Deep Neural Networks

Deep Learning

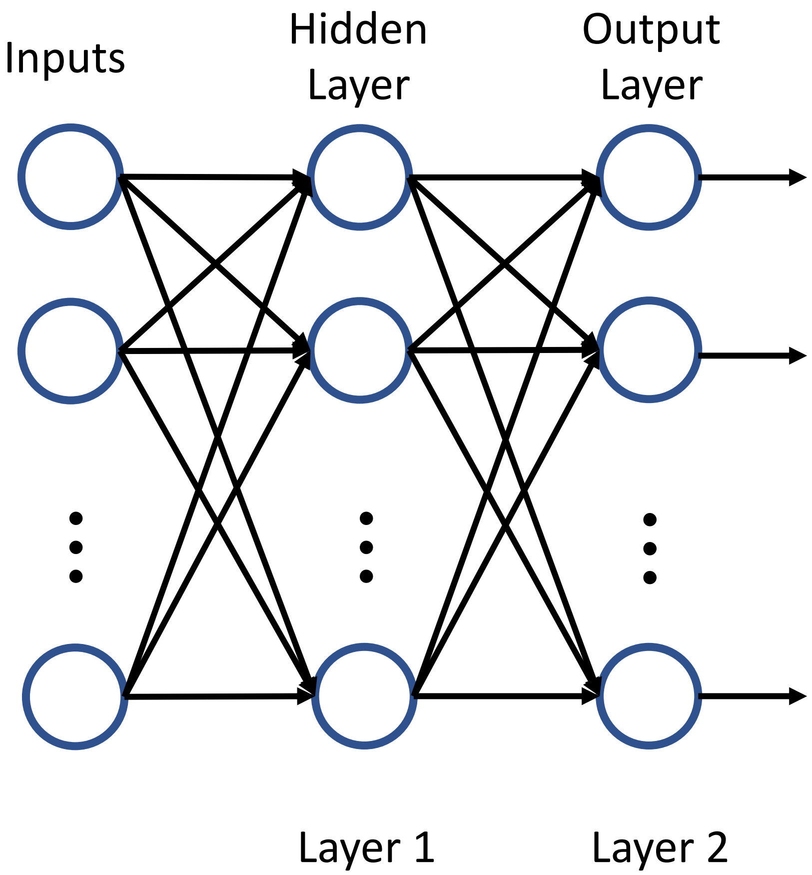
Brad Quinton, Scott Chin

Learning Objectives

- What is a “Deep” Neural Network?
- Intuition of what hidden layers are doing
- Why use deep (vs shallow) networks?
- Backpropagation Through Softmax and Categorical Cross-Entropy Loss

Brad Quinton, Scott Chin

Recap: 2-Layer Neural Network



Output of one activation unit is calculated by:

$$a_i^{[l]} = g\left(\sum_{j=1}^{n_h^{[l-1]}} w_{i,j}^{[l]} a_j^{[l-1]} + b_i^{[l]}\right) = g\left(W_i^{[l]} \cdot A^{[l-1]} + B_i^{[l]}\right)$$

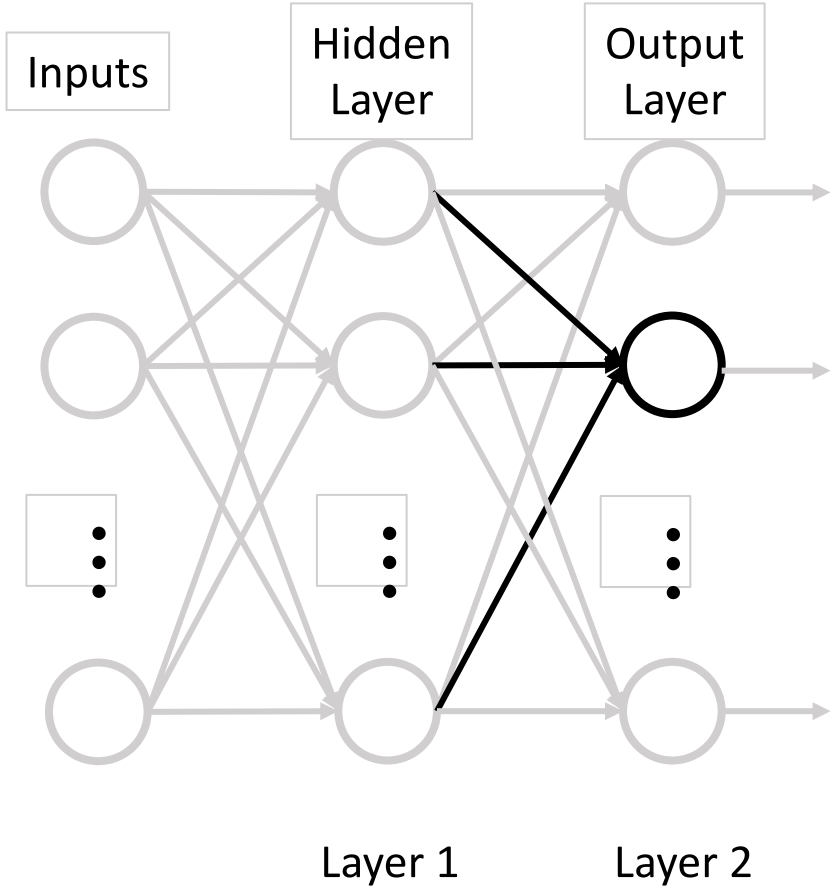
Where

- $g()$ is the activation function
- $w_{i,j}^{[l]}$ is weight parameter of input $a_j^{[l-1]}$ going to unit $a_i^{[l]}$
- $n_h^{[l]}$ is the number of units in some layer l

Also note that

- $\hat{y}_i = a_i^{[L]}$ where L is the number of layers
- $a_i^{[0]} = x_i$

Recap: 2-Layer Neural Network



Output of one activation unit is calculated by:

$$a_i^{[l]} = g\left(\sum_{j=1}^{n_h^{[l-1]}} w_{i,j}^{[l]} a_j^{[l-1]} + b_i^{[l]}\right) = g\left(W_i^{[l]} \cdot A^{[l-1]} + B_i^{[l]}\right)$$

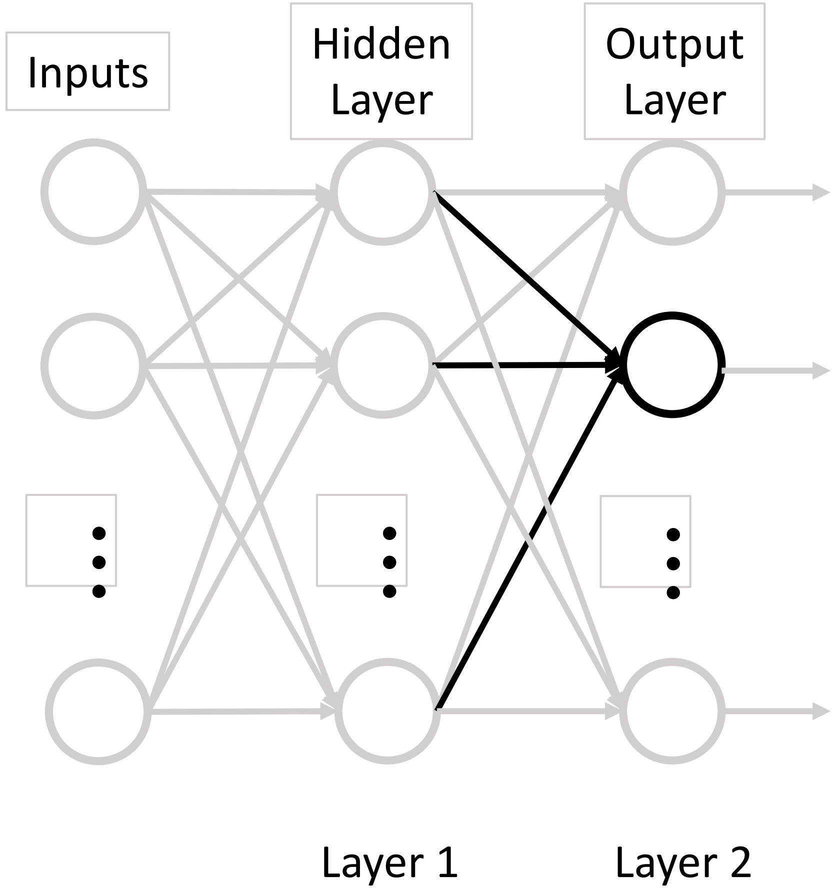
Where

- $g()$ is the activation function
- $w_{i,j}^{[l]}$ is weight parameter of input $a_j^{[l-1]}$ going to unit $a_i^{[l]}$
- $n_h^{[l]}$ is the number of units in some layer l

In this example: $l = 2, i = 1$

Brad Quinton, Scott Chin

Recap: 2-Layer Neural Network



Output of one activation unit is calculated by:

$$a_i^{[l]} = g\left(\sum_{j=1}^{n_h^{[l-1]}} w_{i,j}^{[l]} a_j^{[l-1]} + b_i^{[l]}\right) = g\left(W_i^{[l]} \cdot A^{[l-1]} + B_i^{[l]}\right)$$

Where

- $g()$ is the activation function
- $w_{i,j}^{[l]}$ is weight parameter of input $a_j^{[l-1]}$ going to unit $a_i^{[l]}$
- $n_h^{[l]}$ is the number of units in some layer l

In this example: $l = 2, i = 1$

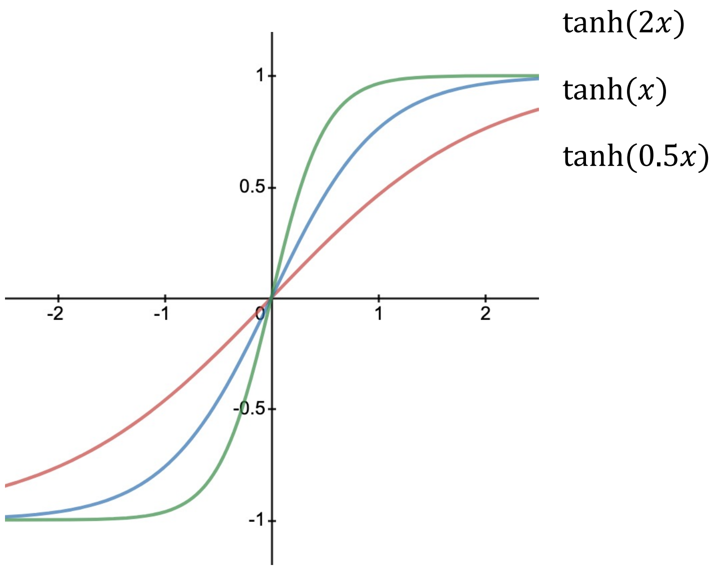
Number of parameters for ONE unit:

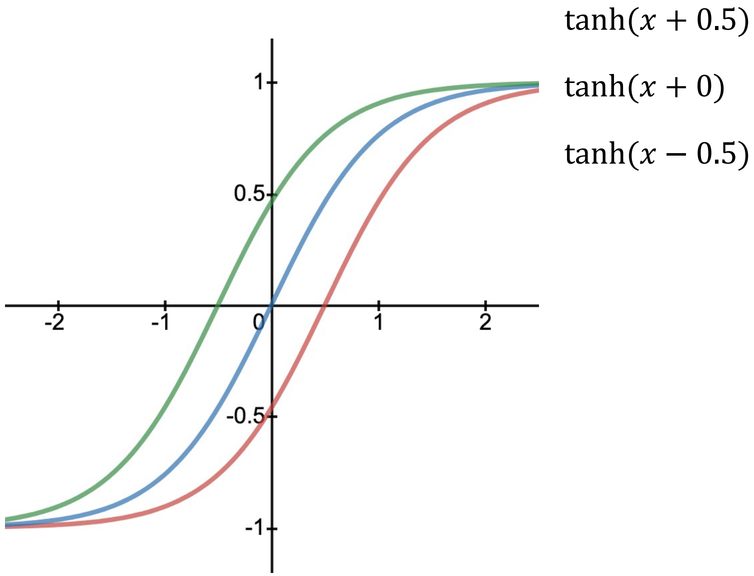
- $n_h^{[l-1]} + 1$

weights bias

A note about the bias

- Just another parameter that helps shape the unit's function
- Same idea as to how a straight line of the form $y = mx + b$ needs the b term to shift the function. Without it, the only lines you could model must pass through the origin.

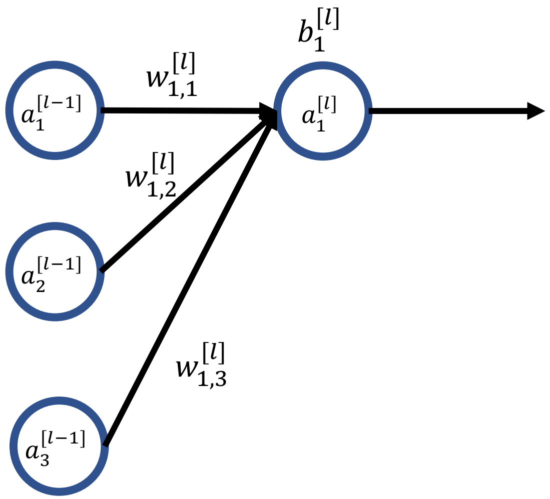


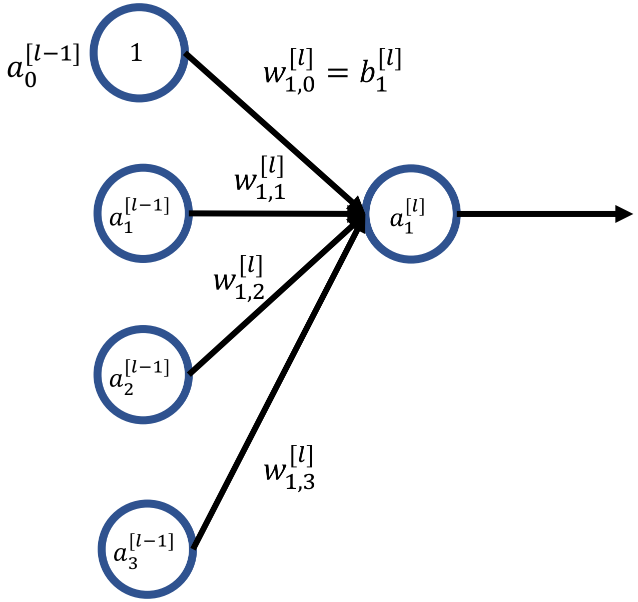


Brad Quinton, Scott Chin

A note about the bias

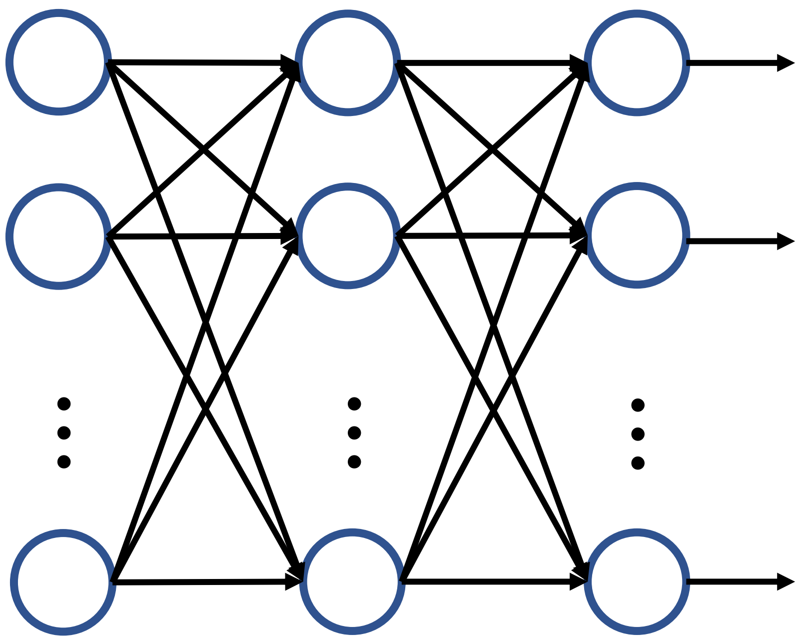
- Some literature includes the bias term in the weights so you don't need to talk about it separately
- To do this, they model an extra node in each layer that outputs a constant one. This node is then weighted by the bias





Brad Quinton, Scott Chin

Back to our 2-Layer Neural Network



Output of one activation unit is calculated by:

$$a_i^{[l]} = g\left(\sum_{j=1}^{n_h^{[l-1]}} w_{i,j}^{[l]} a_j^{[l-1]} + b_i^{[l]}\right) = g\left(W_i^{[l]} \cdot a^{[l-1]} + b_i^{[l]}\right)$$

Where

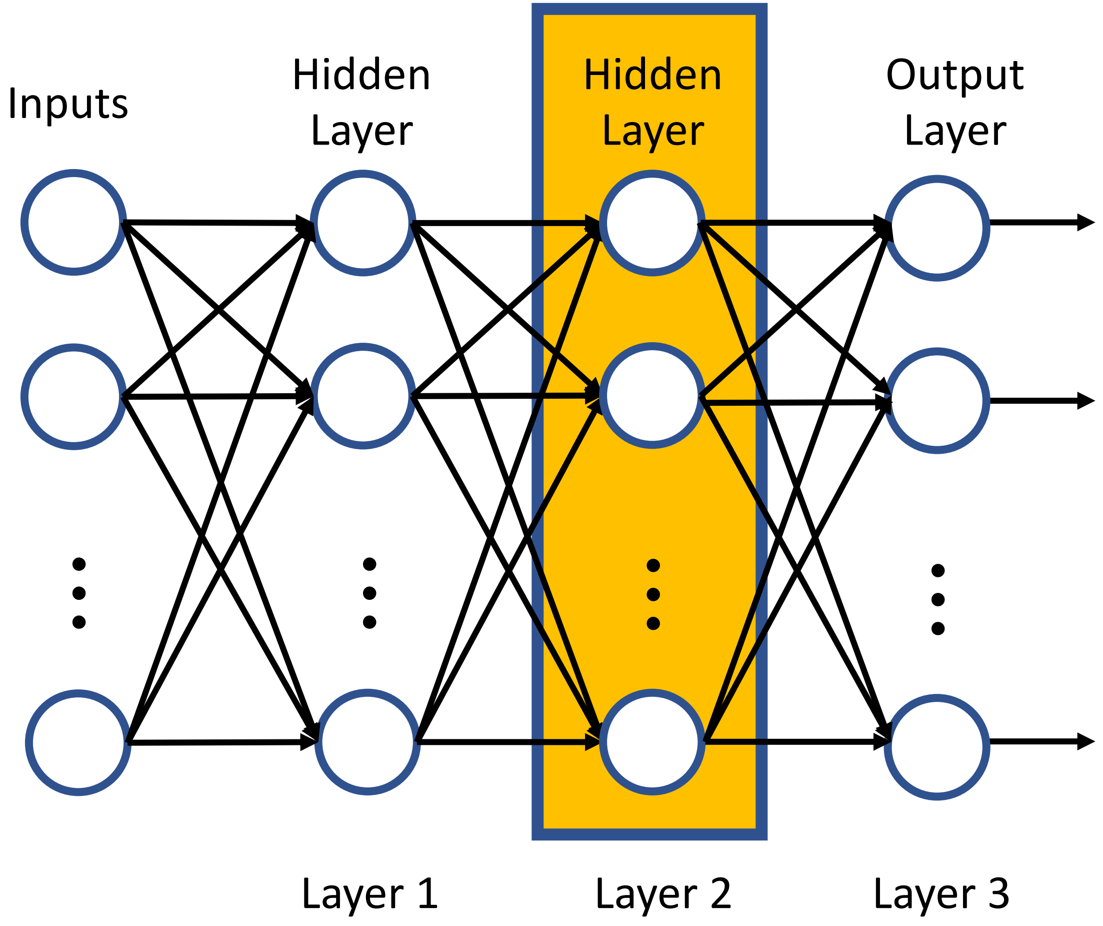
- $g()$ is the activation function
- $w_{i,j}^{[l]}$ is weight parameter of input $a_j^{[l-1]}$ going to unit $a_i^{[l]}$
- $n_h^{[l]}$ is the number of units in some layer l

Also note that

- $\hat{y}_i = a_i^{[L]}$ where L is the number of layers
- $a_i^{[0]} = x_i$

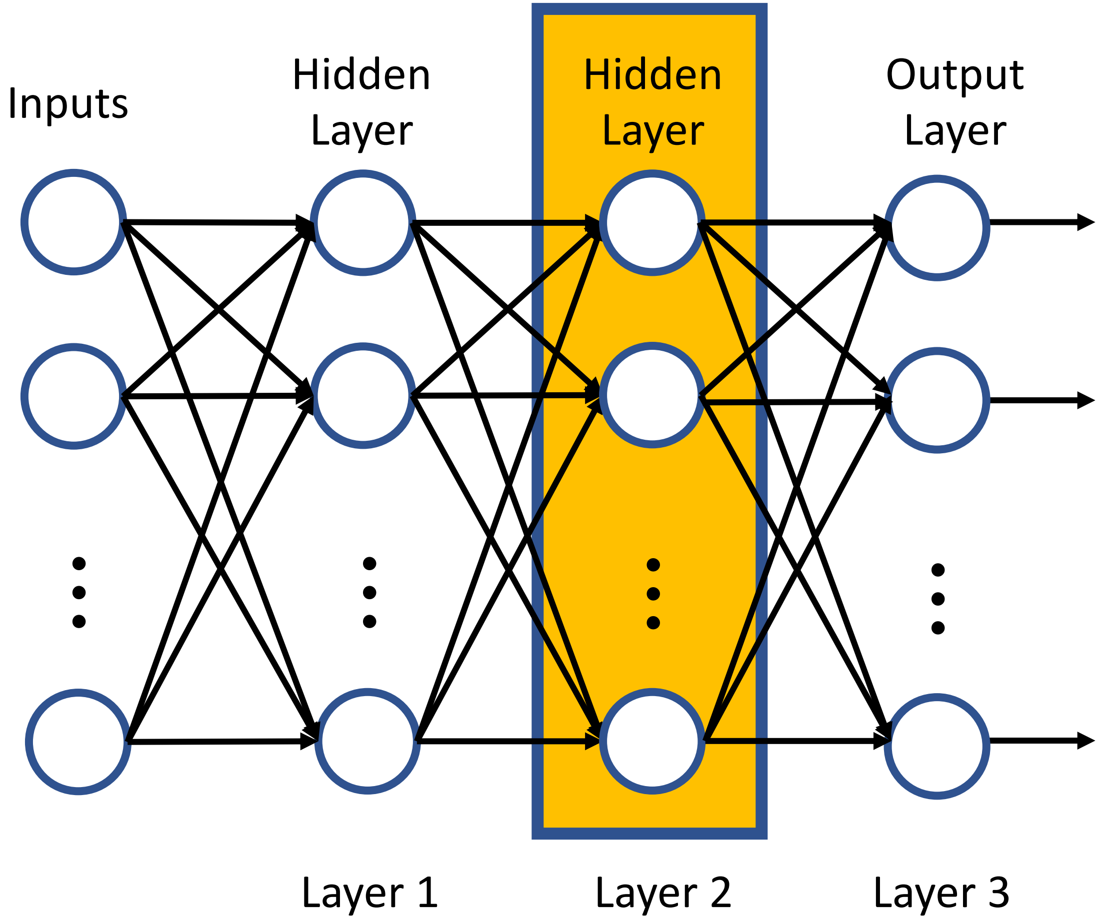
Brad Quinton, Scott Chin

Add another layer: 3-Layer Neural Network



Brad Quinton, Scott Chin

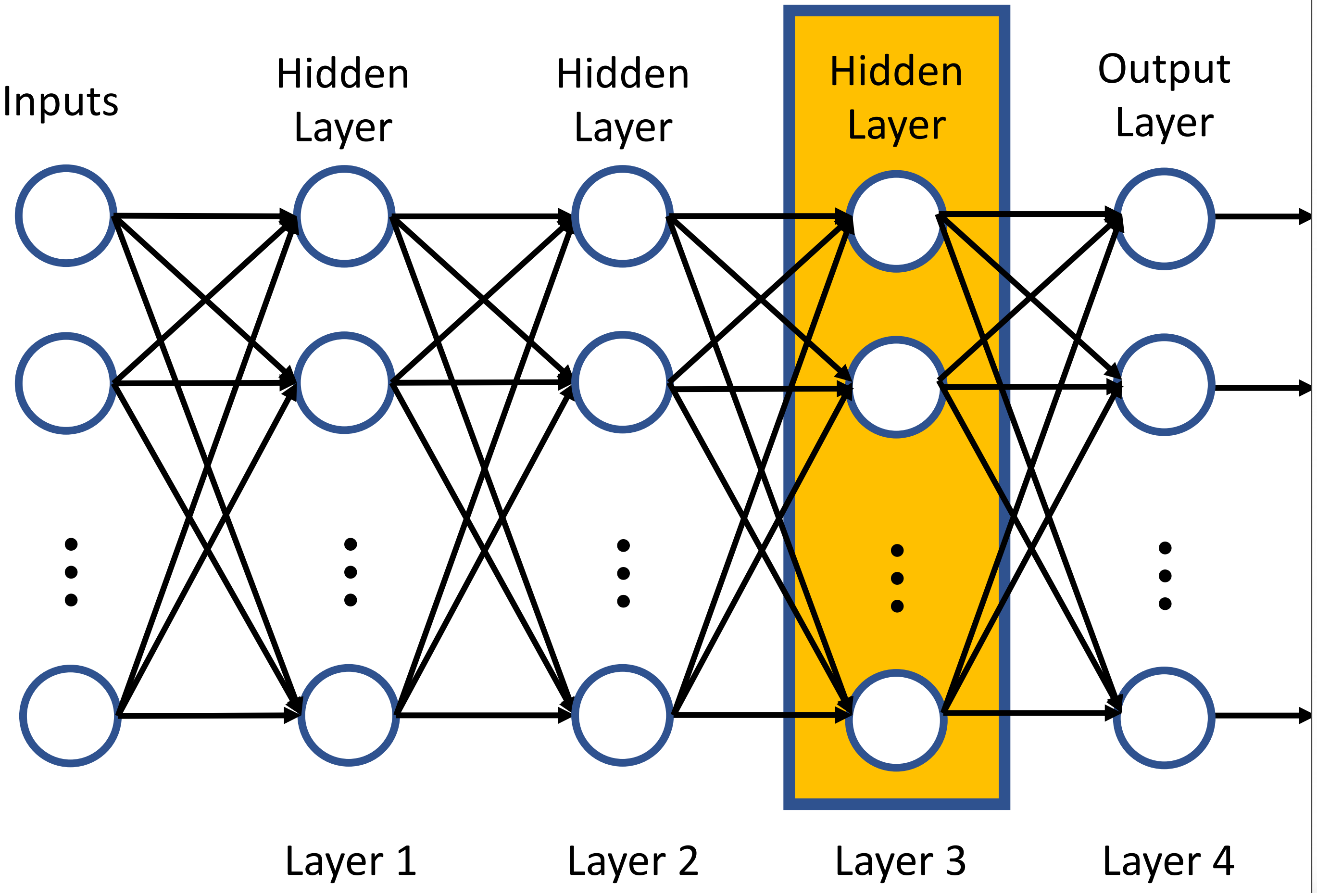
Add another layer: 3-Layer Neural Network



Deep Learning!

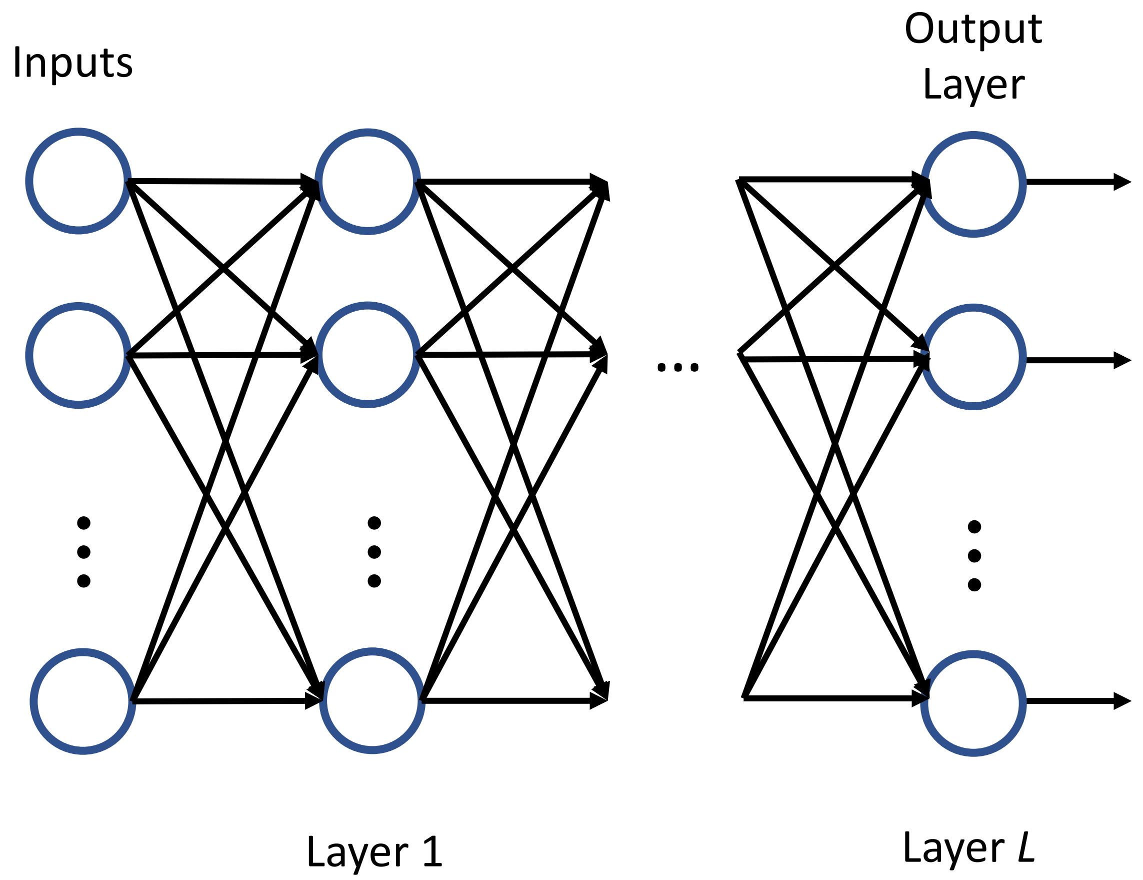
Brad Quinton, Scott Chin

Deep Learning: 4-Layer Neural Network



Brad Quinton, Scott Chin

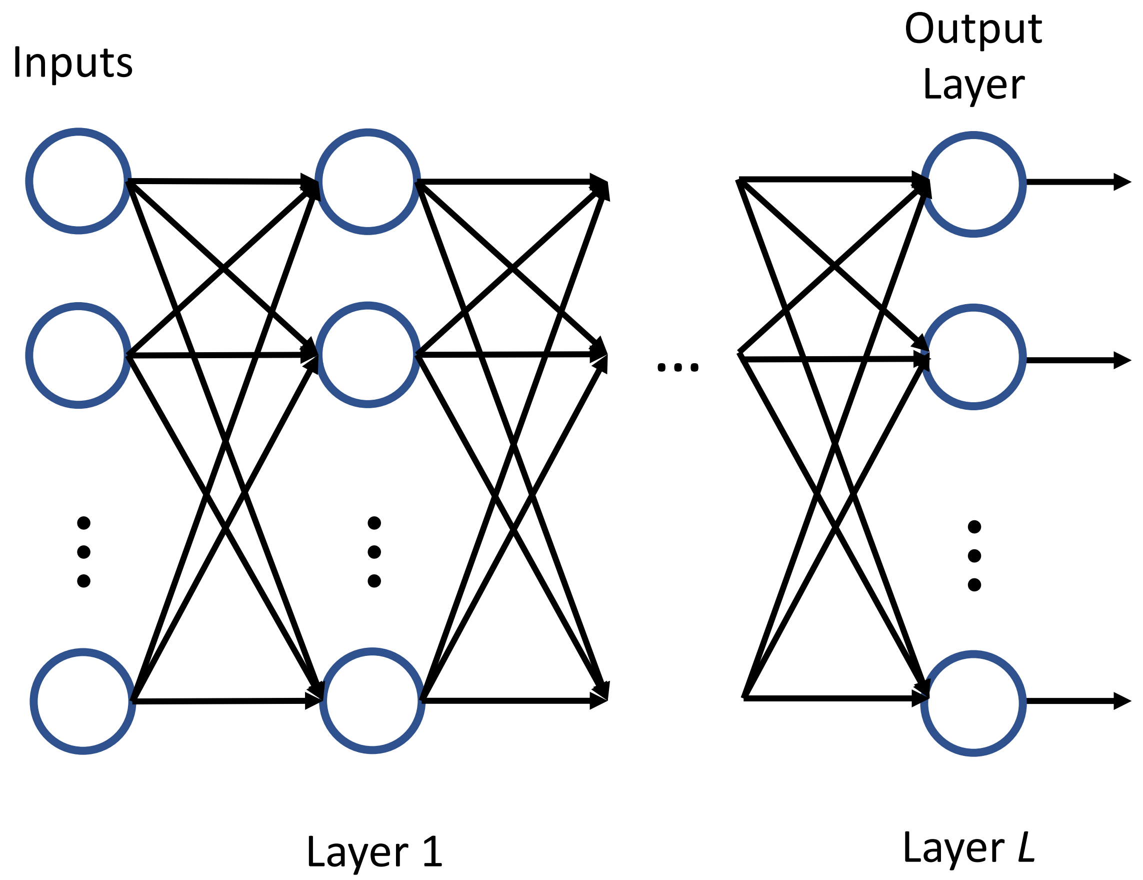
A Few General Points



- Count layers that have parameters
  - Input layer does not have parameters
- Called Fully Connected (FC), or Dense layers
  - Each input “connects” to each node
- Each FC layer can have different number of units (a.k.a. neurons)
- Number of Parameters per FC layer:
  - Weights: ? ? ?
  - Biases: ? ? ?

Brad Quinton, Scott Chin

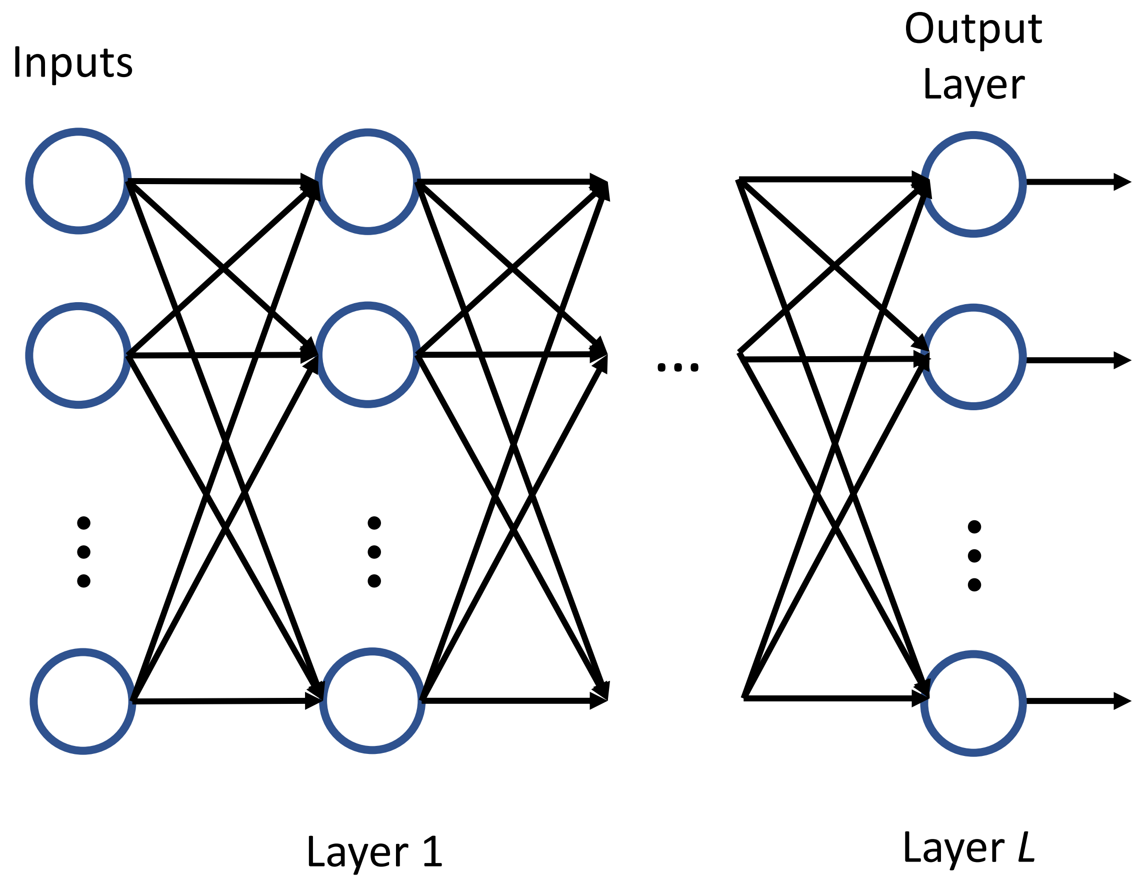
In General



- Count layers that have parameters
  - Input layer does not have parameters
- Called Fully Connected (FC), or Dense layers
  - Each input "connects" to each node
- Each FC layer can have different number of units (a.k.a. neurons)
- Number of Parameters per FC layer:
  - Weights: $n^{[l-1]} * n^{[l]}$
  - Biases: $n^{[l]}$

Brad Quinton, Scott Chin

Layers and Vectorized Forward Propagation



- Arranged in layers for vectorized computation.

```
X = # A single vector of features
A1 = np.tanh(np.dot(W1, X)+B1)
A2 = np.tanh(np.dot(W2, A1)+B2)
...
```

- Due to arrangement for vectorized computing, every unit in the same layer uses the same activation function. But this isn't fundamentally required.

Brad Quinton, Scott Chin

Other Layer Types

- Convolution
- Pooling
- Recurrent
- Normalization
- Regularization

“Classic” Neural Network Types

- You may also see fully-connected neural networks referred to as Multilevel Perceptrons (MLP)
- Later we will look at Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs)

Intuition of What Hidden Layers are doing

Brad Quinton, Scott Chin

Intuition of What Hidden Layers Are Doing

Some Interpretations that I have found useful:

- Increase capacity of the approximation function
- Feature Space Transform
- Automatic Feature Engineering
- Feature composition (or latent variables)

Brad Quinton, Scott Chin

Increase capacity of the approximation function

A neural network with one hidden layer provides the following mapping function:

$$Y(X) = \sigma\left(W^{[2]}\tanh\left(W^{[1]}X + B^{[1]}\right) + B^{[2]}\right)$$

This is a **class** of functions and each member function of this class is realized by a specific set of values for the parameters

Increase capacity of the approximation function

A neural network with one hidden layer provides the following mapping function:

$$Y(X) = \sigma\left(W^{[2]}\tanh\left(W^{[1]}X + B^{[1]}\right) + B^{[2]}\right)$$

This is a **class** of functions and each member function of this class is realized by a specific set of values for the parameters

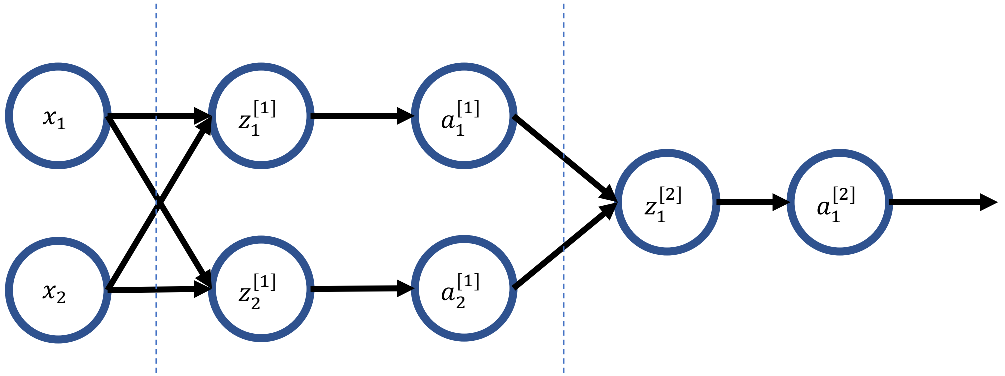
For a neural network with two hidden layers:

$$Y(X) = \sigma\left(W^{[3]}\tanh\left(W^{[2]}\tanh\left(W^{[1]}X + B^{[1]}\right) + B^{[2]}\right) + B^{[3]}\right)$$

Brad Quinton, Scott Chin

Feature Space Transformation

- Consider the following trained Neural Network and training data

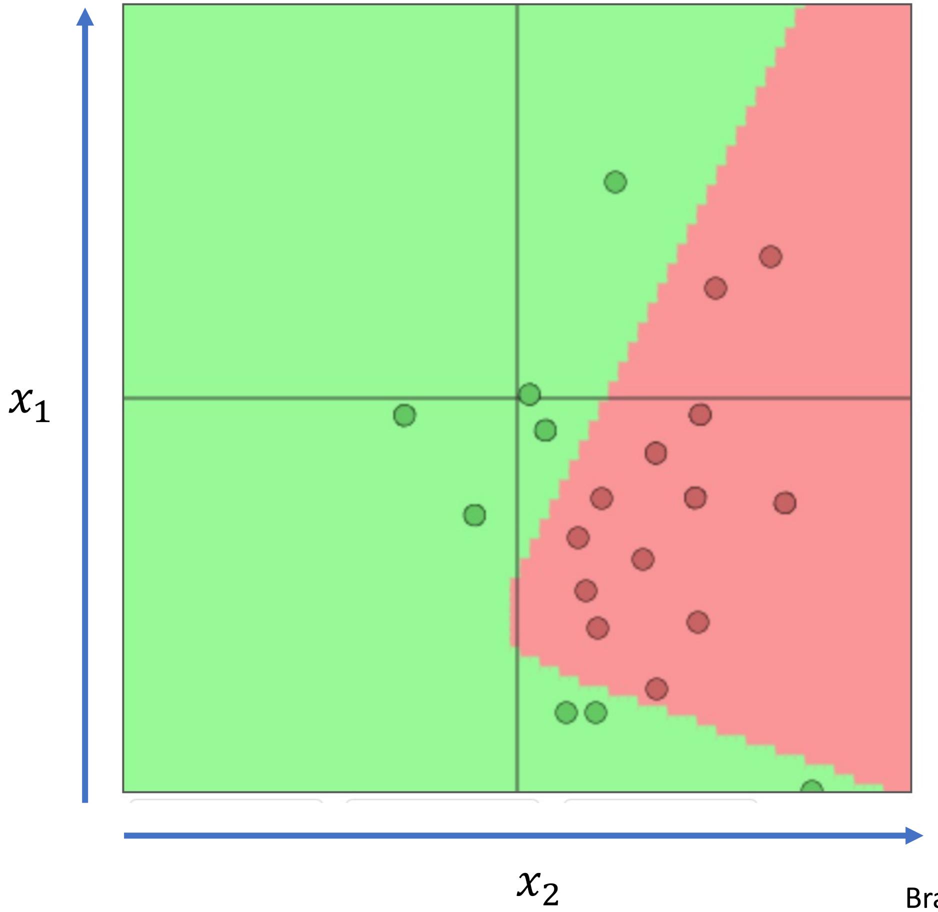


tanh() Layer

$$z_i^{[1]} = W^{[1]}x + b_i^{[1]}$$
$$a_i^{[1]} = \tanh\left(z_i^{[1]}\right)$$

sigmoid Layer

$$z_1^{[2]} = W^{[2]}a^{[1]} + b^{[2]}$$
$$a_1^{[2]} = \sigma\left(z_1^{[2]}\right)$$

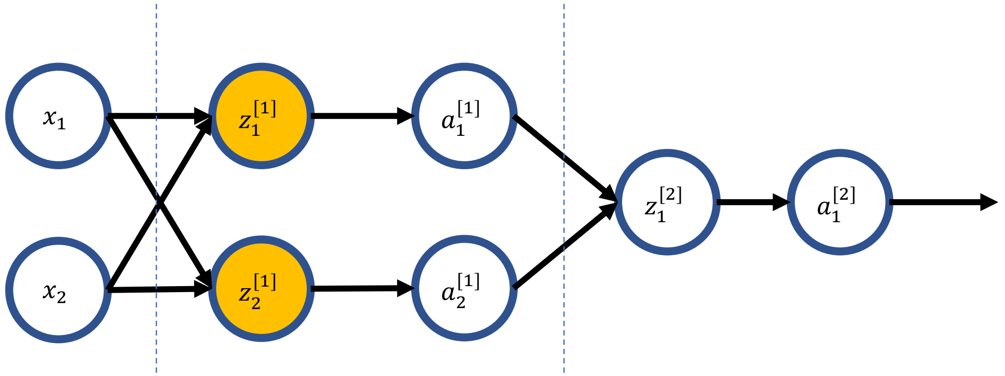


https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html

Brad Quinton, Scott Chin

Feature Space Transformation

- Instead of plotting x on axis, plot $z^{[1]}$ to see the linear transform



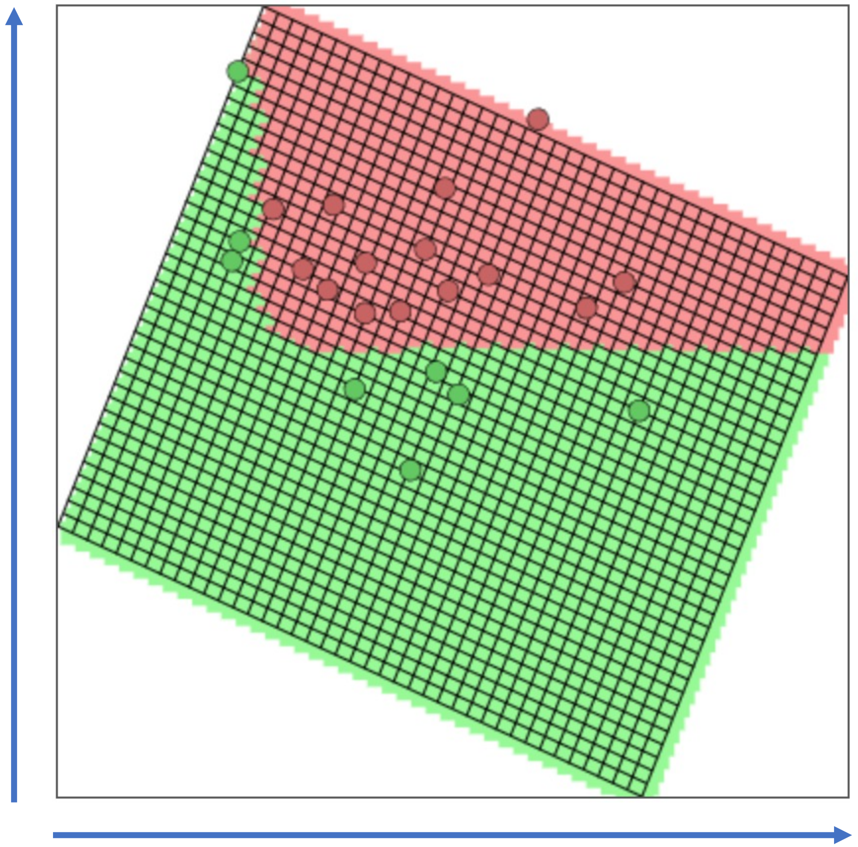
tanh() Layer

sigmoid Layer

$$z_i^{[1]} = W^{[1]}x + b_i^{[1]}$$
$$a_i^{[1]} = \tanh\left(z_i^{[1]}\right)$$

$$z_1^{[2]} = W^{[2]}a^{[1]} + b^{[2]}$$
$$a_1^{[2]} = \sigma\left(z_1^{[2]}\right)$$

Note: We use 2 units in the hidden layer for easy visualization.
With 3 units, we would be looking at a 3D space, etc.



$z_1^{[1]}$

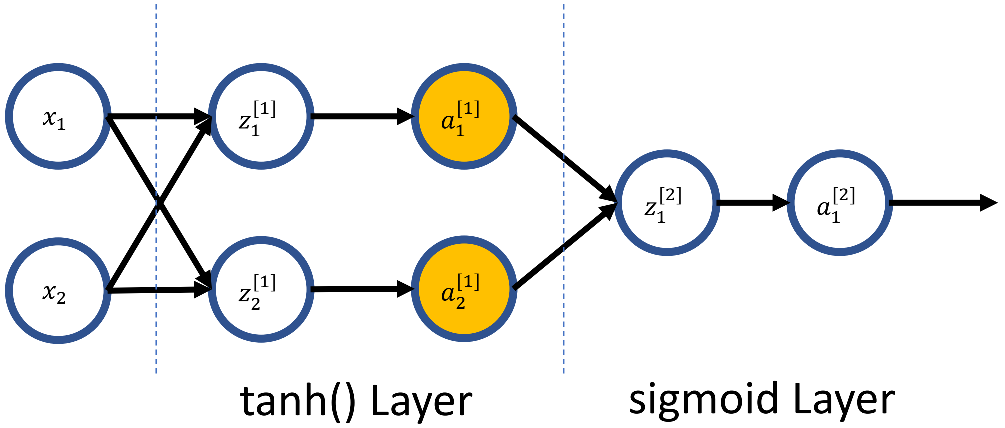
$z_2^{[1]}$

https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html

Brad Quinton, Scott Chin

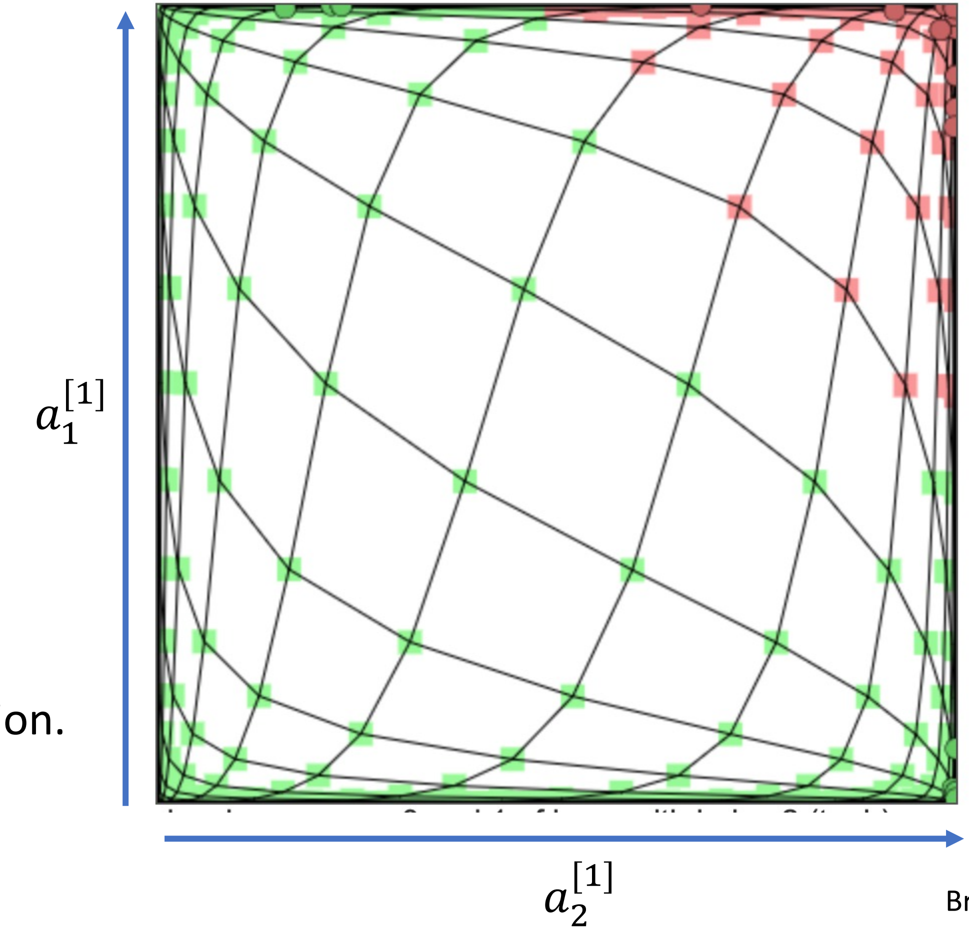
Feature Space Transformation

- Now plot $a^{[1]}$ to see the subsequent tanh() transform



$$z_i^{[1]} = W^{[1]}x + b_i^{[1]}$$
$$a_i^{[1]} = \tanh\left(z_i^{[1]}\right)$$

$$z_1^{[2]} = W^{[2]}a^{[1]} + b^{[2]}$$
$$a_1^{[2]} = \sigma\left(z_1^{[2]}\right)$$



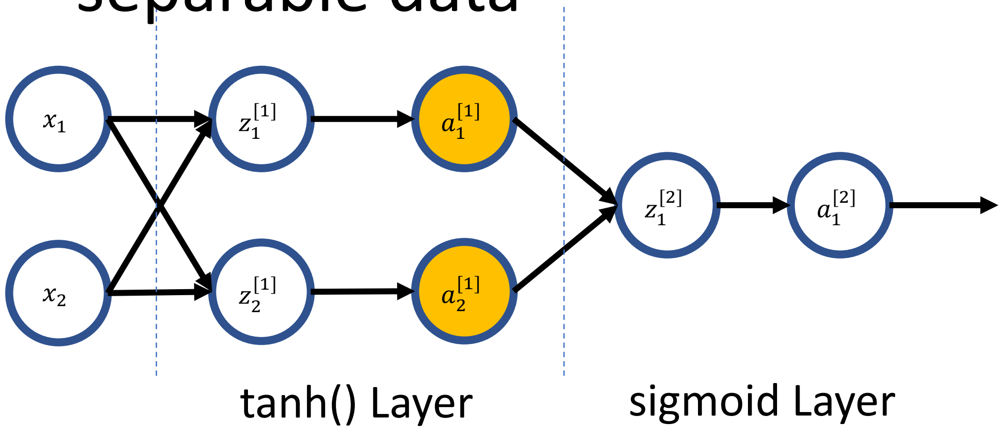
Note: We use 2 units in the hidden layer for easy visualization.
With 3 units, we would be looking at a 3D space, etc.

https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html

Brad Quinton, Scott Chin

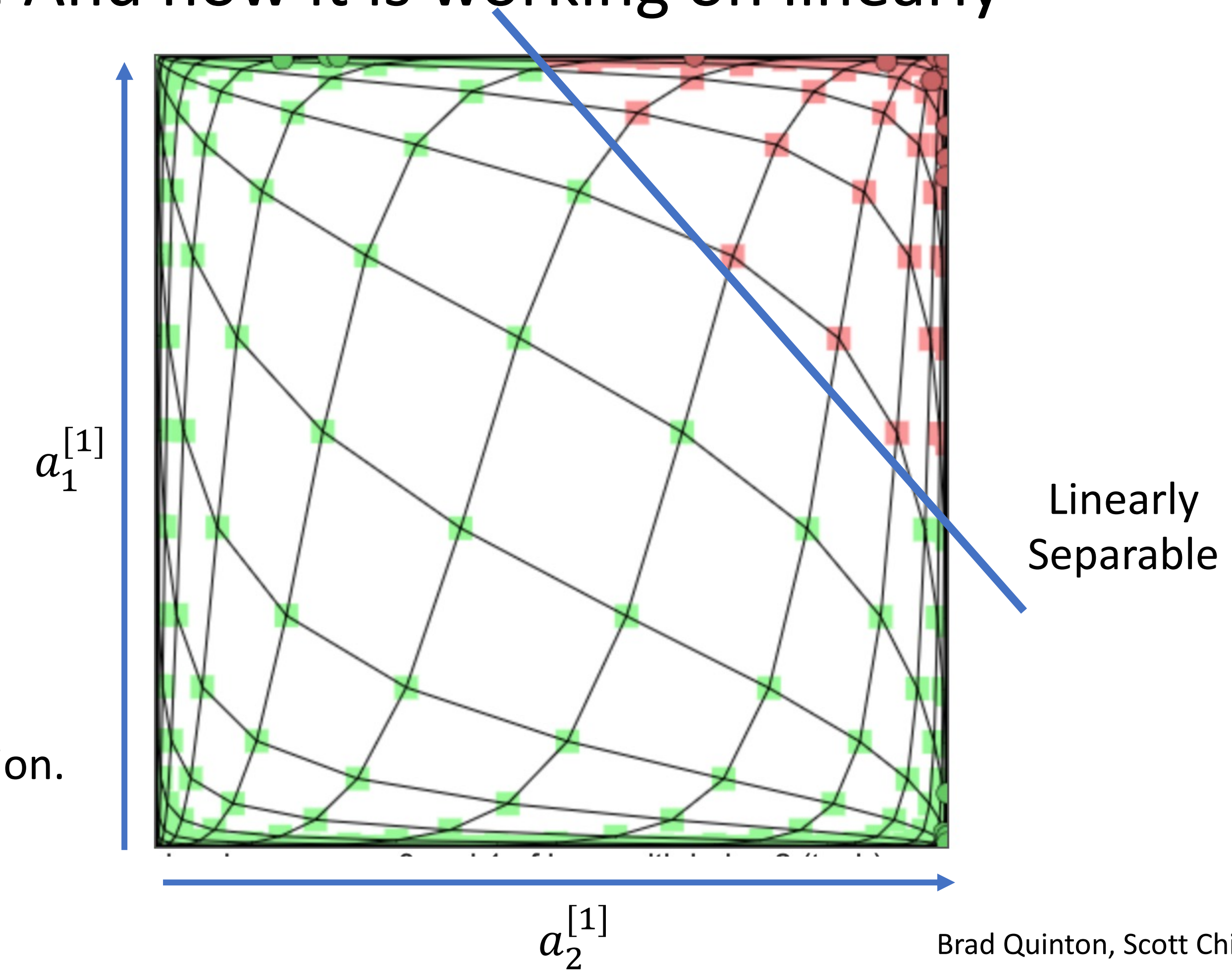
Feature Space Transformation

- The final unit is logistic regression. And now it is working on linearly separable data



$$z_i^{[1]} = W^{[1]}x + b_i^{[1]}$$
$$a_i^{[1]} = \tanh\left(z_i^{[1]}\right)$$

$$z_1^{[2]} = W^{[2]}a^{[1]} + b^{[2]}$$
$$a_1^{[2]} = \sigma\left(z_1^{[2]}\right)$$



Note: We use 2 units in the hidden layer for easy visualization.
With 3 units, we would be looking at a 3D space, etc.

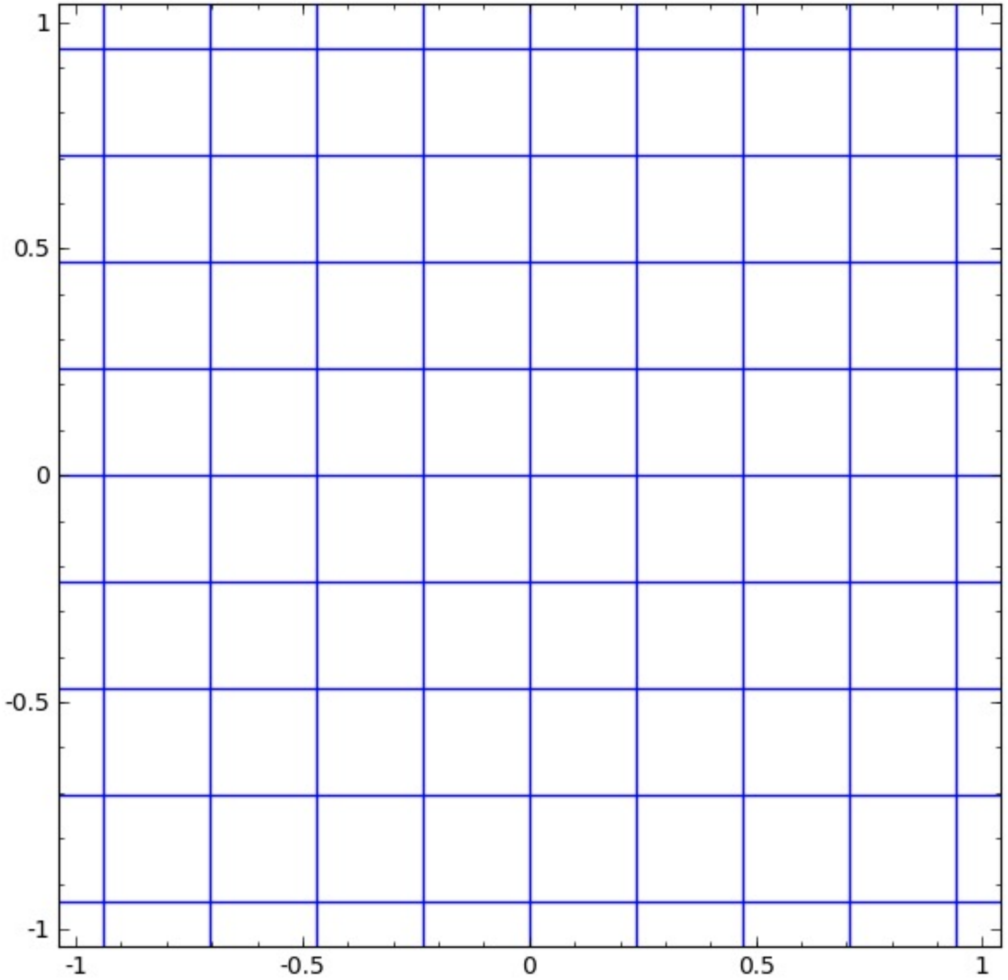
https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html

Brad Quinton, Scott Chin

Feature Space Transformation

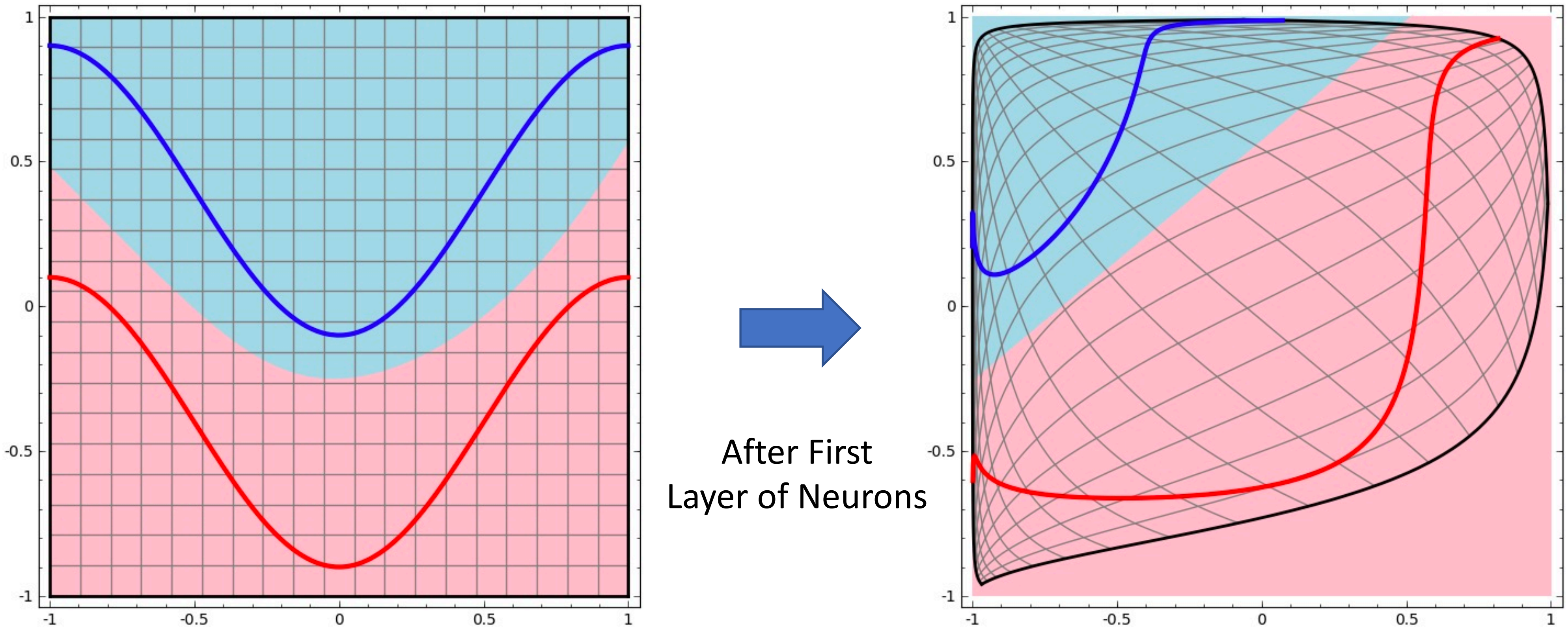
For $\tanh(Wx + b)$

1. A linear transformation of W
2. A translation of b
3. An application of tanh()



Brad Quinton, Scott Chin

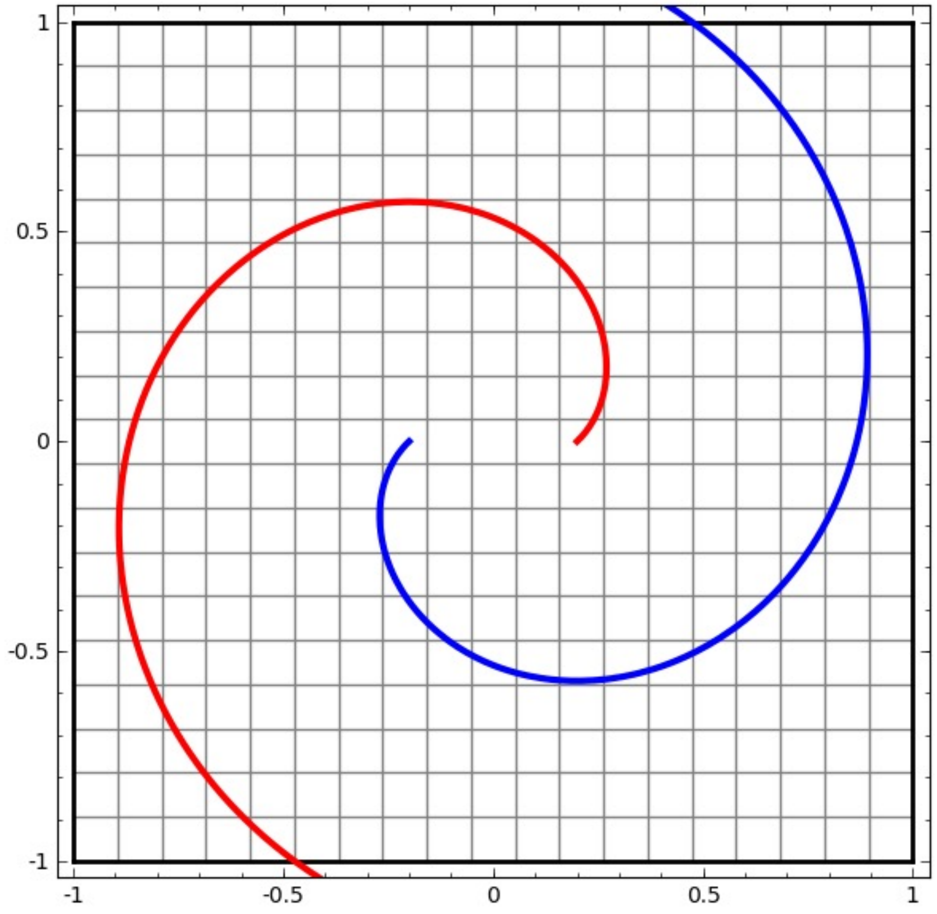
Feature Space Transformation



http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Brad Quinton, Scott Chin

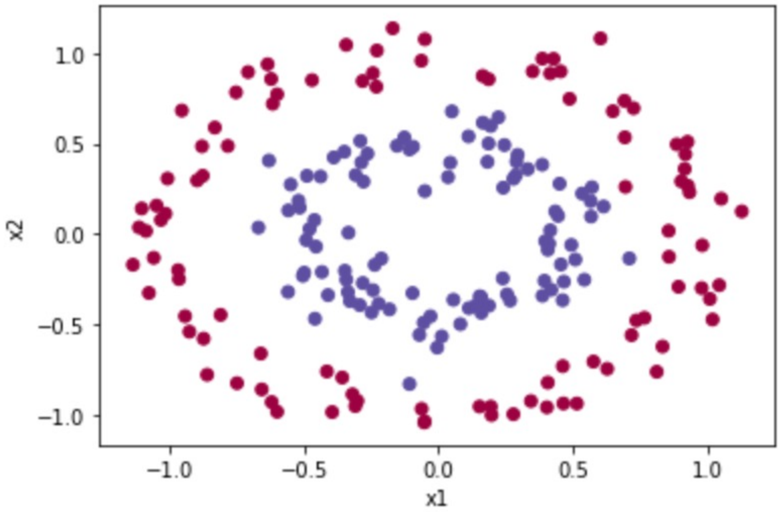
Feature Space Transformation



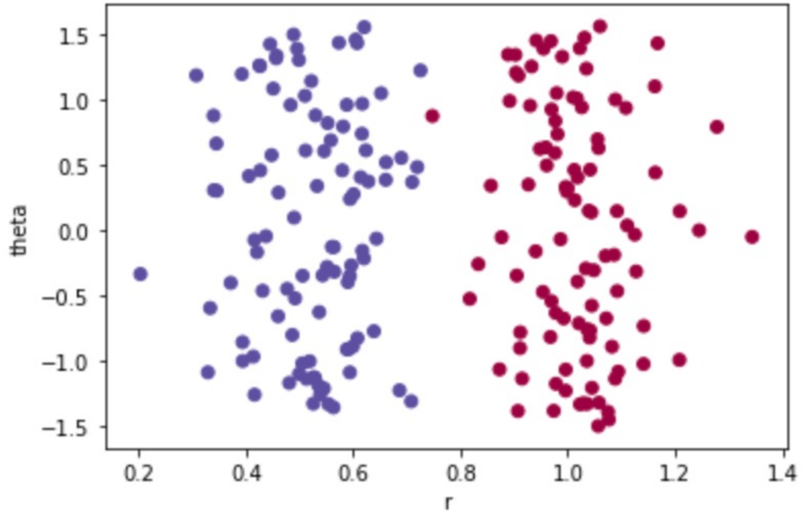
http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Feature Engineering

- With Logistic regression (or any linear classifier), you can manually transform your features to encode non-linearity.
- This is called Feature Engineering and requires analysis and human effort
  - i.e. you engineer new synthetic features from the real ones



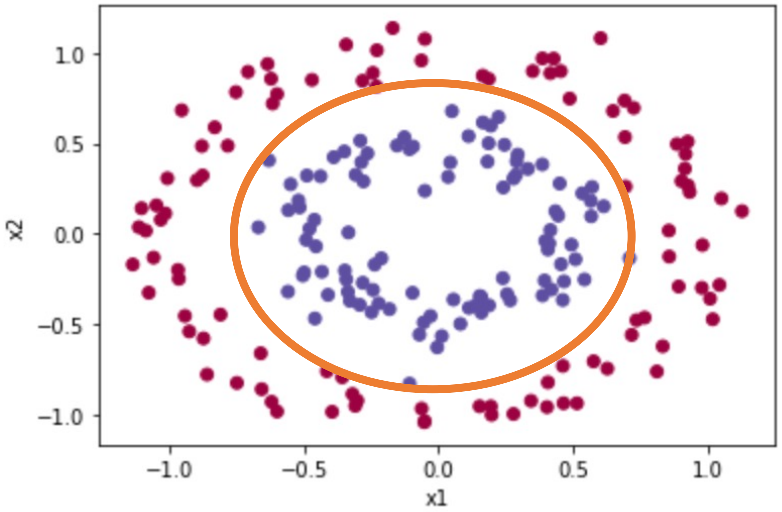
$$r = \sqrt{x_1^2 + x_2^2} \qquad \theta = arctan\left(\frac{x_1}{x_2}\right)$$



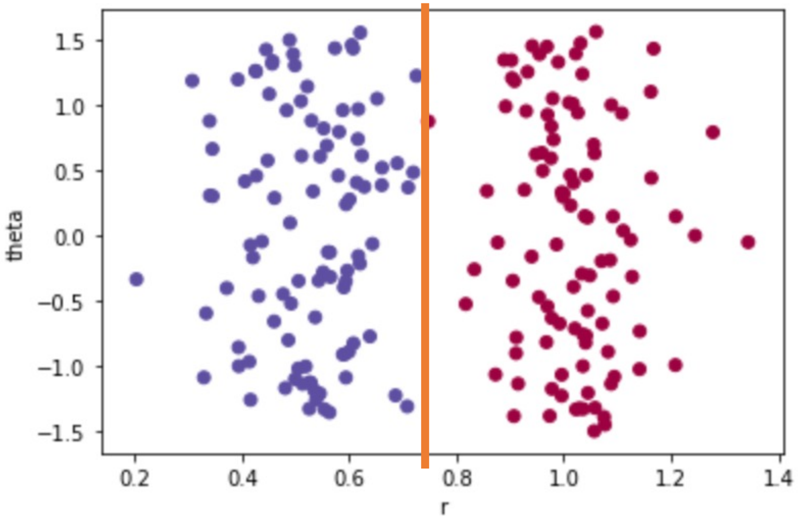
Brad Quinton, Scott Chin

Feature Engineering

- With Logistic regression (or any linear classifier), you can manually transform your features to encode non-linearity.
- This is called Feature Engineering and requires analysis and human effort
  - i.e. you engineer new synthetic features from the real ones
- Data is now linearly separable



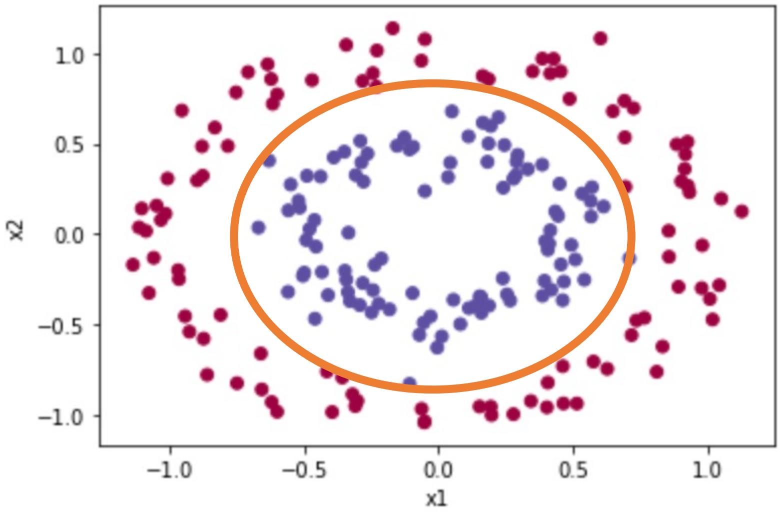
$$r = \sqrt{x_1^2 + x_2^2} \qquad \theta = arctan\left(\frac{x_1}{x_2}\right)$$



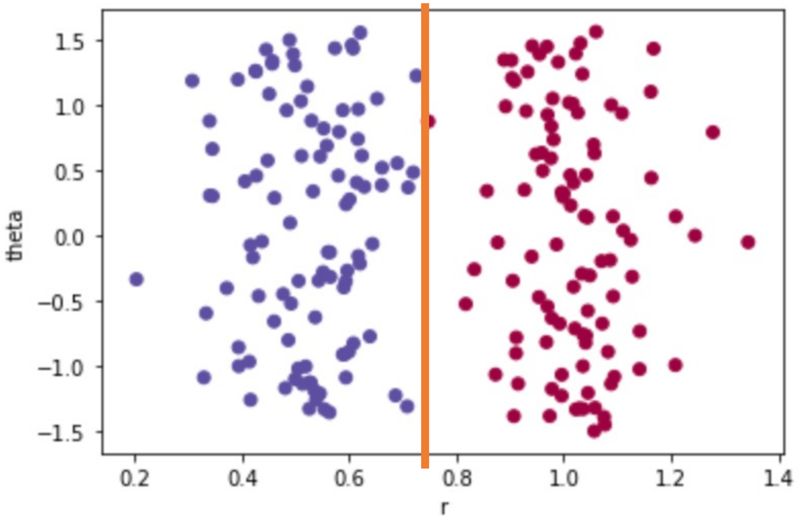
Brad Quinton, Scott Chin

Feature Engineering

- With Logistic regression (or any linear classifier), you can manually transform your features to encode non-linearity.
- This is called Feature Engineering and requires analysis and human effort
  - i.e. you engineer new synthetic features from the real ones
- Data is now linearly separable



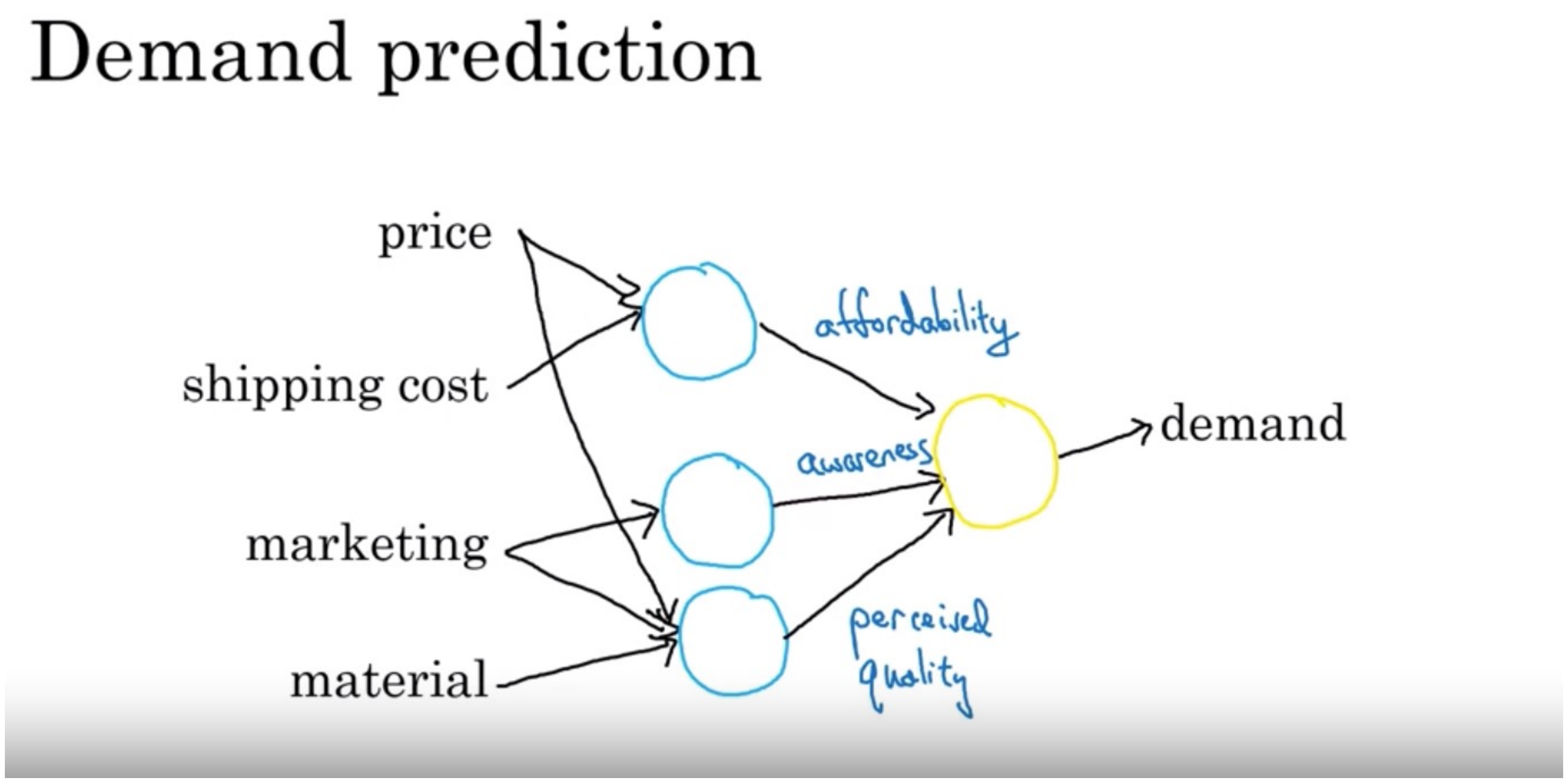
$$r = \sqrt{x_1^2 + x_2^2} \qquad \theta = arctan\left(\frac{x_1}{x_2}\right)$$



Neural networks learn the "best" transform from the data without human assistance

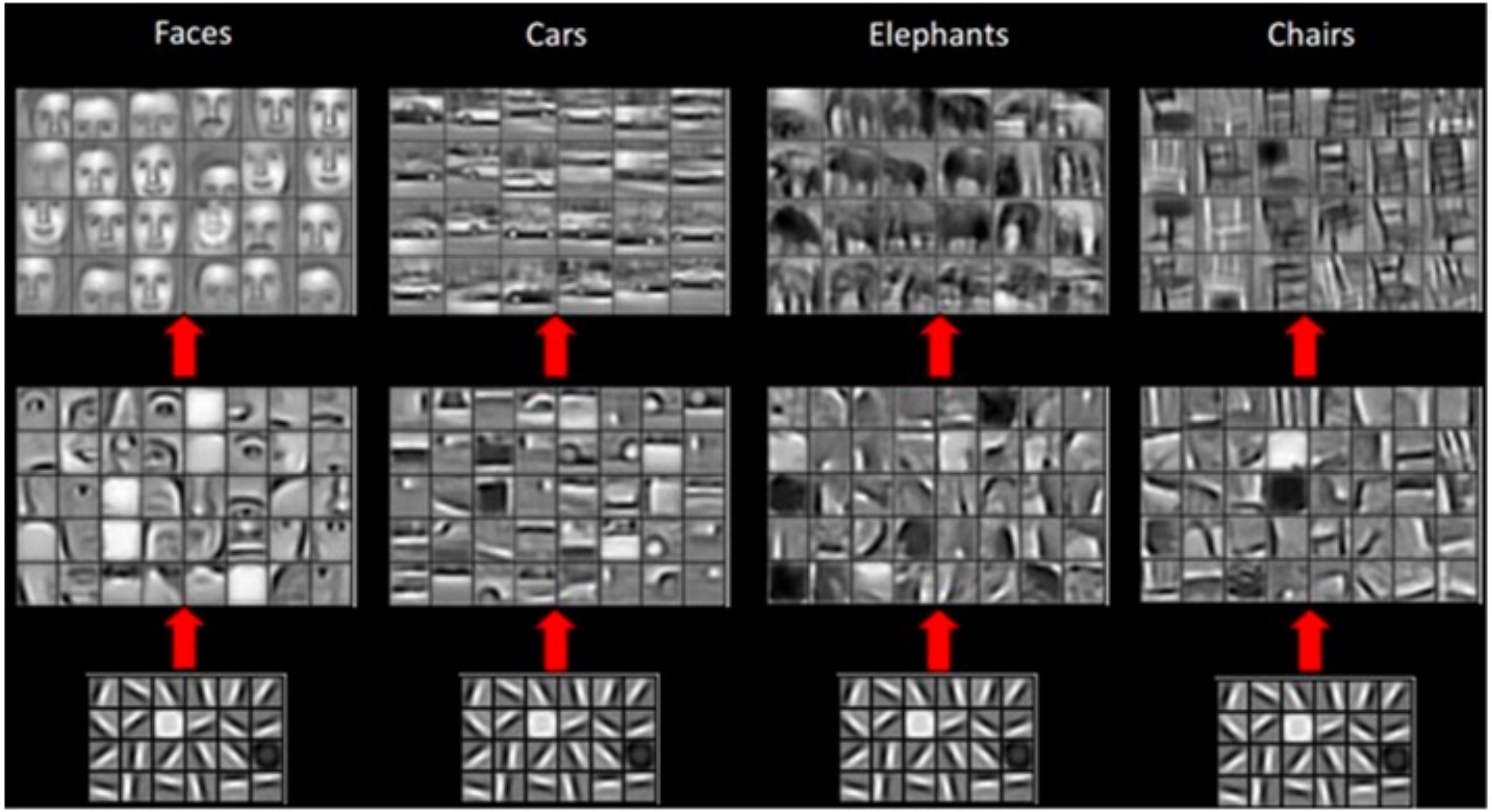
Brad Quinton, Scott Chin

Interpret as Feature Composition



A.I. For Everyone, Andrew Ng, Coursera

Interpret as Feature Composition



Brad Quinton, Scott Chin

Deep Neural Networks

Brad Quinton, Scott Chin

What is a “Deep” Neural Network (DNN)

What is a "Deep" Neural Network (DNN)

- Hmmm… a neural network with many hidden layers….

Brad Quinton, Scott Chin

What is a "Deep" Neural Network (DNN)

- Hmmm… a neural network with many hidden layers….
- What is "many"?

What is a “Deep” Neural Network (DNN)

- Hmmm… a neural network with many hidden layers….
- What is “many”?
- Hmmm…. Technically anything more than 1?

Brad Quinton, Scott Chin

What is a “Deep” Neural Network (DNN)

- Hmmm… a neural network with many hidden layers….
- What is “many”?
- Hmmm…. Technically anything more than 1?

There is no formal definition.

Brad Quinton, Scott Chin

Examples of CNN Depth

Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.713	0.901	138,357,544	23
VGG19	549 MB	0.713	0.900	143,667,240	26
ResNet50	98 MB	0.749	0.921	25,636,712	-
ResNet101	171 MB	0.764	0.928	44,707,176	-
ResNet152	232 MB	0.766	0.931	60,419,944	-
ResNet50V2	98 MB	0.760	0.930	25,613,800	-
ResNet101V2	171 MB	0.772	0.938	44,675,560	-
ResNet152V2	232 MB	0.780	0.942	60,380,648	-
InceptionV3	92 MB	0.779	0.937	23,851,784	159
InceptionResNetV2	215 MB	0.803	0.953	55,873,736	572
MobileNet	16 MB	0.704	0.895	4,253,864	88
MobileNetV2	14 MB	0.713	0.901	3,538,984	88
DenseNet121	33 MB	0.750	0.923	8,062,504	121
DenseNet169	57 MB	0.762	0.932	14,307,880	169
DenseNet201	80 MB	0.773	0.936	20,242,984	201
NASNetMobile	23 MB	0.744	0.919	5,326,716	-
NASNetLarge	343 MB	0.825	0.960	88,949,818	-

Universal Approximation Theorem

In mathematical theory, a neural network with one hidden layer can **approximate** any continuous function!

$$Y(X) = \sigma\left(W^{[2]}\tanh\left(W^{[1]}X + B^{[1]}\right) + B^{[2]}\right)$$

"A visual proof that neural nets can compute any function", Michael Nielsen, http://neuralnetworksanddeeplearning.com/chap4.html
"Universal Approximation Theorem", Wikipedia, https://en.wikipedia.org/wiki/Universal_approximation_theorem

Universal Approximation Theorem

In mathematical theory, a neural network with one hidden layer can **approximate** any continuous function!

$$Y(X) = \sigma\left(W^{[2]}\tanh\left(W^{[1]}X + B^{[1]}\right) + B^{[2]}\right)$$

But whether the suitable parameters for such a neural network can be easily found, or how many units you'd need are unanswered question.

"A visual proof that neural nets can compute any function", Michael Nielsen, http://neuralnetworksanddeeplearning.com/chap4.html
"Universal Approximation Theorem", Wikipedia, https://en.wikipedia.org/wiki/Universal_approximation_theorem

Deep or Shallow?

- The Universal Approximation Theorem proves that a shallow network can approximate any function
- In practice however, we find that deep networks generally perform better than shallow ones. Especially on unstructured data with wide variation (e.g. image, speech)

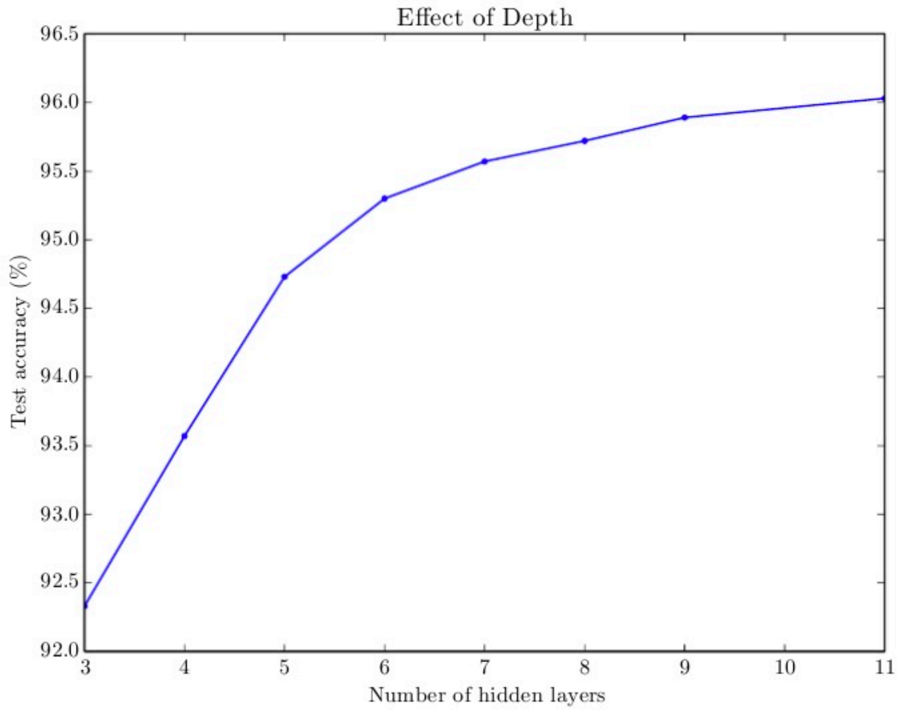


Figure 6.6: Empirical results showing that deeper networks generalize better when used to transcribe multi-digit numbers from photographs of addresses. Data from Goodfellow *et al.* (2014d). The test set accuracy consistently increases with increasing depth. See Fig. 6.7 for a control experiment demonstrating that other increases to the model size do not yield the same effect.

Deep Learning Book, Goodefellow, Bengio, Courville, http://www.deeplearningbook.org/

Why is Deep often Better

No one knows for sure, and it really depends on the problem, but …

1. Maybe shallow networks need an infeasibly larger number of parameters (i.e. units) than a deep one?
2. Maybe shallow networks are just harder to train?
   - lack of good techniques?
   - more local minima?
   - shallow gradients?
3. Maybe shallow networks overfit more
4. Maybe the problems being tackled these days are inherently hierarchical and lend better to deep networks.

Deep Learning Book, Chp6.4.1, Goodefellow, Bengio, Courville, http://www.deeplearningbook.org/

Brad Quinton, Scott Chin

Where does Deep Learning work well?

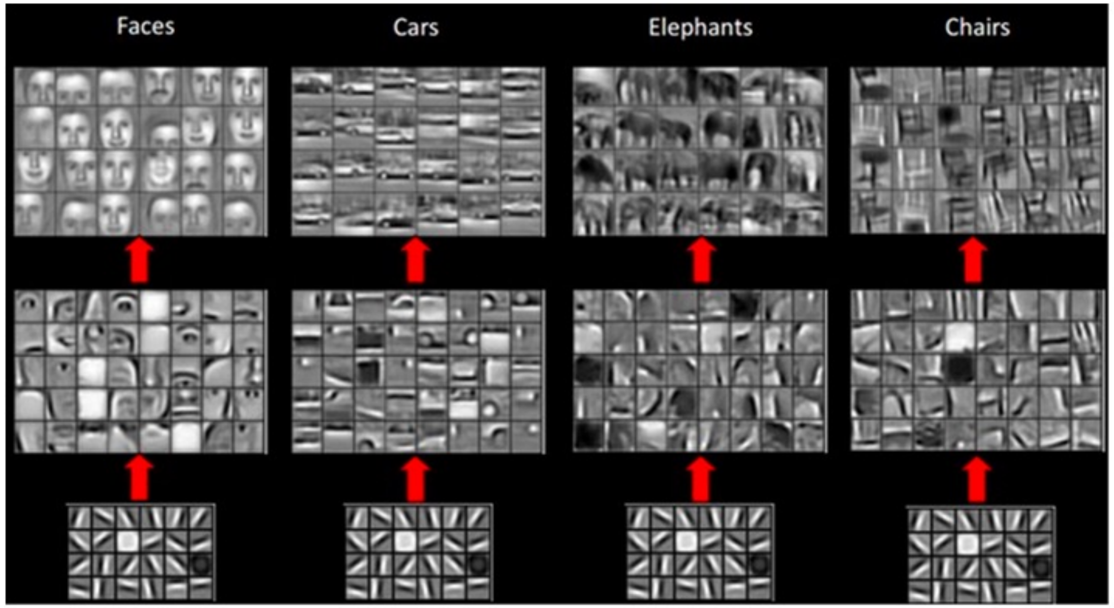
1. Problems where the input is *unstructured data*
   - Images/video
   - Radar/lidar/sonar
   - X-ray/MRI/ultrasound/EEG/EKG
   - Audio/voice
   - Natural Language
   - ‘mixed’ data sources (radar and image data, 3D MRI, etc.)

2. Problems with complex relationships but clear goals
   - Classifying images
   - Identifying objects
   - Winning chess
   - Predicting consumer behavior

Brad Quinton, Scott Chin

Problems that are Amenable to Composition?

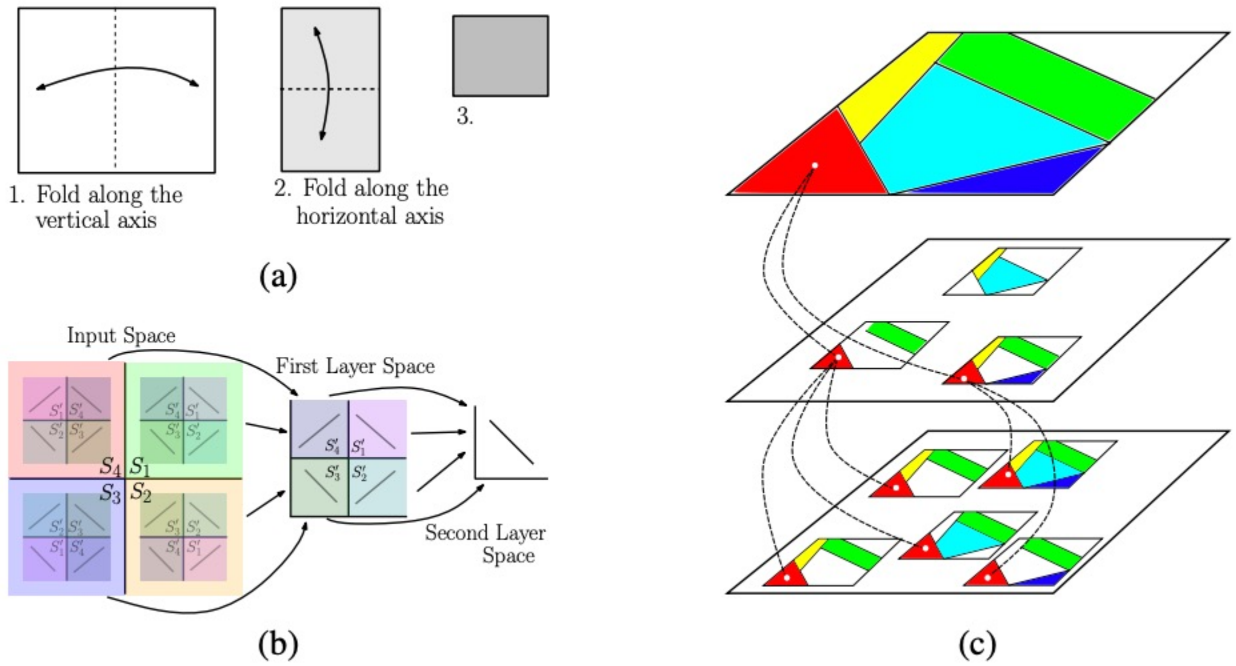
- Intuitively, a lot of problems have a hierarchical nature and require composition for efficient learning.
- For example, computer vision, you may want to start by understanding pixels → edges → simple geometric patterns → larger and larger patterns

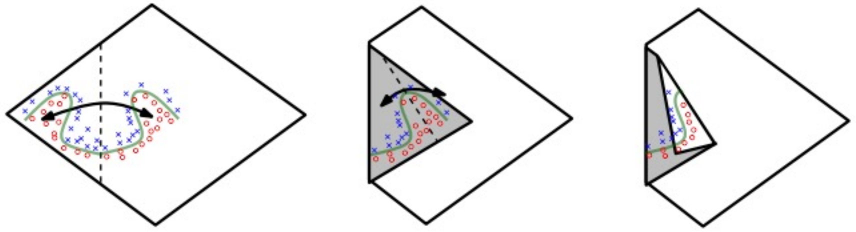


Brad Quinton, Scott Chin

Feature Space Transform – Revisted

- Montufar describes the feature space as having many symmetries
- Looks at each network layer as a folding of the feature space to collapse symmetries
- Final layer operate on collapsed space which simplifies the problem

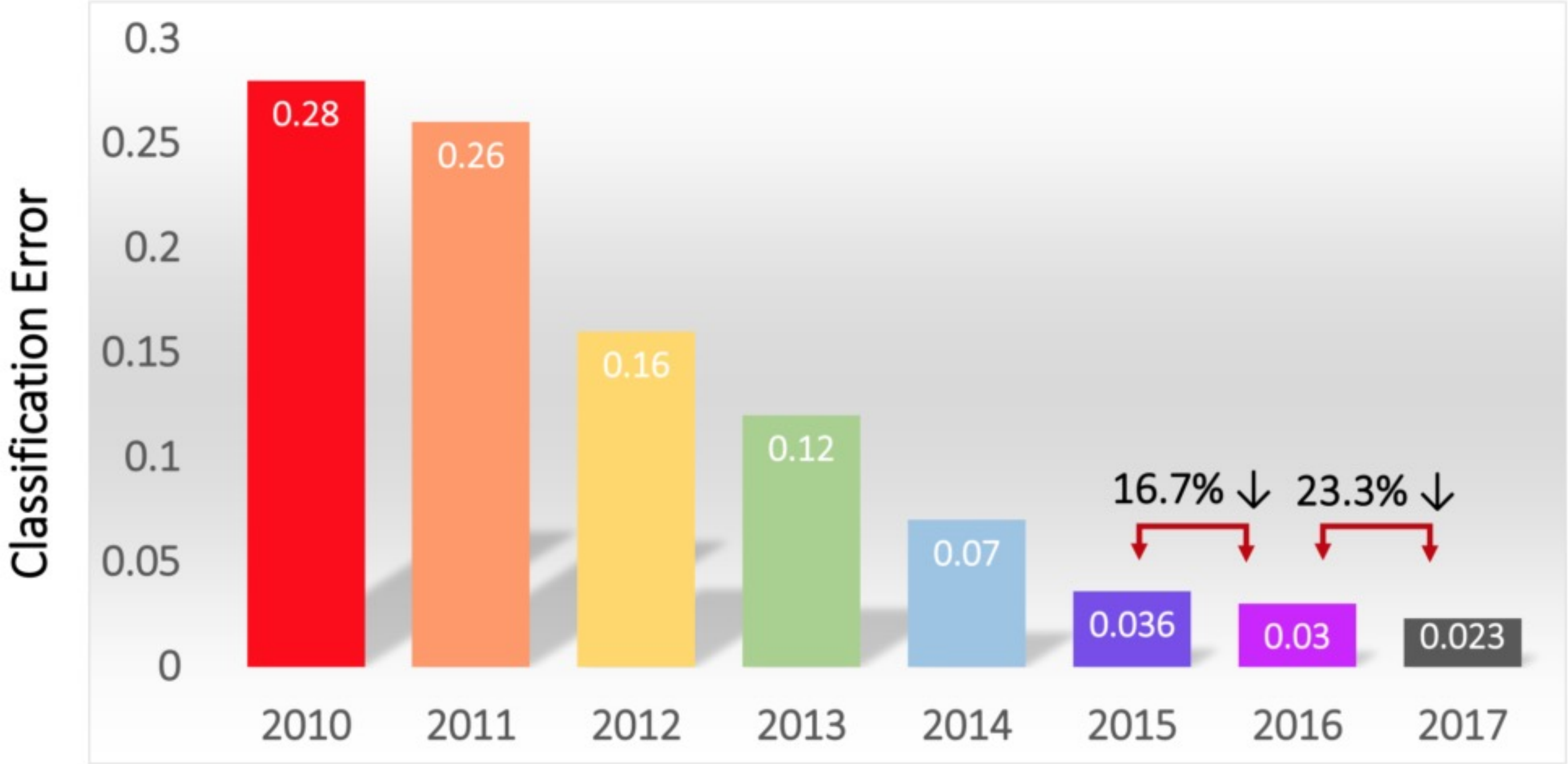




"On the Number of Linear Regions of Deep Neural Networks", Montufar et al, 2014, https://arxiv.org/pdf/1402.1869.pdf

Brad Quinton, Scott Chin

ImageNet (ILSVRC) Competition Winners



# Layers			8	8	22	152

Image Source: https://www.kaggle.com/getting-started/149448

Backpropagation Through Softmax and Categorical Cross-Entropy Loss

Brad Quinton, Scott Chin

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j)$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = \frac{\partial}{\partial z_1}\left(-y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)\right)$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) \ = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = \frac{\partial}{\partial z_1}\left(-y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) \ - y_3 log(\hat{y}_3)\right)$$

$$= -y_1 \frac{\partial}{\partial z_1}\left(log(\hat{y}_1)\right) - y_2 \frac{\partial}{\partial z_1}\left(log(\hat{y}_2)\right) - y_3 \frac{\partial}{\partial z_1}\left(log(\hat{y}_3)\right)$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = \frac{\partial}{\partial z_1}\left(-y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)\right)$$

$$= -y_1 \frac{\partial}{\partial z_1}\left(log(\hat{y}_1)\right) - y_2 \frac{\partial}{\partial z_1}\left(log(\hat{y}_2)\right) - y_3 \frac{\partial}{\partial z_1}\left(log(\hat{y}_3)\right)$$

$$= -y_1 \frac{d}{d\hat{y}_1}\left(log(\hat{y}_1)\right)\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{d}{d\hat{y}_2}\left(log(\hat{y}_2)\right)\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{d}{d\hat{y}_3}\left(log(\hat{y}_3)\right)\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\frac{d}{dx}(logx) = \frac{1}{x}$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\begin{aligned}
\frac{\partial L(\hat{y}, y)}{\partial z_1} &= \frac{\partial}{\partial z_1}\left(-y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)\right) \\
&= -y_1 \frac{\partial}{\partial z_1}\left(log(\hat{y}_1)\right) - y_2 \frac{\partial}{\partial z_1}\left(log(\hat{y}_2)\right) - y_3 \frac{\partial}{\partial z_1}\left(log(\hat{y}_3)\right) \\
&= -y_1 \frac{d}{d\hat{y}_1}\left(log(\hat{y}_1)\right)\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{d}{d\hat{y}_2}\left(log(\hat{y}_2)\right)\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{d}{d\hat{y}_3}\left(log(\hat{y}_3)\right)\frac{\partial \hat{y}_3}{\partial z_1} \\
&= -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}
\end{aligned}$$

$$\frac{d}{dx}(log x) = \frac{1}{x}$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1} \frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2} \frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3} \frac{\partial \hat{y}_3}{\partial z_1}$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1} \frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2} \frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3} \frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1} \frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2} \frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3} \frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \frac{e^{z_1}(e^{z_1} + e^{z_2} + e^{z_3}) - e^{z_1} e^{z_1}}{(e^{z_1} + e^{z_2} + e^{z_3})^2}$$

$$\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{\left(\frac{d}{dx} f(x)\right) g(x) - f(x) \left(\frac{d}{dx} g(x)\right)}{\left(g(x)\right)^2}$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \frac{e^{z_1}(e^{z_1} + e^{z_2} + e^{z_3}) - e^{z_1}e^{z_1}}{(e^{z_1} + e^{z_2} + e^{z_3})^2}$$

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{\left(\frac{d}{dx}f(x)\right)g(x) - f(x)\left(\frac{d}{dx}g(x)\right)}{\left(g(x)\right)^2}$$

$$= \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} - \left(\frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}\right)^2$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) \; = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \frac{e^{z_1}(e^{z_1} + e^{z_2} + e^{z_3}) - e^{z_1}e^{z_1}}{(e^{z_1} + e^{z_2} + e^{z_3})^2}$$

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{\left(\frac{d}{dx}f(x)\right)g(x) - f(x)\left(\frac{d}{dx}g(x)\right)}{\left(g(x)\right)^2}$$

$$= \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} - \left(\frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}\right)^2$$

$$= \hat{y}_1 - (\hat{y}_1)^2$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1} \frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2} \frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3} \frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \frac{e^{z_1}(e^{z_1} + e^{z_2} + e^{z_3}) - e^{z_1} e^{z_1}}{(e^{z_1} + e^{z_2} + e^{z_3})^2}$$

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{\left(\frac{d}{dx} f(x)\right) g(x) - f(x)\left(\frac{d}{dx} g(x)\right)}{\left(g(x)\right)^2}$$

$$= \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} - \left(\frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}\right)^2$$

$$= \hat{y}_1 - (\hat{y}_1)^2$$

$$= \hat{y}_1(1 - \hat{y}_1)$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) \; = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1)$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1)$$

$$\frac{\partial \hat{y}_2}{\partial z_1} = \frac{0(e^{z_1}) - e^{z_2}e^{z_1}}{(e^{z_1} + e^{z_2} + e^{z_3})^2}$$

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{\left(\frac{d}{dx}f(x)\right)g(x) - f(x)\left(\frac{d}{dx}g(x)\right)}{\left(g(x)\right)^2}$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1)$$

$$\frac{\partial \hat{y}_2}{\partial z_1} = \frac{0(e^{z_1}) - e^{z_2}e^{z_1}}{(e^{z_1} + e^{z_2} + e^{z_3})^2}$$

$$= \frac{-e^{z_2}e^{z_1}}{(e^{z_1} + e^{z_2} + e^{z_3})^2}$$

$$= -\frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \cdot \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$= -\hat{y}_2\hat{y}_1$$

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{\left(\frac{d}{dx}f(x)\right)g(x) - f(x)\left(\frac{d}{dx}g(x)\right)}{\left(g(x)\right)^2}$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1)$$

$$\frac{\partial \hat{y}_2}{\partial z_1} = -\hat{y}_2\hat{y}_1$$

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1} \frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2} \frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3} \frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1)$$

$$\frac{\partial \hat{y}_2}{\partial z_1} = -\hat{y}_2 \hat{y}_1$$

$$\frac{\partial \hat{y}_3}{\partial z_1} = -\hat{y}_3 \hat{y}_1$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1} \frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2} \frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3} \frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1) \qquad \frac{\partial \hat{y}_2}{\partial z_1} = -\hat{y}_2 \hat{y}_1 \qquad \frac{\partial \hat{y}_3}{\partial z_1} = -\hat{y}_3 \hat{y}_1$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1) \qquad \frac{\partial \hat{y}_2}{\partial z_1} = -\hat{y}_2\hat{y}_1 \qquad \frac{\partial \hat{y}_3}{\partial z_1} = -\hat{y}_3\hat{y}_1$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\hat{y}_1(1 - \hat{y}_1) - y_2 \frac{1}{\hat{y}_2}(-\hat{y}_2\hat{y}_1) - y_3 \frac{1}{\hat{y}_3}(-\hat{y}_3\hat{y}_1)$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1) \qquad \frac{\partial \hat{y}_2}{\partial z_1} = -\hat{y}_2\hat{y}_1 \qquad \frac{\partial \hat{y}_3}{\partial z_1} = -\hat{y}_3\hat{y}_1$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\hat{y}_1(1 - \hat{y}_1) - y_2 \frac{1}{\hat{y}_2}(-\hat{y}_2\hat{y}_1) - y_3 \frac{1}{\hat{y}_3}(-\hat{y}_3\hat{y}_1)$$

$$= -y_1 \frac{1}{\hat{y}_1}\hat{y}_1(1 - \hat{y}_1) - y_2 \frac{1}{\hat{y}_2}(-\hat{y}_2\hat{y}_1) - y_3 \frac{1}{\hat{y}_3}(-\hat{y}_3\hat{y}_1)$$

$$= \hat{y}_1(y_1 + y_2 + y_3) - y_1$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log\left(\hat{y}_j\right) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1) \qquad \frac{\partial \hat{y}_2}{\partial z_1} = -\hat{y}_2\hat{y}_1 \qquad \frac{\partial \hat{y}_3}{\partial z_1} = -\hat{y}_3\hat{y}_1$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\hat{y}_1(1 - \hat{y}_1) - y_2 \frac{1}{\hat{y}_2}(-\hat{y}_2\hat{y}_1) - y_3 \frac{1}{\hat{y}_3}(-\hat{y}_3\hat{y}_1)$$

$$= -y_1 \frac{1}{\hat{y}_1}\hat{y}_1(1 - \hat{y}_1) - y_2 \frac{1}{\hat{y}_2}(-\hat{y}_2\hat{y}_1) - y_3 \frac{1}{\hat{y}_3}(-\hat{y}_3\hat{y}_1)$$

$$= \hat{y}_1(y_1 + y_2 + y_3) - y_1 \qquad\qquad y_1 + y_2 + y_3 = 1$$

$$= \hat{y}_1 - y_1$$

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = -y_1 \frac{1}{\hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_1} - y_2 \frac{1}{\hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_1} - y_3 \frac{1}{\hat{y}_3}\frac{\partial \hat{y}_3}{\partial z_1}$$

$$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \hat{y}_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1} = \hat{y}_1(1 - \hat{y}_1) \qquad \frac{\partial \hat{y}_2}{\partial z_1} = -\hat{y}_2\hat{y}_1 \qquad \frac{\partial \hat{y}_3}{\partial z_1} = -\hat{y}_3\hat{y}_1$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = \hat{y}_1 - y_1$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = \hat{y}_1 - y_1$$

Brad Quinton, Scott Chin

Consider $n_c = 3$

$$L(\hat{y}, y) = -\sum_{j=1}^{n_c} y_j log(\hat{y}_j) = -y_1 log(\hat{y}_1) - y_2 log(\hat{y}_2) - y_3 log(\hat{y}_3)$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1}, \frac{\partial L(\hat{y}, y)}{\partial z_2}, \frac{\partial L(\hat{y}, y)}{\partial z_3}$$

$$\frac{\partial L(\hat{y}, y)}{\partial z_1} = \hat{y}_1 - y_1 \qquad \frac{\partial L(\hat{y}, y)}{\partial z_2} = \hat{y}_2 - y_2 \qquad \frac{\partial L(\hat{y}, y)}{\partial z_3} = \hat{y}_3 - y_3$$

Brad Quinton, Scott Chin

What was the point of that!?

- Deriving a closed-form analytical formula can be error-prone, tedious, and may not lead to a simple form.
- What if you wanted to use a different loss function?
- What if you wanted to use something that isn’t mainstream?

- Next lecture we will introduce a systematic, modular, numerical approach that you can apply to compute all the partial derivatives via backpropagation.

Brad Quinton, Scott Chin

Learning Objectives

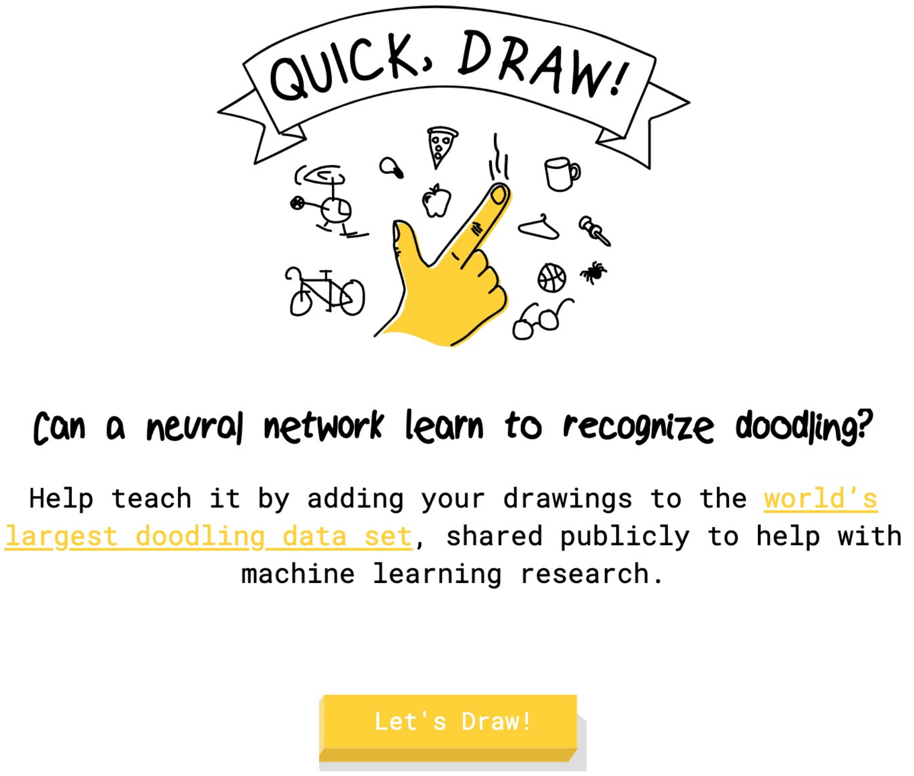
- What is a “Deep” Neural Network?
- Intuition of what hidden layers are doing
- Why use deep (vs shallow) networks?
- Backpropagation Through Softmax and Categorical Cross-Entropy Loss

Brad Quinton, Scott Chin

Extra "Homework"

Go play this fun little game https://quickdraw.withgoogle.com/

Think about how this could be done with a neural network



Brad Quinton, Scott Chin