28: Intro to Deep Learning

Jerry Cain
June 2, 2021
Deep Learning
Innovations in deep learning

Deep learning (neural networks) is the core idea behind the current revolution in AI.

Errata:
- Checkers is the last solved game (from game theory, where perfect player outcomes can be fully predicted from any gameboard). [https://en.wikipedia.org/wiki/Solved_game](https://en.wikipedia.org/wiki/Solved_game)
Computers making art

The Next Rembrandt

A Neural Algorithm of Artistic Style
https://arxiv.org/abs/1508.06576
https://github.com/jcjohnson/neural-style

Google Deep Dream
https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Detecting skin cancer

**Deep learning**

**def Deep learning** is maximum likelihood estimation with neural networks.

**def A neural network** is (at its core) many logistic regression pieces stacked on top of each other.

\[
\begin{align*}
\[1,0, \ldots, 1\] & \quad \text{x, input} \\
\text{LOL} & \quad \text{Lots of Logistic (regressions)} \\
\hat{y}, \text{output} & \quad P(Y = 1|X = x) \\
> 0.5? & \quad \text{Yes. Predict 1}
\end{align*}
\]
Logistic Regression Model

Let's focus on the model up to $\hat{y}$. 

$$y = \arg\max_{y \in \{0,1\}} P(Y = 1|X)$$
Logistic Regression Model

\[ \hat{y} = P(Y = 1 | X) \]

\[ \hat{y} = \arg \max_{y \in \{0, 1\}} P(Y = y | X) \]

Let’s focus on the model up to \( \hat{y} \).
One neuron = One logistic regression
Biological basis for neural networks

A neuron

Your brain

(actually, probably someone else’s brain)

Neural network = many logistic regressions
## Digit recognition example

<table>
<thead>
<tr>
<th>Input image</th>
<th>Input feature vector</th>
<th>Output label</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image0.png" alt="Image of digit 0" /></td>
<td>( x^{(i)} = [0,0,0,0, \ldots, 1,0,0,1, \ldots, 0,0,1,0] )</td>
<td>( y^{(i)} = 0 )</td>
</tr>
<tr>
<td><img src="image1.png" alt="Image of digit 1" /></td>
<td>( x^{(i)} = [0,0,1,1, \ldots, 0,1,1,0, \ldots, 0,1,0,0] )</td>
<td>( y^{(i)} = 1 )</td>
</tr>
</tbody>
</table>

We make feature vectors from (digitized) pictures of numbers.
Logistic Regression

\[ \hat{y}, \text{output} \]

\[ P(Y = 1|X = x) \]

\( x, \text{input features} \)

(pixels, on/off)
Logistic Regression

\[ P(Y = 1|X = x) \]

\( x \), input features
(pixels, on/off)

\( \hat{y} \), output

\( > 0.5? \)

No.
Predict 0
Logistic Regression

\[ \hat{y}, \text{output} \]

\[ > 0.5? \]

Yes. Predict 1

\[ x, \text{input features} \]

\[ 1 \]

indicates logistic regression connection
Logistic Regression

$x$, input features

indicates logistic regression connection

$\hat{y}$, output

Yes. Predict 1

> 0.5?

What can we do to increase complexity of our model?
Take two big ideas from Logistic Regression

**Big idea #1**  \[ P(Y|X = x) \]
Model conditional probability \( \hat{y} \) of class label given input

**Big idea #2**  \[ \sigma(\theta^T x) \]
Non-linear transform of multiple values into one value, using parameter \( \theta \)

\( x \), input features

\( \hat{y} \), output

\( \sigma \) indicates logistic regression connection
Introducing: The Neural network

$x$, input features

$h$, hidden layer

$\hat{y}$, output

> 0.5? No. Predict 0

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Neural network

**Big idea #1** \( P(Y|X = x) \)

Model conditional probability \( \hat{y} \) of class label given input

\( x \), input features

\( h \), hidden layer

\( \hat{y} \), output

> 0.5? 

No. Predict 0
Feed neurons into other neurons

$\mathbf{x}$, input features

Hidden neuron

$\sigma$, hidden neuron

$\mathbf{y}$, output

Neuron = logistic regression

Big idea #2 $\sigma(\theta^T \mathbf{x})$

Non-linear transform of multiple values into one value, using parameter $\theta$

$> 0.5$?

No. Predict 0

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Feed neurons into other neurons

- Neuron = logistic regression
- Different parameters for every connection

$x$, input features
$h$, hidden layer
$\hat{y}$, output

$\sigma$, hidden layer

$\sigma$, hidden neuron

Another hidden neuron

$> 0.5$?

No. Predict 0

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Feed neurons into other neurons

- Neuron = logistic regression
- Different parameters for every connection
Feed neurons into other neurons

- Neuron = logistic regression
- Different parameters for every connection

\[ \hat{y}, \text{output} \]

\[ > 0.5? \]

\[ \text{No. Predict 0} \]

- \[ x, \text{input features} \]

- \[ |h| \text{ logistic regression connections} \]

- \[ |x| \cdot |h| \text{ parameters} \]

- \[ h, \text{hidden layer} \]
Feed neurons into other neurons

\[ x, \text{ input features} \]

\[ h, \text{ hidden layer} \]

\[ |h| \text{ logistic regression connections} \]

\[ |x| \cdot |h| \text{ parameters} \]

\[ \tilde{y}, \text{ output} \]

\[ 1 \text{ logistic regression connection} \]

- Neuron = logistic regression
- Different parameters for every connection
Why doesn’t a linear model introduce “complexity”?

Neural network:
1. for $j = 1, \ldots, |h|$: 
   $$h_j = \sigma \left( \theta_j^{(h)}^T x \right)$$
2. $$\hat{y} = \sigma \left( \theta^{(\hat{y})}^T h \right) = P(Y = 1 | X = x)$$

Linear network:
1. for $j = 1, \ldots, |h|$: 
   $$h_j = \theta_j^{(h)}^T x$$
2. $$\hat{y} = \sigma \left( \theta^{(\hat{y})}^T h \right) = P(Y = 1 | X = x)$$
Why doesn’t a linear model introduce “complexity”?

Neural network:
1. for $j = 1, ..., |h|$:  
   
   \[ h_j = \sigma \left( \theta_j^{(h)^T} x \right) \]

2.  
   \[ \hat{y} = \sigma \left( \theta^{(\hat{y})^T} h \right) = P(Y = 1 | X = x) \]

Linear network:
1. for $j = 1, ..., |h|$:  
   
   \[ h_j = \theta_j^{(h)^T} x \]

2.  
   \[ \hat{y} = \sigma \left( \theta^{(\hat{y})^T} h \right) = P(Y = 1 | X = x) \]

The linear model is effectively a single logistic regression with $|x|$ parameters.
Demonstration

http://scs.ryerson.ca/~aharley/vis/conv/
Neural networks

A neural network (like logistic regression) gets intelligence from its parameters $\theta$.

**Training**

- Learn parameters $\theta$
- Find $\theta_{MLE}$ that maximizes likelihood of training data (MLE)

**Testing/ Prediction**

For input feature vector $X = x$:
- Use parameters to compute $\hat{y} = P(Y = 1|X = x)$
- Classify instance as: $\begin{cases} 1 & \hat{y} > 0.5 \\ 0 & \text{otherwise} \end{cases}$
Neural networks

A neural network (like logistic regression) gets intelligence from its parameters $\theta$.

- Learn parameters $\theta$
- Find $\theta_{MLE}$ that maximizes likelihood of training data (MLE)

How do we learn the $|x| \cdot |h| + |h|$ parameters?

Gradient ascent + chain rule!
Training: Logistic Regression

1. Optimization problem:
\[
\theta_{MLE} = \arg \max_{\theta} \prod_{i=1}^{n} f(y^{(i)}|x^{(i)}, \theta) = \arg \max_{\theta} LL(\theta)
\]
\[
LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)})
\]
\[
\hat{y} = \sigma(\theta^T x^{(i)}) = P(Y = 1|X = x)
\]

2. Compute gradient
   Find $|x|$ parameters

3. Optimize
   initialize params
   repeat many times:
   compute gradient
   params += $\eta \times$ gradient
1. Optimization problem:

\[ \theta_{MLE} = \arg \max_{\theta} \prod_{i=1}^{n} f(y^{(i)} | x^{(i)}, \theta) = \arg \max_{\theta} LL(\theta) \]

2. Compute gradient

3. Optimize
1. Same output $\hat{y}$, same log conditional likelihood

$\theta_{MLE} = \arg \max_\theta \prod_{i=1}^{n} f(y^{(i)} | x^{(i)}, \theta) = \arg \max_\theta LL(\theta)$

$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)} | X = x^{(i)}, \theta)$

$= \prod_{i=1}^{n} (\hat{y}^{(i)})^{y^{(i)}} (1 - \hat{y}^{(i)})^{1-y^{(i)}}$

$LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)})$

Binary class labels: $Y \in \{0, 1\}$

For $j = 1, \ldots, |h|$: $h_j = \sigma(\theta_j^{(h)^T} x)$

$\hat{y} = \sigma(\theta^{(h)^T} h) = P(Y = 1 | X = x)$
(model is a little more complicated)

\[
\theta_{\text{MLE}} = \arg \max_{\theta} \prod_{i=1}^{n} f(y^{(i)} \mid x^{(i)}, \theta) = \arg \max_{\theta} LL(\theta)
\]

\[
LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)})
\]

To optimize for log conditional likelihood, we now need to find:

\[|h| \cdot |x| + |h| \quad \text{parameters}\]

\[
h_j = \sigma \left( \theta_j^{(h)^T} x \right) \quad \text{dimension} \ |x|
\]

\[
\hat{y} = \sigma \left( \theta^{(\odot)}^T h \right) = P(Y = 1 \mid X = x) \quad \text{dimension} \ |h|
\]
2. **Compute gradient**

1. **Optimization problem**:

   \[ \theta_{MLE} = \arg \max_{\theta} \prod_{i=1}^{n} f(y^{(i)} | x^{(i)}, \theta) = \arg \max_{\theta} LL(\theta) \]

   \[ LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \]

   \[ h_j = \sigma \left( \theta_j^{(h)^T} x \right) \quad \text{for } j = 1, \ldots, |h| \]

   \[ \hat{y} = \sigma \left( \theta^{(y)^T} h \right) \]

2. **Compute gradient**

   Take gradient with respect to all \( \theta \) parameters

3. **Optimize**

   **Calculus refresher #1:** Derivative(sum) = sum(derivative)

   **Calculus refresher #2:** Chain rule ⭐⭐⭐
3. Optimize

1. Optimization problem:
   \[ \theta_{MLE} = \arg \max_{\theta} \prod_{i=1}^{n} f(y^{(i)} | x^{(i)}, \theta) = \arg \max_{\theta} LL(\theta) \]
   \[ LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)}) \]
   \[ h_j = \sigma (\theta_j^{(h)} x) \quad \text{for} \ j = 1, ..., |h| \quad \hat{y} = \sigma (\theta (\hat{y})^T h) \]

2. Compute gradient
   
   Take gradient with respect to all \( \theta \) parameters

3. Optimize
   
   initialize params
   repeat many times:
   compute gradient
   params += \( \eta \) * gradient
Training a neural net

1. Optimization problem:

\[
\theta_{MLE} = \arg \max_{\theta} \prod_{i=1}^{n} f(y^{(i)}|x^{(i)}, \theta) = \arg \max_{\theta} LL(\theta)
\]

\[
LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)})
\]

2. Compute gradient

3. Optimize

\[
\theta_{\text{new}} = \theta_{\text{old}} + \eta \times \text{gradient}
\]

Wait, did we just skip something difficult?

initialize params
repeat many times:
    compute gradient
    params += \eta \times \text{gradient}
2. Compute gradient via backpropagation

1. Optimization problem:
   \[ \theta_{MLE} = \arg \max_{\theta} \prod_{i=1}^{n} f(y^{(i)}|x^{(i)}, \theta) = \arg \max_{\theta} LL(\theta) \]
   \[ LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)}) \]

2. Compute gradient
   \[ h_j = \sigma \left( \left[ \theta_j^{(h)} \right]^T x \right) \text{ for } j = 1, \ldots, |h| \]
   \[ \hat{y} = \sigma \left( \left[ \theta^{(y)} \right]^T h \right) \]
   Take gradient with respect to all \( \theta \) parameters

3. Optimize

Learn the tricks behind backpropagation in CS229, CS231N, CS224N, etc.
Beyond the basics
Shared weights?

It turns out if you want to force some of your weights to be shared over different neurons, the math isn’t much harder.

**Convolution** is an example of such weight-sharing and is used a lot for vision (Convolutional Neural Networks, CNN).
Neural networks with multiple layers
Neurons learn features of the dataset

Neurons in later layers will respond strongly to high-level features of your **training data**.

If your training data is faces, you will get lots of face neurons.

If your training data is all of YouTube...

...you get a cat neuron.

Le, et al., Building high-level features using large-scale unsupervised learning. ICML 2012
Hire the smartest people in the world

Invent cat detector
Multiple outputs?

**Softmax** is a generalization of the sigmoid function.

sigmoid($z$): value in range [0, 1]  
$z \in \mathbb{R}$:  
$$P(Y = 1|X = x) = \sigma(z)$$
(equivalent: Bernoulli $p$)

softmax($z$): $k$-dimensional values in range [0,1] that add up to 1  
$z \in \mathbb{R}^k$:  
$$P(Y = i|X = x) = \text{softmax}(z)_i$$
(equivalent: Multinomial $p_1, ..., p_k$)
## Softmax test metric: Top-5 error

| $Y = y$ | $P(Y = y|X = x)$ |
|--------|-----------------|
| 5      | 0.14            |
| 8      | 0.13            |
| 7      | 0.12            |
| 2      | 0.10            |
| 9      | 0.10            |
| 4      | 0.09            |
| 1      | 0.09            |
| 0      | 0.09            |
| 6      | 0.08            |
| 3      | 0.05            |

**Top-5 classification error**

What % of datapoints did *not* have the correct class label in the top-5 predictions?
ImageNet classification

22,000 categories

14,000,000 images

Hand-engineered features (SIFT, HOG, LBP), Spatial pyramid, SparseCoding/Compression

smoothhound, smoothhound shark, Mustelus mustelus American smooth dogfish, Mustelus canis Florida smoothhound, Mustelus norrisi whitetip shark, reef whitetip shark, Triaenodon obesus Atlantic spiny dogfish, Squalus acanthias Pacific spiny dogfish, Squalus suckleyi hammerhead, hammerhead shark smooth hammerhead, Sphyrna zygaena smalleye hammerhead, Sphyrna tudes shovelhead, bonnethead, bonnet shark, Sphyrna angel shark, angelfish, Squatina squatina, monkfish electric ray, crampfish, numbfish, torpedo smalltooth sawfish, Pristis pectinatus guitarfish roughtail stingray, Dasyatis centroura butterfly ray eagle ray spotted eagle ray, spotted ray, Aetobatus narinari cow-nose ray, cow-nosed ray, Rhinoptera bonasus manta, manta ray, devilfish

Atlantic manta, Manta birostris devil ray, Mobula hypostoma grey skate, gray skate, Raja batis little skate, Raja erinacea...

Le, et al., Building high-level features using large-scale unsupervised learning. ICML 2012
ImageNet classification challenge

22,000 categories
14,000,000 images

1000 categories
200,000 images in train set
200,000 images in test set

Hand-engineered features
(SIFT, HOG, LBP),
Spatial pyramid,
SparseCoding/Compression

smooth hammerhead, Sphyrna zygaena
smalleye hammerhead, Sphyrna tudes
shovelhead, bonnethead, bonnet shark, Sphyra tiburo
angel shark, angelfish, Squatina squatina, monkfish
electric ray, crampfish, numbfish, torpedo
smalltooth sawfish, Pristis pectinatus
guitarfish
rougthail stingray, Dasyatis centoura
butterfly ray
eagle ray
spotted eagle ray, spotted ray, Aetobatus narinari
cownose ray, cow-nosed ray, Rhinoptera bonasus
manta, manta ray, devilfish
Atlantic manta, Manta birostris
devil ray, Mobula hypostoma
grey skate, gray skate, Raja batis
little skate, Raja erinacea
...

Le, et al., Building high-level features using large-scale unsupervised learning. ICML 2012
ImageNet challenge: Top-5 classification error

99.5%

Random guess

\[ P(\text{true class label not in 5 guesses}) = \frac{\binom{999}{5}}{\binom{1000}{5}} = \frac{995}{1000} \]
## ImageNet challenge: Top-5 classification error

<table>
<thead>
<tr>
<th>Method</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random guess</td>
<td>99.5%</td>
</tr>
<tr>
<td>Pre-Neural Networks</td>
<td>25.8%</td>
</tr>
<tr>
<td>Humans (2014)</td>
<td>5.1%</td>
</tr>
<tr>
<td>GoogLeNet (2015)</td>
<td>16.4%</td>
</tr>
</tbody>
</table>

*lower is better*

---

Russakovsky et al., ImageNet Large Scale Visual Recognition Challenge. IJCV 2015
Szegedy et al., Going Deeper With Convolutions. CVPR 2015
## ImageNet challenge: Top-5 classification error

<table>
<thead>
<tr>
<th></th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random guess</td>
<td>99.5%</td>
</tr>
<tr>
<td>Pre-Neural Networks</td>
<td>25.8%</td>
</tr>
<tr>
<td>Humans (2014)</td>
<td>5.1%</td>
</tr>
<tr>
<td>GoogLe Net (2015)</td>
<td>16.4%</td>
</tr>
<tr>
<td>SENet (2017)</td>
<td>2.25%</td>
</tr>
</tbody>
</table>

(lower is better)

Russakovsky et al., ImageNet Large Scale Visual Recognition Challenge. IJCV 2015  
Szegedy et al., Going Deeper With Convolutions. CVPR 2015  
GoogLeNet (2015)

1 Trillion Artificial Neurons
(btw human brains have 1 billion neurons)

Multiple, Multi class output
22 layers deep!

Szegedy et al., Going Deeper With Convolutions. CVPR 2015
Speeding up gradient descent minimizes loss (a function of prediction error)

initialize $\theta_j = 0$ for $0 \leq j \leq m$
repeat many times:
gradient[j] = 0 for $0 \leq j \leq m$

for each training example $(x, y)$:
for each $0 \leq j \leq m$:
compute gradient

$\theta_j \leftarrow \eta \times \text{gradient}[j]$ for all $0 \leq j \leq m$

1. What if we have 1,200,000 images in our training set?
2. How can we speed up the update?

Our batch gradient descent (over the entire training set) will be slow + expensive.

1. Use stochastic gradient descent (randomly select training examples with replacement).
2. Momentum update (Incorporate “acceleration” or “deceleration” of gradient updates so far)
Good ML = Generalization

**Overfitting**
Fitting the training data too well, such that we lose generality of model for predicting new data

**Dropout**
During training, randomly leave out some neurons each training step.
It will make your network more robust.
Making decisions?

Not everything is classification.

Deep Reinforcement Learning

Instead of having the output of a model be a probability, you make output an expectation.

http://cs.stanford.edu/people/karpathy/convnetjs/demo/rldemo.html
Deep Reinforcement Learning

http://cs.stanford.edu/people/karpathy/convnetjs/demo/rldemo.html

Deep Mind Atari Games
Score compared to best human