Emotion Recognition from Facial Expressions Using Siamese Network

Naga Venkata Sesha Saiteja Maddula, Lakshmi R. Nair, Harshith Addepalli, and Suja Palaniswamy(✉)

Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru 560035, India
maddulasaitej@gmail.com, lakshmirajnairl@gmail.com, addepalliharshith@gmail.com, p_suja@blr.amrita.edu

Abstract. The research on automatic emotional recognition has been increased drastically because of its significant influence on various applications such as treatment of the illness, educational practices, decision making, and the development of commercial applications. Using Machine Learning (ML) models, we have been trying to determine the emotion accurately and precisely from the facial expressions. But it requires a colossal number of resources in terms of data as well as computational power and can be time-consuming during its training. To solve these complications, meta-learning has been introduced to train a model on a variety of learning tasks, which assists the model to generalize the novel learning tasks using a restricted amount of data. In this paper, we have applied one of the meta-learning techniques and proposed a model called MLARE(Meta Learning Approach to Recognize Emotions) that recognizes emotions using our in-house developed dataset AED-2 (Amrita Emotion Dataset-2) which has 56 images of subjects expressing seven basic emotions viz., disgust, sad, fear, happy, neutral, anger, and surprise. It involves the implementation of the Siamese network which estimates the similarity between the inputs. We could achieve 90.6% of overall average accuracy in recognizing emotions with the state-of-the-art method of one-shot learning tasks using the convolutional neural network in the Siamese network.

Keywords: Emotional recognition · Meta-learning · Machine Learning · Siamese network

1 Introduction

1.1 Challenges of ML in Emotion Recognition

Emotions are generated unconsciously to the extrinsic and intrinsic events, which express some particular physiological states. Our ability to understand and control emotions play a crucial role in taking appropriate decisions and actions in life. Using ML models, researchers have been trying to determine the emotion accurately and precisely from facial expressions. Though ML could reduce the complexity in the computational problems, challenges emerge about the business application like facial emotion recognition. For instance, a simple face or emotion recognition system

S. M. Thampi et al. (Eds.): SoMMA 2020, CCIS 1366, pp. 63–72, 2021.
https://doi.org/10.1007/978-981-16-0419-5_6

demands a large number of well prepared and organized images in diverse format during training. Even an accurate ML cannot create viable results on limited amounts of data, since the data has dispersed and proliferated. So if a model trained on Asian subjects, it is challenging for the model to perform well on African subjects. Also, there is a great need for storage requirements for housing all the data and require a large amount of computational power, thereby leading to a large amount of money and may turn economically infeasible at times.

The distinct methods discussed in papers [1, 2], and [3] to recognize the emotions need a good number of resources in terms of data as well as computational power. Work presented in [1] is an extension of [4] using a layer-based CNN method from the CMU Multi-PIE dataset. The proposed model called DPIIER could achieve an average accuracy of 96.55% in 3 - fold cross-validation on 7,50,000 images taken under 15 different viewpoints and 19 illumination conditions. The requirement and management of this huge amount of data to obtain acceptable performance is a laborious task. Instead of images, 4D videos consist of 60,600 frame models used in [2] to recognize emotions by analyzing surface normals and curves and multiclass Support Vector Machine(SVM) used for classification. Though the proposed algorithm could give an appreciable average recognition rate for curves and surface normal, only few emotions have the foremost recognition rate. A geometric approach proposed in [3] using BU3DFE dataset could achieve an overall accuracy of 83.3% on emotional recognition. The method proposed in [5] applied feature level fusion technique on the AED2 dataset by considering facial expression, gestures, and both in the emotional recognition system. The method could achieve a discernible result on the limited data and suggested to leverage the deep neural network for preferable performance.

The work discussed in this paper is an extension of [5] by analyzing solely facial expression using meta-learning technique and in-house dataset having limited data. We designed a model based on the Siamese network algorithm which is the state-of-the-art method in the field of emotion recognition and obtained fine accuracy.

1.2 Meta-learning

In the scenario where the data is inadequate to constrain the problem, the paradigm called Meta-Learning has great significance. Meta-Learning at an abstract level refers to the set of problem-solving strategies that involves adapting to the new environment easily and also can train with very few examples by the procedure of learning how to learn. A fine meta-learning model has to be trained over distinct learning tasks and optimized for the leading performance on the distribution of tasks, which includes the unprecedented tasks also.

The subsequent section explains the different approaches of meta-learning and the work associated with face recognition using the Siamese network which is an incentive for us to implement the Siamese network in the emotional recognition task. Section 3 elaborates the design and implementation of the proposed model which is termed as-MLARE(Meta-Learning Approach to Recognize Emotions). Experimental results comprehend in Sect. 4 followed by conclusion and future scope.

2 Related Work

2.1 Meta-learning Framework

A Meta-Learning framework consists of training tasks and testing tasks, where we learn how to learn to classify given a set of training tasks and evaluate by using a set of test tasks. Each task in the meta-learning framework is associated with a support set and a query set. The model estimates the efficiency of learning on the query set for each task during the training period. The framework uses completely distinct tasks at test time to evaluate the model.

One-shot, K-shot and few-shot are few terminologies related to meta-learning framework. The intention of few-shot learning is to train the learning model with a few training examples. If the model learns a single example of each class, it is known as one-shot Learning. We can generalise few shot learning as N-way-k-shot classification where the model learns k examples from N classes.

2.2 Approaches to Meta-learning

The classification of meta-learning approaches is quite subtle since groundbreaking algorithms are being pioneered in this field. Recent research in meta-learning like Automatic Domain Randomization [6], has been focusing on data augmentation which democratizes the deep neural network in the domain of a limited amount of data. In fact, we classify meta-learning approaches into model-based learning, metric-based learning, optimization-based learning, and data-based learning.

Metric Based Learning
In metric based learning, the model compares the data in relevant feature space to discriminate unseen classes by learning the embeddings in the training task. Koch et al. propose a Siamese network [7] discriminate against the two unseen classes while models like Matching network [8], and Prototypical network [9], etc., discriminate many unprecedented classes by exploiting the prior knowledge about similarity. The designed models in [7, 8] and [9] are based on the nearest neighbor principle, which cannot derive the correlation between the inputs if the dimension of the data is high. In this scenario, Sung et al. introduced RelationNet [10] and Allen et al. proposed an Infinite Mixture of Prototypes [11] to eradicate the issue with the high dimensionality of data. The model in [11] is an extension of the prototypical network in [9] since it sets the model capacity in an adaptive manner based on the complexity of data. Metric based algorithms are computationally fast and maintain a fine learning consistency in most of the cases.

Model-Based Learning
The fundamental idea of model-based learning is fast parameterization for rapid generalization on the new task by training limited data. Santoro et al. proposed Memory Augmented Neural Networks in [12] by combining Neural Turing Machine(NTM) and Long-short-Term-memory(LSTM) with external memory to keep unassailable details or knowledge obtained from the training tasks and use this knowledge during deduction. Meta networks in [13] also use external memory with fast and slow weight for the

rapid generalization. Though the architecture of the models is complex, it maintains a very high learning capacity.

Optimization-Based Learning

The objective of Optimization-based learning algorithms is to optimize the procedure of acquiring task-specific parameters to show the finest performance. Finn et al. introduced Model-Agnostic Meta-Learning (MAML) in [14], to train the initial parameters of the model so that model can achieve optimal fast learning on the novel tasks by updating the parameters in a few gradient steps. One of the main challenges related to MAML is that it requires a deep neural network architecture to get a fine gradient update. Kim et al. came up with Auto-Meta in [15] to solve this issue by selecting optimal architecture for MAML. Another limitation of the MAML is the unreliability in the second order optimization method. The algorithms like Meta-SGD [16], Alpha-MAML [17], etc. can eradicate the aforementioned problem. The optimization-based models have high learning capacity as well as good learning consistency. But models are computationally expensive since they require the second-order optimization method.

Data-Based Learning

Meta-learning algorithms in the data space allow the model to expand and enhance the correlation between the data by creating diverse data. For instance, Automatic Domain Randomization (ADR) [6] that controls the distribution of the training data in simulation and assists the neural network to generalize to the limited real-world data.

2.3 Siamese Network and One-Shot Learning

Siamese network is another type of neural network that employs a different way to triage inputs based on the similarity. In general, a neural network demands a colossal number of data to build the model which shows good performance. But if the data is inadequate to constrain the problem like signature verification [18], it is impossible to obtain many copies of data samples per person. So, this kind of one-shot learning problem is the principle behind designing the Siamese network, consisting of two symmetrical neural networks with the same parameters. The symmetrical networks joined at the end using an energy function as shown in Fig. 1.

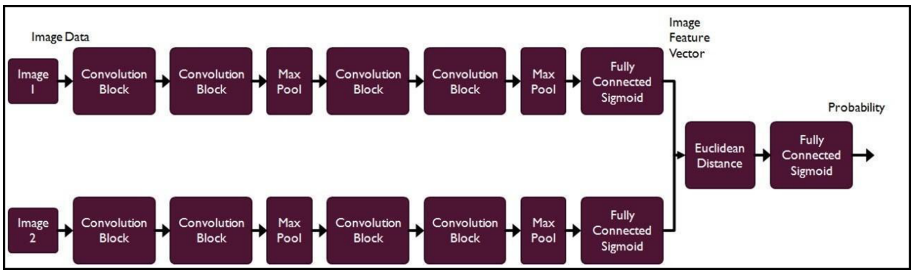

Fig. 1. Siamese network architecture

This energy function serves as the distance function which scrutinizes whether two inputs are in the same class or not. [19, 20] and [21] gives the insight about the Siamese network architecture to implement in the emotional recognition task.

In the classical face recognition system, there are 3 modules Viz face detection, face alignment, and face recognition. The paper [19] ignored the face alignment part to reduce the complexity of programming. The authors adopt cascade classification based on Haar features to detect the faces and use the convolutional Siamese network to recognize the face. This network trains layer by layer gives 92% accuracy on the ORL face database. The model estimates a high value if the sampled image and candidate image are in the same class otherwise it gives a low value. The authors in [20] demonstrate the ability of CNN to deal with the issues in recognizing the face in unconstrained nature. The authors made a slight modification in the architecture of the Siamese network by replacing the cost function with a simple Multi-Layer Perceptron classifier. The dataset CelebFaces Attributes Dataset(CelebA) used for the modeling has more than 200 celebrity images and covers clutter and pose variations. Images that come under the same class are known as genuine pairs, and images of two discrete people are called impostor pairs. The CNN model with the Siamese network gave an accuracy of 85.74% on the CelebA dataset.

In [21], the authors proposed a method to detect face liveliness through 2 stages called offline training stage and online testing stage. During the training period, the authors feed pairwise images in the database to the Siamese network. The pairwise images can be actual images or one hoax and one real image also. If the model estimates the two images are actual, then that images are known as a positive pair or else, it is called a negative pair. In the testing time, the face recognizer in the model identifies the test images and extracts the identity information of the client. Followed by retrieving the actual face image and the test images feed to the Siamese network for the face liveness detection. Euclidean distance used to compare the feature space of the input images in the model. If two images are real face images, it shows the values of 1 otherwise 0. The convolutional neural networks in the model are basically Alex Net architecture that consists of 5 convolutional layers and 3 pooling layers. It is important to note that the aforementioned methods in [19, 20] and [21] require a handful images to obtain viable results. Instead of classifying, Siamese Network demonstrates to differentiate the images by learning similarity functions.

3 Approach

3.1 Dataset and Preprocessing

We have used an in-house dataset named Amrita Emotion Dataset-2 (AED-2) [5] comprising 7 classes where each class represents a specific emotion which includes Anger, Sad, Happy, Disgust, Surprise, Neutral, and Fear. The number of objects posed for this dataset are 8 and hence 56 images in total. A few sample images of AED-2 dataset is shown in Fig. 2. Since the size of the images are different from each other, we

converted each image into a 256 × 256 grayscale format i.e. the portable gray map (.pgm). Thus, the model needs to be trained only on these images to classify and identify the emotion.

Fig. 2. Sample image of AED-2 dataset

3.2 Experimental Setup

One of the implementation challenges faced with the implementation of meta-learning is that there are no available libraries for implementing the algorithms. All the implementations need to be carried out from scratch which is time-consuming and also difficult in carrying out the operations efficiently. The implementation framework is similar to that of neural networks. The Fig. 3 shows the framework of activities that are needed to be carried out for the implementation of the project. In the preprocessing step we cropped the images only for the region of interest and converted into uniform size (256 * 256) for faster processing. In the next step we are converting the images into a grayscale format(.pgm) to manage the data in a simple and convenient manner.

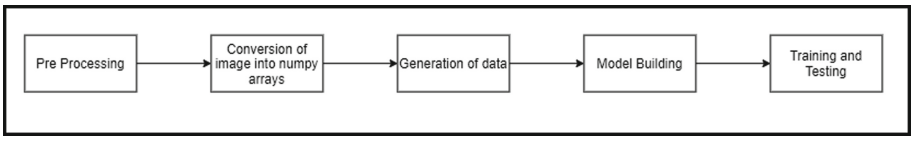

Fig. 3. Block diagram for the implementation of MLARE model

To enhance the performance of the predictive model, python converts the images into NumPy arrays in the format of [height, width, channel]. To explicate the buffer as a

one-dimensional array, the image has been converted to NumPy array using np. frombuffer. Once the conversion of data into a numpy array is done the next step is the generation of data. The Siamese network takes data in the form pairs(genuine and imposter). Initially, the function reads the images (img1, img2) from the same directory as shown in Fig. 4 and stores them in the x_genuine array and assigns y_genuine to 1. Subsequently, the same function reads the images (img1, img2) from the different directory as shown in Fig. 5 and stores them in the x_imposter pair and assigns y_imposter to 0.

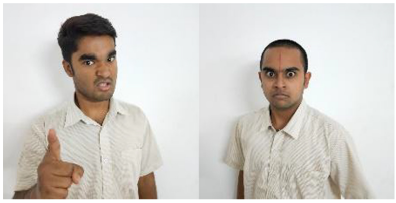

Fig. 4. Example of Genuine pairs

Fig. 5. Example of Imposter pair

Finally, it is required to concatenate both x_imposter, x_genuine to X and y_imposter, y_genuine to Y. Each time this function is executed it generates 112 genuine pairs and 112 imposter pairs which determine the batch size of 224 data points.

MLARE model comprised of 2 convolutional layers with 32 number of filters and kernel_size 3. Border_mode has been assigned as 'valid' to get an output that is smaller than the input. The parameter, dim_ordering given as 'th', since we used TensorFlow as backend. In MLARE rectified linear unit (ReLU) is used as an activation function in hidden layers and max-pooling of size 2X2 followed by a flat layer with an activation function of the sigmoid. After designing the model, feed the image pair to the neural network, which transforms these images into feature vectors. Then the feature vectors, featvec_a, and featvec_b feed to the similarity function to estimate the resemblance between the two input images. In the MLARE model, Euclidean distance is used as the similarity or energy function. Rather than generating all the data at once here we are using batch training for every iteration or epoch. The function has been called to generate data and to feed into the network to train. Number of epochs is set to 15 and Root Mean Square Propagation(RMS prop) is used for optimization. RMS prop is one of the gradient optimization algorithms that take away the need to adjust the learning rate and does it automatically. Binary cross-entropy loss function(BCE) is used along with sigmoid activations as the cost function to achieve better accuracy.

4 Results and Analysis

After carrying out 420 epochs, the training loss becomes less and accuracy reaches about 94.26% and with this, we stop training the model. The next step is validating the proposed model. So to validate the model we created one batch of data points again by

calling the get_data function. It is observed that the validation set gives an accuracy of about 93.75% on the test batch generated for data. So this is the ability of the MLARE model to discriminate the images by learning similarity functions. The subsequent step is to determine the label of the image. For that, we take a test image for which the label needs to be specified, and then an image from each of the classes(anger, neutral, sad, etc.) will be drawn to compare the feature space of the test image and other images. For whichever images it matches more or is close that is the label associated with the image. The authors in [5] proposed two methods to recognize emotions. In the earliest approach, emotions are recognized by considering facial expression and gestures independently. In the next approach, authors combined the features extracted from gestures and facial expression to recognize emotions. The overall average accuracy to recognize emotions from the facial expression in [5] is 86% while that of using the MLARE model is 90.6%. A comparison of the performance of the two models is pictorially represented in Fig. 6.

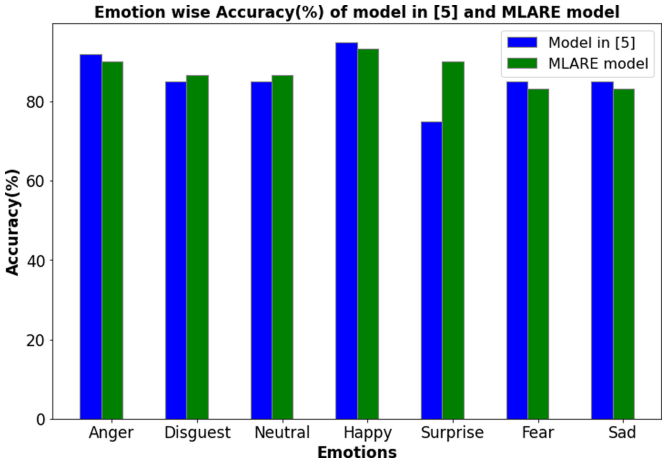

Fig. 6. Bar chart represents the emotion-wise accuracy of two models

We could observe that the MLARE model executed better than the model used in [5] on all emotions and give an extraordinary result on the 'Surprise' emotion. Both of the models could obtain the highest accuracy in 'Happy' emotion as compared to other emotions. It is evident that we could improve the overall average accuracy to recognize emotions from facial expression by enhancing the ability of the MLARE model to differentiate the images.

Figure 6, we can observe that emotions like anger and happiness results in relatively high accuracy as compared to others. In the real world, anger and happiness are the most essential emotions to recognize the feelings of others in a communication. Since the model, even if it is machine learning or meta-learning it actually tries to emulate the human cognitive functions. That is the reason why most of the models show high performance in these aforementioned emotions.

5 Conclusion and Future Scope

In this paper, we have introduced a model called MLARE consisting of two CNNs, for emotional recognition using an in-house dataset and achieved an encouraging performance on the limited data. Along with recognizing emotion from an image, the MLARE is learning similarity functions and manifesting how similar the two images are. In this model, we have chosen BCE+ sigmoid function as the loss function since we use pairs of training data. The MLARE can be modified by changing the loss function to triplet function in future work.

Through data augmentation methods, we can generate more data images from the given 56 images in the different formats to increase the diversity for the training model. Normally simple transformations like cropping, padding and horizontal flipping are used in data augmentation. Further, we can attempt the state-of-the-art methods in data augmentation- like Population-Based Augmentation(PBA) or AutoAugment, to enhance the learning ability of the MLARE on emotional recognition tasks.

Acknowledgement. We express our sincere gratitude to Mr. Pranav B. Sreedhar, for his constant support extended towards this work. His suggestions and exceptional knowledge in the field of meta-learning helped us to explore meta-learning and to complete the work successfully. We are indebted to Chiranjiv, Pranav, Ronak, Sailakshmi, Srilakshmi, Vinayak, Vasuman, and Srivathsan for expressing the required emotions for the dataset and we are thankful to Anupam for coordinating the dataset preparation.

References

1. Palaniswamy, S., Suchitra: A robust pose & illumination invariant emotion recognition from facial images using deep learning for human-machine interface. In: 4th International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, pp. 1–6 (2019)
2. Sai Prathusha, S., Suja, P., Tripathi, S., Louis, L.: Emotion recognition from facial expressions of 4D videos using curves and surface normals. In: Basu, A., Das, S., Horain, P., Bhattacharya, S. (eds.) IHCI 2016. LNCS, vol. 10127, pp. 51–64. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52503-7_5
3. Swetha, K.M., Suja, P.: A geometric approach for recognizing emotions from 3D images with pose variations. In: International Conference on Smart Technologies for Smart Nation (SmartTechCon), pp. 805–809. IEEE (2017)
4. Suja, P., Tripathi, S.: Emotion recognition from facial expressions using images with pose, illumination and age variations for human-computer/robot interaction. J. ICT Res. Appl. **12** (1), 14–34 (2018)
5. Keshari, T., Palaniswamy, S.: Emotion recognition using feature-level fusion of facial expressions and body gestures. In: 2019 4th International Conference on Communication and Electronics Systems, Coimbatore, TamilNadu, India, pp. 1184–1189 (2019)
6. Akkaya, I., et al.: Solving Rubik's cube with a robot hand. arXiv preprint arXiv:1910.07113 (2019)
7. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop (2015)

8. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016), pp. 3637–3645. Curran Associates Inc., Red Hook (2016)
9. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), pp. 4080–4090. Curran Associates Inc., Red Hook (2017)
10. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: relation network for few-shot learning. In: Computer Vision and Pattern Recognition (CVPR), pp. 1199–1208 (2018)
11. Allen, K.R., Shelhamer, E., Shin, H., Tenenbaum, J.B.: Infinite mixture prototypes for few-shot learning. arXiv preprint arXiv:1902.04552 (2019)
12. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: Meta-learning with memory-augmented neural networks. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning, ICML 2016, vol. 48, pp. 1842–1850. JMLR.org (2016)
13. Munkhdalai, T., Yu, H.: Meta networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, vol. 70, pp. 2554–2563. JMLR.org (2017)
14. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, vol. 70, pp. 1126–1135. JMLR.org (2017)
15. Kim, J., et al.: Auto-meta: automated gradient based meta learner search. arXiv preprint arXiv:1806.06927 (2018)
16. Li, Z., Zhou, F., Chen, F., Li, H.: Meta-SGD: learning to learn quickly for few-shot learning. arXiv preprint arXiv:1707.09835 (2017)
17. Behl, H.S., Baydin, A.G., Torr, P.H.S.: Alpha MAML: adaptive model-agnostic meta-learning. arXiv preprint arXiv:1905.07435 (2019)
18. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: Advances in Neural Information Processing Systems, pp. 737–744 (1994)
19. Wu, H., Xu, Z., Zhang, J., Yan, W., Ma, X.: Face recognition based on convolution siamese networks. In: 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, pp. 1–5 (2017) https://doi.org/10.1109/CISP-BMEI.2017.8302003
20. Bukovčiková, Z., Sopiak, D., Oravec, M., Pavlovičová, J.: Face verification using convolutional neural networks with Siamese architecture. In: International Symposium ELMAR, Zadar, pp. 205–208 (2017). https://doi.org/10.23919/ELMAR.2017.8124469
21. Hao, H., Pei, M., Zhao, M.: Face liveness detection based on client identity using siamese network. In: Lin, Z., et al. (eds.) PRCV 2019. LNCS, vol. 11857, pp. 172–180. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31654-9_15