Introduction to Deep Learning (I2DL)

Exercise 6: Hyperparameter Tuning

Today's Outline

1. *Review Solution Exercise 5*
   Sigmoid Activation Function
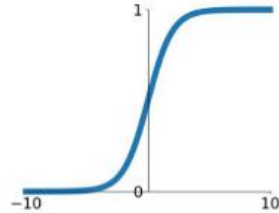

2. *Introduction Exercise 6*
   Hyperparameter Tuning
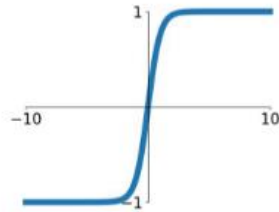
Activation functions

Activation functions

Sigmoid
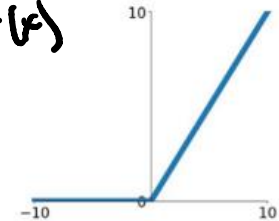
$\sigma(x) = \frac{1}{1+e^{-x}}$

tanh

$\tanh(x)$

$f'(x) = \frac{1}{\cosh^2(x)} = \operatorname{sech}^2(x)$
$\qquad \qquad \qquad = 1 - \tanh^2(x)$

ReLU

$\max(0, x)$

Leaky ReLU

$\max(0.1x, x)$

$f \nearrow^{0.2} \searrow_{1}$
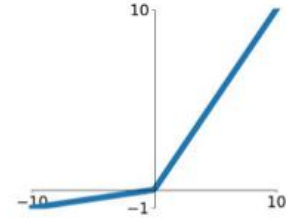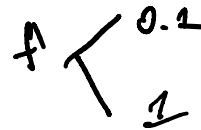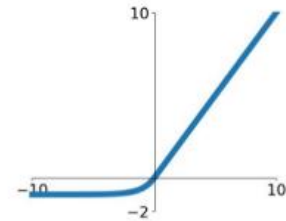
Maxout
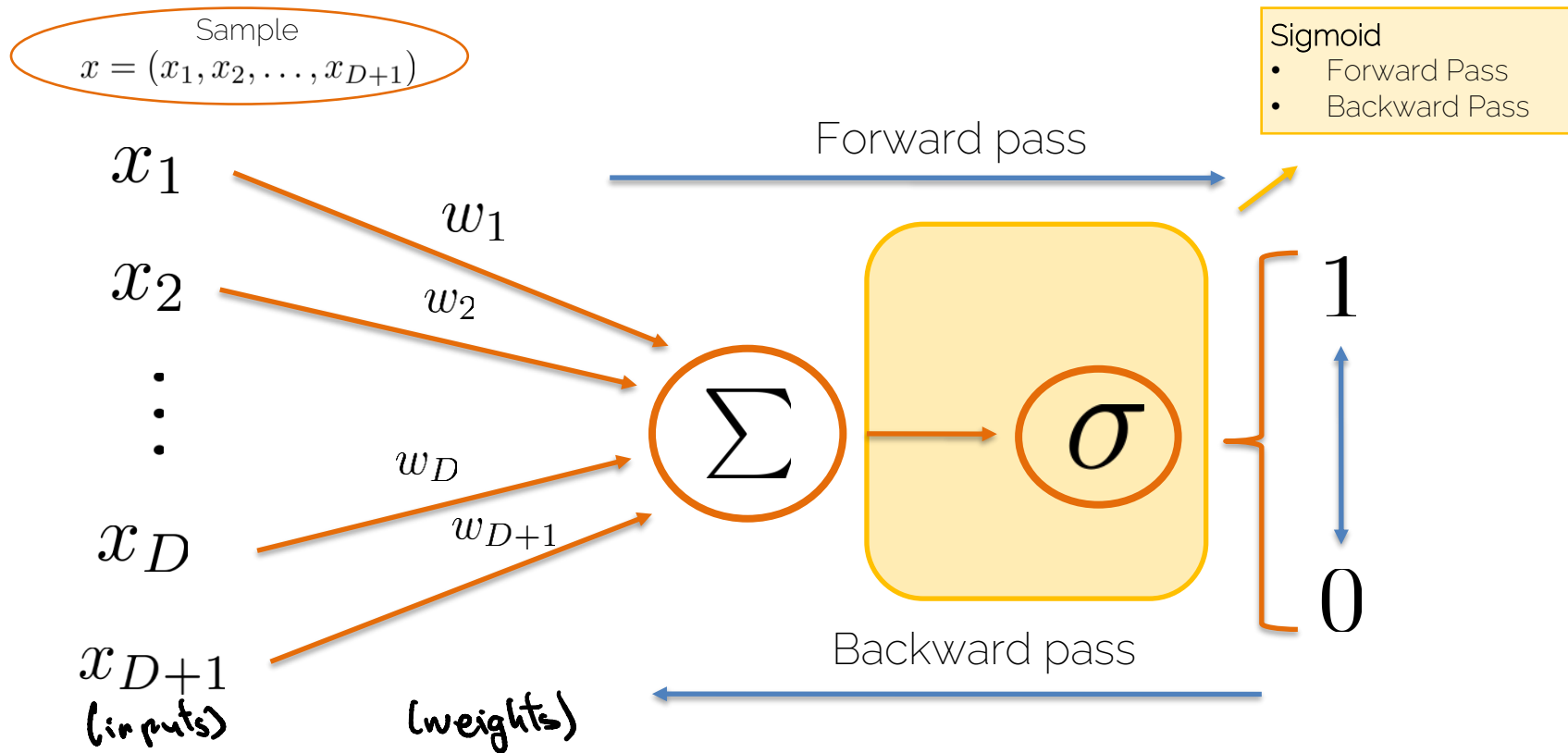
$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU

$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$

Activation function: Sigmoid

Sample
$$x = (x_1, x_2, \ldots, x_{D+1})$$

x_1

x_2

$\vdots$

x_D

x_{D+1}

(inputs)

w_1

w_2

w_D

w_{D+1}

(weights)

Forward pass

$\sum$

σ

1

0

Backward pass

Sigmoid
- Forward Pass
- Backward Pass

Sigmoid: Forward pass

- Definition of the Sigmoid function:

$$\sigma : \mathbb{R} \to \mathbb{R}, \sigma(x) = \frac{1}{1 + e^{-x}}$$
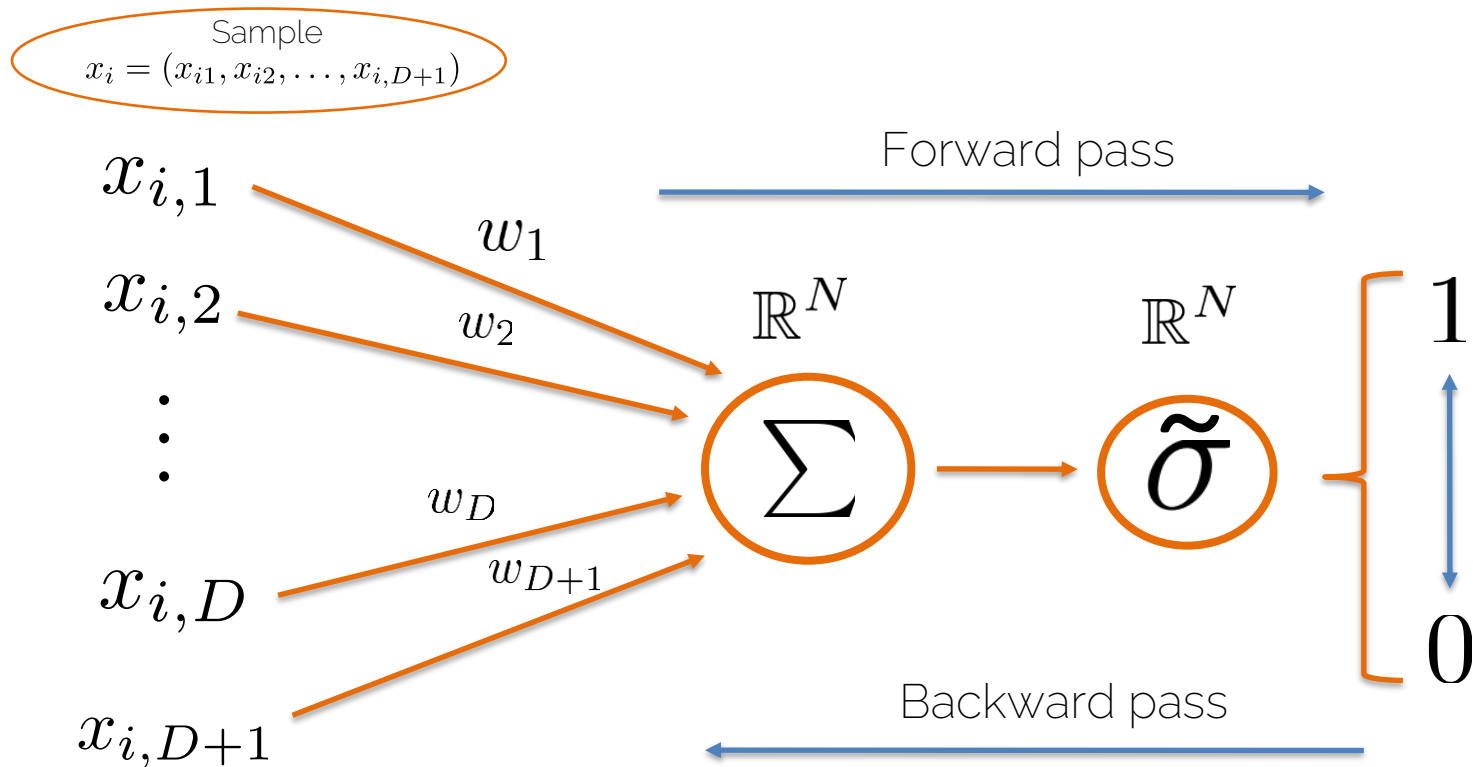
- Derivative of the sigmoid function:

$$\frac{\partial \sigma(x)}{\partial x} = \sigma(x) \cdot (1 - \sigma(x))$$

- Application of the Sigmoid function in higher dimension:

$$\tilde{\sigma} : \mathbb{R}^N \to \mathbb{R}^N, \tilde{\sigma}(x) = \begin{pmatrix} \sigma(x_1) \\ \sigma(x_2) \\ \vdots \\ \sigma(x_N) \end{pmatrix}$$
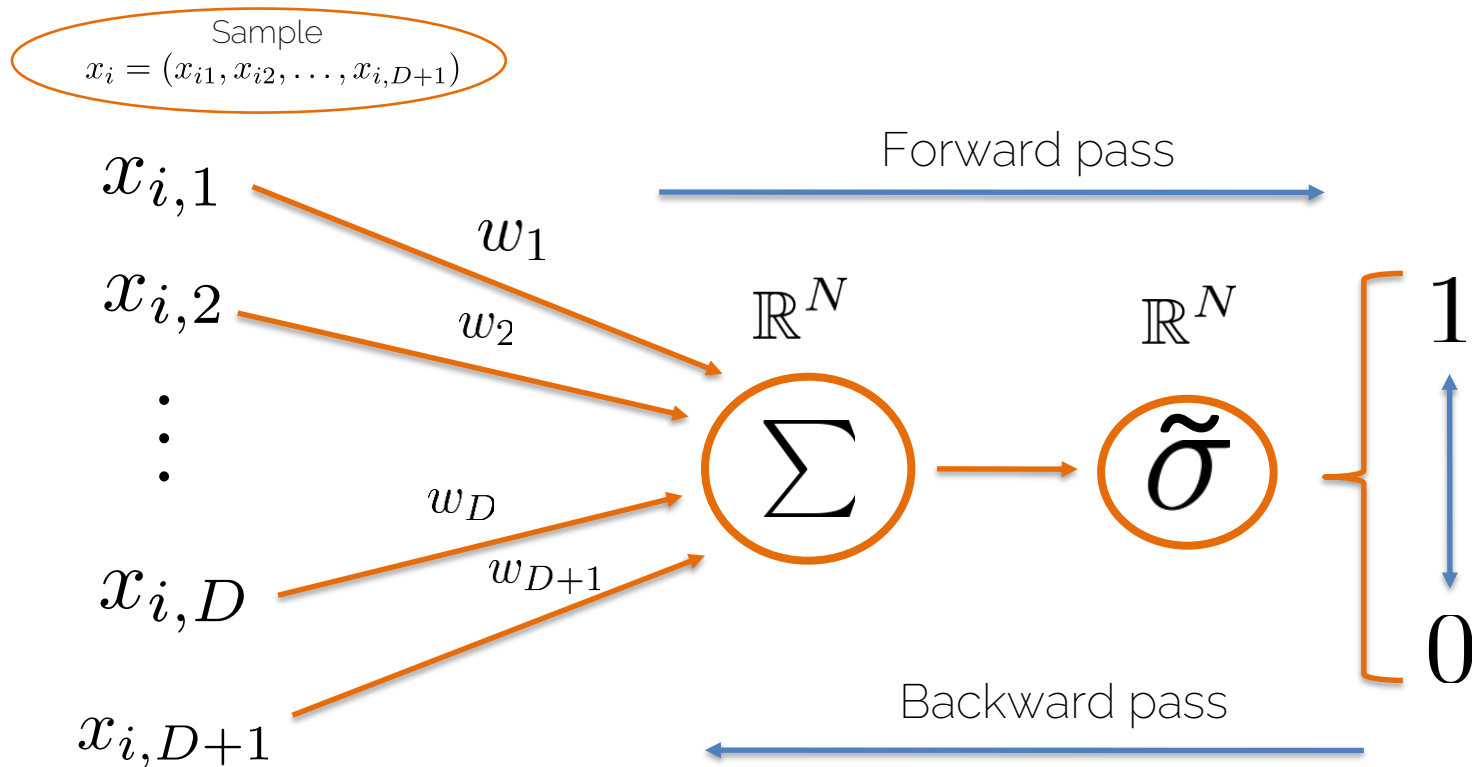
Activation function: Sigmoid

Sigmoid: Forward pass

```python
def forward(self, x):
    """
    :param x: Inputs, of any shape.

    :return out: Outputs, of the same shape as x.
    :return cache: Cache, stored for backward computation, of the same shape as x.
    """
    shape = x.shape
    out, cache = np.zeros(shape), np.zeros(shape)
    ########################################################################
    # TODO:                                                                #
    # Implement the forward pass of Sigmoid activation function            #
    ########################################################################
    # out = np.ones_like(x) / (np.ones_like(x) + np.exp(-x))
    out = 1 / (1 + np.exp(-x))
    cache = out
    ########################################################################
    #                        END OF YOUR CODE                              #
    ########################################################################
    return out, cache
```

$\longrightarrow$ x Dimensions

$\longrightarrow$ 1 Dimension

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Activation function: Sigmoid

Sample
$$x_i = (x_{i1}, x_{i2}, \ldots, x_{i,D+1})$$

$x_{i,1}$

$x_{i,2}$

$\vdots$

$x_{i,D}$

$x_{i,D+1}$

w_1

w_2

w_D

w_{D+1}

Forward pass

$\mathbb{R}^N$

$\mathbb{R}^N$

$\sum$

$\tilde{\sigma}$

1

0

Backward pass

Sigmoid: Backward pass

- The derivative of the sigmoid function is thus given a N x N - sized Jacobian matrix.

$$\tilde{\sigma} : \mathbb{R}^N \to \mathbb{R}^N, \tilde{\sigma}(x) = \begin{pmatrix} \sigma(x_1) \\ \sigma(x_2) \\ \vdots \\ \sigma(x_N) \end{pmatrix}$$

$$J_{\sigma} : \mathbb{R}^N \to \mathbb{R}^{N \times N}, J_{\sigma} = \begin{pmatrix} \frac{\partial \sigma(x_1)}{\partial x_1} & \frac{\partial \sigma(x_1)}{\partial x_2} & \cdots & \frac{\partial \sigma(x_1)}{\partial x_N} \\ \frac{\partial \sigma(x_2)}{\partial x_1} & \frac{\partial \sigma(x_2)}{\partial x_2} & \cdots & \frac{\partial \sigma(x_2)}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \sigma(x_N)}{\partial x_1} & \frac{\partial \sigma(x_N)}{\partial x_2} & \cdots & \frac{\partial \sigma(x_N)}{\partial x_N} \end{pmatrix} = \begin{pmatrix} \frac{\partial \sigma(x_1)}{\partial x_1} & 0 & \cdots & 0 \\ 0 & \frac{\partial \sigma(x_2)}{\partial x_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\partial \sigma(x_N)}{\partial x_N} \end{pmatrix}$$

Sigmoid: Backward pass

$$\frac{\partial \sigma(x)}{\partial x} = \sigma(x) \cdot (1 - \sigma(x))$$

```python
def backward(self, dout, cache):
    """

    :return: dx: the gradient w.r.t. input X, of the same shape as X
    """

    dx = None
    #######################################################################
    # TODO:                                                               #
    # Implement the backward pass of Sigmoid activation function          #
    #######################################################################
    dx = dout * cache * (1 - cache)
    #######################################################################
    #                          END OF YOUR CODE                           #
    #######################################################################
    return dx
```

$$J_\sigma = \begin{pmatrix} \frac{\partial \sigma(x_1)}{\partial x_1} & 0 & \cdots & 0 \\ 0 & \frac{\partial \sigma(x_2)}{\partial x_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\partial \sigma(x_N)}{\partial x_N} \end{pmatrix}$$

$$J_\sigma = \begin{pmatrix} \frac{\partial \sigma(x_1)}{\partial x_1} \\ \frac{\partial \sigma(x_2)}{\partial x_2} \\ \vdots \\ \frac{\partial \sigma(x_N)}{\partial x_N} \end{pmatrix}$$

On paper
- Cache is an N x 1 vector
- Derivative of Sigmoid is N x N matrix
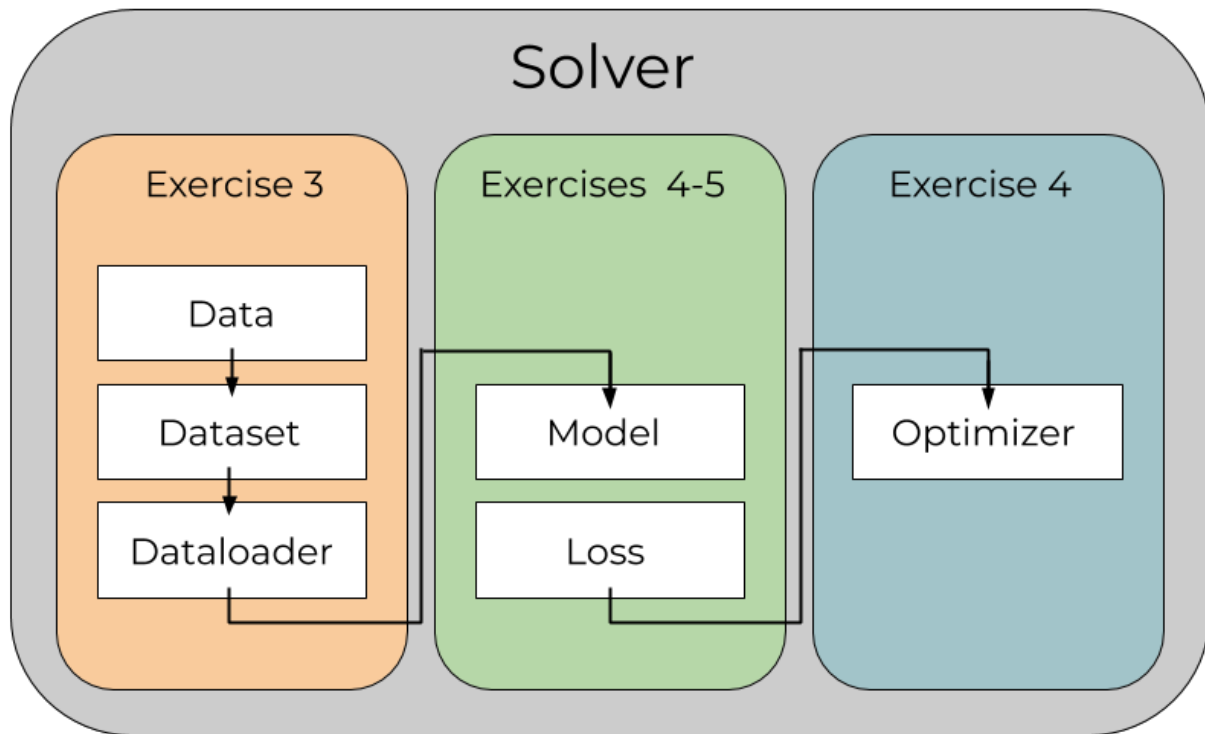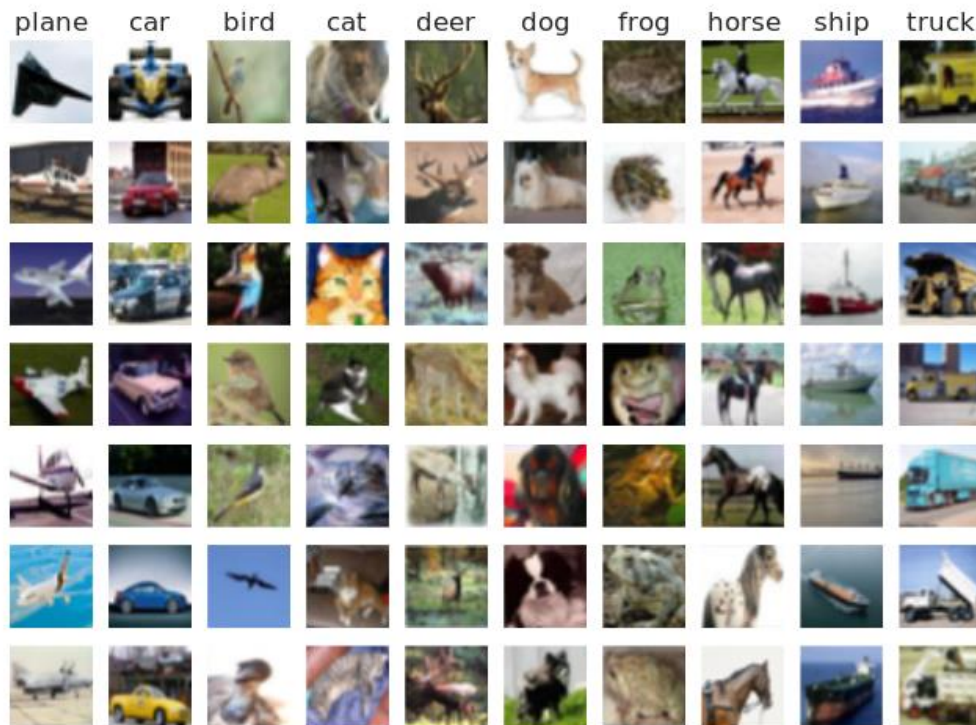- Multiplication is normal matrix multiplication

Numpy arrays
- Cache is a N x 1 vector
- Derivative of Sigmoid is given as N x 1 vector
- Multiplication: Numpy.multiply() which is componentwise multiplication

Exercise 6: Hyperparameter Tuning

Recap: Pillars of Deep Learning

Goal of exercise 6

Cifar10

Goal of exercise 6

- Use existing implementations
  - Reworked implementations of previous exercises
  - We will provide you with additional implementations of all required tools to run sample methods proposed in the lecture

- Learn about neural network debugging strategies and hyperparameter search

ONE DOES NOT SIMPLY

"PICK" HYPERPARAMETERS
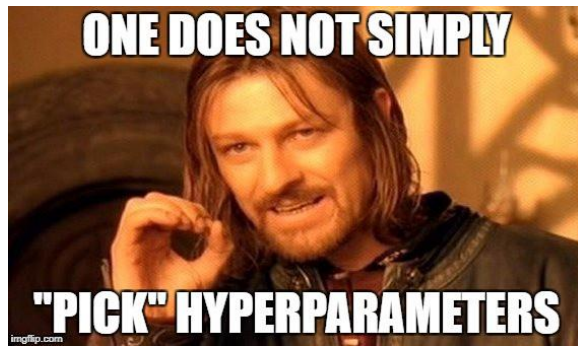
Leaderboard

- Your model's accuracy is all that counts!
  - At least 48% to pass the submission
  - There will be a leaderboard of all students!

Leaderboard

The leaderboard shows for each exercise the highest scoring submission from each user. Only valid submissions are displayed.

Exercise 1	Exercise 3	Exercise 4	Exercise 5	Exercise 6	Exercise 7	Exercise 8	Exercise 9	Exercise 10	Exercise 11

#	User	Score
1	a0008	100.00
2	a0001	100.00
3	a0003	100.00
4	u0306	100.00
5	u1540	100.00

Previously: Dataset

```python
class ImageFolderDataset(Dataset):
    """CIFAR-10 dataset class"""
    def __init__(self, transform=None, mode='train',
        limit_files=None,
        split={'train': 0.6, 'val': 0.2, 'test': 0.2},
        *args, **kwargs): ...

    @staticmethod
    def _find_classes(directory): ...

    def select_split(self, images, labels, mode): ...

    def make_dataset(self, directory, class_to_idx, mode): ...

    def __len__(self): ...

    @staticmethod
    def load_image_as_numpy(image_path): ...

    def __getitem__(self, index): ...
```

```python
# Create a train, validation and test dataset.
datasets = {}
for mode in ['train', 'val', 'test']:
    crt_dataset = ImageFolderDataset(
        mode=mode,
        root=cifar_root,
        download_url=download_url,
        transform=compose_transform,
        split={'train': 0.6, 'val': 0.2, 'test': 0.2}
    )
    datasets[mode] = crt_dataset
```

Previously: Data Loader

```python
class DataLoader:
    """
    Dataloader Class
    Defines an iterable batch-sampler over a given dataset
    """
    def __init__(self,
        dataset,
        batch_size=1,
        shuffle=False,
        drop_last=False): ...

    def __iter__(self): ...

    def __len__(self): ...
```

```python
# Create a dataloader for each split.
dataloaders = {}
for mode in ['train', 'val', 'test']:
    crt_dataloader = DataLoader(
        dataset=datasets[mode],
        batch_size=256,
        shuffle=True,
        drop_last=True,
    )
    dataloaders[mode] = crt_dataloader
```

Previously: Solver

```python
class Solver(object):
    """
    A Solver encapsulates all the logic necessary for training classification
    or regression models.
    The Solver performs gradient descent using the given learning rate.
    """

    def __init__(self, model, train_dataloader, val_dataloader,
        loss_func=CrossEntropyFromLogits(), learning_rate=1e-3,
        optimizer=Adam, verbose=True, print_every=1,
        lr_decay = 1.0, **kwargs): ...

    def _reset(self): ...

    def _step(self, X, y, validation=False): ...

    def train(self, epochs=100, patience = None): ...

    def get_dataset_accuracy(self, loader): ...

    def update_best_loss(self, val_loss, train_loss): ...
```

```python
solver = Solver(model,
                dataloaders['train'],
                dataloaders['val'],
                learning_rate=0.001,
                loss_func=MSE(),
                optimizer=SGD)

solver.train(epochs=epochs)
```

Previously: Classification Network

```python
class ClassificationNet(Network):
    """
    A fully-connected classification neural network with configurable
    activation function, number of layers, number of classes, hidden size and
    regularization strength.
    """

    def __init__(self,
        activation=Sigmoid(), num_layer=2,
        input_size=3 * 32 * 32, hidden_size=100,
        std=1e-3, num_classes=10, reg=0, **kwargs): ...

    def forward(self, X): ...

    def backward(self, dy): ...

    def save_model(self): ...

    def get_dataset_prediction(self, loader): ...
```

```python
# Instantiate a new model.
model = ClassificationNet(activation=Sigmoid(),
                          num_layer=num_layer,
                          reg=reg,
                          num_classes=10)

# X is a batch of training features
# X.shape = (batch_size, features_size)
y_out = model.forward(X)

# dout is the gradient of the loss function
#    w.r.t the output of the network.
# dout.shape = (batch_size, )
model.backward(dout)
```

Previously: Binary Cross Entropy Loss

$$BCE\left(\hat{y}, y\right) = \frac{1}{N} \sum_{i=1}^{N} \left[-y_i \log\left(\hat{y}_i\right) - (1 - y_i)\log(1 - \hat{y}_i) \right]$$

Where

- N is the number of samples

- $\hat{y}_i$ is the network's prediction for sample i

- y_i is the ground truth label (0 or 1)

New: Multiclass Cross Entropy Loss

$$CE\left(\hat{y}, y\right) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{C} \left[-y_{ik} \log\left(\hat{y}_{ik}\right)\right]$$

Where

- N is the number of samples

- $\hat{y}_{ik}$ is the network's predicted probability for the kth class when given the sample i

- y_{ik} is the ground truth label which is either 1 if the ith sample is of class k or zero otherwise

We implemented this for you! More on this topic in the next lecture.

Basic Recipe for Machine Learning

- Split your data

60%	20%	20%
train	validation	test

Find your hyperparameters   Only used once at the end!
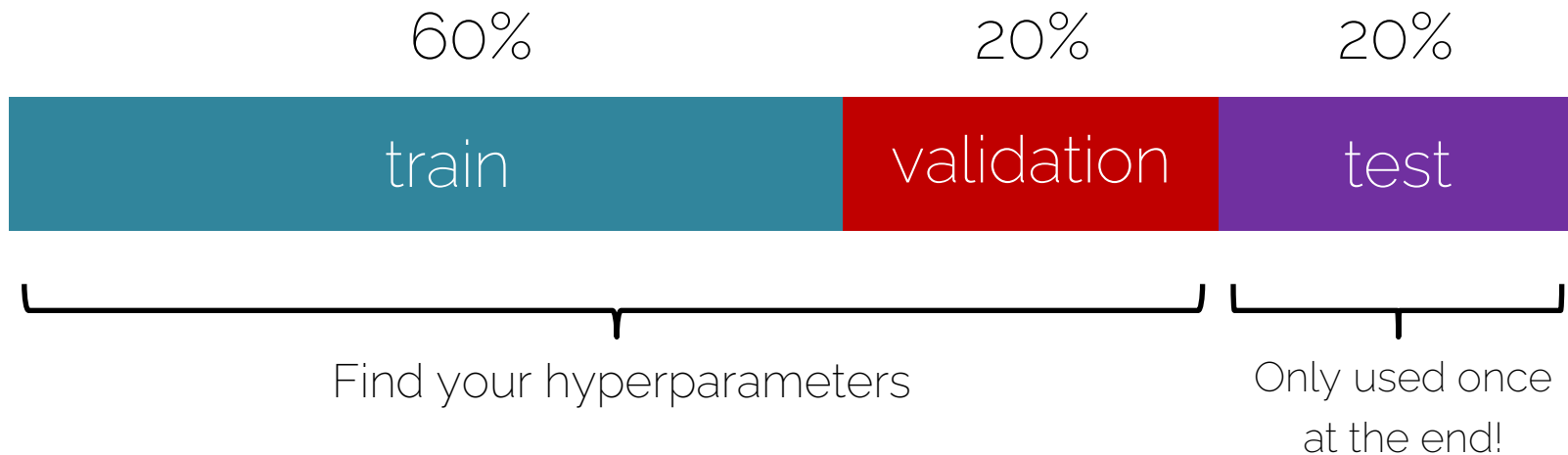
hidden test → Benchmarking: On our submission system

How to Start

- Start with single training sample
  - Check if output correct
  - Overfit →train accuracy should be 100% because input just memorized

- Increase to handful of samples

- Move from overfitting to more samples
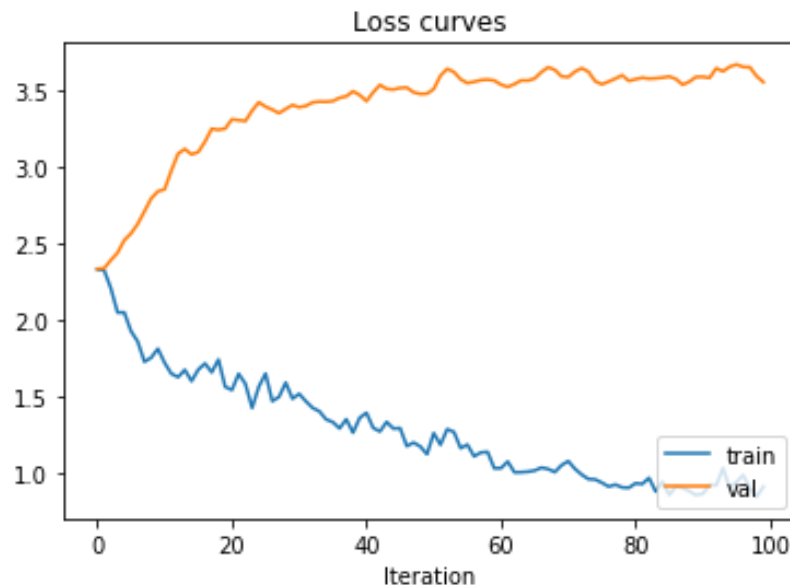 - At some point, you should see generalization

ONE

FEW

MANY
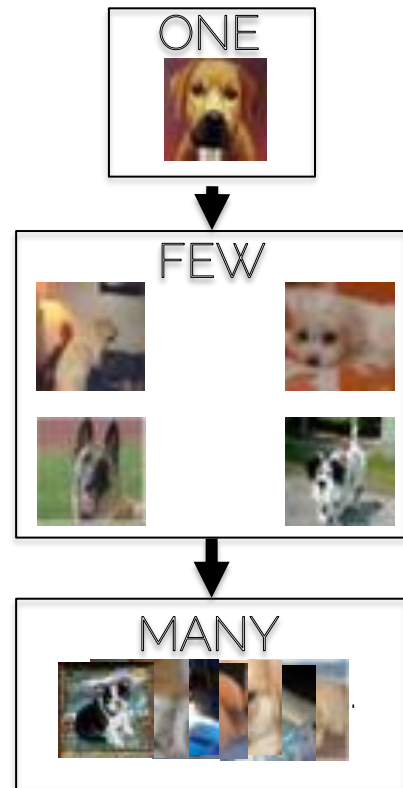
How to Start

- Overfit a single training sample
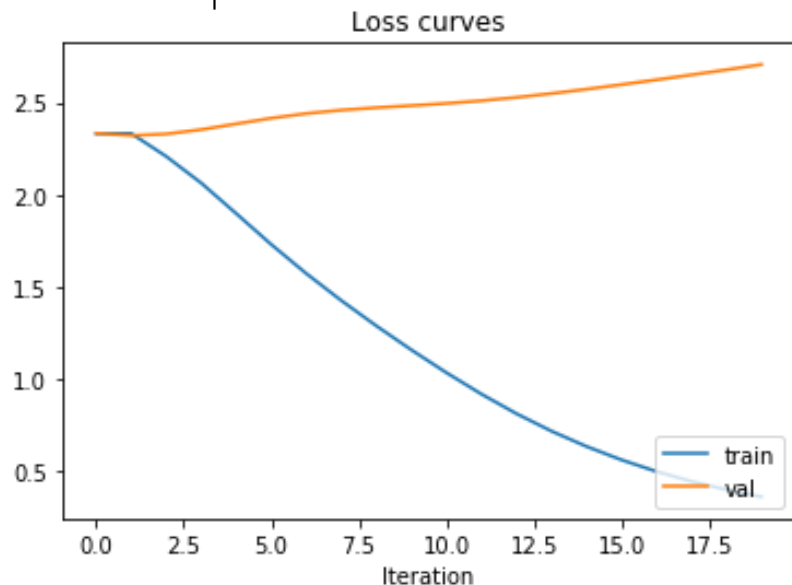
- Then a few samples

Hyperparameters

- Network architecture (e.g., num layers, hidden layer, activation function)

- Number of iterations

- Learning rate(s) (i.e., solver parameters, decay, etc.)

- Regularization (more later next lecture)

- Batch size

- …

Hyperparameter Tuning

Hyperparameters

$n\_layers = 3$
$n\_neurons = 512$
$learning\_rate = 0.1$

$n\_layers = 3$
$n\_neurons = 1024$
$learning\_rate = 0.01$

$n\_layers = 5$
$n\_neurons = 256$
$learning\ rate = 0.1$

Parameters

run optimize()

Weights optimization

Weights optimization

Weights optimization

Score

85%

80%

92%

Source: https://images.deepai.org/glossary-terms/05c646fe1676490aa0b8cab0732a02b2/hyperparams.png

How to find good Hyperparameters?

- Manual Search (trial and error)
- Automated Search:
  - Grid Search
  - Random Search

```python
from exercise_code.hyperparameter_tuning import grid_search

best_model, results = grid_search(
    dataloaders['train_small'], dataloaders['val_500files'],
    grid_search_spaces = {
        "learning_rate": [1e-2, 1e-3, 1e-4, 1e-5, 1e-6],
        "reg": [1e-4, 1e-5, 1e-6]
    },
    epochs=10, patience=5,
    model_class=ClassificationNet)
```

- Think about how different hyper parameters affect the model
  - E.g. Overfitting? -> Increase Regularization Strength, decrease model capacity

Exercise plan: Recap and Outlook

Exercise 03: Dataset and Dataloader
Exercise 04: Solver and Linear Regression
Exercise 05: Neural Networks
Exercise 06: Hyperparameter Tuning

Numpy
(Reinvent the wheel)

Exercise 07: Introduction to Pytorch
Exercise 08: MNIST with Pytorch

Pytorch/Tensorboard

Exercise 09: Convolutional Neural
        Networks
Exercise 10: Semantic Segmentation
Exercise 11: Recurrent Neural Networks

Applications
(Hands-off)

Summary

- **Tuesday, December 6**: Lecture 7 (*Training NN's 2*)

- **Wednesday, December 7, 15:59:59**: Deadline Ex6
  - Pass it by achieving required accuracy on our hidden test set.

- **Thursday, December 8**: Tutorial Session 7 *(Pytorch)*

Good luck &
see you next week ☺