Attempts to Interpret a Neural Network

Attempts to Interpret a Neural Network

We can

- Understand the model architecture
- **Visualize the filters / weights**
- Extract the output of intermediate neurons / layers
- Locate important parts of the image according to the model

What do you mean by parameters?

What do you mean by parameters?

In [0]:
```
# defining the architecture of the model
model=Sequential()
model.add(InputLayer(input_shape=(224*224*3,)))
model.add(Dense(100, activation='sigmoid'))
model.add(Dense(units=1, activation='sigmoid'))
```

What do you mean by parameters?

```
In [0]: # summary of the model
        model.summary()
```

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 100)               15052900
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 101
=================================================================
```

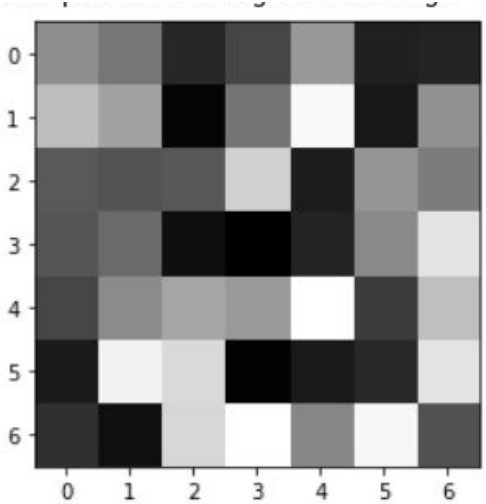
What do you mean by parameters?

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 218, 218, 32)      4736
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 54, 54, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 48, 48, 32)        50208
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 12, 12, 32)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 6, 6, 32)          50208
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 1, 1, 32)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 32)                0
_________________________________________________________________
preds (Dense)                (None, 1)                 33
=================================================================
Total params: 105,185
Trainable params: 105,185
Non-trainable params: 0
_________________________________________________________________
```

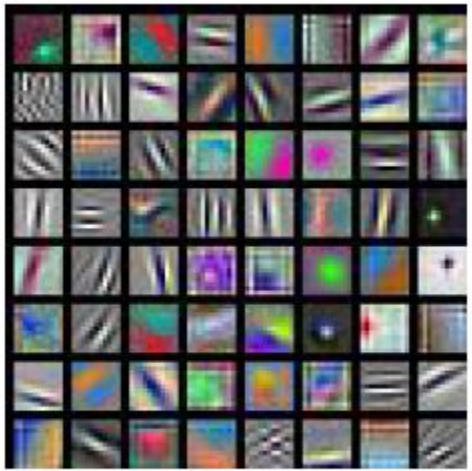
Attempt 2 - Visualizing the weights / filters

Attempt 2 - Visualizing the weights / filters

Attempt 2 - Visualizing the weights / filters



Attempt 2 - Visualizing the weights / filters



AlexNet:
64 x 3 x 11 x 11

ResNet-18:
64 x 3 x 7 x 7

ResNet-101:
64 x 3 x 7 x 7

DenseNet-121:
64 x 3 x 7 x 7

Thank You