibliography

Bibliography

- Yu et al. *Neural Lineage*. CVPR, 2024
- Yax et al. *PhyloLM: Inferring the Phylogeny of Large Language Models and Predicting their Performances in Benchmarks*. ICLR, 2025