MATH 4995 Project 2:
Pawpularity Prediction Using Machine Learning with Tabular Metadata and Images

Presenter: SHAO Zhihao
Date: 11/23/2021

Table of Contents:

Table of Contents:

Introduction:

Millions of stray animals suffer on the streets or are euthanized in shelters every day around the world. You might expect pets with attractive photos to generate more interest and be adopted faster.

PetFinder.my is Malaysia's leading animal welfare platform, featuring over 180,000 animals with 54,000 happily adopted. PetFinder collaborates closely with animal lovers, media, corporations, and global organizations to improve animal welfare.

But what makes a good picture?

With the help of data science, you may be able to accurately determine a pet photo's appeal and even suggest improvements to give these rescue animals a higher chance of loving homes.

Evaluation metric:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Table of Contents:

Dataset Description:

Photo metadata:

- Manually labeling each photo for key visual quality and composition parameters
- 12 predictors

The **train.csv** and **test.csv** files contain metadata for photos in the training set and test set, respectively. Each pet photo is labeled with the value of **1** (Yes) or **0** (No) for each of the following features:

- **Focus** - Pet stands out against uncluttered background, not too close / far.
- **Eyes** - Both eyes are facing front or near-front, with at least 1 eye / pupil decently clear.
- **Face** - Decently clear face, facing front or near-front.
- **Near** - Single pet taking up significant portion of photo (roughly over 50% of photo width or height).
- **Action** - Pet in the middle of an action (e.g., jumping).
- **Accessory** - Accompanying physical or digital accessory / prop (i.e. toy, digital sticker), excluding collar and leash.
- **Group** - More than 1 pet in the photo.
- **Collage** - Digitally-retouched photo (i.e. with digital photo frame, combination of multiple photos).
- **Human** - Human in the photo.
- **Occlusion** - Specific undesirable objects blocking part of the pet (i.e. human, cage or fence). Note that not all blocking objects are considered occlusion.
- **Info** - Custom-added text or labels (i.e. pet name, description).
- **Blur** - Noticeably out of focus or noisy, especially for the pet's eyes and face. For Blur entries, "Eyes" column is always set to 0.

Dataset Description:

Image data:

- The training data contains 9912 pictures in total, each with a pawpularity score.
- The Pawpularity Score is derived from each pet profile‘s page view statistics at the listing pages

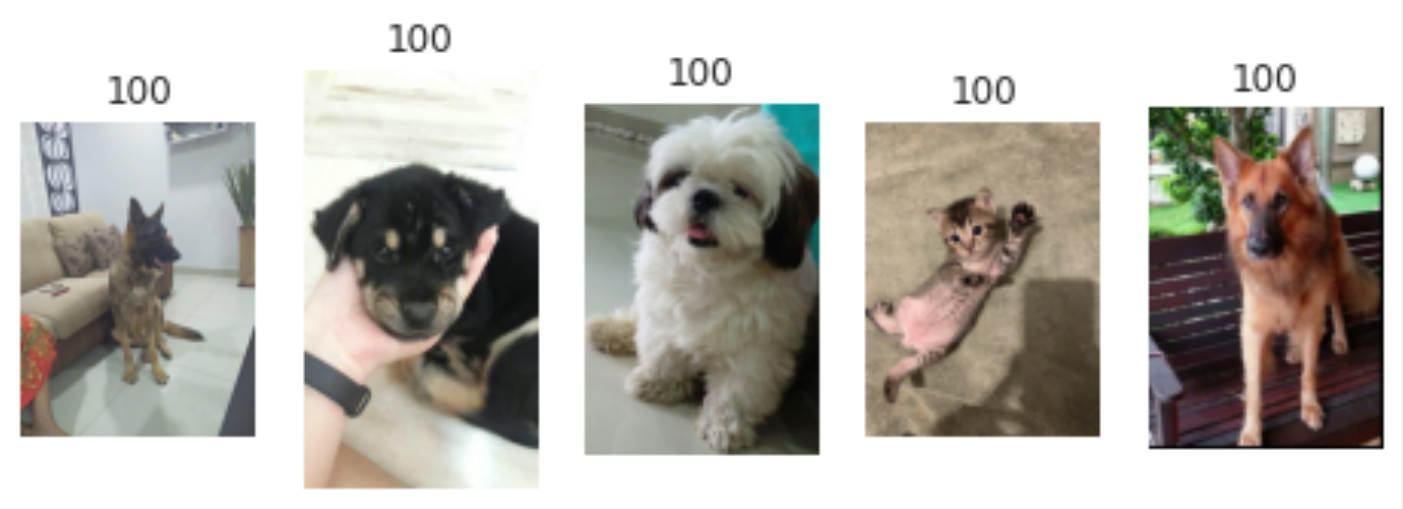


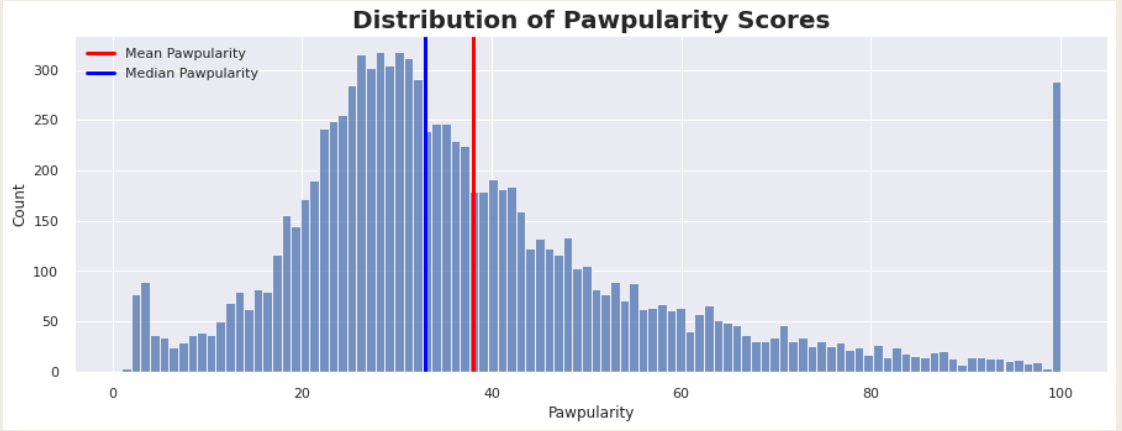
Table of Contents:

Exploratory Data Analysis:

Photo metadata

- **Response variable – Pawpularity score:**

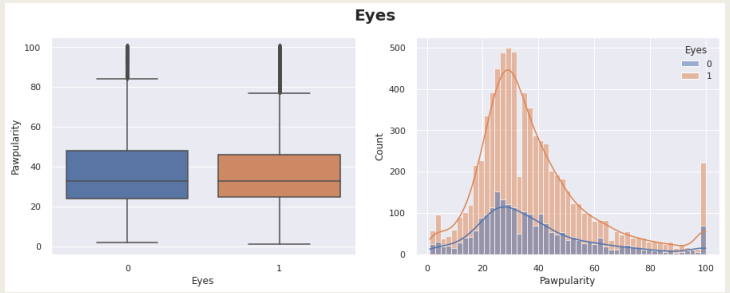
The distribution of target variable – pawpularity score is slight skewed to the left, and plenty of samples have 100 score.



- **Predictors:**

The distribution of pawpularity score is similar for each predictor and class

(Changing the features doesn't seem to influence the pawpularity scores that much)



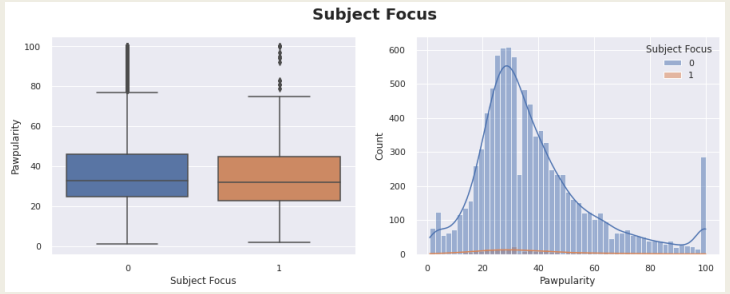
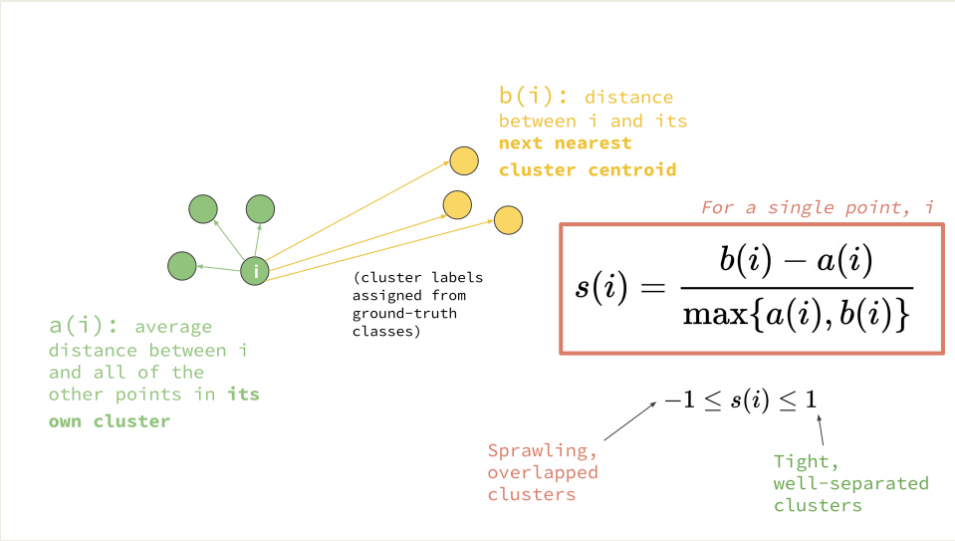


Table of Contents:

Regression with Tabular Metadata:

Feature Engineering

- Add normalized size and shape of images
  - *Normalized by MinMaxScaler*
- Add K-means clustering results
  - *K = 8 by* ***Silhouette score***
- Add PCA features
  - *Top 15 PCA features explained about 100% variance*



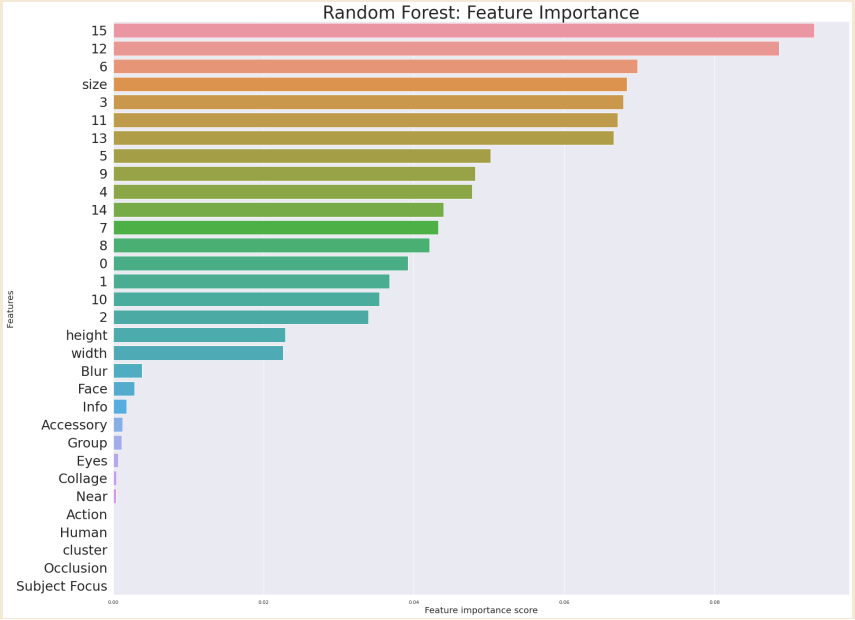
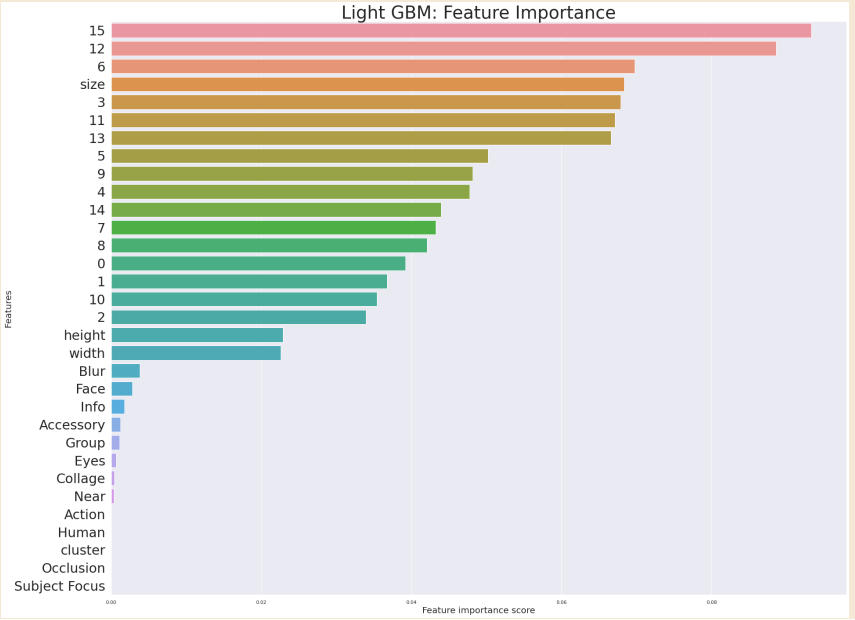
Model Selection

- Random Forest
- Light Gradient Boosting Machine (GBM)
- Voting regression – Ensemble method

Regression with Tabular Metadata:

Hyperparameter tuning

- Method: Five-fold cross validation

Model	Random Forest	Light GBM
Feature Importance		

- Equally weighted for voting regression

Regression with Tabular Metadata:

Results

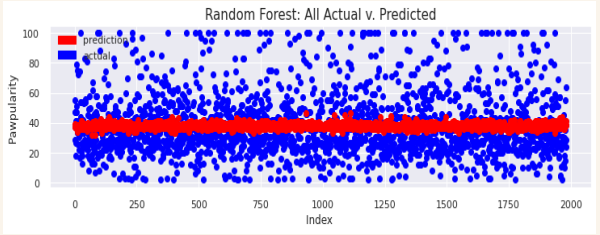
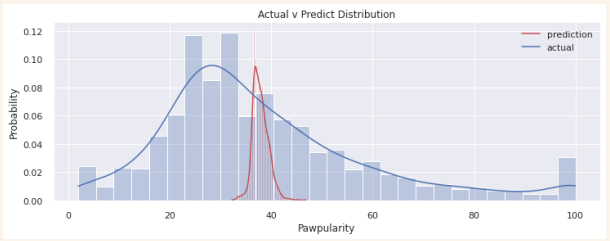
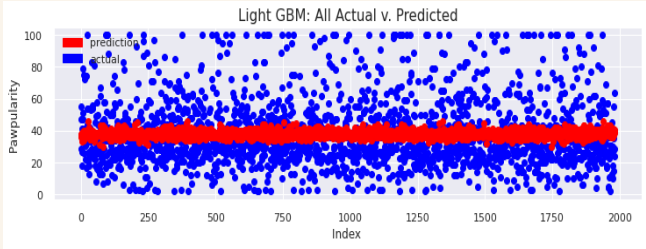
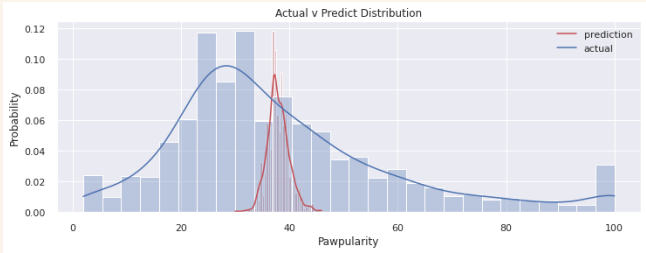
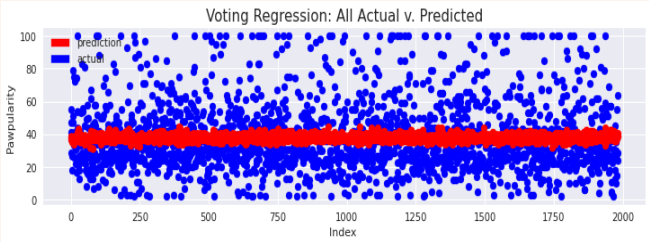
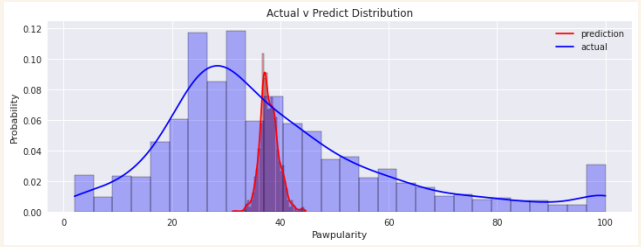
Model	Random Forest	Light GBM	Voting Regression
RMSE (test split)	20.787	20.725	20.746
Prediction v.s. Actual scores	 	 	 

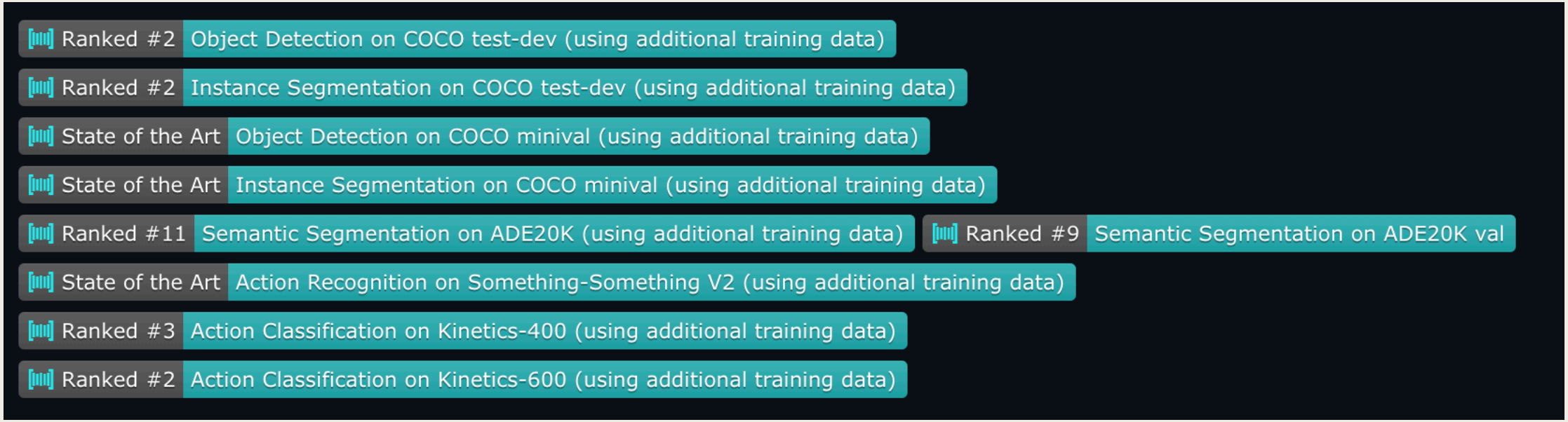
Table of Contents:

Swin Transformer:

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Ze Liu[†*] Yutong Lin[†*] Yue Cao[*] Han Hu[*‡] Yixuan Wei[†]
Zheng Zhang Stephen Lin Baining Guo
Microsoft Research Asia
{v-zeliu1,v-yutlin,yuecao,hanhu,v-yixwe,zhez,stevelin,bainguo}@microsoft.com

ICCV 2021 best paper award



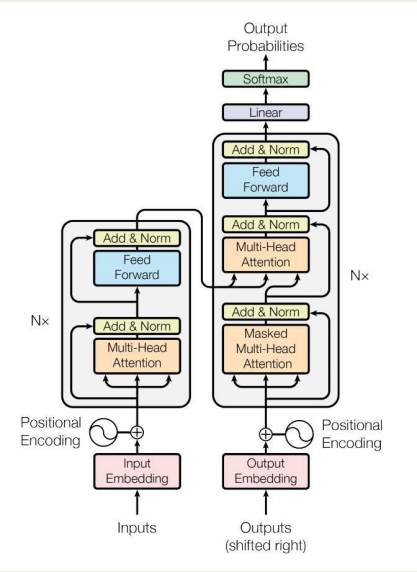
https://arxiv.org/pdf/2103.14030.pdf

Swin Transformer:

Disadvantages of Transformers for image data:

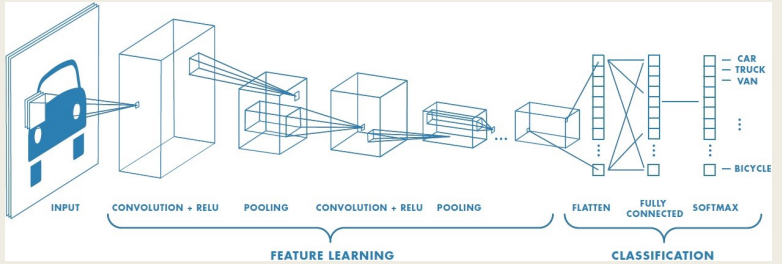
- Unlike word tokens in NLP, visual elements have various scales. Especially in object detection;
- Images are in much higher resolution and vision tasks that require pixel-level predictions are intractable for transformers because the computation complexity(+memory usage) of self-attention is quadratic to image size.

Transformer



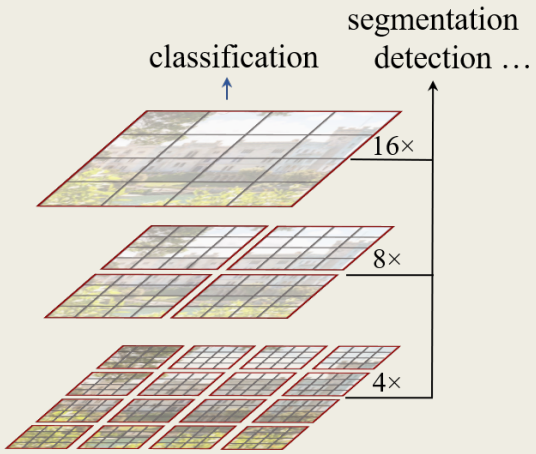
VS

CNN

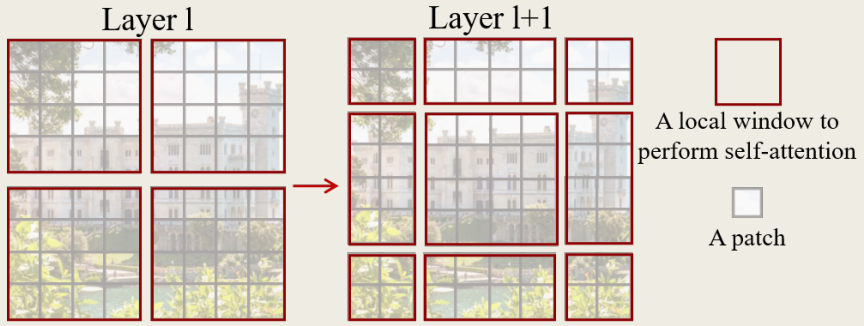


Swin Transformer:

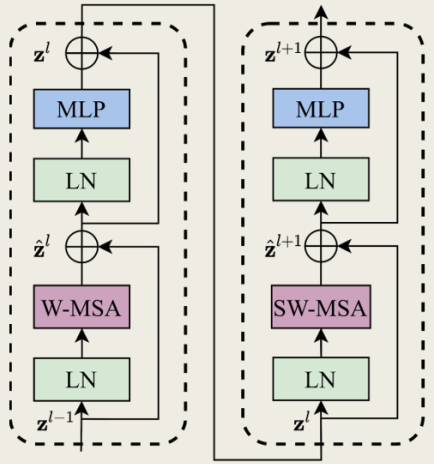
Novelties



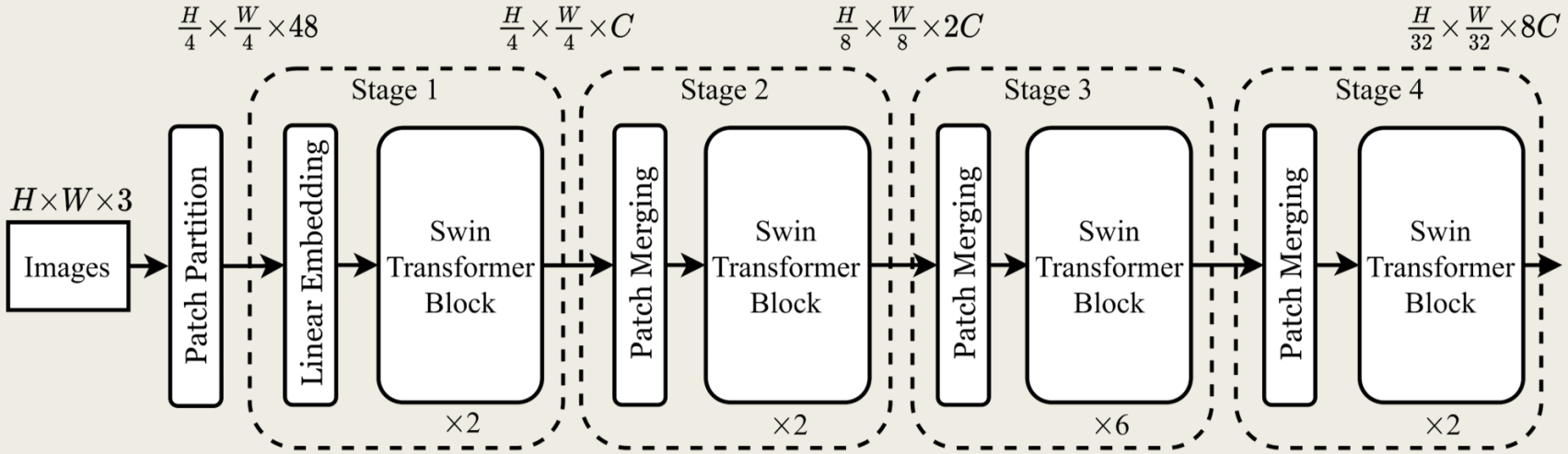
(a) Swin Transformer



(b) Shifted Window

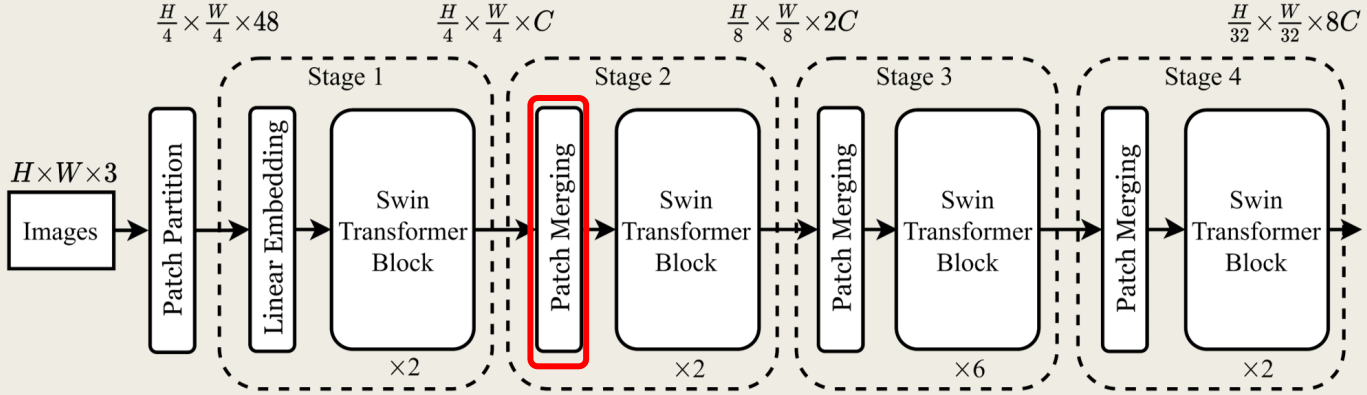


(c) Two Successive Swin Transformer Blocks



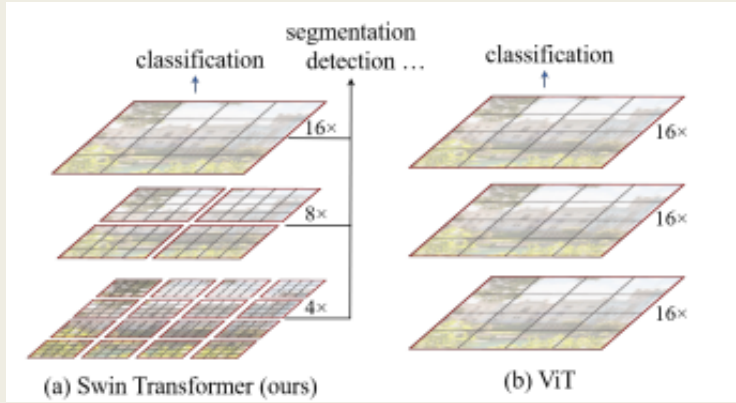
(d) Architecture

Swin Transformer:



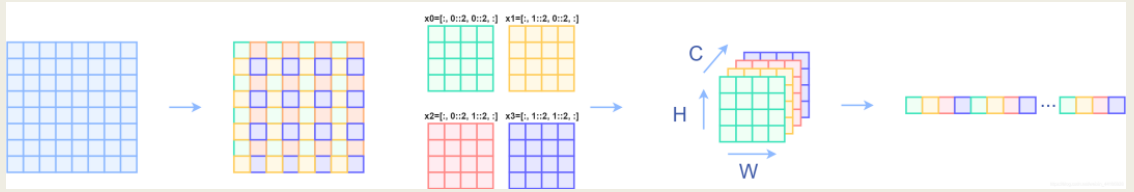
1. Hierachical Representation

- Detect objects of various scales efficiently



The number of tokens is reduced by patch merging layers which concatenates the features of 2 × 2 neighboring patches as the network gets deeper.

In contrast, previous vision Transformers produce feature maps of a single low resolution.



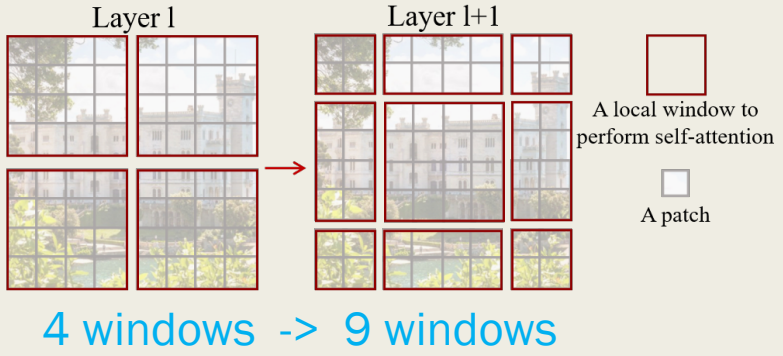
Patch merging layers in Transformer ⇔ Pooling in CNN

https://cloud.tencent.com/developer/article/1819813

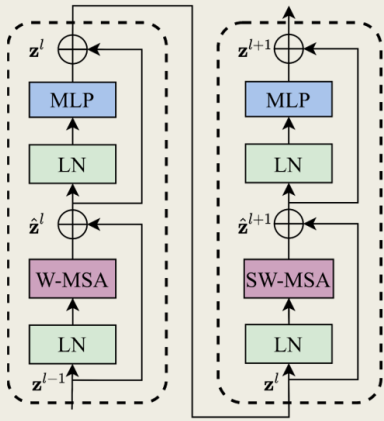
Swin Transformer:

2. Shifted-Window Multi-head Self-Attention (SW-MSA)

- Lower computation complexity



4 windows -> 9 windows



The shifted window aims to compute self-attention within local windows.

During the Swin transformer blocks, the network alternates between standard window configuration(W-MSA) and shited window configuration(SW-MSA).

This approach introduces connections between neighboring overlapping windows just like how deep convolutions work.

$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C,$$
$$\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC,$$

Swin Transformer:

3. Efficient shifting

- Lower computation complexity

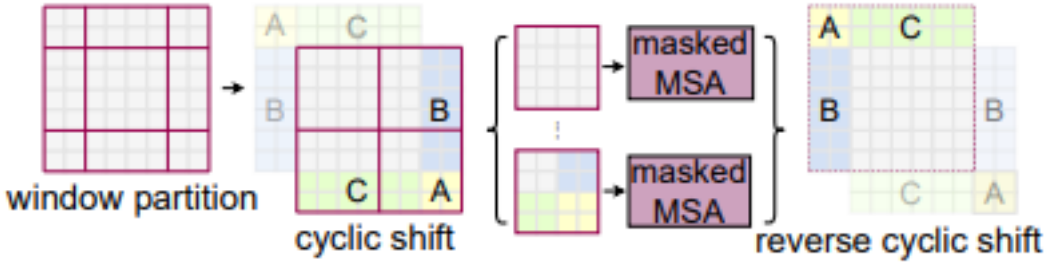
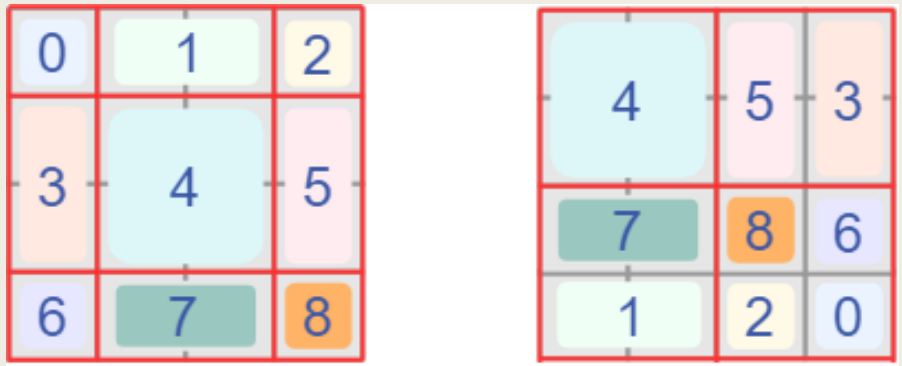


Figure 4. Illustration of an efficient batch computation approach for self-attention in shifted window partitioning.

For efficient processing of edge windows smaller than M × M, the paper applies cyclic-shifting before computing self-attention.

A masking mechanism is applied to the partitions so that computation is limited within each original window.

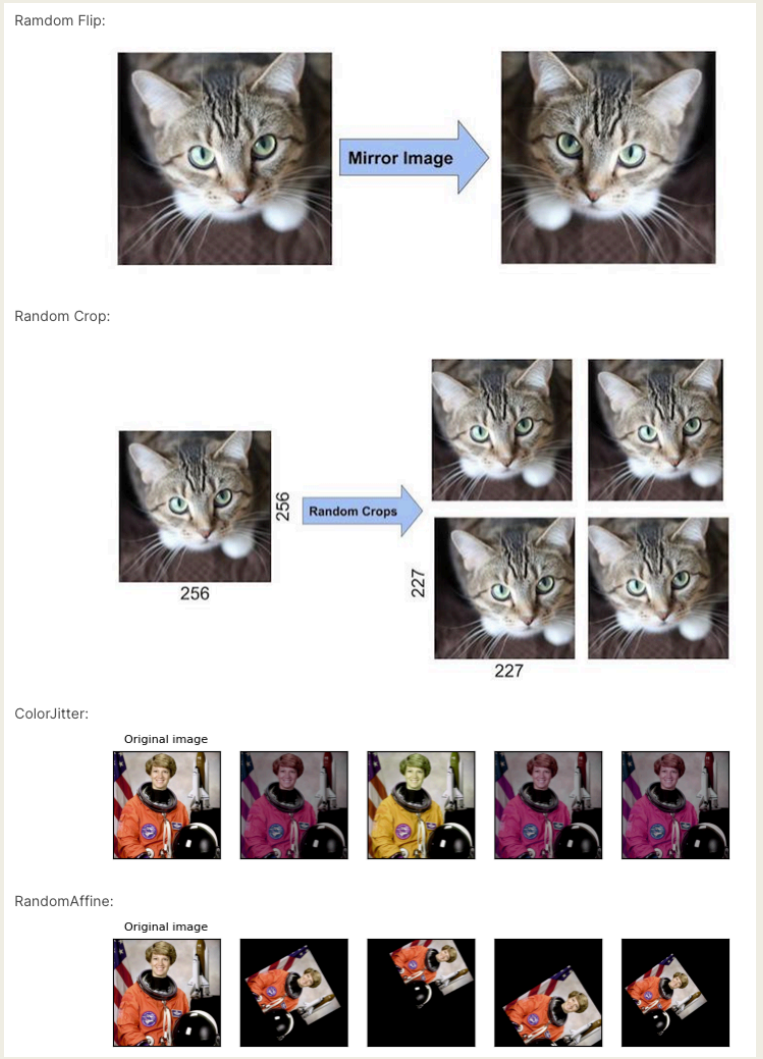


https://cloud.tencent.com/developer/article/1819813

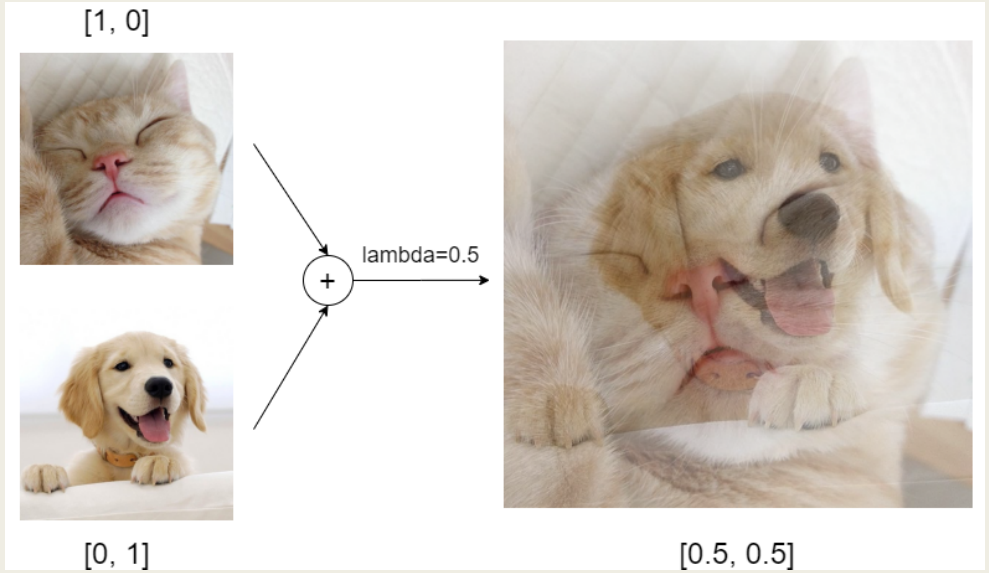
Swin Transformer with Image data:

Image data augmentation

- Random flip/crop/affine, ColorJitter



- MixUp



Swin Transformer with Image data:

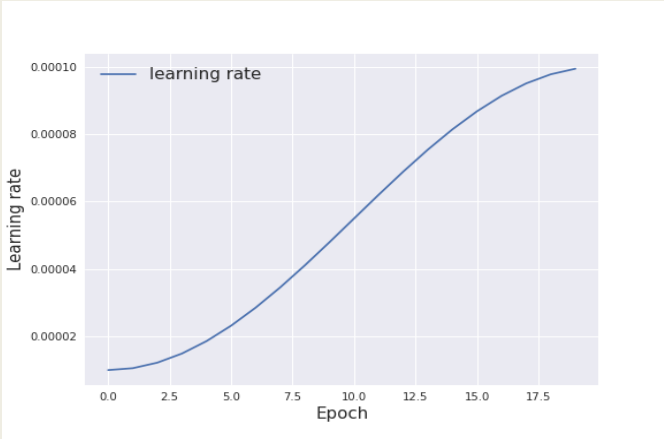
Model Training

- **LR scheduler**: CosineAnnealingWarmRestarts
- **Optimizer**: AdamW
  AdamW yields better training loss, and the models generalize much better than models trained with Adam
- **Train loss**: BCEWithLogitsLoss
  If we directly use the output of the model as our predicted Pawpularity, then this value can be negative (which is not reasonable). However, if we use BCEWithLogitsLoss, its embedded Sigmoid function will transform the model output to [0,1], which is a reasonable region for Pawpularity Score (normalized by 100).
- Max. epoch: 20
- Five-fold Cross Validation

Swin Transformer with Image data:

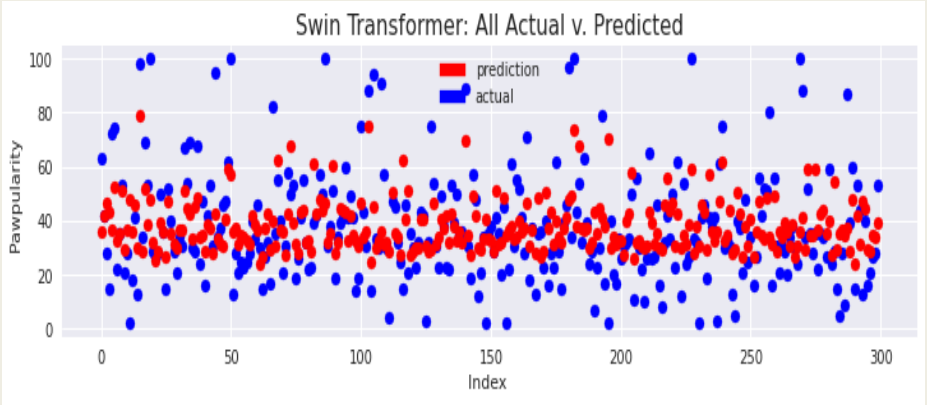
Results: RMSE = 18.366 on Kaggle (1053/2260)

Learning rate



Training & Validation loss

Prediction v.s. Acutual scores



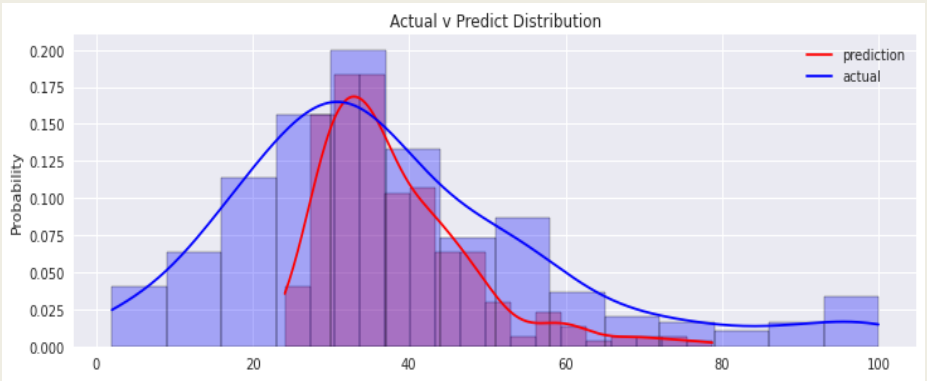


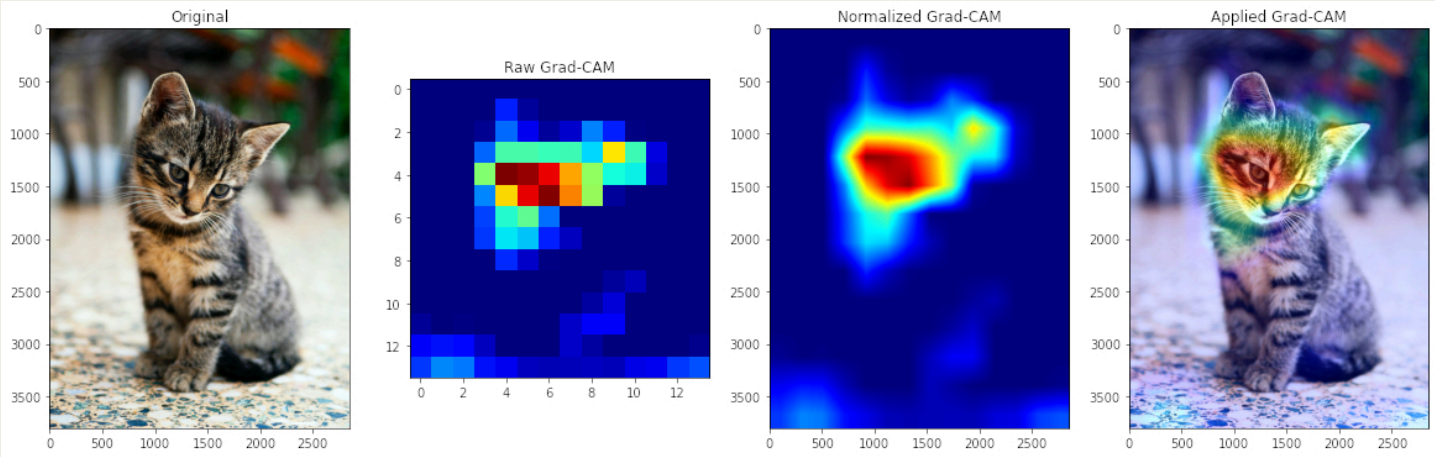
Table of Contents:

Analysis:

Grad-Cam:

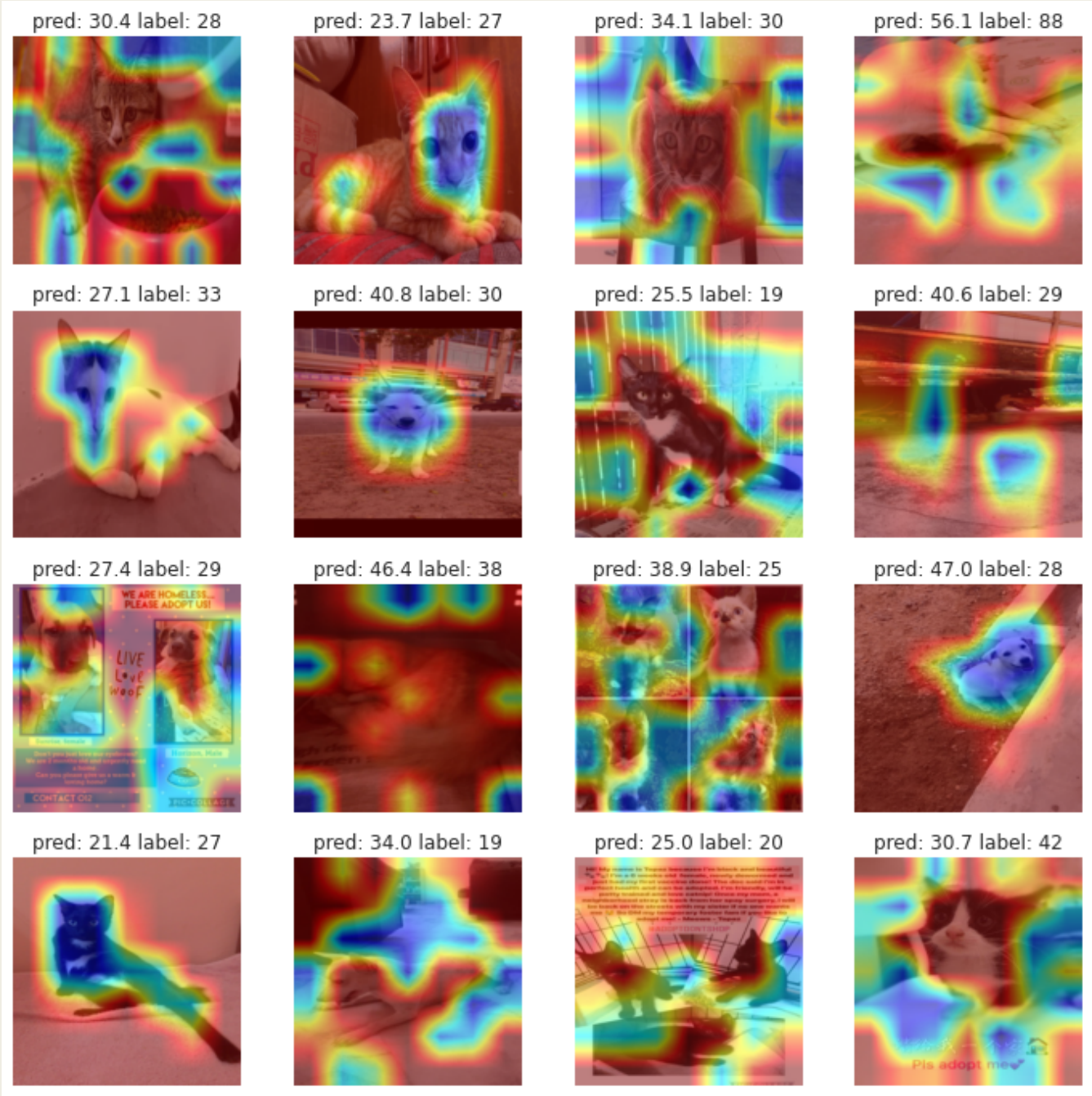
While CNN enable superior performance, their lack of decomposability into intuitive and understandable components makes them hard to interpret.

- Use the gradient information flowing into the last convolutional layer of the CNN to understand each neuron for a decision of interest.
- Explain why they predict what they predict.
- SwinT: model.layers[-1].blocks[-1].norm1



Analysis:

Grad-Cam:



Analysis:

- Only two weak learners for the voting ensemble
- Predictions based on image data are genuinely better than those based on tabular features
- Methods integrating these two data sources may maximize their prediction power.

The end.

Any question is welcomed!