Improving Facial Keypoints Detection Through Image Augmentation, Split Pipelining, and Stacked Genrealization of Weak Learners

Cristopher Benge
Columbia University
New York, NY
cb3704@columbia.edu

Abstract

In this paper, we present a novel approach to improving facial keypoint detection performance by combining image augmentation techniques, stacked generalization of weak learners, and exploitation of diverse training data subsets via targeted splitting of data preparation and training pipelines. Our methodology entails careful data preprocessing, including thorough cleaning of provided images and labels to ensure high-quality training data is available to all learners. Further, We leverage a stacked ensemble architecture comprising multiple simple neural networks, each trained on a subset of the data arranged by different data organizers, with a final linear meta-model that utilizes multiplication-based feature interactions for the final prediction. This approach resulted in a final RMSE of 1.28637, good for second place in a Kaggle competition and only 0.004 RMSE off from a tie for first place.[1]

1. Introduction

Facial keypoint detection, the task of localizing and identifying key facial landmarks in images, serves as a fundamental component in various computer vision applications such as facial recognition, emotion analysis, and facial expression synthesis. Further, the accurate localization of key facial landmarks such as eyes, nose, and mouth corners is crucial for tasks like face alignment and tracking, enabling advancements in human-computer interaction, biometrics, and augmented reality. Despite significant advancements in recent years, accurate and robust facial keypoint detection remains a challenging problem due to variations in facial expressions, poses, lighting conditions, and occlusions. [7]

Traditional approaches to facial keypoints detection often rely on handcrafted features and shallow learning models, which may struggle to capture the complex and subtle patterns present in facial images or generalize well to diverse and complex datasets. With the emergence of deep learning, convolutional neural networks (CNNs) have demonstrated remarkable success in automatically learning hierarchical representations directly from raw pixel data. [2] However, the success of CNN-based methods heavily depends on the quality and diversity of the training data, as well as the effectiveness of the training process itself.

In this paper, we propose a novel approach to enhancing facial keypoints detection performance by integrating several key strategies: image correction and label fine-tuning, stacked generalization of weak learners leveraging a K-fold strategy, and a split pipeline for data processing and training through detection of individual contributing organizers, allowing for exploitation of variance in labeling patterns. To further enhance the robustness of our system, we incorporate image augmentation techniques, artificially expanding the training dataset, thereby improving diversity and generalized performance of each weak learner.

To evaluate the effectiveness of our approach, we conduct extensive experiments on a benchmark facial keypoints detection dataset from the 2017 Kaggle competition: Facial Keypoints Detection. Competition submissions are scored on the root mean squared error as the standard metric for comparing solution quality:

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

Our results demonstrate that the proposed method achieves state-of-the-art performance. Overall, this paper presents a comprehensive framework for improving facial keypoints detection, offering valuable insights into the synergy between data augmentation, ensemble learning, and data processing techniques in the context of computer vision tasks.

[1] Solution available on GitHub: cbenge509 - W4732 Final Project

1.1. Related Work

Facial keypoint detection has been a subject of active research in computer vision for several decades, with a variety of approaches proposed. Traditional methods often relied on handcrafted features and shallow learning models, such as support vector machines (SVMs) coupled with techniques like Haar cascades or local binary patterns (LBP) for feature extraction [1, 6].

In recent years, the rise of deep learning has transformed facial keypoints detection work by enabling end-to-end learning directly from raw (lcalized) pixel data. CNNs have emerged as the dominant architecture for this task due to their ability to automatically learn hierarchical representations, capturing both low-level features like edges and high-level semantic information [3]. Early CNN-based approaches demonstrated promising results on facial keypoints detection tasks, insporing subsequent research in the domain. [11]

Several studies have focused on enhancing the robustness and accuracy of facial keypoints detection by exploring various aspects of deep learning architectures and training methodologies. Zhang et al. [10] introduced a multi-task learning framework that jointly predicts facial keypoints and facial attributes, leveraging shared representations to improve performance. Similarly, Sun et al. [5] proposed a cascaded CNN architecture that refines keypoints predictions iteratively, enabling better handling of occlusions and variations in facial expressions.

Ensemble learning techniques have also been applied to facial keypoints detection, aiming to leverage the diversity of multiple models to improve overall performance. Zhang et al. [9] proposed a boosting-based approach that combines weak learners trained on different subsets of the data, effectively reducing generalization errors and enhancing robustness to outliers. Similarly, Liu et al. [4] explored the use of stacked generalization, combining predictions from multiple base models with a meta-learner to achieve superior performance.

While these prior works have made significant contributions to the field, they often focus on individual aspects of the facial keypoints detection pipeline, such as model architecture or training strategy, without fully exploiting the potential synergy between different techniques. In contrast, our proposed method integrates image augmentation, stacked generalization, and targeted data preprocessing to achieve state-of-the-art performance, highlighting the importance of a holistic approach to addressing the challenges of facial keypoints detection.

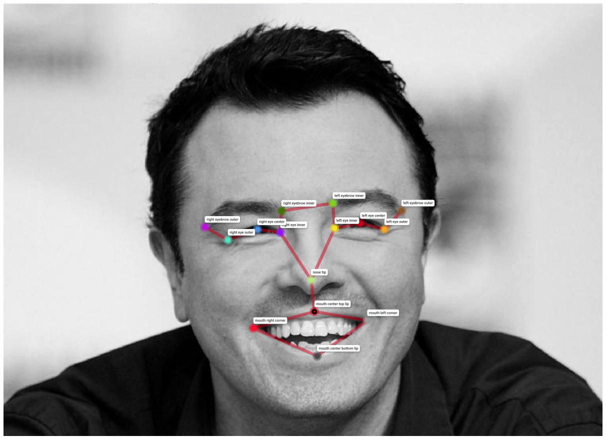


Figure 1. Example of 15 facial landmarks labeled on American actor and animator Seth MacFarlane. This image was not a part of the competition training or test data; rather, it was created to serve as an exemplar inference outcome of our solution.

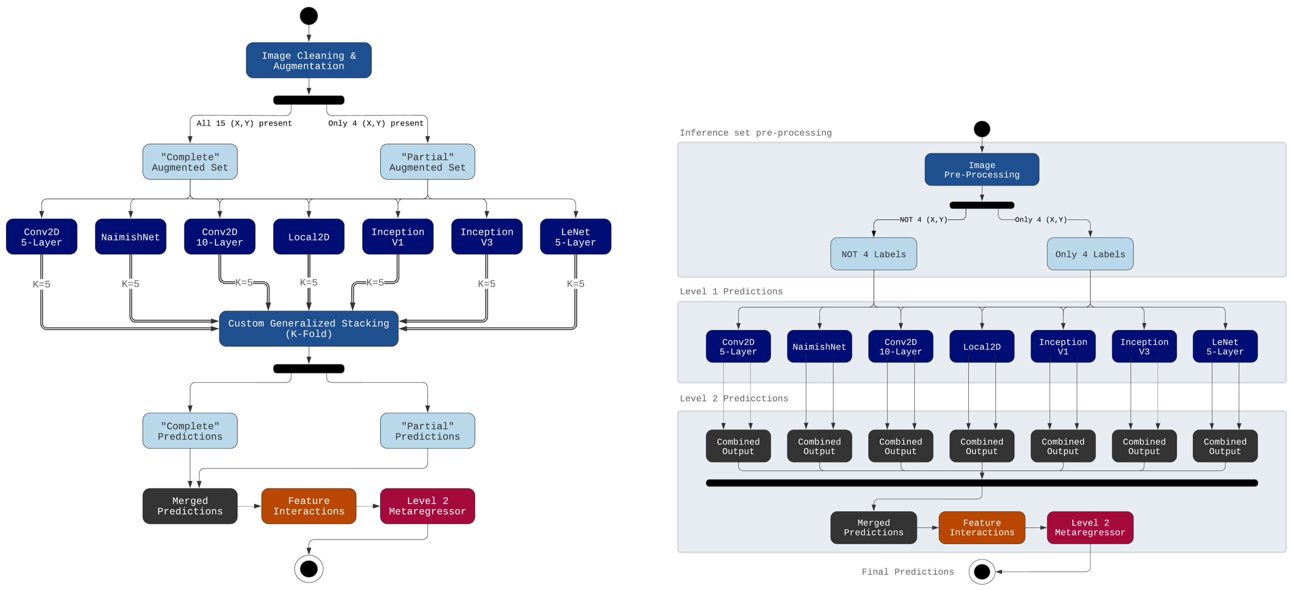
2. Methodology

Our proposed methodology for enhancing facial keypoints detection encompasses a comprehensive framework that integrates several aforementioned strategies: image correction and label fine-tuning, stacked generalization of weak learners leveraging a K-fold $(K = 5)$ strategy, and a split pipeline for data processing and training through detection of individual contributing organizers, allowing for exploitation of variance in labeling patterns. The following sections outline each component of our methodology in detail.

2.1. Data Preprocessing and Augmentation

Prior to model training, we perform extensive data preprocessing to improve quality of labels. Following extensive exploratory data analysis, we performed the following pre-processing steps:

- Removal of two training images that did not contain faces or facial keypoint labels.
- 56 images had incorrect keypoint labels. All received manually corrected labels to establish a better ground truth.
- 277 duplicate images identified through hashing; labels were mean averaged and duplicates images removed to reduce bias.
- 186 images were dropped due to containing neither all nor only-four keypoints (8 labels). This distinction is how we differntiate our split pipeline later.
- all pixel values were normalized $(\bar{x} = \frac{x}{255})$



(a) Training pipeline consists of splitting on partial and complete keypoint examples, training eight separate learners twice (once for each split set), performing generalized stacking on K-flod strategy, and emitting predictions for level-2 metaregressor.

(b) Inferencing works much the same as our training pipeline, with prediction patterns being split on asks for 15 or 4 labels, generating predictions for each weak learner, and merging those predictions into a single output for input into our level-2 metaregressor.

Figure 2. Our solution training and inference pipeline that results in state-of-the-art performance for facial keypoints detection.

To augment the training dataset and improve model generalization, we employ various image augmentation techniques. These include positive and negative rotations, elastic transformation of images, injection of random gaussian noise, brightening and dimming, constrast stretching, image sharpening, horizontal flipping, and Contraste Limited AHE (adaptive histogram equalization). This process, outlined in Fig. 3 below, produced an 18-fold increase in training data available to our weak learners.

2.2. Stacked Generalization of Weak Learners

Our methodology adopts a stacked ensemble architecture [8], which combines predictions from multiple weak learners to make final predictions. We employ a K-fold strategy to divide the dataset into training and validation sets, ensuring that each model is trained on a diverse subset of the data and evaluated on unseen samples. Predictions of the held-out values are saved to train the final level-2 metaregressor after multiplication-based feature interactions are generated. See Fig. 4 for depiction of this process.

We evaluated dozens of simple neural network architectures for individual performance and eliminate those that either under-perform or have predictions with a high pearson correlation to another model already included in our ensemble. Resultingly, sevel models were chosen as final candidates for our level-1 weak learners. See Tab. 1.

Model	**Description**
Conv2D 5-layer	A simple 5-layer 2d CNN.
NaimishNet	A 7.4M parameter 2D CNN that learns only one keypoint at a time (paper).
Conv2D 10-layer	A deeper version of Conv2D 5-layer above.
Local2D	A modified version of Conv2D 5-layer whose final layer is a Local2D with global average pooling.
Inception V1	A modified version of Google's Inception V1 model.
Inception v3	A modified version of Google's Inception V3 model.
LeNet5	A slightly modified 5-layer version of the classic LeNet model.
ResNet50	A slightly modified version of Microsoft's classic ResNet50 model.
ResNet (custom)	A customized, greatly simplified ResNet model architecture.
ResNeXt50	An implementation of the full ResNeXt50 model.

Table 1. The seven weak-learner models used in our ensemble.

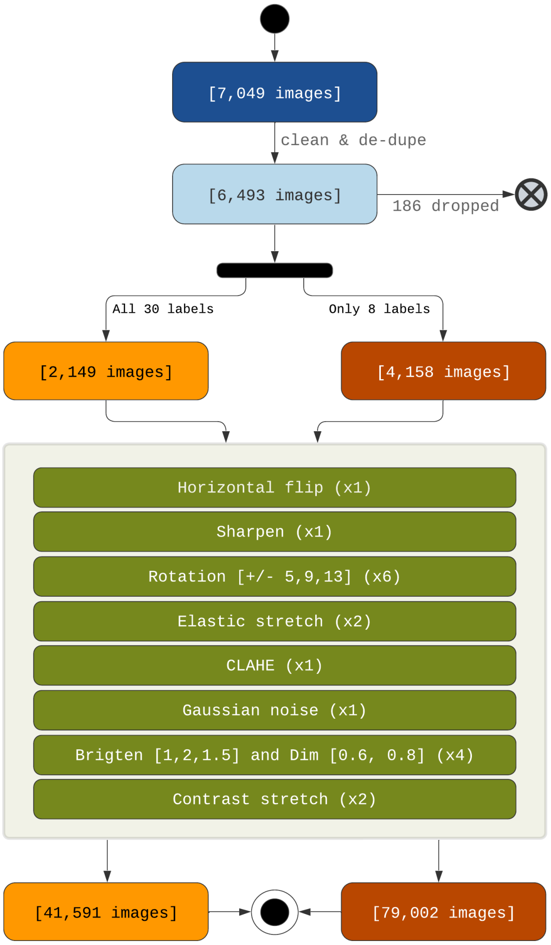


Figure 3. Data augmentation pipeline results in 18x training images and better generalization.

2.3. Split Pipeline for Data Processing and Training

Recognizing that individual data organizers label facial keypoints in slightly different ways was a key insight for improving the performance of this solution. The organizer who prepared the labels with only four landmarks had a slightly different idea of where the tip of the nose or upper lip were located, for example, than the organizer who prepared the examples with all landmarks present. To further exploit the variance in labeling patterns and enhance the diversity of the trained models, we employ a split pipeline for data processing and training. This involves splitting and training images with only four keypoints separate of those with all keypoints. By training separate models on data subsets organized by different criteria, we aim to capture the underlying patterns specific to each organizer and improve the overall performance of the ensemble.

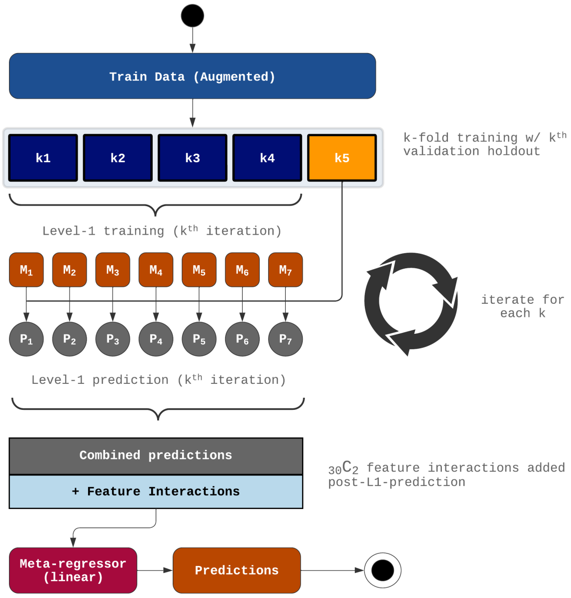

Train Data (Augmented)
k1
k2
k3
k4
k5
k-fold training w/ kth validation holdout
Level-1 training (kth iteration)
M1
M2
M3
M4
M5
M6
M7
P1
P2
P3
P4
P5
P6
P7
iterate for each k
Level-1 prediction (kth iteration)
Combined predictions
+ Feature Interactions
30C2 feature interactions added post-L1-prediction
Meta-regressor (linear)
Predictions

Figure 4. K-fold stacked generalization strategy usuing $(K = 5)$ is used to train our level-2 (linear) metaregressor.

2.4. Evaluation Metrics

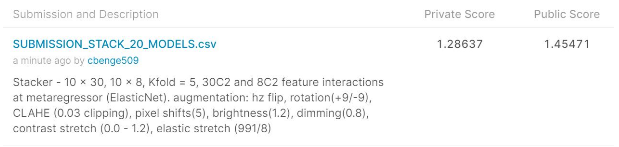
We evaluate the performance of our methodology using the Root Mean Squared Error (RMSE) metric, which quantifies the average deviation between the predicted and ground truth keypoints across all facial images in the test set. Additionally, we analyze the performance of individual keypoints to assess the robustness and accuracy of the trained models across different facial landmarks.

2.5. Experiment Setup

We conducted hundreds of experiments on candidate models against our benchmark facial keypoint detection dataset provided in the Facial Keypoints Detection Kaggle competition. The original dataset was comprised of 7,049 labeled training samples and 1,783 test images. Of the 7,049 labeled images, only 2,140 ($\approx 30.4\%$) were provided with annotated labels for all 15 facial landmarks. Following our methodology listed above, we pre-processed all image data and pre-trained the seven selected models and final level-2 metaregressor model. Finally, all test images were run through the pipeline to emit a final (X, Y) coordinate for each landmark and were submitted to the private leaderboard on Kaggle. All data processing, training, inference, and evaluation were performed on a local desktop with a single SSD and 24GiB of RAM provided thorugh an NVidia 4090 RTX card.

3. Results

Our results demonstrate the effectiveness of the proposed methodology in improving facial keypoints detection performance, yielding competitive results in the Facial Keypoints Detection Kaggle competition. We achieved a final RMSE of 1.28637, securing second place in the competition. Notably, our solution was within an RMSE of 0.00401 of tying for first place, showcasing the effectiveness of our approach.



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

Figure 5. A final submission scoring 1.28637 RMSE on the private leaderboard

3.1. Ensemble Performance

Our winning solution was constructed as an ensemble of seven weak learners, each contributing to the final prediction. Through stacked generalization, we combined predictions from multiple models to improve overall performance. The ensemble approach proved highly effective, improving our score by 0.06546 RMSE, allowing us to achieve a significant reduction compared to individual models and putting us within a very small marin for a first-place solution.

3.2. Best Performing Weak Learner

Among the ensemble of models, the Conv2D 10-layer model emerged as the single best-performing weak learner. This model, leveraging a convolutional neural network architecture with ten layers, achieved an RMSE of 1.35183. Despite being outperformed by the ensemble, this model's performance highlights the effectiveness of deep learning techniques in facial keypoints detection.

3.3. Comparison with First Place Solution

While our solution secured second place in the competition, we were narrowly edged out by the first place winner, whose solution achieved an RMSE of 1.28236 and was significantly more complicated. The small difference of 0.00401 RMSE points underscores the competitiveness of our approach, showcasing its effectiveness in addressing the challenges of facial keypoints detection.

3.4. Limitations

Despite our competitive performance, our methodology has limitations. The reliance on a diverse ensemble of models introduces additional computational complexity and may require significant resources for training and inference. Additionally, a lot of pre-processing and selective data augmentation was required to achieve state-of-the-art performance. Regardless, careful data augmentation and cleaning is a necessity - our pre-processing was worth ≈ 0.25 RMSE along on the best performing single model. Finally, while our solution achieved high accuracy on the benchmark dataset, its generalizability to real-world scenarios and diverse populations will require further validation.

4. Conclusion & Future Direction

Moving forward, there are several avenues for further research and improvement. Fine-tuning model architectures, exploring novel ensemble techniques, evaluating transfer learning options on larger pre-trained models, and incorporating additional data sources could all contribute to enhancing the robustness and generalization capabilities of our approach. Additionally, evaluating our methodology on larger and more diverse datasets could provide valuable insights into its performance across different demographic groups and environmental conditions.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006. 2

[2] Savina Colaco and Dong Seog Han. Facial keypoint detection with convolutional neural networks. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 671–674, 2020. 1

[3] Bengio Yoshua LeCun, Yann and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015. 2

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild, 2015. 2

[5] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013. 2

[6] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 2

[7] Nannan Wang, Xinbo Gao, Dacheng Tao, Heng Yang, and Xuelong Li. Facial feature point detection: A comprehensive survey. *Neurocomputing*, 275:50–65, 2018. 1

[8] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. 3

[9] Yue Wu and Qiang Ji. Robust facial landmark detection under significant head poses and occlusion, 2017. 2

[10] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. 2

[11] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 94–108, Cham, 2014. Springer International Publishing. 2