MASTER'S THESIS IN

HUMAN EXPLAINABILITY THROUGH AN AUXILIARY NEURAL NETWORK

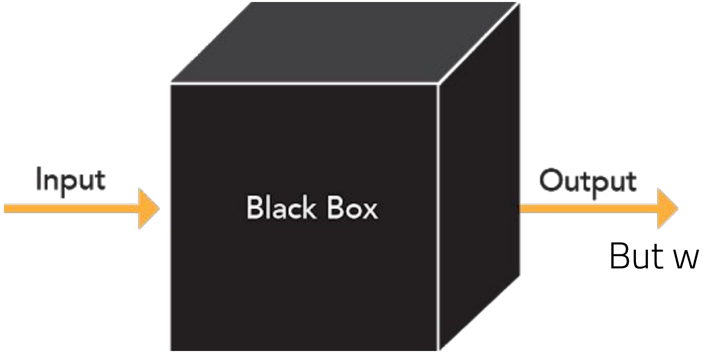
Albert Garcia Sanchez

FINDING THE MASTER'S THESIS

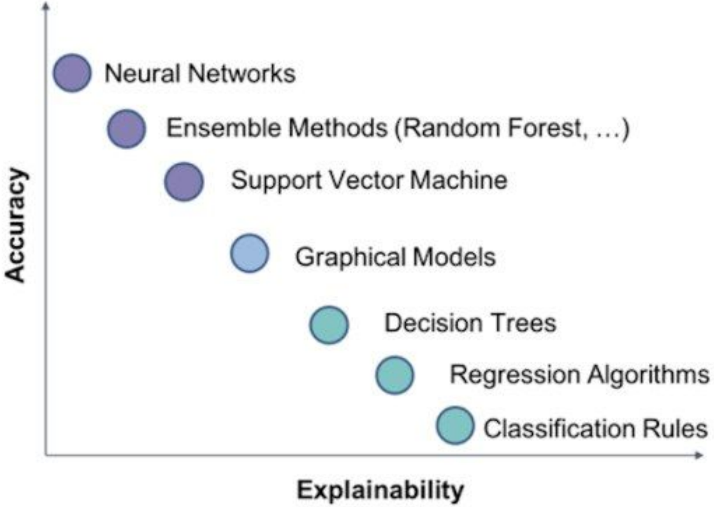
Perform research in Computer Vision

I was presented with a hot-topic in Machine Learning and Deep Learning

EXPLAINABILITY



But why...?



METHODOLOGIES FOLLOWED

- **Visualisation methods:** highlight most impactful input features

- **Model distillation:** mimic a DNN with a white box model (ML)

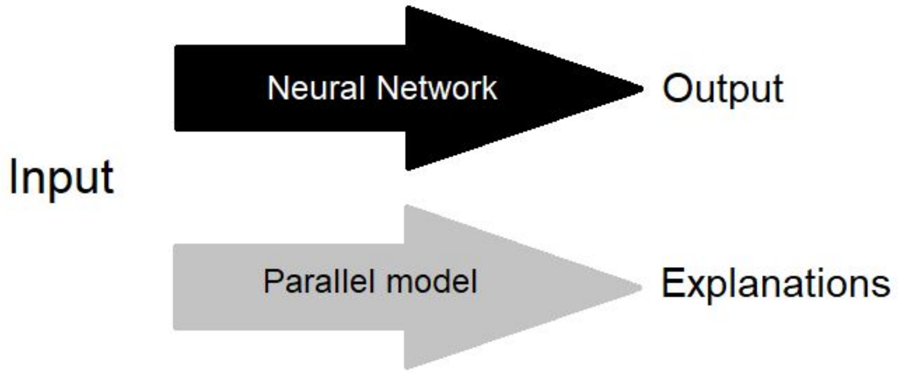
- **Intrinsic methods:** design a DNN capable of yielding explanations at the same time

KEY IDEA THAT LEAD TO THE THESIS

"I really think the development of Deep Learning should be parallel to that of its explainability"

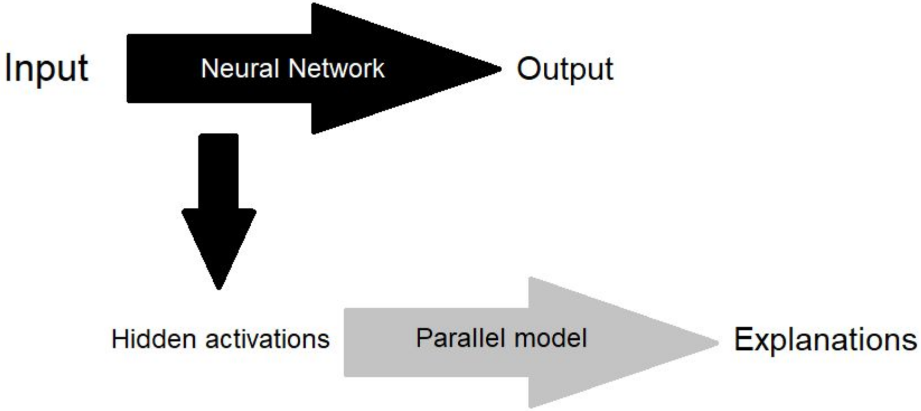
MY REASONING – INITIAL PROPOSAL

What if instead of directly finding explanations within the predictor Neural Network I find these through a parallel model



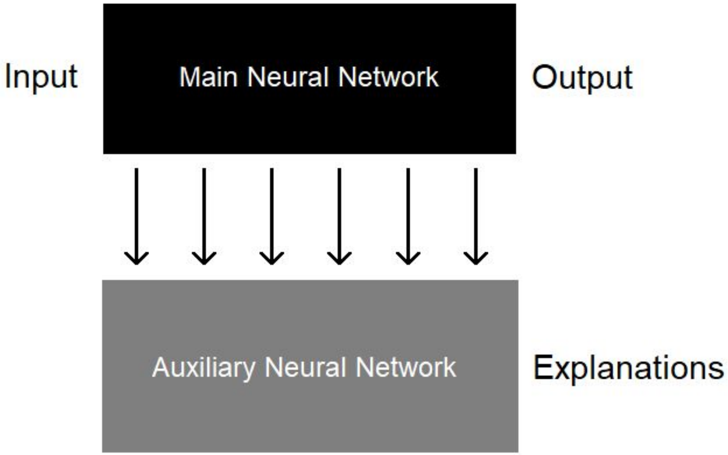
MY REASONING – THE BLACK BOX

My proposal to deal with the Neural Network being a black box is by using its hidden activations as the input to the parallel model



MY REASONING – FINAL PROPOSAL

We can take advantage of the layers and connect them in a sequential way. This provides the Auxiliary Network with the capability of selecting the best features to make the explanations.



BUT WHY A NEURAL NETWORK

- Model capable of handling and working with hidden activations

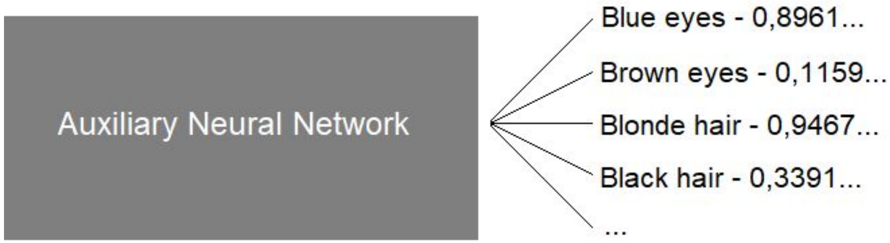
- Hidden activations cannot be directly translated into human-interpretable explanations

- A Neural Network can do both things as long as there is data with the needed labels or information

MODELING THE EXPLANATIONS

A set of essential and meaningful explanations, which can justify the possible outputs, must be carefully designed.

This renders the Auxiliary Neural Network to deal with a multi-label classification task (*N* binary classifications)



Auxiliary Neural Network

Blue eyes - 0,8961...

Brown eyes - 0,1159...

Blonde hair - 0,9467...

Black hair - 0,3391...

...

AUXILIARY NETWORK LEARNING

The Auxiliary Neural Network can achieve high accuracy without actually capturing the notions behind each explanation

Nevertheless, there is a property which can *presumably* help with this

VARIABILITY IN THE TRAINING DATA

- Common explanations from different classes
- Different explanations from same class
- These comparisons could guide the auxiliary network towards learning what we want it to learn

WRAPPING EVERYTHING UP

- Auxiliary Network feeds from the hidden activations of the Main one

- A set of explanations has to be carefully designed (e.g. by a field professional)

- **Hypothesis:** the high variability of explanations in the training data can guide the Auxiliary Network into learning how to extract the explanations

APPLICATIONS

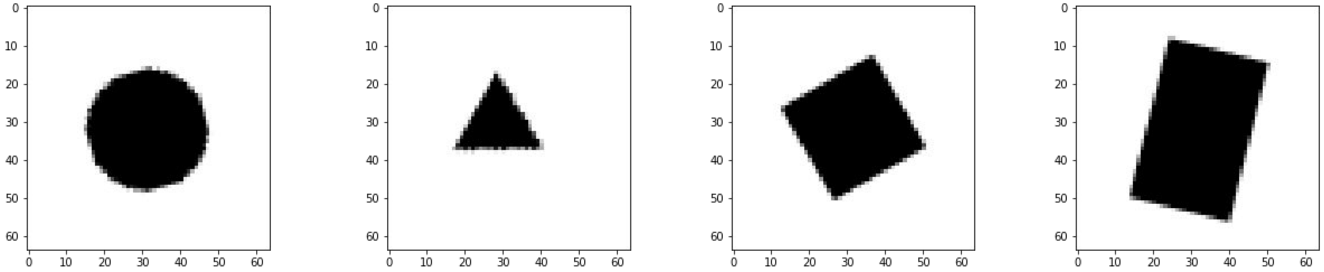
- Comparison between outputs for verification
- Detection of out-of-distribution or anomaly
- Detection of unwanted biases
- And many more…

FIRST PROJECT

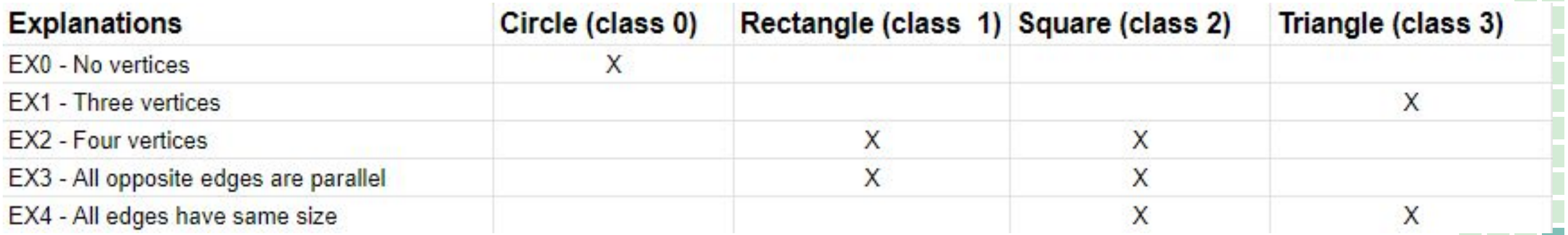
- Classification of 2D grayscale shapes: Circle, triangle, square, rectangle

- Explanations based on fundamental properties of each shape

- Synthetic dataset (3000 samples per class) + simple CNN implemented with Keras

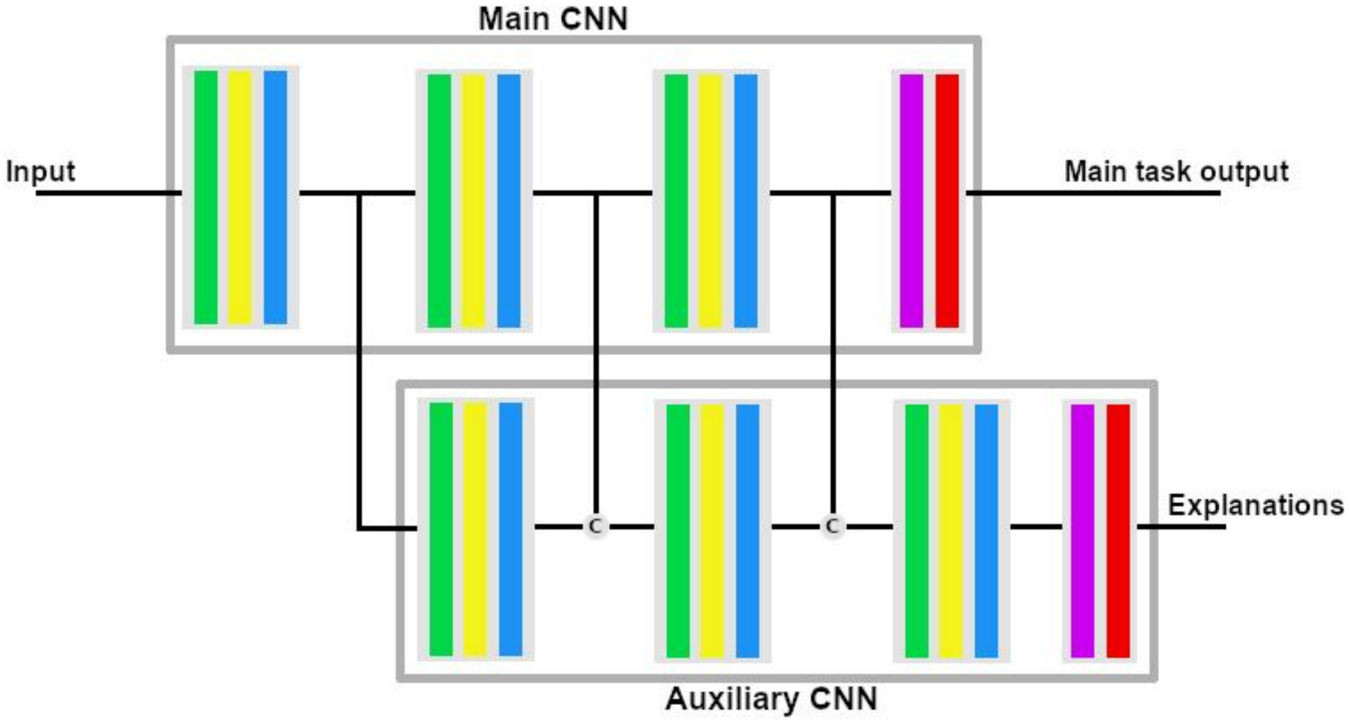


SET OF EXPLANATIONS DEFINED



Explanations	Circle (class 0)	Rectangle (class 1)	Square (class 2)	Triangle (class 3)
EX0 - No vertices	X			
EX1 - Three vertices				X
EX2 - Four vertices		X	X	
EX3 - All opposite edges are parallel		X	X	
EX4 - All edges have same size			X	X

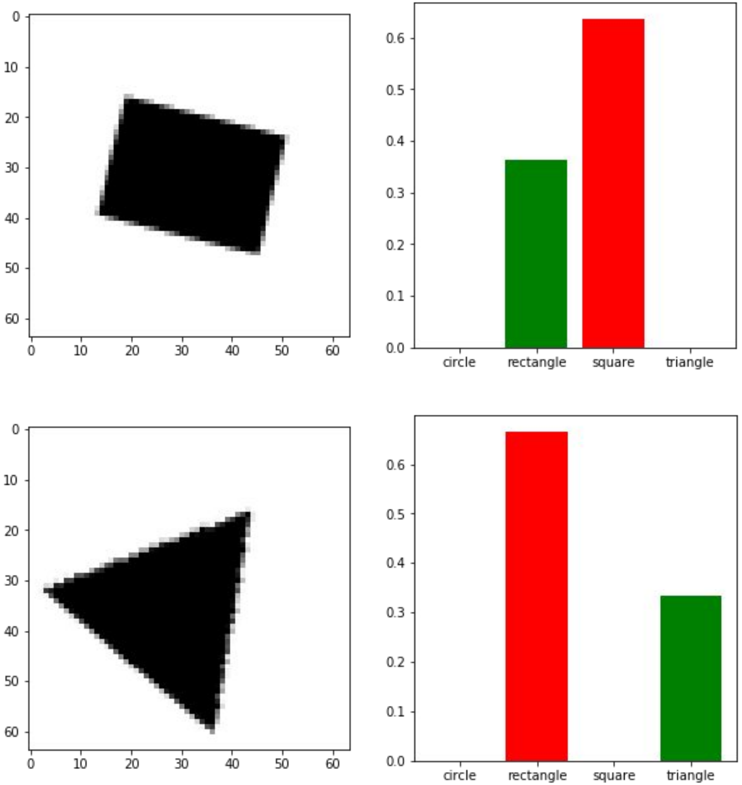
MODEL ARCHITECTURE – MIRROR CNN



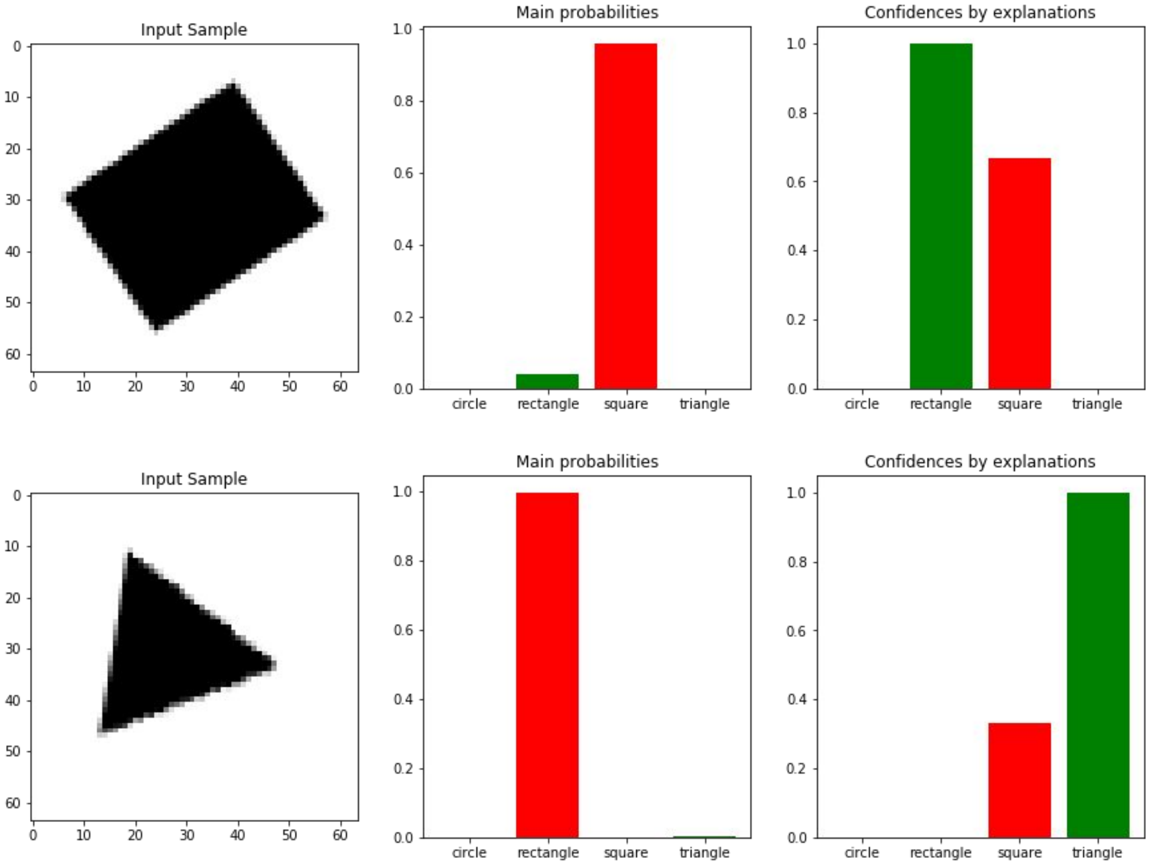
TRAINING RESULTS – EARLY STOPPING

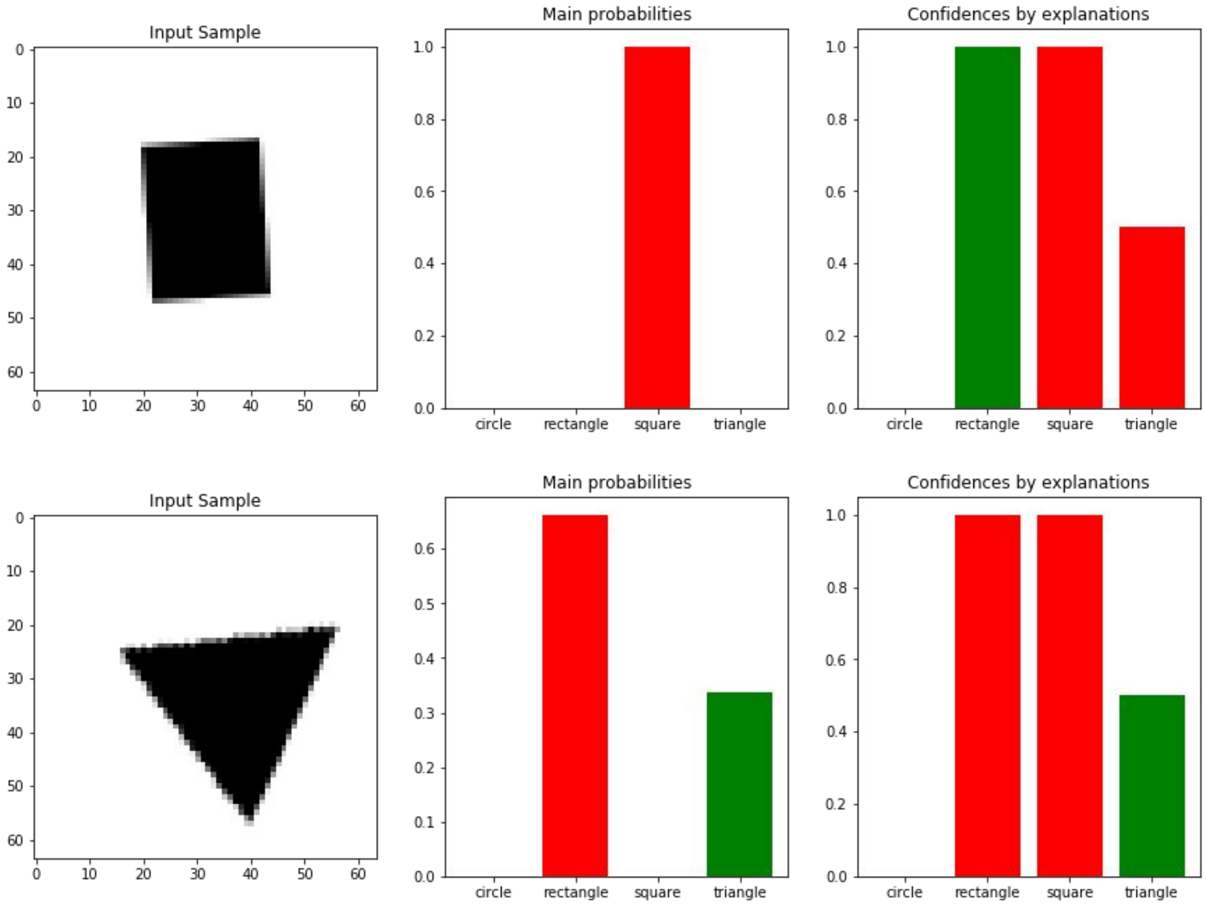
	Main CNN **(Simple CNN architecture)**	**Explanations CNN** **(Mirror CNN architecture)**
Train accuracy	100%	99.99%
Validation accuracy	99.15%	99.74%

CHECKING MISCLASSIFICATIONS

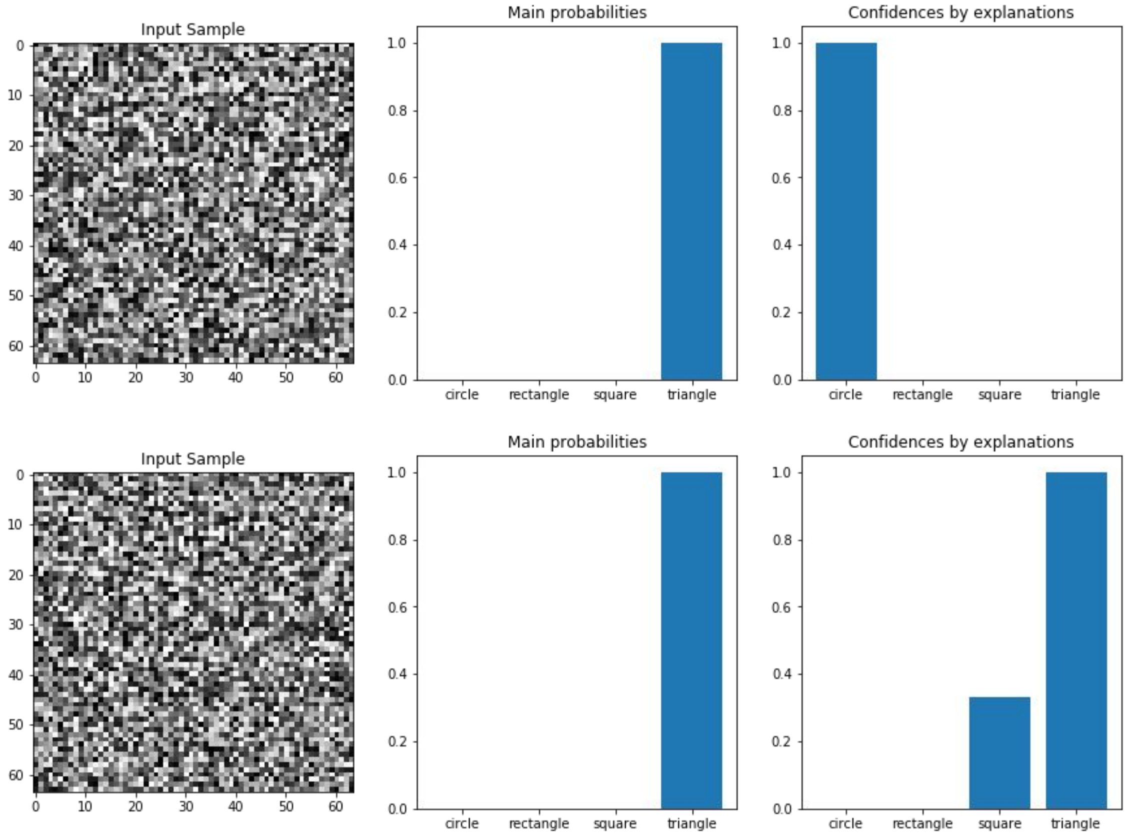


MISCLASSIFICATIONS WITH EXPLANATIONS




Input Sample
0
10
20
30
40
50
60
0
10
20
30
40
50
60
Main probabilities
1.0
0.8
0.6
0.4
0.2
0.0
circle
rectangle
square
triangle
Confidences by explanations
1.0
0.8
0.6
0.4
0.2
0.0
circle
rectangle
square
triangle
Input Sample
0
10
20
30
40
50
60
0
10
20
30
40
50
60
Main probabilities
0.6
0.5
0.4
0.3
0.2
0.1
0.0
circle
rectangle
square
triangle
Confidences by explanations
1.0
0.8
0.6
0.4
0.2
0.0
circle
rectangle
square
triangle

DETECTING OUT-OF-DISTRIBUTION



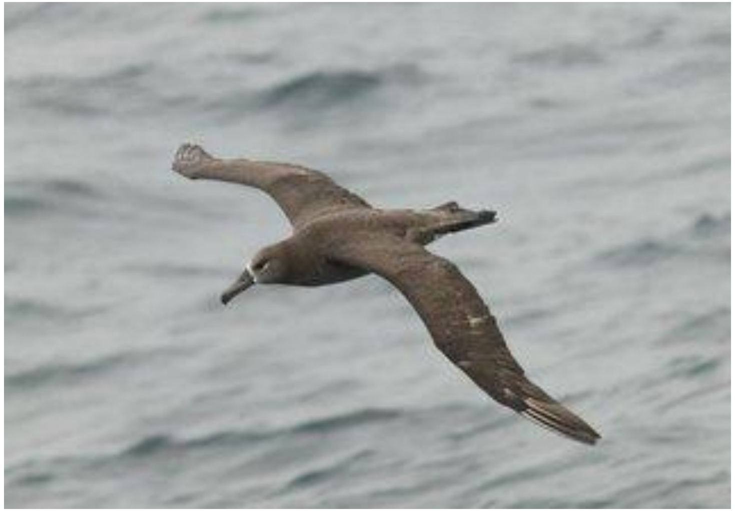
COMMENTS ON THE FIRST PROJECT

- ~80% of the out-of-distributions are questioned
- ~57% of the misclassified samples are questioned while only 0.13% of correctly classified samples are questioned
- Cannot ensure the Auxiliary Network has learned the notions behind each explanation. **No variability within the same class**
- Not a real dataset and extremely easy task to optimise

SECOND PROJECT

- Real dataset with 200 classes of birds (60 samples per class)
- 50/50 split taken from another project
- 312 explanations based on the birds visual attributes in each image
- Using a state-of-the-art CNN architecture, ResNet, with PyTorch
- Transfer Learning from ImageNet to avoid overfitting

DATASET EXAMPLE



Bird class: Black footed Albatross

Explanations

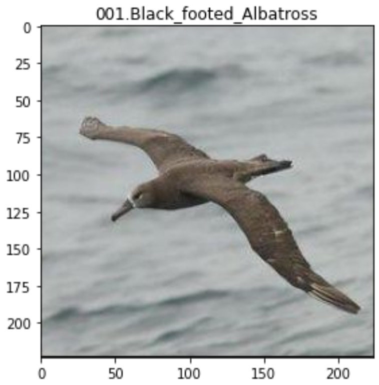
- Wing color: Brown
- Bill shape: Spatulate
- Upperparts color: Brown
- Underparts color: Brown
- Breast pattern: Solid
- Back color: Brown
- Breast color: Brown
- Tail shape: Squared
- Upper tail color: Brown
- Head pattern: Plain
- Throat color: Brown
- And more…

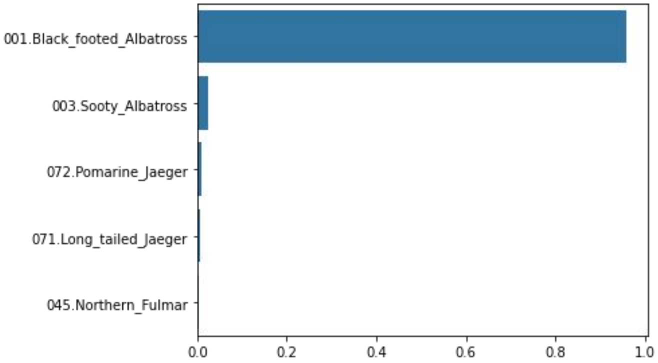
Rest of the explanations not present

TRAINING THE MAIN NETWORK

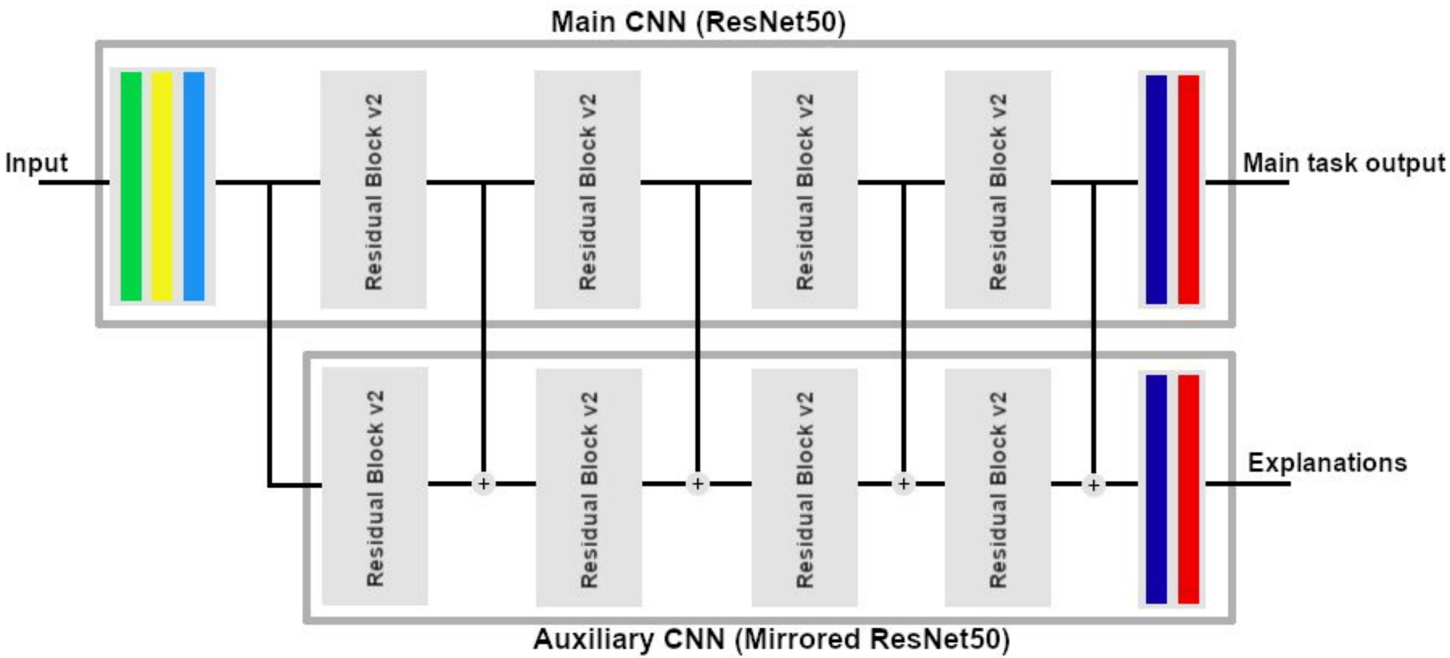
	ResNet18	ResNet50
Train accuracy	78.43%	89.81%
Validation accuracy	77.05%	81.03%

MAIN PREDICTION EXAMPLE





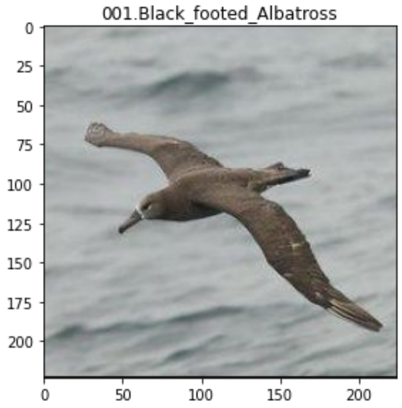
MODEL ARCHITECTURE – MIRROR CNN

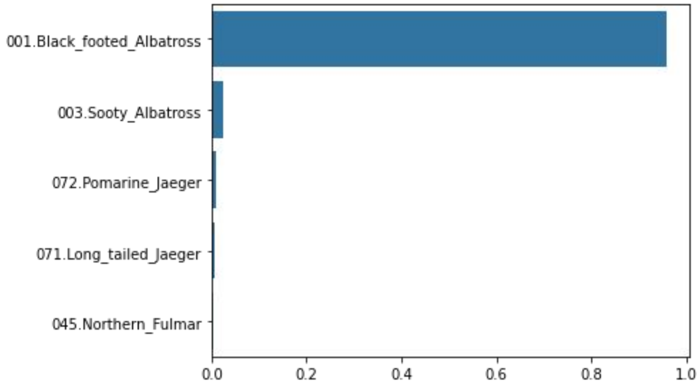


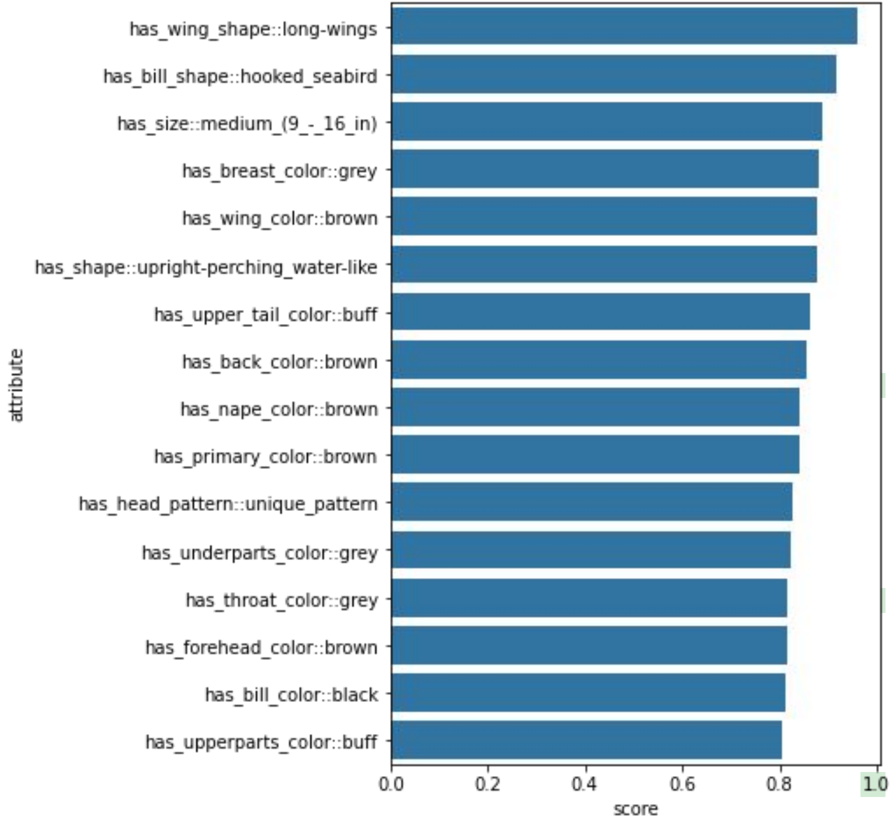
TRAINING THE AUXILIARY NETWORK

	Main CNN **(ResNet50)**	**Explanations CNN** **(ResNet50 mirror)**
Train accuracy	89.81%	85.43%
Validation accuracy	81.03%	85.19%

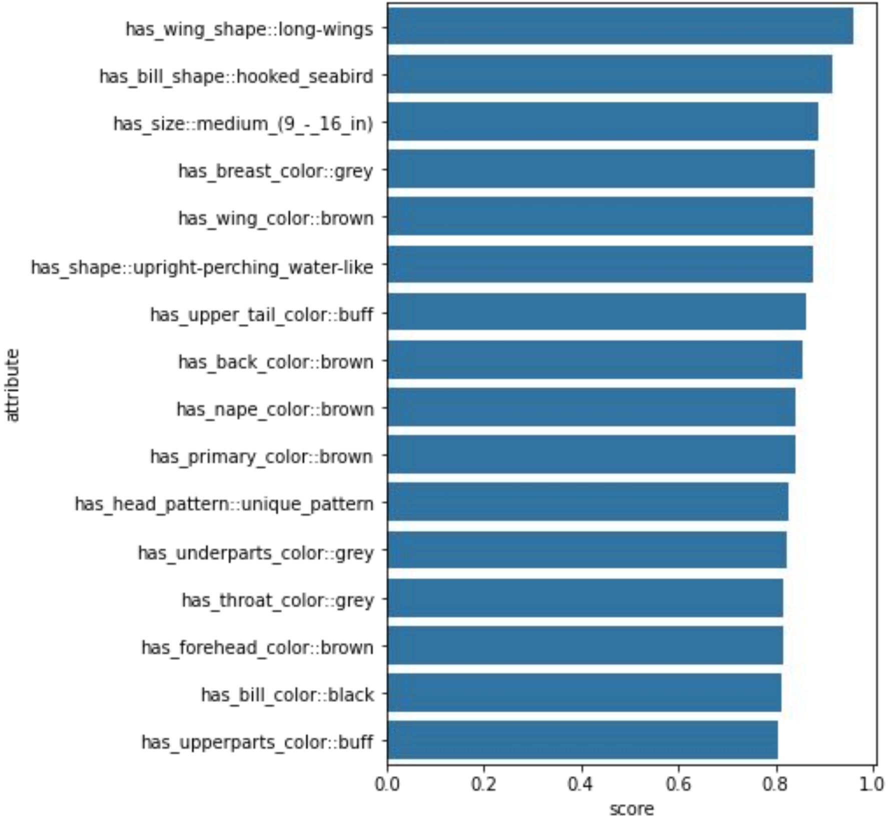
PREDICTION EXAMPLE WITH EXPLANATIONS

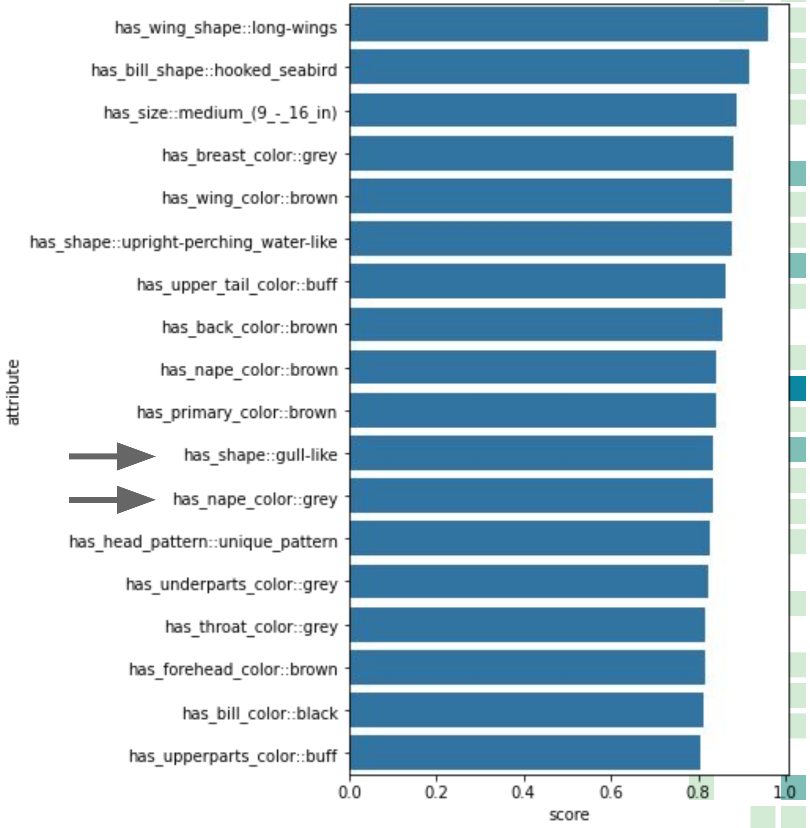






CLEANING REPEATED EXPLANATIONS





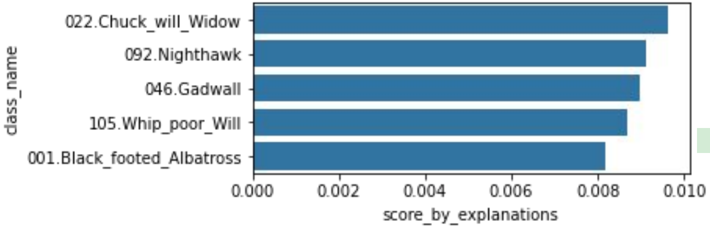
PREDICTIONS BY EXPLANATIONS

Predicted explanations vector
(312 scores)

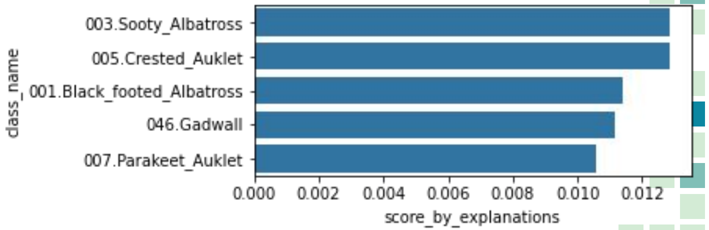
+

Explanation probability per class
(200 classes x 312 explanations)

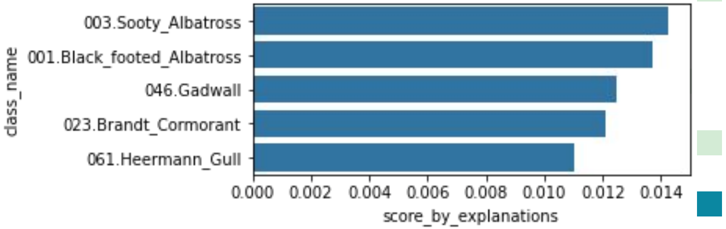
No cleaning
No threshold



Cleaning
No threshold

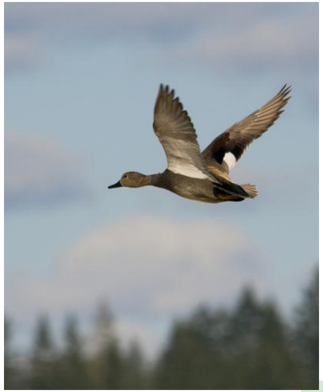


Cleaning
Threshold

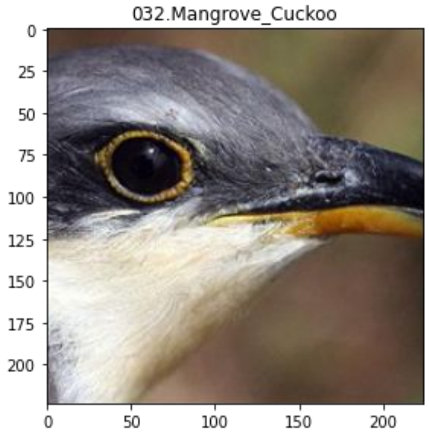


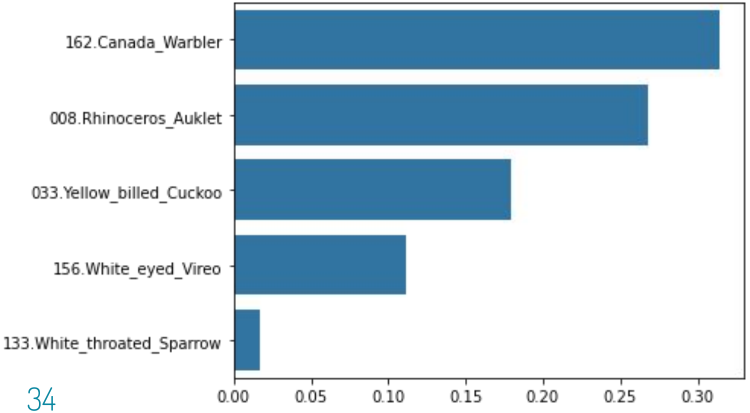
DO THEY MAKE SENSE?

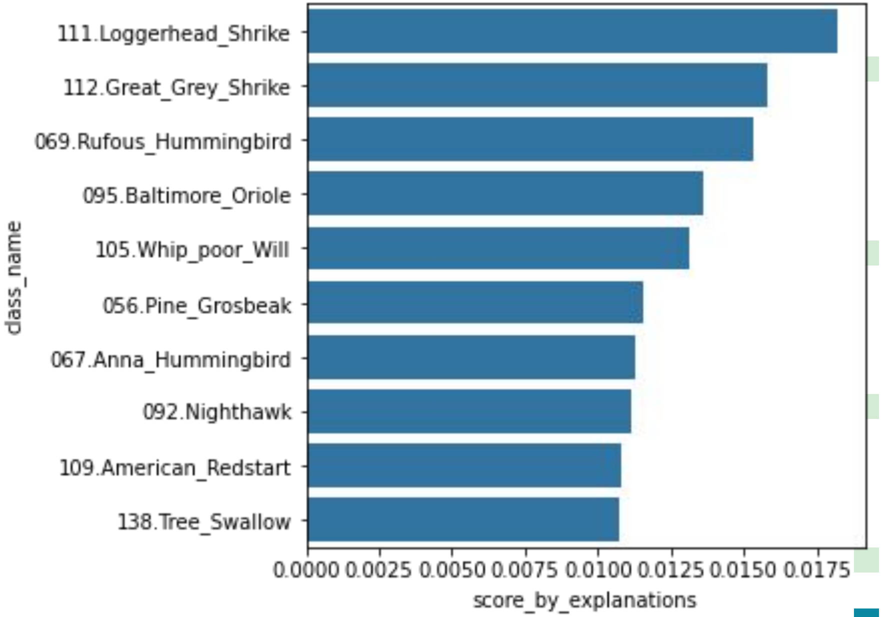
©2004 W. Ryan Holliday



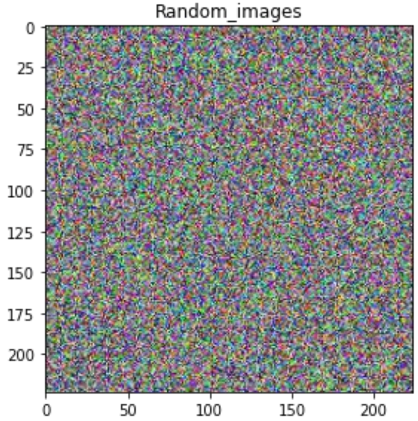
MISCLASSIFIED EXAMPLE

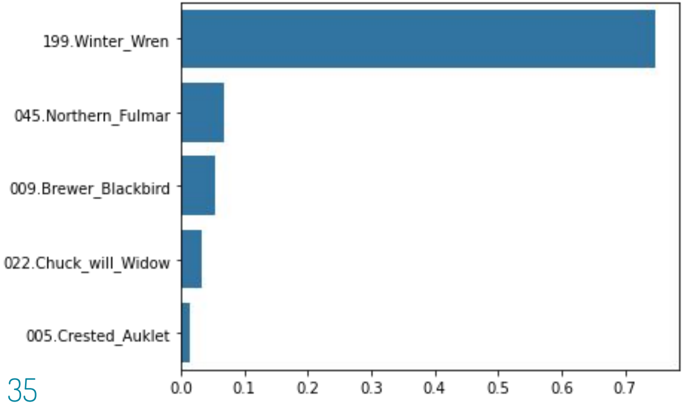


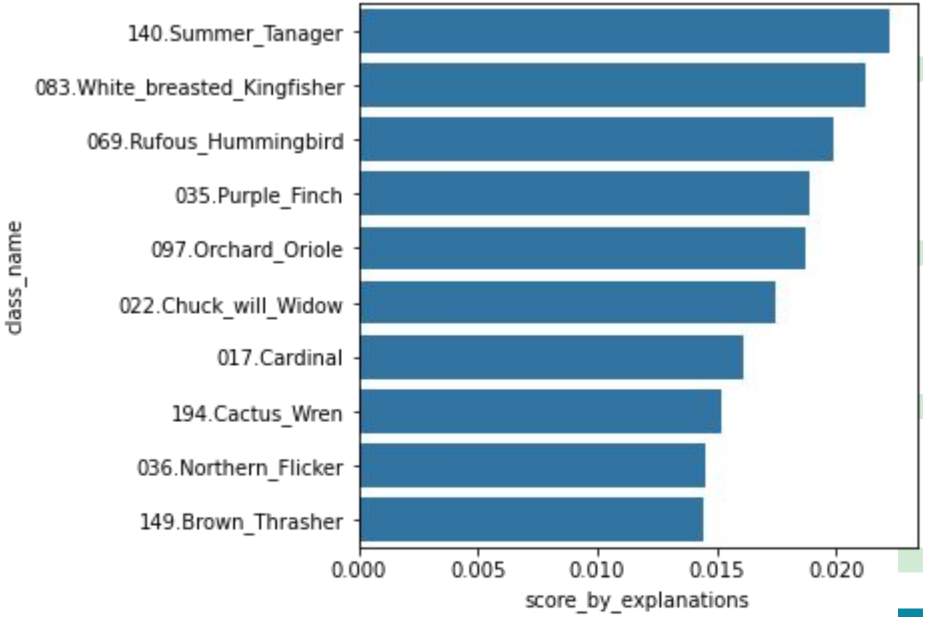




OUT-OF-DISTRIBUTION EXAMPLE







COMMENTS ON THE SECOND PROJECT

- 100% of the out-of-distributions are questioned

- 35% of the misclassified samples are questioned while 15% of correctly classified samples are questioned

- The predicted explanations are, by themselves, extremely valuable and useful

FINAL REMARKS

- **By no means** explanation predictions replace the main predictions
- The birds classification task is hard: lots of visual similarities and low amount of samples per class (heavy data augmentation)
- Further work could be performed such as trying to detect unwanted biases or training both networks simultaneously
- To conclude, **results are truly promising as well as appealing** at the cost of acquiring a labeled dataset with explanations

Thank you
for your time