Bridge the Domain-Gap: Facial Landmarks with Fake-it Dataset

Chaoyi Hu, 301559775

Simon Fraser University, CMPT 732, October 2022

1 Introduction

1.1 The Domain Gap

Access to high-quality datasets is crucial to success in deep learning projects, yet data availibility has always been a problem in human-related computer vision projects due to ethic restraints such as privacy protection. Synthetic human facial image using computer graphics provides an ideal data source, as it offers theoretically infinite amount of data with extra benefits such as elimination of ethic concerns, full control over the diversity of the dataset, as well as perfect labels[1]. However, crossing the gap between synthetic and real world data remains a challenging task.

1.2 Demand for Efficient Models

Large pre-trained convolutional neural networks are impractical for uses on consumer devices, where memory and computational power are usually limited. For example, VGG-16, one of the earlier benchmark models, has a total of 138 million parameters and occupies over 552 MB of space[2]. Running such a huge network on a typical smartphone with 4 to 8 Gigabytes of RAM will be extremely inefficient. To perform evaluation of deep neural network at a rate of at least 30 times per second on consumer devices, the network needs to be highly efficient and compact[3]. Series of light-weight models have been proposed to suit the increasing demand. Featuring compactness and high efficiency, these models are designed to be more suitable for real-time uses on consumer devices, aiming at smaller model size that is memory-efficient, and faster inferencing that is computationally efficient.

1.3 Our Project

In this project, we divide our tasks in a progressive manner. For the basic task, as described in section 2.1, we aim to design a facial landmark detection model that closes the domain gap, i.e. a model that trains on synthetic human facial images and generalizes to real human facial images. Comparative tests of model accuracy and inference speed will be conducted to evaluate the performances of different models. For the bonus task, as described in section 2.2, we will

experiment with various network architectures and model compression techniques in order to construct an improved model with higher efficiency. Followingly, as described in section 2.3, we seek to develop a prototype based on our improved model. This prototype should perform inferences in real time on consumer devices.

2 Project Design

The project requirements recommended the following datasets. For the synthetic training data, we will use the CelebA-HQ dataset consisting of 30,000 high-resolution synthetic human facial images[4]. For the real-world test data, we will use real human facial images provided by Flickr-Faces-HQ Dataset (FFHQ)[5]. Before the main experiment, we will inspect the dataset and conduct some preliminary experiments to determine the minimum viable amount of data we need to use in our project.

2.1 Model Design: Closing the Domain Gap

This subsection corresponds to our basic task. We propose to build the network based on the Pytorch framework. As for tools for necessary image manipulations, our primary option is OpenCV. To measure the accuracy of the landmark locations, we propose to use widely adopted metrics including Normalized Mean Error (NME) or Falure Rate (FR) for the ease of comparison. Additionally, we will calculate the number of parameters as a metrics for model size, and measure the inference time (ms) as a metric for inference speed, in order to evaluate the potential of models for real-time inference on consumer devices.

More research needed to be done before we determine the design of our network architecture. For now, our starting point would be established facial landmark detection backbones such as ResNet[6] or ShuffleNetV2[3]. Deep learning libraries such as Dlib can hopefully provide a baseline for landmarks localization[7]. On top of the proposed model and baseline models, we will select 2-3 benchmark models for performance comparison.

2.2 Model Optimization: Lighter Model, Higher Efficiency

This subsection correspondes to our bonus task. We seek to improve our model by reducing the inference time and computational power required, making it more suitable for real-time inference on consumer devices.

Improvement in model performances usually comes as a result of optimizations in two aspects: network architecture design, and model compression techniques. SqueezeNet[8] features parameter compression in convolutional neural networks using Fire modules, and has been demonstrated to be able to run on low-power processing platforms such as smartphones and custom processors. Other networks such as SqueezeNext[9], ShuffleNet[3], MobileNet[10] also offer great references for light-weight network design. Novel approaches such as Wing Loss[11], Practical Facial Landmark Detector (PFLD)[12] can potentially be applied to improve our network as well. Model compression, such as quantization and pruning of model parameters, can be applied to a deep neural network after it has been trained to make the trained model more compact.

We will experiment with different network architectures attained by adjusting network components or applying novel approches on top of the proposed network. We will observe the changes in accuracy and efficiency and determine the best improved model.

2.3 Development of a Real-time Prototype

Based on the improved model, we plan to build a desktop application for real-time facial landmark detection that runs on a personal laptop. The prototype should be able to detect and display facial landmarks on a 30 fps video file or camera input. We propose to build the application in Python, with tkinter library for the interface.

3 Timeline

- From Oct 1 to Oct 15, we will explore the datasets, review existing methods and network architectures for facial landmarks detection, and select as well as reproduce representative light-weight models suitable for inference on consumer hardware for future comparative experiments.
- From Oct 16 to Oct 31, we plan to finish tasks described in 2.1.
- From Nov 1 to Nov 15, based on the results we get in the previous stage, we will seek to improve the efficiency of the proposed model. Employing the improved model, we will develop a prototype that runs in real time. Ideally, we will build a interface for demo purposes.
- By Nov 30, we will wrap up the project, summarizing our work in the form of report, presentation, and source files.

References

[1] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, M. Johnson, V. Estellers, T. J. Cashman, and J. Shotton, "Fake it till you make it: Face analysis in the wild using synthetic data alone." [Online]. Available: https://arxiv.org/pdf/2109.15102

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/pdf/1409.1556

[3] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design." [Online]. Available: https://arxiv.org/pdf/1807.11164

[4] H. Huang, Z. Li, R. He, Z. Sun, and T. Tan, "Introvae: Introspective variational autoencoders for photographic image synthesis." [Online]. Available: https://arxiv.org/pdf/1807.06358

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks." [Online]. Available: https://arxiv.org/pdf/1812.04948

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/pdf/1512.03385

[7] K. Khabarlak and L. Koriashkina, "Fast facial landmark detection and applications: A survey," *Journal of Computer Science and Technology*, vol. 22, no. 1, p. e02, 2022.

[8] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size." [Online]. Available: https://arxiv.org/pdf/1602.07360

[9] A. Gholami, K. Kwon, B. Wu, Z. Tai, X. Yue, P. Jin, S. Zhao, and K. Keutzer, "Squeezenext: Hardware-aware neural network design," *Design Automation Conference*, 2018. [Online]. Available: https://arxiv.org/pdf/1803.10615

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications." [Online]. Available: https://arxiv.org/pdf/1704.04861

[11] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks." [Online]. Available: https://arxiv.org/pdf/1711.06753

[12] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling, "Pfld: A practical facial landmark detector." [Online]. Available: https://arxiv.org/pdf/1902.10859