Domain Adaptation for Emotion Detection from Face Expressions

This is the project report for course Deep learning 2020 offered at Information Technology University Lahore supervised by Dr. Mohsen Ali.

Muhammad Suleman Khan
msds19011@itu.edu.pk

Khaqan Ashraf
msds19019@itu.edu.pk

Muhammad Ahmad
msds19023@itu.edu.pk

Muhammad Khubaib Raza
msds19064@itu.edu.pk

Abstract

Humans have seven distinct facial emotions. Facial expression recognition algorithms have applications in healthcare, entertainment, criminal justice and more. Deep learning algorithms are efficient for facial expression classification but these algorithms demand high amount of data. Domain Adaptation can be used to address the lack of sufficient data. Right Now, we don't have much data of Pakistani facial expressions. In this Project, We created data set of Pakistani facial expressions and used domain adaptation to developed efficient facial expression algorithm of Pakistani faces. We have achieved 58% accuracy with baseline 32 %.

1. Introduction

The analysis of human facial behavior is a very complex and challenging problem. Understanding the mental satisfaction of people is important for intelligent agents and robots to operate be working within social environments. Making the chat bots able to detect the satisfaction of user while communicating. In other words, feeding facial expression along with text adds more information for a ML model to process as human beings do. The communication achieved by text only comprises of less than 50% of the total information which we want to communicate. So, a facial expression recognition mechanism must be implanted in text manipulation algorithms to generate much more effective results.

Identifying human expression is named as Emotion Recognition. Humans have a capability to express emotions like sadness , anger , happiness etc. So learning these expressions would be helpful in a bundle of applications. Human emotions can be detected using multiple modalities and combination of multiple modalities works best. Facial emotion recognition is an active area of research. It has interesting applications in education, the technology would be helpful for teachers in teaching like measuring the level of students learning, as well as monitor the student attitude in specific class and students assessments, As we know face expression is most important medium in communication of humans, If the machine can identify the human expressions of the face and really understand the emotion of humans. It will improve the efficiency of Human Computer Interaction(HCI) and be important in the healthcare system as well.

In many computer vision tasks, there is a huge amount of data with ground truth labels for training besides that face expression carrying character differences in gender age culture factors , illumination conditions viewpoint of camera occlusion and pose variation of persons that influence the intensity of expressions. Transfer learning and Domain adaptation are similar techniques. In transfer learning we change the target labels but, In Domain adaptation input source and target distribution different but same target class. For example we want models to work on Facial expression analysis of Pakistan people. For that we will need a sample of images from all our the Pakistan but domain adaptation has a solution for this problem. We can use a model trained on Western images and map to Pakistan images, so that we reduce the distribution of both dataset.

Domain Adaptation methods are based on deep learning and enable end-to-end training. Learning model leads us to domain in-variance by learning more general features from both source and target domains. Domain adaptation can be categorized into three approaches known as Discrepancy-based, adversarial based and other techniques which minimizes reconstruction error. Discrepancy-based techniques learns the domain invariant representations while adversarial based techniques utilize the minimization maximization technique. These techniques introduce domain discriminator.

In this paper we present the extension of the self-ensembling method for visual domain adaptation[3] that helps to recognize the facial expression. They proposed an approach to deal with semi supervised problem but our main

goal is unsupervised domain adaptation.

The aim of this work is to classify the object with superior performance over the generic model learned using training data. The datasets that we are using for evaluation : Fer2013, Ck dataset and our own scraped dataset from YouTube videos as target dataset.

2. Related Work

In Literature tackle the problem of domain adaptation, here we mainly focus on the deep learning based methods that are relevant to our proposed solution. Generative Adversarial Network (GANs) are deep unsupervised generative models that composed of two networks one is generator network that will be trained and generate images samples that matches the distribution of training data and second network is discriminator network that will differentiate the samples from real and fake samples that generated by generator. Many of the GANs based approaches to learn domain invariant embeddings. So, that minimizes the distribution of target and source domain using generator network As [2] adapts GANs based approach for source uses synthetic images and in target uses the real images so that matches the distribution of target and source domain. Generator takes the sample image and noisy vector and returns new adapted images. After that train the classifier on that images to predict classes In the same way another approach Bi-directional GANs [1] uses two generators that maps source domain to target domain and target domain to source domain, So that distribution between source and target domain is minimized.

Our work is based on mean teacher architecture [5], As we present our modification that purely works on domain adaptation. The internal structure of the model is consist of supervised algorithm as we choose VGG-B[4] like convolution neural network that extract basic emotion from a single frame. One is a student network and the other is a copy of student network called teacher. One more thing in teacher network weights change through student weights. That is called weighting moving average. At first pass target image and source image to stochastic augmentation after that source pass to the children and the other two target pass through the student and teacher network. The total loss is then calculated as the sum of unsupervised and supervised components at the time of test not passing through the augmentation block. Furthermore, It is not rely on the generative networks that generate the target samples and it takes time. So we make it more efficient with time-efficient and also applicable for real-time learning

3. Data

The databases used in this project are Extended Cohn Kanade(CK) , Facial Expression Recognition (FER2013)and Pakistani dramas and Talk shows (PDTS) . Ck and FER2013 are used in many deep learning approaches and performed experiments on it. CK includes 123 subjects that have 593 images and Fer2013 data set contains 35,887 grey scale images with 7 basic emotions. Both data set are used as a source domain and for target domain PDTS dataset scraped from YouTube videos using pytube python package, extracted frames and detected faces suing opencv library and annotated them manually name as PDTS. All four data sets are changed into 48 X 48 single channel image. All group members equally labeled the dataset. PDTS(G2G) dataset is labeled by group G2G and dataset PDTS(G3G) is labeled by us. We performed multiple experiments in different setting which are explained in experiments section.

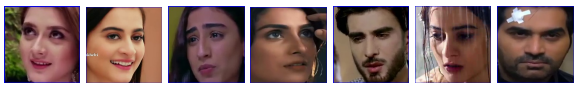

Figure 1. Some example images of Pakistani Dramas and talk shows data set

Figure 2. Some example images of Western Data set

4. Methods

In our entire procedure, we adopted different methods and techniques to reach our goal. Discussing the techniques:

4.1. Mean Teacher

In order to learn the key features of data, Computer Vision usually takes a very large number of learn able parameters which tends towards over-fitting and high computation cost. If a slight change is made in a picture, the human still consider it as the previous picture through which it was generated but the computer vision algorithms consider them a separate example. To make computer minimize difference between these two images, new samples are generated by adding a little noise in true data and this difference is minimized.

Keeping in mind the previous strategy, our model exhibit two roles i.e. teacher and student. As a student, it learns the data points. After that, as a Teacher, it generates new sample points by adding a little noise and use these new sample points for learning purpose.

The student model is continuously updating itself through new sample points, i.e. its mean of the previous contiguous student models hence known as Mean Teacher Method.

Dataset	Angry	Fear	Happy	Neutral	Sad	Surprise	Disgust
CK	135	75	207	10	84	249	177
FER2013	4593	5121	8989	6198	6077	4002	547
PDTS(G2G)	90	122	950	1300	309	339	28
PDTS(G3G)	129	85	276	271	332	106	160

Table 1. We mainly used three datasets CK, FER2013 and Pakistan Drama and Talk shows(PDTS). PDTS(G2G) is tagged by group G2G and dataset PDTS(G3G) is tagged by us.

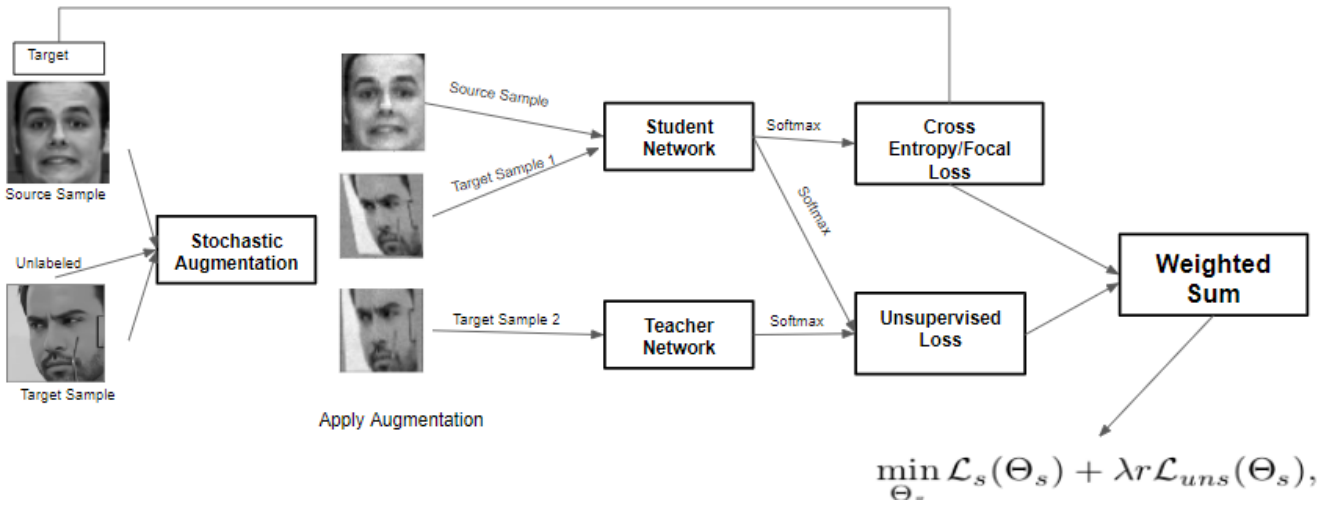

Figure 3. Source and target samples pass to the training, Apply stochastic augmentation on each image separately, Passed augmented images to student and teacher architecture. Apply cross entropy or loss function on student and unsupervised loss on teacher and student ensemble. computed combined weighted sum.

4.2. Self-ensembling

Keeping the Mean Teacher method in mind, we moved further by making a little modification in our model as depicted in diagram.See Figure 3

In Mean Teacher method, the input for the teacher and student model was same but know, an unlabeled image is passed to the teacher along with a labeled one to match the predictions of the teacher and to train the student.

4.3. Domain Adaption

Now coming to the technique used by us. In order to make the model to learn only the facial expression and not to learn the faces themselves, we were in need of a dataset which compose of different people (different from the training set). For that, we produce our own dataset by using Pakistani Dramas as source. All of the classes were well filled and a clean and precise dataset was formed.

After the completion of dataset, the FER and CK datasets were used for student for learning purpose and our own dataset was used as teacher input in order to make the model to learn only facial expressions but not the pictures or faces.

The output of the Student model is passed to cross entropy loss as it is labeled sample. Whereas, the outputs of student and teacher are subtracted and squared. Now these two quantities are add after some weights are multiplied to each.

5. Experiments and Results

We used some code from this repository available at GitHub.[1] and our code is available at GitHub.[2345] The results can be seen in Table 2. First we trained FER2013 dataset that contains 48x48 grey-scale images belonging to 7 classes of emotions and trained VGG-B architecture another data set PDTS that contains 7 emotions of classes as well. All images passed to the trained model and got 15% accuracy. Another experiment we have done, uses pre-trained model FER2013 and after that uses Fer2013 as Source domain and PDTS as Target domain. But this time we saw improvement in results. That's a good sign for us as we see in the table. Number of examples in one class is less

[1] https://github.com/Britefury/self-ensemble-visual-domain-adapt

[2] https://github.com/msulemannkhan/msds19011_Project_DLSpring2020

[3] https://github.com/khaqanashraf/msds19019_Project_DLSpring2020

[4] https://github.com/doubleacbt/msds19023_Project_DLSpring2020

[5] https://github.com/imkhubaibraza/msds19064_Project_DLSpring2020

Classes	Source Dataset	Target Dataset	Loss	Error	Accuracy	Pretrained
7	FER2013	–	1.0844	–	23%	No
7	Fer2013	PDTS(G3G)	1.095	71.3%	28%	No
7	FER2013	PDTS(G2G)	1.407	55%	40%	Yes
6	CK	PDTS(G2G)	1.38	75.3%	18%	No
6	FER2013	–	0.80687	–	31%	No
6	FER2013	PDTS(G3G)	0.9832	45%	53.21%	Yes
6	FER2013	PDTS(G2G)	—	–	43%	Yes
5	FER2013	–	0.8687	–	32%	No
5	FER2013	PDTS(G2G)	0.844	46%	51%	Yes
5	FER2013	PDTS(G3G)	1.092	62%	41.12%	No
5	FER2013	PDTS(G2G)	0.7068	42%	58.12%	Yes

Table 2. We performed multiple experiments on FER2013, CK as source dataset and used PDTS as target dataset.

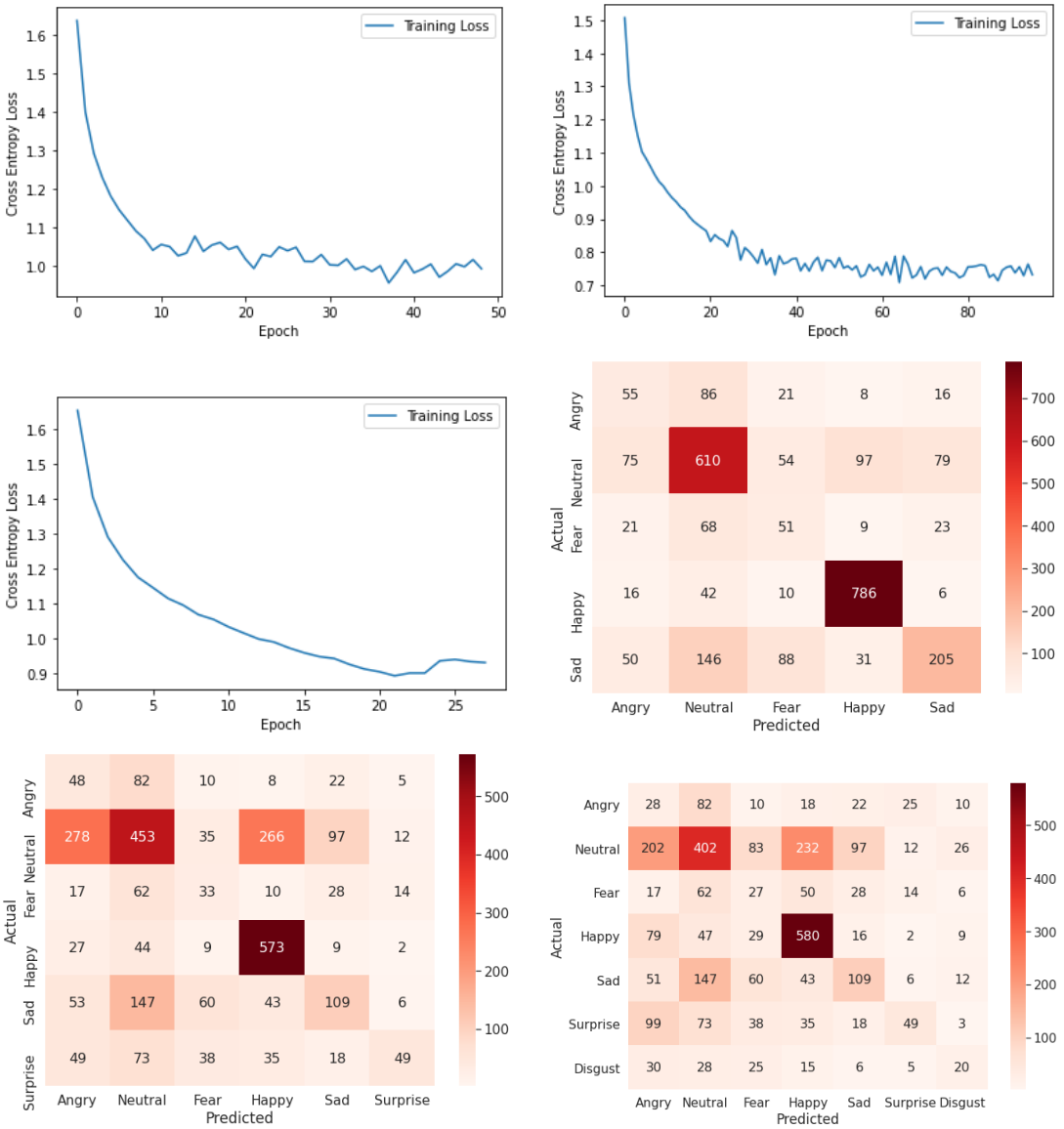

Figure 4. second figure is loss and in confusion matrix first one is 5 class expression and other two have 6 and 7 expressions

so we try focal loss on it. Focal loss handles imbalanced classes efficiently. One more thing that we observed, when we change cross entropy loss with focal loss, accuracy is not much improved because True positive classes with high classes send to other classes. Running different hyper parameters. One solution is, we selected 6 classes and remove disgust that is causing the problem and another data set use CK with 6 class but the number of examples Trained our model on 6 classes same as the pre-trained FER2013 model on 6 classes, Source domain Fer2013 and Target domain PDTS but now we are getting much higher accuracy as compared to 7 classes. In our data set PDTS examples points are less. So we again remove 1 class fear that causes a problem. First we trained without any pre-trained model and compared with baseline accuracy. The last experiment we have done with focal loss and pre-trained model on FER2013. In results we get best accuracy on 7 class , 6 class and 5 class respectively as mentioned in table Plot the error% and loss curves of three best models Figure 2

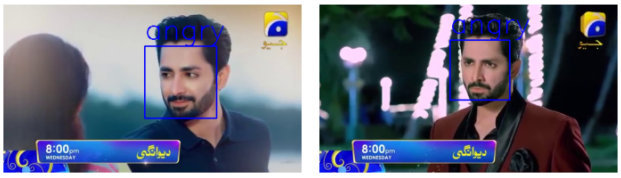

Figure 5. Labels

6. Conclusion

We have been presented an domain adaption model extended from mean teacher a role model. The resulting network achieved 58 % accuracy with baseline 32 %. We used Binary cross entropy loss and Focal loss use for class imbalance. We observed that FER2013 data set is given good results as source data set because it has more variations in images and pose. We extract our own data set using pytube python package and extract frames after that faces from images. We performed different experiment as we can see in Table2 Facial expressions can be extended to video.Th aim of this project to show that Domain adaptation is worked fine with source domain and target domain, It solved the problem of small data set in target domain.

References

[1] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[2] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016.

[3] G. French, M. Mackiewicz, and M. Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

[4] C. Pramerdorfer and M. Kampel. Facial expression recognition using convolutional neural networks: state of the art. *arXiv preprint arXiv:1612.02903*, 2016.

[5] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.