BOSTON UNIVERSITY

Deep Learning for Data Science
DS 542

https://dl4ds.github.io/fa2025/

Introduction and Course Overview

Original slides by Tom Gardos,
other content from Understanding Deep Learning unless otherwise cited

Plan for Today

- Applications of Deep Learning
- Why Deep Learning?
- How We Figured out Deep Learning
- Course Logistics

What Interested You about this Class?

???

Predictions, forecasting

Backends of NLP

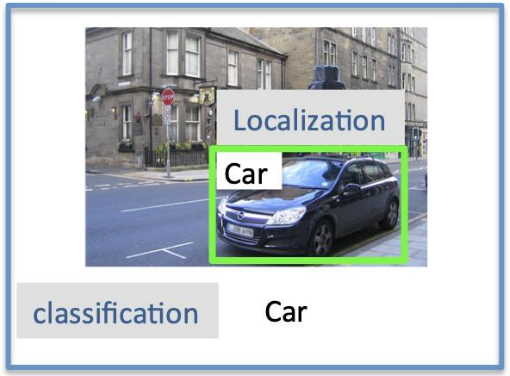
Architectures + capabilities + why?

ImageNet Challenge 2012

Task 1: Classification

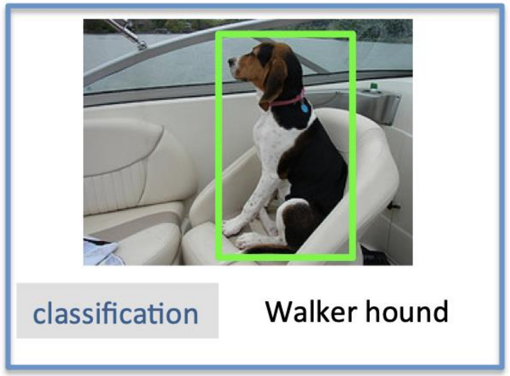
- Predict a class label
- 5 predictions / image
- 1000 classes
- 1,200 images per class for training
- Bounding boxes for 50% of training.

Task 2: Detection (Classification + Localization)



- Predict a class label and a bounding box
- 5 predictions / image
- 1000 classes
- 1,200 images per class for training
- Bounding boxes for 40% of training.

Task 3: Fine-grained classification



- Predict a class label given a bounding box in test
- 1 prediction / image
- 120 dog classes (subset)
- ~200 images per class for training (subset)
- Bounding boxes for 100% of training

Source: https://www.image-net.org/static_files/files/ilsvrc2012.pdf

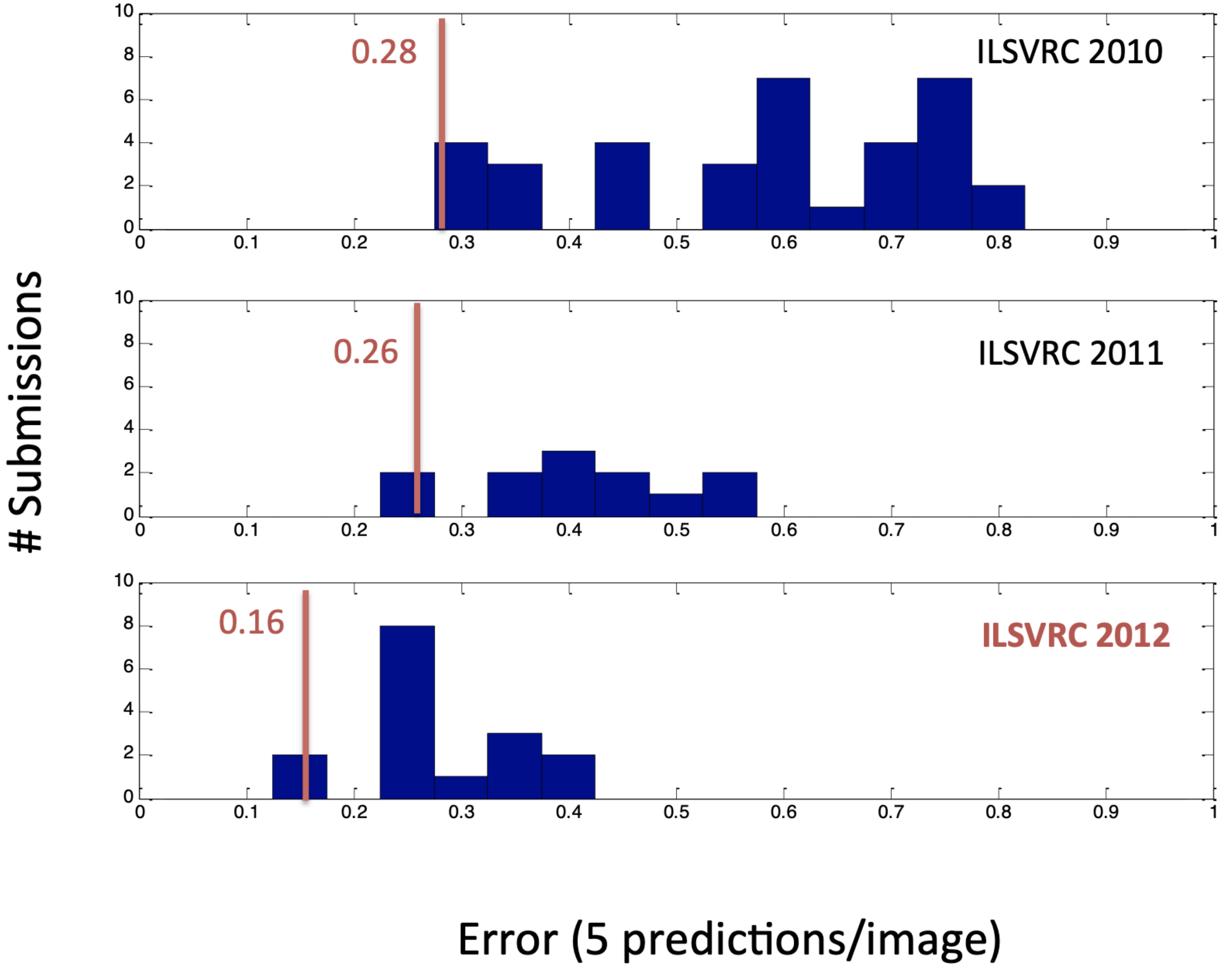
ImageNet Categories

Breakdown of categories (samples):

Subtree	# of leaf categories
Instrument	358
Canine	130
Covering	90
Vehicle	67
Invertebrate	61
Bird	59
Structure (construction)	58
Food	27
…	

Source: https://www.image-net.org/static_files/files/ilsvrc2012.pdf

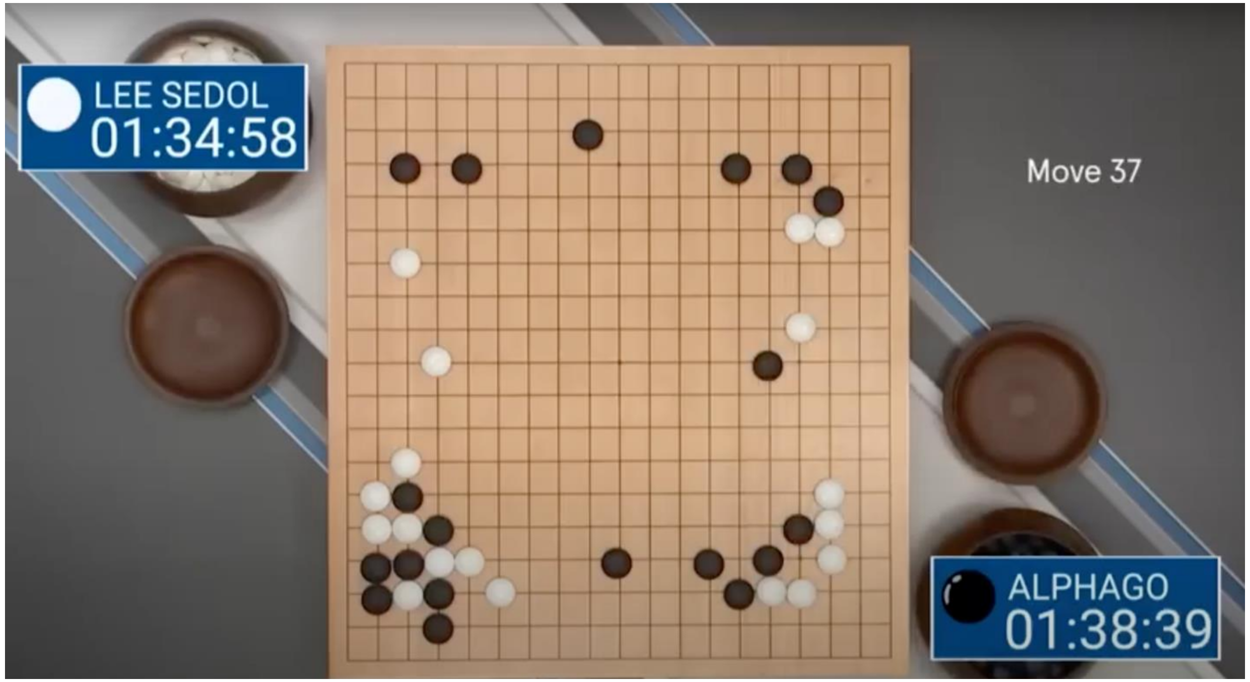
2012 Results (Classification)



Source: https://www.image-net.org/static_files/files/ilsvrc2012.pdf

AlphaGo (2016)

Google's Computer Program Beats Lee Se-dol in Go Tournament



Sources:

https://www.nytimes.com/2016/03/10/world/asia/google-alphago-lee-se-dol.html

Lee Sedol vs AlphaGo Move 37 reactions and analysis

Emotion Identification (2019)

- 42 muscles control all possible expressions
- Restrictions on how faces and heads look subject to physics of illumination and reflectance, etc.

- The “manifold” of possible faces is much, much smaller than the combinatoric collection of pixel values

C. A. C. Holland, N. C. Ebner, T. Lin, and G. R. Samanez-Larkin, “Emotion identification across adulthood using the Dynamic FACES database of emotional expressions in younger, middle aged, and older adults,” *Cognition and Emotion*, vol. 33, no. 2, pp. 245–257, Feb. 2019, doi: 10.1080/02699931.2018.1445981.

GPT-2 (2019)

System Prompt (human-written)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model Completion (machine-written, 10 tries)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

https://openai.com/index/better-language-models/

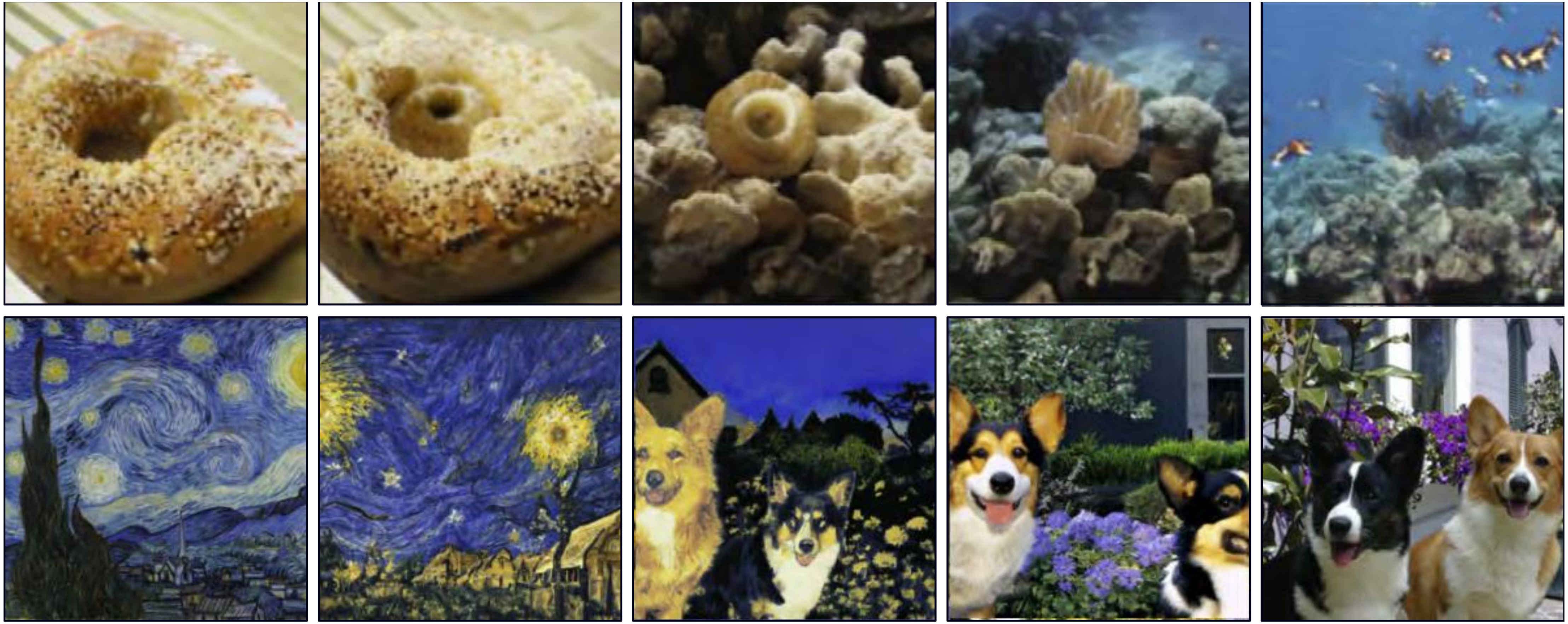
This Person Does Not Exist (2020)

https://thispersondoesnotexist.com

Latent Diffusion (2021)

https://arxiv.org/abs/2112.10752

Image Interpolation (2022)



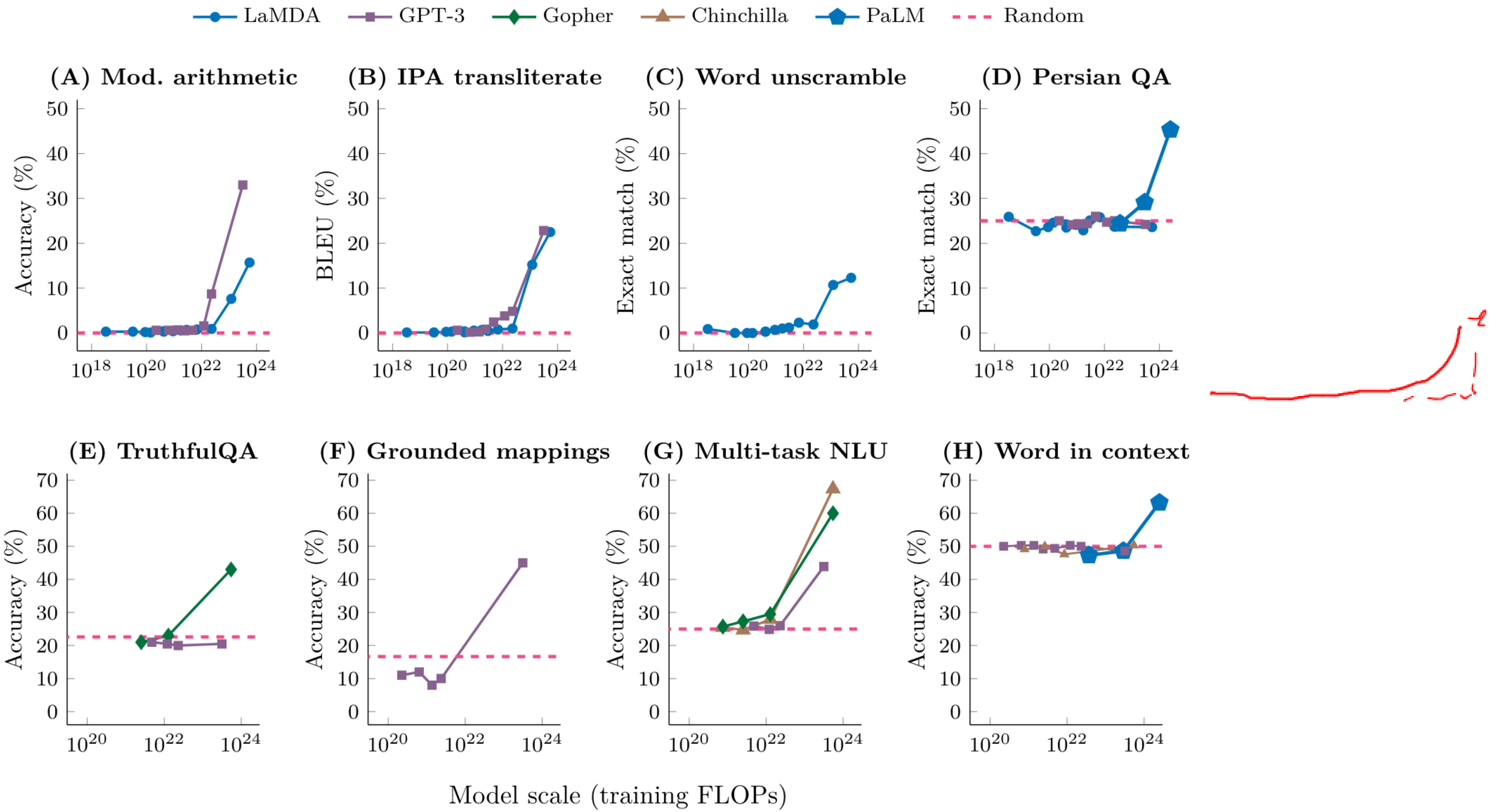
Axel Sauer, Katja Schwarz, and Andreas Geiger. 2022. StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets. In ACM SIGGRAPH 2022 Conference Proceedings (SIGGRAPH '22). Association for Computing Machinery, New York, NY, USA, Article 49, 1–10. https://doi.org/10.1145/3528233.3530738

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP Latents. arXiv:2204.06125

Conditional synthesis (2022)

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., & Norouzi, M. (2022a). Palette: Image-to-image diffusion models. ACM SIGGRAPH, (link)

Emergent Abilities of Language Models (2022)

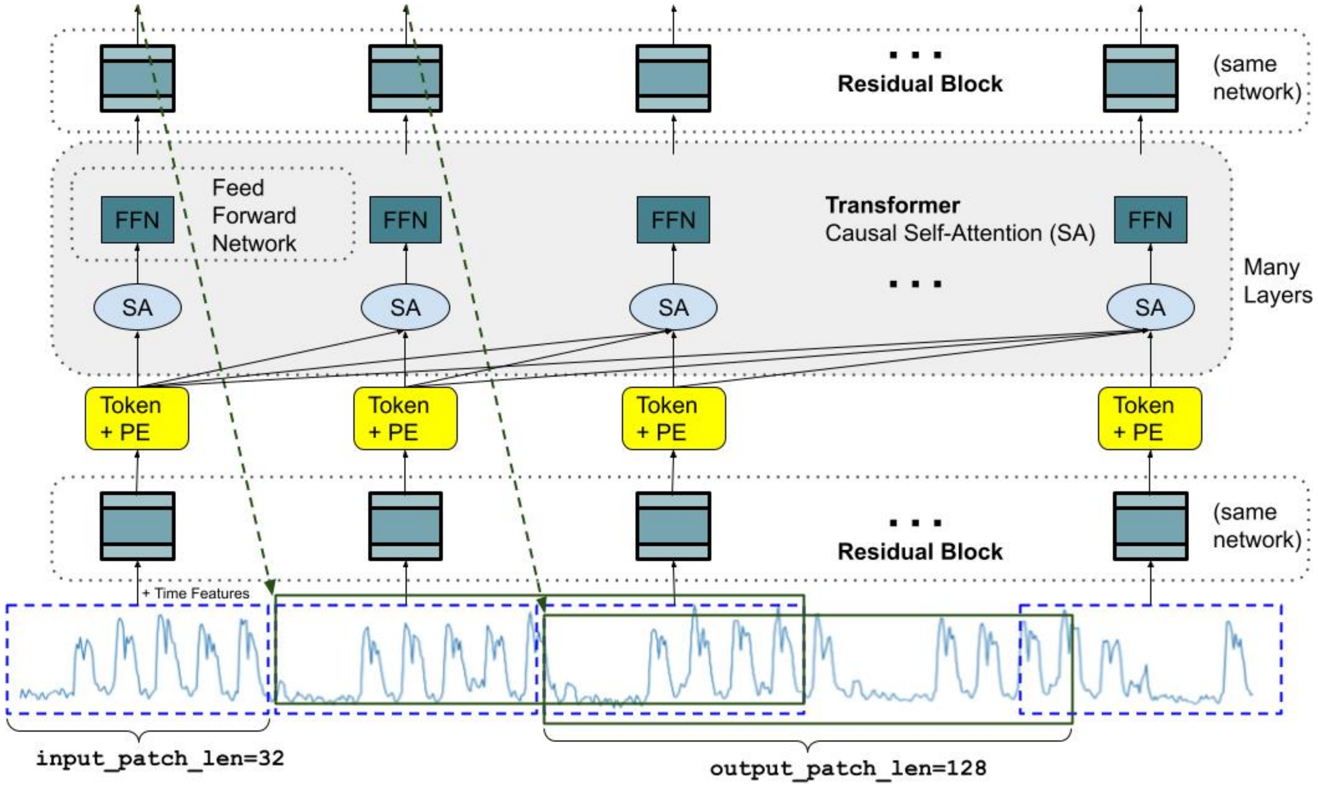


"Emergent Abilities of Large Language Models."(https://arxiv.org/abs/2206.07682) J. Wei et al., Oct. 26, 2022.

Segment Anything (2023)

Source: https://segment-anything.com

Time Series Forecasting (2024)



https://research.google/blog/a-decoder-only-foundation-model-for-time-series-forecasting/

Image/Video/Music Generation (2024)

A teenage superhero fighting crime in an urban setting shown in the style of claymation.

https://sora.com

Write a short pop song about students wanting to learn about neural networks and do great things with them.

What do these examples have in common?

- Very complex relationship between input and output
- Sometimes may be many possible valid answers
- But outputs (and sometimes inputs) obey rules

"A Kazakh man on a horse holding a bird of prey"

Language obeys grammatical rules

Natural images also have "rules"

Any Questions?

Why Deep Learning?

- Why do we need deep learning for these problems?

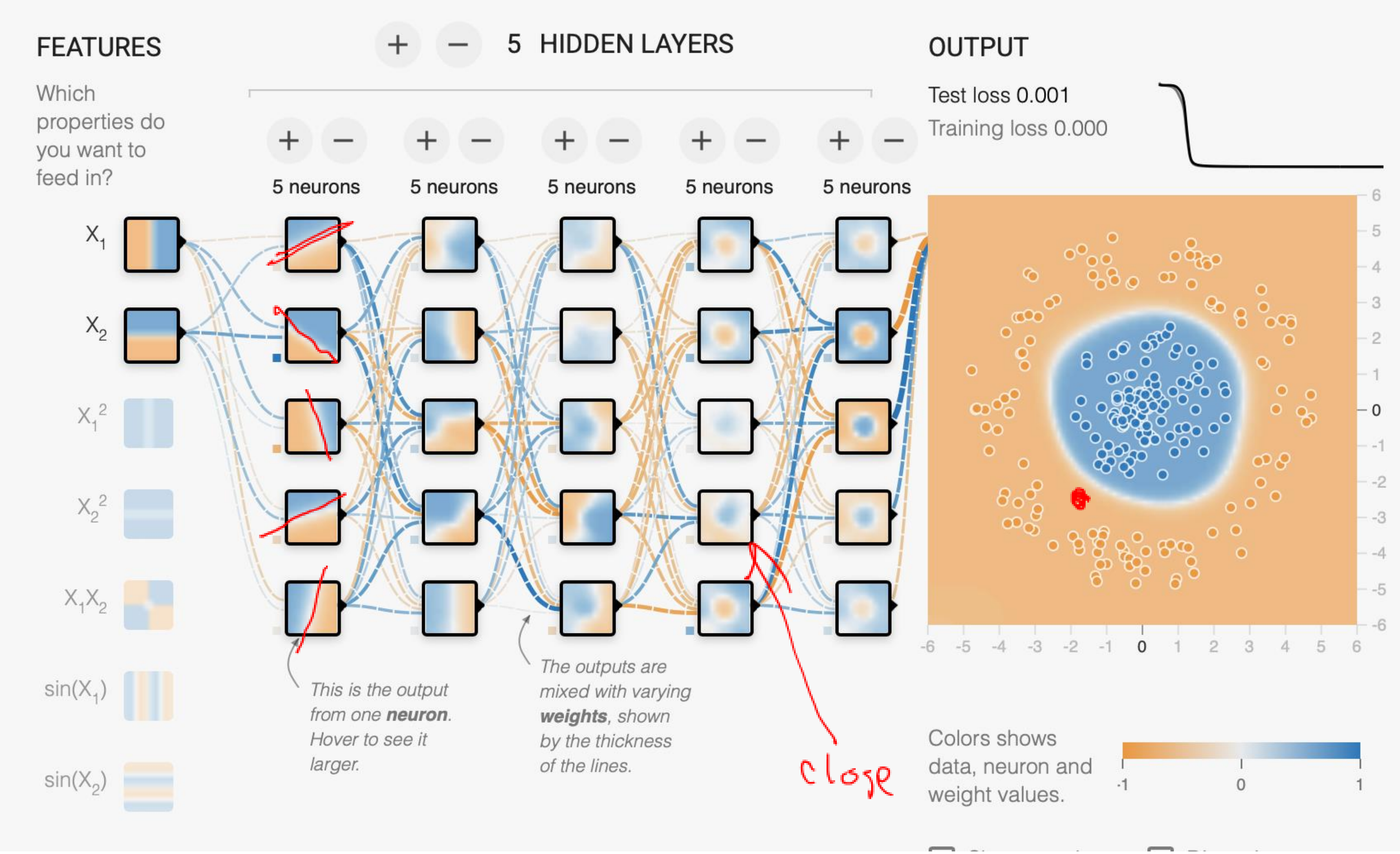
Universal Function Approximation

- A big enough neural network can approximate any function.
  - Does not require deep learning.

- To fit or overfit?

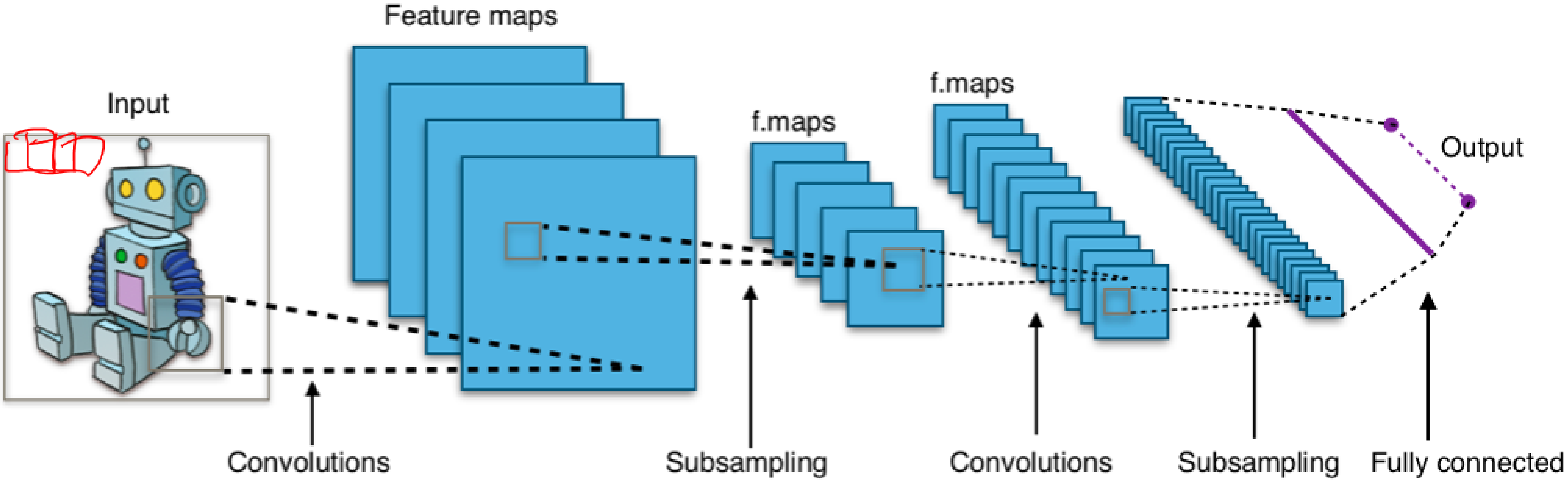
Layers of Abstraction

- Many layers may map to progressively developing concepts or abstractions.



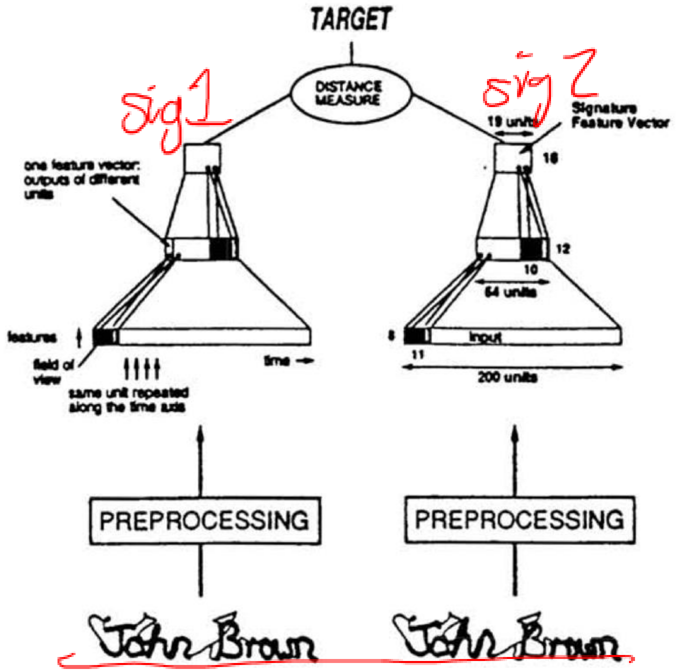
http://playground.tensorflow.org/

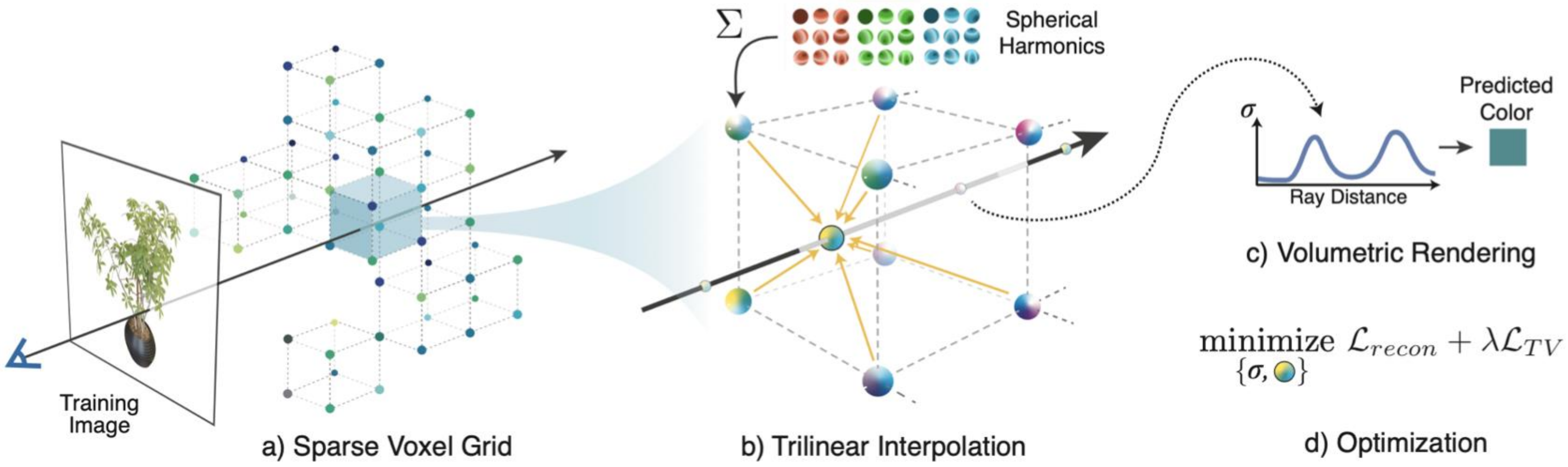
Parameter Sharing



Source: https://en.wikipedia.org/wiki/Convolutional_neural_network

Composability





Sources:

[Signature Verfication using a “Siamese” Time Delay Neural Network (1993)]

[Plenoxels: Radiance Fields without Neural Networks (2022)]

Joint Optimization

Do all of these at once –

- Universal function approximation
- Layers of abstraction
- Parameter sharing
- Composability

Differentiable Computations

- Can the partial derivative of “output quality” be computed with respect to every input and every parameter in the system?
  - Parameter sharing
  - Composability
  - Joint Optimization

- Gradient descent as a universal algorithm!

Deep Learning with Neural Networks

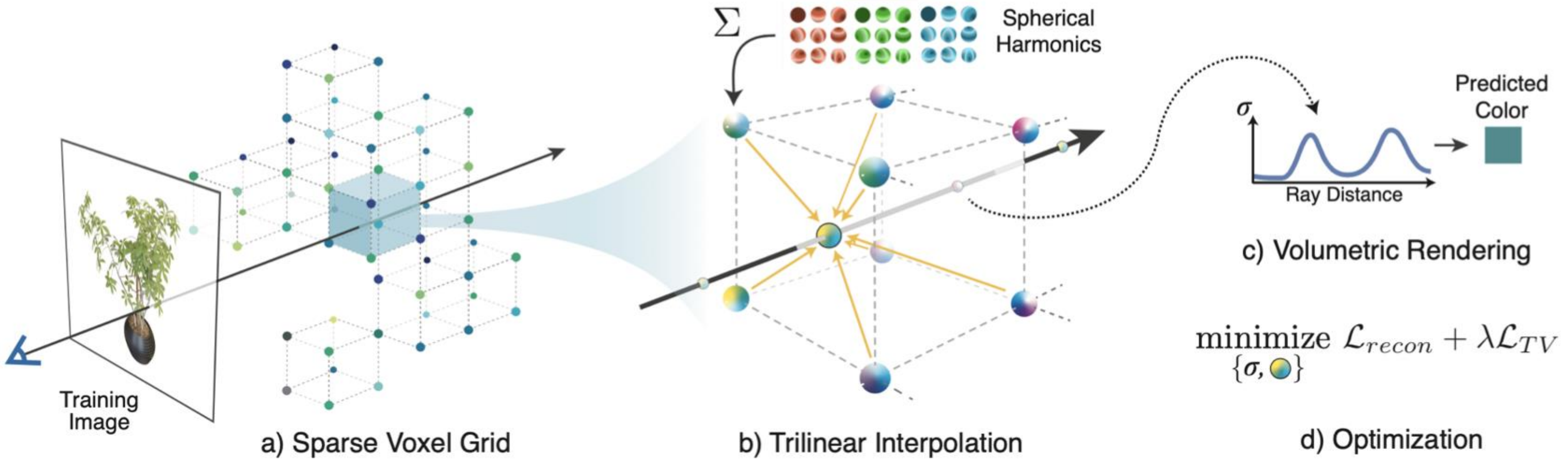
- Universal function approximation
  - Deep neural networks tend to be able to fit more complex patterns with fewer parameters.
  - Some toy problems have probable exponential gaps between shallow and deep network sizes.

- Layers of abstraction
  - Happens automatically but many parts not well understood.

- Train with gradient descent!

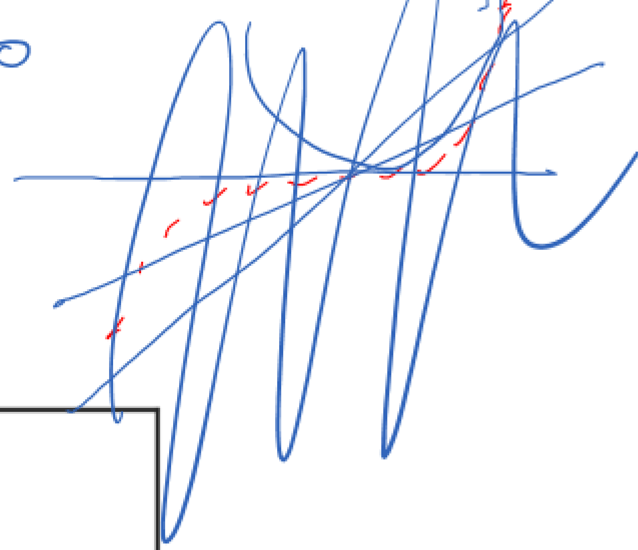
Deep Learning without Neural Networks

- Universal function approximation and layers of abstraction must be designed?

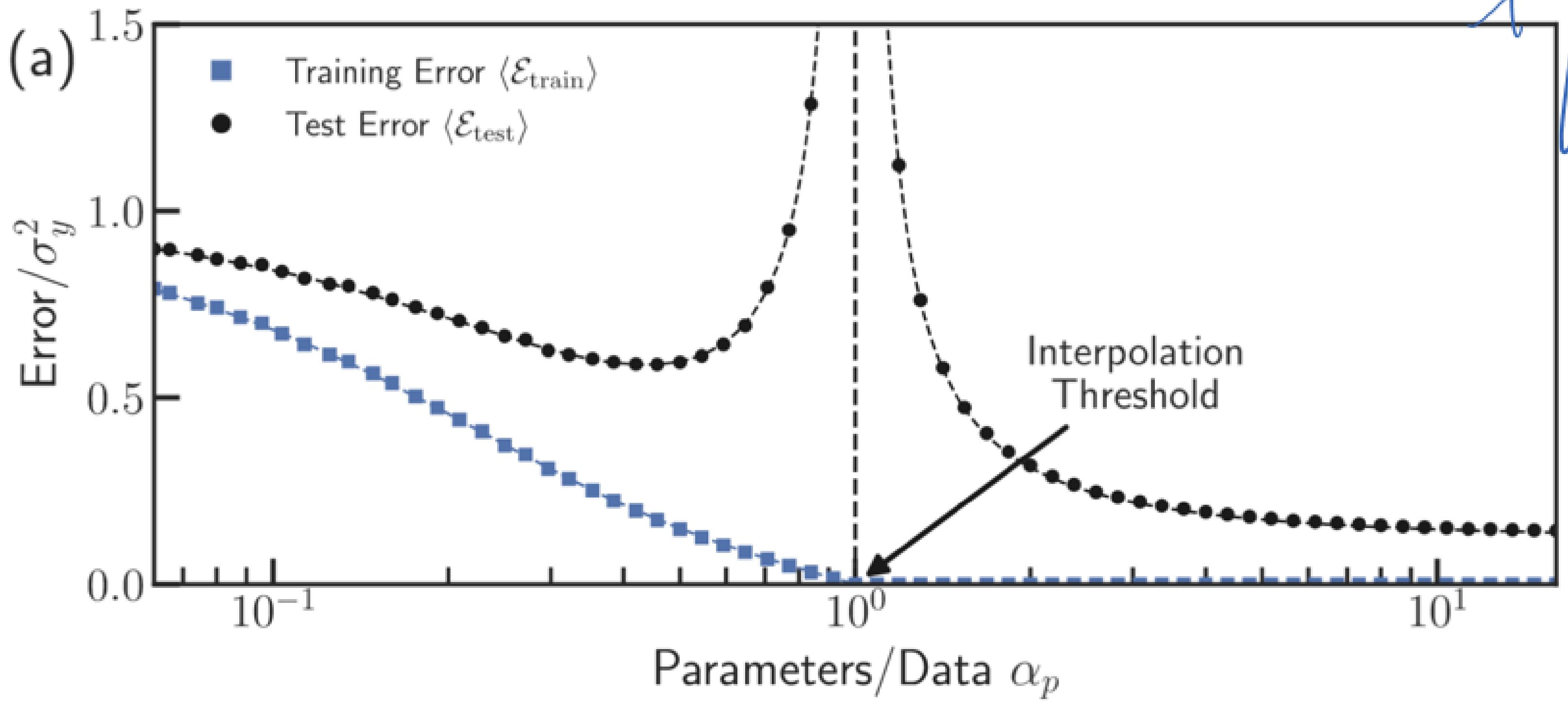


Source: [Plenoxels: Radiance Fields without Neural Networks (2022)]

$f(x) = 30\cancel{10}\, x^{100}$



Better Generalization?



Source: [Double Descent Demystified: Identifying, Interpreting & Ablating the Sources of a Deep Learning Puzzle]

Any Questions?

Perceptrons

"The Perceptron: A Probabilistic Model For Information Storage And Organization in the Brain"
by Rosenblatt (1958)

- Simple formula
- Simple updates

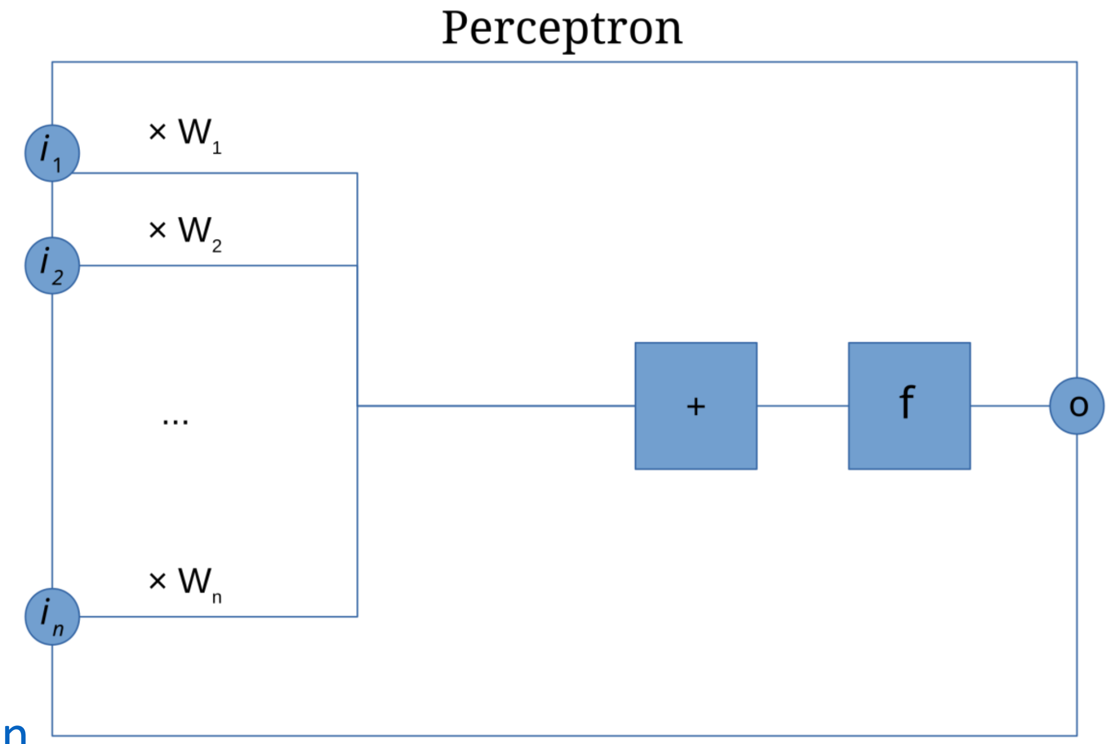


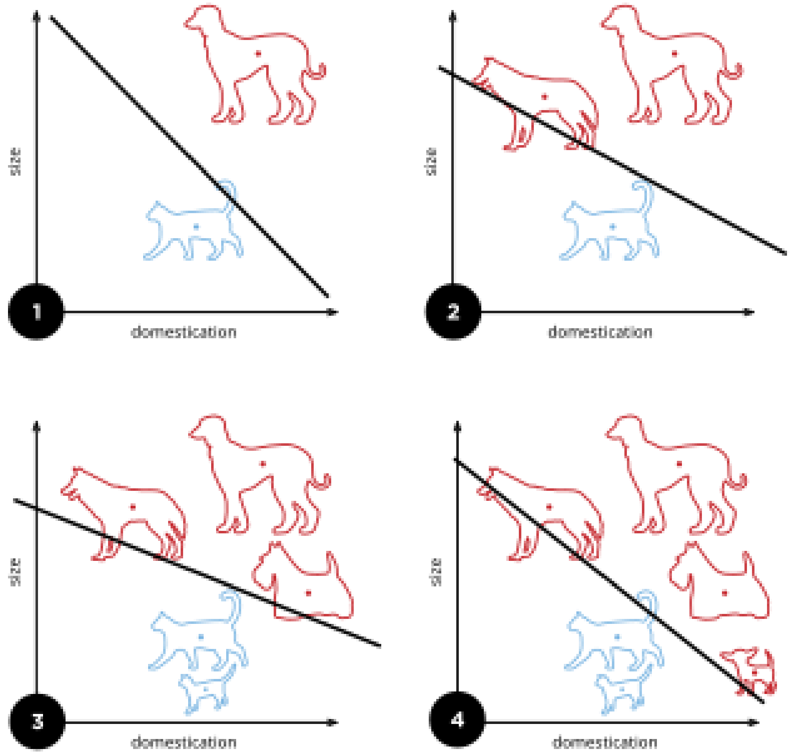
Image Source: https://en.wikipedia.org/wiki/Perceptron

$$o = f\left(\sum_{k=1}^{n} i_k \cdot W_k\right)$$

linear

Positive Perceptron Results

- Can solve any problem where the classes are linearly separable.



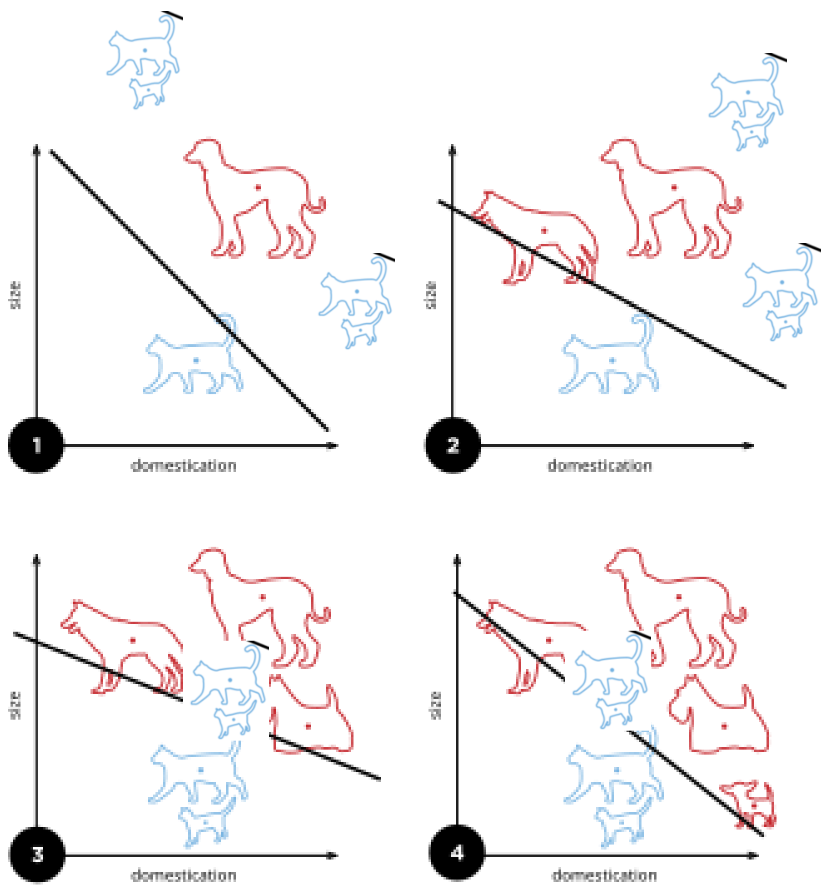
Sources: https://en.wikipedia.org/wiki/Perceptron

Negative Perceptron Results

- Can only solve problems where the classes are linearly separable.

parity

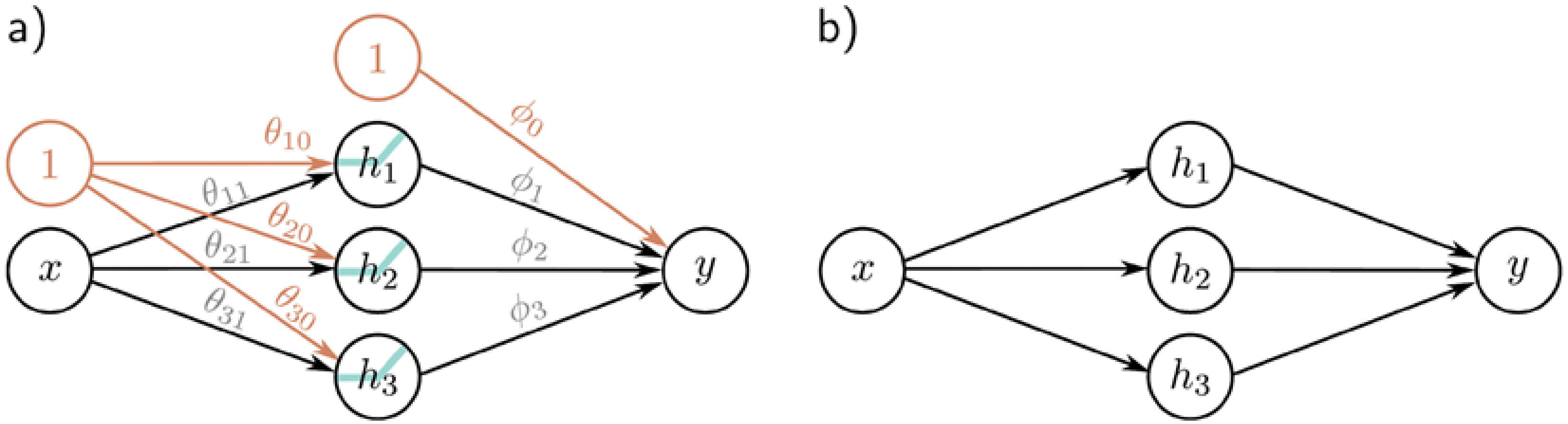
- This result led to an "AI winter".



Sources: https://en.wikipedia.org/wiki/Perceptron (modified)

Multilayer Perceptrons

- Limits of Perceptrons were misunderstood
- Did not apply to general neural network configurations



- One layer sufficient for universal approximation.

Image source: Understanding Deep Learning

Lingering Issue

- Negative result does not apply, but how exactly do we wire and train them?

Image source: https://www.reddit.com/r/aww/comments/236k8s/hurumph/

How to Train ~~Multilayer Perceptrons~~ Neural Networks

???

Backpropagation

- We can efficiently calculate gradients and update our models.
  - This is "gradient descent".

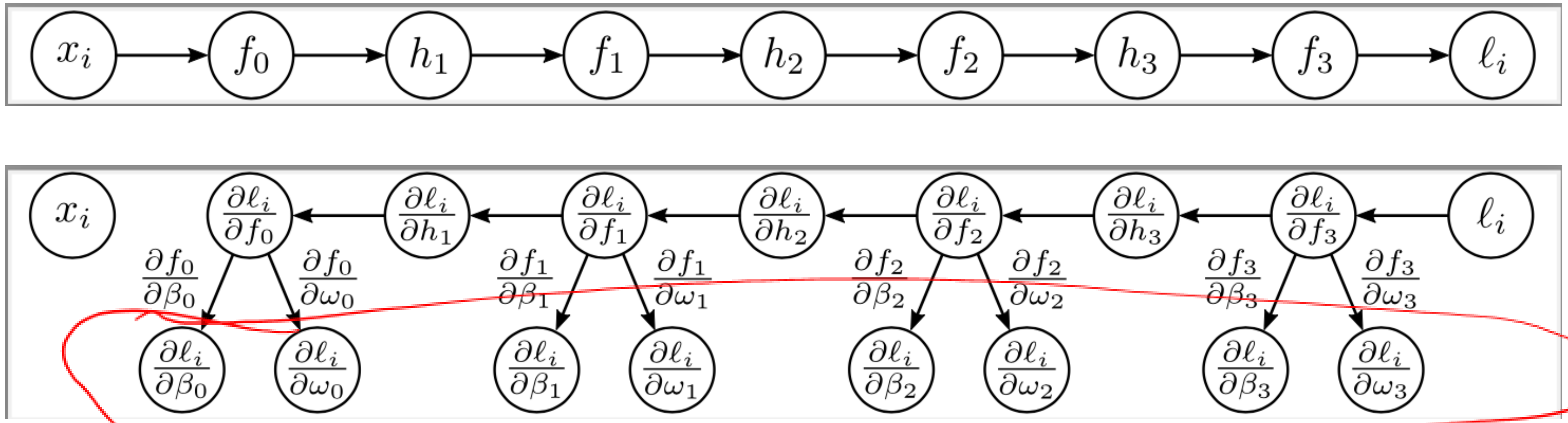


Image source: Understanding Deep Learning

partial derivatives of loss w/respect to parameter

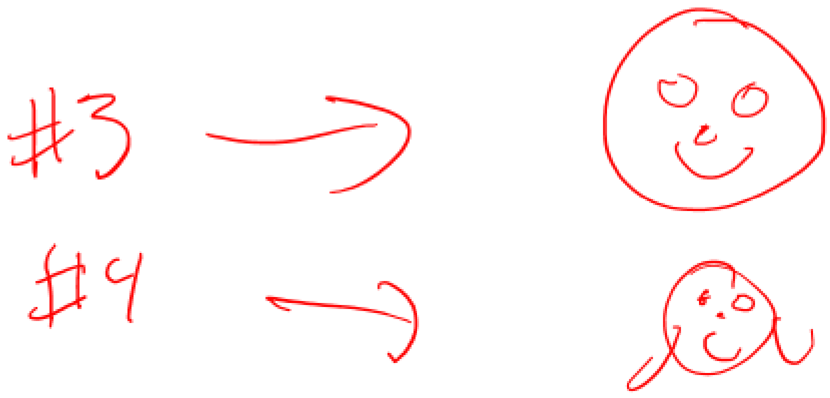
Difficulties Training 5-Layer Neural Networks

- But this did not work reliably for deeper networks…

Computer Vision as Inverse Computer Graphics (not a tangent)

Paraphrase of Geoffrey Hinton -

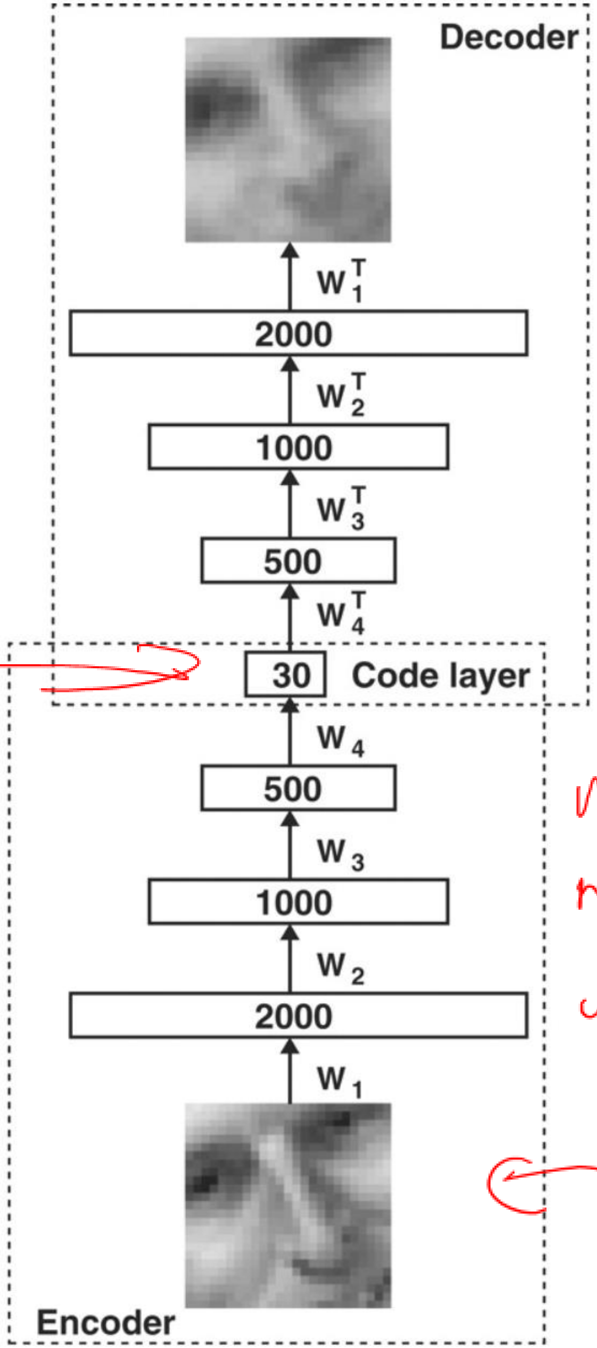
If we can briefly describe a short description of an object well enough to draw it, then we must have captured the essence of the object?



Auto-Encoder Idea

- If we can build an architecture like this, then the small “code layer” must have most of the important information?

Source: [Reducing the Dimensionality of Data with Neural Networks (2006)]()



Unrolling

reverse

small

narrower

narrow

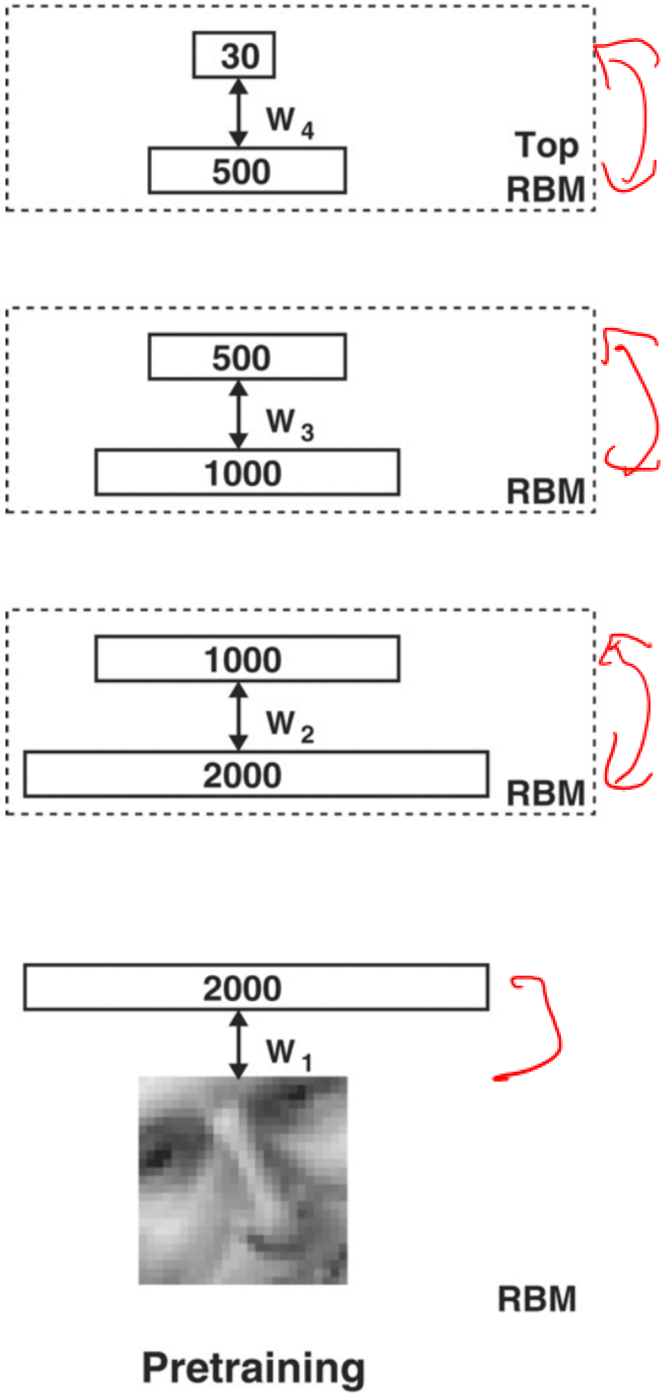
wide

input image

Auto-Encoders and Boltzmann Machines

- Instead of training the neural networks like in the last slide, they used restricted Boltzmann machines to build one layer at a time.
  - Mapping had to work both ways.
  - Then made neural networks imitating those mappings.

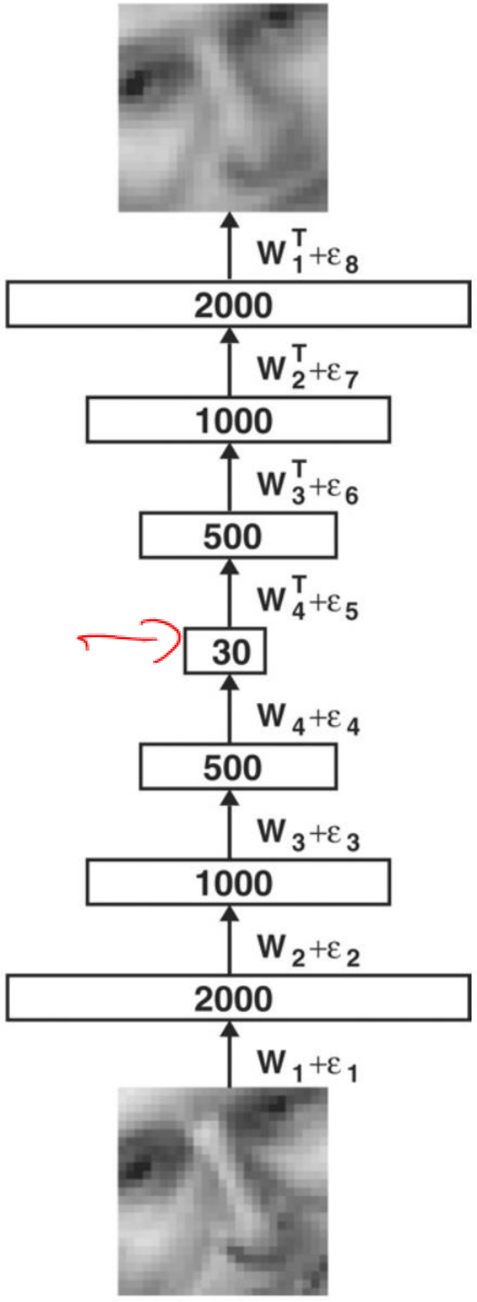
Source: [Reducing the Dimensionality of Data with Neural Networks (2006)]()



Fine-Tuning Pre-Trained Networks

- After the conversion from restricted Boltzmann machines to neural networks, gradient descent can be used to make the blurry decoding sharper.

Source: [Reducing the Dimensionality of Data with Neural Networks (2006)]



Fine-tuning

Auto-Encoder Latents for Classification

- Applied this technique to MNIST data set.

- Results (error rates):
  - 1.6% for randomly initialized neural networks (previous)
  - 1.4% for support vector machines (previous)
  - 1.2% using w/linear model based on code layer. 30

Source: [Reducing the Dimensionality of Data with Neural Networks (2006)]

Deep Learning from Scratch

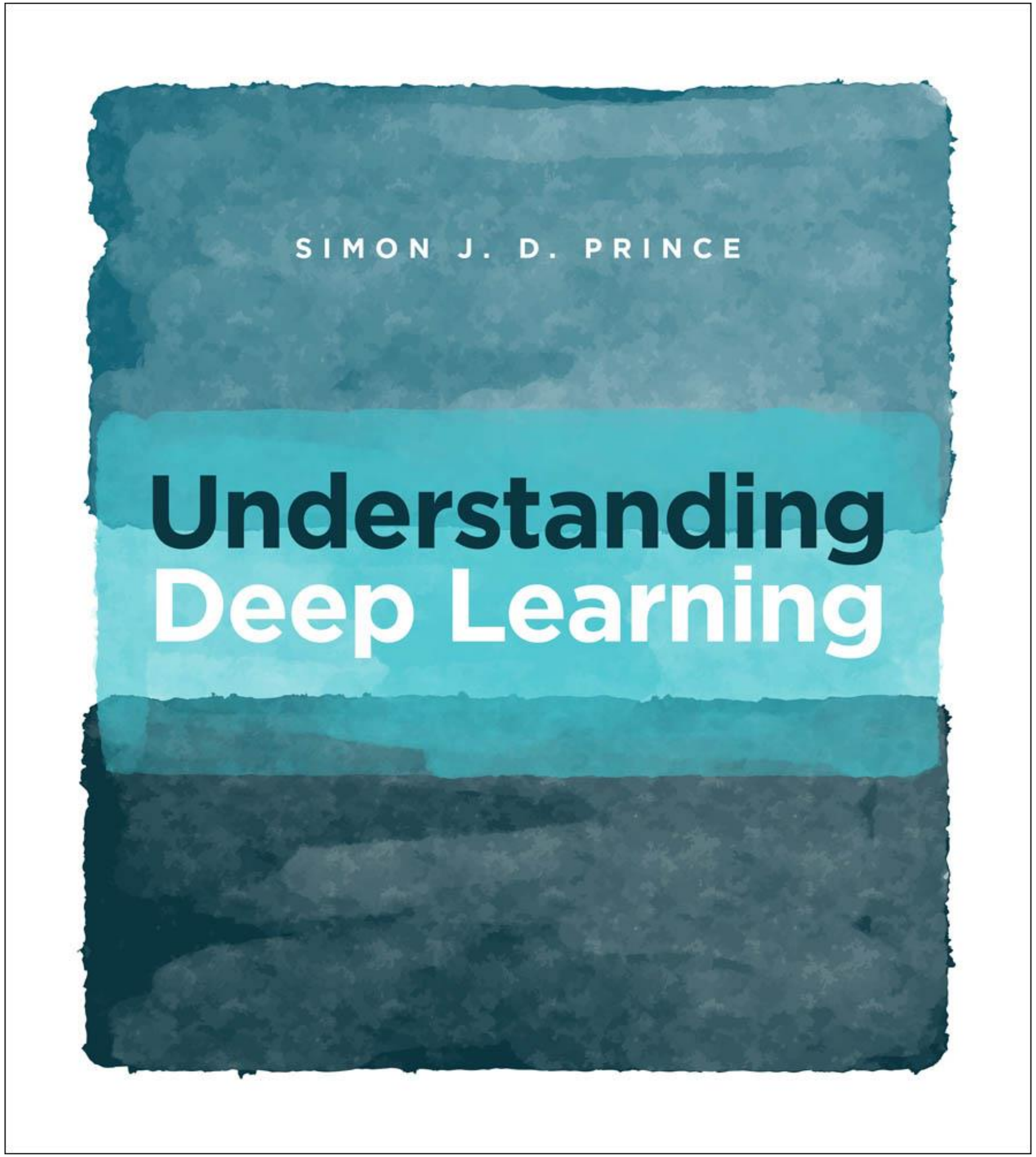
- After the previous work, we started figuring out what we were doing wrong and learned how to train deep neural networks directly.

- This will be the first 1/3 of this course…

Any Questions

Book

- Published December 2023
- http://udlbook.com
  - Free PDF there or buy at BU bookstore
- Used heavily for 1st half of the course, and a bit at the end too



Lecture Time

- Nominally 1h45m.

- Practically 1h15m?
  - Depends on the lecture. 🤷
  - I will be available if the lecture does not take the whole slot.

Course Web Site

- https://dl4ds.github.io/fa2025/
  - Piazza and Gradescope links here.
  - Syllabus here too (scroll to the bottom)

Generative AI Assistance (GAIA) Policy

https://dl4ds.github.io/fa2025/index.html#gaia-policy

1. Give credit to AI tools whenever used, even if only to generate ideas rather than usable text, illustrations or code.

…

3. When using AI tools on _coding_ assignments, unless prohibited
    1. Add the prompt text and tool used as comments before the generated code. Clarify whether the code was used as is, or modified somewhat, moderately or significantly.

…

5. Use AI tools wisely and intelligently, aiming to deepen understanding of subject matter and to support learning.

Focus on your learning objectives!

Please, no AI Slop or Verbosity!

Most assignments will be focused on implementing techniques covered in class, but you will sometimes be asked questions with text answers. For example, you may be asked to explain, motivate or otherwise argue for an approach. In those cases, you are expected to give a concise and direct answer and not be unnecessarily verbose. Points may be deducted for poorly written responses.

^^ in the formal syllabus.

You may use generative AI but use it well!

- Don’t waste the grader’s time.
- Don’t lose points for
  - Saying it is such an interesting question…
  - Answering a couple other questions!
  - Answering the next question too.
  - Making your answer pretty with redundant headings or emojis.
  - Repeating something an LLM hallucinated.
- Be concise and to the point.
- Submit something better than ChatGPT!

Grades

Item	Percentage
Discussions (due at end of discussion)	35%
Homework (~1 week each)	25%
Projects (~2 weeks each)	40%

- Late submissions up to 2 days late
- 1% penalty per hour

Feedback