Machine Learning and How Physicists Can Think About It

Jared Kaplan
Johns Hopkins University

hep-ai
high energy physics (hep) – artificial intelligence (ai)

OpenAI

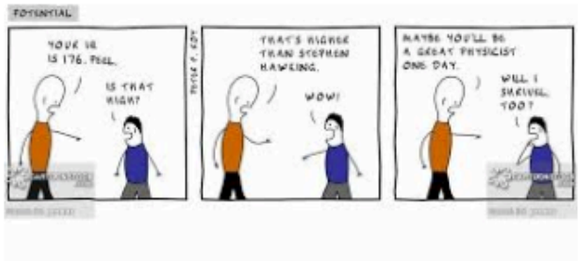
(Blueshift@Google — Institute led by former HEP physicists)

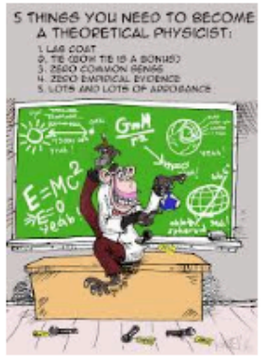
GOOGLE

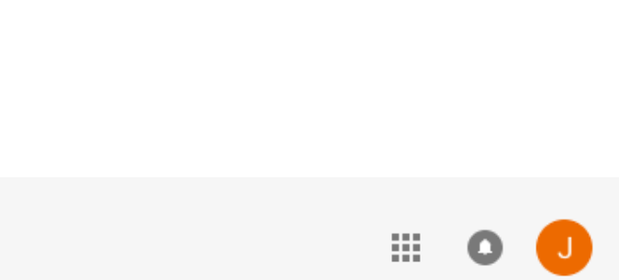
arrogant physicist cartoons

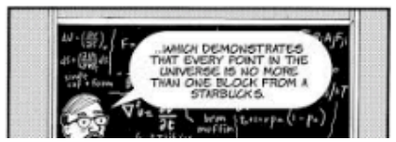
All | Images | Videos | News | Shopping | More | Settings | Tools | View saved | SafeSearch

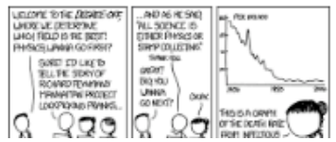
gravity | tiphaine riviere | xkcd | philosophy | thesis | empirical | cycle | purity | funny | smbc | graphic novel | biology | applied | meme | math









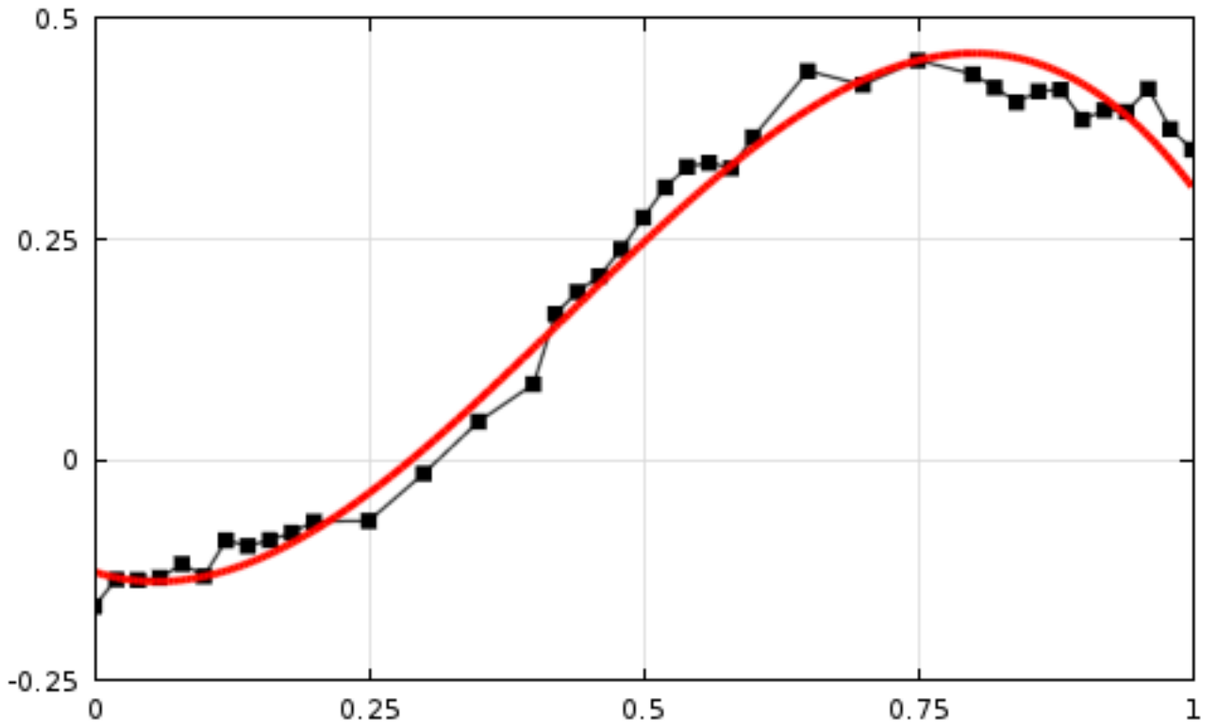


Outline

- What is (Contemporary) Machine Learning?
- Should physicists work on it?
- Projections, Comparisons, & Outlook

What is Contemporary ML?

- Just **curve fitting** with a general function approximation



What is Contemporary ML?

- Just **curve fitting** with a general function approximation

- Models learn “intuition” ~ correlations, rather than logic

- Much like a condensed matter system — hopeless to understand all of the constituents — instead we need a statistical understanding of emergent behavior

Function Approximation

We need a very general and versatile way to express extremely complicated functions. Then we can do “curve fitting” in all of their parameters to “learn” about the data distributions.

Neural networks build complicated functions from high-dimensional matrix multiplication + simple non-linearities.

Neural Network “Layers”

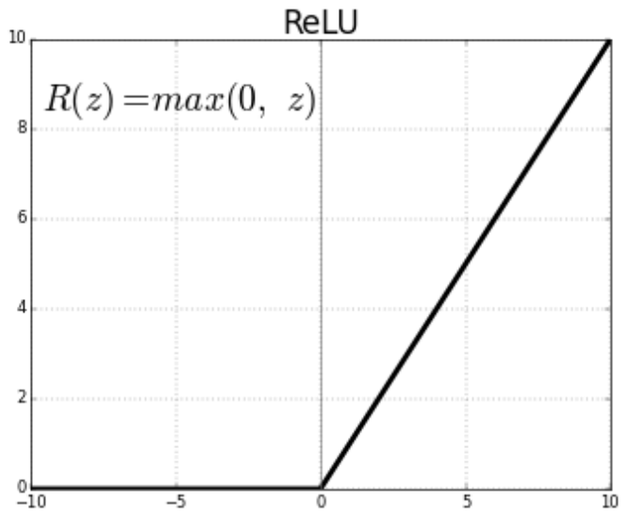
Start with a vector (a data point), usually in a high dimensional space:

$$\vec{x}$$

Apply a linear (affine) transformation to it

$$\mathbf{w} \cdot \vec{x} + \vec{b}$$

Using a (usually very large) matrix and an affine shift. Now apply:



component-wise. NN parameters are weights and biases: $\mathbf{w}, \vec{b}$

A Very Explicit Example

Toy Neural Network parameters:

$$w = \begin{pmatrix} 4 & 5 & 6 \\ 3 & 2 & 1 \\ -9 & -8 & -7 \end{pmatrix} \qquad b = \begin{pmatrix} 12 \\ 3 \\ 4 \end{pmatrix}$$

A single item of data:

$$x = \begin{pmatrix} 3 \\ -2 \\ 1 \end{pmatrix}$$

The neural network layer will process the data as:

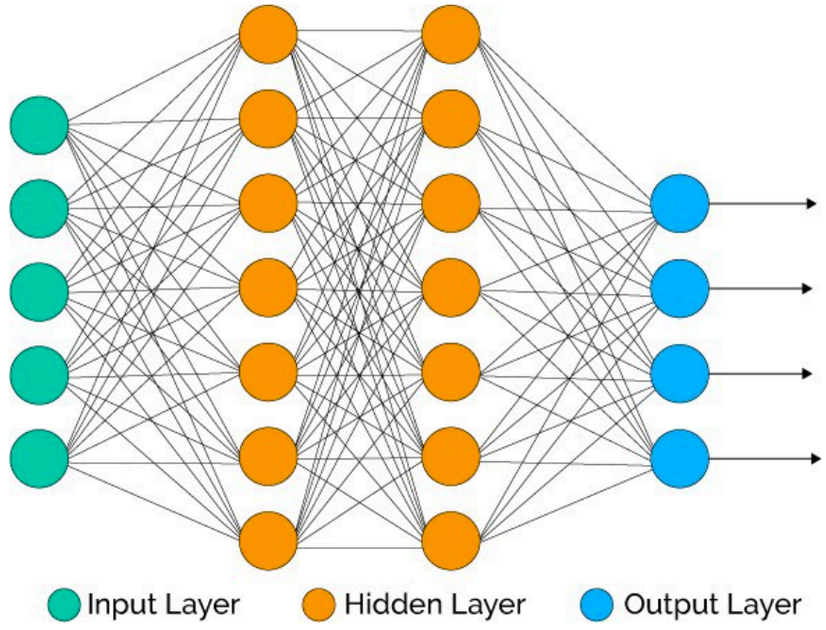
$$\mathrm{ReLU}\left[w \cdot x + b = \begin{pmatrix} 20 \\ 9 \\ -14 \end{pmatrix} \right] = \begin{pmatrix} 20 \\ 9 \\ 0 \end{pmatrix}$$

A full neural network is simply a **composition of many of these layers**, with different parameters for each layer.

Neural Network

$$f(\vec{x}; \mathbf{w}, \vec{b}) = \max(0, \mathbf{w} \cdot \vec{x} + \vec{b})$$

A ‘deep’ neural network is just a composition of these building blocks.



Pictures like this are a silly way of writing some matrix multiplications.

Hello World - MNIST

Classifier NN is just a function from images to probabilities:

$$f(\text{3};\Theta) = (p_0, p_1, \cdots, p_9)$$

It depends on the neural network parameters:

$$\Theta = \{\mathbf{w}_i, \vec{b}_i\}$$

Learning means determining the NN parameters w & b. How?

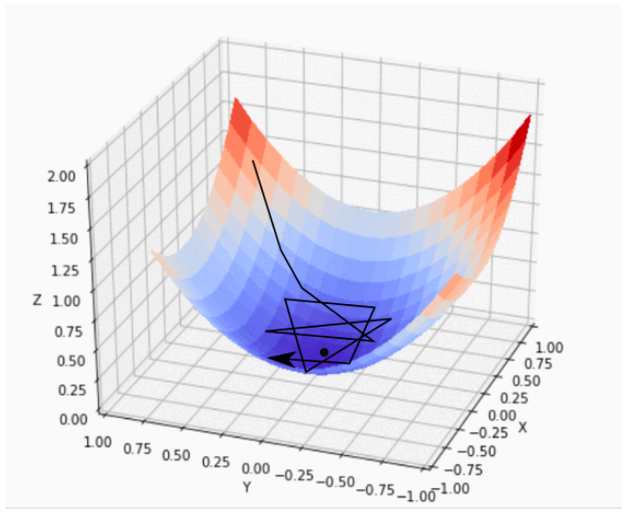
How does it learn?

Our loss function is a negative log-likelihood; for this particular image:

$$L(\Theta) = -\log p_3(\text{3}\,; \Theta)$$

Optimize — minimize the loss — via gradient descent with mini-batches:

$$\Theta_{n+1} = \Theta_n - \epsilon \sum_i \nabla_\Theta L(\vec{x}_i, y_i; \Theta)$$



It's called "stochastic" gradient descent because we use mini-batches of data rather than the full dataset.

MNIST by the Numbers

The MNIST dataset has 50,000 images:

Images are 28x28 pixels, so 784-dimensional vectors. Each layer would have

$$615,440$$

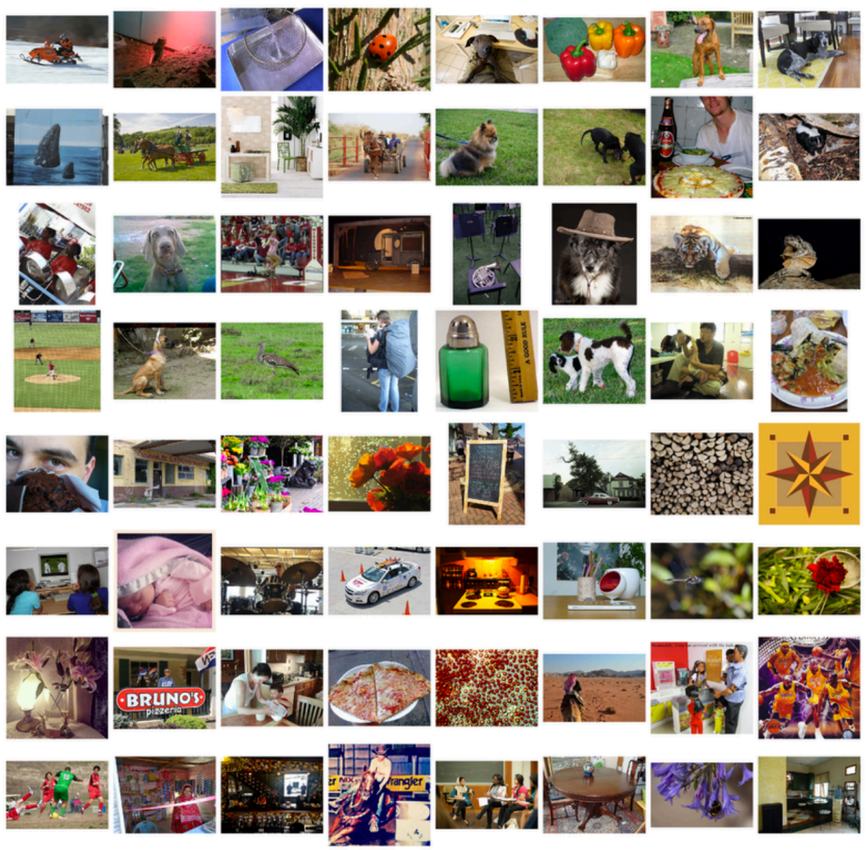
parameters. Might pass over the full dataset ~20 times in training. Note that:

$$N_{param} \gg N_{data}$$

which might surprise you! But this is typical of supervised learning.
What about more realistic data…

ImageNet

A more challenging image recognition dataset… major progress by "AlexNet" in 2012 was a key event in contemporary ML.



10^6 images, 10^3 classes

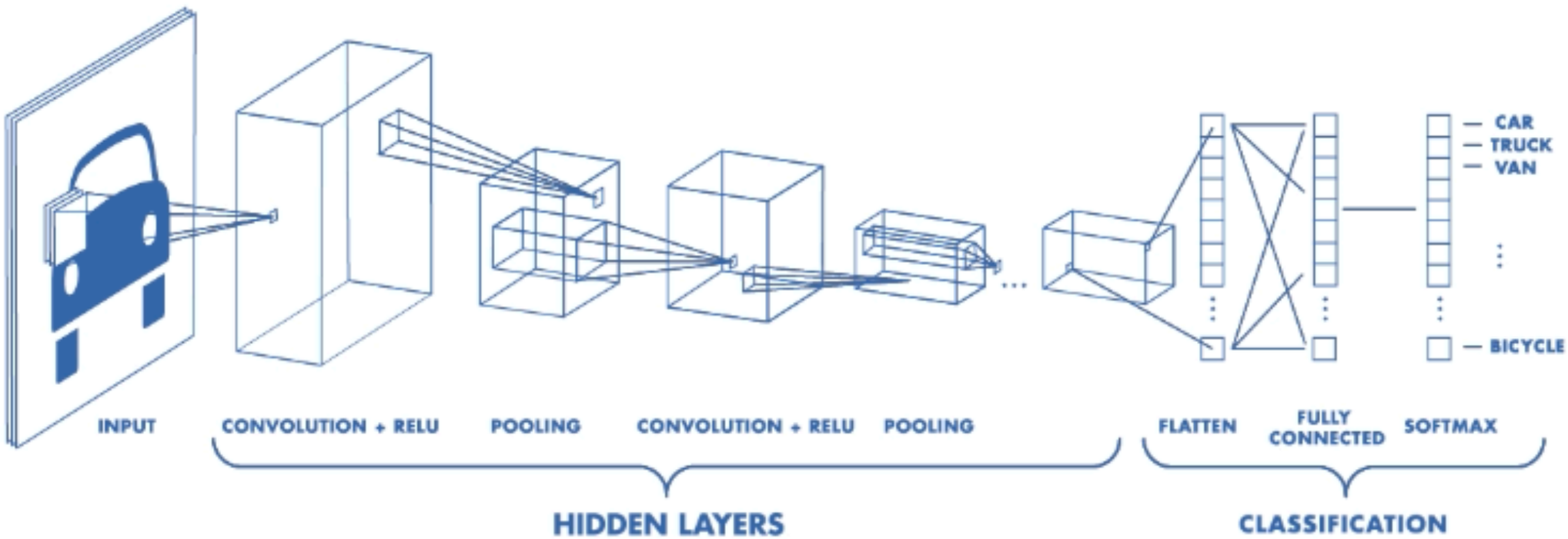
$256 \times 256 \approx 6 \times 10^5$ pixels each

Shouldn't use huge matrices!
How can we do better?

Much like physical systems, improved Neural Network architectures are organized by **symmetry**…

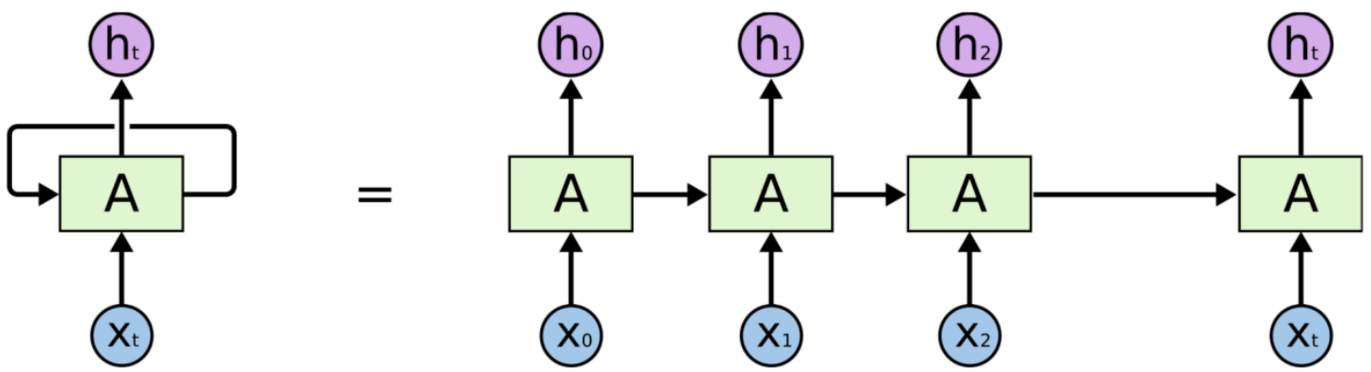
Architecture and Symmetry

(Spatial) Translational Symmetry - Convolutional NNs



Architecture and Symmetry

For sequential data, such as language — Recurrent Neural Networks:



A popular variant is the LSTM = Long Short-Term Memory…
it's designed to store information for more "steps" of the sequence.

Roughly speaking, LSTMs may help to solve the problem that many sequential matrix multiplications may be dominated by large eigenvalues.

What Else & How?
Three Levels of Abstraction

1. List different interesting domains (images, language, games, robotics) and ML accomplishments

2. Describe how to convert interesting & important problems into complicated function approximation tasks.

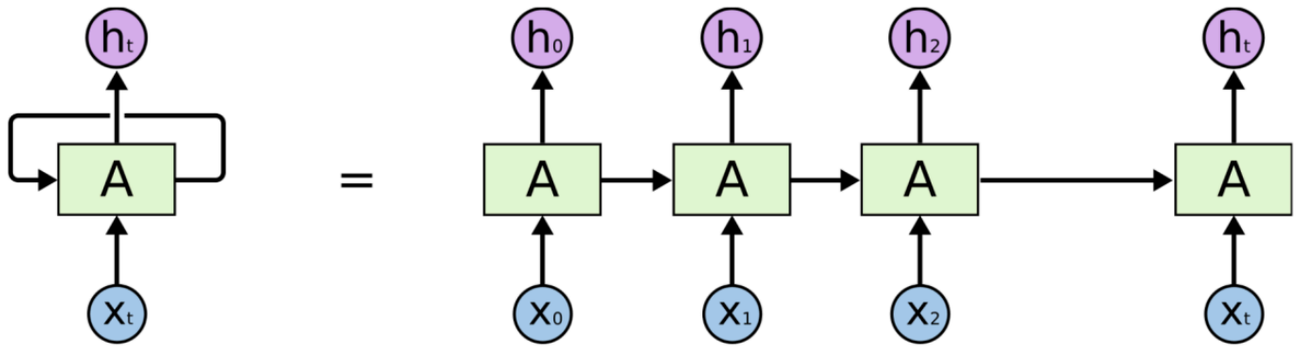
3. Most abstractly, classify tasks based on the type of connection between “success” and the model parameters

Language Models

Learn language by “autoregressive” prediction:

$$P(w_{n+1}|w_n, \cdots w_1)$$

Optimize the log-likelihood of real world text, which provides the “right answer”.
This is “supervised learning” just like image classification.



Model takes (prior context) & (current word) and produces (new context) & (prediction).

Language Model Sample

With its world-renowned faculty and state-of-the-art facilities, the Henry A. Rowland Department of Physics and Astronomy combines the best aspects of a top research university with the more intimate learning environment typical of small liberal arts colleges. Through small classes and numerous opportunities to participate in research, students gain a solid foundation in the physics and astronomy areas of their choosing.

The department has a number of specialty laboratories and the center is the university's main laboratory, which is in the center of the university. There, each lab is run by a student, which serves its purpose. All the departments of science and technology are divided into individual divisions. Each department is divided into four sections. The main laboratory is the largest and has the most advanced technology available. The lab has a single research station, which is located in front of the university. That is a major part of the facility. There, students can conduct their research by the main lab.

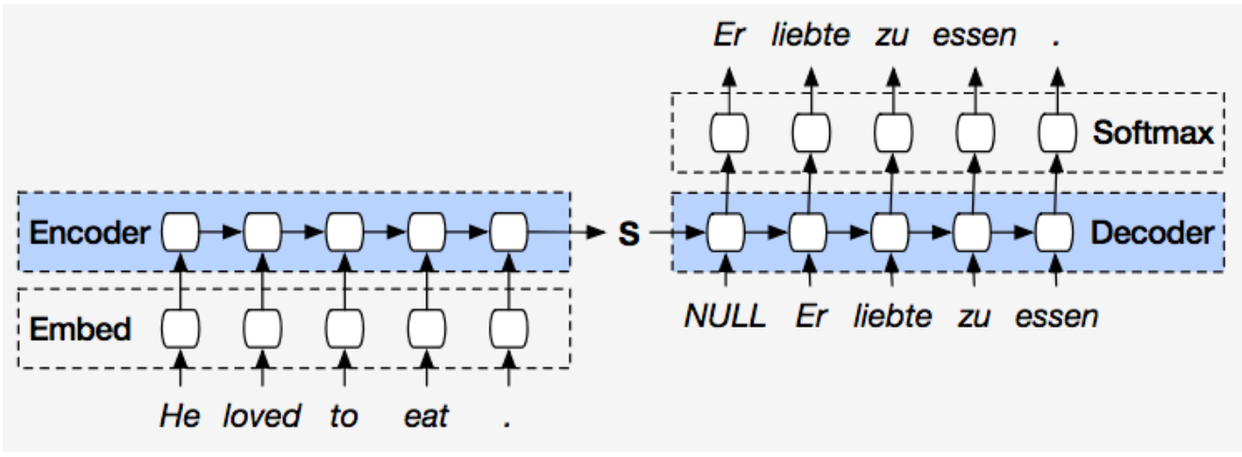
In the center of the university, a large central office is used as an administrative office. The room consists of a large desk and a large, black leather chair. In front of the desk, a large computer system and a large screen are on the wall.

The front of the lab is occupied by a man. He has medium brown hair and dark glasses that frames a face that is very serious. He seems to have a great deal of energy. In fact, I can see an aura radiating all around him. I am sure that many of the professors of the department have a strong energy that is very strong here in the school. It is hard to understand how such a large area of this world-wide science hasn't evolved by now with the technology and resources used by the military…

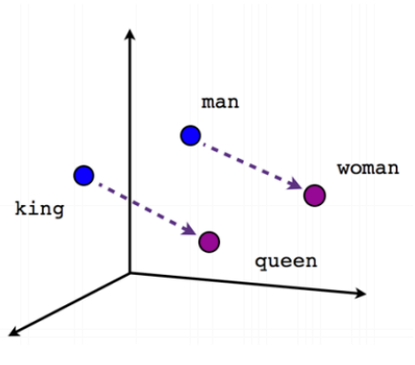
[LM from Radford et al]

Language Models

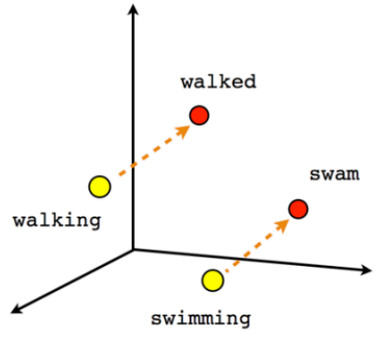
A more complicated task is translation, which “encodes” one language and then “decodes” into another.



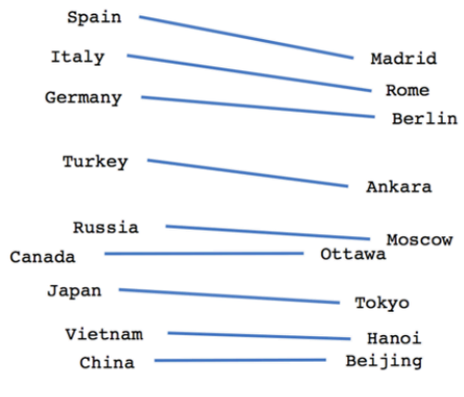
The learned “Embedding Space” places all words into a single high-dimensional (eg ~500) vector space — it’s often interesting on its own:



Male-Female

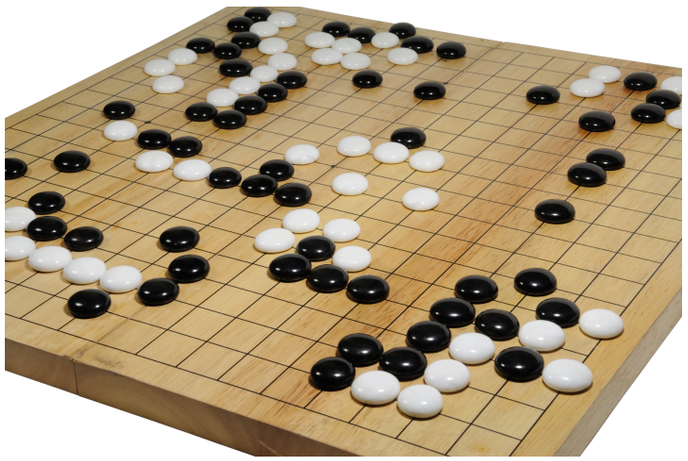


Verb tense

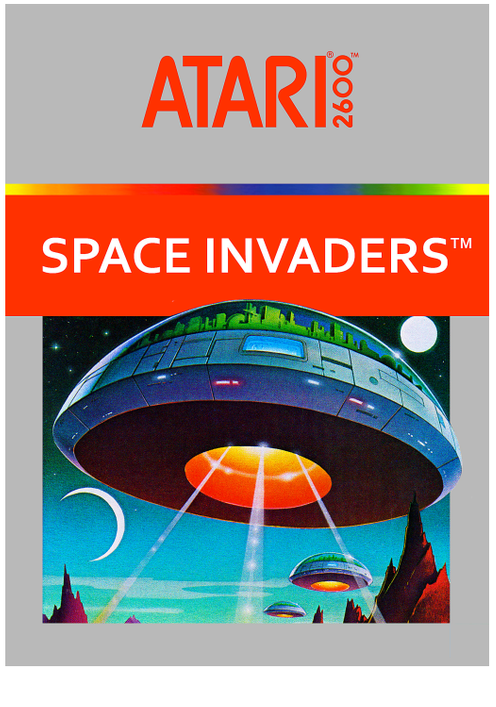


Country-Capital

Games



Formalize games in terms of a state
and a set of possible actions.
We want to choose actions to win.
Abstractly, this is what life does.



One approach is to simply learn a policy $\pi(s)$, a map from states —> actions.

Alternatively, we can try to learn the total value $Q(s, a)$ of actions in any given state. Then we can simply choose the highest value action.

This is Reinforcement Learning. Much harder to gauge success!

Self-play has proven to be a powerful way to train in games like Go.

Understand the World

Perhaps the hardest problem is simply to understand the world.
This is “unsupervised learning”.

Understand the World

Perhaps the hardest problem is simply to understand the world.
This is “unsupervised learning”.

Formalize as determining the underlying probability distribution of the world.

Understand the World

Perhaps the hardest problem is simply to understand the world.
This is “unsupervised learning”.

Formalize as determining the underlying probability distribution of the world.

Clearly a deep and important problem.

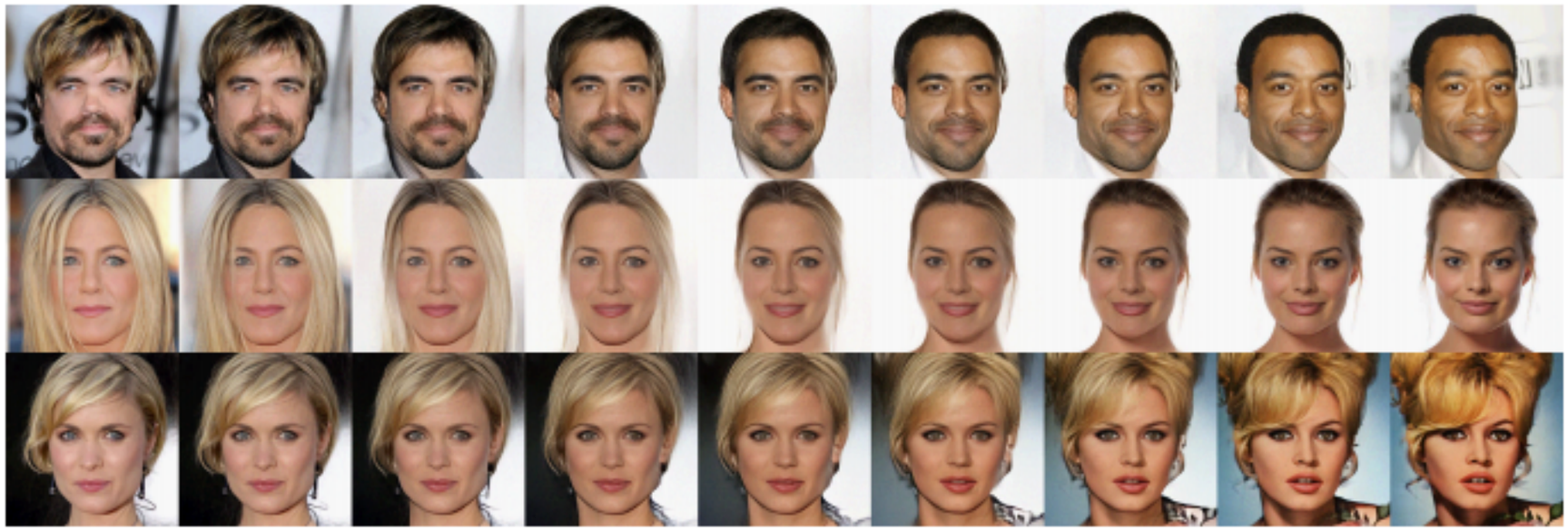
Understand the World

Perhaps the hardest problem is simply to understand the world.
This is “unsupervised learning”.

Formalize as determining the underlying probability distribution of the world.

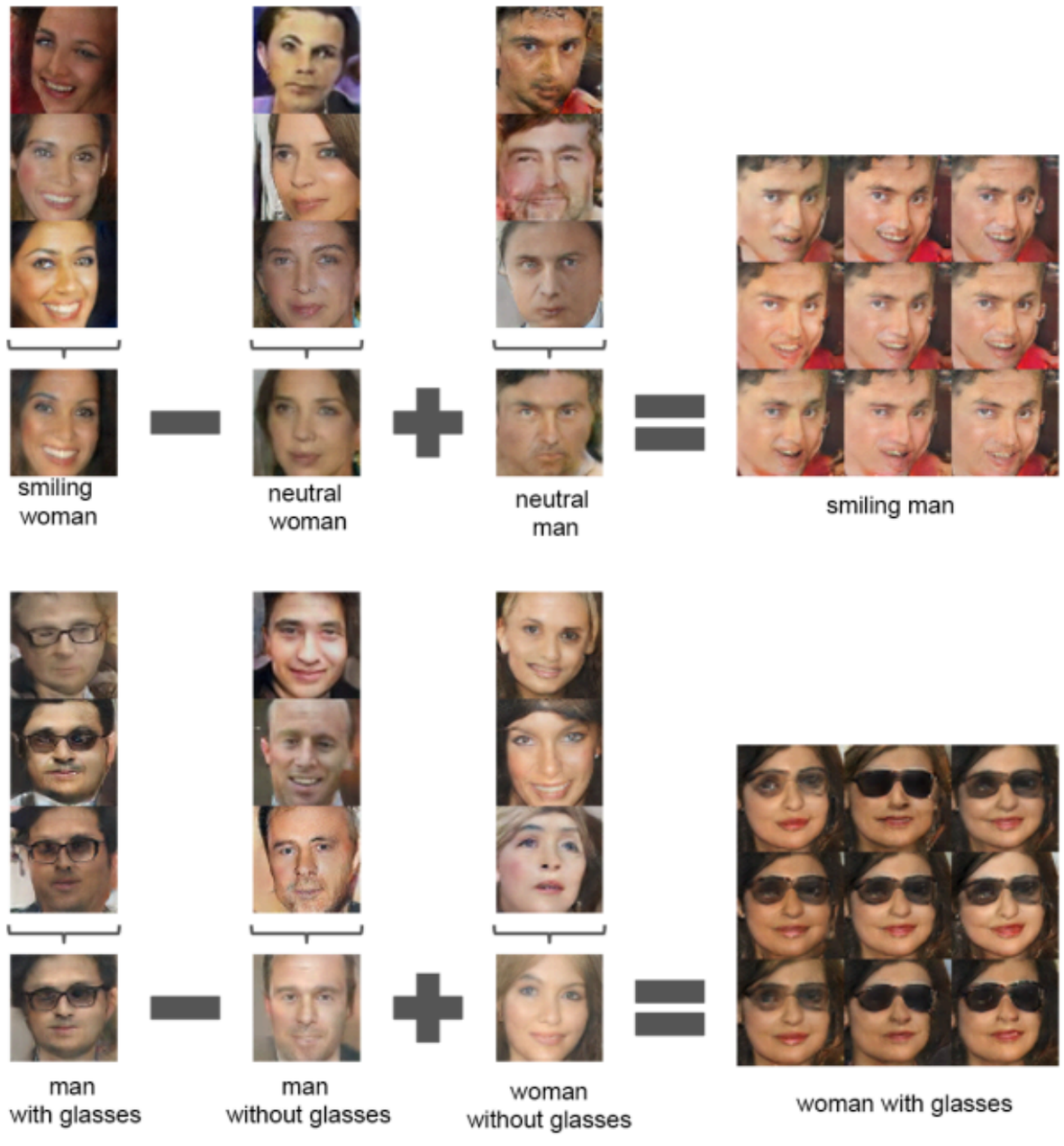
Clearly a deep and important problem.

ML researchers have solved it so they can… interpolate between celebrities!?



[Glow, Kingma & Dhariwal]

Understand the World



[DCGAN, Radford et al]

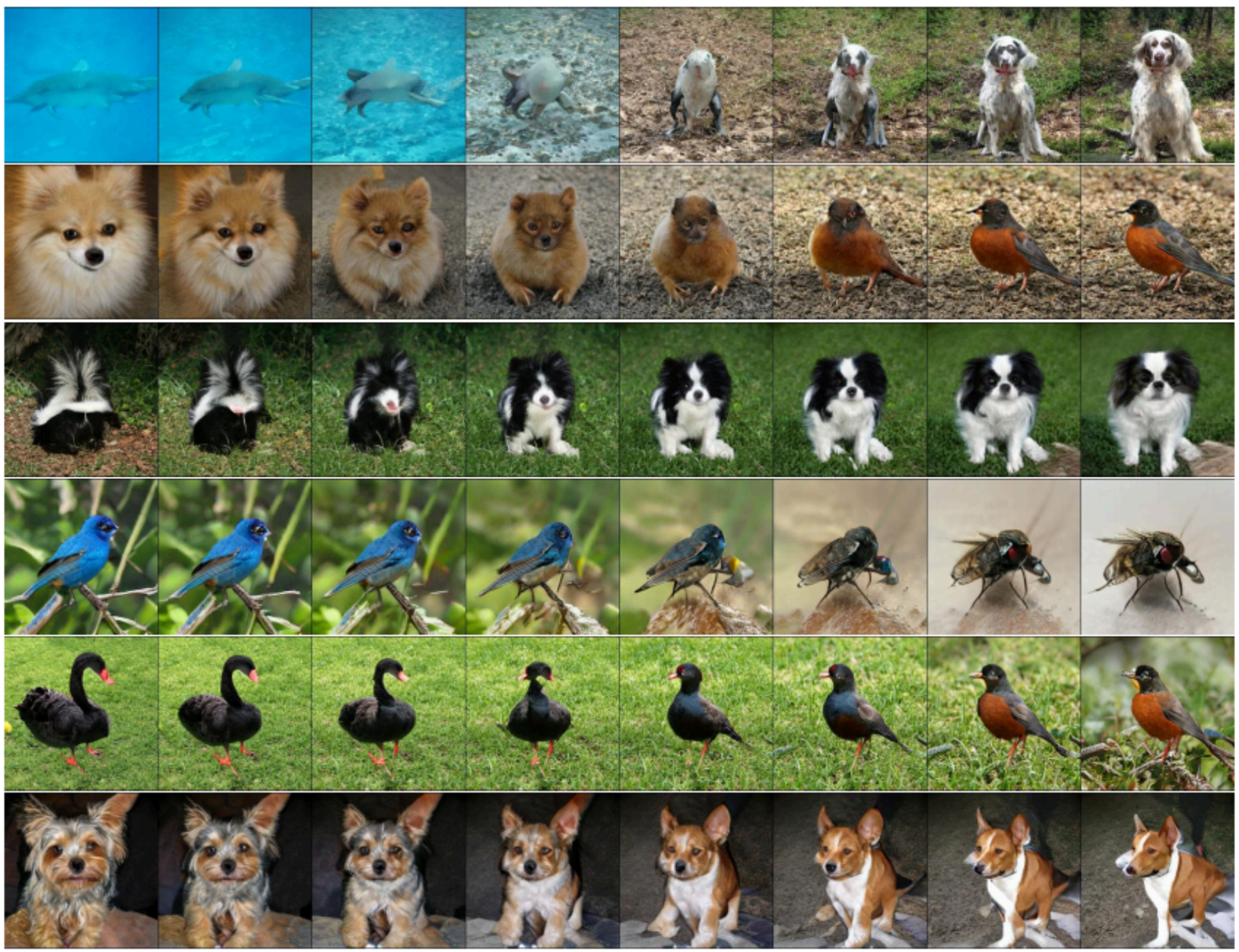
One approach just parameterizes a map from a high-dimensional vector space to the “space” of images or other data.

“Latent” vector space can then acquire interesting meaning.

Loss function either tries to model probabilities (VAE/Glow) or plays a real/fake discrimination game.

Image Generation with More Compute, More Data

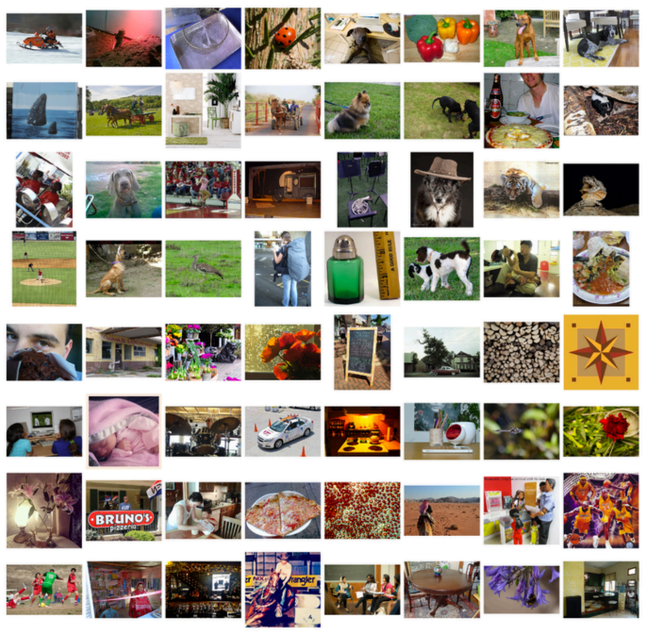
All of ImageNet and more…



[https://arxiv.org/abs/1809.11096]

Relevant Scales in Machine Learning

ML by the Numbers - ImageNet



10^6 images, 10^3 classes

$256 \times 256 \approx 6 \times 10^5$ pixels each

First models to “do well” had 6×10^7 parameters.

Models for this sort of data typically have 10s of layers.

In total, training takes roughly 10^{19} flop, or around a petaflop-hour.

ML Scales - Language

A common language dataset is the Billion Word Benchmark.

A billion words would take a typical person about 10 years of continuous reading. Language models usually “read” this dataset 10-100 times when training; this occurs in a few days to a month.

Some heavy duty state-of-the-art translation systems $\gtrsim 10^{21}$ flop to train.

Model	Test Perplexity	Number of Params [billions]
Sigmoid-RNN-2048 (Ji et al., 2015a)	68.3	4.1
Interpolated KN 5-gram, 1.1B n-grams (Chelba et al., 2013)	67.6	1.76
Sparse Non-Negative Matrix LM (Shazeer et al., 2015)	52.9	33
RNN-1024 + MaxEnt 9-gram features (Chelba et al., 2013)	51.3	20
LSTM-512-512	54.1	0.82
LSTM-1024-512	48.2	0.82
LSTM-2048-512	43.7	0.83
LSTM-8192-2048 (No Dropout)	37.9	3.3
LSTM-8192-2048 (50% Dropout)	32.2	3.3
2-Layer LSTM-8192-1024 (BIG LSTM)	30.6	1.8
BIG LSTM+CNN Inputs	**30.0**	**1.04**
BIG LSTM+CNN Inputs + CNN Softmax	39.8	**0.29**
BIG LSTM+CNN Inputs + CNN Softmax + 128-dim correction	35.8	**0.39**
BIG LSTM+CNN Inputs + Char LSTM predictions	47.9	**0.23**

[Exploring the Limits of Language Modeling, 2016]

ML by the Numbers - Atari

Atari is “easy” and uses simple CNNs to interpret the screen pixels plus a relatively small NN to choose policy or action-values, eg ~1M parameters.

Look at millions of “frames" (ie views of the screen) to learn to do well at Atari. This is ~days to weeks of straight play, ie $\gtrsim 10^5$ seconds.

Note that knowledge of one game **does not transfer** to other games, even if they seem similar to a human.

Learning from one level of Sonic the Hedgehog so that the AI can beat other levels is a state of the art research problem.

ML by the Numbers - Dota

Online multi-player strategy game.

Training:

256 GPUs and 10^5 CPUs.

$\sim 10^{10}$ seconds of game play per day

Trains with mini-batches of $\sim 10^6$ observations, each is 20k dimensional vector

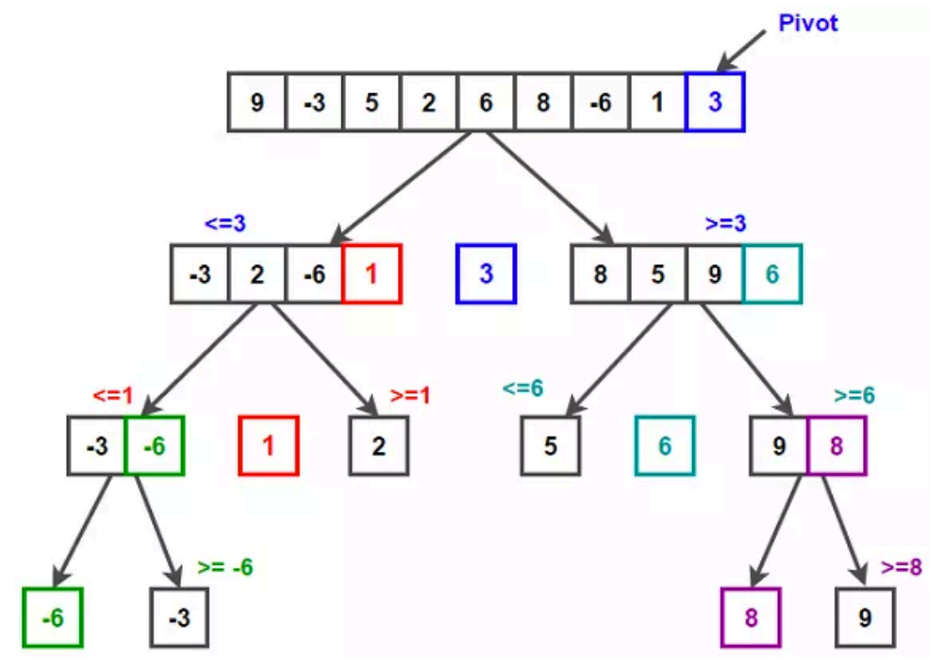
A game is about 30 minutes, which means there are ~20k moves per game.
Thus the AI needs to think at a large range of timescales.

The Neural Network itself for Dota is not very large,
just a 1024 or 2048 LSTM, like a medium-sized language model.

Does it have anything to do with Physics?

Are NNs Algorithms?

Sorting is a prototypical example of a classic algorithmic task. Quicksort:



You can easily discuss worst case, best case, average case performance…
and you can prove theorems about it.

Ingredients are simple, visible, and amenable to detailed analysis.

…or Snowflakes?

“If you look at the fractal structure of a snowflake, you might think that whoever made it did something impossibly intricate and difficult, but that building it piece by piece must somehow be possible, since someone did it. In fact, both statements are false: the way to make a snowflake is not to think in terms of its pieces but to know the laws of physics, have enough raw material (data) and a large enough chamber (parameters), set the temperature, pressure, and humidity correctly (training target), and wait for long enough (compute). Furthermore, this is your only way to make snowflakes; trying to piece together a single one from little bits of ice is hopeless.” —Dario Amodei, OpenAI

ML and Experimental Physics / Phenomenology

- ML involves **experiments** on systems that are under good control, but too complicated to reason about except via statistics, symmetries, & toy/sub-problems

- No one knows (quantitatively & systematically) **why** it works, or at what scale it should have / will work

- Many observed “fixed points” and “scaling rules” but engineering approach doesn’t tend to focus on them

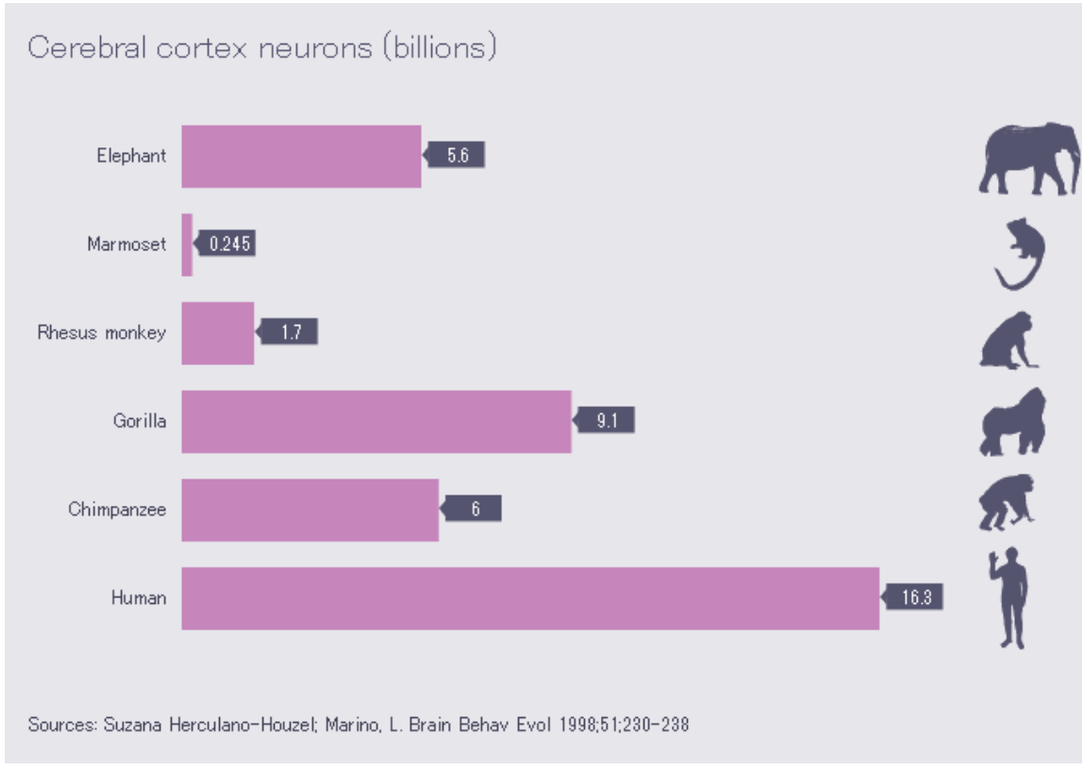
Physics, Astrophysics, Biophysics… Machine Learning Physics?

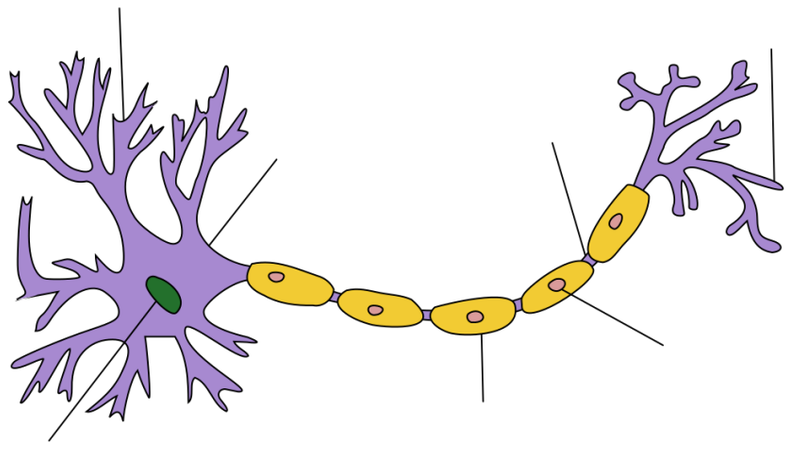
Why Now?

Why Now?

Let's **compare NN to the brain** at the order of magnitude level.

$\lesssim 10^{11}$ Neurons





$\sim 10^4$ connections per neuron. Synapses spike about once per second.

Roughly speaking, human cortex runs at about a **petaflop** $\sim 10^{15}$ flops.

Why Now?

But we really care about the amount of computation needed to **train a human brain.** This takes ~30 years, or

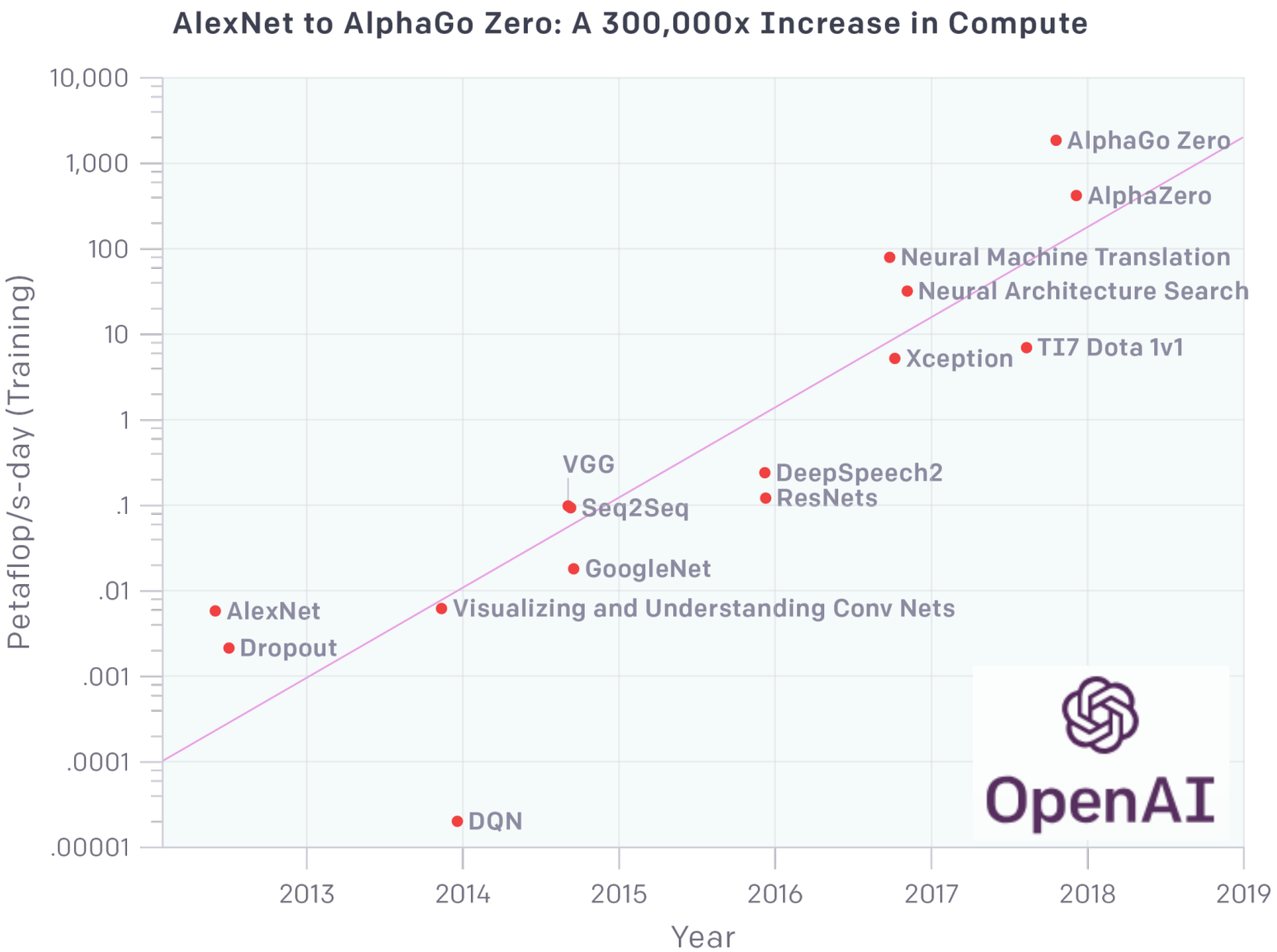
$$\sim 10^9 \quad \text{seconds}$$

So you can summarize human life as $\lesssim 10^{24}$ flop computations.

That level of computation per day will be accessible relatively soon for very large organizations.

Need it per day to experiment with models, as researchers run hundreds or thousands of experiments before creating a working model.

Compute in ML



2020

2022

~1000 human lifetimes of compute

Safety & Ethics

Even if ML didn't improve much from current state of the art, it would continue to permeate society. And most likely it'll improve a lot more…

- Will ML systems be deployed ethically and be used to improve human welfare?

- At a technical level, how do we "teach" complex human goals to obtain AI **aligned** with human values, and that respond to new situations in a **robust**, cautious way?

Conclusions

- ML looks to be a important, largely experimental new science where the methods of estimation, scaling, statistical physics, and “phenomenology”… ie the careful and quantitative but frequently non-rigorous thinking that physicists employ… may be applied to great effect

- It may change the world a lot, and quickly, and so it’ll be very important to get that right. We need to build safe systems and campaign for safe policies.

Language Model Sample

With its world-renowned faculty and state-of-the-art facilities, the Henry A. Rowland Department of Physics and Astronomy combines the best aspects of a top research university with the more intimate learning environment typical of small liberal arts colleges. Through small classes and numerous opportunities to participate in research, students gain a solid foundation in the physics and astronomy areas of their choosing.

The department has a number of specialty laboratories and the center is the university's main laboratory, which is in the center of the university. There, each lab is run by a student, which serves its purpose. All the departments of science and technology are divided into individual divisions. Each department is divided into four sections. The main laboratory is the largest and has the most advanced technology available. The lab has a single research station, which is located in front of the university. That is a major part of the facility. There, students can conduct their research by the main lab.

In the center of the university, a large central office is used as an administrative office. The room consists of a large desk and a large, black leather chair. In front of the desk, a large computer system and a large screen are on the wall.

The front of the lab is occupied by a man. He has medium brown hair and dark glasses that frames a face that is very serious. He seems to have a great deal of energy. In fact, I can see an aura radiating all around him. I am sure that many of the professors of the department have a strong energy that is very strong here in the school. It is hard to understand how such a large area of this world-wide science hasn't evolved by now with the technology and resources used by the military…

[LM from Radford et al]

Extra Slides

Some Open Problems

- Meta-Learning and Transfer Learning
- Unsupervised + Reinforcement = Model-based RL, ie agents that plan for the future?
- Why does all of this work, and can understanding it more quantitatively lead to further improvements?