

# EEP 596: LLMs: From Transformers to GPT || Lecture 3

Dr. Karthik Mohan

Univ. of Washington, Seattle

January 15, 2025

# Outline for Lecture

- Can 2 layer perceptron learn XOR?

# Outline for Lecture

- Can 2 layer perceptron learn XOR?
- Activation functions

# Outline for Lecture

- Can 2 layer perceptron learn XOR?
- Activation functions
- Tensorflow Demo



# Outline for Lecture

- Can 2 layer perceptron learn XOR?
- Activation functions
- Tensorflow Demo
- Training and Back-propagation

# Outline for Lecture

- Can 2 layer perceptron learn XOR?
- Activation functions
- Tensorflow Demo
- Training and Back-propagation
- Over-fitting and Hyper-parameters

# Outline for Lecture

- Can 2 layer perceptron learn XOR?
- Activation functions
- Tensorflow Demo
- Training and Back-propagation
- Over-fitting and Hyper-parameters
- Other DL architectures

# House Keeping Items

- Office Hours and Review Hours
- Assignment 1 due this weekend
- Assignment 2 to be assigned this Friday
- Any questions?

# Deep Learning Reference

## Deep Learning

Great reference for the theory and fundamentals of deep learning: Book by Goodfellow and Bengio et al [Bengio et al](#)

## [Deep Learning History](#)

# Recap from Last time!

- Introduction to Perceptron

# Recap from Last time!

- Introduction to Perceptron
- Perceptron and Logistic Regression

# Recap from Last time!

- Introduction to Perceptron
- Perceptron and Logistic Regression
- OR and AND functions

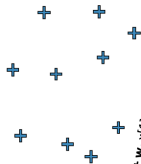


# Recap

# Perceptron

$$\text{Score}(x) = w_0 + w_1 x[1] + w_2 x[2] + \dots + w_d x[d]$$

$\text{Score}(x) > 0$

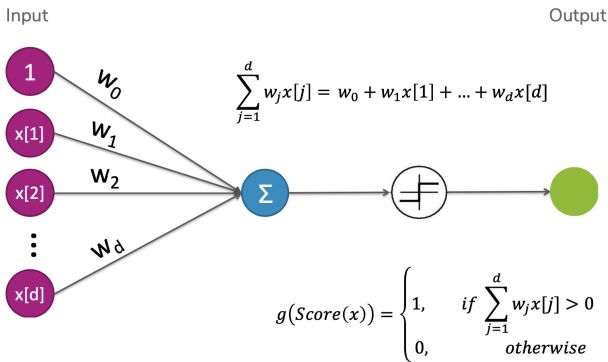


$$w_0 + w_1 x[1] + w_2 x[2] + \dots + w_d x[d] = 0$$

$\text{Score}(x) < 0$

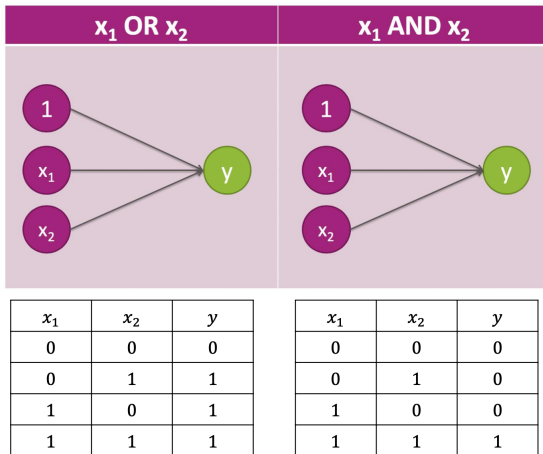


# Perceptron

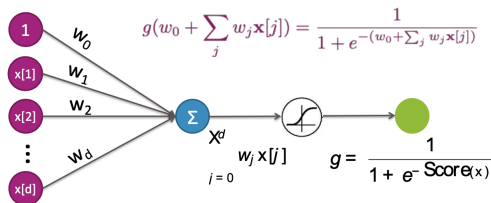


# OR and AND Functions

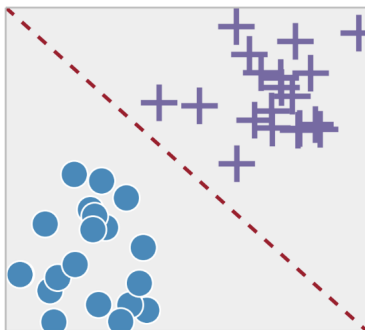
What can a perceptrons represent?



# Perceptron to Logistic Regression



# Logistic Regression



## LR fundamentals

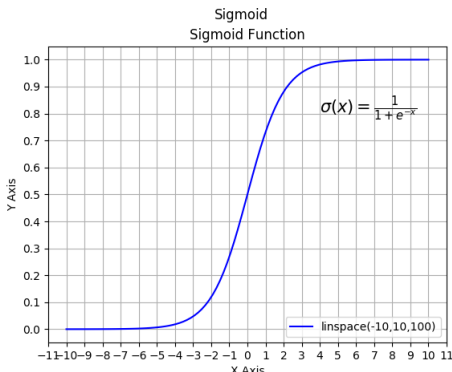
- Linear Model
- Want score  $w^T x^i > 0$  for  $y_i = +1$  and  $w^T x_i < 0$  for  $y_i = -1$ !
- If linearly separable data, above is feasible. Else, minimize error in separability!!

# Logistic Regression

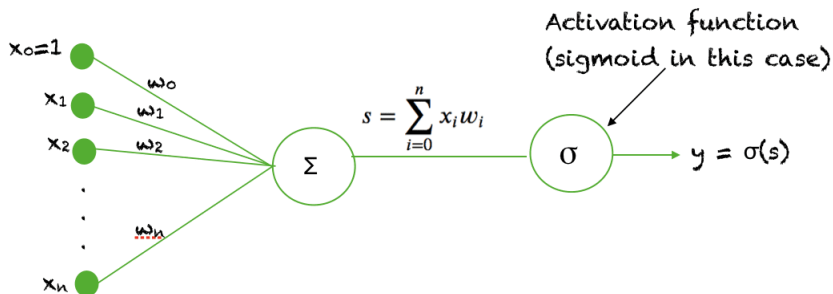
## Probability for a class

In LR, the score,  $w^T x$  is converted to a probability through the sigmoid function. So we can talk about  $P(\hat{y}^i = +1)$  or  $P(\hat{y}^i = -1)$

## Sigmoid Function



# LR represented Graphically





# Logistic Regression

## LR Prediction

$$\hat{y}_i = \frac{1}{1 + e^{-\hat{w}^T x^i}}$$

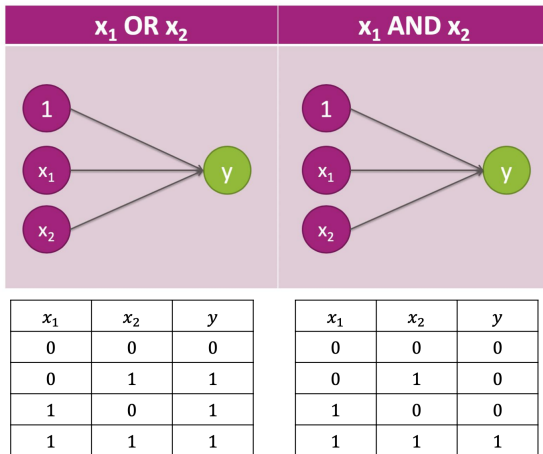
## LR Loss

Assume that  $y_i = 0$  or  $y_i = 1$  (i.e. the negative class has a label 0). Then the binary cross-entropy loss applies to LR:

$$\min_w y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

# OR and AND Functions

What can a perceptrons represent?



# Learning XOR

# XOR through 2 Layer perceptron

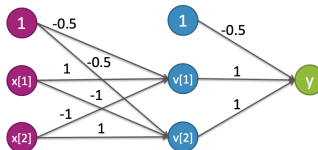
This is a 2-layer neural network

$$y = x[1] \text{ XOR } x[2] = (x[1] \text{ AND } \neg x[2]) \text{ OR } (\neg x[1] \text{ AND } x[2])$$

$$\begin{aligned} v[1] &= (x[1] \text{ AND } \neg x[2]) \\ &= g(-0.5 + x[1] - x[2]) \end{aligned}$$

$$\begin{aligned} v[2] &= (\neg x[1] \text{ AND } x[2]) \\ &= g(-0.5 - x[1] + x[2]) \end{aligned}$$

$$\begin{aligned} y &= v[1] \text{ OR } v[2] \\ &= g(-0.5 + v[1] + v[2]) \end{aligned}$$



# Why does Linear Activation not work?

# How Step Function activation works?

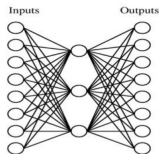
# ICE #1

Which methods can learn the XOR function?

- ① Logistics Regression
- ② Naive Bayes Classifier
- ③ Decision Trees
- ④ Support Vector Machines

## 2 Layer Neural Network

Two layer neural network (alt. one hidden-layer neural network)



Single

$$out(x) = g\left(w_0 + \sum_j w_j x[j]\right)$$

1-hidden layer

$$out(x) = g\left(w_0 + \sum_k w_k g\left(w_0^{(k)} + \sum_j w_j^{(k)} x[j]\right)\right)$$

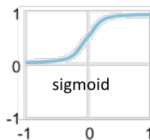


# Deep Learning: Activations, FFN and more

# Choices for Non-Linear Activation Function

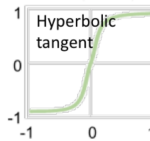
- **Sigmoid**

- Historically popular, but (mostly) fallen out of favor
- Neuron's activation saturates (weights get very large  $\rightarrow$  gradients get small)
- Not zero-centered  $\rightarrow$  other issues in the gradient steps
- When put on the output layer, called “softmax” because interpreted as class probability (soft assignment)



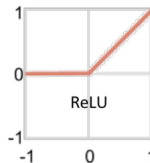
- **Hyperbolic tangent**  $g(x) = \tanh(x)$

- Saturates like sigmoid unit, but zero-centered



- **Rectified linear unit (ReLU)**  $g(x) = x^+ = \max(0, x)$

- Most popular choice these days
- Fragile during training and neurons can “die off”... be careful about learning rates
- “Noisy” or “leaky” variants



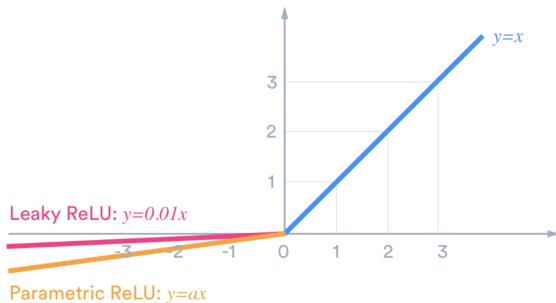
- **Softplus**  $g(x) = \log(1 + \exp(x))$

- Smooth approximation to rectifier activation

# Gradient of Sigmoid and RELU

# Sigmoid vs RELU

# RELU vs Leaky RELU



# Multi-Layer Perceptron (MLP)

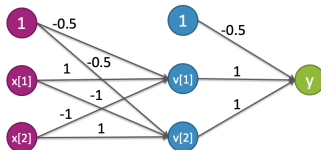
This is a 2-layer neural network

$$y = x[1] \text{ XOR } x[2] = (x[1] \text{ AND } \neg x[2]) \text{ OR } (\neg x[1] \text{ AND } x[2])$$

$$\begin{aligned} v[1] &= (x[1] \text{ AND } \neg x[2]) \\ &= g(-0.5 + x[1] - x[2]) \end{aligned}$$

$$\begin{aligned} v[2] &= (\neg x[1] \text{ AND } x[2]) \\ &= g(-0.5 - x[1] + x[2]) \end{aligned}$$

$$\begin{aligned} y &= v[1] \text{ OR } v[2] \\ &= g(-0.5 + v[1] + v[2]) \end{aligned}$$



# Breakout Session: Would ReLU work?

## Work in your breakouts

What weights would give a  $y$  value  $> 0.7$  for  $(1, 0)$ ,  $(0, 1)$  inputs and a value of  $y < -0.7$  for  $(0, 0)$ ,  $(1, 1)$  for the ReLU function?

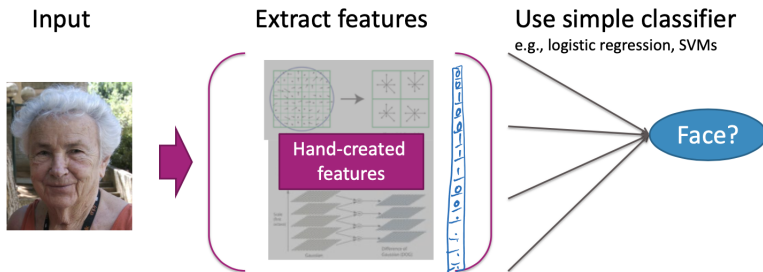
# The phenomenon of Overfitting



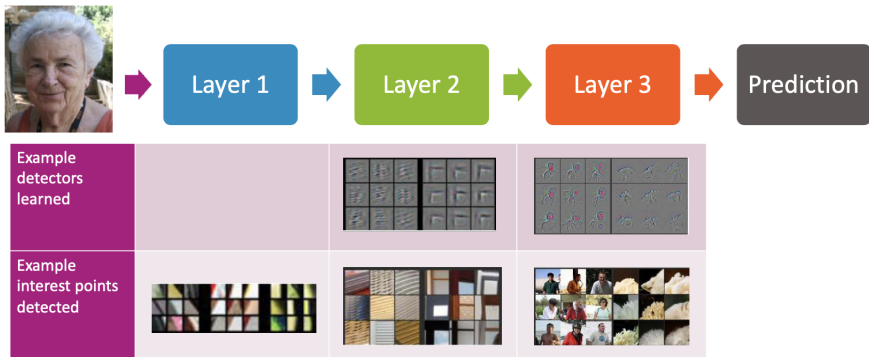
# Tensorflow Playground Demo

Tensorflow Playground Demo

# Computer vision before deep learning

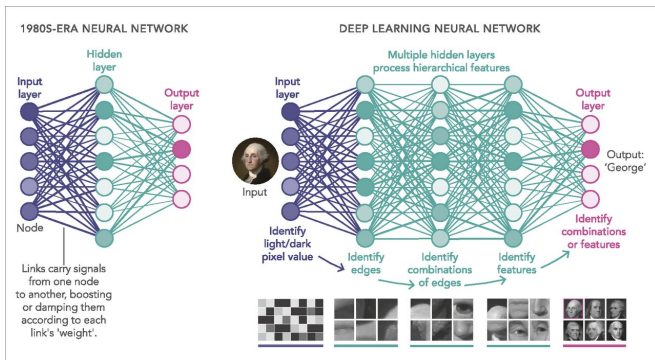


# Computer vision after deep learning

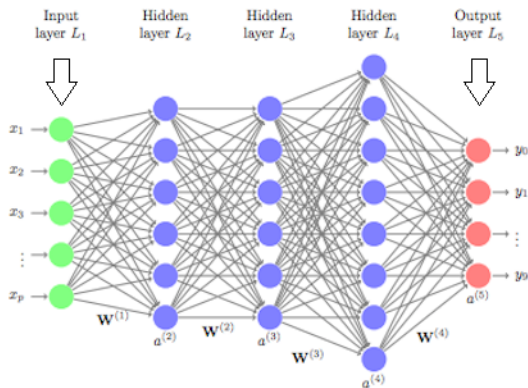


[Zeiler & Fergus '13]

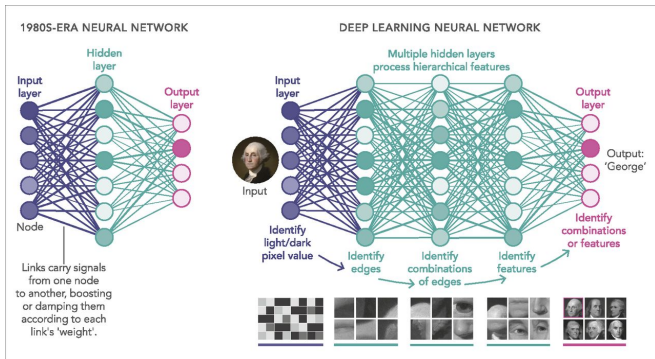
# Feed-forward Deep Learning Architecture Example



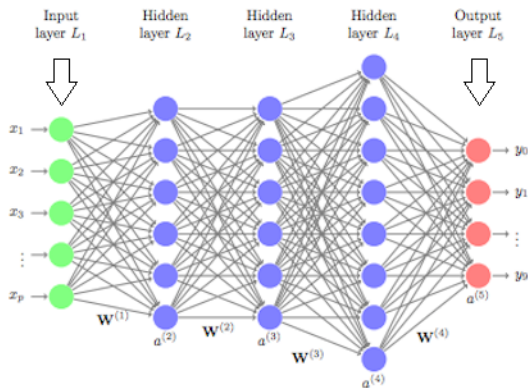
# Feed-forward Deep Learning Architecture Example



# Feed-forward Deep Learning Architecture Example



# Feed-forward Deep Learning Architecture Example



# ICE #2

## Compute the number of parameters in DNN model

Consider a DNN model with 3 hidden layers where each hidden layer has 1000 neurons. Let the input layer be raw pixels from a 100x100 image and the output layer has 10 dimensions, let's say for a 10 class image classification example. How many total parameters exist in the DNN model?

- ① 10 million parameters
- ② 11 million parameters
- ③ 12 million parameters
- ④ 13 million parameters



# Training a DNN

## SGD with mini-batch

SGD mini-batch is the staple diet. However there are some **learning rate schedulers** that are known to work better for DNNs - Such as Adagrad and more recently, ADAM. ADAM adapts the learning rate to each individual parameter instead of having a global learning rate.

# Training a DNN

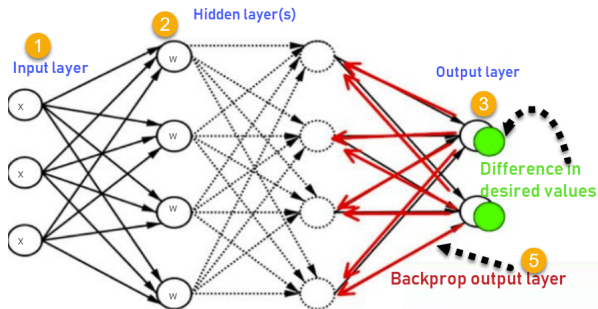
## SGD with mini-batch

SGD mini-batch is the staple diet. However there are some **learning rate schedulers** that are known to work better for DNNs - Such as Adagrad and more recently, ADAM. ADAM adapts the learning rate to each individual parameter instead of having a global learning rate.

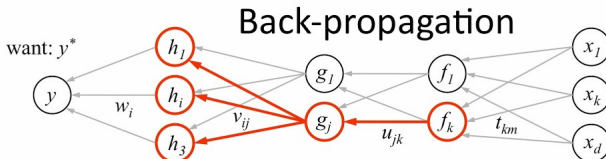
## How do we compute gradient in a DNN?

Back-propagation!

# Forward Propagation vs Back-propagation



# Back Propagation explained



1. receive new observation  $\mathbf{x} = [x_1, \dots, x_d]$  and target  $y^*$
2. **feed forward:** for each unit  $g_j$  in each layer  $1 \dots L$   
compute  $g_j$  based on units  $f_k$  from previous layer:  $g_j = \sigma \left( u_{j0} + \sum_k u_{jk} f_k \right)$
3. get prediction  $y$  and error  $(y - y^*)$
4. **back-propagate error:** for each unit  $g_j$  in each layer  $L \dots 1$

(a) compute error on  $g_j$

$$\frac{\partial E}{\partial g_j} = \sum_i \underbrace{\sigma'(h_i)}_{\text{should } g_j \text{ be higher or lower?}} \underbrace{v_{ij}}_{\text{how } h_i \text{ will change as } g_j \text{ changes}} \underbrace{\frac{\partial E}{\partial h_i}}_{\text{was } h_i \text{ too high or too low?}}$$

(b) for each  $u_{jk}$  that affects  $g_j$

(i) compute error on  $u_{jk}$

$$\frac{\partial E}{\partial u_{jk}} = \underbrace{\frac{\partial E}{\partial g_j}}_{\text{do we want } g_j \text{ to be higher/lower}} \underbrace{\sigma'(g_j) f_k}_{\text{how } g_j \text{ will change if } u_{jk} \text{ is higher/lower}}$$

(ii) update the weight

$$u_{jk} \leftarrow u_{jk} - \eta \frac{\partial E}{\partial u_{jk}}$$

Copyright © 2014 Victor Lavrenko

# Back Propagation Summary

## Back Prop

Back prop is one of the fundamental backbones of the training modules behind deep learning and beyond (including for example ChatGPT). What exactly is back prop? It is just a way to unravel gradient computation in the neural network. Back prop is how we would **compute the gradient** in a neural network.

# Back Propagation Summary

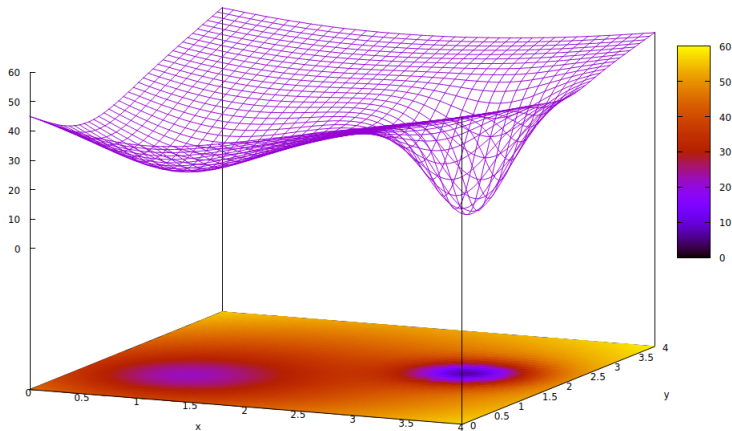
## Back Prop

Back prop is one of the fundamental backbones of the training modules behind deep learning and beyond (including for example ChatGPT). What exactly is back prop? It is just a way to unravel gradient computation in the neural network. Back prop is how we would **compute the gradient** in a neural network.

## Back Prop as information flow

It can also be thought of as flow information from the error in the output (the loss function) down to the weights. Update the weights so we don't make **this error** next time around. Back prop is a way to do **gradient descent in neural networks!**

# Good vs Bad Local minima



# Hyper-parameters in Deep Learning

ICE #3: Which of the following is not a hyper-parameter in deep learning?

- ① Learning rate
- ② Number of Hidden Layers
- ③ Number of neurons per hidden layer
- ④ All of the above



# Hyper-parameters in Deep Learning

## Hyper-parameters

- ① Learning rate
- ② Number of Hidden Layers
- ③ Number of neurons per hidden layer

# Hyper-parameters in Deep Learning

## Hyper-parameters

- ① Learning rate
- ② Number of Hidden Layers
- ③ Number of neurons per hidden layer
- ④ Type of non-linear activation function used

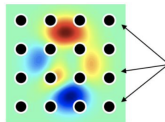
# Hyper-parameters in Deep Learning

## Hyper-parameters

- ① Learning rate
- ② Number of Hidden Layers
- ③ Number of neurons per hidden layer
- ④ Type of non-linear activation function used
- ⑤ Anything else?

# Hyper-parameter tuning methods

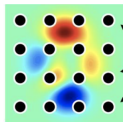
Grid search:



Hyperparameters  
on 2d uniform grid

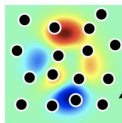
# Hyper-parameter tuning methods

Grid search:



Hyperparameters  
on 2d uniform grid

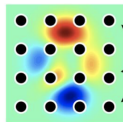
Random search:



Hyperparameters  
randomly chosen

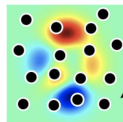
# Hyper-parameter tuning methods

Grid search:



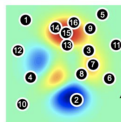
Hyperparameters  
on 2d uniform grid

Random search:



Hyperparameters  
randomly chosen

Bayesian Optimization:



Hyperparameters  
***adaptively*** chosen

# Over-fitting in DNNs

## How to handle over-fitting in DNNs

- ① A DNN model with 100 million parameters and only 100k data points or even a million data points will overfit unless we take care of over-fitting.

# Over-fitting in DNNs

## How to handle over-fitting in DNNs

- 1 A DNN model with 100 million parameters and only 100k data points or even a million data points will overfit unless we take care of over-fitting.
- 2 Weight regularization can help -  $\ell_1, \ell_2$



# Over-fitting in DNNs

## How to handle over-fitting in DNNs

- 1 A DNN model with 100 million parameters and only 100k data points or even a million data points will overfit unless we take care of over-fitting.
- 2 Weight regularization can help -  $\ell_1, \ell_2$
- 3 More common over-fitting strategy for DL?

# Over-fitting in DNNs

## How to handle over-fitting in DNNs

- ① A DNN model with 100 million parameters and only 100k data points or even a million data points will overfit unless we take care of over-fitting.
- ② Weight regularization can help -  $\ell_1, \ell_2$
- ③ More common over-fitting strategy for DL?
- ④ Dropouts!

# Over-fitting in DNNs

## How to handle over-fitting in DNNs

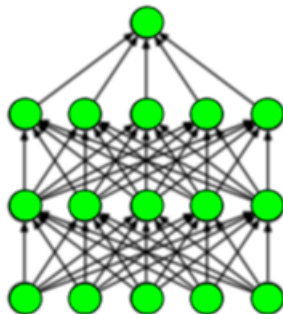
- 1 A DNN model with 100 million parameters and only 100k data points or even a million data points will overfit unless we take care of over-fitting.
- 2 Weight regularization can help -  $\ell_1, \ell_2$
- 3 More common over-fitting strategy for DL?
- 4 Dropouts!
- 5 Early stopping is also a great strategy! Stop training the DL model when the validation error starts increasing. How's this different from regular validation we were doing earlier??

# Over-fitting in DNNs

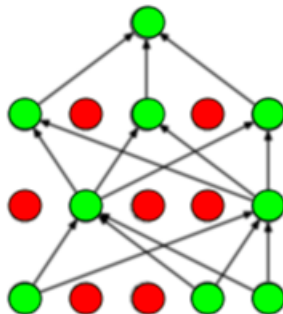
## How to handle over-fitting in DNNs

- 1 A DNN model with 100 million parameters and only 100k data points or even a million data points will overfit unless we take care of over-fitting.
- 2 Weight regularization can help -  $\ell_1, \ell_2$
- 3 More common over-fitting strategy for DL?
- 4 Dropouts!
- 5 Early stopping is also a great strategy! Stop training the DL model when the validation error starts increasing. How's this different from regular validation we were doing earlier??
- 6 Book by Yoshua Bengio has tons of details and great reference for Deep Learning!

# Taking care of Over-fitting: Dropouts



(a) Standard Neural Net



(b) After applying dropout.

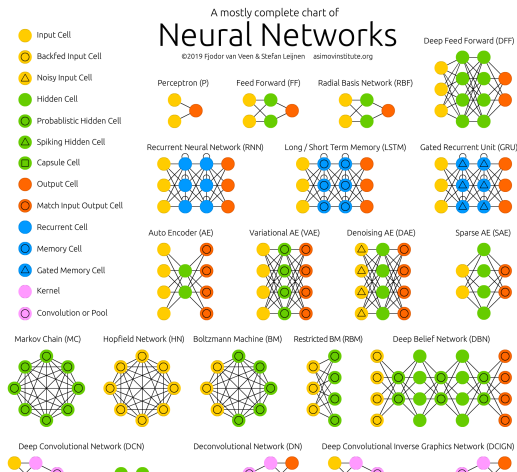
# Tensorflow Playground Demo

Tensorflow Playground Demo

# More DL Architectures

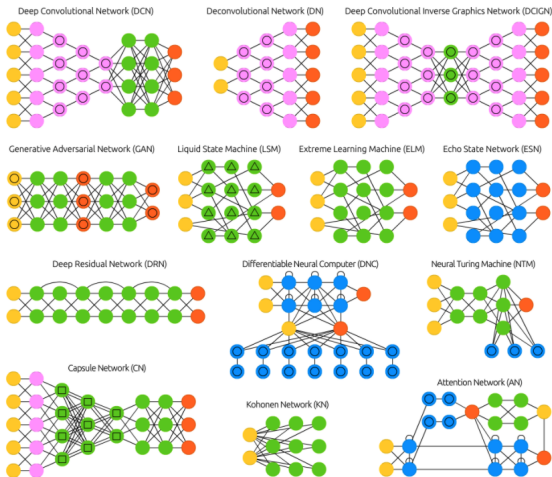
## Neural Networks Zoo

### Zoo Reference



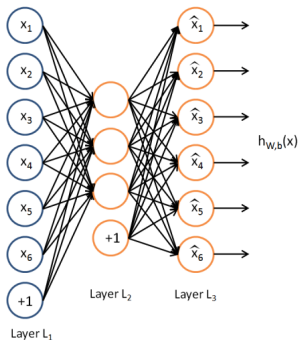
# More DL Architectures

## Neural Networks Zoo





# Auto Encoders



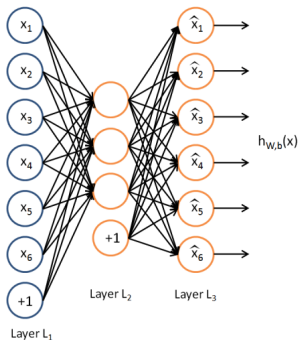
# ICE #4

## PCA vs Auto Encoder

Which of the following statements are true ?

- ① Both PCA and Auto Encoders serve the purpose of dimensionality reduction
- ② They are both linear models but one uses a neural nets architecture and the other is based on projections
- ③ PCA is robust to outliers while Auto Encoders are not
- ④ Auto Encoders are as better than Glove Embeddings to find low-dim embeddings for words

# PCA vs Auto-Encoders



# AutoEncoders and Dimensionality Reduction

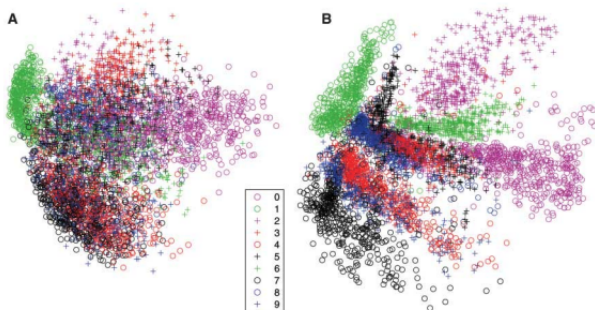
Visualization Performance

[Auto Encoder Reference Paper](#)

# AutoEncoders and Dimensionality Reduction

## Reading Reference for AE Dimensionality Reduction

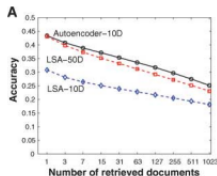
**Fig. 3.** (A) The two-dimensional codes for 500 digits of each class produced by taking the first two principal components of all 60,000 training images. (B) The two-dimensional codes found by a 784-1000-500-250-2 autoencoder. For an alternative visualization, see (B).



# AutoEncoders and Dimensionality Reduction

## Reading Reference for AE Dimensionality Reduction

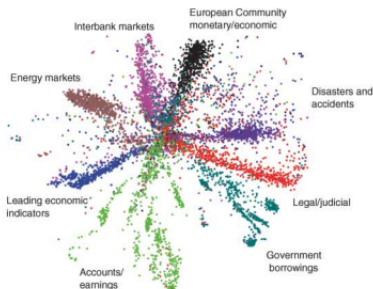
**Fig. 4.** (A) The fraction of retrieved documents in the same class as the query when a query document from the test set is used to retrieve other test set documents, averaged over all 402,207 possible queries. (B) The codes produced by two-dimensional LSA. (C) The codes produced by a 2000-500-250-125-2 autoencoder.



**B**



**C**



# AutoEncoders Summary

- ① Auto-Encoders are a method for dimensionality reduction and can do better than PCA for visualization

# AutoEncoders Summary

- ① Auto-Encoders are a method for dimensionality reduction and can do better than PCA for visualization
- ② Use Neural Networks architecture and hence can encode non-linearity in the embeddings



# AutoEncoders Summary

- ① Auto-Encoders are a method for dimensionality reduction and can do better than PCA for visualization
- ② Use Neural Networks architecture and hence can encode non-linearity in the embeddings
- ③ Anything else?

# AutoEncoders Summary

- 1 Auto-Encoders are a method for dimensionality reduction and can do better than PCA for visualization
- 2 Use Neural Networks architecture and hence can encode non-linearity in the embeddings
- 3 Anything else?
- 4 Auto Encoders can learn convolutional layers instead of dense layers - Better for images! More flexibility!!

# Removing obstacles in images

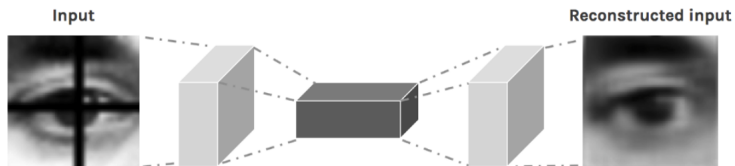


Figure 12: Reconstructed image from missing image [14]

# Removing obstacles in images

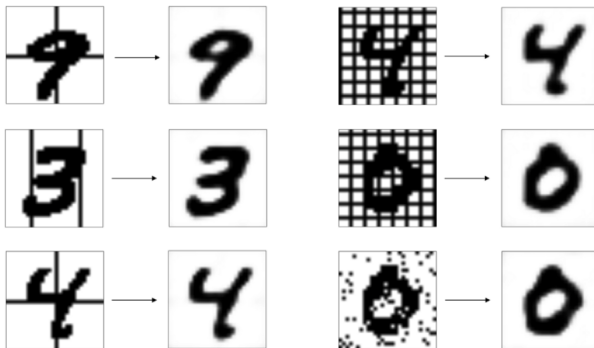


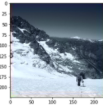
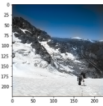



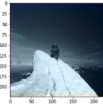
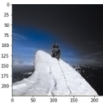



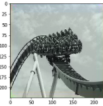
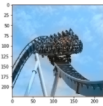

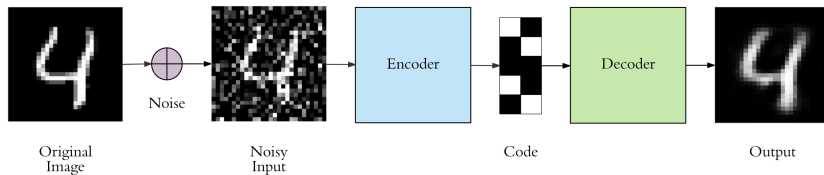


Figure 13: Source [15]

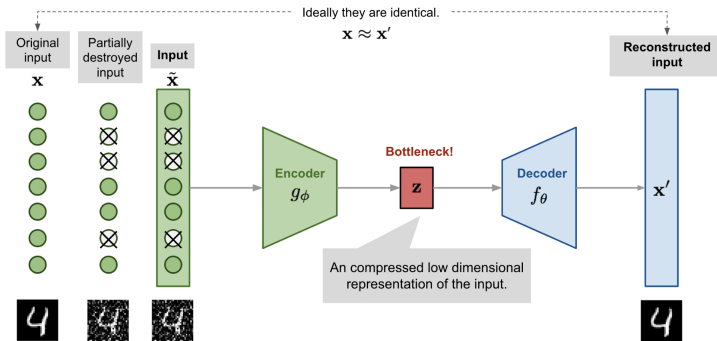
# Coloring Images

Gray Image	Vanilla Autoencoder	Merge Model (YCbCr)	Merge Model (LAB)	Original
				
				
				

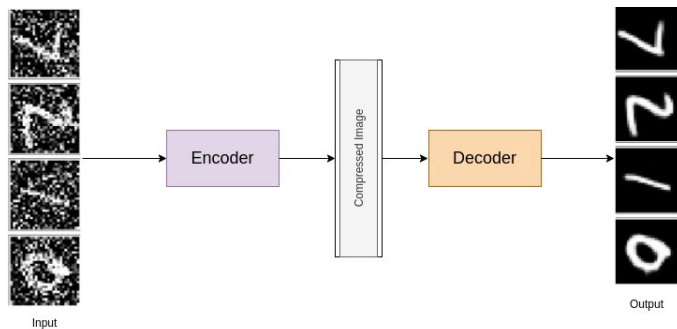
# De-noising Auto Encoders



# De-noising Auto Encoders



# De-noising Auto Encoders





# De-noising Auto Encoders

## Details

- Just like an Auto Encoder

# De-noising Auto Encoders

## Details

- Just like an Auto Encoder
- Difference: Noise is injected in the inputs on purpose but output is a clean data point.

# De-noising Auto Encoders

## Details

- Just like an Auto Encoder
- Difference: Noise is injected in the inputs on purpose but output is a clean data point.
- This forces the Auto Encoder to “de-noise” data, esp. useful for images!

# De-noising Auto Encoders

## Details

- Just like an Auto Encoder
- Difference: Noise is injected in the inputs on purpose but output is a clean data point.
- This forces the Auto Encoder to “de-noise” data, esp. useful for images!
- Esp. useful for a category of objects or images (e.g. digit recognition or face recognition, etc)

# ICE #5

## Unsupervised Learning

Which of these is NOT an example of unsupervised learning?

- ① Perceptron
- ② Auto Encoder
- ③ De-noising Auto Encoder
- ④ K-means++
- ⑤ None of the above
- ⑥ All of the above

# Breakouts Time 1

5 mins

Discuss in your groups what are some real-world applications of any or many of the Auto Encoder Architectures we discussed so far you can think of in your area of work or in a standard context e.g. images.

# Sequence structure in NLP

## Example

I love this car! Positive Sentiment

# Sequence structure in NLP

## Example

I love this car! Positive Sentiment

## Example

I am not sure I love this car! Negative Sentiment



# Sequence structure in NLP

## Example

I love this car! Positive Sentiment

## Example

I am not sure I love this car! Negative Sentiment

## Example

I don't think its a bad car at all! → Positive Sentiment

# Sequence structure in NLP

## Example

I love this car! Positive Sentiment

## Example

I am not sure I love this car! Negative Sentiment

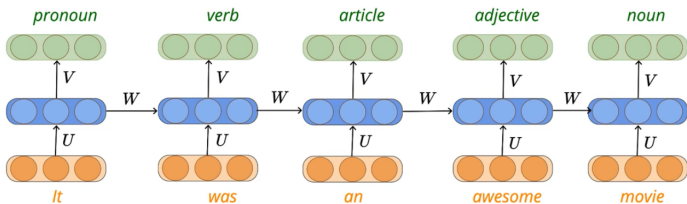
## Example

I don't think its a bad car at all! → Positive Sentiment

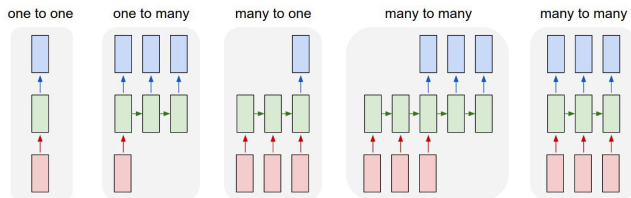
## Example

Have to carry the **context(state)** from some-time back to fully understand what's happening!

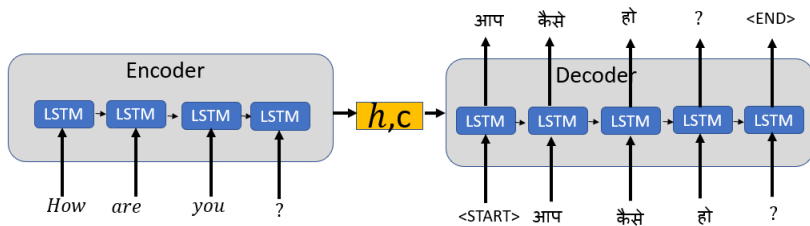
# Sequence to Sequence Model (LSTM) Applications



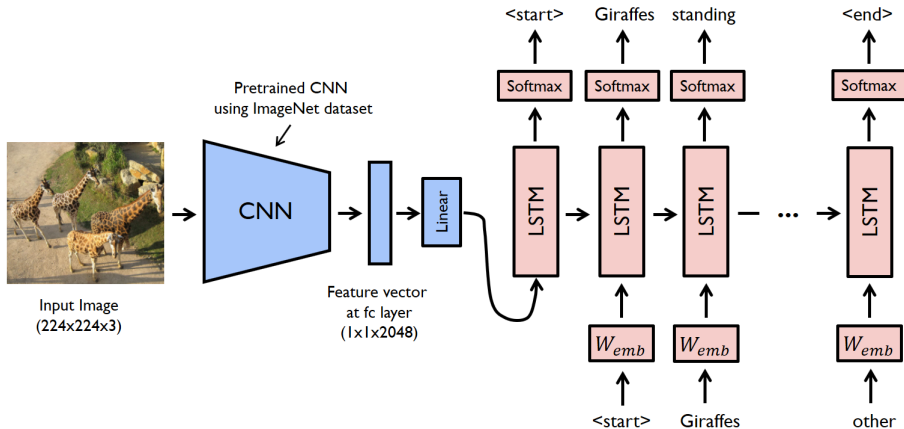
# Sequence to Sequence Model (LSTM) Applications



# Sequence to Sequence Model (LSTM) Applications



# Sequence to Sequence Model (LSTM) Applications



# Breakouts Time #2

## Auto-complete — 5 mins

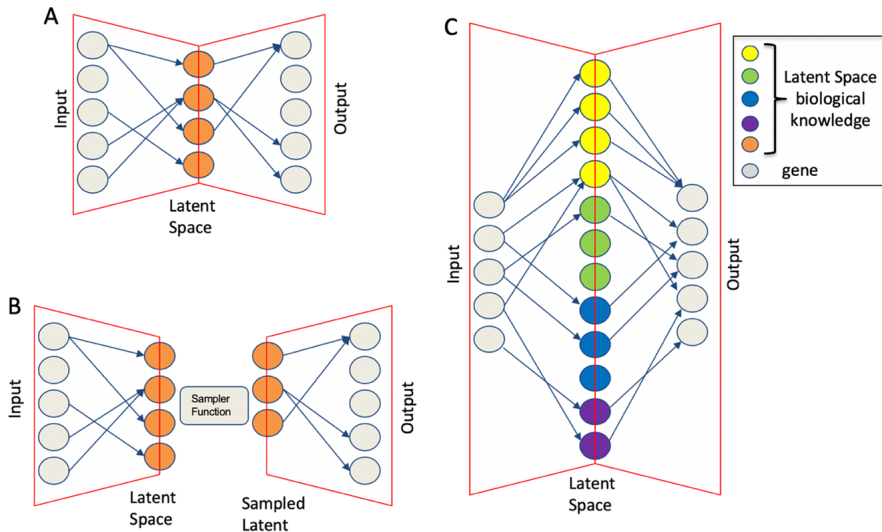
Let's say you are tasked with building an in-email auto-completion application, which can help complete partial sentences into full sentences through suggestions (auto-complete). How would you use what we have learned so far to model this? What architecture would you use? What would be your data? And what are some pitfalls or painpoints your model should address?

# Extra Slides



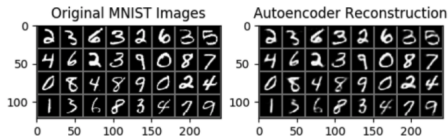
# Sparse Auto Encoders

## Sparse AE



# Sparse Auto Encoders

## Sparse AE Reference

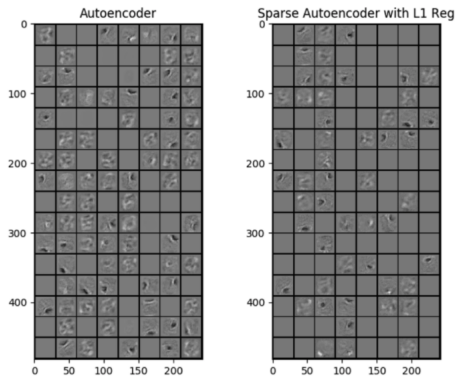


Methods	Best MSE Loss (MNIST or CIFAR-10)
Simple Autoencoder	0.0318 (MNIST)
Sparse Autoencoder (L1 reg)	0.0301 (MNIST)

Experiment Results

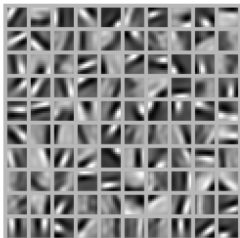
# Sparse Auto Encoders

## Sparse AE Reference



# Sparse Auto Encoders

Input Image that maximizes activations for each neuron in hidden layer!



# Sparse De-noising Auto Encoders

