Introduction to Deep Learning

Tutorial 6

Gabriel Deza
Department of Industrial Engineering
Tel Aviv University

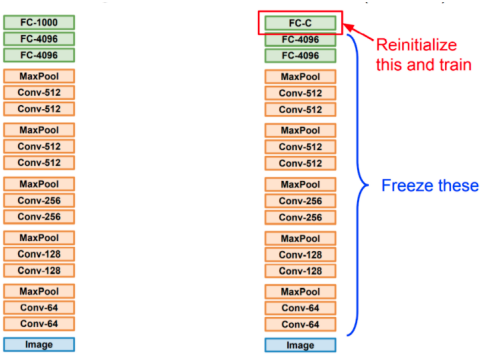
Deceember 9 & 11, 2025

Overview

Structure for next two tutorials:

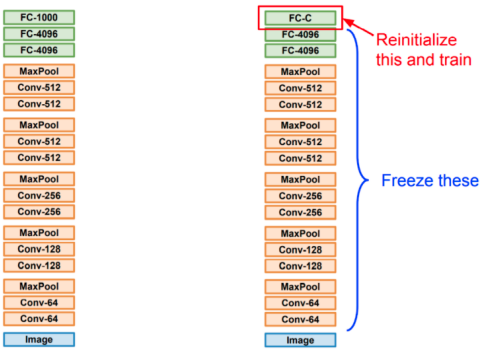
- Finetuning & Pretraining (concept)
- Finetuning & Pretraining (code)
- Brief history of how we got here

Transfer Learning



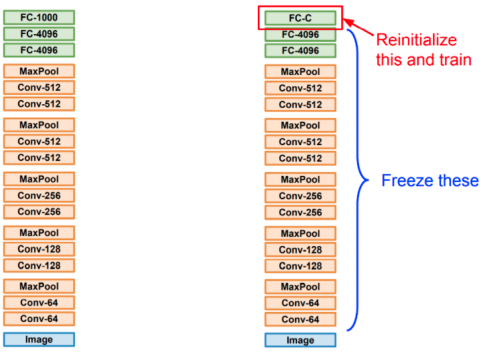
- **Phase 1: Pre-training**
  - Train a model $f(\cdot;\theta)$ on a large dataset (e.g., ImageNet).
  - The backbone learns robust feature representations.

Transfer Learning



- **Phase 1: Pre-training**
  - Train a model $f(\cdot;\theta)$ on a large dataset (e.g., ImageNet).
  - The backbone learns robust feature representations.
- **Phase 2: Transfer (Feature Extraction)**
  - Remove the final FC layer $W_{old} \in \mathbb{R}^{d\times 1000}$.
  - Initialize $W_{new} \in \mathbb{R}^{d\times K}$ (where K is target classes).
  - Freeze θ_{body}; train only W_{new}.

Transfer Learning



- **Phase 1: Pre-training**
  - Train a model $f(\cdot;\theta)$ on a large dataset (e.g., ImageNet).
  - The backbone learns robust feature representations.
- **Phase 2: Transfer (Feature Extraction)**
  - Remove the final FC layer $W_{old} \in \mathbb{R}^{d\times 1000}$.
  - Initialize $W_{new} \in \mathbb{R}^{d\times K}$ (where K is target classes).
  - Freeze θ_{body}; train only W_{new}.
- **Phase 3: Fine-tuning (Optional)**
  - Unfreeze θ_{body}.
  - Train all parameters, but use $\text{LR}_{body} \ll \text{LR}_{head}$.

ResNet

- 2015 CVPR paper by Microsoft Research.
- Trained on ImageNet and beat SOTA across benchmarks.
- $\sim$300k citations.
- Paper: *Deep Residual Learning for Image Recognition* (He et al., 2015)
- We are going to fine-tune ResNet-18 (one of several sizes: 18, 34, 50, 101, 152).
- Code: Colab notebook for ResNet-18 fine-tuning
- (Next) brief overview of how did we even get to pretrained models

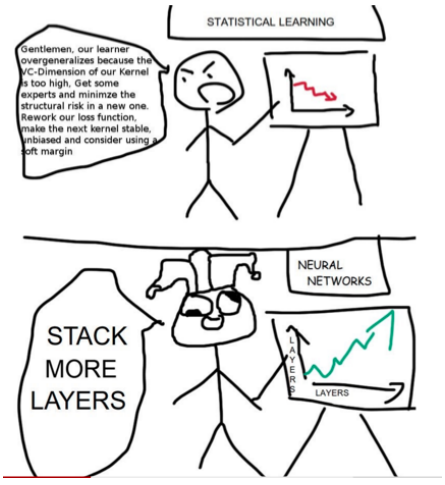
How did we even get to pretrained models?

ML/Stats dominance (Pre-2012)

- **Stats models**: SVMs, Random Forests, XGB etc
- Carefully designing features and kernels to minimize risk.
- **The Problem:** NNs were hard to train (vanishing gradients).

The "Deep"/NN Era (2012–2016)

- **AlexNet (2012):** Halves ImageNet error via *Deep CNNs + GPUs*.
- Takeaway was add more layers. Intense competition.
- **The Models:** VGG, GoogLeNet, and **ResNet**.
- By 2017, ImageNet is consider **solved**. Generative images are good.



Natural language processing (NLP)

2012–2017

- SOTA for NLP was **RNN/LSTM**.
- Processing was sequential (token-by-token). Hard to scale, hard to parallelize.

2017

- **"Attention Is All You Need"** (Google): Introduces the Transformer.
- **Key Win:** Allows parallel processing of entire sequences. *Scale becomes possible.*

The "ImageNet Moment" for NLP (2018)

- **BERT** demonstrates the power of **Transfer Learning** in text.
- Pre-train a massive "World Model" on the internet → Fine-tune for your specific task (we are here in the class)

Scale + multimodal learning

2019–2022

- Don't be clever, just add more data and GPUs (read "The Bitter Lesson" (Sutton) for more).
- **Homogenization:** Everything becomes a Transformer.
  - Vision $\rightarrow$ Vision Transformers (ViT).
  - Audio $\rightarrow$ Audio Transformers.
- *Result:* Emergent behaviors in Large Language Models (LLMs).

Scale + multimodal learning

2019–2022

- Don't be clever, just add more data and GPUs (read "The Bitter Lesson" (Sutton) for more).
- **Homogenization:** Everything becomes a Transformer.
  - Vision → Vision Transformers (ViT).
  - Audio → Audio Transformers.
- *Result:* Emergent behaviors in Large Language Models (LLMs).

2022–Present

- LLMs went from text completion to what we have now
- Instruction Tuning & **RLHF** (Reinforcement Learning from Human Feedback).
- remaining work on reasoning (Chain of Thought), multimodality (Gemini/GPT-4o), and agents.