Course Introduction

[CSED490X] Recent Trends in ML: A Large-Scale Perspective

Jungtaek Kim
jtkim@postech.ac.kr

POSTECH
Pohang 37673, Republic of Korea
https://jungtaek.github.io

February 23, 2022

Table of Contents

Course Introduction

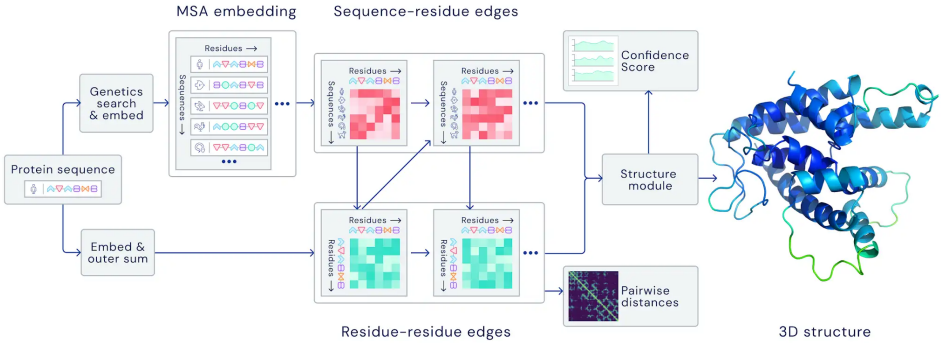
Recent Trends in Machine Learning: A Large-Scale Perspective

- Introduce recent trends in machine learning.
- In particular, focus on a large-scale machine learning model.
- Help to understand cutting-edge technologies in artificial intelligence.
- Provide an understanding of primary requirements in computer science and artificial intelligence.
- Additionally, have special sessions by Kakao Brain.

Why Do We Need to Study a Large-Scale Model?

- A large-scale model is a cutting-edge technology in artificial intelligence.
- It is developed by utilizing a collection of techniques in diverse subfields of computer science.
- It will be applied to novel use-cases in many areas, e.g., chemistry, biology, and mathematics.
- It will demonstrate the effectiveness of machine learning and deep learning.

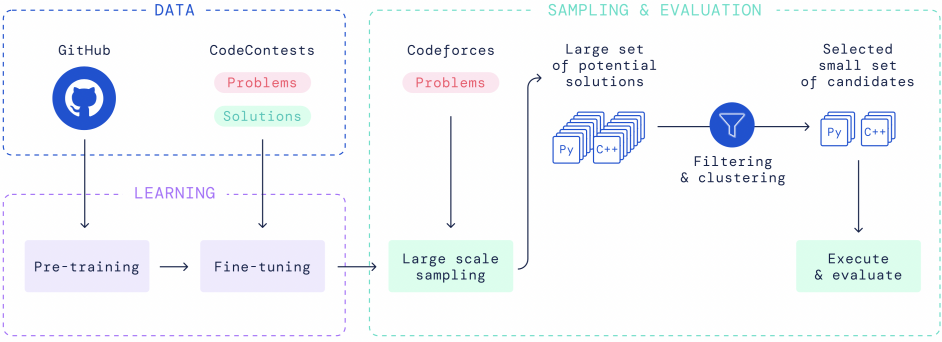
Recent Progress: AlphaFold



Taken from https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

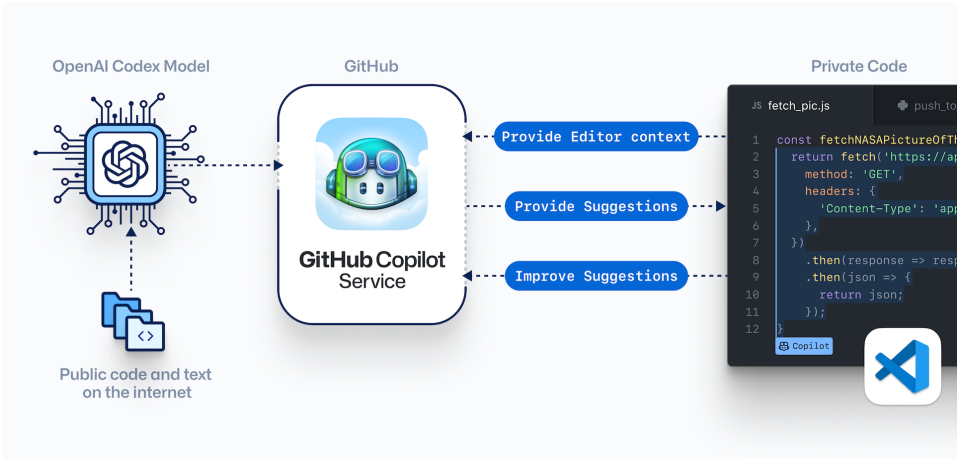
Recent Progress: AlphaCode

- https://alphacode.deepmind.com/



Taken from https://deepmind.com/blog/article/Competitive-programming-with-AlphaCode.

Recent Progress: GitHub Copilot

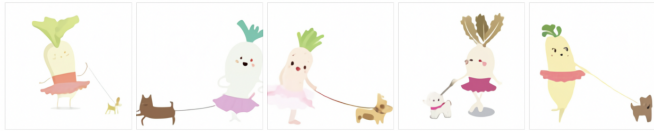


Taken from https://copilot.github.com.

Recent Progress: DALL·E

TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES



Edit prompt or view more images↓

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

Taken from https://openai.com/blog/dall-e/.

Drawbacks and Future Concerns on Large-Scale Models

- A large-scale model is too large to execute, e.g., GPT-3 has 175 billion parameters.
- It requires huge computational resources to train a model, e.g., the cost for training GPT-3 using a cloud service is about $4.6M.
- A large-scale model is vulnerable to ethical issues, e.g., racism.
- Carbon footprint, caused by training and testing a large-scale model, accelerates climate change.

https://lambdalabs.com/blog/demystifying-gpt-3/

Course Introduction

- Class time: Wednesday 15:30 - 16:20
- Instructor: Jungtaek Kim (Email: jtkim@postech.ac.kr)
- Assessment: Letter grade, 70% homeworks and 30% class participation
- Language: Korean
- Teaching assistant: Dayoung Kong (Email: dayoung.kong@postech.ac.kr)

Schedule (Tentative)

Date	Contents
02/23	Course introduction
03/02	Basics of machine learning
03/09	Presidential election day (No lecture)
03/16	Basics of development tools & environments
03/23	Advances in large-scale language models: Transformer
03/30	Advances in large-scale language models: BERT
04/06	Advances in large-scale language models: GPT-1, GPT-2, GPT-3
04/13	Midterm exams period (No lecture)

Schedule (Tentative)

Date	Contents
04/20	Advances in large-scale language models: DistilBERT, RoBERTa, T-NLG
04/27	Advances in large-scale vision models: Vision Transformer, Swin Transformer
05/04	Advances in large-scale vision-and-language models: TBD (by Kakao Brain)
05/11	Advances in large-scale vision-and-language models: TBD (by Kakao Brain)
05/18	Advances in large-scale vision-and-language models: TBD (by Kakao Brain)
05/25	Advances in large-scale vision models: TBD
06/01	Other large-scale models
06/08	Final exams period (No lecture)

Assessment

- Homeworks (70%): Six one-page reports during this semester, submitting them through `https://plms.postech.ac.kr`.
- Class participation (30%): Turning on your video.

Brief Introduction

What Is Artificial Intelligence?

Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985)
"[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998)
"The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Figure 1.1 Some definitions of artificial intelligence, organized into four categories.

Figure 1: Four definitions of artificial intelligence.

Figure 1 is taken from [Russell and Norvig, 2010].

[Russell and Norvig, 2010] S. J. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Pearson Education, 3 edition, 2010.

What Is Artificial Intelligence?

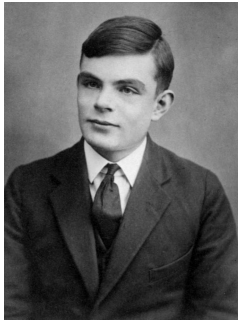


Figure 2: Alan Turing.

- Acting humanly: The Turing Test approach
  1. **natural language processing** to enable it to communicate successfully in human language;
  2. **knowledge representation** to store what it knows or hears;
  3. **automated reasoning** to use the stored information to answer questions and to draw new conclusions;
  4. **machine learning** to adapt to new circumstances and to detect and extrapolate patterns;
  5. **computer vision** to perceive objects;
  6. **robotics** to manipulate objects and move about.

Figure 2 is taken from Wikipedia.

[Russell and Norvig, 2010] S. J. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Pearson Education, 3 edition, 2010.

What Is Artificial Intelligence?

- Thinking humanly: The cognitive modeling approach

  1. Through introspection – trying to catch our own thoughts as they go by;
  2. Through psychological experiments – observing a person in action;
  3. Through brain imaging – observing the brain in action.

- Thinking rationally: The "laws of thought" approach

  - The Greek philosopher Aristotle was one of the first to attempt to codify "right thinking;"
  - His syllogisms provided patterns for argument structures that always yielded correct conclusions when given correct premises – for example, "Socrates is a man; all men are mortal; therefore Socrates is mortal;"
  - By 1965, programs existed that could, in principle, solve any solvable problem described in logical notation;
  - Although if no solution exists, the program might loop forever.

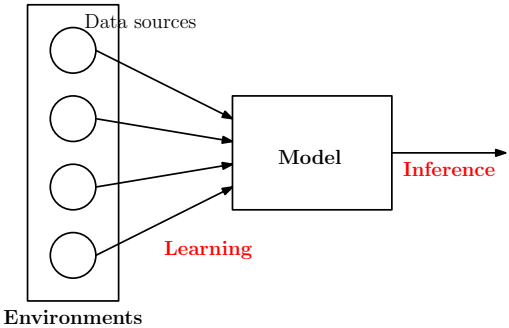
[Russell and Norvig, 2010] S. J. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Pearson Education, 3 edition, 2010.

What Is Artificial Intelligence?

- Acting rationally: The rational agent approach
  - An agent is just something that acts – all computer programs do something, but such an agent is expected to do more;
  - A rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome;
  - Making correct inferences by the "laws of thought" approach is sometimes part of being a rational agent, but correct inference is not all rationality;
  - All the skills needed for the Turing Test also allow an agent to act rationally;
  - Finally, the rational agent approach has two advantages over the other approaches
    1. It is more general than the "laws of thought" approach because correct inference is just one of several possible mechanisms for achieving rationality;
    2. It is more amenable to scientific development than the approaches based on human behavior or human thought.

[Russell and Norvig, 2010] S. J. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Pearson Education, 3 edition, 2010.

What Is Machine Learning?



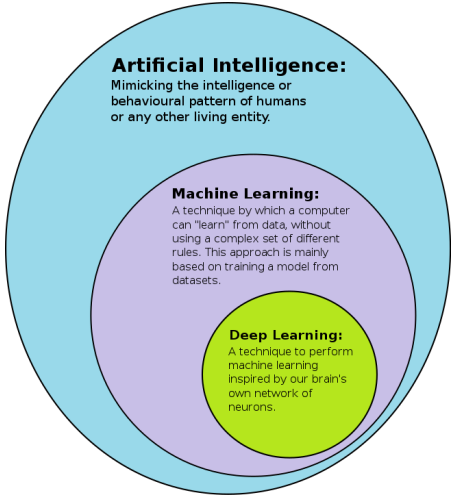
- Machine learning is a data-driven method for artificial intelligence.
- Three key ingredients in machine learning
  1. Data;
  2. A machine learning model;
  3. A learning algorithm.
- Details will be covered in the next lecture.

What Is Machine Learning?

	Feedback	**Goal**
Supervised learning	Instructive feedback	Regression & classification
Unsupervised learning	No feedback	Representation learning & clustering
Reinforcement learning	Evaluative feedback	Sequential decision making

Taken from the slides created by Prof. Seungjin Choi.

Deep Learning in Machine Learning



Taken from Wikipedia.

Deep Learning Revolution

- Resurgence of deep learning has been started in 2006 by [Hinton et al., 2006].
- The following advances allow deep learning to attract great attention from diverse research fields and industries.
  1. Cheap computing resources, i.e., GPU;
  2. Huge data, i.e., social networks and streaming media;
  3. Improvements in machine learning algorithms, i.e., ReLU and ADAM.
- Deep learning starts to show its effectiveness in various fields such as speech recognition [Graves et al., 2013], computer vision [Krizhevsky et al., 2012], machine translation [Kalchbrenner and Blunsom, 2013, Sutskever et al., 2014].

[Hinton et al., 2006] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. Neural Computation, 18(7):1527–1554, 2006.

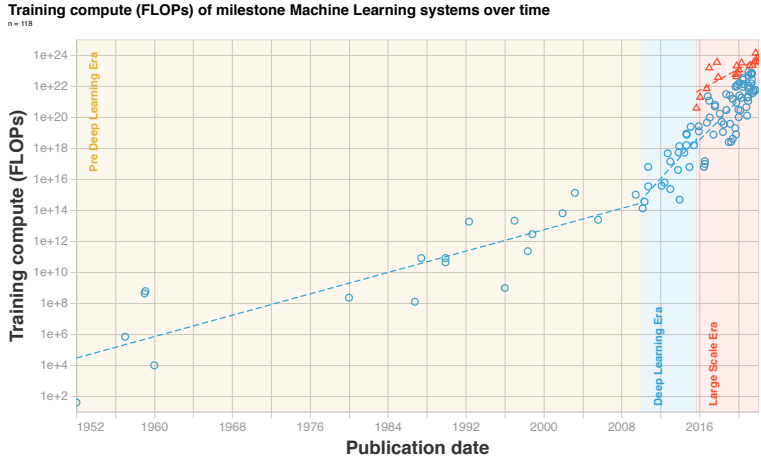
[Graves et al., 2013] A. Graves, A.-r. Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 6645–6649, 2013.

[Krizhevsky et al., 2012] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NeurIPS), volume 25, Lake Tahoe, Nevada, USA, 2012.

[Kalchbrenner and Blunsom, 2013] N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1700–1709, 2013.

[Sutskever et al., 2014] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (NeurIPS), volume 27, Montreal, Quebec, Canada, 2014.

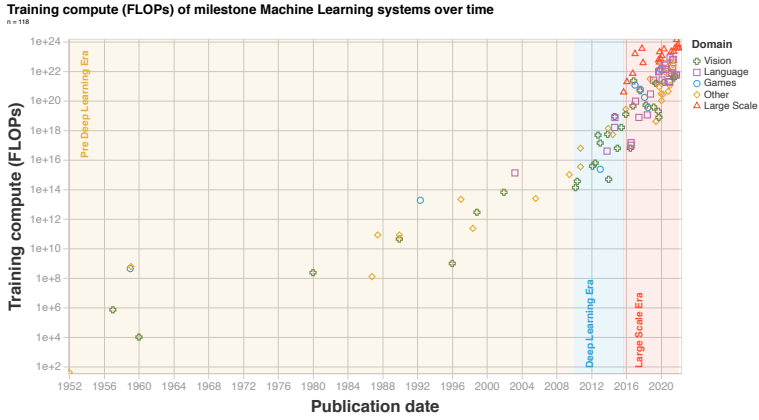
Deep Learning to Large-Scale Models



Taken from [Sevilla et al., 2022].

[Sevilla et al., 2022] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. Compute trends across three eras of machine learning. arXiv preprint arXiv:2202.05924, 2022.

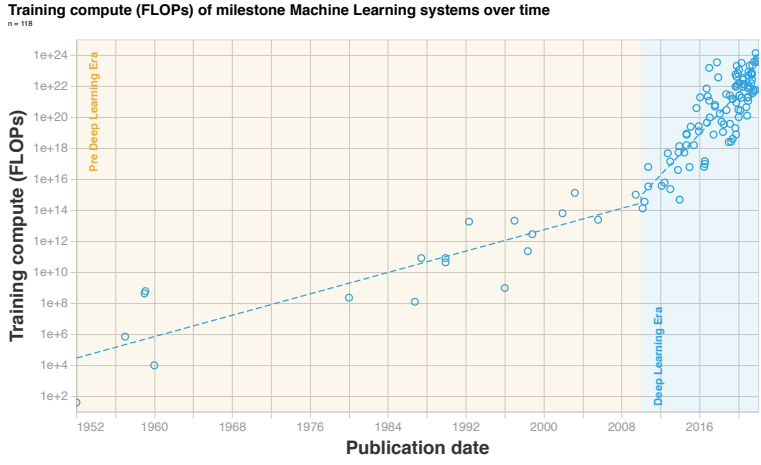
Deep Learning to Large-Scale Models



Taken from [Sevilla et al., 2022].

[Sevilla et al., 2022] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. Compute trends across three eras of machine learning. arXiv preprint arXiv:2202.05924, 2022.

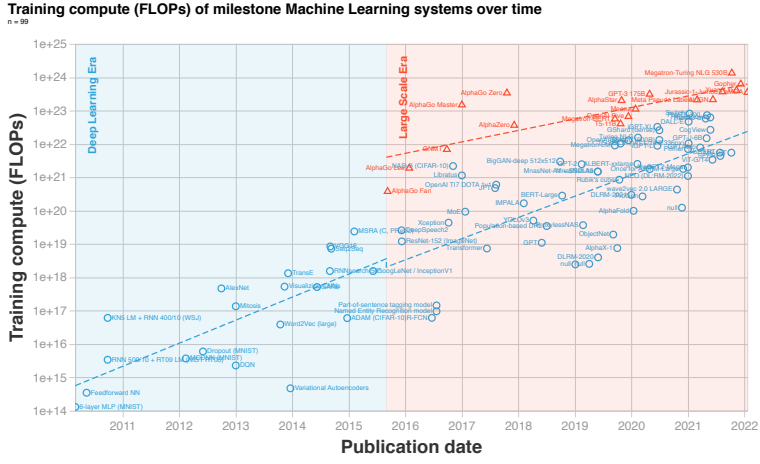
Deep Learning to Large-Scale Models



Taken from [Sevilla et al., 2022].

[Sevilla et al., 2022] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. Compute trends across three eras of machine learning. arXiv preprint arXiv:2202.05924, 2022.

Deep Learning to Large-Scale Models



Taken from [Sevilla et al., 2022].

[Sevilla et al., 2022] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. Compute trends across three eras of machine learning. arXiv preprint arXiv:2202.05924, 2022.

Gradients

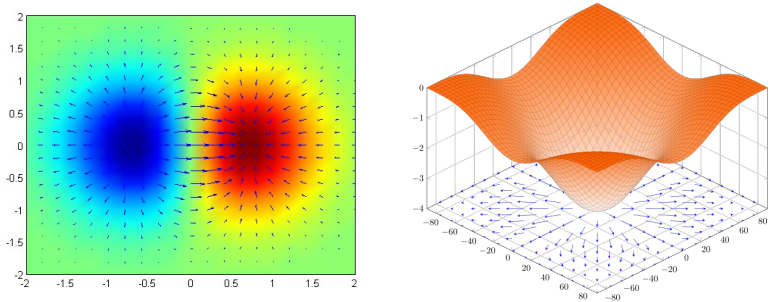


Figure 3: Illustration of gradients.

- Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{x} = [x_1, \ldots, x_d] \in \mathbb{R}^d$. The gradient of a function f at $\mathbf{x}$ is

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_d} \right] \in \mathbb{R}^d. \tag{1}$$

Figure 3 is taken from Wikipedia.

Automatic Differentiation

- Automatic differentiation evaluates the derivative of a function where a function is expressed by a sequence of elementary arithmetic operations (e.g., addition, subtraction, multiplication, and division) and elementary functions (e.g., exp, log, sin, and cos).
- Two distinct modes exist:
  1. Forward accumulation
  2. Reverse accumulation

$$\frac{\mathrm{d}w_i}{\mathrm{d}x} = \frac{\mathrm{d}w_i}{\mathrm{d}w_{i-1}} \frac{\mathrm{d}w_{i-1}}{\mathrm{d}x}, \tag{2}$$

$$\frac{\mathrm{d}y}{\mathrm{d}w_i} = \frac{\mathrm{d}y}{\mathrm{d}w_{i+1}} \frac{\mathrm{d}w_{i+1}}{\mathrm{d}w_i}. \tag{3}$$

Automatic Differentiation

- We have already known many derivatives of simple functions:

$$\frac{\mathrm{d}x^2}{\mathrm{d}x} = ???, \tag{4}$$

$$\frac{\mathrm{d}\sin(x)}{\mathrm{d}x} = ???, \tag{5}$$

$$\frac{\mathrm{d}\cos(x)}{\mathrm{d}x} = ???, \tag{6}$$

$$\frac{\mathrm{d}\tan(x)}{\mathrm{d}x} = ???, \tag{7}$$

$$\frac{\mathrm{d}\exp(x)}{\mathrm{d}x} = ???, \tag{8}$$

$$\frac{\mathrm{d}(f(x)+g(x))}{\mathrm{d}x} = ???, \tag{9}$$

$$\frac{\mathrm{d}f(x)g(x)}{\mathrm{d}x} = ???, \tag{10}$$

$$\frac{\mathrm{d}f(g(x))}{\mathrm{d}x} = ???, \tag{11}$$

$$\frac{\mathrm{d}\sin(f(x))}{\mathrm{d}x} = ???, \tag{12}$$

$$\frac{\mathrm{d}\exp(f(x))}{\mathrm{d}x} = ???. \tag{13}$$

Automatic Differentiation

- We have already known many derivatives of simple functions:

$$\frac{\mathrm{d}x^2}{\mathrm{d}x} = 2x, \tag{14}$$

$$\frac{\mathrm{d}\sin(x)}{\mathrm{d}x} = \cos(x), \tag{15}$$

$$\frac{\mathrm{d}\cos(x)}{\mathrm{d}x} = -\sin(x), \tag{16}$$

$$\frac{\mathrm{d}\tan(x)}{\mathrm{d}x} = \sec^2(x), \tag{17}$$

$$\frac{\mathrm{d}\exp(x)}{\mathrm{d}x} = \exp(x), \tag{18}$$

$$\frac{\mathrm{d}(f(x)+g(x))}{\mathrm{d}x} = f'(x)+g'(x), \tag{19}$$

$$\frac{\mathrm{d}f(x)g(x)}{\mathrm{d}x} = f(x)g'(x)+f'(x)g(x), \tag{20}$$

$$\frac{\mathrm{d}f(g(x))}{\mathrm{d}x} = f'(g(x))g'(x), \tag{21}$$

$$\frac{\mathrm{d}\sin(f(x))}{\mathrm{d}x} = \cos(f(x))f'(x), \tag{22}$$

$$\frac{\mathrm{d}\exp(f(x))}{\mathrm{d}x} = \exp(f(x))f'(x). \tag{23}$$

Automatic Differentiation

- An objective function is

$$f(x_1, x_2) = x_1 x_2 + \sin(x_1). \tag{24}$$

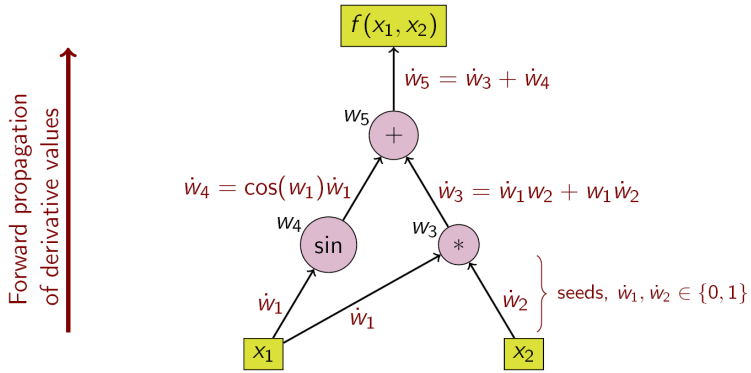


Figure 4: Computational graph of (24).

Figure 4 is taken from Wikipedia.

Automatic Differentiation

- An objective function is

$$f(x_1, x_2) = x_1 x_2 + \sin(x_1). \tag{25}$$

- Suppose that we are given the followings:

$$w_1 = x_1, \quad w_2 = x_2, \quad w_3 = w_1 w_2, \quad w_4 = \sin(w_1), \quad w_5 = w_3 + w_4. \tag{26}$$

- Then,

$$\begin{aligned} \frac{\mathrm{d}f(x_1, x_2)}{\mathrm{d}x_1} &= \dot{w}_5 \\ &= \dot{w}_3 + \dot{w}_4 \\ &= w_1 \dot{w}_2 + \dot{w}_1 w_2 + \cos(w_1)\dot{w}_1 \\ &= x_2 + \cos(x_1). \end{aligned} \tag{27}$$

Automatic Differentiation

- An objective function is

$$f(x_1, x_2) = x_1 x_2 + \sin(x_1). \tag{28}$$

- Suppose that we are given the followings:

$$w_1 = x_1, \quad w_2 = x_2, \quad w_3 = w_1 w_2, \quad w_4 = \sin(w_1), \quad w_5 = w_3 + w_4. \tag{29}$$

- Then,

$$\begin{aligned} \frac{\mathrm{d}f(x_1, x_2)}{\mathrm{d}x_2} &= \dot{w}_5 \\ &= \dot{w}_3 + \dot{w}_4 \\ &= w_1 \dot{w}_2 + \dot{w}_1 w_2 + \cos(w_1)\dot{w}_1 \\ &= x_1. \end{aligned} \tag{30}$$

Any Questions?

References I

A. Graves, A. r. Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6645–6649, 2013.

G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1700–1709, 2013.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, Lake Tahoe, Nevada, USA, 2012.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 3 edition, 2010.

J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. Compute trends across three eras of machine learning. *arXiv preprint arXiv:2202.05924*, 2022.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, Montreal, Quebec, Canada, 2014.