


# PREDICTING "PAWPULARITY" USING IMAGE ANALYTICS

Prepared for PetFinder.my   
December 2022



Prepared By -

Yash Bhirud, Shubham Garg, Ram G S, Saloni Jadhav and Amlendu Kumawat



# Agenda



- Project Background
- Data Sources and Exploratory Data Analytics
- Approach and Model Building
- Model Interpretation and Insights



# Project Background



## Situation and Challenges



- PetFinder.my is a platform that improves animal welfare and **promote adoption**.
- It is expected that pets with more attractive photos get adopted faster.
- Currently, a basic Cuteness Meter ranks pet photos.

## Objective and Decisions



- Our goal is to accurately determine a pet photo's attractiveness metric – "PAWPULARITY".
- Additionally, we aim to suggest recommendations to boost the current algorithm to improve chances of adoption.



# Data Sources

We have two types of data - **raw images** and **metadata**



## Raw Images

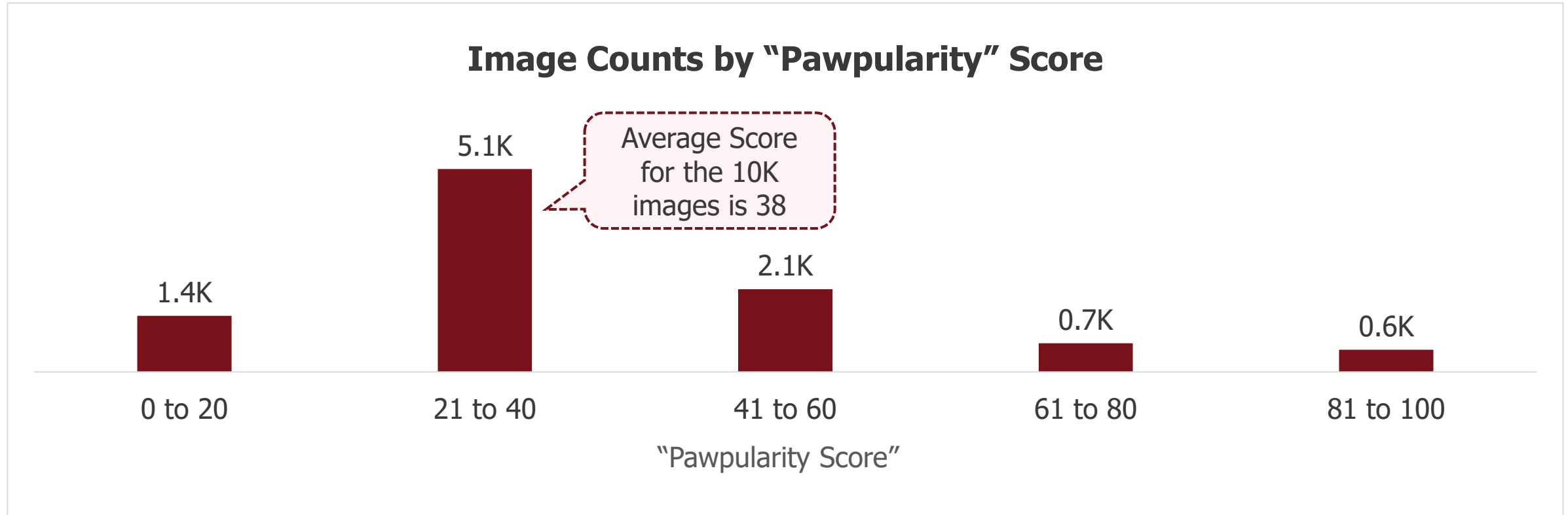


## Metadata

Id	Subject Focus	Eyes	Face	Near	Action	Accessory
0007de188	0	1	1	1	0	0
0009c66b9	0	1	1	0	0	0

Group	Collage	Human	Occlusion	Info	Blur	Pawpularity
1	0	0	0	0	0	63
0	0	0	0	0	0	42

# Exploratory Data Analytics (1/2) – Outcome Variable



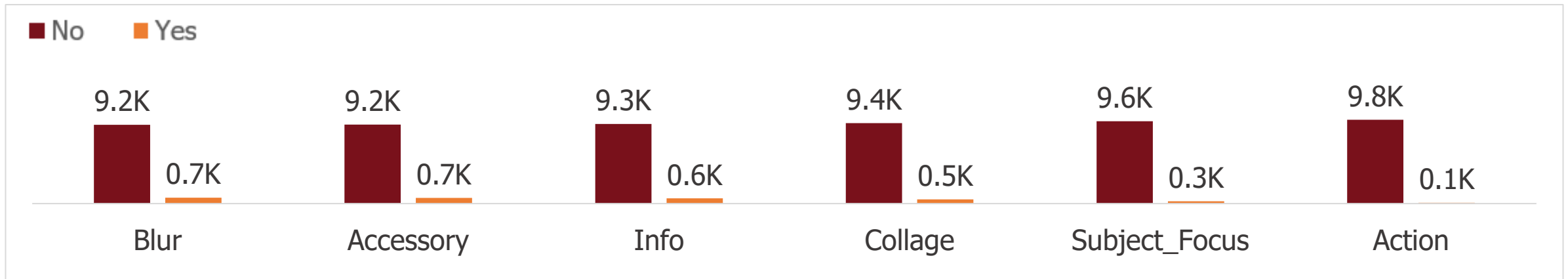
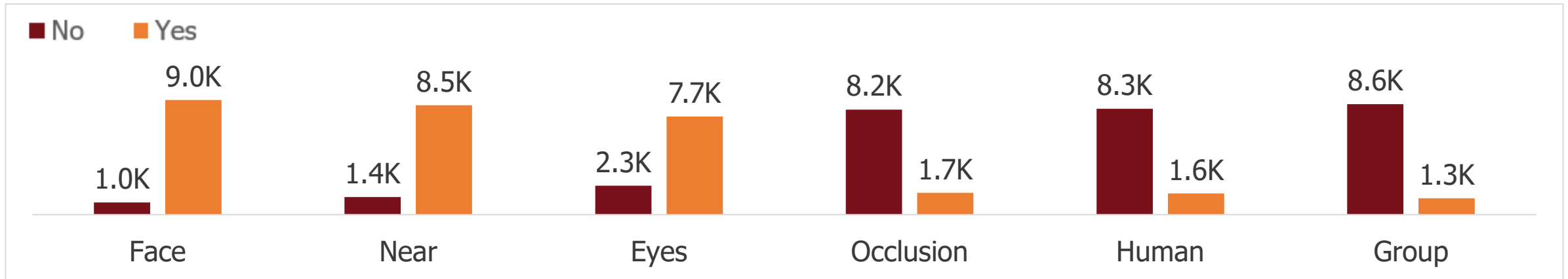
The “Pawpularity” score for the images is the outcome variable we’re trying to predict, and identify the image characteristics which lead to a higher score



# Exploratory Data Analytics (2 of 2) – Metadata



## Image Count by Attribute Presence



# Approach and Model Selection



**Goal 1:**  
**Build Predictive Model for Image**  
**"Pawpularity"**



**1a. CNN for Images** *(Convolutional Neural Network)*

- VGG16, RESNET50 and EfficientNet B7

**1b. Models for Tabular Data**

- Linear Regression, SVR, Random Forest

**Goal 2:**  
**Identify Top Image Features**



**2. Attention Mechanism**

- 2a. Vision Transformer
- 2b. Patch ConvNet



# 1. Predictive Model – Summary



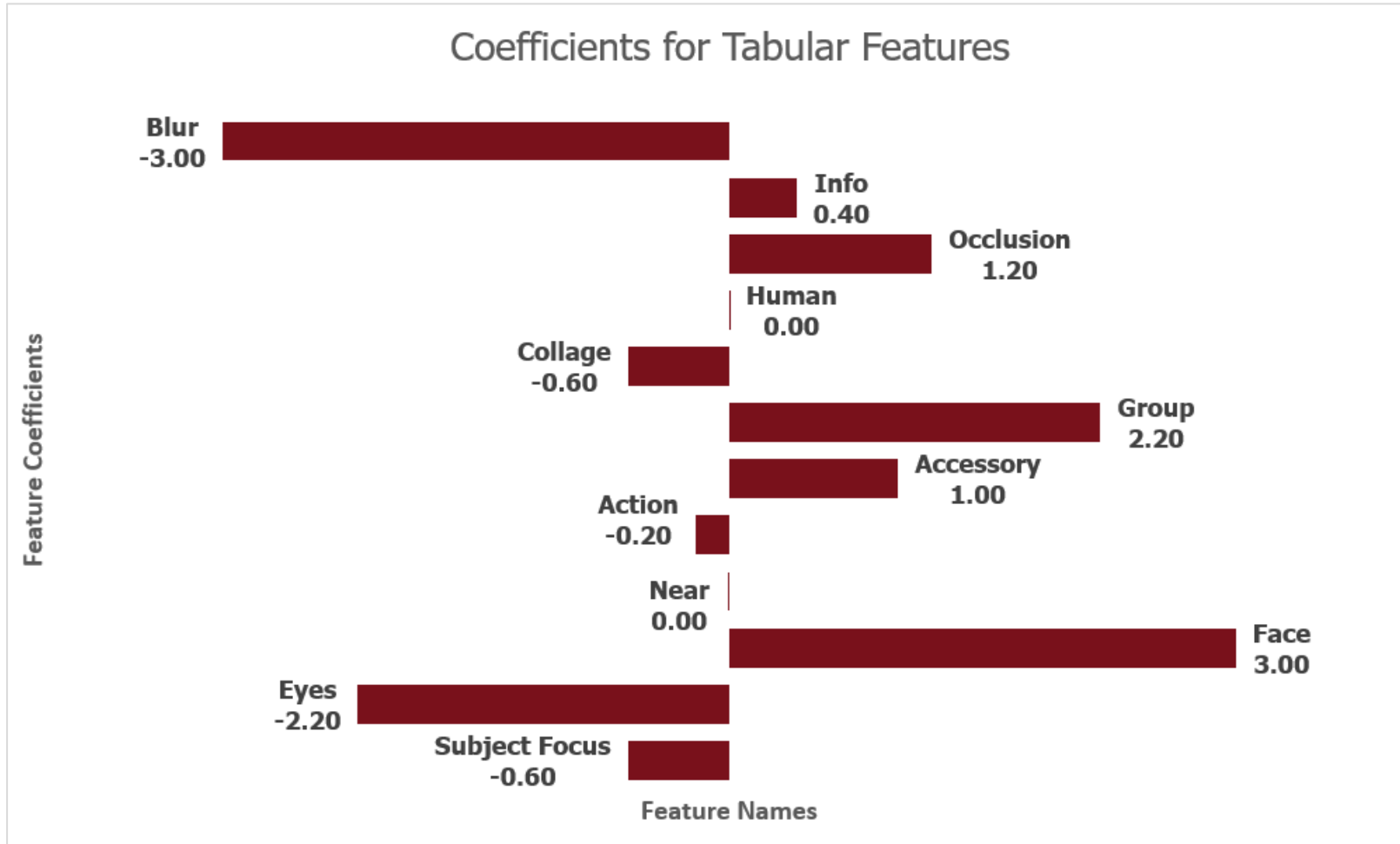
	1a. CNN for Images			1b. Models for Tabular Data		
<i>Model</i>	<b>VGG16</b>	<b>RESNET50</b>	<b>EfficientNet B7</b>	<b>Linear Regression</b>	<b>SVR</b>	<b>Random Forest</b>
<i>Training RMSE</i>	42.30	20.72	<b>20.64</b>	-	-	-
<i>Validation RMSE</i>	41.42	20.44	<b>20.33</b>	20.35	21.02	<b>20.34</b>
<i>Time Taken</i>	21 mins.	18 mins.	22 mins.	<1 min	<1 min	<1 min

“**EfficientNet B7**” and “**Random Forest**” stacked model give us the best models based on Validation RMSE  
This models are stacked which lead to **Test RMSE of 20.86**





# 1. Predictive Model – Insights



- Attributes like – **Face and Group**, have **positive** coefficients and **increase** “Pawpularity” score
- On the other hand, attributes like – **Blur and Collage** affect it **negatively**



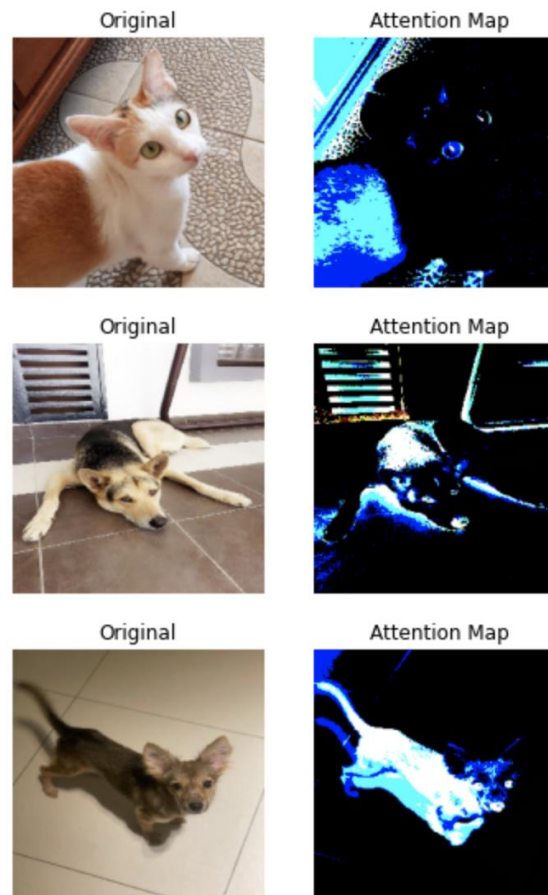


## 2. Attention Mechanism – Model Summary

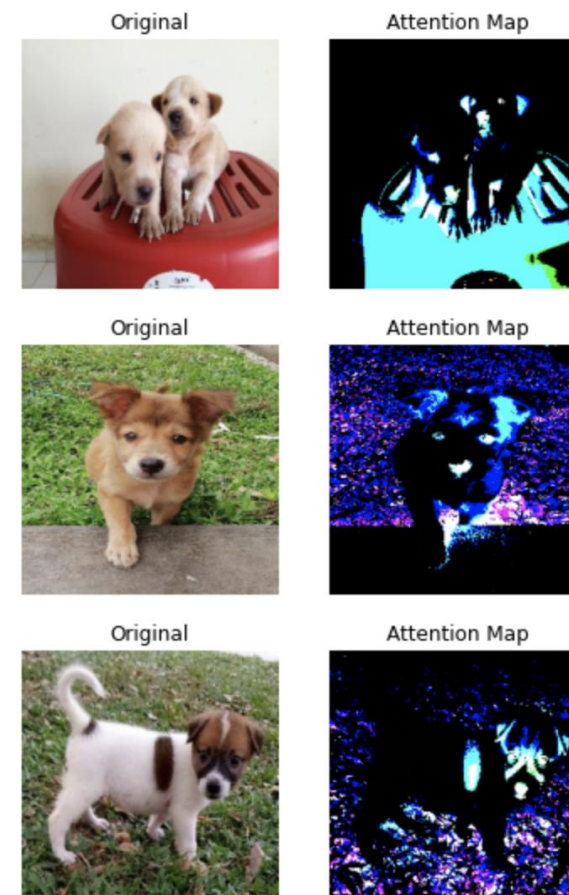
	Attention Mechanism	
Model	Vision Transformer	Patch ConvNet
Training RMSE	21.02	20.75
Validation RMSE	20.35	22.25
Time Taken	51 mins.	21 mins.

***Vision Transformer results were promising – only second to EfficientNet B7***

### Low "Pawpularity"



### High "Pawpularity"



## 2. Attention Mechanism – Insights



- Attention maps were analyzed for images with high and low Pawpularity, and we see that the parts of the image used for machine learning are identical
- Calculating pet popularity scores is tricky as cuteness is very subjective, a difficult task for even the most complicated Neural networks
- A more appropriate target variable to use would be a subjective one like 'Number of clicks on the image'
- Nevertheless, based on attention maps of various images from Vision Transformer, we identified the following features to be attended the most –
  - Eyes 🗨️
  - Nose 🍷
  - Ears 🗨️
  - Paws 🐾



**THANK YOU!**

**That was Machine Learning at it's best!**

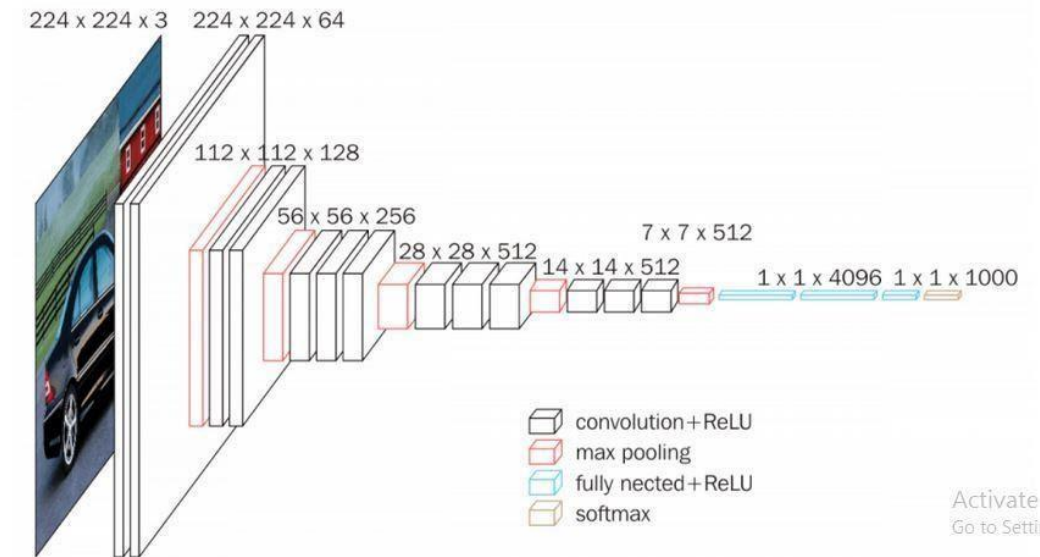


# APPENDIX



# Model Building (1 of 5) – VGG16

## Architecture



## VGG-16

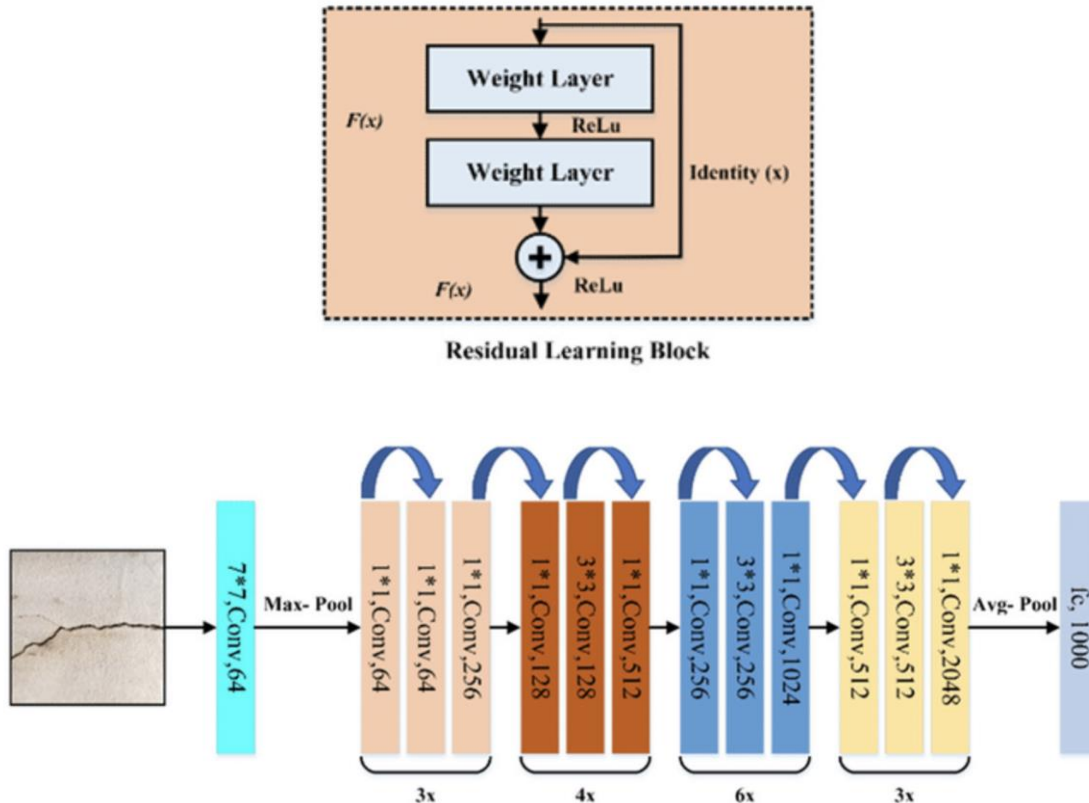


## What is it?

- VGG16 is a type of CNN that is 16 layers deep
- The creators of this model evaluated the networks and increased the depth using an architecture with very small ( $3 \times 3$ ) convolution filters, which showed a significant improvement on the prior-art configurations
- They pushed the depth to 16–19 weight layers making it close to 138 trainable parameters
- **Fact** – VGG16 is object detection and classification algorithm which is able to classify 1000 images of 1000 different categories with 92.7% accuracy.

# Model Building (2 of 5) – ResNet 50

## Architecture



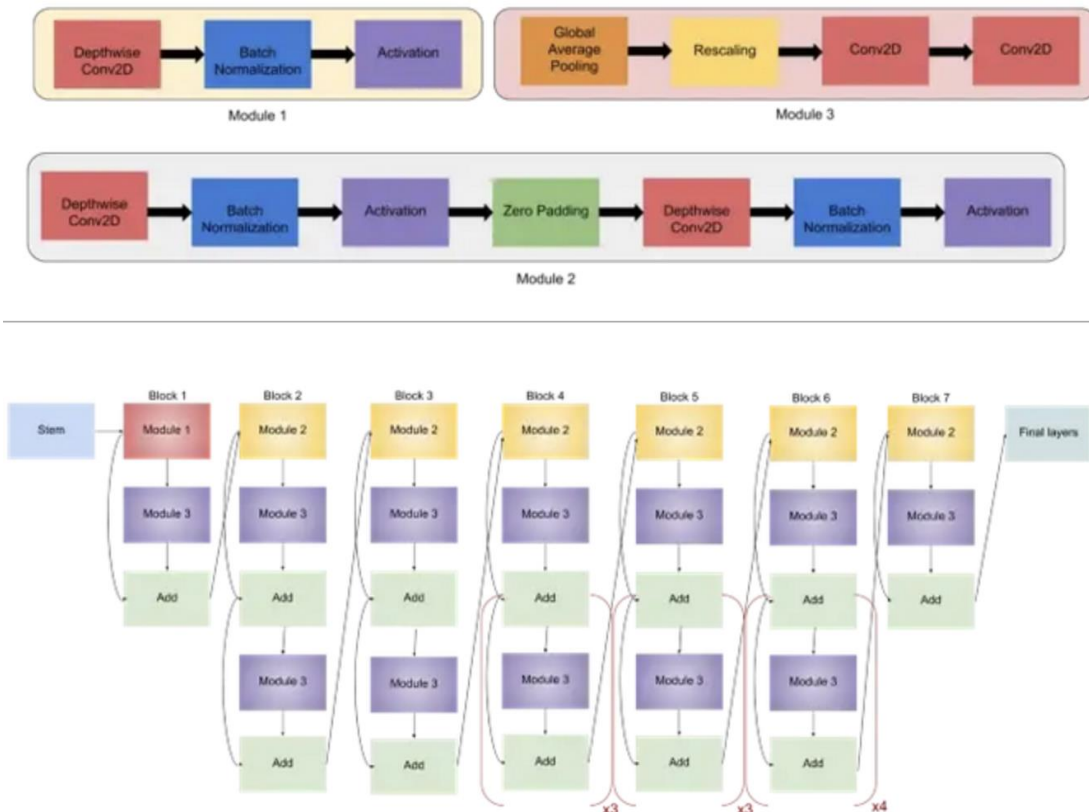
## What is it?

- ResNet stands for Residual Network and is a specific type of CNN.
- It is a 50-layer CNN - 48 convolutional layers, one MaxPool layer, and one average pool layer.
- It uses a residual block of  $1 \times 1$  convolutions, known as a “bottleneck”, which reduces the number of parameters and matrix multiplications. This enables much faster training of each layer.
- **Fact** – ResNet50 model was the winner of ImageNet challenge in 2015. The fundamental breakthrough, was it allowed us to train extremely deep neural networks with 150+layers.



# Model Building (3 of 5) – Efficient Net

## Architecture



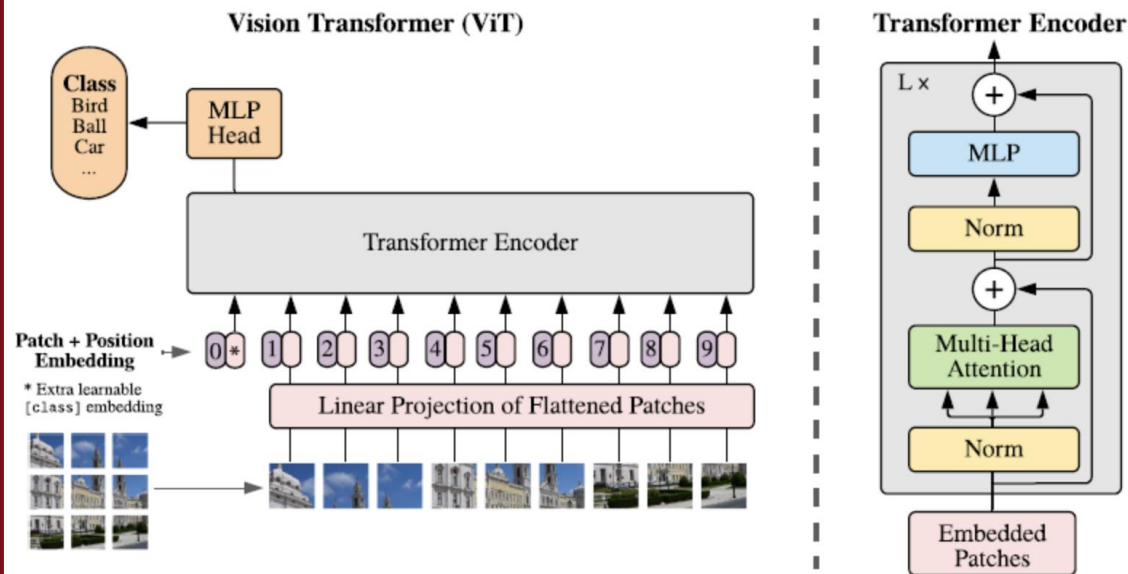
## What is it?

- Efficient Net is a CNN architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient.
- The compound scaling method is justified by the intuition that if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image.
- Fact** – EfficientNet-B7 achieves state-of-the-art 84.4% top-1 / 97.1% top-5 accuracy on ImageNet, being 8.4x smaller and 6.1x faster on inference than the best existing ConvNet.



# Model Building (4 of 5) – Vision Transformer

## Architecture



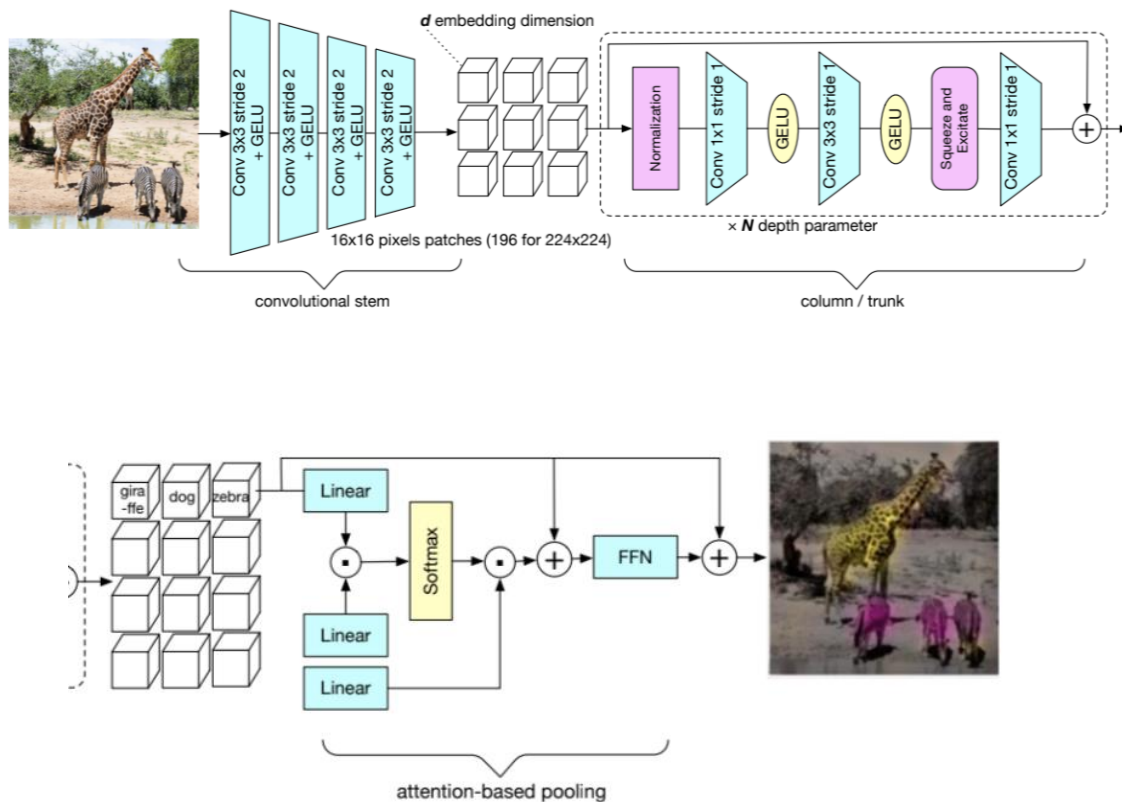
## What is it?

- ViTs overcome some of the drawbacks of CNNs, especially inductive bias associated with pixel positions.
- To overcome this, images are divided into patches, embedded with positional representations and passed through an attention layer that weighs the importance of each patch in an image.

**Fact** - In 2020 Vision Transformers were then adapted for tasks in Computer Vision with the paper "An image is worth 16x16 words". The idea is basically to break down input images as a series of patches which, once transformed into vectors, are seen as words in a normal transformer.

# Model Building (5 of 5) – Patch ConvNet

## Architecture



## What is it?

- Substitute the global average pooling layer of a convnet with a Transformer layer. The self-attention layer of the Transformer would produce attention maps that correspond to the most attended patches of the image for the classification decision.
- We minimally implement the ideas of Augmenting Convolutional networks with attention-based aggregation. The simple design for the attention-based pooling layer, such that it explicitly provides the weights (importance) of the different patches.
- **Fact** – The novel architecture of convnet called the PatchConvNet which deviates from the age old pyramidal architecture.