Evaluation of Explainable Artificial Intelligence techniques

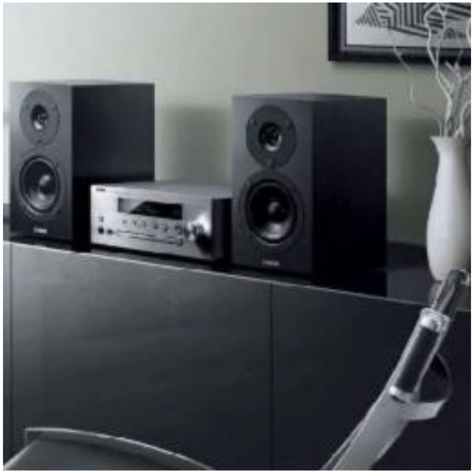
The latest generation of artificial intelligence has proven to be extremely powerful and valuable, with applications in various fields like medicine and computer science. Its learning techniques can be difficult for humans to understand. But can we trust a program that can learn and decide independently without understanding how it works?

To address this issue, a new field of research called XAI (eXplainable Artificial Intelligence) has emerged. Its goal is to make the decisions made by artificial intelligence understandable to humans. This is important not only for understanding how AI works, but also for increasing its transparency and people's trust in it.

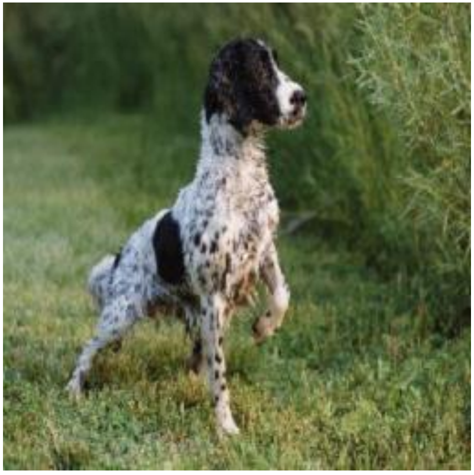
Your task is to answer a series of questions to evaluate different XAI techniques, considering the outputs of these techniques, which are provided in this document.

The AI system that we want to explain with this experiment is a Neural Network capable of classifying images by predicting the correct class. We present four correctly classified examples, which will be used throughout the whole document:

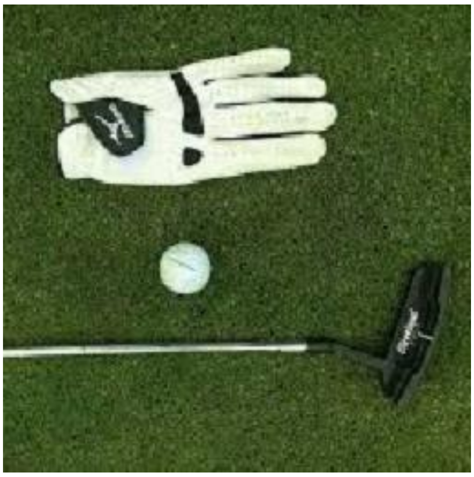
Cassette Player



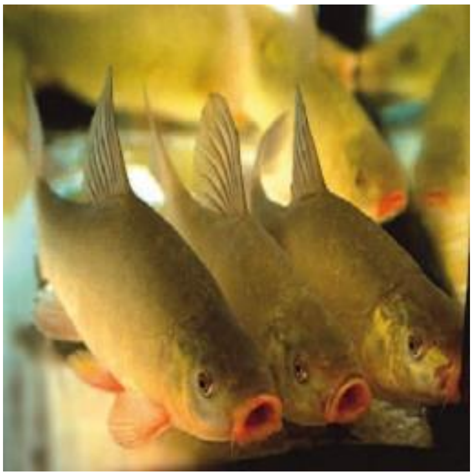
English Springer



Golf Ball



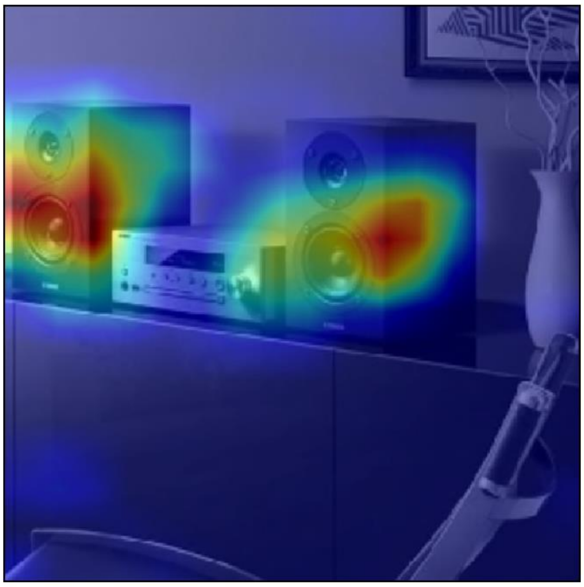
Tench



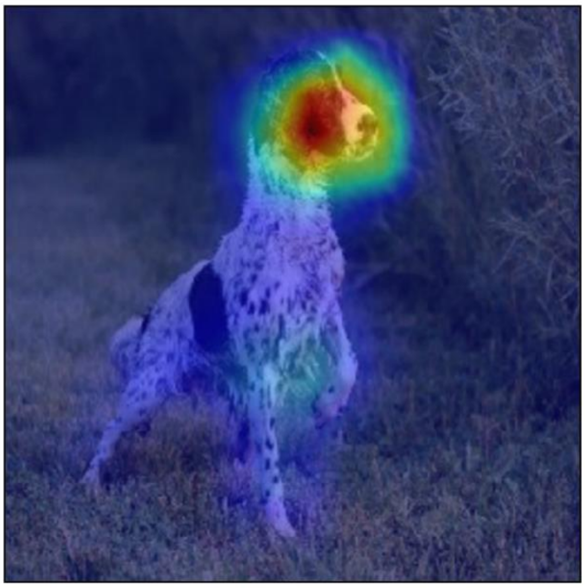
Grad-CAM

This technique produces an explanation in the form of a heatmap where the “hot regions” (the part coloured in red) represent the most important parts towards the prediction.

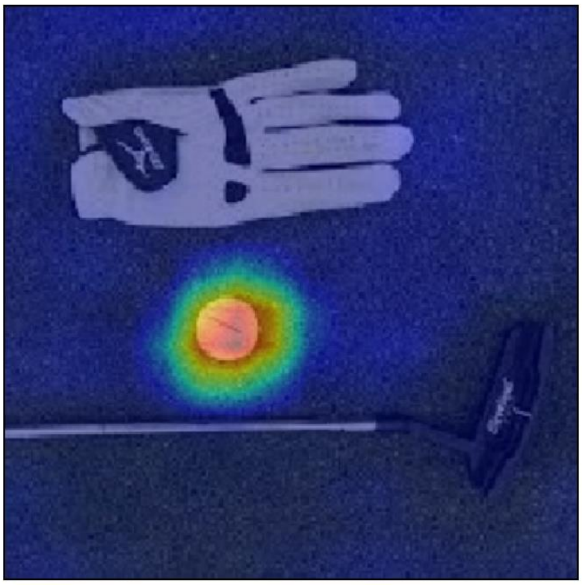
Cassette Player



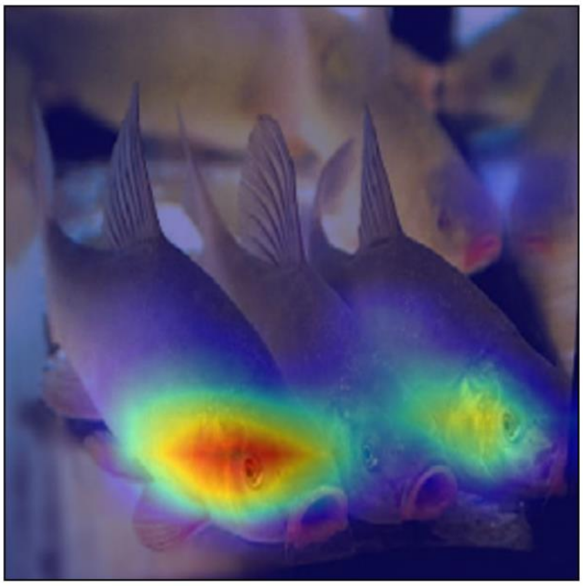
English Springer



Golf Ball



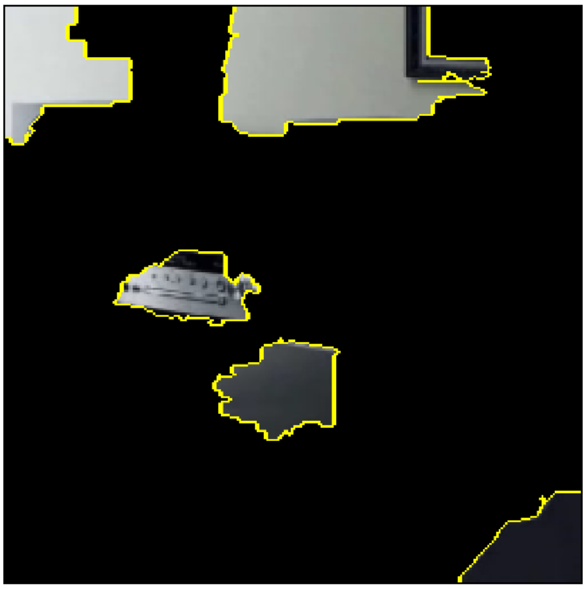
Tench



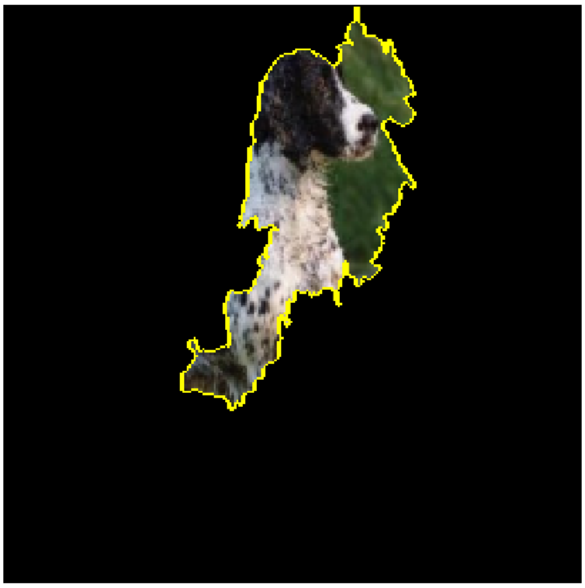
LIME

This method divides an image into different regions, called superpixels, and shows the ones that, if removed, change the network's output the most.

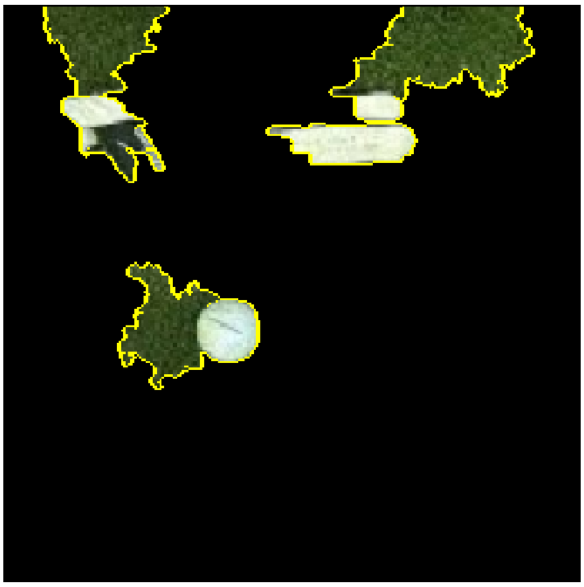
Cassette Player



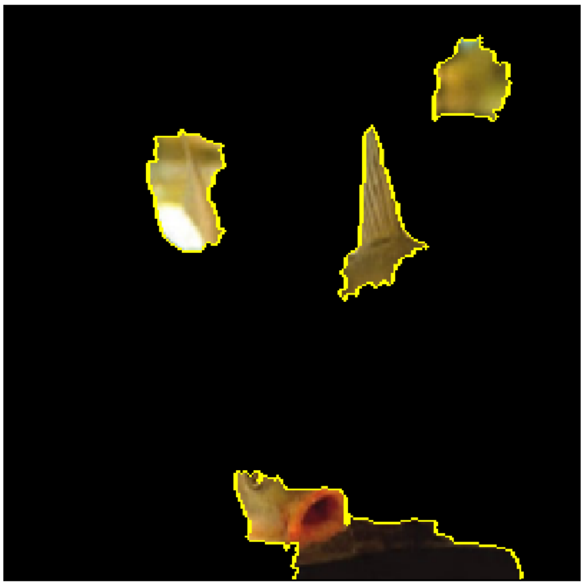
English Springer



Golf Ball



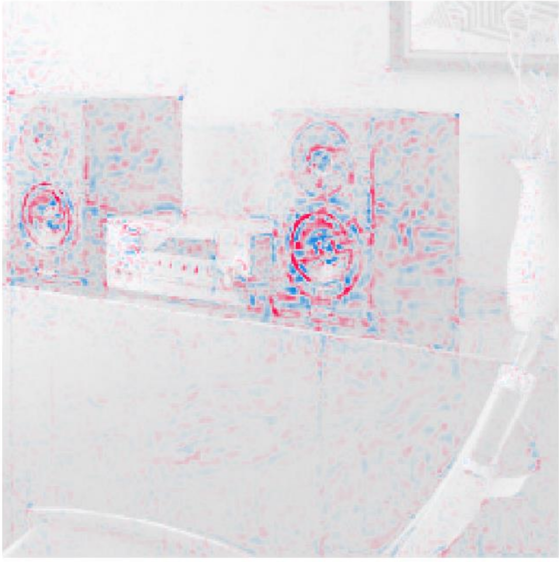
Tench



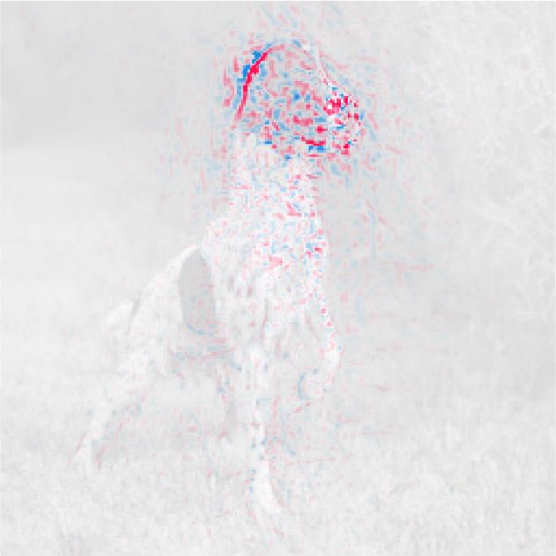
SHAP

It computes the contribution of each pixel of the image, using a game theory approach. More specifically, red pixels represent a positive contribution, while blue pixels represent a negative contribution towards the prediction.

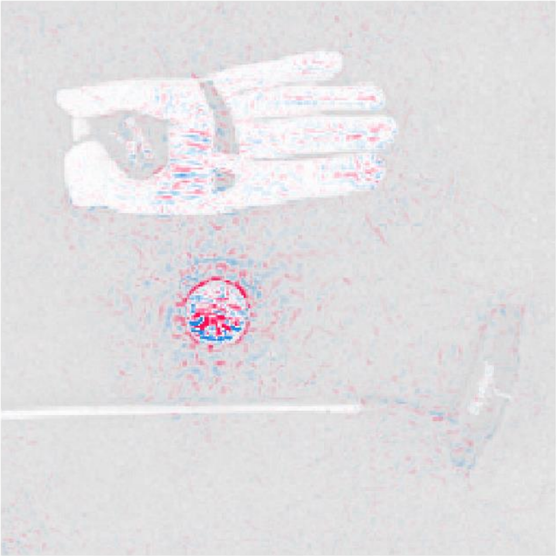
Cassette Player



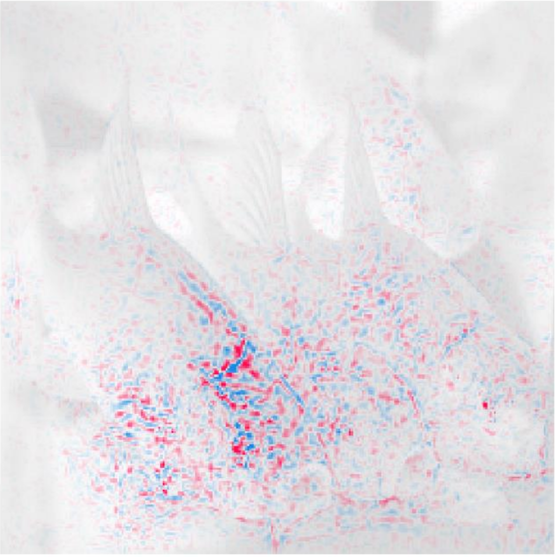
English Springer



Golf Ball



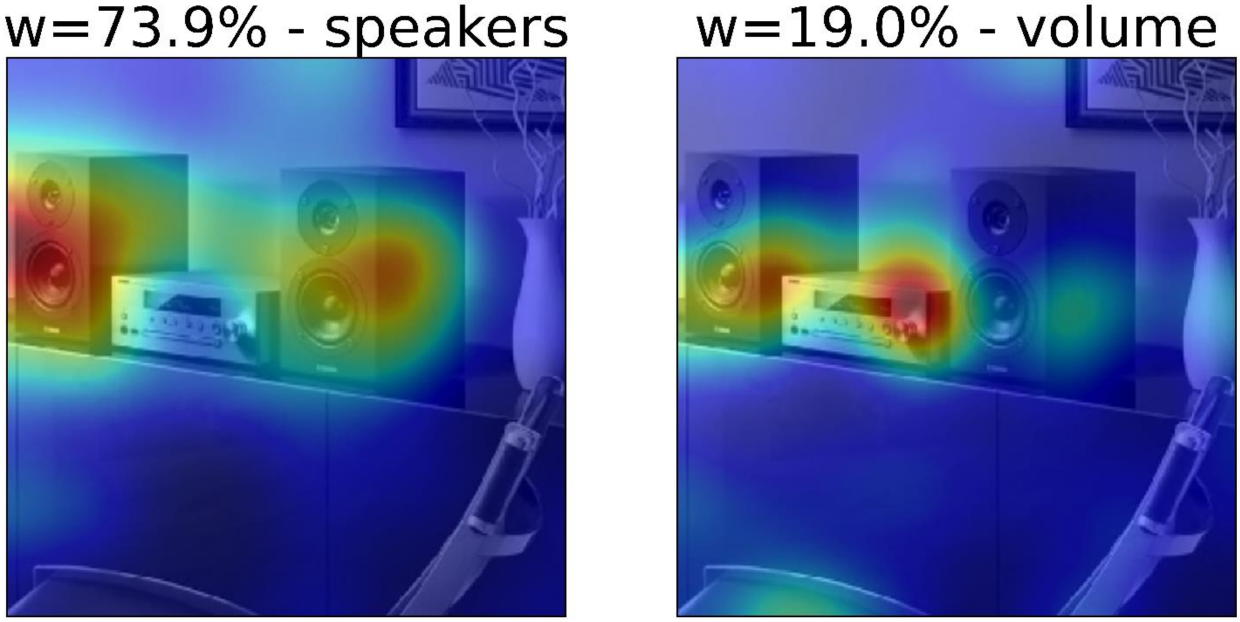
Tench



INV

This method extracts features that the network used to make the prediction, with their corresponding contribution towards the prediction (w). The features are represented as heatmaps, like Grad-CAM, and are associated with a name.

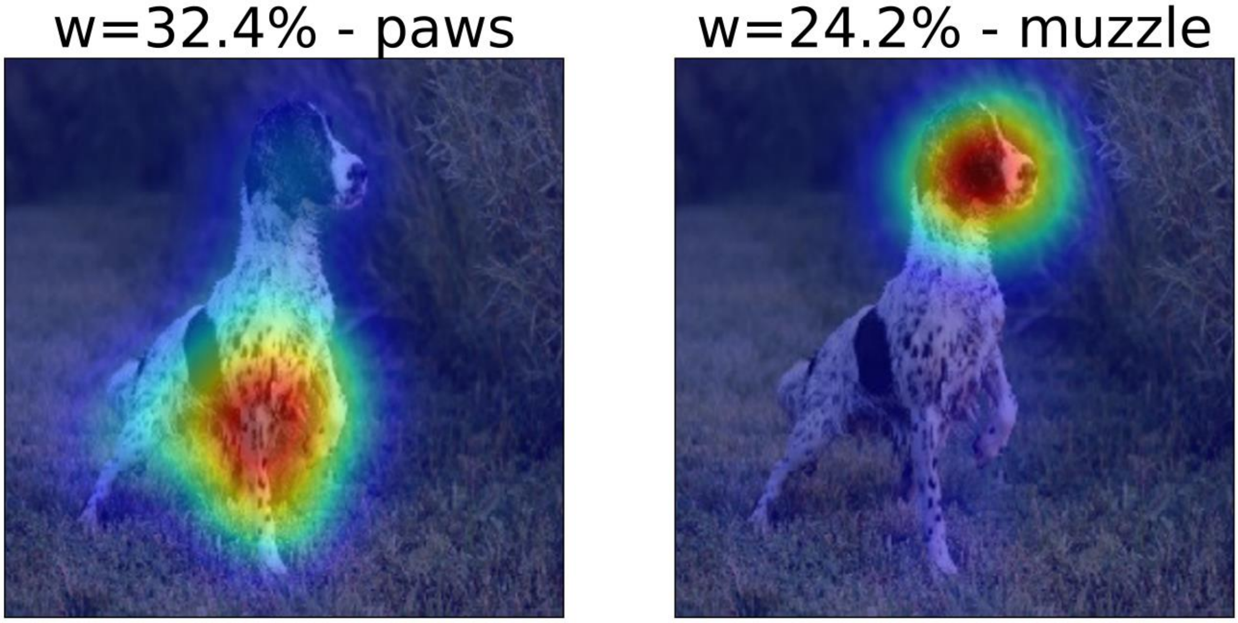
Cassette Player



w=73.9% - speakers

w=19.0% - volume

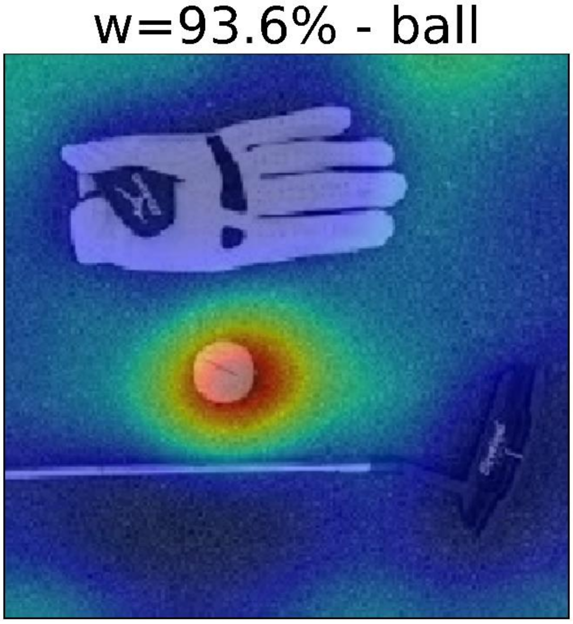
English Springer



w=32.4% - paws

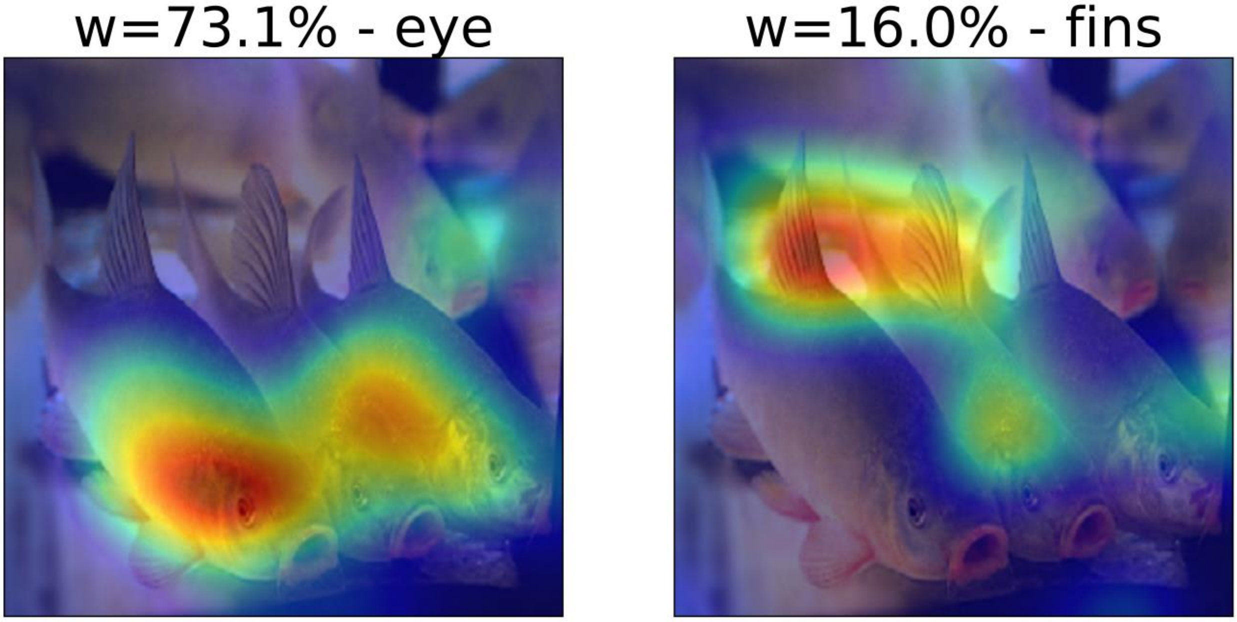
w=24.2% - muzzle

Golf Ball



w=93.6% - ball

Tench



w=73.1% - eye

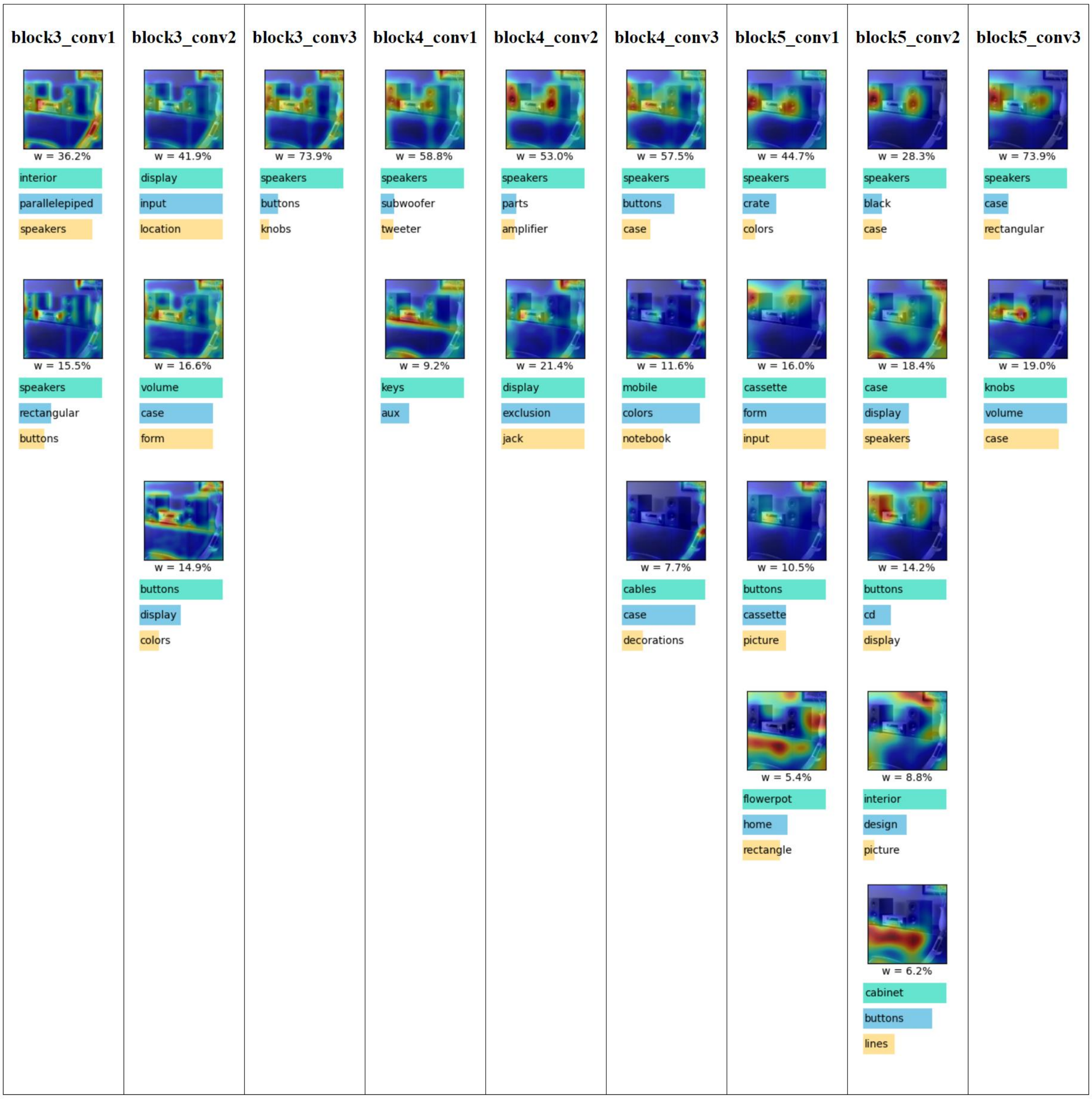
w=16.0% - fins

Extended INV

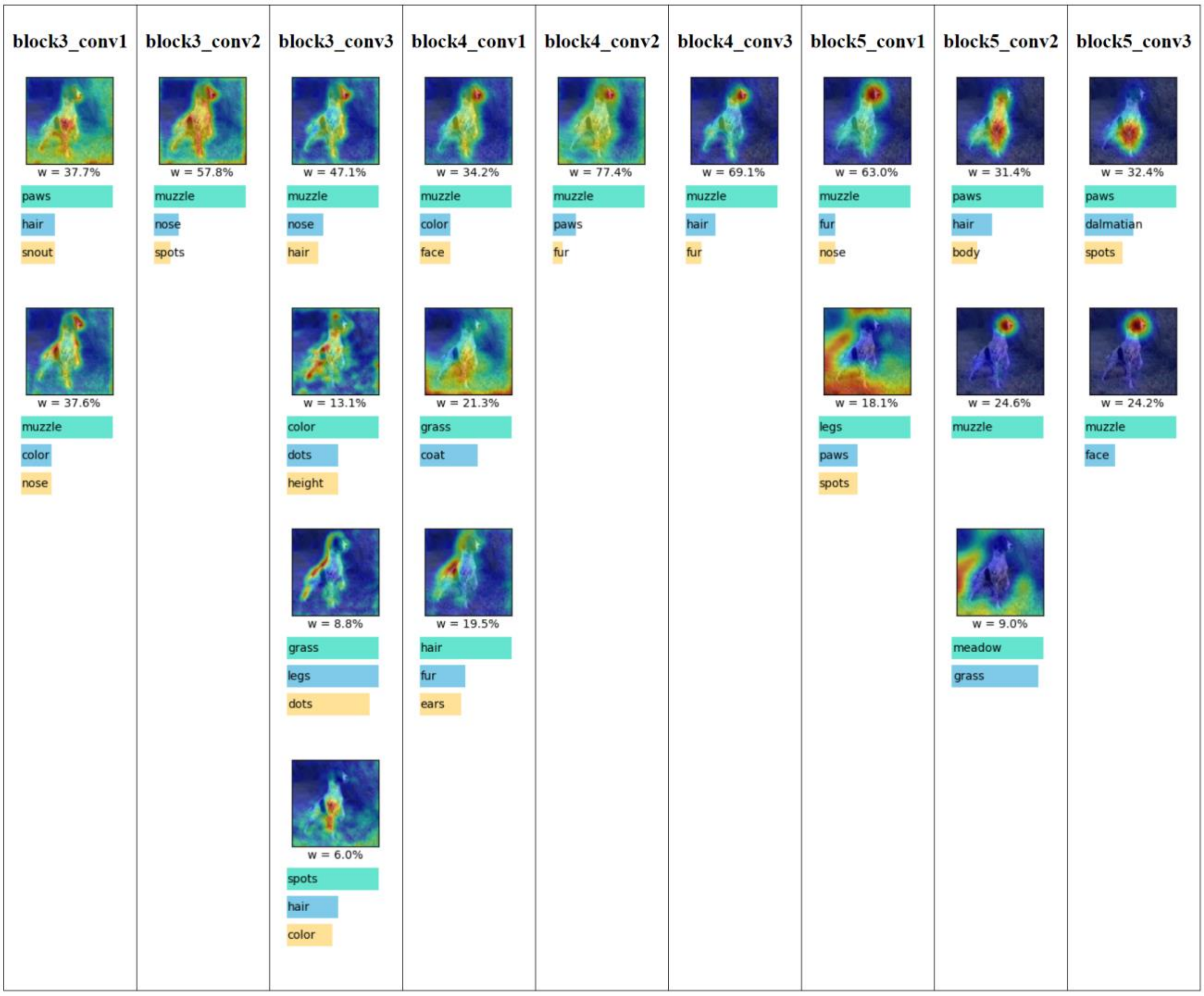
The building blocks of a neural network are called layers. They represent groups of neurons which extract features from the input image. The initial layer extract simple features like shapes, colours, etc., which are sent to the subsequent layer to extract more complex features that are more helpful towards making a correct prediction. This process is repeated for N layers, and finally an output is produced.

This explanation extends the previous technique, which considered only the final layer, by showing the features extracted also in earlier layers, with their corresponding weight and the labels that better describe the portion of image shown. Each column represents a layer, identified by its name (for instance, block5_conv3)

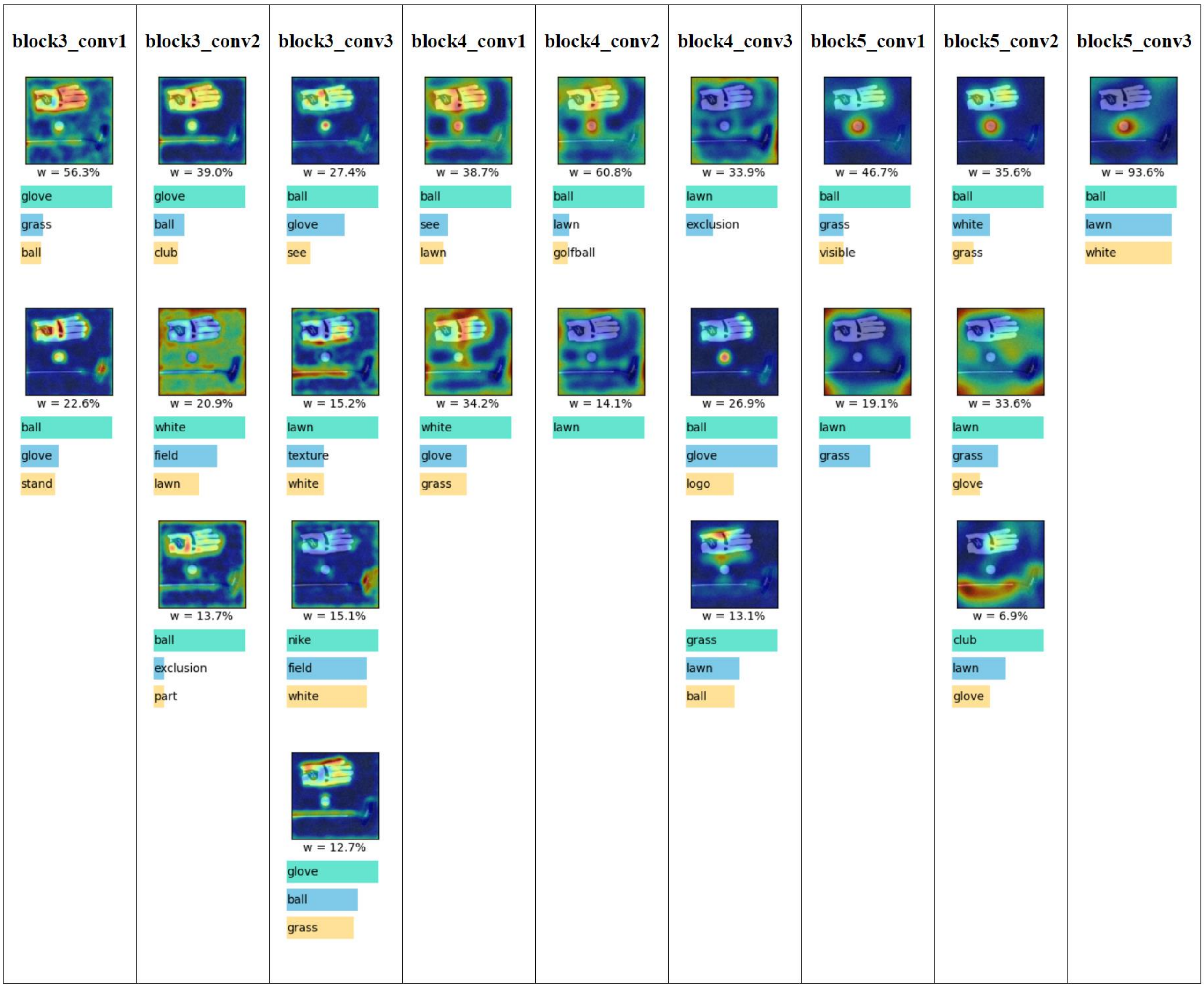
Cassette Player



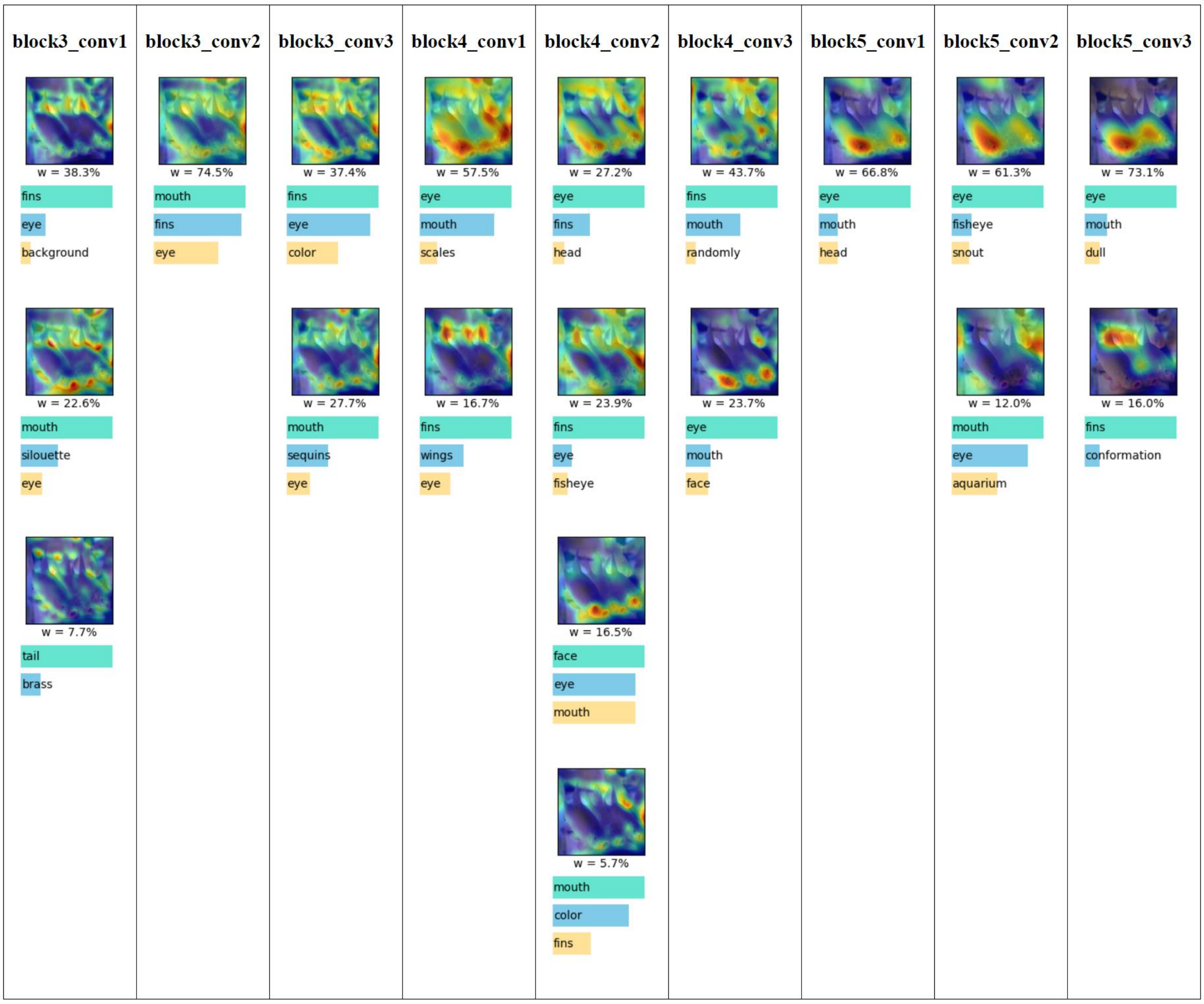
English Springer



Golf Ball


block3_conv1
w = 56.3%
glove
grass
ball
w = 22.6%
ball
glove
stand
block3_conv2
w = 39.0%
glove
ball
club
w = 20.9%
white
field
lawn
w = 13.7%
ball
exclusion
part
block3_conv3
w = 27.4%
ball
glove
see
w = 15.2%
lawn
texture
white
w = 15.1%
nike
field
white
w = 12.7%
glove
ball
grass
block4_conv1
w = 38.7%
ball
see
lawn
w = 34.2%
white
glove
grass
block4_conv2
w = 60.8%
ball
lawn
golfball
w = 14.1%
lawn
block4_conv3
w = 33.9%
lawn
exclusion
w = 26.9%
ball
glove
logo
w = 13.1%
grass
lawn
ball
block5_conv1
w = 46.7%
ball
grass
visible
w = 19.1%
lawn
grass
block5_conv2
w = 35.6%
ball
white
grass
w = 33.6%
lawn
grass
glove
w = 6.9%
club
lawn
glove
block5_conv3
w = 93.6%
ball
lawn
white

Tench


block3_conv1
w = 38.3%
fins
eye
background
w = 22.6%
mouth
silouette
eye
w = 7.7%
tail
brass
block3_conv2
w = 74.5%
mouth
fins
eye
block3_conv3
w = 37.4%
fins
eye
color
w = 27.7%
mouth
sequins
eye
block4_conv1
w = 57.5%
eye
mouth
scales
w = 16.7%
fins
wings
eye
block4_conv2
w = 27.2%
eye
fins
head
w = 23.9%
fins
eye
fisheye
w = 16.5%
face
eye
mouth
w = 5.7%
mouth
color
fins
block4_conv3
w = 43.7%
fins
mouth
randomly
w = 23.7%
eye
mouth
face
block5_conv1
w = 66.8%
eye
mouth
head
block5_conv2
w = 61.3%
eye
fisheye
snout
w = 12.0%
mouth
eye
aquarium
block5_conv3
w = 73.1%
eye
mouth
dull
w = 16.0%
fins
conformation