Week 3: Deep Learning Session

Younesse Kaddar

Philosophy Seminar, University of Oxford

Topics in Minds and Machines: Perception, Cognition, and ChatGPT

1. Introduction

- **Some Milestones:**
  - 1980s-1990s - Backpropagation
  - 1998 - Convolutional Neural Networks (CNNs)
  - 2012 - AlexNet and the ImageNet competition
  - 2013 - Long Short-Term Memory (LSTM) for translation
  - 2014 - Generative Adversarial Networks (GANs)
  - 2015 - Residual Networks (ResNets)
  - 2015 - AlphaGo
  - 2017 - Transformers
  - 2020 - GPT-3

Importance of Deep Learning in ML and AI

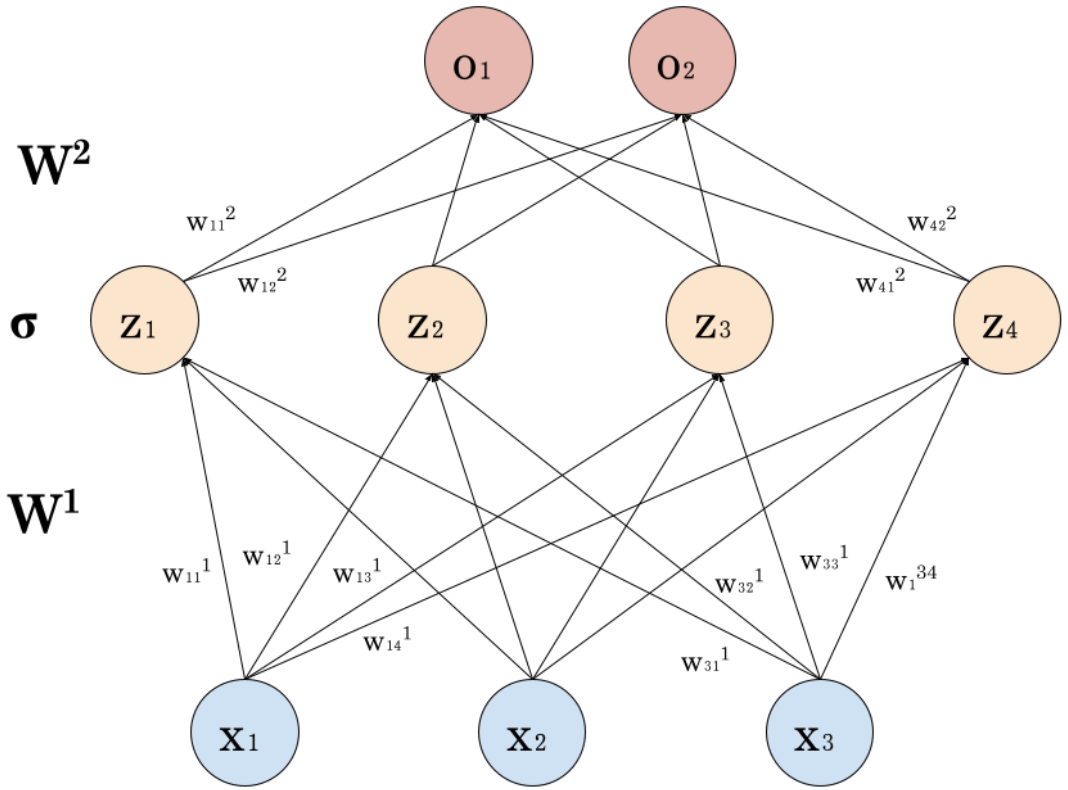
- **Subfield of Machine Learning:** Deep Learning is a subset of Machine Learning, a branch of Artificial Intelligence.
- **Diverse Applications:**
    - Image recognition
    - Natural language processing
    - Self-driving cars
    - Medical diagnosis, etc.

2. Deep Learning Foundations

Neural Networks

- Inspired by biological neurons
- Three types of layers: input $(x_i)_i$, hidden $(z_j)_j$, and output layers $(o_k)_k$
- **Weights and Biases**: Learning parameters in a neural network
  - Weights $W^l \overset{\text{def}}{=} (w^l_{i,j})_{i,j}$: strength of influence between nodes
  - Biases b_j: constant term added to the weighted sum of inputs

$$z_j \overset{\text{def}}{=} \sum_{i=1}^{n} w_{ij} x_i + b_j \qquad o_k \overset{\text{def}}{=} \sum_{j=1}^{m} w_{jk} \sigma(z_j) + b'_k$$



> If X is the design matrix, the NN (without the bias terms) is given by:
>
> $$\sigma(XW^1)W^2$$

Pedagogical sources: Michael Nielsen's book and 3Blue1Brown

- Michael Nielsen's "Neural Networks and Deep Learning"
- 3Blue1Brown's Deep Learning Video Series

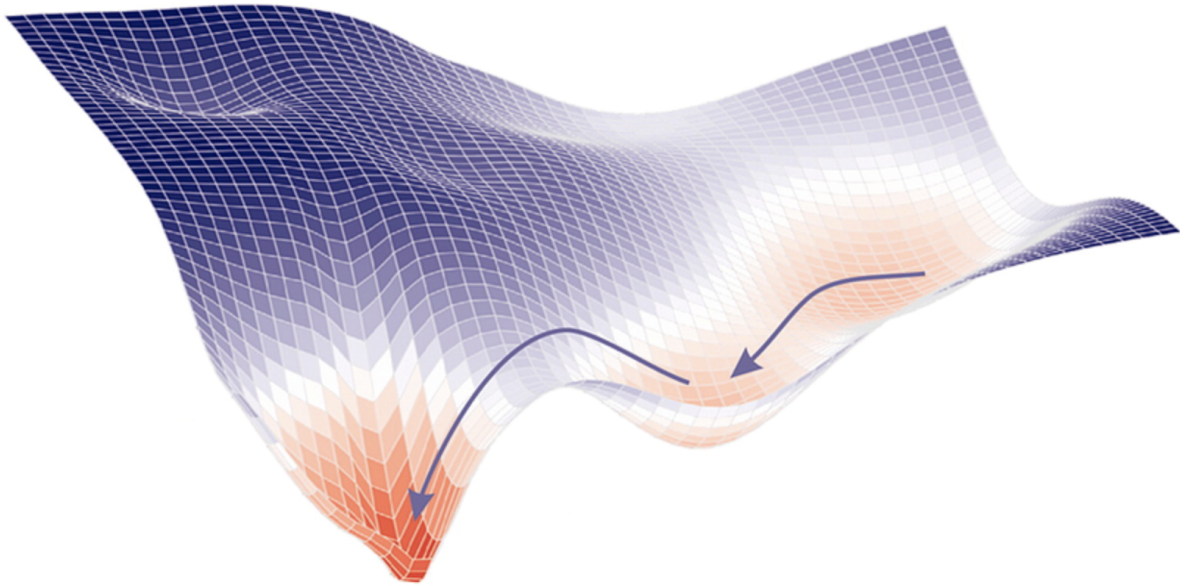
Other references

- Hinton, Osindero, Teh (2006): A fast learning algorithm for deep belief nets. *Neural computation*.
- LeCun, Bengio, Hinton (2015): Deep learning. *Nature*.
- Goodfellow, Bengio, Courville (2016): Deep learning. *MIT press*.

3. Backpropagation

Gradient Descent

- **Optimization algorithm** used to minimize the loss function
  - Iteratively moves in the direction of steepest descent
  - Core algorithm in DL for training models

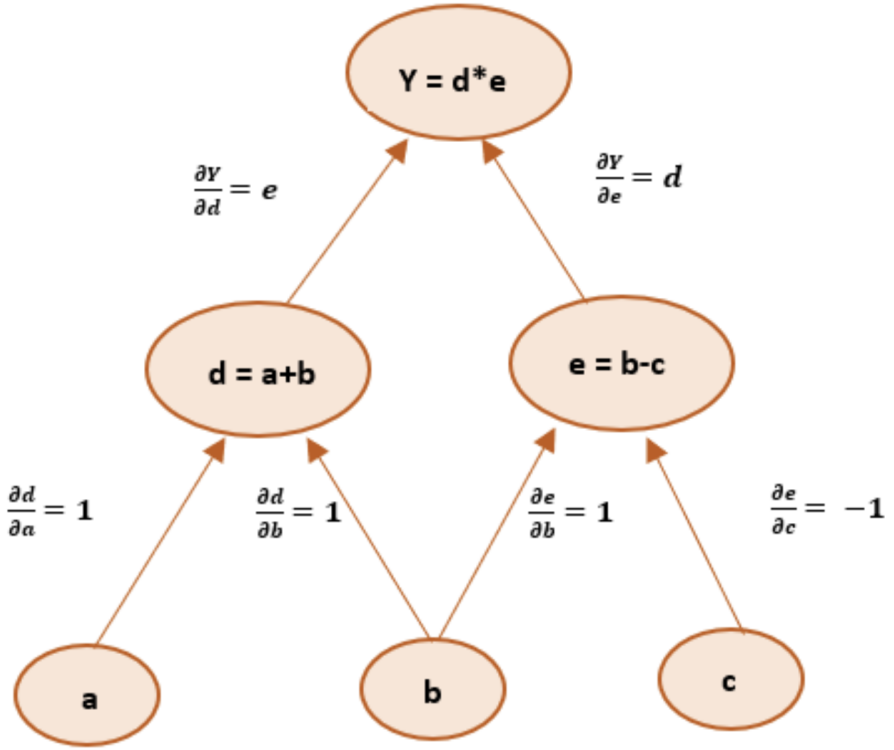


Flavours of Gradient Descent (GD)

- **Batch gradient descent:** Entire training dataset used to calculate the gradient at each step
- **Stochastic gradient descent:** Single training example used to calculate the gradient at each step
- **Mini-batch gradient descent:** Small batch of training examples used to calculate the gradient at each step
- *Other variants:* Momentum, Adaptive GD

Automatic Differentiation (AD)

- Techniques to compute derivatives of numeric functions expressed as computer programs
- Two modes of AD: *forward mode* and *reverse mode*



Backpropagation: a special case of revers mode AD

- Used to train neural networks by minimizing the prediction error
- Efficiently computes gradients for updating neural network weights
- Introduced in “Learning representations by back-propagating errors” (Rumelhart, Hinton, Williams, 1986)

Cognitive Sciences: Machine Learning as a Form of Automated Reasoning?

- Rescorla-Wagner rule (1972): Pavlovian conditioning
- Backpropagation and AD automate learning from errors
- Reflect on implications for understanding learning and intelligence

4. Scaling Neural Networks

Bias-Variance Tradeoff in Deep Learning

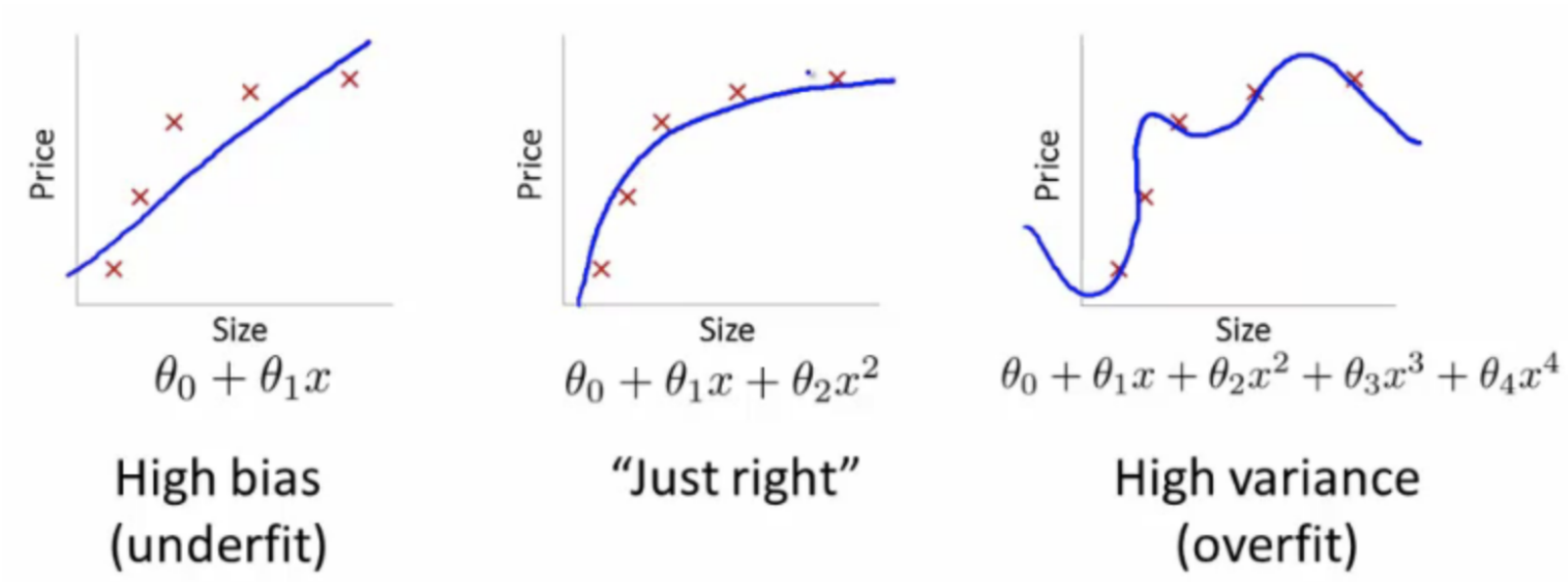
- **Bias:** Simplifying assumptions made by a model
- **Variance:** Estimate variability due to different training data
- **Bias-Variance Tradeoff:** Striking a balance between bias and variance to minimize total error

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

$$E\Big(\underbrace{y}_{\stackrel{\text{def}}{=} f(x) + \varepsilon} - \hat{f}(x) \Big)^2 = E(\hat{f}(x) - f(x))^2 + \text{Var}(\hat{\text{f}}(\text{x})) + \sigma^2$$

Introduction to Bias and Variance

- Two fundamental sources of error in predictive models
  - *High bias*: underfitting
  - *High variance*: overfitting



Bias-Variance Tradeoff in Deep Learning

- Tradeoff appears in the form of model capacity, regularization, and the amount of training data
- Increase in network size decreases bias but increases variance

Strategies for Handling Bias-Variance Tradeoff

- Techniques: cross-validation, dropout, batch/layer normalization, early stopping, or gathering more data
- Ensemble methods: combine multiple models' predictions for lower overall error

Optimizers and Learning Rate Schedulers

Introduction to Optimizers

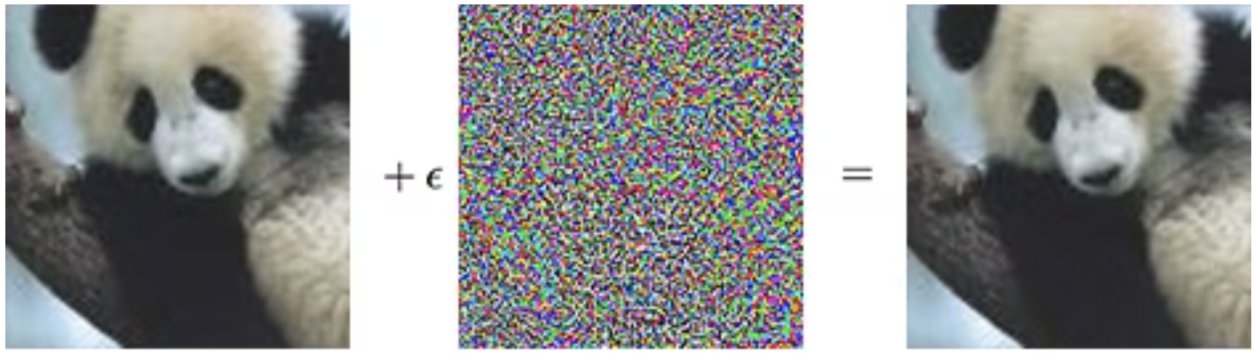
- Momentum: term added to the update rule, proportional to the previous update.
- Adaptive GD: different learning rate for each parameter:
  - frequently updated ⇒ smaller learning rate
  - not been updated frequently ⇒ larger learning rate

Learning Rate Schedulers

- Adjust learning rate during training
- Determine step size at each iteration while moving toward a (local) minimum of the loss function

6. Adversarial Attacks

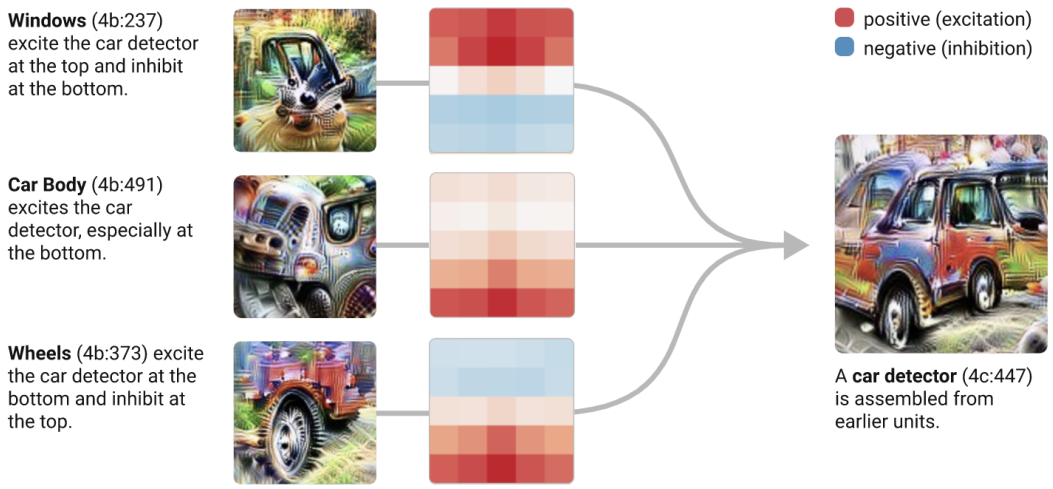
Small modifications to input data can cause wildly incorrect outputs: Goodfellow, Shlens, Szegedy (2014)



- Serious implications in sensitive areas such as facial recognition, autonomous vehicles, and cybersecurity
- Philosophically: what do NNs recognize?

7. Understanding Neural Networks: Visualisations and Type Theory

- Chris Olah's blog, Distill.pub
- Techniques like feature visualization show what a neural network learns and how it understands images



Functional programming for neural networks:

- Formal system to describe function behavior
- Can be used to describe behavior of neural networks in a precise, formal way

8. “Dark Matter of Intelligence”

Self-Supervised Learning (SSL)

- Model is trained on unlabeled data: learns to predict missing parts of the input data (eg. next word in a sentence or part of an image)

SSL has several advantages over supervised learning:

1. Does not require labeled data (expensive and time-consuming)
2. Can be used to train models on very large datasets (better performance)
3. Can be used to train models on data that is not easily labeled (eg. natural language text)

⟶ Yann LeCun: “Dark Matter of Intelligence”

9. Introduction to Attention Mechanisms

Attention Mechanisms in a Nutshell

- Focus on relevant parts of input data to make decisions
- Improves performance in various applications, especially in NLP tasks

How Attention Works

- Computes a score for each item in the input sequence
- Scores determine focus when producing