Emotion Recognition for Human-Robot Interface

You Li (u6430173)
& $YexiaoLin(u6257745)$
& $ZeruoLiu(u6277483)$
& $HanshanZhang(u5975228)$

June 4, 2019

Abstract

This paper focused on multiple Deep Learning models for emotion recognition and classification from facial expressions. Implementation of ResNet(18), DenseNet, VGG and Attention Model on Japanese Female Facial Expression database were discussed. Variable experiment result represented the effectiveness of learning model, and this proved robust feature extraction with minimized over-fitting in training process. The present study, therefore, proved the feasibility of computer vision based facial expression recognition for practical applications like AI agents and human computer interaction. Keywords: Facial Expression Detection; Deep Learning; ResNet; DensNet; VGG; Neural Network; Feature Extraction; Data Augmentation

1 Introduction

AI agents, such as Affectiva, could understand the human social and emotional context. The key element of this ability is allowing AI agents to correctly identify the human emotions. Facial expressions might be the most direct way to indicate human feelings and help one to express emotional states. It has been proved that Deep Learning based methods are more successful in predicting emotions than the model with handcrafted features which motivated us to attempt Deep Learning methods in detecting human emotion.

In this project, we developed a model to recognize emotional states based on facial expressions. In this model, we try to train the model under multitask settings, add facial region detection, or facial landmark detection task together. The Japanese Female Facial Expression (JAFFE) Database [12] from 10 Japanese models are used as our main reference dataset, which has 213 images of 7 facial expressions, including 6 basic facial expressions: anger, disgust, fear, happiness, sadness, and surprise, and neutral [12]. Each image has been rated on emotion adjectives by 60 Japanese subjects.

We reviewed models and papers related to Facial Expression Recognition(FER), then designed our own models and experiments. Our models take images to provide the corresponding facial expression labels to represent the emotion states in distinct categories. We have tried several models in Convolutional Neural Network (CNN) and Transfer Learning, such as ResNet, DenseNet, and VGG. To be innovative, our group implemented Attention Net to extract more important features for facial expression detection. We compared performance of each model, the outcome could highlight the use of Deep Learning methods.

2 Related work

2.1 Convolutional Neural Network(CNN)

Convolutional Neural Network (CNN)[14] has been widely and profitably applied in the image classification tasks as one of the most effective deep learning methods and we implement it in our project. CNN is a regularized model of multi-layer perceptrons, which is inspired by the biological processes [5] that are similar to the animals visual cortex neural networks. Typically, CNN contains input layer to input data, several hidden layers that includes convolutional layers, activation function layers, pooling layers, fully-connected layers,and an output layer to process with results, as shown in Figure 1.

In particular, because of the topographic property of our network preserving positional information of respective facial features from the bottom to top layers, the translation invariance in Facial Expression Recognition(FER) is inherent in our convolutional architecture. A CNN feature map is determined by comparing the maximum distance of all facial feature elements in feature graph and the mapping relationship between facial feature elements and becomes sensitive to the facial features if the distance is larger. The approach has been verified to

form a detector for the facial action unit in the training process to realize FER. [23] In this project, we used complex pre-trained models based on CNN.

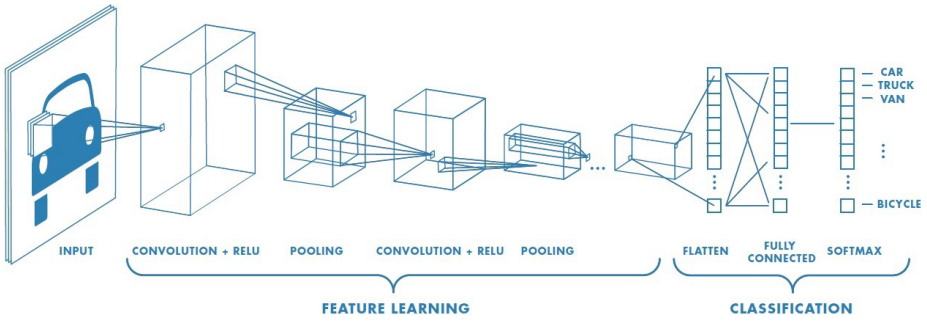


Figure 1: Typical CNN structure

2.2 Literature Review

A relative project applied AffectNet pretrained database of manually annotated images could be a start phase [17]. In that project, different Neural Network architectures were get trained on AffectNet to tackle the tasks of emotion recognition. There were limited dataset, hence they found and collected, annotated, and prepared for public distribution a new database of facial emotions (called AffectNet), which is by far the largest dataset[17].

One related research survey found that in deep FER listed the normal data set and algorithm[11] with recent advance. These papers gave an overview of proposed methodologies and technique for face detection, feature extraction, and classification that uses CNN architecture for improving the performance of the FER and overcome the above challenges. It was highlighted that innovative data pre-processing was the key to improve performance and to avoid over-fitting in FER, which includes detection of face landmarks [16] and face normalization and illumination changes to data set [11]. This provided an inspiration for data augmentation during implementation. Another innovative work was using facial parts as input, rendered a reduction on the size of input and FER processing time, when compared to other CNN architectures that used the whole face for FER[19].

The inputs of original images and landmark processed images are compared. Then, deep learning methods are mostly applied and achieved the best performance[2] which seizes high-level abstractions through architectures to represent and non-linear transform. Considering the data augmentation, it could enlarge the dataset. Finally, they are trained for the classification into emotion categories[11]. We reviewed several pipelines and strategies in terms of FER implemented in this paper and design our own considering the pre-training and fine-tuning in terms of the Transfer leaning.

Another state-of-art method is based on Facial Expression Recognition with Attention Net (FREAtt) [4], we found an attention mechanism to present a new

network architecture. This literature is based on two fundamental models[4] including facial image correction and attention mechanism, as well as facial expression representation and classification. In details, encoder-decoder was applied to convolutional network in order to get an attention map, which performed representation and classification of facial expressions. This paper provided the idea of performing frontal face with proper weight of features [4]. FREAtt could provide obviously better classification performance compared to the conventional models.

3 Implementation

3.1 Models

3.1.1 ResNet

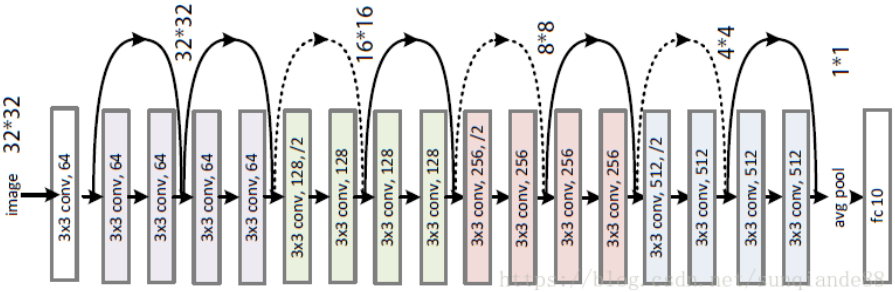


Figure 2: ResNet-18 model structure

Residual Neural Network (ResNet)[22] is one sort of Artificial Neural Network (ANN)[3] that are constructed with two or three layers skips which contain ReLu and Batch Normalization in them[7].

Compared to other traditional Convolutional Network models, the advantages of skipping over layers is to reduce the vanishing of gradients problems[20], because in ResNet a layer l-1 can reuse the activation from the previous layer l-2 and finally the neighbouring layer will learn its weights. Through skips, it simplifies the layers in the original training stages, which can help us speeding up to learn as fewer layers which need to be propagated. In addition, ResNet can reduce the exploration of feature space[10], other network without residual components may appear more vulnerable to perturbations and will cause that it leaves the manifold and this situation will require more training data to fix up. In this project, we applied ResNet-18 to do the facial recognition task, and the model structure can be found in Figure 2.

3.1.2 DenseNet

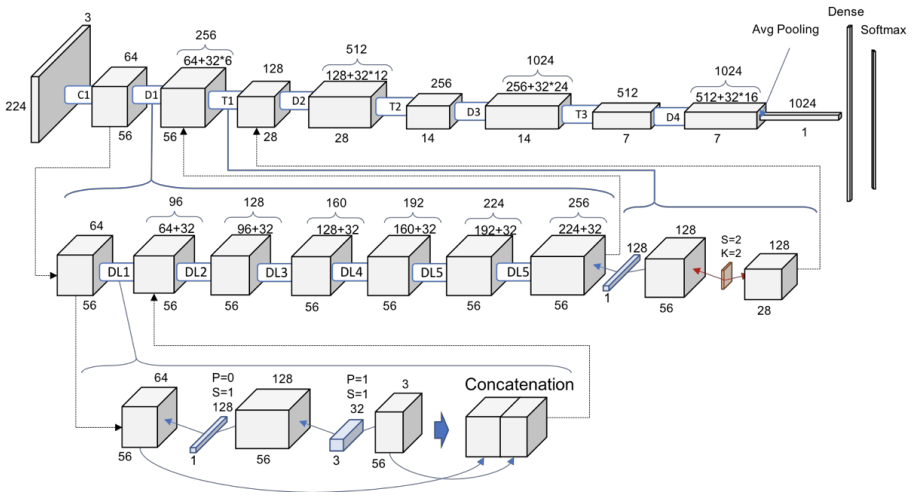


Figure 3: Dense-121 model structure

In Dense Convolutional Network (DenseNet)[8], every layer is connected to each other layers shown in Figure 3. The core attributes of DenseNet are that all layers in networks will receive the collective features from all the previous layers and then add its own feature map as input data pass to all the following layers, and the final classification output will be decided on the whole networks feature maps.

Similar to ResNet as above mentioned in Section 3.1.1, DenseNet also helps to eliminate the vanishing of gradient problem[20], also the error messages can be faster propagated directly to the previous layers. Compared to the other Neural Network models, like ResNet, DenseNet significantly decreases the number of parameters because the reusing of the dense connectivity network features, such efficient use of parameters can help improving the flow of information.

3.1.3 VGG

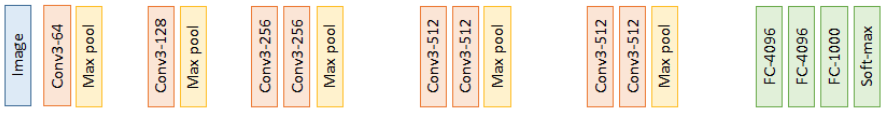


Figure 4: VGG11 model: each convolutional block is in The Conv-BatchNormalize-ReLU structure

VGG is proposed by a group named Visual Geometry Group. VGGNet-16 consists of 8 convolutional layers, 5 max-pooling layers and 3 fully-connected layers, show in Figure 4 and it is very appealing because of its very uniform architecture [6]. VGG is the best choice in extracting features from different figures. The weight configuration of the VGG is publicly available and it has been applied in a wide range of applications and state of affairs as basic feature extractors[1]. However, VGGNet also has shortcomings that it consists of 138 million parameters, which can be a lot resource-consuming and time-consuming to deal with[1].

3.1.4 Attention Model

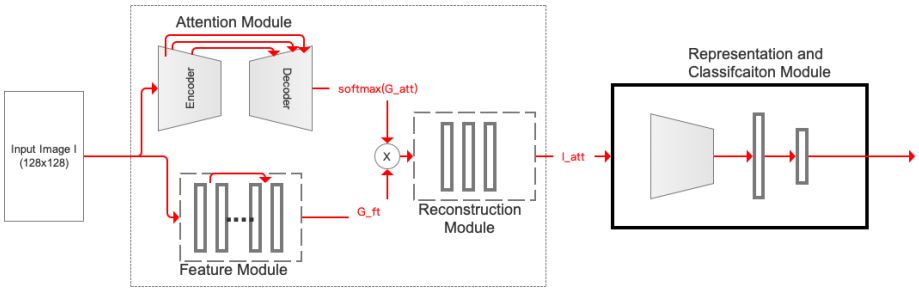


Figure 5: **Architecture of Attention Model**: In this project, we mainly follow [4]'s setting. There are four major modules in our model; attention module $\sigma(G_{att})$, feature module G_{ft}, reconstruction module I_{att} and classification module G_{cls}.

In terms of attention model,we implement the attention mechanism to recognize the frontal face and jointly train from attention and classification. The whole model architecture could be found in 5 and mainly followed the setting of [4]. In this architecture, There are four major modules in our model; attention module $\sigma(G_{att})$, feature module G_{ft}, reconstruction module I_{att} and classification module G_{cls}. Compared to the [4], we change the attention module's encoder to ResNet-18 for the limited number of training data set.

The training object is also changed as following:

$$\lambda \mathcal{L}_{att}(\sigma(I_{att}, I \otimes I_{marsk}) + \mathcal{L}_{cls}(\hat{y}, y) \tag{1}$$

where $\mathcal{L}_{att}$ means square error loss function, $\mathcal{L}_{cls}$ is cross-entropy loss function and λ is the trading off parameter between attention loss and classification loss ($\lambda = 10$ in our project). This focus attention mechanism is to target at the learning the relevant part of the frontal face and not the irrelevant parts of the figures.

3.2 Data Augmentation

3.2.1 Horizontal Flip for RestNet, DenseNet, VGG

As one of data prepossessing step, we did horizontal image flip with random degrees. This step was done within the considerations to allow the Neural Network to learn more realistic faces. In this way, our group extended the expression database for expression happy, sad, disgust, fear, surprise, neutral and angry. Each facial expression was expanded by random flip ten times with expression label and facial landmarks before passed to training.

3.2.2 Add Background for Attention Model

To augment the facial data, we first implemented facial landmark to detect the face edge, and then we gained cropped faces. After that, we transposed the face, and put the transposed faces on the random gray-scale background. Finally, the images were adjusted by brightness and contrast of those compounded images. The goal to augment was letting the Neural Network understand and got trained itself to ignore the irrelevant FER factors, such as background, transposition, and image color attributes. Robust face detection with invariance properties in terms of translation, scale, and pose, inherented in our CNN model brought robustness to dynamical changes both in head movements and in facial expressions. [15]

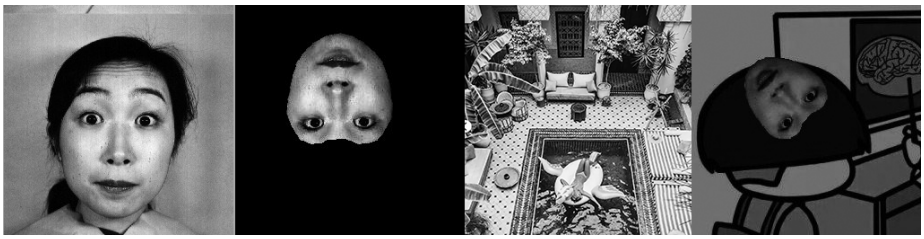


Figure 6: Data Augmentation: Add background to training face image

3.3 Transfer Learning

Transfer learning is a deep learning method developed for a task which is reused as the starting point for a model on a second task. It is a popular in Deep Learning for pre-trained models used as a starting point on computer vision and tasks given the vast compute and time resources required to develop Neural Network models on these problems. Facing the limited number of JAFFE data examples, we tried to use two different transfer learning approaches with the help of large scaled pre-trained models. We applied two transfer learning approaches: feature extractor and fine-tuning with CNN modules mentioned in Section 2.1.

3.3.1 Feature Extractor

Feature Extractor is a method to start on a measured dataset and build derived values as features to facilitate the sequence learning[13]. During this process, we can remove last few layers of the trained network and retrain with fresh layers for target job. It is used in this project because it can process the large-dimension data to reduce the dimensions as well as a set of features number. It is a method only extracting the useful data[13].

3.3.2 Fine-tuning

Fine-tuning is a concept of Transfer learning which is to fine-tune the weights of pre-trained Neural Network by continuing the back-propagation [21]. In deep learning, the last few layers are more likely being related to tasks. Fine-tuned learning experiments require resources to adjust precisely to the observation, but they are still much faster than learning from scratch. Additionally, they are more accurate compared to models trained from scratch.

4 Experiment results

In this section, we tried to perform three experiments with different training settings. In the first two, RestNet-18, DenseNet and VGG were used as main models under different training strategies: training from starch, using pre-traind modes as feature extractor or fine-tuning towards the pre-trained model on JAFFE. The models were trained under different training task settings: facial expression classification task, and facial expression classification with landmark prediction as an additional task. In the end, we tried to introduce attention mechanism followed by [4] exploration of the model abilities of whether it can directly ignore the background information and only focus on the facial area.

4.1 Evaluation Matrix

We reported both accuracy and F1 score as the key evaluation metrics in this project.

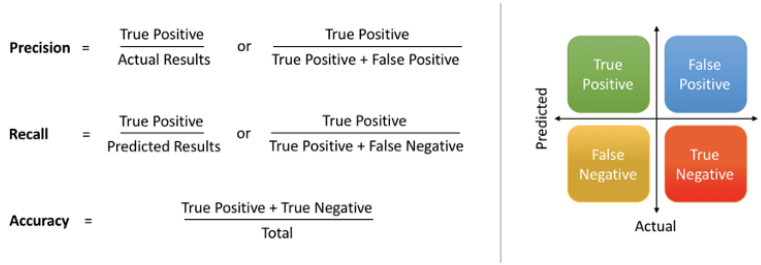


Figure 7: Evaluation Matrix calculation

4.2 Data Set

We randomly split the original data set with 80% for training and 20% for testing to provide a equal chance to train faces of each actress. Models' performance results are reported based on the same testing set.

4.3 Baseline

In this project, we will compare our classification models' performance with following works:

1. Jabid, Kabir and Chae (2010) proposed a appearance-based feature descriptor, the local directional pattern (LDP) to encode the facial geometry information[9]. They tried to use this facial geometry representation approach with two traditional machine learning approaches: Template matching and Support Vector Machine(SVM) as classifier. We will use the best template matching and SVM model from their work as our baseline.

2. Neagoe, Brar, Sebe and Robitu (2013) evaluated two CNN model, one Deep Belief Network (DBN) and one SVM with Linear kernel[18]. In this report, we will include the best performance CNN model from their work. Their CNN model contains 3 ConvNet (Conv-Maxpooling) and 2 fully-connected layers.

3. Nwostu et al. (2018) also proposed a CNN based model but they tried let the model directly focus on eye and mouth[19]. They first extracted face region from original images and then they extracted eye and mouth regions based on face region. By feeding these two region images into two CNN model, finally, one fully-connect layer are used to combine the information.

4.4 Setting 1: No landmark information

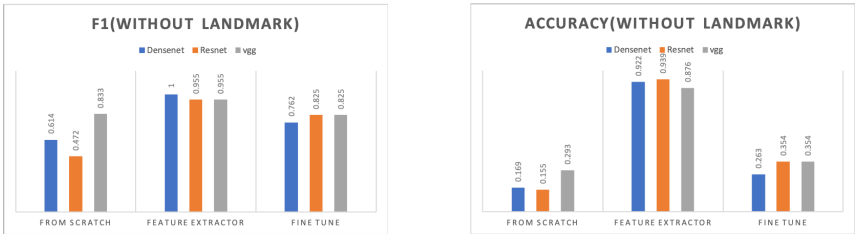
At this stage, we trained the model only with the original image in training set and we tested the model on testing data set. There are three different training settings in all:

1. **Training from starch**. We randomly initialized the model parameters and trained the model on the training set.

2. **Feature Extractor**. We loaded the pre-trained model from Pytorch vision models [1]. We replaced the original classification layer with our own task specific classification layer. During training, we only updated the classification layers and parameters of rest models are remained unchanged.

[1] https://pytorch.org/docs/stable/torchvision/models.html

3. **Fine-tuning**. We also used the pre-trained model, and changed the classification layer, but the whole model's parameters were updated with learning error.



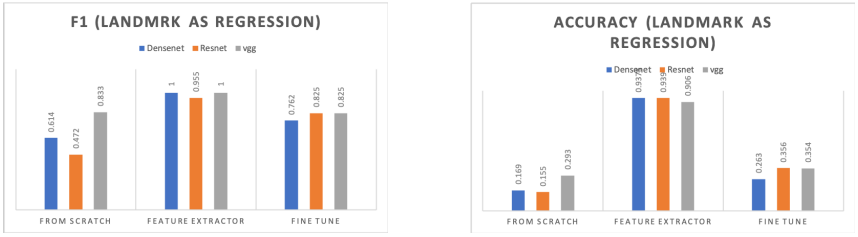
(a) F1 score (b) Accuracy

Figure 8: F1 and Accuracy score for three models with no landmark information

As shown in Figure 8, using models as feature extractor has achieved the best performance among other two settings.

4.5 Setting 2: Landmark as regression task

With the help of *face_recognition* tool[2], we could easily obtain the face landmarks for each image in our data set.
With extra face landmarks information, we firstly considered it as another training task, regression task. In order to predicting the landmarks, we normalized the landmarks information into $[0, 1]$ and added another fully-connect layer for predicting landmarks which used the same features as it was used for facial expression classification layer. The loss function for regression is the mean square error. Under this setting, we wanted the model to capture the more generalizing feature representations for expression tasks. The models were trained under the same three training settings as it in section 4.4.



(a) F1 score (b) Accuracy

Figure 9: F1 and Accuracy score for three models with no landmark information

[2]https://github.com/ageitgey/face_recognition

As shown in Figure 9, Feature extractor setting still is the best training strategy for our data set. Landmark regression task also slightly improve the performance of three models under feature extractor setting.

4.6 Setting 3: Attention model

Instead of letting the model to predict the landmarks for each facial image, we considered the another option that letting the model to learn the facial region area and put more attention on this area by ignoring the backgrounds. With this idea, we have following [4]'s model setting and change the model Resnet component to be a more shallow one(ResNet18) which requires less training data.

Figure 10: Visualization of Attention model. From left to right, the images are from training data example, feature extraction module, reconstruction module, and true attention target

The training results are visualized in Figure 10. From the result, our reconstruction module does rebuild the most important image patterns for faces like eyes, noses, mouths, and hair. Taking our data augment methods into account, the model can successfully capture the facial patterns regardless of their position and rotation.

4.7 Discussion

Model	Acc	F1
Densenet-128	0.9375	1
CNN[19]	0.9771	*
CNN [18]	0.6522	*
LDP[9]	0.869	*

Table 1: Expression recognition performance on JAFFE datase

We have attempted several methods in all to achieve better performance for the expression recognition and we match those different methods together and analyze each performance. It can be seen that the landmark, feature extractor are all perform positively good to train the module. Our attention model also

learned to focusing on the correct facial region although it's F1 score and accuracy source ls lower than other three models. One possible explanation of this is that our data is relatively small which has only around 200 images and the attention model is almost 2 times bigger than other three models. There might be the over-fitting problem performed on this data set.

Compared to the results in Section 4.4 and Section 4.5, it can be concluded that the large data pre-trained model is the best option for the new tasks if there is only a small amount of labeled data set. It is because the pre-trained models can capture the general image patterns which is beneficial for the large scale images and these learned patterns can be used for almost all image related tasks. Compared with the three training settings introduced in Section 4.4, The models under learning from scratch and fine-tuning settings are over-fitted. It suggests that facing limited labeled training set, the best option might be using the pre-trained image classification model and taking it as a feature extractor.

The performances of our best model settings and three baselines are listed in Table 1. Our model can beat two of them except CNN from [19]. The one possible explanation of this is that the proposed CNN model from [19] can be tough as a hard attention implementation , which directly focuses on the eyes and mouth. However the limitation if this setting is that the corresponding region need to be correctly extracted. Our model does not need this kind of pre-possessing and it can achieve similar result which suggests that our model could be applied to more facial databases comparing to [19]'s. Comparing our model to LDP from [9], one advantage is that our model don't need to handcraft the feature and it can learn the needed feature by itself.

5 Conclusion

5.1 Conclusion

Our CNN models and Transfer Learning are performed effectively to increase the effects of the learning faces emotions. We first learn from scratch with and without landmark. After that, Transfer Learning performed on CNN models including ResNet, DenseNet and VGG as feature extractor and fine-tuning have been compared and tested respectively with and without landmark. To be innovative, a model combining attention net (focus on frontal face like eyes and mouth) and CNN together have been attempted to construct a more robust Neural Network. Since it is a small dataset, we also perform the data augmentation to enlarge the dataset respectively by landmarks on different models to get better performance.

After the control experiment, Feature Extractor towards the pre-trained classification performed the best among all the other implementation. The limitations might be caused by the over-fitting problem after our analysis.

5.2 Future Work

More reliable data set is required to perform our different models, since JAFFE database is limited in data size and facial expression classes. Besides, practical modification on training models is considered as a future task. Applications in terms of the real-time detection for different human faces emotions would provide better computer interactive experience of this research task. Overall performance of convoluted neural network could considered as a good approach, less focus required in this criteria. However, more experiment on variable parameters could be experimented to improve the performance Attention model. Innovative learning model and method to avoid over-fitting are related research topics for facial expression detection.

6 Summary of Learning Outcome

1. **Pytorch**. Through this project, we have learned the most commonly used deep learning framework, *Pytorch*. With its help, we can easily build the deep learning model and train the model on GPU servers.

2. **Learning three model architectures for image classification**. We successfully applied three model architectures, ResNet,DenseNet and VGG to our facial expression classification tasks and try to analyze their performance under different training settings. From the experiments, we found that it is extremely useful to only fine-tuning the classification layers facing the small training data set.

3. **Exploring the Landmark information** Instead of using the original images as training set, we also tried to add landmark information into training processing with help of facial_recognition tool. By considering the extra landmark as regression task, we improve the performance of two pre-trained model, DenseNet and ResNet.

4. **Attention** Instead of using landmark information as regression task, we also tried to apply the attention mechanism in to our task, helping the model can ignore the background information and only focusing on the face region. From the result in Section 4.6, even with a small amount of training set , our model can still achieve the expected learning outcome. However, the model's F1 and accuracy are lower than other models. One possible reason is that the data set is not large enough to train the model.

References

[1] S. Das and S. Das. Cnn architectures: Lenet, alexnet, vgg, googlenet, resnet and more ...., Nov 2017. URL https://medium.com/@sidereal/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5.

[2] L. Deng, D. Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[3] A. Elkamel. An artificial neural network for predicting and optimizing immiscible flood performance in heterogeneous reservoirs. *Computers Chemical Engineering*, 22(11):1699 – 1709, 1998. ISSN 0098-1354. doi: https://doi.org/10.1016/S0098-1354(98)00237-3. URL http://www.sciencedirect.com/science/article/pii/S0098135498002373.

[4] P. D. M. Fernandez, F. A. G. Peña, T. I. Ren, and A. Cunha. FERAtt: Facial Expression Recognition with Attention Net. 2019. URL http://arxiv.org/abs/1902.03284.

[5] K. Fukushima. Neocognitron. *Scholarpedia*, 2(1):1717, 2007. doi: 10.4249/scholarpedia.1717. revision #91558.

[6] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*, pages 262–270, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/CVPR.2016.90.

[8] G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, and K. Weinberger. Convolutional networks with dense connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2918284.

[9] T. Jabid, M. H. Kabir, and O. Chae. Robust facial expression recognition based on local directional pattern. *ETRI Journal*, 32(5):784–794, 2010. ISSN 12256463. doi: 10.4218/etrij.10.1510.0132.

[10] B. Li, Z.-T. Fan, X.-L. Zhang, and D.-S. Huang. Robust dimensionality reduction via feature space to feature space distance metric learning. *Neural Networks*, 112:1 – 14, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.01.001. URL http://www.sciencedirect.com/science/article/pii/S0893608019300103.

[11] S. Li and W. Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[12] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. The japanese female facial expression (jaffe) database. In *Proceedings of third international conference on automatic face and gesture recognition*, pages 14–16, 1998.

[13] E. A. A. Maksoud, S. Barakat, and M. Elmogy. Medical images analysis based on multilabel classification. In *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*, pages 209–245. Elsevier, 2019.

[14] S. Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016. doi: 10.1098/rsta.2015.0203. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2015.0203.

[15] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5):555 – 559, 2003. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(03)00115-1. URL http://www.sciencedirect.com/science/article/pii/S0893608003001151. Advances in Neural Networks Research: IJCNN '03.

[16] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

[17] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017.

[18] V. Neagoe, A. Brar, N. Sebe, and P. Robitu. A Deep Learning Approach for Subject Independent Emotion Recognition from Facial Expressions. *Recent Advances in Image, Audio and Signal Processing*, 978(Dl):960, 2013.

[19] L. Nwosu, H. Wang, J. Lu, I. Unwala, X. Yang, and T. Zhang. Deep Convolutional Neural Network for Facial Expression Recognition Using Facial Parts. *Proceedings - 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 2017 IEEE 15th International Conference on Pervasive Intelligence and Computing, 2017 IEEE 3rd International Conference on Big Data Intelligence and Computing and 2017 IEEE Cyber Science and Technology Congress, DASC-PICom-DataCom-CyberSciTec 2017*, 2018-January:1318–1321, 2018. doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.213.

[20] L. D. Pascale, J. Louet, and F. Santambrogio. The monge problem with vanishing gradient penalization: Vortices and asymptotic profile. *Journal*

de Mathématiques Pures et Appliquées, 106(2):237 – 279, 2016. ISSN 0021-7824. doi: https://doi.org/10.1016/j.matpur.2016.02.009. URL http://www.sciencedirect.com/science/article/pii/S0021782416000106.

[21] Standford. convolutional neural networks for visual recognition. URL http://cs231n.github.io/transfer-learning/.

[22] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu. Residual networks of residual networks: Multilevel residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(6):1303–1314, June 2018. ISSN 1051-8215. doi: 10.1109/TCSVT.2017.2654543.

[23] Y. Zhu, H. Fan, and K. Yuan. Facial expression recognition research based on deep learning. *CoRR*, abs/1904.09737, 2019. URL http://arxiv.org/abs/1904.09737.

Peer Review and Group contribution Statement

Name	ID	Workload distribution	Contribution Ratio
You Li	u6430173	Data Loader, Fera attention Model, **Data Augmentation**, CNN Net, **Report**, Jaffe, Presentation, **Literature review, final editing**	40%
Yexiao Lin	u6257745	Presentation, **Report citation**	8%
Zerou Liu	u6277483	Data Loader, **Presentation**, **Report**	12%
Hanshan Zhang	u5975228	**ResNet and other training models**, **take GPU to train the model,** evaluate matrix, literature review, Presentation, Report.	40%