BST 261: Data Science II
Lecture 5

Convolutional Neural Networks (CNNs):
Data Augmentation
+ Overview so far

Santiago Romero-Brufau
Harvard T.H. Chan School of Public Health
Spring 2

Administrivia

Fill out the final assignment form!

Office hours for the TAs (in the syllabus): Online after labs on Fridays

Might take attendance in labs (same as in lecture)

“

You can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future. You have to trust in something - your gut, destiny, life, karma, whatever. This approach has never let me down, and it has made all the difference in my life.

Steve Jobs (paraphrasing Kierkegaard)

https://youtu.be/UF8uR6Z6KLc

(15 min. Stanford Commencement address)

Overview so far

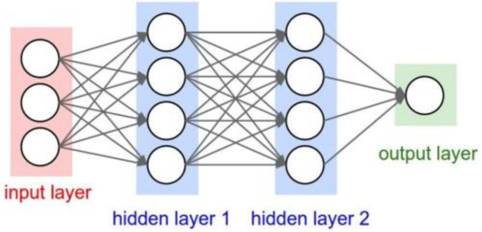
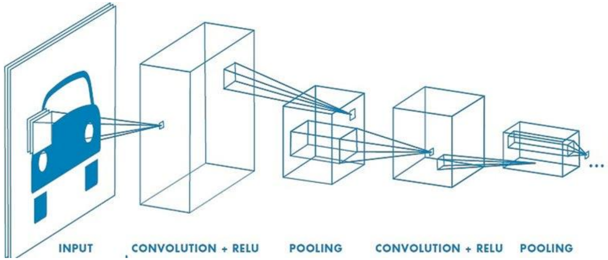
3 Components of a neural network

Network structure: Determines how the inputs are processed

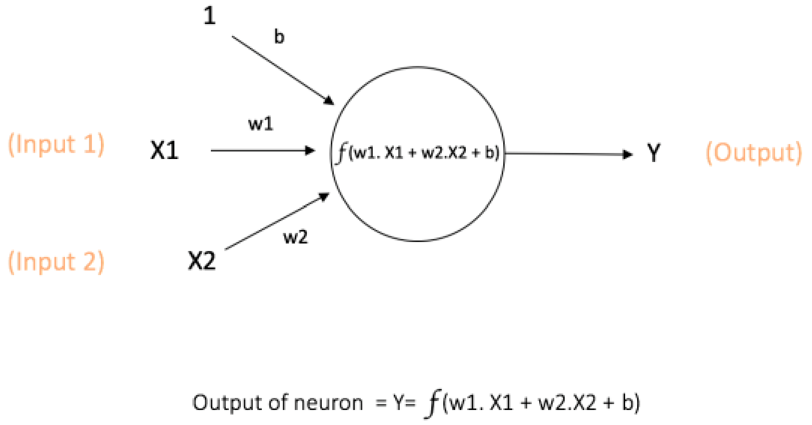
Loss function: How we calculate the “distance” between the prediction and the ground truth (between Yhat and Y). Depends on the output type (classification vs regression)

Optimizer: determines how the parameters are updated. Generally it’s a version of Stochastic Gradient Descent.

Neural network structures

Neural Network type	Structure	Best for
MLP (multilayer perceptron)		Structured data (tables)
CNN (convolutional neural network)		Image data

Forward pass

Forward pass (Inference)	We take a set of X and calculate Yhat	

Key thing to remember: in each layer, the activation function is a **non-linear function.**

Back propagation

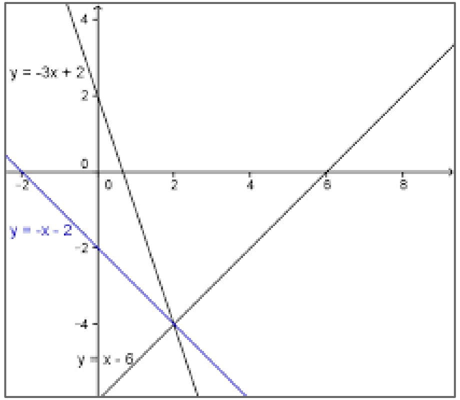
Back propagation	We take a loss and calculate the new parameters (W and b)	Old weight Derivative of Error with respect to weight $^{*}W_x = W_x - a\left(\frac{\partial Error}{\partial W_x}\right)$ New weight Learning rate

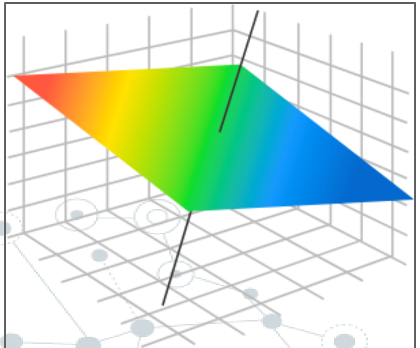
Key thing to remember: in each layer, we move the set of weights in the direction in which the loss function is decreased (downhill).

Other quick thoughts

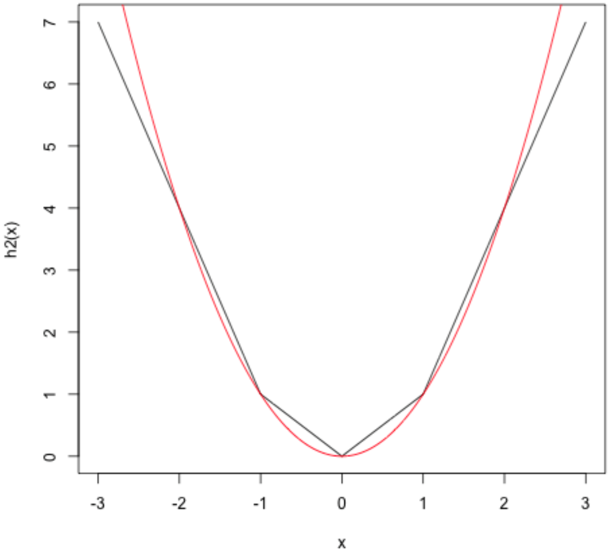
Why non-linearity matters

Without ReLU (only linear functions)





With ReLU (non-linearity!)



f$(x)=g(x) + g(-x) + g(2x-2) + g(-2x+2)$

where $g(x)$ = ReLU(x)

Example MLP: multilayer perceptron

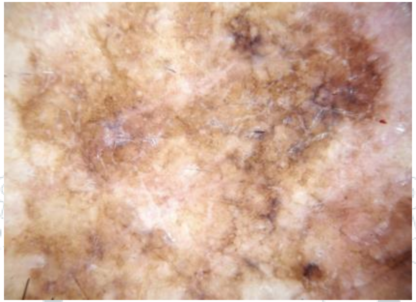
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 60)                2700
_________________________________________________________________
dense_1 (Dense)              (None, 55)                3355
_________________________________________________________________
dense_2 (Dense)              (None, 50)                2800
_________________________________________________________________
dense_3 (Dense)              (None, 45)                2295
_________________________________________________________________
dense_4 (Dense)              (None, 30)                1380
_________________________________________________________________
dense_5 (Dense)              (None, 20)                620
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 21
=================================================================
Total params: 13,171
Trainable params: 13,171
Non-trainable params: 0
```

CNNs in Python

Classifying Skin Lesions

- We’ll be using data from the International Skin Imaging Collaboration: Melanoma Project
- We’ll be classifying images as malignant or benign
- The overarching goal of the ISIC Melanoma Project is to support efforts to reduce melanoma-related deaths and unnecessary biopsies by improving the accuracy and efficiency of melanoma early detection

Malignant



Malignant

Benign

Benign

Classifying Skin Lesions

- This archive contains 23k images of classified skin lesions. It contains both malignant and benign examples
  - We’ll be using a fraction of this
- Each example contains the image of the lesion, meta data regarding the lesion (including classification and segmentation) and meta data regarding the patient
- The data can be viewed in this link (in the gallery section)
- It can be downloaded through the site or by using this repository

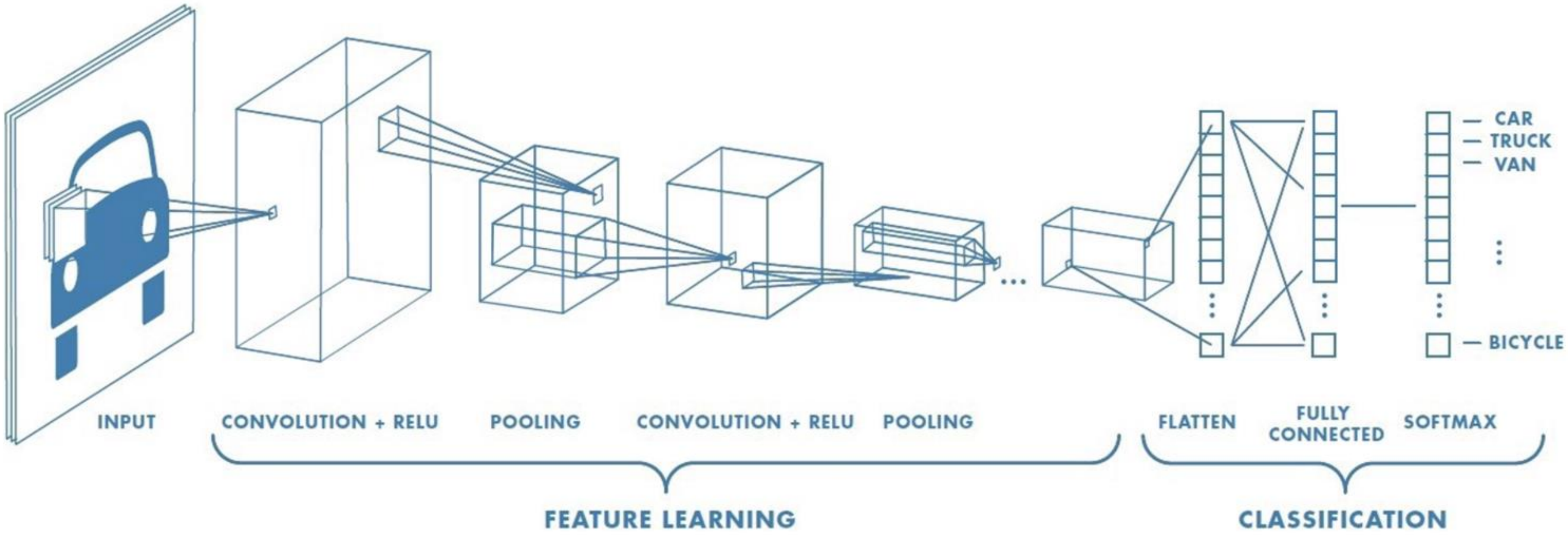
Classification Skin Lesions

- The subsample of the data is available in a Google Drive folder
- You can access it with this code and notebook
- You can also download the images to your machine if you would like
  - There are zip files available on canvas
- We'll start with creating a simple CNN

```python
# Define model
model = keras.Sequential([
  layers.Conv2D(32, (3, 3), activation='relu', input_shape=(150, 150, 3)),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(64, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(128, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(128, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Flatten(),

  layers.Dense(512, activation='relu'),

  layers.Dense(1, activation='sigmoid')
])
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 148, 148, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 74, 74, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 72, 72, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 36, 36, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 34, 34, 128)       73856
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 17, 17, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 15, 15, 128)       147584
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
flatten (Flatten)            (None, 6272)              0
_________________________________________________________________
dense (Dense)                (None, 512)               3211776
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 513
=================================================================
Total params: 3,453,121
Trainable params: 3,453,121
Non-trainable params: 0
```

CNN Schematic



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Number of parameters

width m, **height n**, **previous layer's filters d** and account for all such filters **k in the current layer**. Don't forget the bias term for each of the filter. Number of parameters in a CONV layer would be : **((m * n * d)+1)* k)**, added 1 because of the bias term for each filter.

((shape of width of the filter * shape of height of the filter * number of filters in the previous layer+1)*number of filters)

((3 * 3 * 3) +1)* 32

https://towardsdatascience.com/understanding-and-calculating-the-number-of-parameters-in-convolution-neural-networks-cnns-fc88790d530d

```python
# Define model
model = keras.Sequential([
  layers.Conv2D(32, (3, 3), activation='relu', input_shape=(150, 150, 3)),
```

width m, **height n**, **previous layer's filters d** and account for all such filters **k in the current layer**. Don't forget the bias term for each of the filter. Number of parameters in a CONV layer would be : **((m * n * d)+1)* k)**, added 1 because of the bias term for each filter.

((m * n * d)+1)* k)
((3 * 3 * 3)+1)* 32) = 896

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 148, 148, 32)      896
_________________________________________________________________
```

```python
# Define model
model = keras.Sequential([
  layers.Conv2D(32, (3, 3), activation='relu', input_shape=(150, 150, 3)),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(64, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(128, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(128, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Flatten(),

  layers.Dense(512, activation='relu'),

  layers.Dense(1, activation='sigmoid')
])
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 148, 148, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 74, 74, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 72, 72, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 36, 36, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 34, 34, 128)       73856
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 17, 17, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 15, 15, 128)       147584
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
flatten (Flatten)            (None, 6272)              0
_________________________________________________________________
dense (Dense)                (None, 512)               3211776
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 513
=================================================================
Total params: 3,453,121
Trainable params: 3,453,121
Non-trainable params: 0
```

Convolution and pooling layers - notice how the output size decreases with each layer. Remember what applying a filter, padding, and/or strides does to an input.

```python
# Define model
model = keras.Sequential([
  layers.Conv2D(32, (3, 3), activation='relu', input_shape=(150, 150, 3)),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(64, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(128, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(128, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Flatten(),

  layers.Dense(512, activation='relu'),

  layers.Dense(1, activation='sigmoid')
])
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 148, 148, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 74, 74, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 72, 72, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 36, 36, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 34, 34, 128)       73856
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 17, 17, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 15, 15, 128)       147584
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
flatten (Flatten)            (None, 6272)              0
_________________________________________________________________
dense (Dense)                (None, 512)               3211776
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 513
=================================================================
Total params: 3,453,121
Trainable params: 3,453,121
Non-trainable params: 0
```

We finish the network with 1 hidden dense layer and 1 output layer. Note that most of the parameters in the model come from the hidden dense layer and not the convolution or pooling layers.

```python
# Define model
model = keras.Sequential([
  layers.Conv2D(32, (3, 3), activation='relu', input_shape=(150, 150, 3)),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(64, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(128, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Conv2D(128, (3, 3), activation='relu'),
  layers.MaxPooling2D((2, 2)),

  layers.Flatten(),

  layers.Dense(512, activation='relu'),

  layers.Dense(1, activation='sigmoid')
])
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 148, 148, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 74, 74, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 72, 72, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 36, 36, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 34, 34, 128)       73856
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 17, 17, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 15, 15, 128)       147584
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
flatten (Flatten)            (None, 6272)              0
_________________________________________________________________
dense (Dense)                (None, 512)               3211776
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 513
=================================================================
Total params: 3,453,121
Trainable params: 3,453,121
Non-trainable params: 0
```

Total # of parameters that need to be learned by the network

```python
from keras.preprocessing.image import ImageDataGenerator

# All images will be rescaled by 1./255
train_datagen = ImageDataGenerator(rescale=1./255)
test_datagen = ImageDataGenerator(rescale=1./255)

train_generator = train_datagen.flow_from_directory(
        # This is the target directory
        train_dir,
        # All images will be resized to 150x150
        target_size = (150, 150),
        batch_size = 20,
        # Since we use binary_crossentropy loss, we need binary labels
        class_mode = 'binary')

validation_generator = test_datagen.flow_from_directory(
        validation_dir,
        target_size = (150, 150),
        batch_size = 20,
        class_mode = 'binary')
```

We first scale the data to get values between 0 and 1.

Then we transform the images to be 150x150 pixels in size (this is arbitrary), declare a batch size of 20 (this is also arbitrary), and declare the class mode (i.e. the type of classification we want to do)

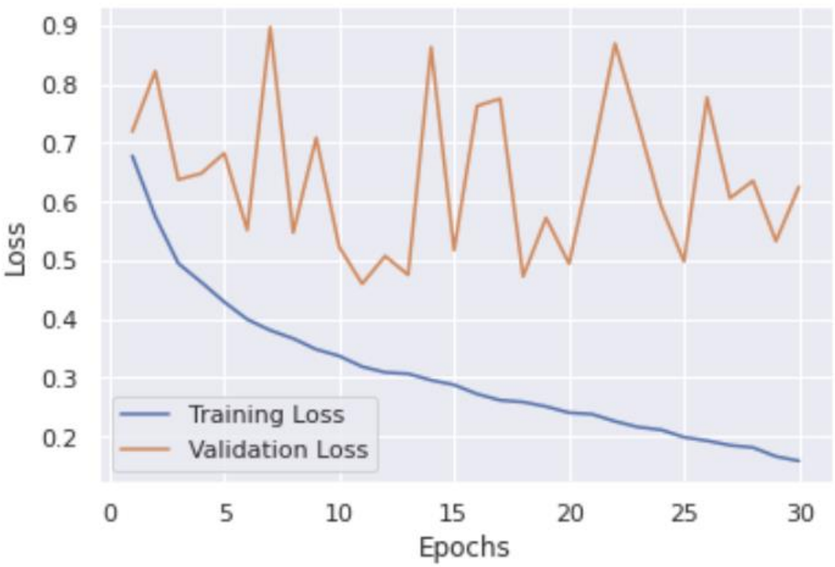
```
for data_batch, labels_batch in train_generator:
    print('data batch shape:', data_batch.shape)
    print('labels batch shape:', labels_batch.shape)
    break
```

```
data batch shape: (20, 150, 150, 3)
labels batch shape: (20,)
```

We can see the shape of the images: they are in batches of 20, with each image being represented by 3, 150x150 tensors: one for R, one for G and one for B color “channels”.

```
history = model.fit(
        train_generator,
        steps_per_epoch = 81, # ceil(1609/20)
        epochs = 30,
        validation_data = validation_generator,
        validation_steps = 22) # ceil(426/20)
```

The number of training examples divided by the batch size, i.e. how many batches we need to go through until the model sees all training data. For both the training and validation sets.


Loss
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
5
10
15
20
25
30
Epochs
Training Loss
Validation Loss


Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
25
30
Epochs
Training Accuracy
Validation Accuracy

Data Augmentation

- As we have seen, overfitting is caused by having too few training examples to learn from

- Data augmentation generates more training data from existing training examples by **augmenting** the samples via a number of random transformations

- These transformations should yield believable images

Data Augmentation

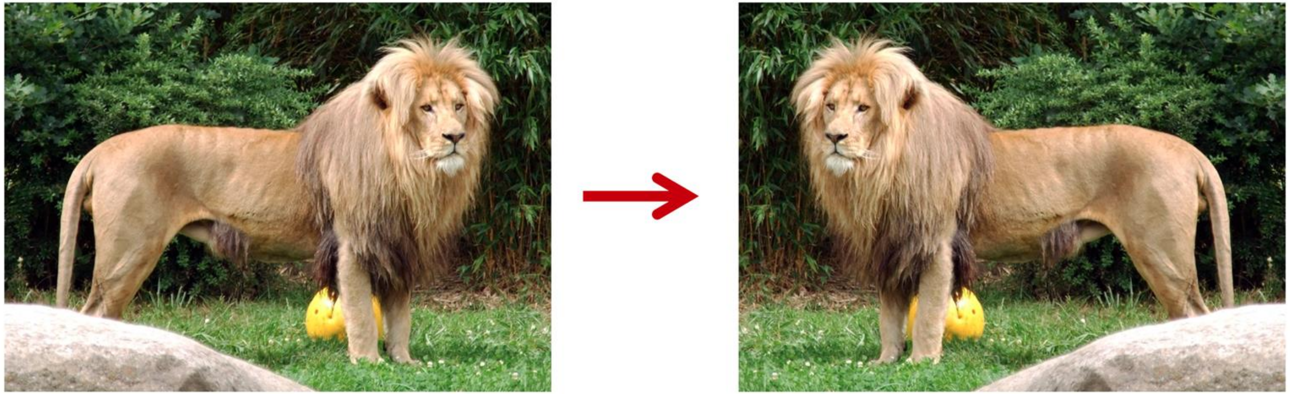
- Types of augmentation:
  - Rotation
  - Horizontal/vertical flip
  - Random crops/scales
    - Zoom
    - Width or height shifts
  - Shearing
  - Brightness, contrast, saturation
  - Lens distortions

Types of data augmentation

1. Rotations

Types of data augmentation

2. Horizontal/Vertical Flips

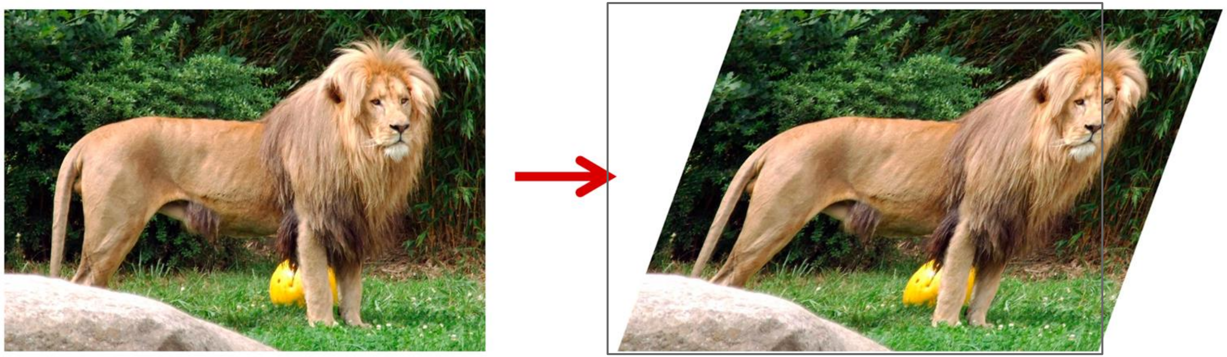


Types of data augmentation

3. Random crops/scales

Types of data augmentation

4. Shearing

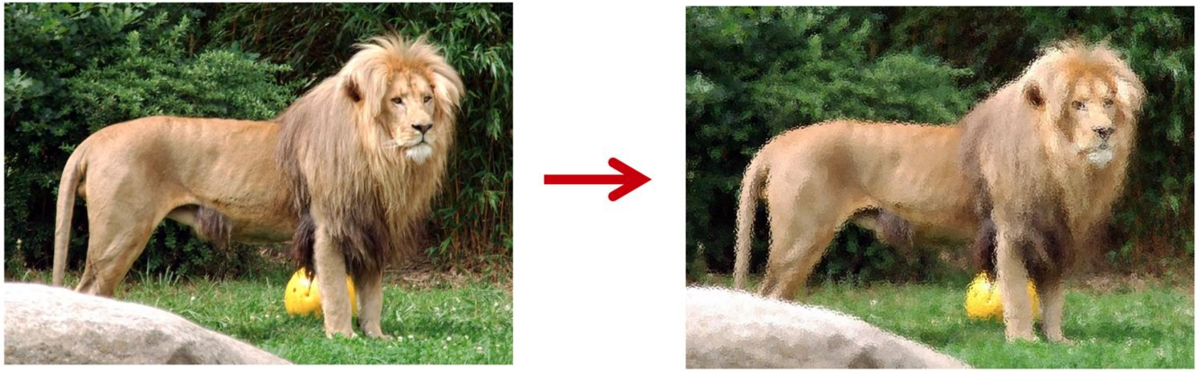


Types of data augmentation

5. Brightness, contrast, saturation

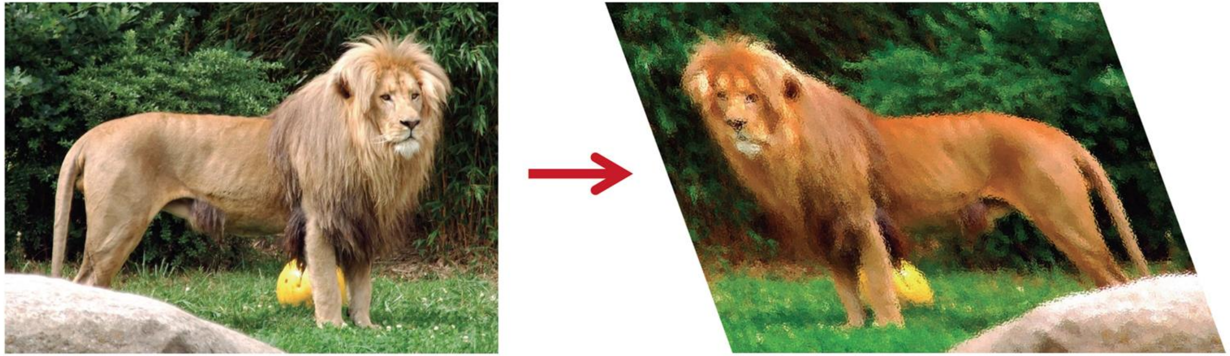
Types of data augmentation

6. Lens distortions



Types of data augmentation

7. Combinations of the above



Data Augmentation

- If you train a network using data augmentation, it will never see the same input twice, but the inputs will still be heavily correlated
    - You're remixing known information, not producing new information
- May not completely escape overfitting due to this correlation
- Adding dropout can also help

Data Augmentation in Keras

```python
from keras.preprocessing.image import ImageDataGenerator
datagen = ImageDataGenerator(
        rotation_range = 40,
        width_shift_range = 0.2,
        height_shift_range = 0.2,
        shear_range = 0.2,
        zoom_range = 0.2,
        horizontal_flip = True,
        fill_mode = 'nearest')
```

You can create your own data generator with any specifications you'd like. The values chosen here are arbitrary.

You can check out the Keras documentation to see all of the available options and values each type of augmentation type can take.

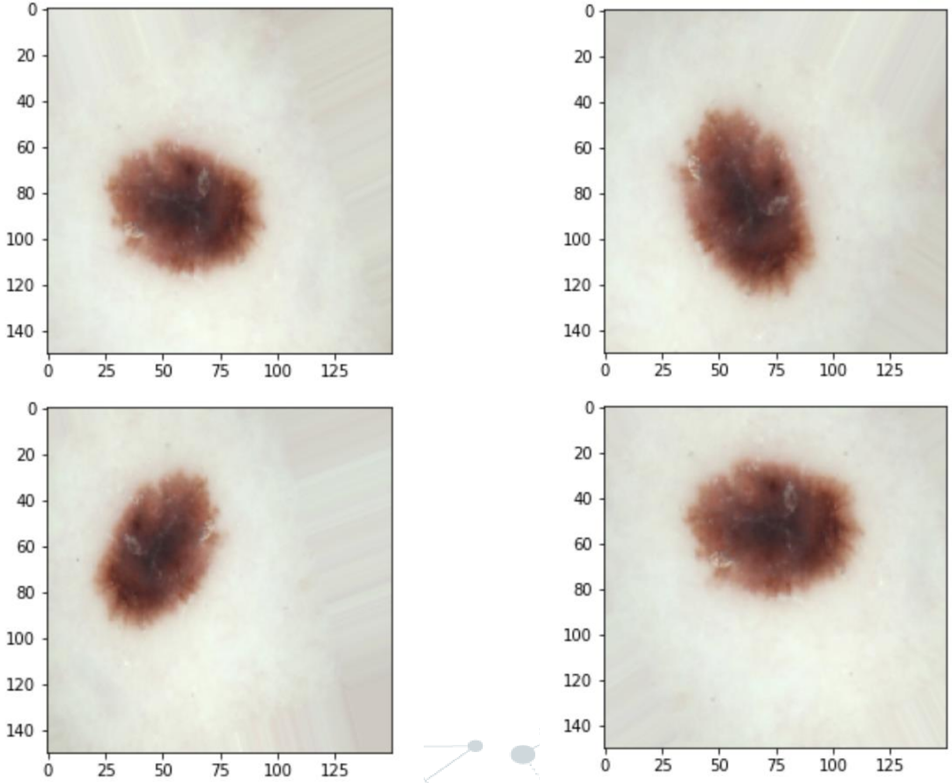
Note that only your training data should be augmented - not the test or validation sets. The point of augmentation is to "increase" your training set size.

```
from keras.preprocessing.image import ImageDataGenerator

# All images will be rescaled by 1./255
train_datagen = ImageDataGenerator(rescale=1./255)
test_datagen = ImageDataGenerator(rescale=1./255)

train_generator = train_datagen.flow_from_directory(
        # This is the target directory
        train_dir,
        # All images will be resized to 150x150
        target_size = (150, 150),
        batch_size = 20,
        # Since we use binary_crossentropy loss, we need binary labels
        class_mode = 'binary')

validation_generator = test_datagen.flow_from_directory(
        validation_dir,
        target_size = (150, 150),
        batch_size = 20,
        class_mode = 'binary')
```

We first scale the data to get values between 0 and 1.

Then we transform the images to be 150x150 pixels in size (this is arbitrary), declare a batch size of 20 (this is also arbitrary), and declare the class mode (i.e. the type of classification we want to do)

Data Augmentation in Keras

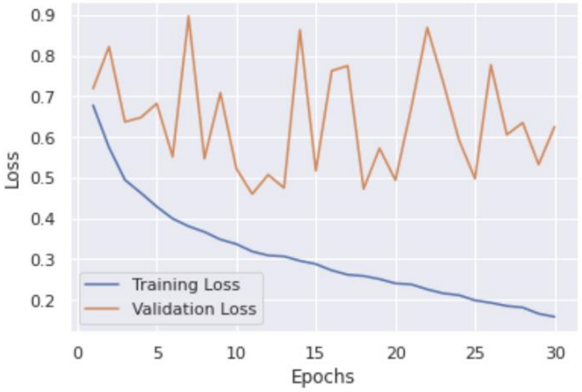


Back to the notebook

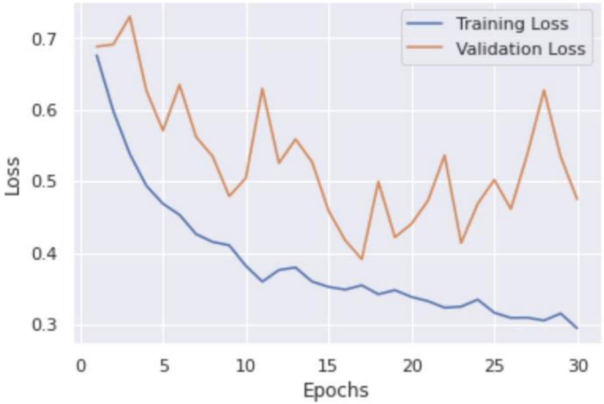
[Skin lesions with data augmentation]

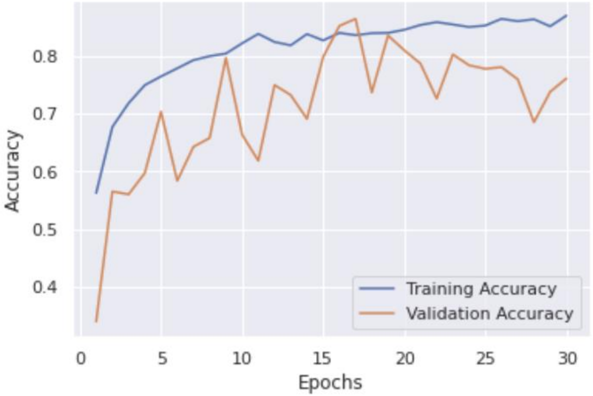
Data Augmentation in Keras

Without augmentation



With augmentation





Change due date for HW1 on Canvas (to Sat), and remove date on the homework itself