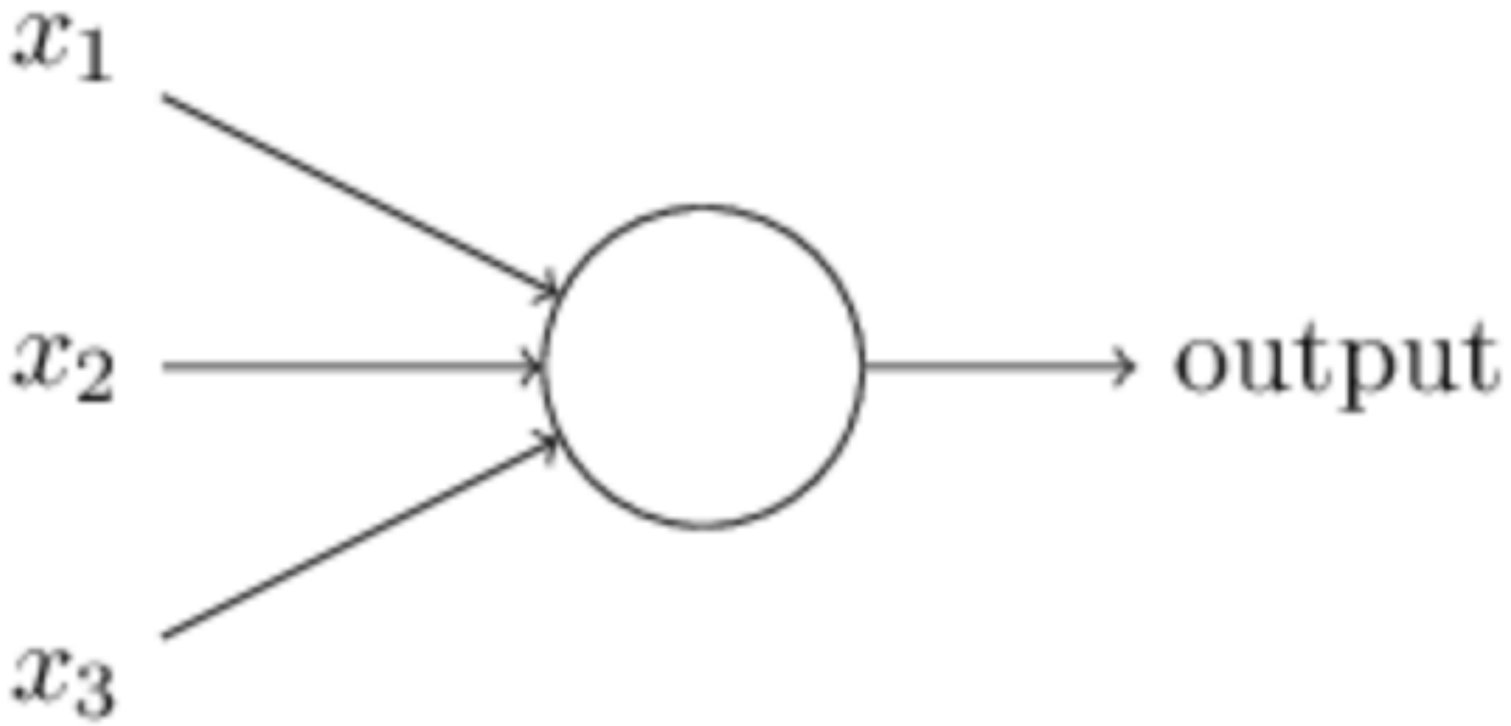
Why use neural networks?

- You've already seen the main limitation of linear approximation: need to hardcode the features.
- This may not even be possible if there are many of them.
- Also: in some sense, this is the crux of the problem.
  — you want an AI to learn to play a game
  — so first you've got figure out what is important to be successful in the game
  — that is, you've got to figure out how to play it first
- An AI will always be limited by the features a human being has hardcoded for them.
- Would be nice if somehow we could *learn* the features as we interact with the MDP.

- We are going to talk about a method that does this. The change we make to what we have done is pretty basic...we just have to move away a little bit from linear approximation.
- Next, we'll talk about neural networks.
  For a while, we'll just be talking about neural networks as if this was a generic ML class.
  Eventually, we'll bring everything back to RL.
- I'll follow Michael Nielsen's notes on the subject, which I think are pretty well written:
  http://neuralnetworksanddeeplearning.com
- I encourage you to read those notes yourselves in parallel with the lectures.



The basic building block of a neural network: the perceptron

Inputs are real numbers x_1, x_2, ... , x_n

The output is also a real number

output = sigma(w_1 x_1 + w_2 x_2 + ... + w_n x_n + b)

w_1, ..., w_n and b are weights. The quantity b is sometimes called the bias.

sigma: R — > R is a function called an "activation function"

A typical choice is sigma(x) = max(x,0) called a ReLu

Other choices for sigma are also possible

- Some simple examples.

- $f(x) = \sigma_{\mathrm{ReLu}}(x) = \max(x,0) = \begin{cases} x & x \geq 0 \\ 0 & \text{else} \end{cases}$

- $f(x) = \sigma_{\mathrm{ReLu}}(x-1) = \max(x-1,0) = \begin{cases} x-1 & x \geq 1 \\ 0 & \text{else} \end{cases}$

- $f(x) = \sigma_{\mathrm{ReLu}}(x+1) = \max(x+1,0) = \begin{cases} x+1 & x \geq -1 \\ 0 & \text{else} \end{cases}$

- $f(x) = \sigma_{\mathrm{ReLu}}(2x+4) = \max(2x+4,0) = \begin{cases} 2x+4 & x \geq -2 \\ 0 & \text{else} \end{cases}$

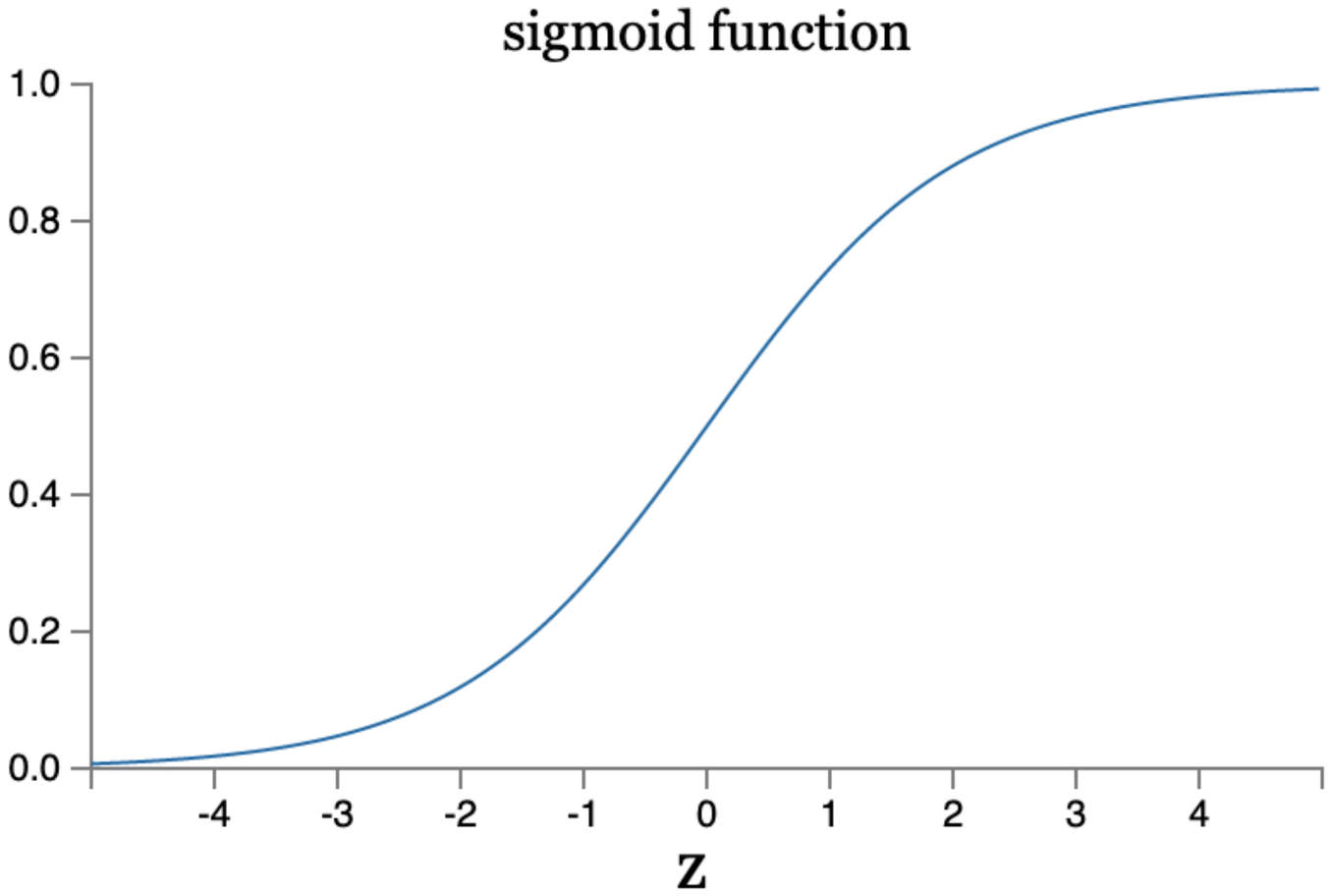
- Let's take a couple of examples of what a perceptron can do.
- We could have $y = \sigma_{\text{ReLu}}(x_1 - x_2 - 1)$, corresponding to weights $[1, -1]$ and bias -1.

  If $x_1 \leq x_2 + 1$, this equals zero.

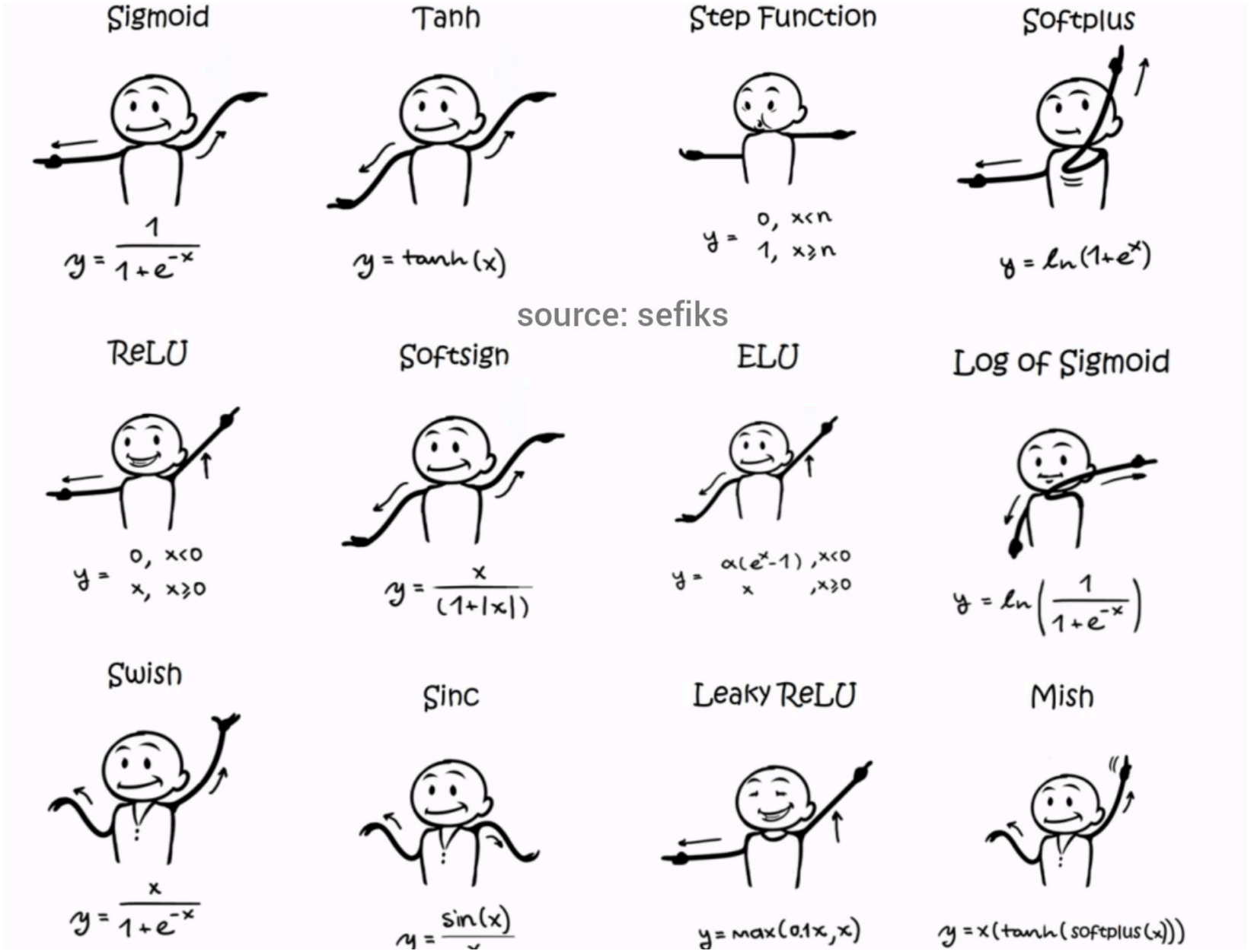
  If $x_1 > x_2 + 1$, this equals $x_1 - x_2 - 1$
- Let's define $\sigma_{\text{threshold}}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else} \end{cases}$
- Suppose $y = \sigma_{\text{threshold}}(x_1 - x_2)$.

  This equals 1 if $x_1 \geq x_2$

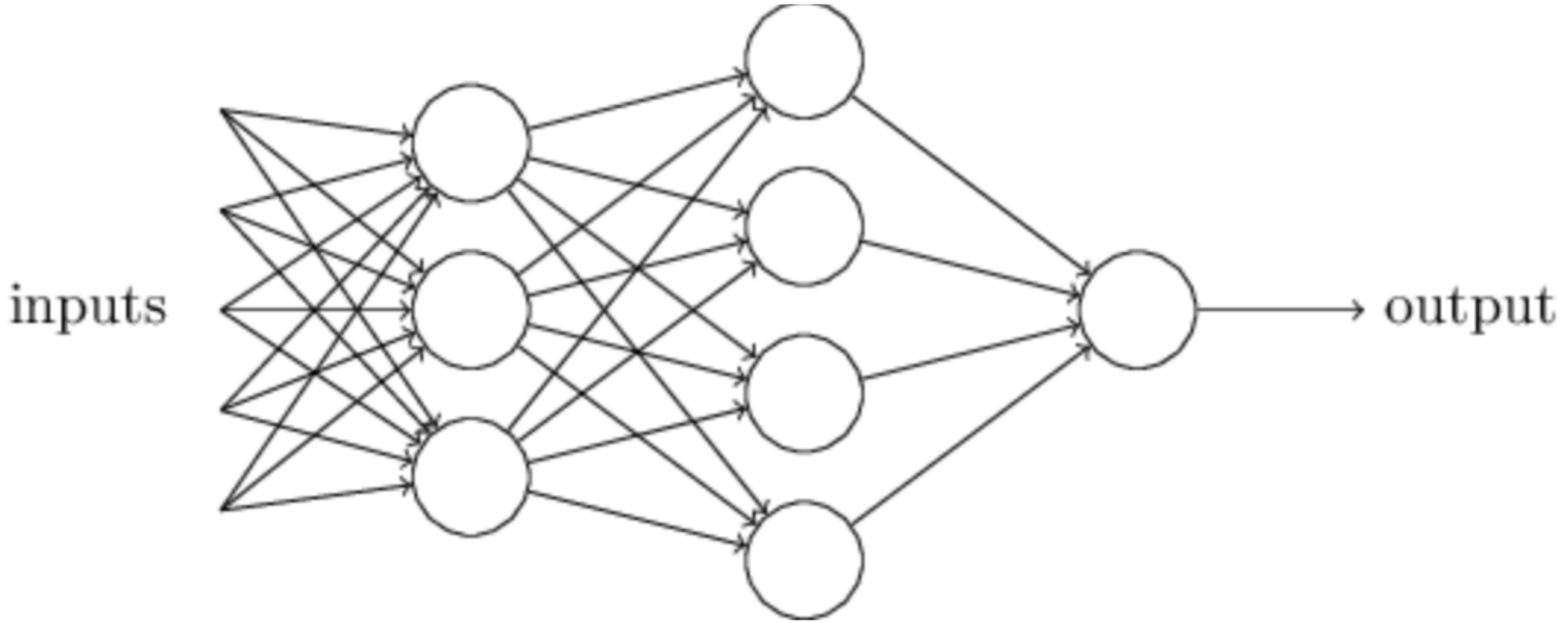
  It equals 0 if $x_1 < x_2$
- The problem with this choice of activation function is that it is not continuous.
- We can mimic it with a continuous function as described on the next slide.



$$\sigma_{sigmoid}(x) = 1/(1+e^{-x})$$



Other activation functions



Each circle is a perceptron

This is what a neural network is: a collection of perceptrons such that the outputs of some perceptron fit into the others

This is called a feedforward architecture

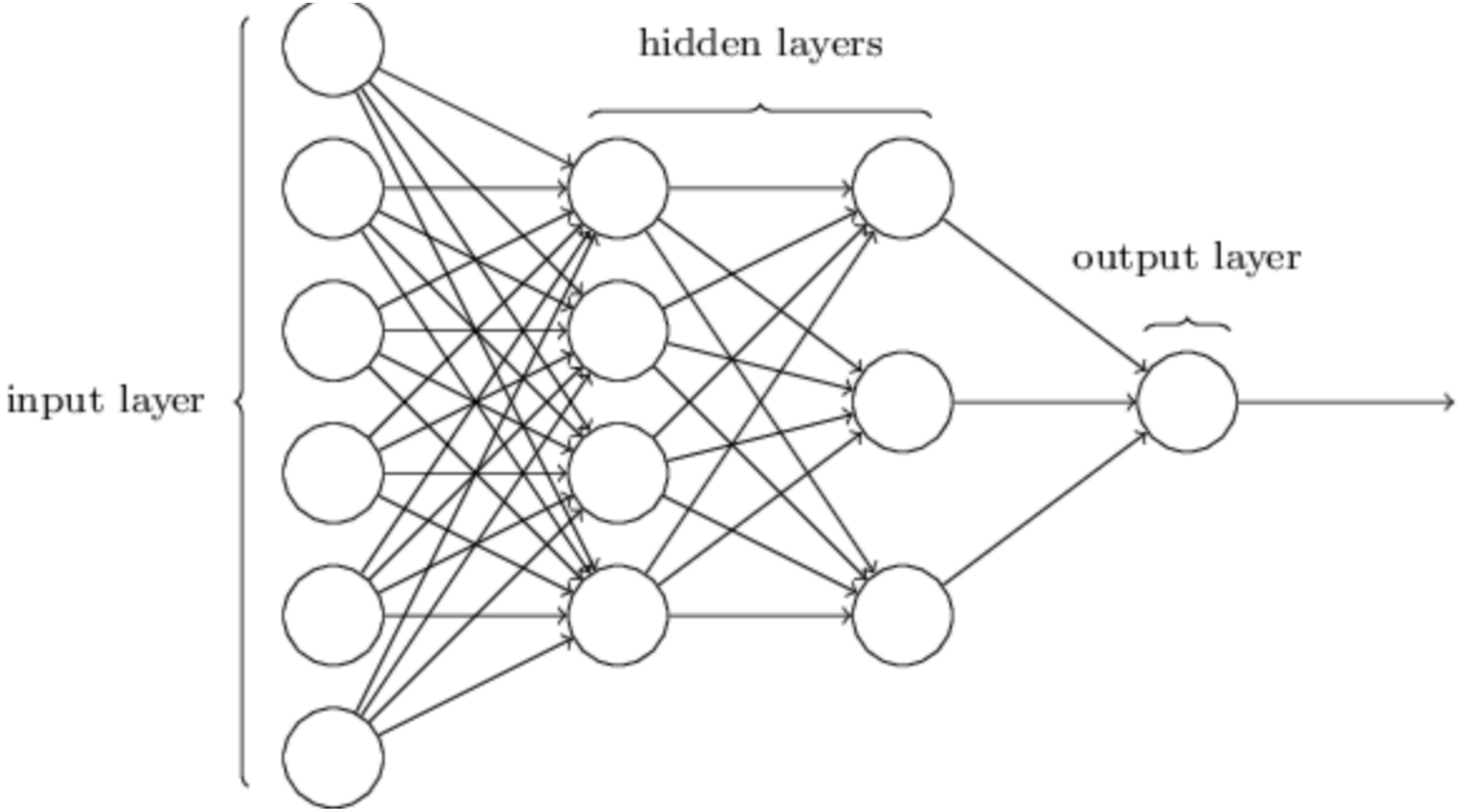
The number of layers is called the depth of the network

The number of perceptrons in each layer is called the width of that layer

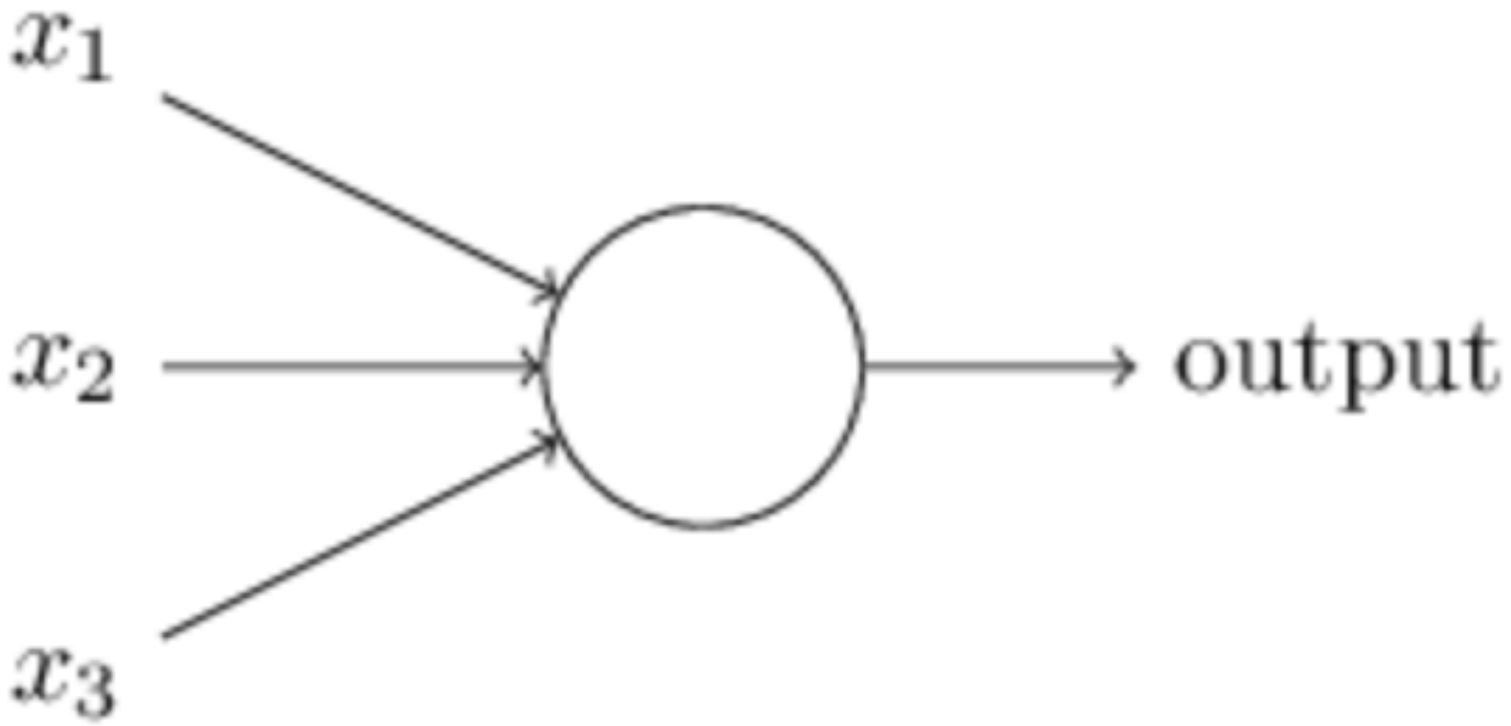
If all the width are the same, we can talk about the width of the network

The depth + widths + activation functions are usually referred to as the architecture of the network

(sometimes activation functions for different neurons are taken to be different)



You usually put the input as a layer as well (convention).
This network takes six numbers and maps them to one number
The perceptrons are also sometimes called neurons



If we fix the weights, we can think of a perceptron (or a neural network) as mapping inputs to outputs

We think of the bias as a weight too

Suppose the weights are [1,1,1,3] and the activation function is ReLU.

Then f_{[1,1,1,3]}(1,1,1) =max(1 x 1 + 1x1 + 1x1 + 3, 0) = 6

On the other hand, f_{[1,1,1,0]}(-10,0,0) = max(1 x (-10) + 1x 0 + 1x0 + 3,0) = 0

- The problem we typically encounter is: given a neural network architecture and a collection of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$

  find the weights so that the resulting NN maps each x_i to y_i.
- Often x_i is a vector while y_i is a scalar. For example, x_i could be the pixel-level representation of an image. The quantity y_i could be 1 or 0 depending on whether the image contains a cat or a dog.
- Regardless of whether y_i is a scalar, it is called the "label."
- Once we've fixed the architecture, let $f_w(x)$ be the output of the neural network with weights w on an input x

  [yes this is confusing because we should perhaps write $f_{w,b}(x)$....but we'll just write f_w because the bias is also a weight]
- If its not possible to find weights w such that $f_w(x_i) = y_i$ for all i, then we want weights such that $f_w(x_i) \approx y_i$ as well as possible. For example, we want to solve $\min_w \sum_{i=1}^{m} \frac{1}{2m}(f_w(x_i) - y_i)^2$.
- How do we solve this?

- OK, so the "data set" is the collection $(x_1, y_1), \ldots, (x_n, y_n)$ and that is fixed.

  The architecture is fixed (we discuss later how to choose).

  But the weights need to fit to the data.

- So we want to solve $\min_w \sum_{i=1}^{m} \frac{1}{2m}(f_w(x_i) - y_i)^2$.

- OK, we know how to do this: we use gradient descent.

  $$w_{t+1} = w_t - \alpha_t \frac{1}{m} \sum_{i=1}^{m} (f_{w_t}(x_i) - y_i) \frac{\partial f_w(x_i)}{\partial w}(w_t)$$

- There are two issues to address here:

  — a gradient requires a single pass through the entire data set

  — this can be really unrealistic if the data set is large (more on this later)

  — but at an even more basic level, what is $\frac{\partial f_w(x_i)}{\partial w}(w_t)$?

  We need to differentiate the NN with respect to the weights.

- Consider Imagenet, a data set released in 2010 which has been a catalyst for much of ML research.
- 14 Million images.
- 20,000 categories.
- About 150GB in total.
- If each iteration of gradient descent meant summing over the entire dataset, you wouldn't be able to do many iterations.
- Solution: for the last term, $w_{t+1} = w_t - \alpha_t \frac{1}{m} \sum_{i=1}^{m} (f_w(x_i) - y_i) \frac{\partial f_w(x_i)}{\partial w}(w_t)$

  pick b data points at random, and just average the terms corresponding to them:

  $$w_{t+1} = w_t - \alpha_t \frac{1}{b} \sum_{i \in \{i_1, i_2, \ldots, i_b\}} (f_w(x_i) - y_i) \frac{\partial f_w(x_i)}{\partial w}(w_t)$$

- Motivation: this has the right expectation. This method is called SGD.

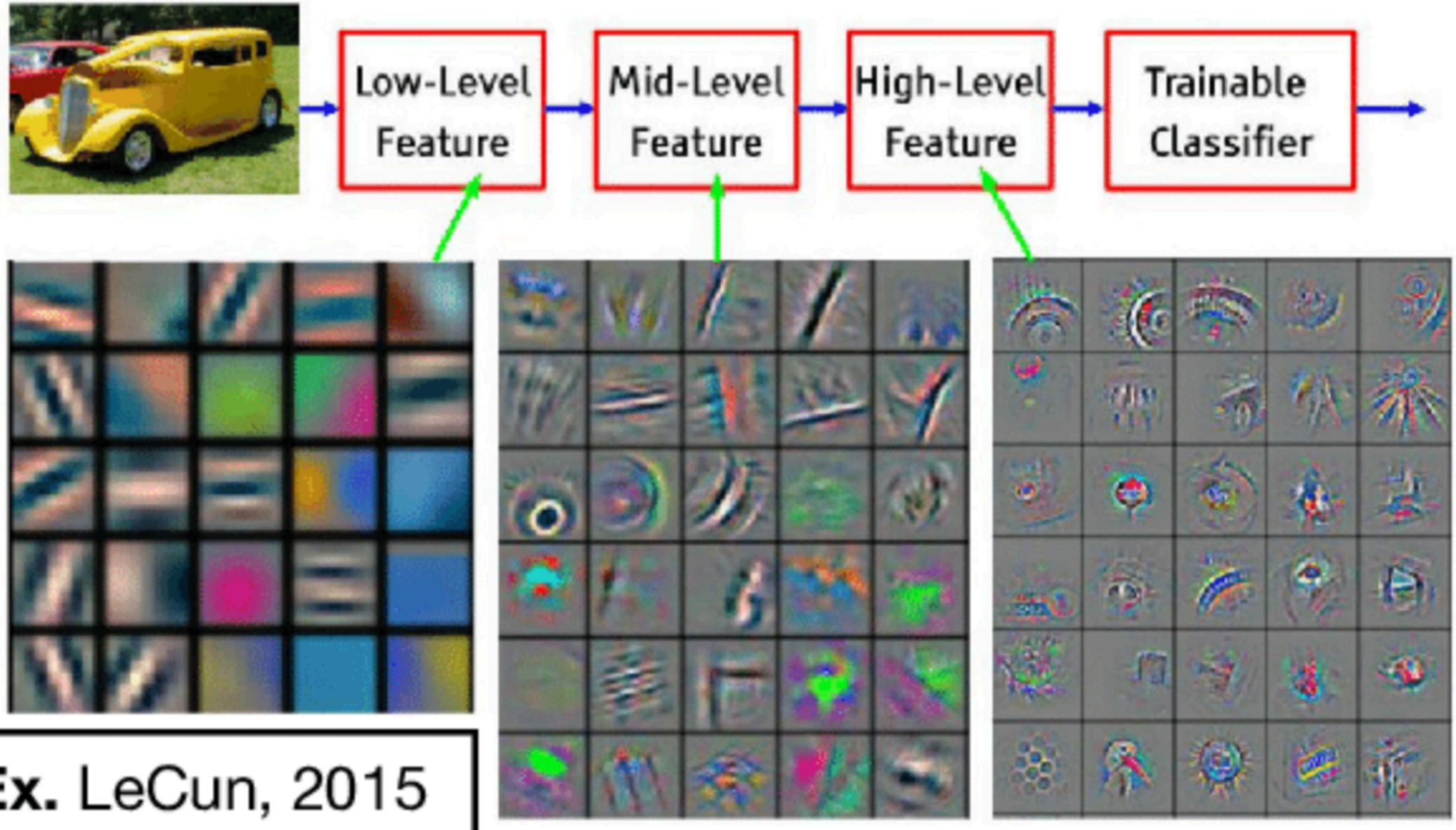
- Context: neural networks began to be the focus on ML research in the early 2010s after decades of neglect.
- Turns out that the main thing they needed was:
  — tons of data (e.g., millions of images)
  — large size (millions of parameters, large depth)
  — lots of SGD iterations (though now there are methods better than SGD).
- It's something of a controversy about who invented all this stuff first, see https://people.idsia.ch/~juergen/who-invented-backpropagation.html for your entertainment

- So let’s give this a try?
- ...not just yet. Need two more tricks which turn out to be really important to making this work.
- [details really matter here]
- First of all...let’s think of about why this might work.
- Key idea: image recognition is a composition.

  You have low-dimensional features (lines)

  These are ingredients of medium-dimensional features (brow, eye, nose)

  These in turn are ingredients of higher dimensional features (face, car)



Ex. LeCun, 2015

- This suggests: let us use something called a one-hot encoding.
- Suppose you are doing recognition of handwritten digits. The label is $\{0,\ldots,9\}$. We will encode 0 as $\begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ and 1 as $\begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ and so on, until 9 as $\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$.
- So the label, instead of being a number, is a vector in $\mathbb{R}^{10}$.
- Why is this better?
- The point is that if the class measures which combination of features, it is much easier to learn when the output is of this form.
- Consider the layer **before** the last one.
  Suppose the first entry = whether there is a brow in the picture
  Second entry = whether there is an eye in the picture
  Third entry = whether there is a wheel in the picture
  Four entry = whether there is a license plate number in the picture
  If we want to output either “face” or “car”, a natural way to do it is to check if
  “First entry + second entry” is bigger than “third entry + fourth entry”
- By contrast, if “car” mapped to output of 6 while face mapped to output of 9, the mapping would be more involved.

- So, for recognition of hand-written digits, we have the output as a vector in $\mathbb{R}^{10}$.
- We will interpret the number at location i as how likely the network thinks the output is $\mathbf{e}_i$ (this is the notation for the vector with a one in the i'th place and zero elsewhere).
  Higher = more likely.
- More formally, for a vector $z \in \mathbb{R}^k$, we define the function

$$f(z) = \frac{1}{e^{z_1} + \cdots e^{z_k}} \begin{pmatrix} e^{z_1} \\ e^{z_2} \\ \cdots \\ e^{z_k} \end{pmatrix}$$

- This is called the softmax function. Why does that name make sense?
- We use the soft-max as an activation function for the last layer.
- Note that, unlike the activation functions we used previously, this one depends on the entire vector.
- We thus interpret the output of the neural network, a vector in $\mathbb{R}^{10}$, as how likely it thinks each of the 10 possibilities is...and this actually makes sense because the output is a probability vector.

- Let’s take this one step further.

  Suppose the neural network outputs the distribution

  $$\begin{pmatrix} 3/4 \\ 1/4 \end{pmatrix}$$

  in the case of two possibilities.

  The correct answer is $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

  What cost should it pay?

- One possibility: $(3/4 - 1)^2 + (1/4 - 0)^2$.
- Actually, this is not a great choice of cost here. Let’s discuss why.

- Suppose your next exam is multiple choice.
- You are allowed to submit a probability distribution over the outcomes.
- For example, if you have a question with two answers A and B — and you think the probability that A is correct is 3/4 — then you are allowed to submit the distribution $(3/4, 1/4)$ as your answer.
- I will grade it as follows: if the answer is A, you will gain 3/4 of a point.
  If the answer is B, you will gain 1/4 of a point.
- Suppose you think the probability that the answer is A is 3/4. What should you do?
- One possibility: (3/4, 1/4). You expect to gain
  $(3/4) \cdot (3/4) + (1/4) \cdot (1/4) = 10/16.$
- Second possibility: $(1,0)$. You expect to gain $(3/4) \cdot 1 + (1/4) \cdot 0 = 3/4 = 12/16.$
- This scheme encourages you to lie!

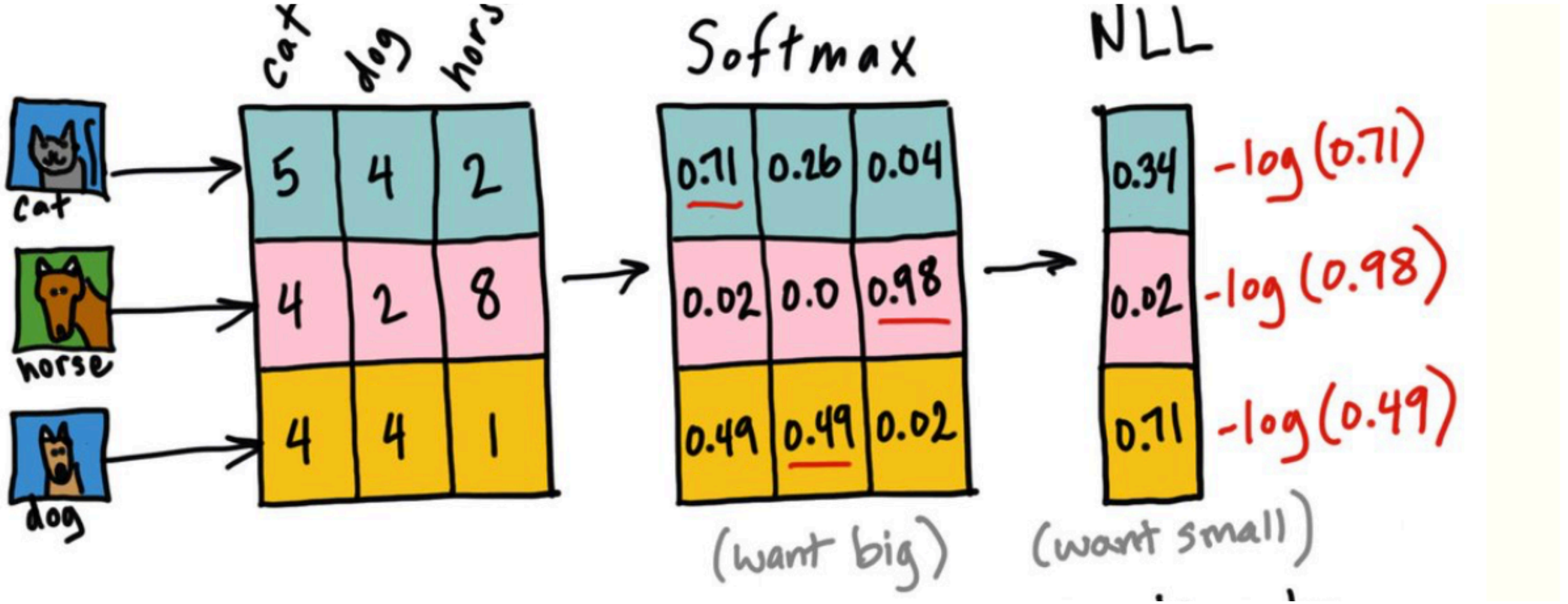
- This gives rise a fun question:

  suppose you, as a student, submit the probability distribution $(q, 1-q)$.
  What grade should I give you to encourage you to tell the truth?
- Suppose I assign you a grade of $f(q)$. Naturally, I announce the function $f(\cdot)$ ahead of time.
- If you think the true distribution is $(p, 1-p)$ you will submit $(q, 1-q)$ that maximizes $pf(q) + (1-p)f(1-q)$ over $q \in [0,1]$.
- You will solve $pf'(q) - (1-p)f'(1-q) = 0$.
- So I need to choose f with the property that $pf'(p) - (1-p)f'(1-p) = 0$ for all p.
- A natural choice: $pf'(p)$ should be constant.
- $f'(p) = C/p$
- $f(p) = C \log p + D$
- Let's further set $f(1) = 1, f(1/2) = 0$. This means $f(p) = \log_2(p) + 1$.
- Equivalently, can choose $f(p) = \log_2(p)$.
- Punchline: I've got to give you points equal to the log of the probability you gave to the correct answer.

- Back to neural networks. Your neural network outputs a vector of probabilities in $\mathbb{R}^{10}$.
- The “credit” you should give it is the log of the entry corresponding to the correct answer.
- But our algorithms minimize the loss the NN suffer.
- So you say the loss it suffers is the negative of the log of the entry corresponding to the correct answer:

$$-\log \frac{e^{z_i}}{e^{z_1} + e^{z_1} + \cdots e^{z_k}}$$

- This loss is nonnegative and equals zero exactly when you assign a 1 to the correct answer and zero to everything else.
- But also, this loss will encourage you never to give a zero probability to any posssibility. Why?



Picture —> Last layer of NN —> after softmax —> log loss suffered

- OK, time to try this on a real world example.
- We'll use the MNIST data set https://en.wikipedia.org/wiki/MNIST_database
- Contains 60,000 images of hand-written digits along with labels. Each image is 28x28 pixels.
- Created in 1998. The classic paper http://yann.lecun.com/exdb/publis/pdf/lecun-98.pdf achieving 99% recognition rate on this.
- https://colab.research.google.com/drive/1ZpOWilfKB9ZpKYumKIEfxGm4wyeLS6b_?usp=sharing
- Let's try on CIFAR-10
- https://colab.research.google.com/drive/1ZpOWilfKB9ZpKYumKIEfxGm4wyeLS6b_?usp=sharing

- ADAM is an alternative to SGD based on two simple ideas.
- Let's take the SGD iteration

  $$w_{t+1} = w_t - \alpha_t \left[\nabla f(w_t) + \text{noise} \right]$$

- In neural networks, the noise comes from the selection of the mini-batch.
- How can we combat the noise?
- First idea: since the gradients don't change that quickly (because α_t is small), let's take linear combination of the gradients:

  Initialize $v_0 = 0$

  $$g_t = \nabla f(w_t) + \text{noise}$$

  $$v_t = \beta v_{t-1} + (1 - \beta) g_t$$

  $$w_{t+1} = w_t - \alpha_t v_t$$

- Let's consider this recursion:

  $v_t = \beta v_{t-1} + (1-\beta) g_t$
- We have $v_1 = \beta v_0 + (1-\beta) g_1$
- Next, $v_2 = \beta v_1 + (1-\beta) g_2 = \beta^2 v_0 + \beta(1-\beta) g_1 + (1-\beta) g_2$

  ...this is a convex combination!
- Next, $v_3 = \beta v_2 + (1-\beta) g_3 = \beta^3 v_0 + \beta^2(1-\beta) g_1 + \beta(1-\beta) g_2 + (1-\beta) g_3$

  ...this also is a convex combination.
- In general, you can show that convex combinations of convex combinations are convex combinations.
- The point: v_t is a convex combination of all the gradients, with the coefficients on the past decaying geometrically.
- In general, $v_t = \beta^t v_0 + \beta^{t-1}(1-\beta) g_1 + \cdots + (1-\beta) g_t$

- Let's look at this equation: $v_t = \beta^t v_0 + \beta^{t-1}(1-\beta)g_1 + \cdots + (1-\beta)g_t$
- A sort of problem is that because $v_0 = 0$, we have that there will be a bias towards zero.
- Not a huge problem: the bias will decrease over time. The coefficients on everything except v_0 add up to $1 - \beta^t$.
- One possibility is to start the method at time one initialized at g_1. That is what I would do.
- What the creators of ADAM did was to choose the approach

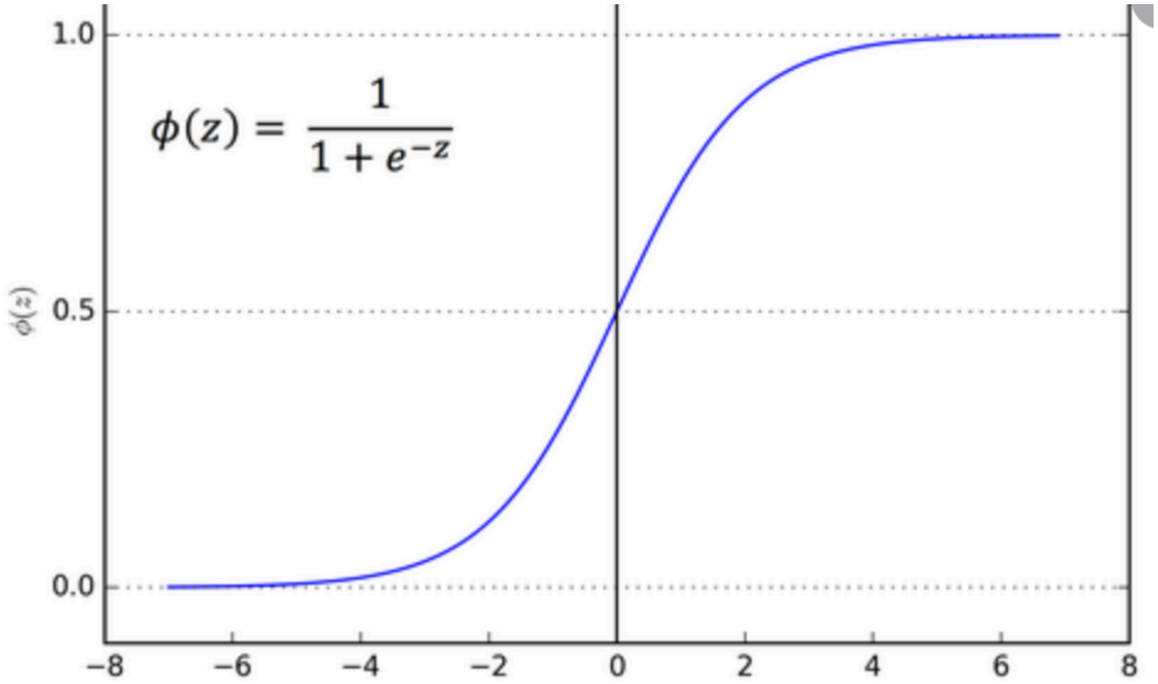
$$g_t = \nabla f(w_t) + \text{noise}$$

$$v_t = \beta v_{t-1} + (1-\beta)g_t$$

$$\hat{v}_t = \frac{v_t}{1-\beta^t}$$

$$w_{t+1} = w_t - \alpha_t \hat{v}_t$$

- Next insight: we can have the situations where the gradients are either huge or close to zero.
- If you get a huge gradient while your step-size α_t is not sufficiently small, your update will be all over the place.
- On the other hand, consider any of our activation functions that were flat at the extremes.



For very large or very small z, the gradient is close to zero

- Solution: try to normalize by the norm of the gradient, i.e., instead of updating as

  $$w_{t+1} = w_t - \alpha_t \left[\nabla f(w_t) + \text{noise} \right]$$

  instead update as

  $$w_{t+1} = w_t - \alpha_t \frac{\nabla f(w_t) + \text{noise}}{|| \nabla f(w_t) + \text{noise} ||_2}$$

- This is a good idea, but there is another complication.
  Suppose everything is perfect...on every minibatch, you classify things perfectly.
  All your gradients are zero, and you shouldn't move anywhere.
- Unfortunately, sometimes gradients which are zero are evaluated to something like 10^{-12} due to round-off error.
- Solution:

  $$w_{t+1} = w_t - \alpha_t \frac{\nabla f(w_t) + \text{noise}}{\sqrt{|| \nabla f(w_t) + \text{noise} ||_2^2 + \epsilon}}$$

  where ϵ is usually taken to be 10^{-8}.
- So a gradient of 10^{-6} will get normalized into a direction to follow. But a gradient of 10^{-10} will get ignored.
- OK, but this is SGD...we are using a linear combination of the gradients.
- No problem, just normalize that.

- This finally brings us to ADAM:

  Initialize $v_t = 0, q_t = 0$

  $g_t = \nabla f(w_t) + \text{noise}$

  $v_t = \beta_1 v_{t-1} + (1 - \beta_1) g_t$

  $\hat{v}_t = \frac{v_t}{1 - \beta_1^t}$

  $q_t = \beta_2 q_{t-1} + (1 - \beta_2) || g_t ||_2^2$

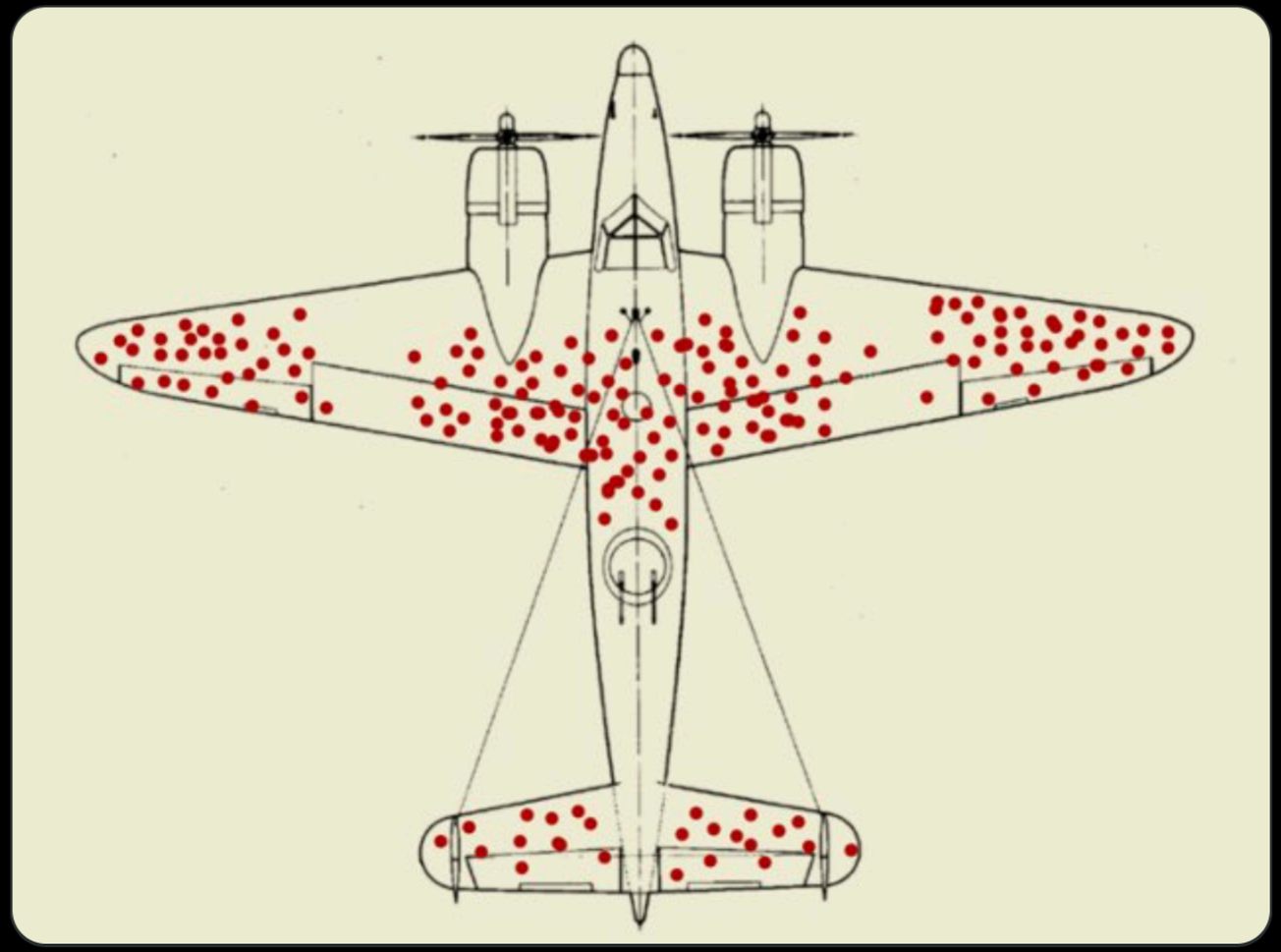
  $\hat{q}_t = \frac{q_t}{1 - \beta_2^t}$

  $w_{t+1} = w_t - \alpha_t \frac{1}{\sqrt{\hat{q}_t + \epsilon}} \hat{v}_t$

- Default values $\beta_1 = 0.9, \beta_2 = 0.999$. Why are they different? I have no idea.
- There is some skepticism that ADAM is a good method: https://parameterfree.com/2020/12/06/neural-network-maybe-evolved-to-make-adam-the-best-optimizer/

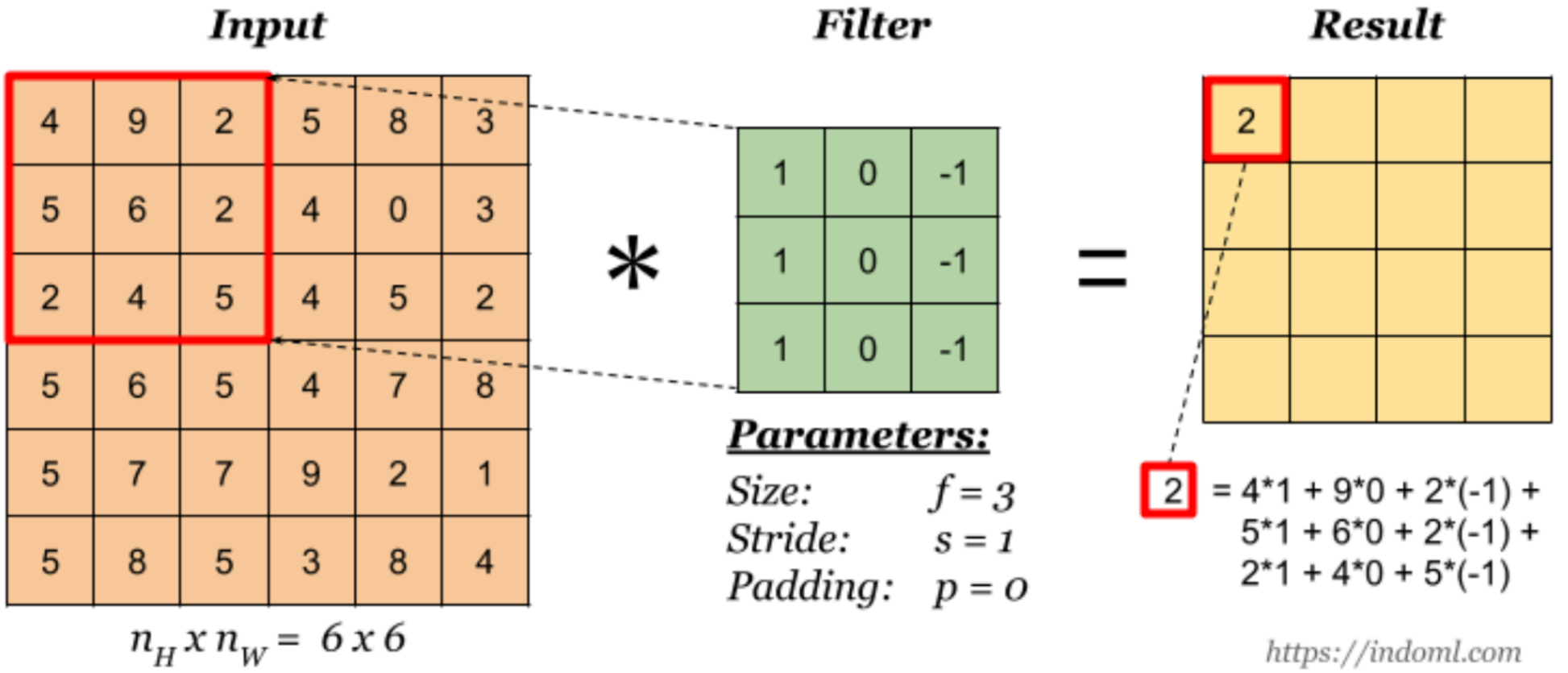
Roger Grosse
@RogerGrosse

SGD and Adam are good enough for training most neural net architectures.



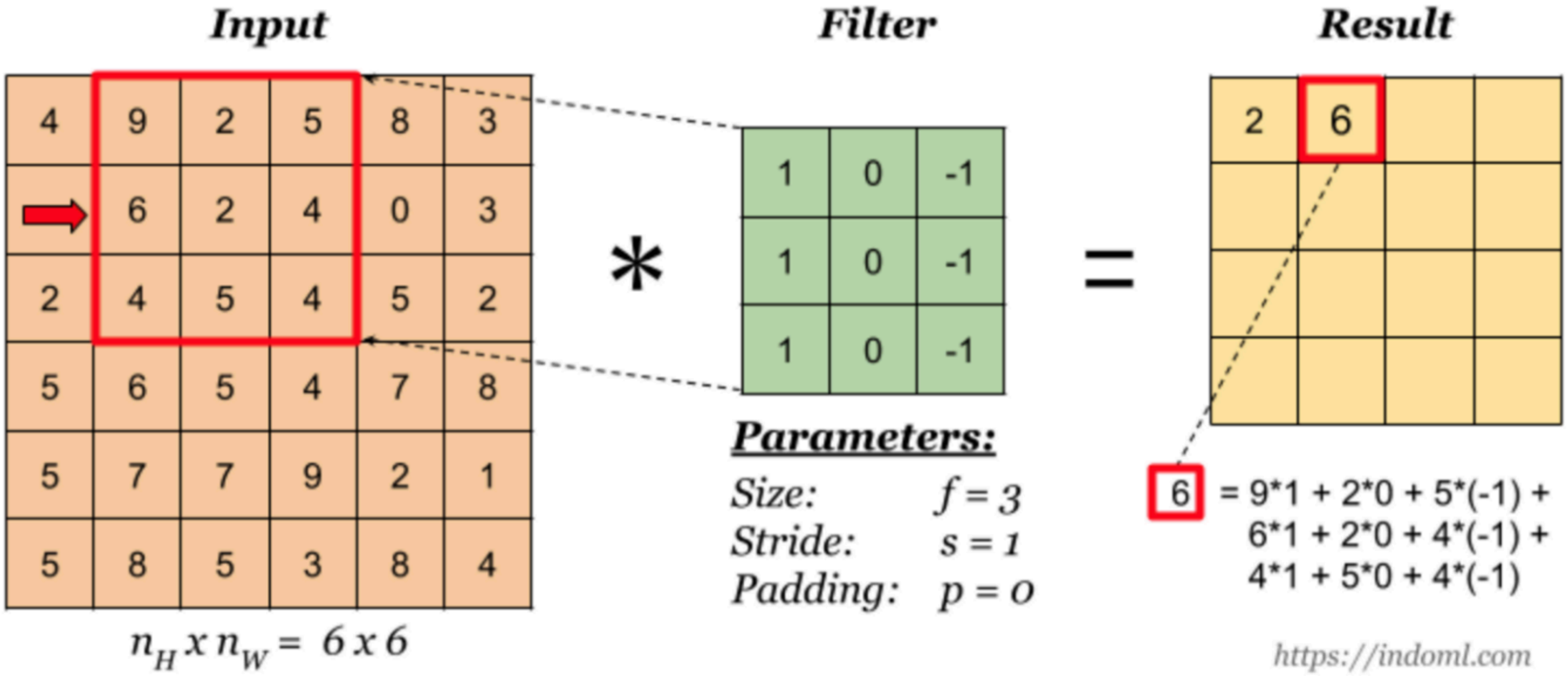
2:24 PM · Oct 8, 2020 · Twitter Web App

- Further: image recognition is “shift invariant.”
  If you take a picture of a 3, shift it to the right, it’s still recognizable as a 3.
- ...but a neural network as we have described it may not be shift invariant.
- Counterpoint: if you give it enough 3s (and other digits) shifted all over the place, it will learn to be shift invariant.
- In practice: not quite. Part of the problem is that it is not clear what, in the end, the training process converges to — the problem of local vs global minima.
- Better approach: code shift invariance INTO the arhitecture.



Convolutional operation.

We take the "inner product" by multiplying the corresponding elements and adding to get the two

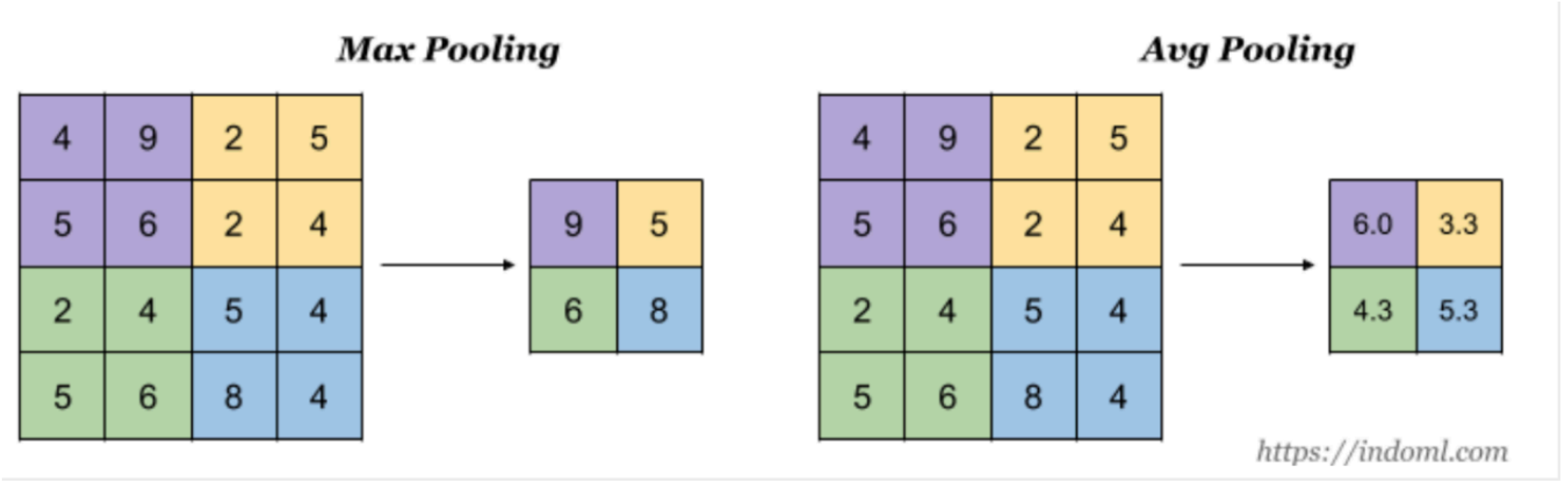


Stride Length = 1

We translate the input by one unit to the right and redo the operation
We'll then translate it one unit to the right and redo
And then one more unit to the right and redo
Then we'll reset to the starting position and move one unit DOWN
This is called a convolutional layer

- A convolutional layer is just taking linear combination of the pixel values...

  ....except it is constrained to take linear combinations of pixels that are close.
- As indicated earlier, if we flattened the image into vectors and added all to all connections, in principle the NN could reproduce a convolutional layer.
- But this way there is a “bias” towards the things it learns, and that bias makes sense.
- Convolutional layers are usually followed by pooling layers.



Pooling layers

- A convolutional layer is shift invariant.
- Indeed, suppose you shift the image i to right by one pixel to obtain i_s.

  If o is the output of i through a convolutional layer, then the output of i_s will be

  o shifted to the right by one pixel.

  [here we are ignoring border effects — imagine the imagine goes on to infinity in each direciton].
- The usual architecture for images

  — convolutional layer followed by pooling layer

  — after these are stacked up, they are followed by one or more fully connected layers.
- Let's try it out: With this in mind, let us try this out: https://colab.research.google.com/drive/1AdeopLi_9trH-amct5kpq9k0KJkEoGew?usp=sharing

- OK, now let's come back to RL. The last thing we did was try to approximate the state by a linear function

  $V \approx \Phi\theta$

  or

  $V(s) = \sum_i \theta_i \phi_i(s)$

  which required knowing the features $\phi_i(s)$ in advance — but then we learned the weights θ over the course of a TD iteration.
- We will now try to approximate

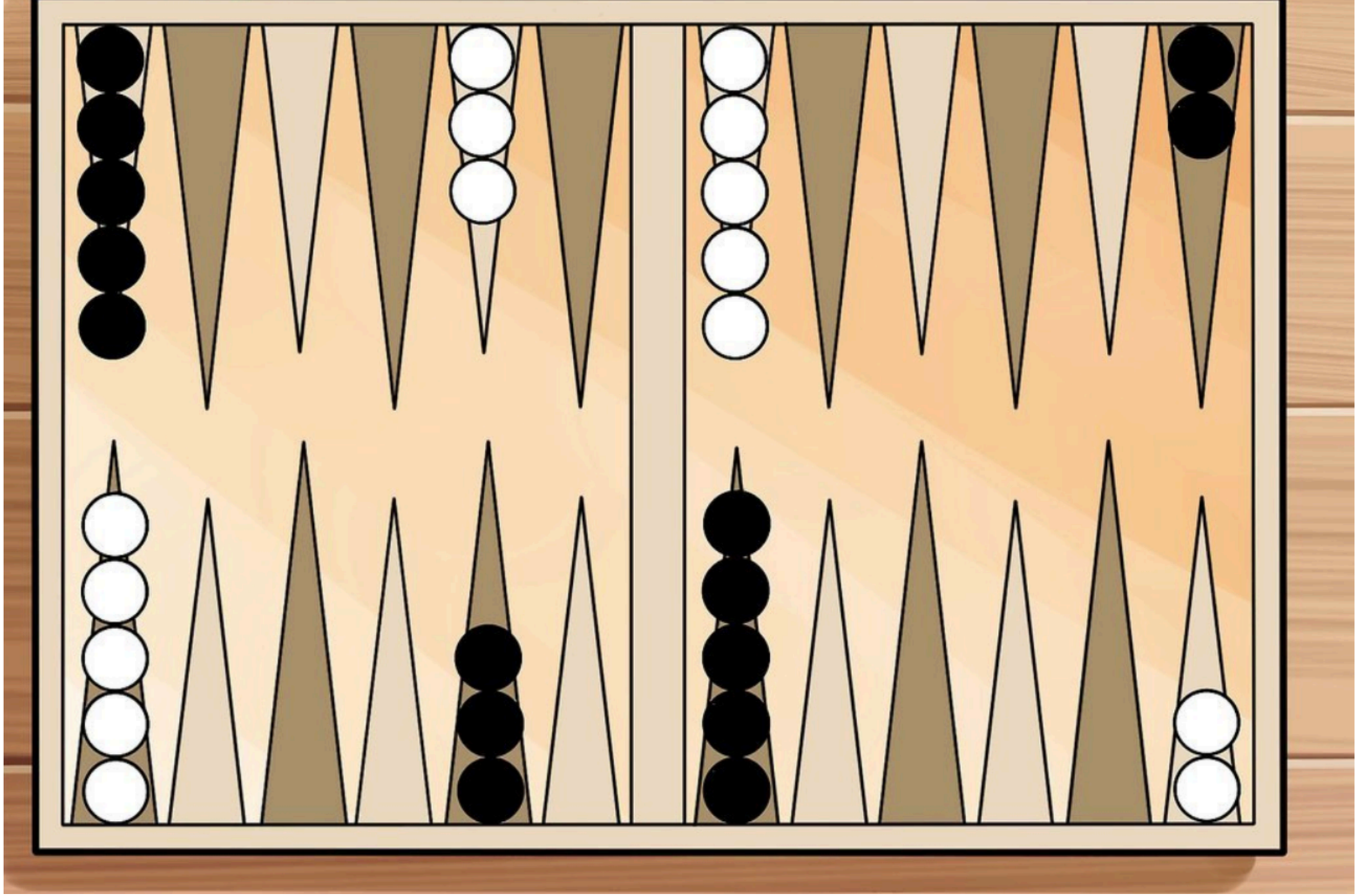
  $V(s) \approx f_{NN,\theta}(s)$

  where $f_{NN,\theta}(\cdot)$ is a neural network with weights θ.

  s here will be fed into the input of the neural network.
  If the state is an image, then it will be flattened as before.
- Over the course of the algorithm, we will try to learn the best weights θ.
- If this works, it solves the problem of having to hard-code features.

- Let's take this idea and make it more concrete.
- Let's come back to the TD setting. We have an MDP and a policy π. Our goal is to compute V_π.
- Two constraints:
  — we don't know the MDP and can only interact with it through experience
  — the state is very high-dimensional and so we want to keep track of a low-dimensional approximation.
- We define the function V_θ by fixing some NN architecture that will take the state as input and output a real number, and we initialize all the weights θ randomly.
- Now we run the method. We start at some state s. We take action a by sampling from $\pi(s)$. We obtain reward r and move to state s'.
- We update as:
- This is $\theta_{t+1} = \theta_t + \alpha_t(r + \gamma V_\theta(s') - V_\theta(s))\,\nabla_\theta V_\theta(s)$.
- The quantity $r + \gamma V_\theta(s') - V_\theta(s)$ is called the temporal difference.
- If the temporal difference is positive, this update tries to increase $V_\theta(s)$.
- If the temporal difference is negative, this update tries to decrease $V_\theta(s)$
- This is called **Neural TD.**

- The first major success for this came in the late 1980s, when Gerald Tesauro used it to develop a computer program that played backgammon.
- In the 1980s, Tesauro’s program was rated as the top computer program. W=
- By the mid 1990s, it was playing about as well as the best human players.
- By mid 2000s, it was so much better that the best human players that most players were trying to emulate it’s style.



Each player rolls a dice and can make a move that depends on the outcome of that dice
The winner is the first person to move pieces across the board

- Tesauro's method: set reward equals 1 if black wins and zero if white wins.
- The train a NN to predict the value of a position.
- Input: how many blocks at each location, and whether they are white and black (with some clever tricks, he got this to 16 numbers).
- The network: one hidden layer, about 80 hidden units.
- He then used TD learning to try to learn the value of each position with $\gamma = 1$.
- Interpretation: value is the probability that black wins.
- To turn that into a program that plays moves is easy: simply go through all the moves and select the ones leading to a position with the highest value.

- This was before the era of multi-layer, deep, networks, so it was natural to train a NN with only one hidden layer.
  Also, think about how slow computers were back in the late 80s....
- What I outlined ended up doing OK, it played like an intermediate player. This was the late 80s.
- Tesauro improved it in several ways:
  — use six networks depending on the stage of the game (beginning, early, middle, etc).
  — but all six networks had a single layer
- Second, he began using a few hard-coded features which he fed into the input of the network (e.g., how close each side's pieces are to the end).

- Next step: use a little look-ahead.
- Go through all the moves, select, say, the top 5 or 6 the NN likes.
- For each of the moves, pretend you've made it, go through every dice roll your opponent makes, and look at all of their moves.
- Evaluate the position after the best moves of the opponent for every dice roll.
- To estimate the value of the position for your opponent, you want to average the result over all the dice rolls.
- This is called 2-ply search.
- The version that was superior to humans actually used 3-ply search.

Opinion: Have Computers Ruined Backgammon?

16th September 2015 By Julia Hayward

John Hurst writes:

Ex World Champion Mike Sovodny once said that the emergence of computer programs such as Snowie had, for him, to some extent lessened the appeal of backgammon. Certainly playing props, at which he was an expert, is now pointless as all you need to do to find the correct move or doubling decision is to stick the position in the computer.

"It says that two on the bar is better by far, Professor"

In the past even experienced players would differ wildly in their opinions of the right move and nobody could be sure who was correct; that uncertainty made the game intriguing. Some enthusiasts would spend countless hours rolling out positions to find the right answer and presumably that work would be rewarded by their improved results. Now all a player has to do is ask Snowie or XG. Different players would develop different playing styles, now all players who take the game seriously try to ape the computer as closely as possible.There is no point in trying to develop your own strategies, if your move differs from the computer then it is wrong – it's as simple as that.

- Let's come back to RL again. We've talked about using TD learning with a neural network. Why not Q-learning?
- Let's think about how this works. We fix an architecture and at each step of the algorithm we update the weight θ such that $Q_\theta(s, a)$ is our approximation to the optimal value function.
- Here s and a are the input to the neural network (we can stack up). The θ is the weight of the neural network.
- So we are at state s, we choose action a, obtain a reward r, and transition to s'.

  Define $V_{\text{target}} = \max_{a'} Q_\theta(s', a')$.
- Update $\theta_{t+1} = \theta_t + \alpha_t(r + \gamma \max_{a'} Q_\theta(s', a') - Q_\theta(s, a)) \nabla_\theta Q_\theta(s, a)$
- This is called **Deep Q-learning**. Somehow, it's called that even when the architecture you choose is not all that deep...
- Deep Q-learning + convolutional neural networks are responsible for some of the more recent successes of RL.

- One way to improve this is to observe that there's no reason to generate new states all the time.
- Instead, every time you encounter (s, a, r, s') add it to an "experience replay buffer."
- Then, at each step, you could either update

$$\theta_{t+1} = \theta_t + \alpha_t(r + \gamma \max_{a'} Q_\theta(s', a') - Q_\theta(s, a)) \nabla_\theta Q_\theta(s, a)$$

or you could also sample a random batch of size b from the buffer and and update as

$$\theta_{t+1} = \theta_t + \alpha_t \frac{1}{b} \sum_{s,a,s',r \text{ in batch}} (r + \gamma \max_{a'} Q_\theta(s', a') - Q_\theta(s, a)) \nabla_\theta Q_\theta(s, a)$$

- Experience replay significantly improves performance in practice.

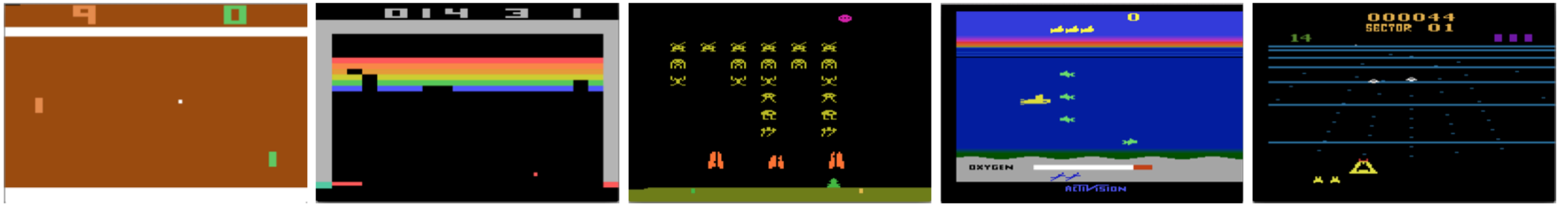


Figure 1: Screen shots from five Atari 2600 Games: (*Left-to-right*) Pong, Breakout, Space Invaders, Seaquest, Beam Rider

The 2013 paper of Mnih et al. showed really good performance for DQN

Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih **Koray Kavukcuoglu** **David Silver** **Alex Graves** **Ioannis Antonoglou**

Daan Wierstra **Martin Riedmiller**

DeepMind Technologies

{vlad,koray,david,alex.graves,ioannis,daan,martin.riedmiller} @ deepmind.com

Abstract

	B. Rider	Breakout	Enduro	Pong	Q*bert	Seaquest	S. Invaders
Random	354	1.2	0	−20.4	157	110	179
Sarsa [3]	996	5.2	129	−19	614	665	271
Contingency [4]	1743	6	159	−17	960	723	268
DQN	**4092**	**168**	**470**	**20**	**1952**	**1705**	**581**
Human	7456	31	368	−3	18900	28010	3690
HNeat Best [8]	3616	52	106	19	1800	920	**1720**
HNeat Pixel [8]	1332	4	91	−16	1325	800	1145
DQN Best	**5184**	**225**	**661**	**21**	**4500**	**1740**	1075

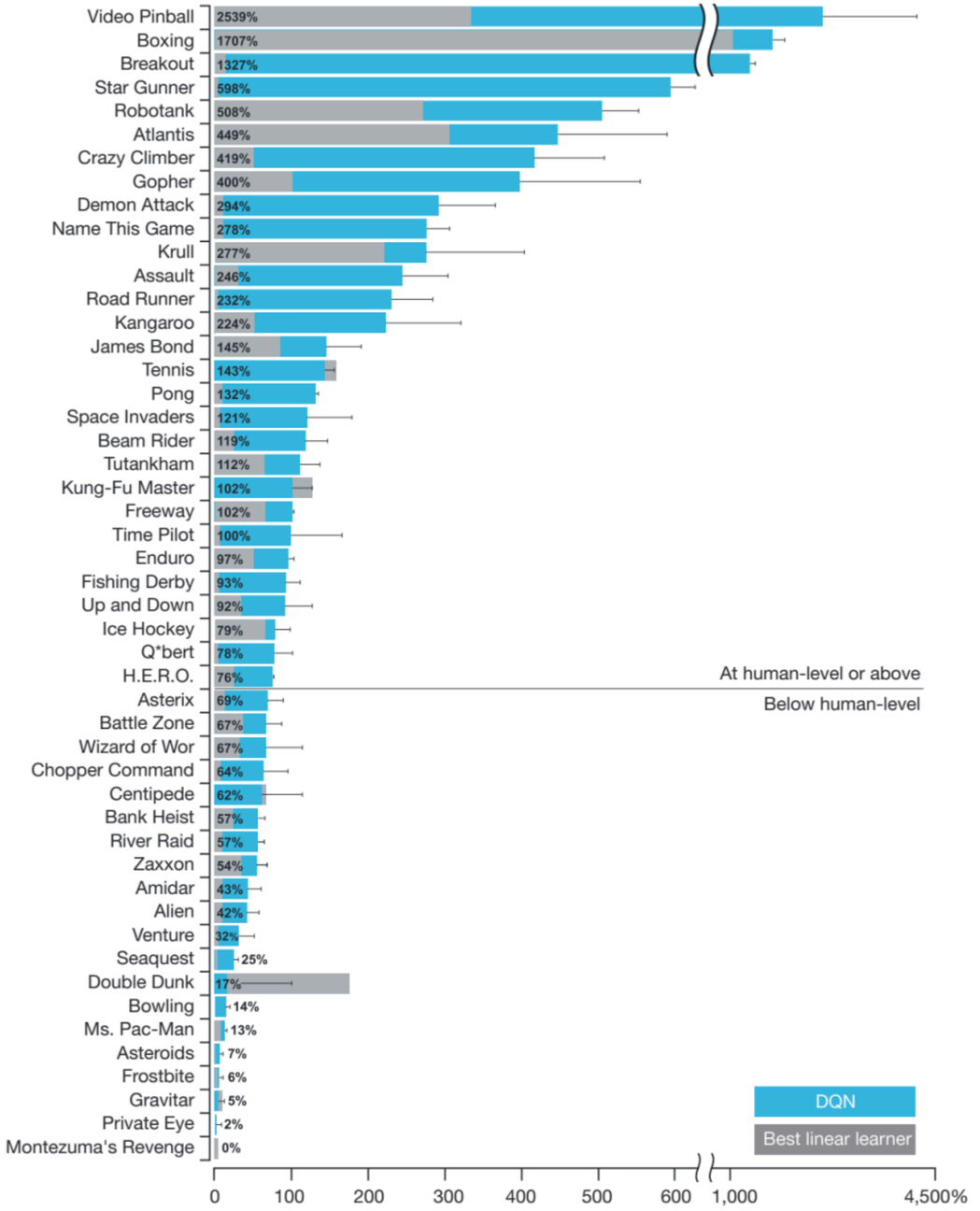
Their main result

- How they did it: first step is to pre-process the data (very important in practice).
  They take the images, make them grayscale, downscale them, and crop to show the “playing area” only.
  In the end, a 210x160 image is reduced to 84x84.
- They stack up the last four frames (after pre-processing) into the state. So in this case the state has dimension $4 \cdot 84^2$.
- Two convolutional layers (16 and 64 filters), followed by a fully connected layer (512 neurons), all with ReLu activations.

Algorithm 1 Deep Q-learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to equation 3
  end for
end for

The full paper may be looked at https://arxiv.org/abs/1312.5602



Two years later the same group was able to get more impressive results

- To explain how they did, let's talk about a core problem of this method: DQN consistently over-estimates true Q-values.
- Suppose you take a simple RL problem where the state space is sufficiently small that you can solve it exactly.

  You find the true Q-values $Q^*(s, a)$.
- You then solve it with DQN. You have your final Q-values $Q_{\theta*}(s, a)$ where $\theta*$ is the set of weights when you stop your method.
- Empirical finding: you will typically have $Q_{\theta*}(s, a) > Q^*(s, a)$.
- Think about why this is....

- Let's take a simple and wrong model.
  When we use DQN, there will be errors.
  Let's assume these errors are Gaussian and independent across nodes.
- That is, $Q_{\theta^*}(s,a) = Q^*(s,a) + w_{sa}$ where the random variables w_{sa} are all unit normal and independent.
- Again, this is clearly wrong....but let's go with it.
- Do you see why $r + \max_{a'} Q_{\theta^*}(s',a')$ will tend to over-estimate $r + \gamma \max_{a'} Q^*(s',a')$?
- Suppose there are five actions with roughly equal value, i.e.,
  $Q^*(s,a_1) = Q^*(s,a_2) = \cdots = Q^*(s,a_5).$
  If we add random noise to its, and THEN pick the largest, this will tend to over-estimate $\max_{a_1,\ldots,a_5} Q(s,a_i)$

- We can crystallize the issue by observing that our estimate of the value of the next state is

  $$V_{\text{target}} = r + \gamma Q_{\theta_t}(s', \arg\max_{a'} Q_{\theta_t}(s', a'))$$

- Solution: use two different networks!
- Ideally one might train to train $Q^1(\theta_t^1)$ and $Q^2(\theta_t^2)$ starting from independent random initial conditions and estimate the target as

  $$V_{\text{target}}^1 = r + \gamma Q_{\theta_t^1}(s', \arg\max_{a'} Q_{\theta_t^2}(s', a'))$$

- In practice: two time consuming. Usually, you fix some large K and you use $Q_{\theta_{t-K}}$ as your second network.
- Another possibility: Start with a random θ_0^2 for your second network and update

  $$\theta_{t+1}^2 = \tau\theta_{t+1}^1 + (1-\tau)\theta_t^2.$$

- People call this double Q-learning.
- Popular belief: target network strongly significantly improves performance.

- Worth reflecting on how RL is more difficult that supervised learning (e.g., image labeling).
- In supervised learning, you typically have labels.
  In RL, you use your method to generate your labels.
- In other words, your labels change as you run your method.
- This can lead to instability and divergence.
- Again, can think about interpreting the world through the lease of your beliefs.
- No good fix for this. Sometimes things work, sometimes they don’t.
- Example: https://www.youtube.com/watch?v=fv-oFPAqSZ4

- Let's talk about **Continuous Problems**. So your state space is a vector and your action space is a vector.
- No real problem in implementing the Q-value iteration

  $$\theta_{t+1} = \theta_t + \alpha_t(r + \gamma \max_{a'} Q_\theta(s', a') - Q_\theta(s, a)) \nabla_\theta Q_\theta(s, a)$$

  (of course, you have to fix an architecture first).
- What's the problem?
- Solution: use a neural network to compute the max.
- We will have two neural networks now. One neural networks will we referred to as $Q_{s_q}(s, a)$ (called the critic network) and will be update as above.

  The second neural network $\mu_{\theta_\mu}(s)$ (called the actor network) will decide which action to choose.
- How should we update θ_μ?
- Natural choice: $\theta_\mu(t+1) = \theta_\mu(t) + \beta(t) \nabla_{\theta_u} Q_{\theta_q}(s, \mu_{\theta_\mu}(s))$
- Can use chain rule: $\nabla_{\theta_\mu} Q_{\theta_q}(s, \mu_{\theta_\mu}(s)) = \nabla_a Q_{\theta_q}(s, a) \nabla_{\theta_\mu} \mu_{\theta_\mu}(s)$

Algorithm 1 DDPG algorithm

Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ.
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer R

Receive initial observation state s_1
for t = 1, T **do**
Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise
Execute action a_t and observe reward r_t and observe new state s_{t+1}
Store transition (s_t, a_t, r_t, s_{t+1}) in R
Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from R
Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$
Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$$

Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

end for
end for

The DDPG method
https://arxiv.org/pdf/1509.02971.pdf

- When it came out, DDPG was state of the art in a variety of robotics tasks: https://www.youtube.com/watch?v=tJBIqkC1wWM
- https://static-content.springer.com/esm/art%3A10.1038%2Fnature14236/MediaObjects/41586_2015_BFnature14236_MOESM124_ESM.mov
- Can still get OK results with it: https://www.youtube.com/watch?v=rd2pM43zzTo
- Many method since that (claim to) obtain better performance.
- Nice blog post: https://lilianweng.github.io/lil-log/2018/04/08/policy-gradient-algorithms.html