REPRODUCTION OF MLCR

Jing Chi, Zeyu Chen & Mengting Liu
School of Electronics and Computer Science
University of Southampton
{jc1r19,zc4y19,ml2u18}@soton.ac.uk

1 INTRODUCTION

We chose 'Multi-label Co-regularization (MLCR) for Semi-supervised Facial Action Unit Recognition' (Niu et al., 2019) from NeuroIPS 2019 to complete our reproducibility challenge. This paper mainly aims at the problem that the existing Action Units (AUs) recognition method requires a large amount of data with accurate AU labels but annotating AU manually is a tedious and time-consuming work. Hence this paper proposed a novel semi-supervised learning method for AU recognition, which is called multi-label co-regularization. Only a small part of the data in its dataset has the label obtained through co-training, while more and larger part of the data is unlabeled. The key points of this method are as follows, first it generates multi-view features from two different views, including labeled and unlabeled data. In order to ensure the independence of the features generated by different perspectives, a multi-view loss function is introduced. And then they proposed a multi-label co-regularization loss function to minimize the difference in the predicted distribution of the two views. In the end, in order to fully apply the previous experience, that is, the relationship between individual AUs, the author uses a graph convolutional network (GCN) to develop as much useful information as possible for unlabeled data. The entire structure of this approach be presented in Figure 1.

In our reproducibility challenge, we reproduce the above method according to the description in the paper. In the beginning, we used the widely used and public CK + action units database (Lucey et al., 2010) to evaluate the method. And then we conducted an ablation learning test on the CK + dataset to further evaluate the effectiveness of the core components used in the article. Finally, we also carried out the general ability assessment of the method in CelebA database (Liu et al., 2015). The source code for all reproduction experiments is available on GitHub[1].

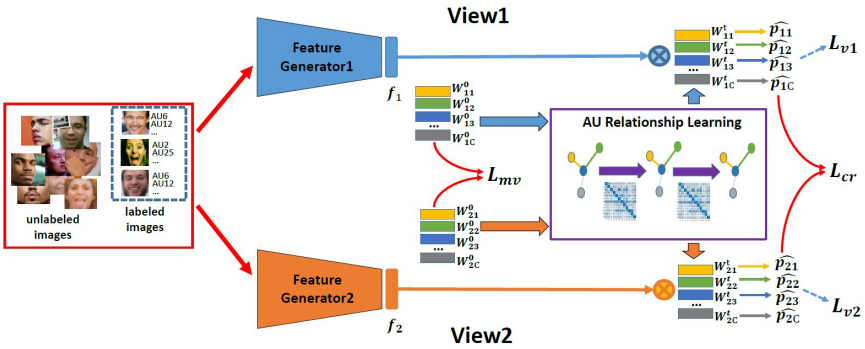

Figure 1: An overview of the proposed multi-label co-regularization method for semi-supervised AU recognition (Niu et al., 2019).

2 REPRODUCTION OF EXPERIMENTS RESULTS

2.1 DATABASE PREPROCESSING

We evaluate this method on the Extended Cohn-Kanade Dataset (CK+) which is a public and widely used AU database. CK+ dataset includes 593 expression sequences of 123 people. These expression sequences are from calm to peak expression. The peak expression is coded with FACS. Apart from these image sequences, this database also includes a list of FACS coded files. For each sequence

[1]https://github.com/COMP6248-Reproducability-Challenge/Multi-Label-Co-regularization

there is only one FACS file which recoeds the AU value and intensity of the last frame (the peak frame). A total of 64 different AU labels appear in the CK + dataset, but in order to be consistent with the author, we only choose 12 of them (1,2,4,5,6,9,12,17,20,25,26,43) to evaluate.

First, we extract the images containing these 12 AUs in CK + and classify them according to the labels. However, the amount of image data of different AU labels extracted from CK + varies greatly. For example, AU43 class only has 9 images, but AU25 class has 324 images. Therefore, in order to solve the huge number difference between different classes, we expanded the data for the classes with relatively small data volume. Specifically, we collected and labeled image data with different classes in constrained scenarios to expand the classes which have a small amount of data. In the end, there are about 200 data in each class. At the same time, we also randomly selected 2000 images from the famous face database LFW (Huang et al., 2007) as the unlabeled training set to help us train.

For all images, we use the SeetaFace detector like the author to perform preliminary face detection using five face landmarks. Then we perform face alignment and crop the face to 240×240. Finally, for the train set, we randomly cropped the corrected image to 224×224 and then started the train. For the test set, we center cropped the images from the initial aligned images and then used them for testing.

2.2 Training Datails

According to the method proposed in the paper, firstly, we need two feature generators to generate two-view features from both labeled and unlabeled data. As deep neural networks have been proven to be effectively in terms of feature extraction, we choose two ResNet-34 models as the feature generators, and the last fully connected layers of these two models are used to predict the probability of 12 AUs using the two-view features which is a common multi-label problem. The loss function of the proposed method is as follows:

$$L = \frac{1}{2}\sum_{i=1}^{2} L_{vi} + \lambda_{mv} L_{mv} + \lambda_{cr} L_{cr} \tag{1}$$

Here, L_{vi} is a binary cross-entropy loss used to calculate the losses for two-view features, L_{mv} is a multi-view loss utilized to ensure that two feature generators obtain conditional independent features instead of being completely similar, and L_{cr} is a co-regularization loss aiming to enforce the classifiers from two views to generate consistency predictions for labeled and unlabeled data. λ_{mv} and λ_{cr} are two hyper-parameters which are used to balance the influence of L_{mv} and L_{cr}. In the first step of training, we pre-train the two ResNet-34 models by setting $\lambda_{mv} = 400$ and $\lambda_{cr} = 100$ (same as the paper setting). Then, during the training process, we find that it takes about one or two epochs for the L_{cr} to be very small (a few tenths). Therefore, we adjust $\lambda_{cr} = 100$ to increase the impact of L_{cr} and the final value of λ_{cr} is 500. The effectiveness of L_{mv} and L_{cr} will be discussed in detail in Section 3.

Secondly, we take GCN into consideration. GCN is a key component of the multi-label co-regularization method which is utilized to extract useful information from the large amount of unlabeled data. After the pre-training of two ResNet-34 models, these two feature generators have been able to generate useful two-view features. In the second step of training, we jointly train the feature generators and GCN on both labeled and unlabeled data. In order to reduce the influence of overfitting, we plot the training curves (the plot of loss against epochs) as shown in Figure 2. From the figure, we find that epoch=30 is the optimal choice since after epoch=30, the loss for test set increases with the increase of epoch.

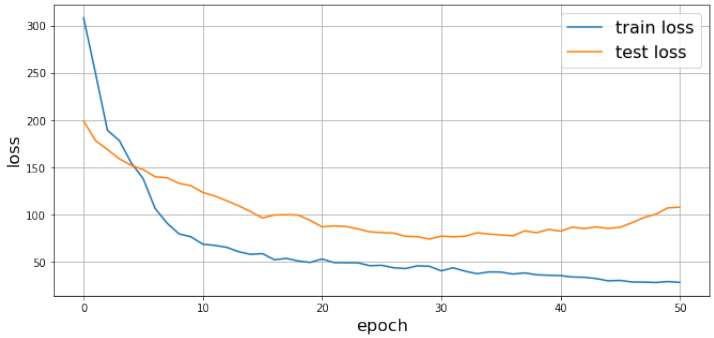

Figure 2: Training curves with both the training and test data.

2.3 Evaluation Experiments

For the evaluation experiments of the method, we choose F1 score as the metric since F1 score takes into account both the precision and recall of the classification model and it can intuitively measure the quality of a model. As the train and test sets are randomly split, we choose to run the model three times and calculate the average performance of the three tests to reduce the influence of the dataset bias. We record both F1 score for each AUs and average F1 score over all AUs in Table 1.

The baseline method represents two ResNet-34 models which only be trained on labeled data. From the results of Table 1, we can see that the multi-label co-regularization method outperforms baseline method as this kind of semi-supervised methods can extract useful information from a large amount of unlabeled data. It is worth mentioning that the method performs significantly better on AU6, AU12 and AU25 than on other AUs. These three AUs represent the facial muscle movement in three large regions, which are Cheek Raiser, Nose Wrinkler and Lips Part respectively, which makes it helpful for the recognition of these AUs.

Table 1: F1 score (in %) for 12 AUs recognition on CK+ database.

Method \ AUs	1	2	4	5	6	9	12	17	20	25	26	43	Avg.
Baseline	45.7	33.5	58.1	37.2	69.8	48.7	76.2	36.8	36.1	79.3	46.2	51.3	51.6
Proposed	**51.7**	**42.5**	**63.4**	**45.1**	**72.9**	**57.3**	**78.4**	**41.9**	**45.2**	**80.3**	**53.1**	**58.3**	**57.5**

3 Reproduction of Ablation Study

As in the original article, we also conduct ablation study, which is to remove certain components from the model, to study the contribution of each part of the model to the result. The key innovation of this model is the design of loss function, that is, the addition of the co-regularization loss L_{cr} and the multi-view loss L_{mv}. Also, the Graph Convolutional Network (GCN) is added to capture the relationship between different AU labels, so the ablation study will also start from these two components.

The specific method is to use ResNet-34 trained with labeled images without adding L_{mv}, L_{cr} and CGN as the baseline, then add L_{cr} and observe the results, then add L_{mv}, and finally add CGN to achieve proposed model. For four different models, the results are shown in Table 2.

Table 2: F1 score (in %) in the ablation study on CK+ database.

Method \ AUs	1	2	4	5	6	9	12	17	20	25	26	43	Avg.
Baseline	45.7	33.5	58.1	37.2	69.8	48.7	76.2	36.8	36.1	79.3	46.2	51.3	51.6
Baseline + L_{cr}	50.9	39.3	61.9	44.1	70.9	55.6	77.8	40.8	43.5	79.7	52.6	56.7	56.2
Baseline + L_{cr} + L_{mv}	50.6	41.2	62.0	44.2	71.3	56.0	77.5	41.4	44.6	79.9	52.9	57.1	56.6
Baseline + L_{cr} + L_{mv} + CGN	51.7	42.5	63.4	45.1	72.9	57.3	78.4	41.9	45.2	80.3	53.1	58.3	57.5

3.1 Effectiveness of L_{mv} and L_{cr}

It can be seen from the results in Table 2 that the addition of these three components has brought benefits to the model, among which the addition of L_{cr} is the key. The average F1 scores increase from 51.6% to 56.2% due to the addition of L_{cr}, which is a great improvement. Besides, the addition of L_{mv} also contribute to the increase of the average F1 scores by nearly 0.4%.

This is consistent with the results in the original paper. In theory, since L_{cr} measures the difference of the two distribution predicted based on the two views, it is reasonable that L_{cr} plays a distinct role. For L_{mv}, it measures the similarity of the two features extracted from the images, and it can extract information from the unlabeled data, so its contribution in semi-supervised learning is also obvious.

3.2 Effectiveness of GCN

As for the role of GCN, on the whole, the addition of GCN makes the average F1 score increase from 56.6% to 57.5%. There is some improvement. As for the analysis in the original paper, the

improvement that GCN leads to is mainly reflected in some F1 score of AU labels which have the association with other AU labels, for example, AU4 (Brow Lowerer) and AU9 (Nose Wrinkler). This can also be seen from table 2: because of the addition of GCN, the F1 score of AU4 and AU9 increase 1.4% (from 62.0% to 63.4%) and 1.3% (from 56.0% to 57.3%) respectively.

4 Reproduction of Generalization Ability

We choose the same dataset —— CelebA (Liu et al., 2015), as in the original paper, as training material to evaluate the generalization capabilities of the model. In more detail, each image in this dataset has 40 binary tags. As the number of samples in this dataset is more than 200,000, a semi-supervised learning multi-label classification task, similar to Facial Action Unit Recognition, can be easily constructed by selecting some samples with their tags as labeled images and some samples removed their tags as unlabeled images. We randomly select 30,000 labeled images and 100,000 unlabeled images as a training set, and 10,000 labeled images are selected randomly as a test set. Our baseline model is ResNet-34, and its training data is only 30,000 labeled images and no unlabeled images.

The F1 scores of the two models on different tags are shown in the following Figure 3. It is obvious that the proposed model for processing semi-supervised learning achieves more or less higher F1 scores on each tag, compared with the ResNet-34 for processing supervised learning. At this point, our reproduction results are consistent with the original paper.

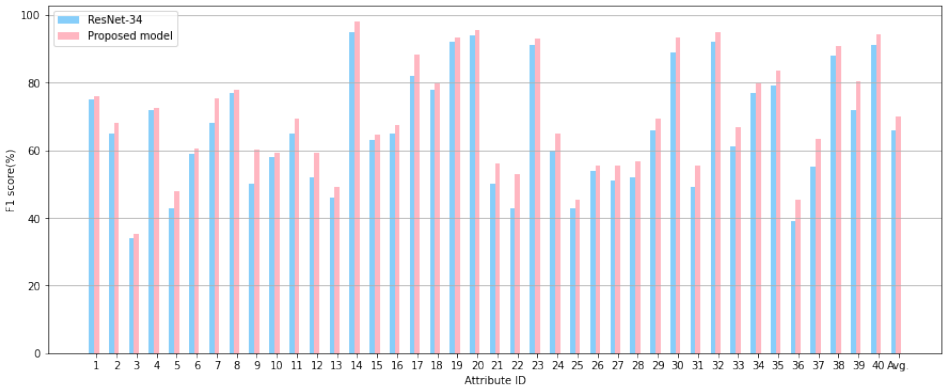

Figure 3: F1 score (in %) for the 40 attributes in CelebA dataset.

5 Conclusion

Through this reproduction task, we believe that the model proposed in the original paper is indeed valuable in the Facial Action Unit Recognition task. In addition, this model can be appropriately used in other semi-supervised learning tasks. In the field of machine learning, labeling often requires higher costs, so the study of semi-supervised learning is very cost-effective. We learned much in this task, and it also provides some thoughts for our subsequent study.

References

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101. IEEE, 2010.

Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. In *Advances in Neural Information Processing Systems*, pp. 907–917, 2019.