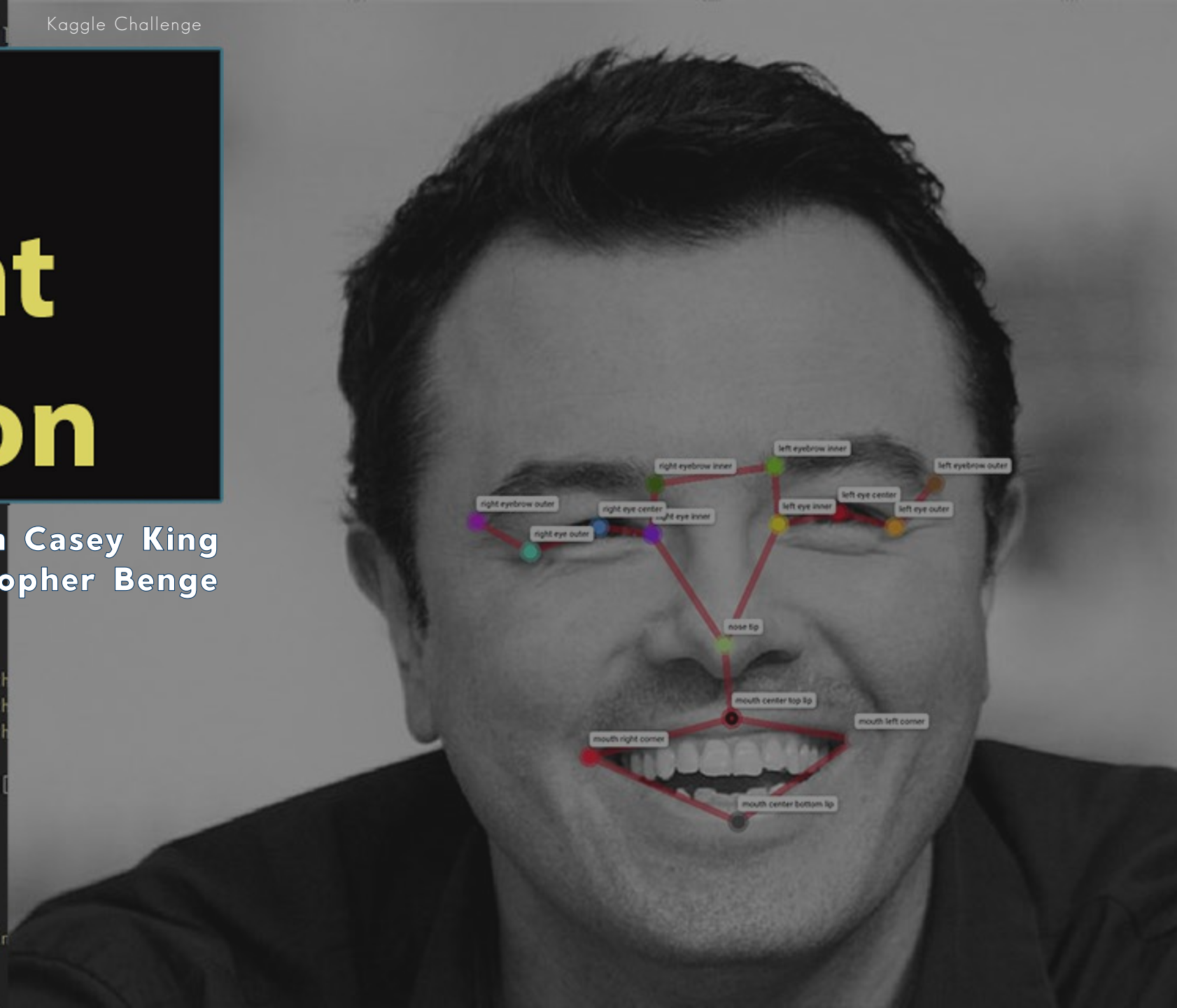
Kaggle Challenge

Facial Keypoint Detection

William Casey King
Cristopher Benge



Your Mission, Should You Choose To Accept…

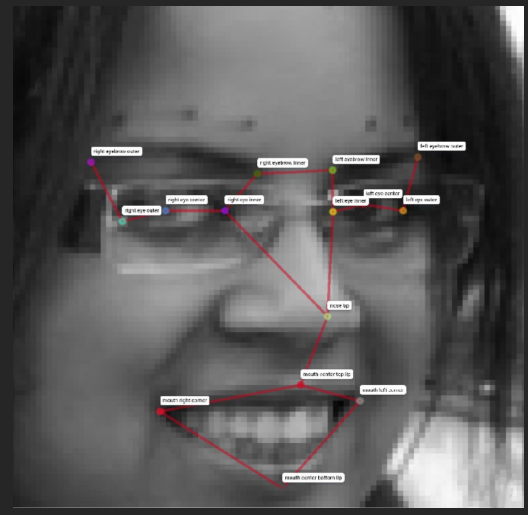
Detection of [up to] 15 facial landmarks on 1,783 images

Dataset

- 96*h* x 96*w* x 1*c* grayscale images
- Pixel values [0..255]
- Up to 30 labels of 15 keypoints in (x,y) coordinates for each image.
- Images provided as space-separated integer array (1 x 9,216)

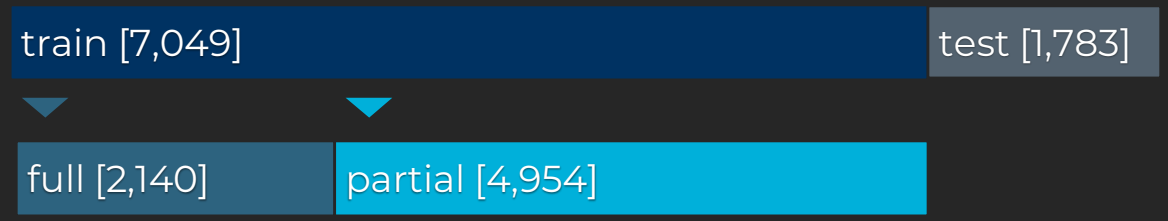
Unlabeled Image

Keypoints Detected



Find (x,y) coordinate for…

- eye center (left, right)
- eye inner corner (left, right)
- eye outer corner (left, right)
- eyebrow inner corner (left, right)
- eyebrow outer corner (left, right)
- mouth corner (left, right)
- mouth center (top, bottom)
- nose tip



Performance measured by RMSE:

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

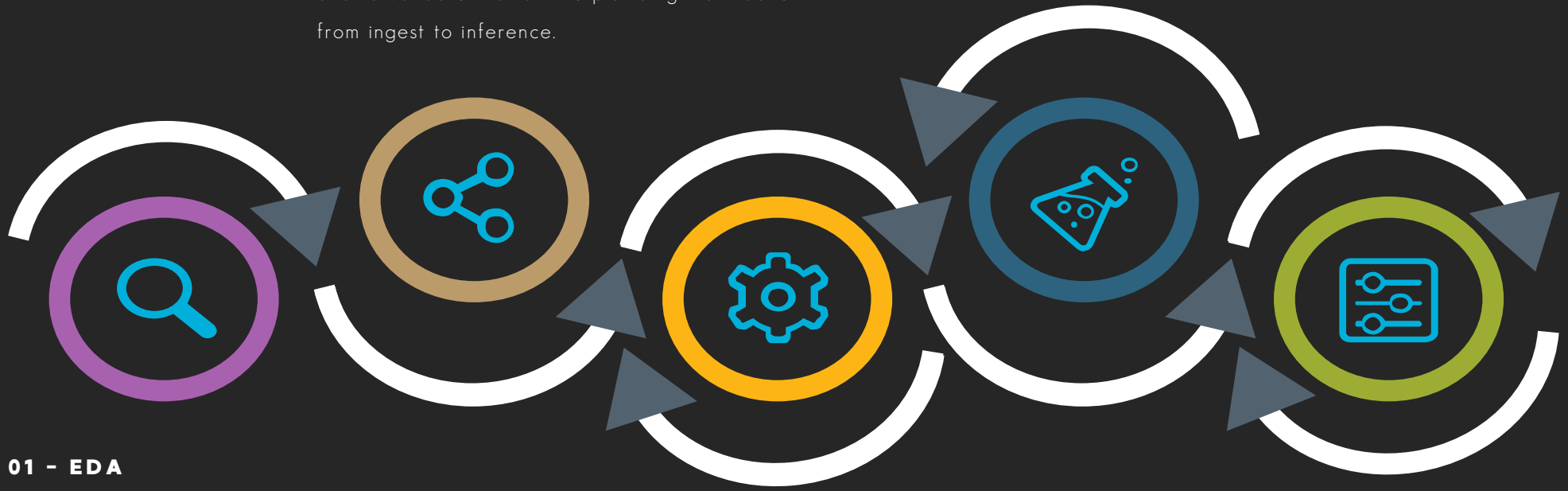
Project Lifecycle

02 – Naïve Model

A naïve model is created to serve as a baseline and to validate that all the plumbing work done from ingest to inference.

04 – Model Experimentation

Various model architectures are evaluated, compared to baseline, and minor tuning is performed to identify the most suitable strategy.



01 – EDA

Exploratory data analysis is performed to understand the data and the challenges we'll have to address to achieve acceptable model performance.

03 – Cleaning & Augmentation

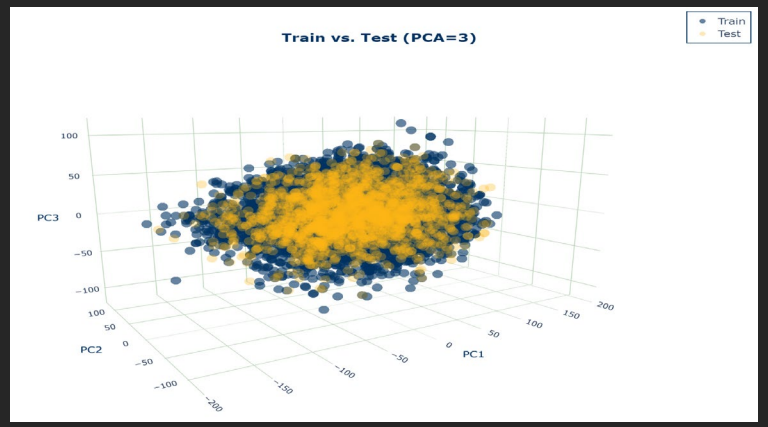
Cleaning and augmentation of the training data set is a significant part of any ML project; this one was no exception.

05 – Tuning

Hyperparameter tuning, regularization, and other optional techniques (such as generalized stacking) are performed to fine-tune for ideal performance.

Exploratory Data Analysis

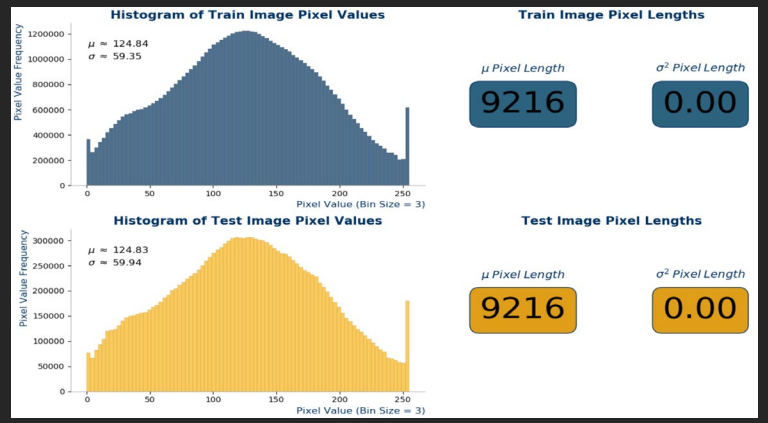
Preliminary assessment of data quality in both train and test datasets



Adversarial Validation

Principal Component Clusters

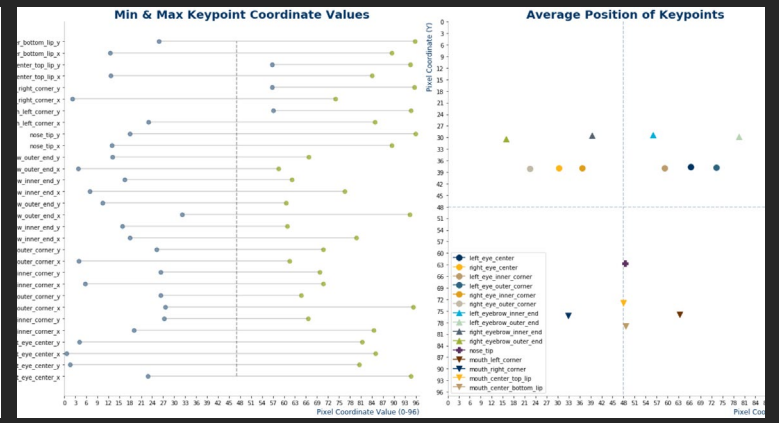
PCA was performed on train and test and compared as a means to quickly assess if their distributions were similar enough to make reasonable predictions from.



Distribution Analysis

Pixel Intensity Distribution

The distributions of pixel intensity values are nearly identical between train and test. Further, there are no images in either set "missing" pixel values.



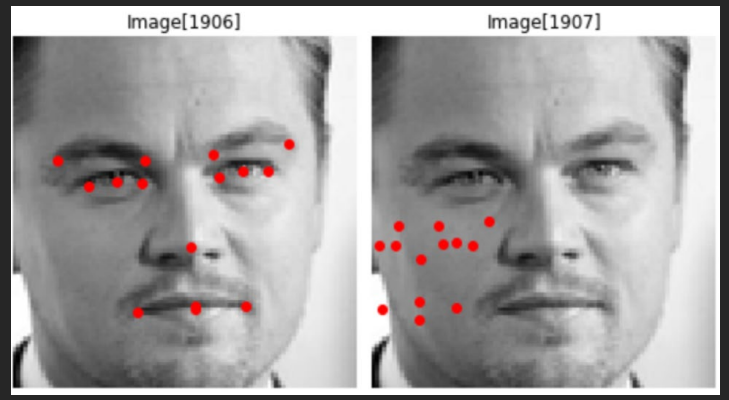
(X, Y) Extreme Position Values

Coordinate Validation

Some of the images have extreme locations along the X or Y axis (i.e. a nose tip that appears at the "top" of the image), implying some images are not centered or face-forward. Regardless, the average coordinate positions for each keypoint forms what looks like a centered face.

Exploratory Data Analysis (Contd.)

Preliminary assessment of data quality in both train and test datasets

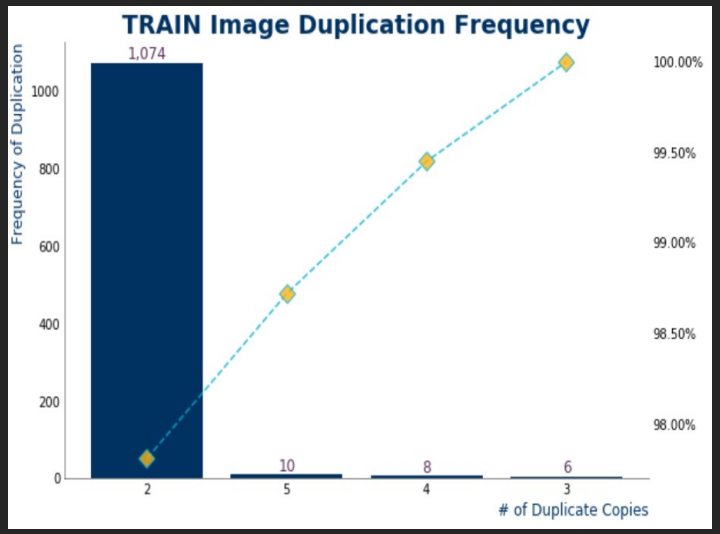


Label Quality

A visual inspection of labels revealed the presence of both duplicate images and incorrect labels. These will need to be addressed to improve the quality of keypoint predictions.

Label Availability

A large number of training images possess only a fraction of the labeled keypoints.

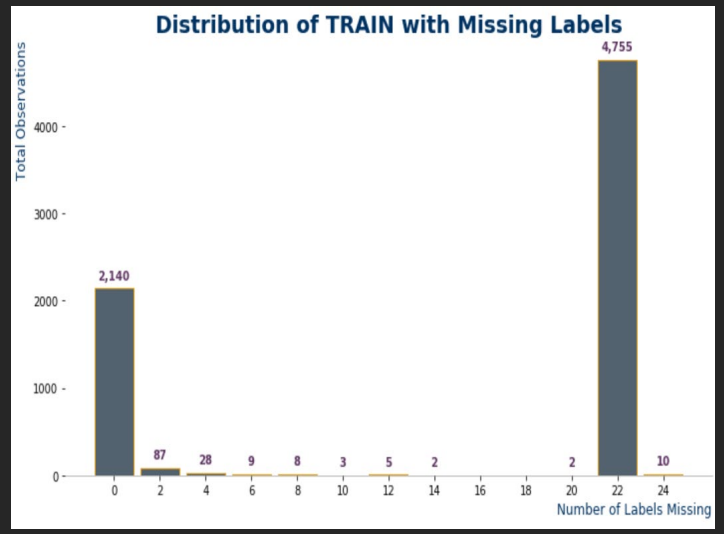


Duplicate Images

There are a large number of duplicate images in train. Reconciling these will be essential to attaining a top score.

Exploratory Data Analysis (Contd.)

Preliminary assessment of data quality in both train and test datasets

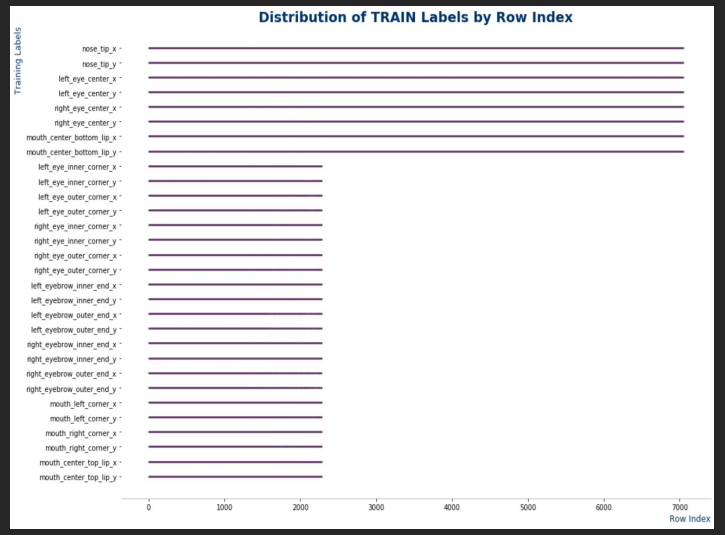


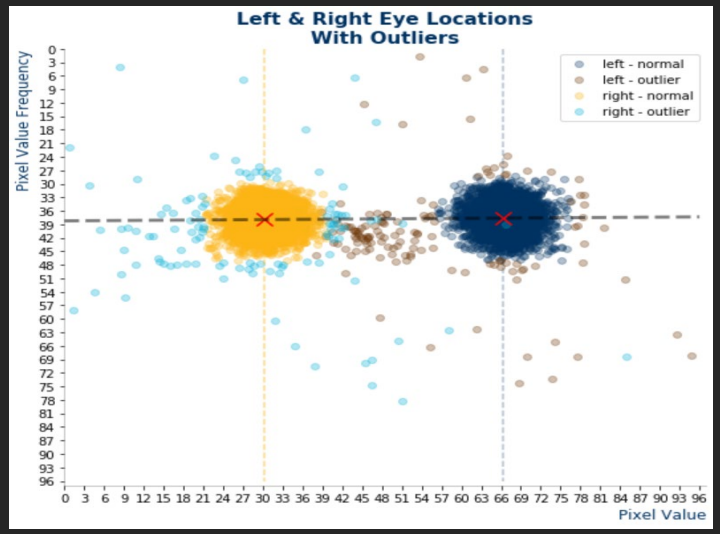
Missing Labels Analysis

It appears that there are roughly two types of images provided by the organizers: 2,140 have all 15 keypoints, and 4,755 have only 4 keypoints.

Missing Labels (Contd.)

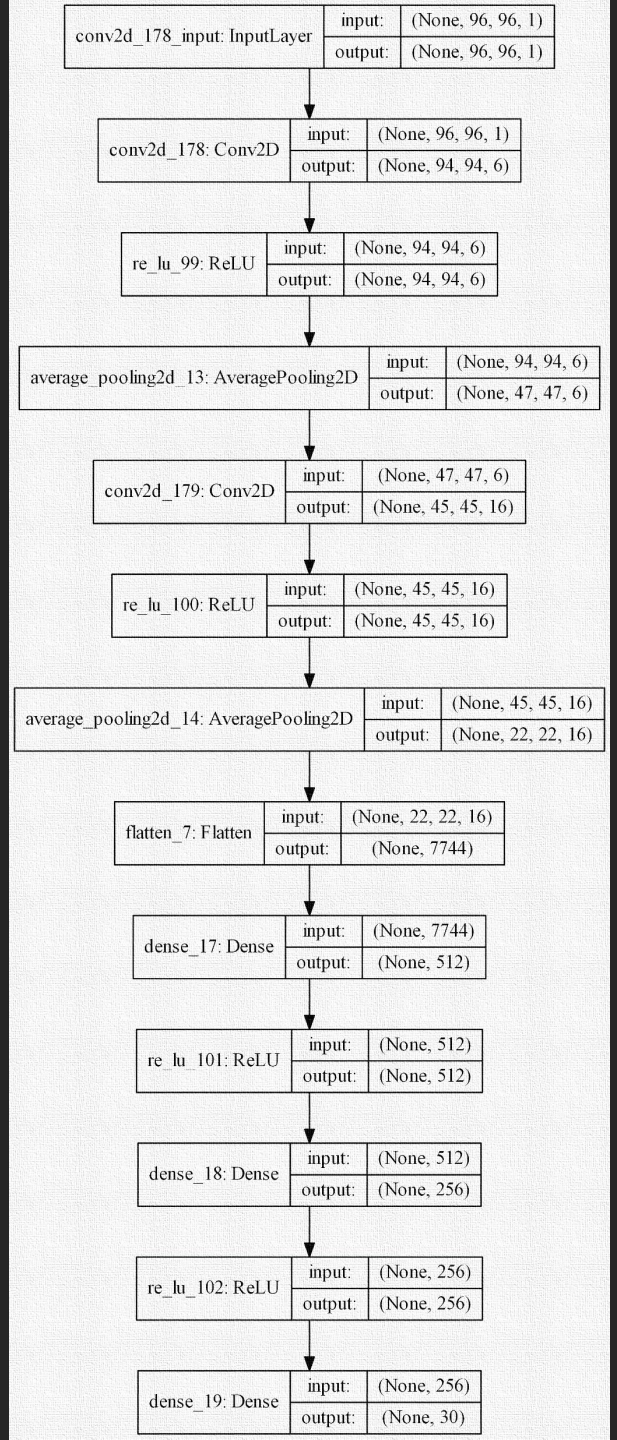
When ordered by physical row index, a pattern emerges that suggests train originates from two distinct datasets.





Eigeneyes

Most of the eyes are picture centered, but a handful (those in light blue and maroon) are positioned more than 3σ from μ.



Naïve Model (Baseline)

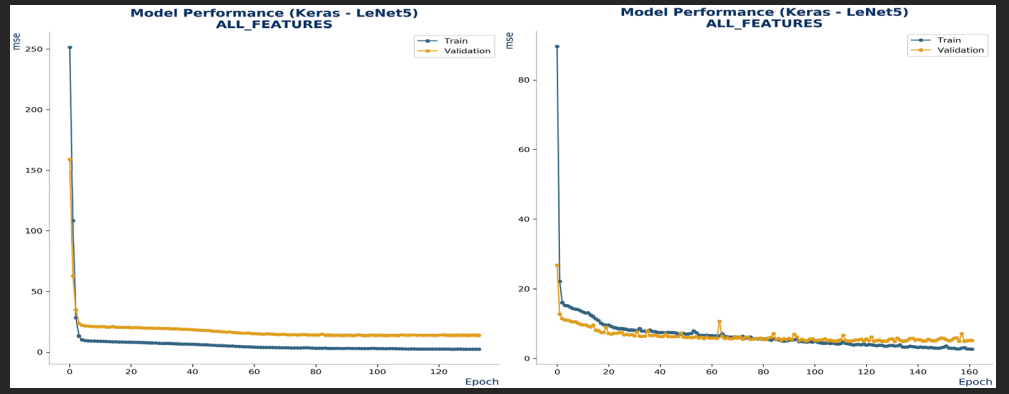
LeNet-5

The classic 5-layer LeNet-5 (2 x convolution, 3 x dense) was used as the project baseline. No augmentation was used for our baseline, and only image de-duplication, dropping of two low quality images, and manually fixed labels for 55 training images was applied.

All training is fixed at 128 batch size, 300 epochs with patience set at 30. Adam optimization is used with an initial learning rate of 0.001, beta1 of 0.9, beta2 of 0.999, and epsilon of 1e-8.

Training and test was split into two separate tasks : Path #1 learns and predicts all images where all 15 keypoints are present. Path #2 learns and predicts all images where only 4 keypoints are present. At submission time, these predictions are combined.

Under this scheme, and with no added regularization, LeNet-5 scored 2.33068 RMSE on the private leaderboard. This is good for 51st place.



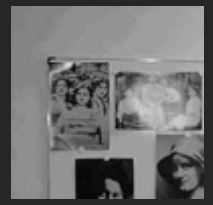
Data Cleaning

Some light cleaning goes a long way…

Drop "Bad" Images

Two train images are removed:

[6492,6493]



x 2

Load "Fixed" Labels

56 manually selected and fixed train image labels.

Deduplication

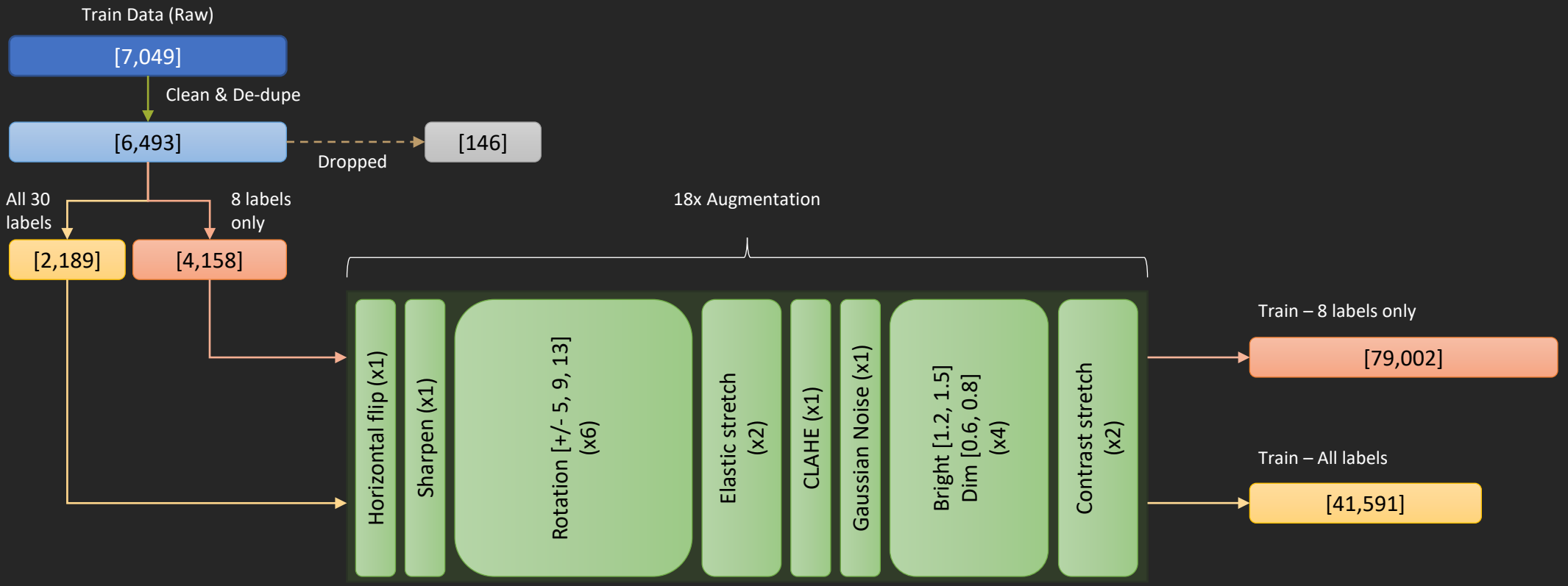
Hashing to identify duplicate images; taking the mean average of label values.

Scale Pixel Values

$\bar{x} = \frac{x}{255}$ → [0..1]

Data Augmentation Pipeline

The unreasonable effectiveness of data…



Model Architectures

Models, models… everywhere!

01 **LeNet-5**

4.1M parameters (ref)

02 **Conv2D 5-Layer**

3.2M parameters

03 **Conv2D 10-Layer**

7.2M parameters

04 **Local2D 5 Layer**

2.4M parameters

05 **ResNet V1 (18, 34)**

11.2M / 21.3M parameters (ref)

06 **ResNet V2 (50)**

23.6M parameters (ref)

07 **Inception V1**

6.8M parameters (ref)

08 **Inception V3**

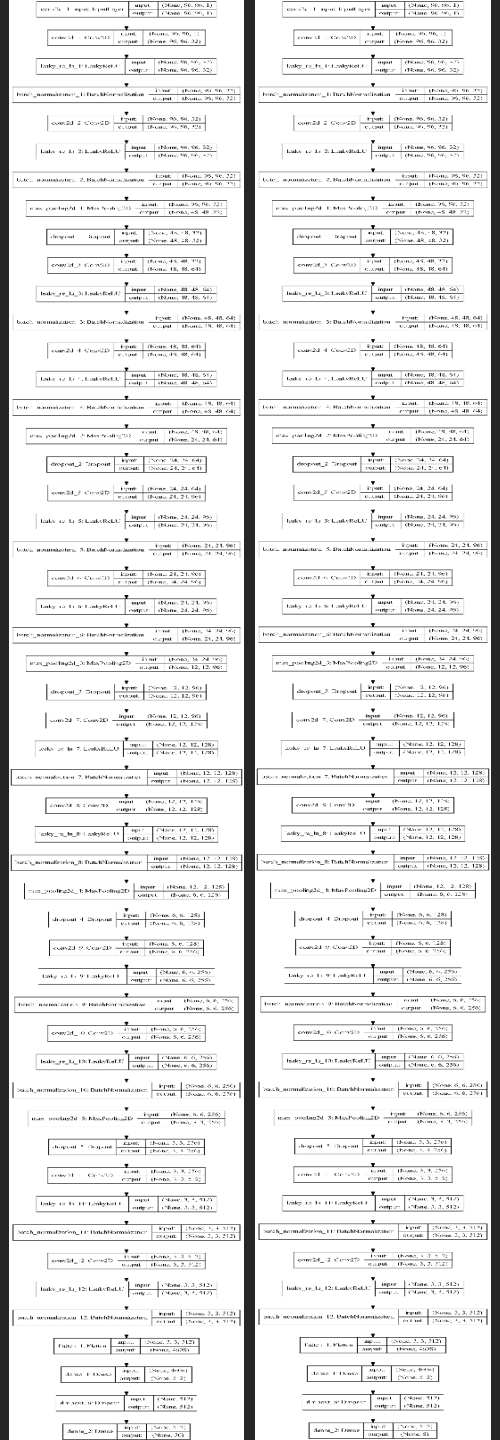
21.8M parameters (ref)

09 **ResNeXt101**

241M parameters (ref)

10 **NaimishNet**

7.4M parameters (ref)



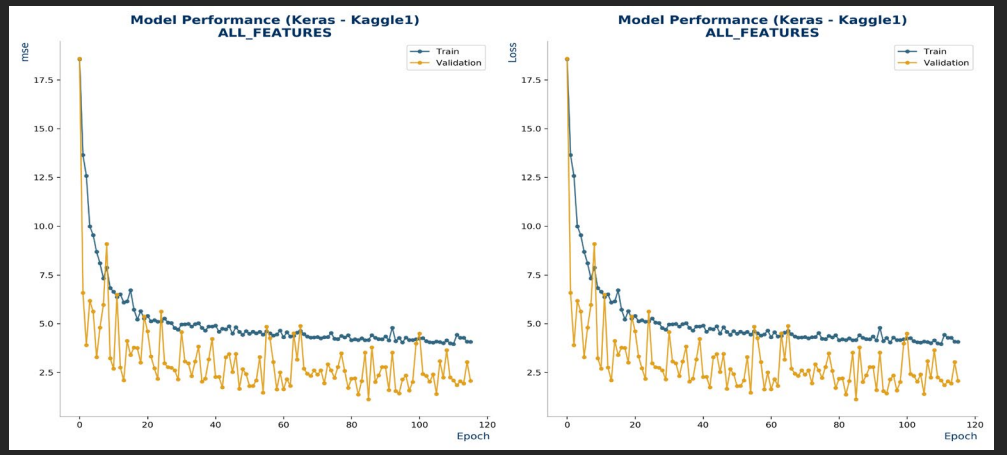
Best Single Model

Conv2D-10 Layer

A best single model in our ensemble was a straight-forward 10-layer CNN that follows a pattern of ever-increasing feature maps (filters) via: conv2d -> LReLU -> batch norm -> conv2d -> LReLU -> batch norm -> max pool -> dropout (X 5). The final layer consist of flatten -> fully connected (512) -> dropout -> fully connected (30 or 8).

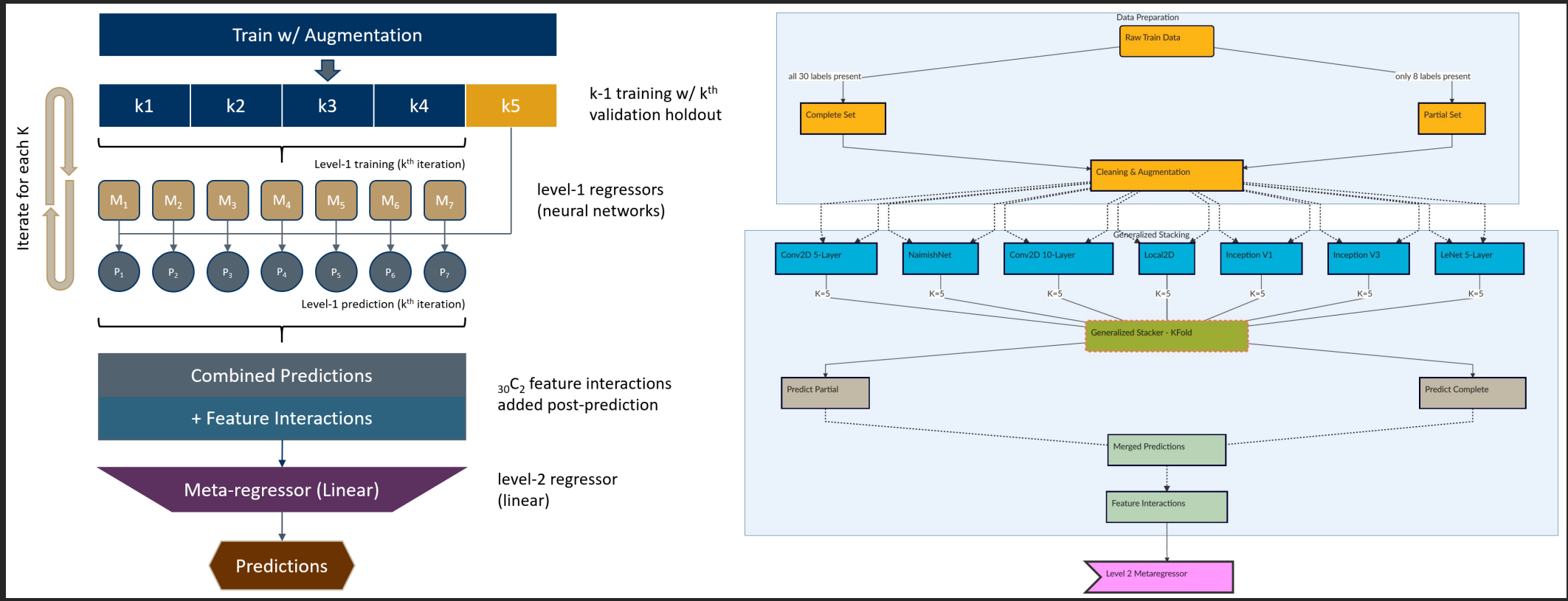
All training is fixed at 128 batch size, 300 epochs with patience set at 30. Adam optimization is used with an initial learning rate of 0.001, beta1 of 0.9, beta2 of 0.999, and epsilon of 1e-8.

Conv2D-10 Layer scored 1.35183 RMSE on the private leaderboard. This is good for 2nd place, and is 0.06947 RMSE away from a tie with the 1st place submission.



Generalized Stacking

"…a technique whose purpose is to achieve a generalization accuracy (as opposed to learning accuracy) which is as high as possible."



Final Results

Our final submission, consisting of a stacked generalization of 20 models scored a private-leaderboard rating of 1.28637 RMSE or just 0.00401 away for a tie with the 1st place submission.

Submission and Description	Private Score	Public Score
SUBMISSION_STACK_20_MODELS.csv a minute ago by cbenge509 Stacker - 10 × 30, 10 × 8, Kfold = 5, 30C2 and 8C2 feature interactions at metaregressor (ElasticNet). augmentation: hz flip, rotation(+9/-9), CLAHE (0.03 clipping), pixel shifts(5), brightness(1.2), dimming(0.8), contrast stretch (0.0 - 1.2), elastic stretch (991/8)	1.28637	1.45471

Key Takeaways

"Science is what we understand well enough to explain to a computer… art is everything else."

01 **Cleaning is [almost] everything**

With just basic cleaning we got 50th place!

02 **The last 10% is 90% of the work**

Augmentation, tuning, and stacking - oh my!

03 **Don't allow any assumption to go unchallenged**

Sorry, your intuition is [most likely] wrong.

04 **Success is non-linear**

Be "ok" with failure… a whole lot of failure.

05 **When in doubt… put it down**

There is a point where progress is best had after a break.