Lecture 3A
Lightning Introduction to Keras/TF

Training a DL Model for a Structured Data Problem

15.S04: Hands-on Deep Learning
Spring 2024
Farias, Ramakrishnan

MANAGEMENT
SLOAN SCHOOL

Recap: Minimizing a (multi-variate) function

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

Minimize $g(w)$

$$\begin{bmatrix} \partial g / \partial w_1 \\ \partial g / \partial w_2 \end{bmatrix}$$

1) Start at some point w^0

2) Compute the gradient of the function at that point $\nabla g(w)$

3) Move in the direction of the gradient

$$w \leftarrow w - \alpha \nabla g(w)$$

Minimizing a loss function

Minimize $$\frac{1}{n}\sum_{i=1}^{n} -y^i \log\left(model(x^i)\right) - (1-y^i)\log(1-model(x^i))$$

$model(w; x^i)$

1) Start at some point w^0

2) Compute the gradient of the function at that point $\nabla g(w)$

3) Move in the direction of the gradient

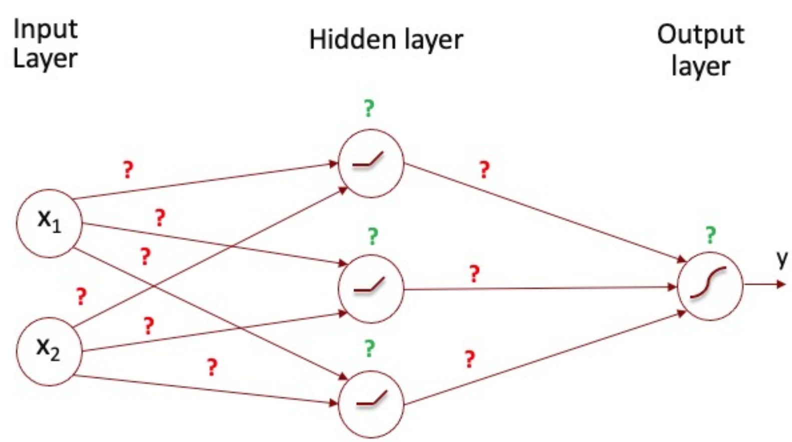
$$w \leftarrow w - \alpha \nabla g(w)$$

Minimizing a loss function

Minimize $$\frac{1}{n}\sum_{i=1}^{n} -y^i \log\left(model(x^i)\right) - (1-y^i)\log(1-model(x^i))$$

$$model(x^i) = \frac{1}{1+e^{-(w_1+w_2 max(0,w_3+w_4x_1^i+w_5x_2^i)+w_6 max(0,w_7+w_8x_1^i+w_9x_2^i)+w_{10} max(0,w_{11}+w_{12}x_1^i+w_{13}x_2^i))}}$$

Recall this model and the NN it represents



Gradient Descent → Stochastic Gradient Descent

Making Gradient Descent work with large datasets

- Problem: For large datasets (e.g., *n* in the millions), computing the gradient of the loss function can be very expensive

$$\frac{1}{n}\sum_{i=1}^{n} -y^i \log\left(model(x^i)\right) - (1 - y^i)\log(1 - model(x^i))$$

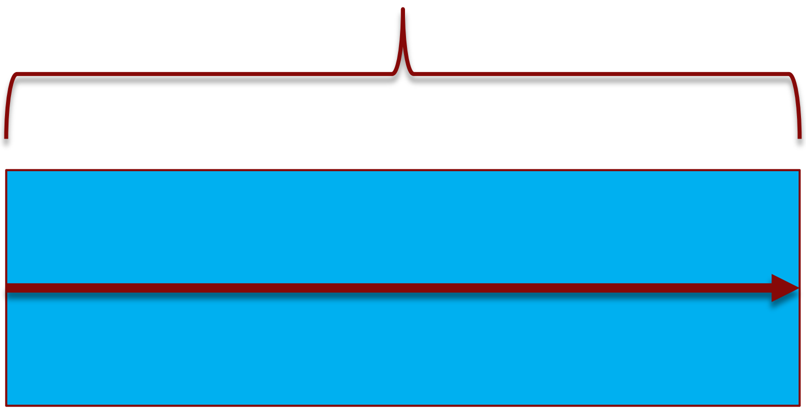
Making Gradient Descent work with large datasets

- Problem: For large datasets (e.g., *n* in the millions), computing the gradient of the loss function can be very expensive

- The Solution:
  - At each iteration, instead of using all the n data points in the calculation of the gradient of the loss function, randomly choose just a few of the n observations (called a *minibatch*) and use only these observations to compute the partial derivatives.

The Gradient Descent ‘Epoch’

Training set

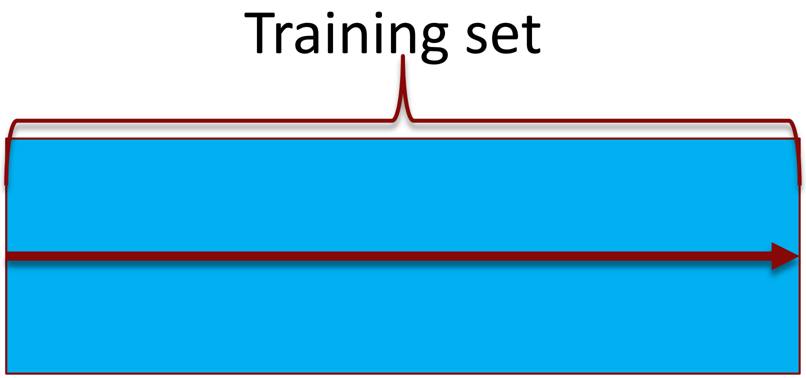


$$\frac{1}{n}\sum_{i=1}^{n} -y^i \log\left(model(x^i)\right) - (1-y^i)\log(1-model(x^i))$$

An epoch is one pass through the full training set.

But this plays out differently for Gradient Descent vs *Stochastic* Gradient Descent.

An epoch in Gradient Descent



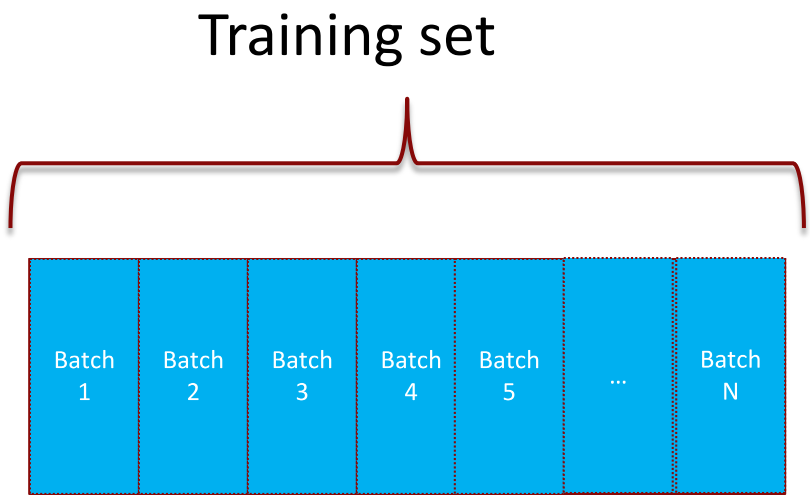
$$\frac{1}{n}\sum_{i=1}^{n} -y^i \log\left(model\left(x^i\right)\right) - \left(1-y^i\right)\log\left(1-model\left(x^i\right)\right)$$

- We run every training sample through the network to get the predictions
- We calculate the gradient of the loss
- We update the parameters

$$\boxed{w \leftarrow w - \alpha \frac{dLoss(w)}{dw}}$$

This is done just once at the end of the epoch

An epoch in Stochastic Gradient Descent



But when we do Stochastic Gradient Descent (SGD), we process the data in minibatches, one after the other*

*we will refer to minibatches as batches from now on for simplicity

An epoch in Stochastic Gradient Descent

Training set

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5	…	Batch N

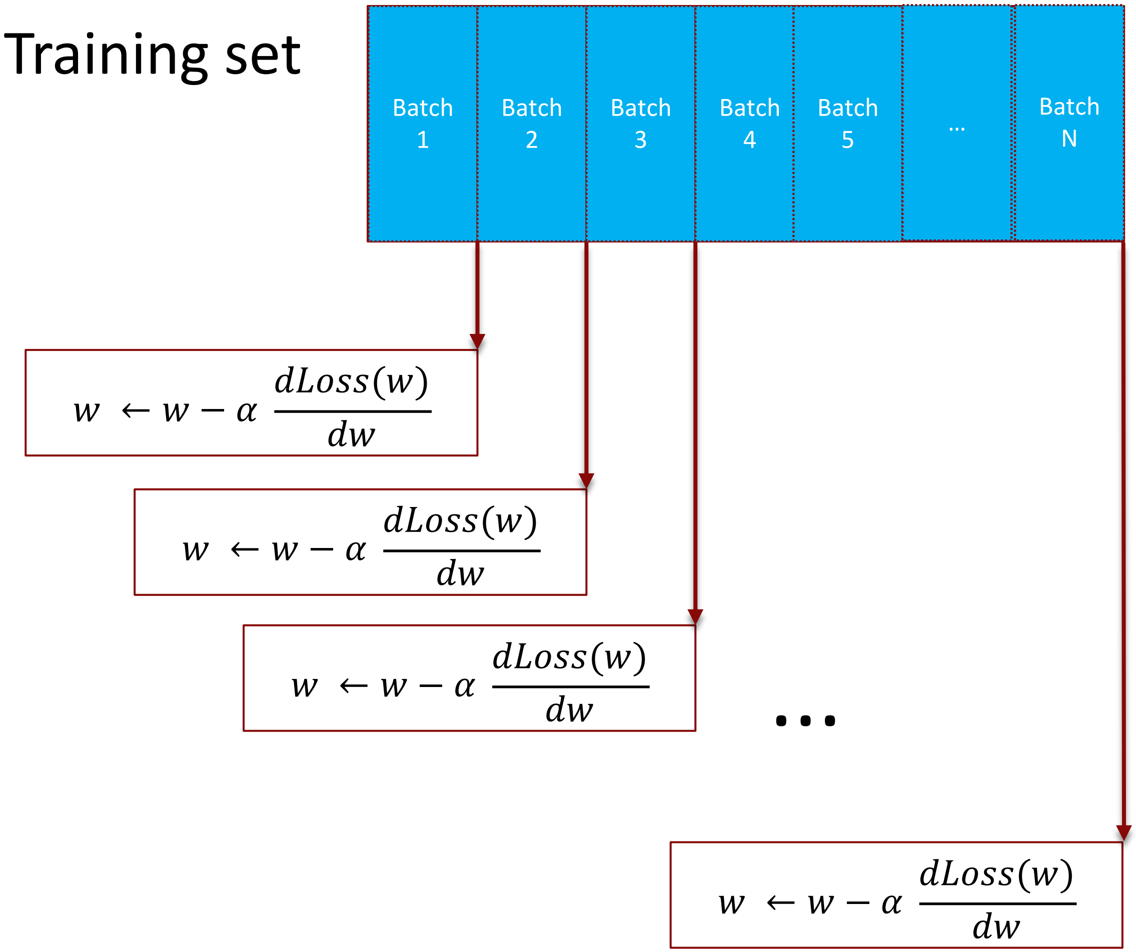
For each batch:

- *We run the training samples in that batch through the network to get predictions*
- *We calculate the gradient of the loss*
- *We update the parameters*

$$\boxed{w \leftarrow w - \alpha \frac{dLoss(w)}{dw}}$$

An epoch in <u>Stochastic</u> Gradient Descent

Training set



Lightning Intro to Tensorflow/Keras

What's a Tensor?

What’s a Tensor?

Tensor of rank 0 (Scalar)

42

What’s a Tensor?

Tensor of rank 0 (Scalar)

42

Tensor of rank 1 (aka Vector)

(42, 23.4, 11.2)

What's a Tensor?

Tensor of rank 0 (Scalar)

42

Tensor of rank 2 (aka Matrix)

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]	[,28]
[1,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[3,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	0	0	0	0	0	0	0	3	18	18	18	126	136	175	26	166	255	247	127	0	0	0	0
[7,]	0	0	0	0	0	0	0	0	30	36	94	154	170	253	253	253	253	253	225	172	253	242	195	64	0	0	0	0
[8,]	0	0	0	0	0	0	0	49	238	253	253	253	253	253	253	253	253	251	93	82	82	56	39	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	18	219	253	253	253	253	253	198	182	247	241	0	0	0	0	0	0	0	0	0	0
[10,]	0	0	0	0	0	0	0	0	80	156	107	253	253	205	11	0	43	154	0	0	0	0	0	0	0	0	0	0
[11,]	0	0	0	0	0	0	0	0	0	14	1	154	253	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[12,]	0	0	0	0	0	0	0	0	0	0	0	139	253	190	2	0	0	0	0	0	0	0	0	0	0	0	0	0
[13,]	0	0	0	0	0	0	0	0	0	0	0	11	190	253	70	0	0	0	0	0	0	0	0	0	0	0	0	0
[14,]	0	0	0	0	0	0	0	0	0	0	0	0	35	241	225	160	108	1	0	0	0	0	0	0	0	0	0	0
[15,]	0	0	0	0	0	0	0	0	0	0	0	0	0	81	240	253	253	119	25	0	0	0	0	0	0	0	0	0
[16,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	186	253	253	150	27	0	0	0	0	0	0	0	0
[17,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	93	252	253	187	0	0	0	0	0	0	0	0
[18,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	249	253	249	64	0	0	0	0	0	0	0
[19,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	130	183	253	253	207	2	0	0	0	0	0	0	0
[20,]	0	0	0	0	0	0	0	0	0	0	0	0	39	148	229	253	253	253	250	182	0	0	0	0	0	0	0	0
[21,]	0	0	0	0	0	0	0	0	0	0	24	114	221	253	253	253	253	201	78	0	0	0	0	0	0	0	0	0
[22,]	0	0	0	0	0	0	0	0	23	66	213	253	253	253	253	198	81	2	0	0	0	0	0	0	0	0	0	0
[23,]	0	0	0	0	0	0	18	171	219	253	253	253	253	195	80	9	0	0	0	0	0	0	0	0	0	0	0	0
[24,]	0	0	0	0	55	172	226	253	253	253	253	244	133	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[25,]	0	0	0	0	136	253	253	253	212	135	132	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[26,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[27,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[28,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tensor of rank 1 (aka Vector)

(42, 23.4, 11.2)

What’s a Tensor?

Tensor of rank 0 (Scalar)

42

Tensor of rank 1 (aka Vector)

(42, 23.4, 11.2)

Tensor of rank 2 (aka Matrix)

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]	[,28]
[1,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[3,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	0	0	0	0	0	0	0	3	18	18	18	126	136	175	26	166	255	247	127	0	0	0	0
[7,]	0	0	0	0	0	0	0	0	30	36	94	154	170	253	253	253	253	253	225	172	253	242	195	64	0	0	0	0
[8,]	0	0	0	0	0	0	0	49	238	253	253	253	253	253	253	253	253	251	93	82	82	56	39	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	18	219	253	253	253	253	253	198	182	247	241	0	0	0	0	0	0	0	0	0	0
[10,]	0	0	0	0	0	0	0	0	80	156	107	253	253	205	11	0	43	154	0	0	0	0	0	0	0	0	0	0
[11,]	0	0	0	0	0	0	0	0	0	14	1	154	253	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[12,]	0	0	0	0	0	0	0	0	0	0	0	139	253	190	2	0	0	0	0	0	0	0	0	0	0	0	0	0
[13,]	0	0	0	0	0	0	0	0	0	0	0	11	190	253	70	0	0	0	0	0	0	0	0	0	0	0	0	0
[14,]	0	0	0	0	0	0	0	0	0	0	0	0	35	241	225	160	108	1	0	0	0	0	0	0	0	0	0	0
[15,]	0	0	0	0	0	0	0	0	0	0	0	0	0	81	240	253	253	119	25	0	0	0	0	0	0	0	0	0
[16,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	186	253	253	150	27	0	0	0	0	0	0	0	0
[17,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	93	252	253	187	0	0	0	0	0	0	0	0
[18,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	249	253	249	64	0	0	0	0	0	0	0
[19,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	130	183	253	253	207	2	0	0	0	0	0	0	0
[20,]	0	0	0	0	0	0	0	0	0	0	0	0	39	148	229	253	253	253	250	182	0	0	0	0	0	0	0	0
[21,]	0	0	0	0	0	0	0	0	0	0	24	114	221	253	253	253	253	201	78	0	0	0	0	0	0	0	0	0
[22,]	0	0	0	0	0	0	0	0	23	66	213	253	253	253	253	198	81	2	0	0	0	0	0	0	0	0	0	0
[23,]	0	0	0	0	0	0	18	171	219	253	253	253	253	195	80	9	0	0	0	0	0	0	0	0	0	0	0	0
[24,]	0	0	0	0	55	172	226	253	253	253	253	244	133	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[25,]	0	0	0	0	136	253	253	253	212	135	132	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[26,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[27,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[28,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tensor of rank 3 (aka “cube”)

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	147	131	138	144	131	134	144	135	133	145

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	251	232	233	237	230	243	255	255	250	246
[2,]	248	234	239	245	238	246	255	251	246	243
[3,]	255	241	238	236	229	241	253	249	238	234
[4,]	255	252	243	233	228	237	242	234	218	205
[5,]	255	255	249	231	228	231	224	215	204	166
[6,]	255	255	230	192	189	202	205	205	204	147
[7,]	231	231	188	140	138	152	156	159	177	136
[8,]	155	172	149	114	113	111	93	82	119	115
[9,]	107	130	108	93	113	100	67	66	81	95
[10,]	84	104	90	69	69	61	52	63	59	46

Can you give an example of a rank-4 tensor?

What’s a Tensor?

See Chapter 2.2 of text for more detail

Tensorflow

Tensorflow (TF) is a library that provides

- Automatic calculation of the gradient of (complicated) loss functions and models

Tensorflow

Tensorflow (TF) is a library that provides

- <u>Automatic</u> calculation of the gradient of (complicated) loss functions and models
- Library of state-of-the-art optimizers

Tensorflow

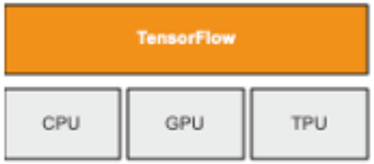
Tensorflow (TF) is a library that provides

- Automatic calculation of the gradient of (complicated) loss functions and models
- Library of state-of-the-art optimizers
- Automatic distribution of computational load across servers

Tensorflow

Tensorflow (TF) is a library that provides

- Automatic calculation of the gradient of (complicated) loss functions and models
- Library of state-of-the-art optimizers
- Automatic distribution of computational load across servers
- Automatic adaptation of code to work on parallel hardware (GPUs and TPUs)



Keras "sits on top of" Tensorflow …

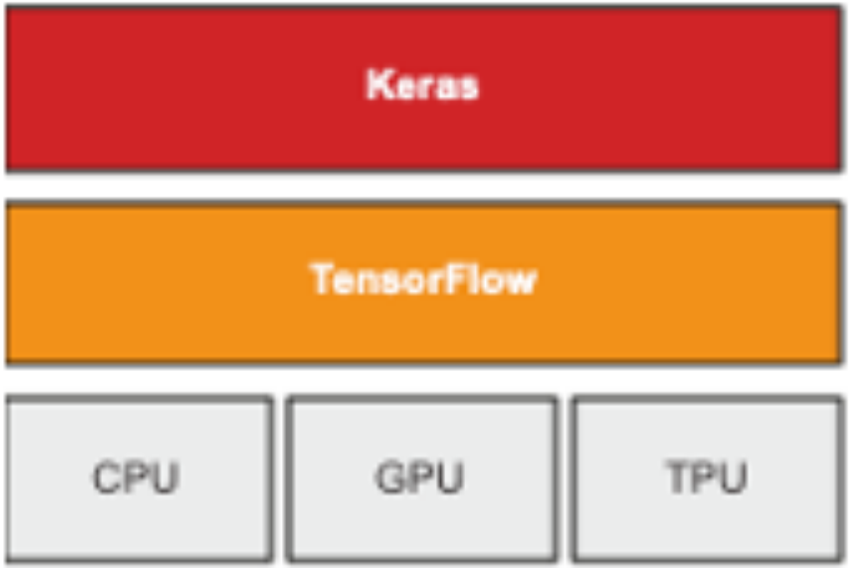


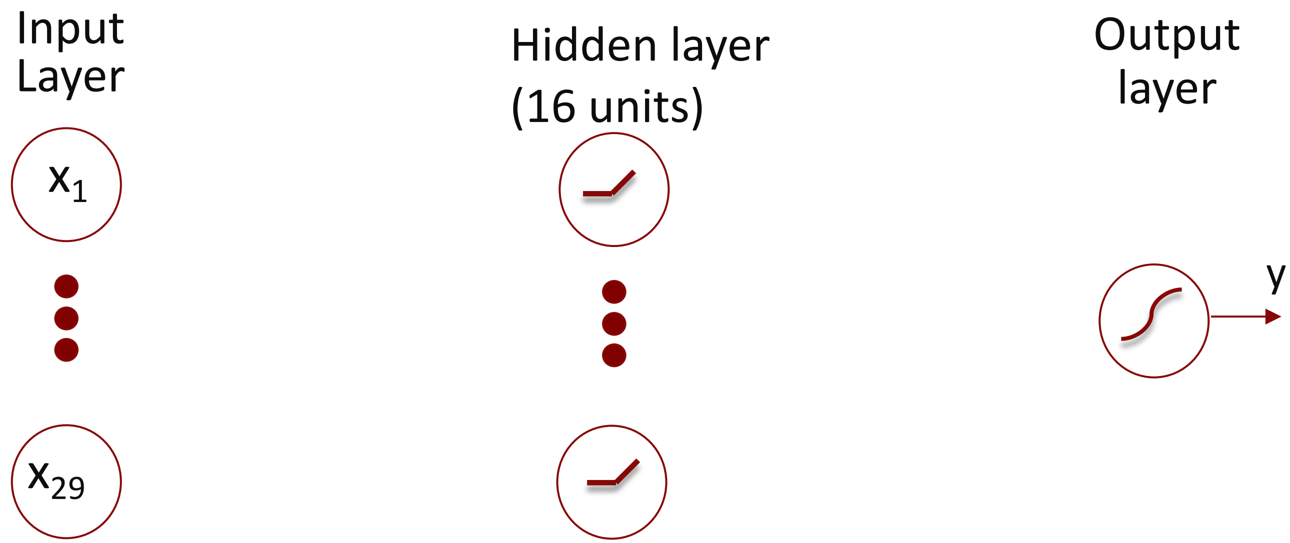
Image: Page 70 of textbook

… and provides "convenience" features

- Pre-defined layers
- Incredibly flexible ways to specify network architectures
- Easy ways to preprocess data
- Easy ways to train models and report metrics
- Pre-trained models you can download and customize

Check out the wealth of introductory and advanced material, with accompanying colabs, at tensorflow.org and keras.io

Let’s revisit the Heart Disease Prediction Model we defined earlier




```
input = keras.Input(shape=29)
h = keras.layers.Dense(16, activation=“relu”)(input)
output = keras.layers.Dense(1, activation="sigmoid”)(h)
model = keras.Model(input, output)
```

Training Checklist

- We get the data ready (will cover in the colab)

- We design i.e., "lay out" the network 1 hidden layer with 16 ReLU neurons
  - Choose ***the number of hidden layers*** and ***the number of 'neurons' in each layer***
  - Pick the ***right output layer*** based on the type of the output Sigmoid

- We pick
  - An appropriate ***loss function*** based on the type of the output binary crossentropy
  - An ***optimizer from the many SGD flavors*** that are available"adam"

- We set things up in Keras/Tensorflow and start training!

Colab

[Predicting Heart Disease]

Before we start coding …

- Don't worry if you don't understand <u>every</u> detail of what we will do in class.

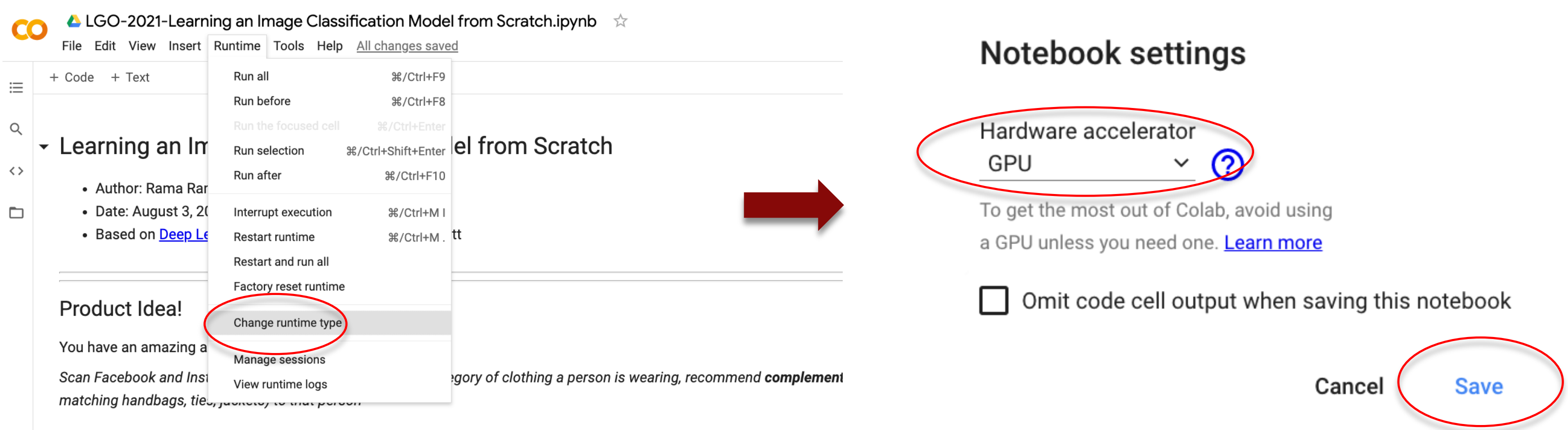
- But go through the Colab notebooks carefully later, play around with the code and make sure you understand <u>every</u> line

Colab Instructions

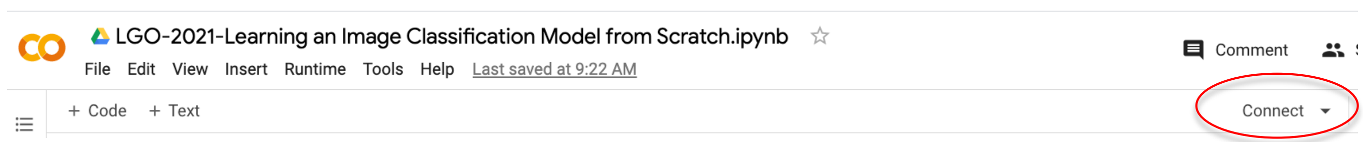
https://colab.research.google.com/drive/1bR9Tx87L4HxB94rlp-mhy4Aj4XJtpcGS?usp=sharing

Step 1 Make your own copy of the notebook

Step 2 Request a GPU for your notebook



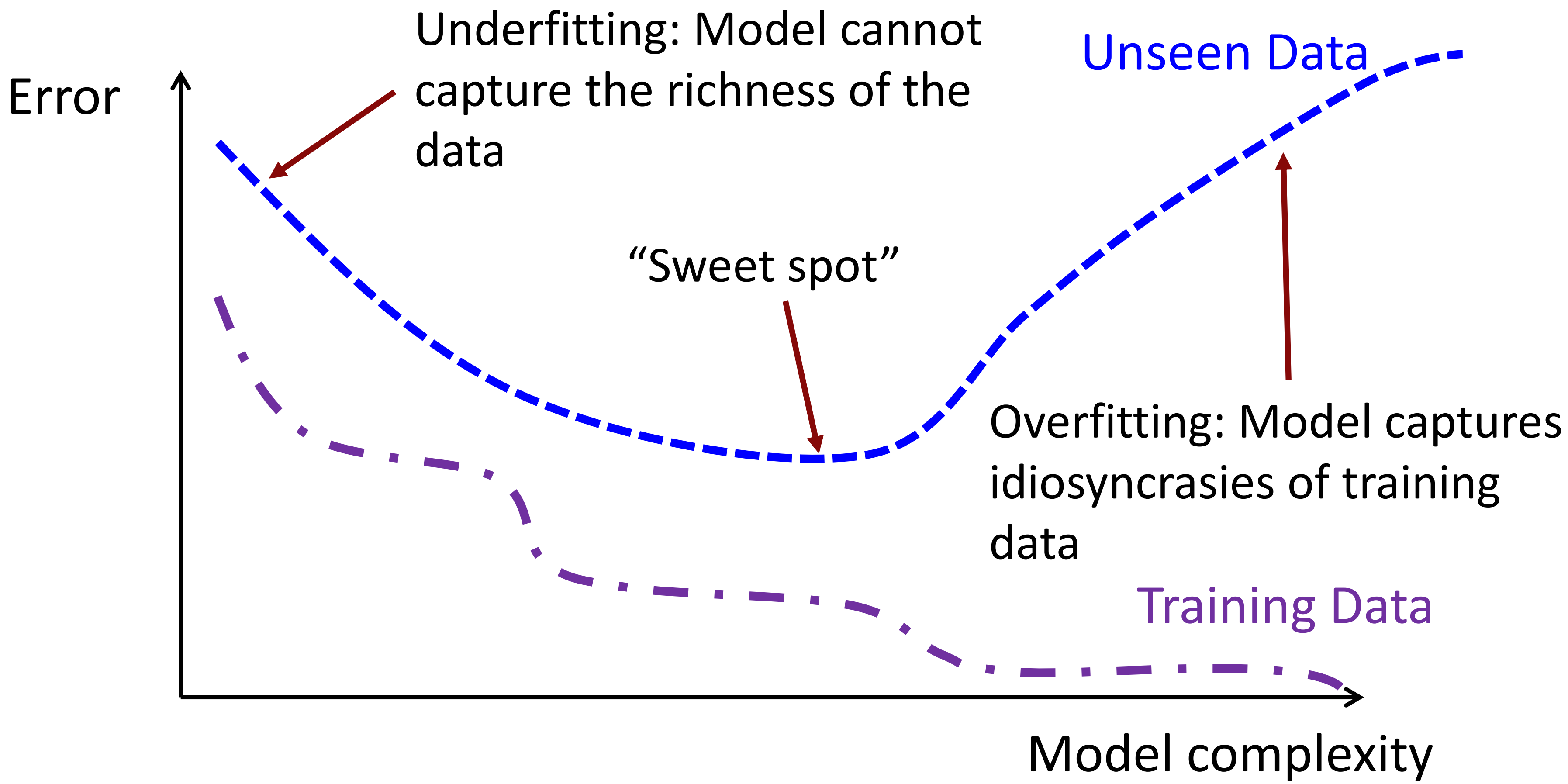
Step 3 Start your notebook



You need to do steps 1 and 2 just the first time you use a notebook. From the second time onwards, jump to Step 3.

Overfitting and Regularization

Recall Underfitting vs. Overfitting



Overfitting in Neural Networks

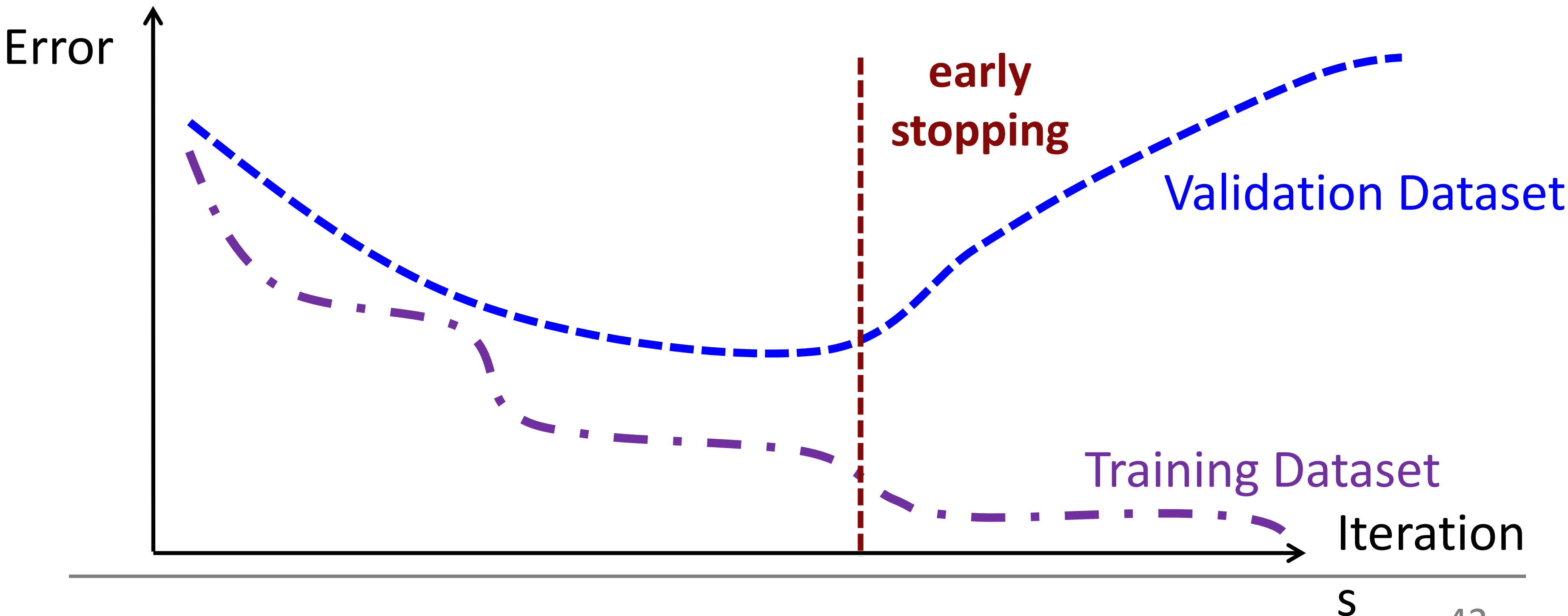
- To learn smart representations of complex, unstructured data, the NN needs to have large “capacity” i.e., many layers and many neurons in each layer

- But this raises the likelihood of overfitting so we need to add *regularization*

- Several regularization methods have been developed to address this problem

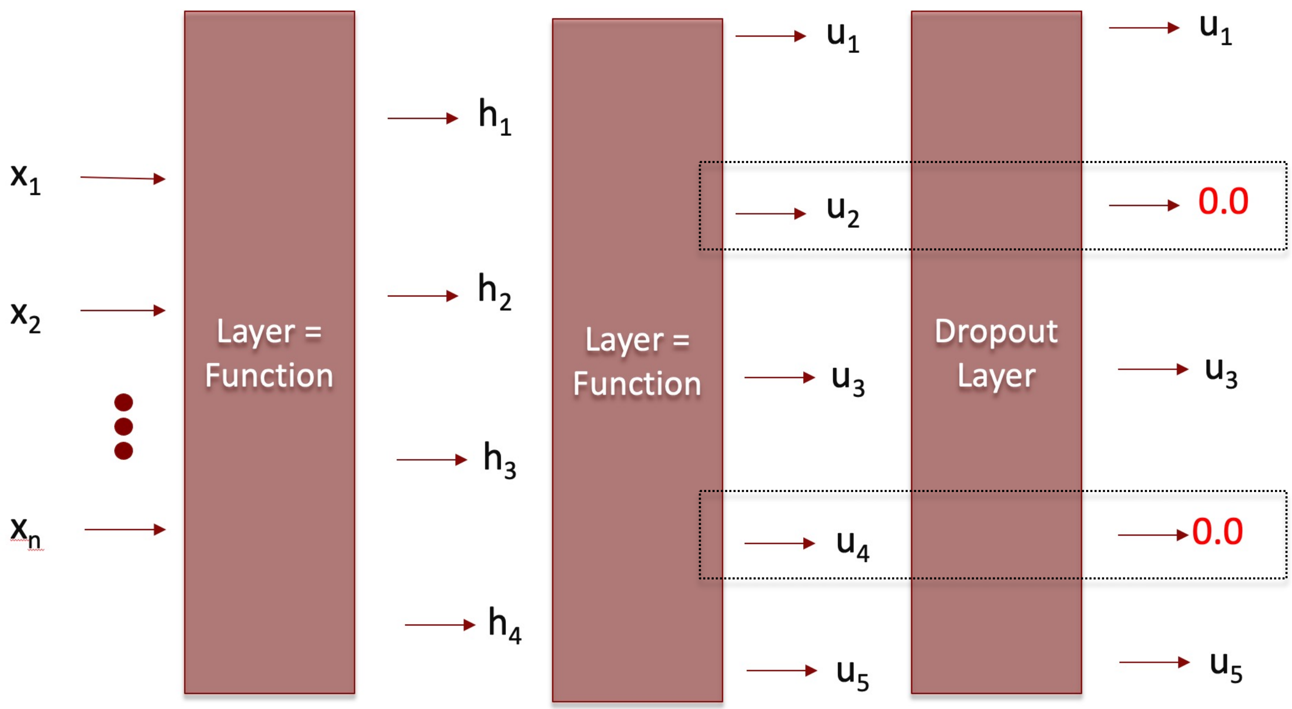
Regularization strategy: *Early Stopping*

Stop the training early before the training loss is minimized by monitoring the loss on a validation dataset.



Regularization strategy: *Dropout*

Randomly zero out the output from some of the nodes (typically 50% of the nodes) in a hidden layer (implemented as a “dropout layer” in Keras)



<“Bank teller” analogy>

Training Checklist

- We get the data ready (will cover in the colab)
- We design i.e., “lay out” the network 1 hidden layer with 16 ReLU neurons
  - Choose ***the number of hidden layers*** and ***the number of ‘neurons’ in each layer***
  - Pick the ***right output layer*** based on the type of the output Sigmoid
- We pick
  - An appropriate ***loss function*** based on the type of the output binary crossentropy
  - An ***optimizer from the many SGD flavors*** that are available“adam"
- We decide on a ***regularization strategy*** Early stopping
- We set things up in Keras/Tensorflow and start training!