Hey Human, If your Facial Emotions are Uncertain, You Should Use Bayesian Neural Networks!

Maryam Matin[1] and Matias Valdenegro-Toro[2]

[1] Hochschule Bonn-Rhein-Sieg, 53757 Sankt Augustin, Germany
maryam.matin.1987@gmail.com
[2] German Research Center for Artificial Intelligence, 28359 Bremen, Germany
matias.valdenegro@dfki.de

Abstract. Facial emotion recognition is the task to classify human emotions in face images. It is a difficult task due to high aleatoric uncertainty and visual ambiguity. A large part of the literature aims to show progress by increasing accuracy on this task, but this ignores the inherent uncertainty and ambiguity in the task. In this paper we show that Bayesian Neural Networks, as approximated using MC-Dropout, MC-DropConnect, or an Ensemble, are able to model the aleatoric uncertainty in facial emotion recognition, and produce output probabilities that are closer to what a human expects. We also show that calibration metrics show strange behaviors for this task, due to the multiple classes that can be considered correct, which motivates future work. We believe our work will motivate other researchers to move away from Classical and into Bayesian Neural Networks.

Keywords: Facial Emotion Recognition, Uncertainty Quantification, Bayesian Deep Learning

1 Introduction

Emotion recognition in facial images is the task of classifying the face of a person into a set of emotions. One important characteristic of this task is its high degree of aleatoric uncertainty [2], which presents itself as ambiguity in defining what is the correct emotion class given an image [22] [11]. Most state of the art neural networks used for this task do not model any kind of uncertainty, which makes them ill-posed for emotion recognition.

In this paper we evaluate three scalable methods for uncertainty quantification in neural networks, namely Monte Carlo Dropout/DropConnect, and Deep Ensembles, on the FER+ dataset [2] using three different neural network architectures. This dataset is a variation of the FER dataset [6] where each image is labeled with a crowd-sourced distribution of emotion classes, instead of a single class annotation per image. Our results show that Bayesian neural networks are better able to model this kind of problem, even as only one label is used during training.

We believe that our results show that metrics for this task need to be rethought, and that only methods able to model at least aleatoric uncertainty should be used for emotion recognition. It makes little sense to obtain high accuracy on this task, given the visual ambiguity and the multiple correct answers that are possible.

2 Related Work

There is a rich literature on emotion recognition from facial images [11]. The FER+ dataset [2] is one dataset used to evaluate progress in this task. Two defining characteristics of this dataset are being grayscale images at 64×64 resolution, and labels might indicate multiple emotions, as defined by a crowd-sourced probability distribution. Classes are 'neutral', 'happiness', 'surprise', 'sadness','anger', 'disgust', 'fear', and 'contempt'. The baseline reported in [2] is 84.7% accuracy with a custom VGG13 network and a standard cross-entropy loss and data augmentation from [23].

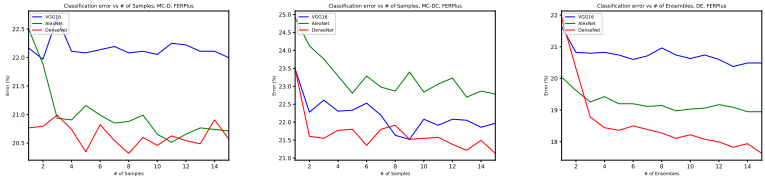
Georgescu et al. [5] use CNNs with bag of visual words to obtain 87.7% accuracy. Other baselines presented in this paper are 84.4% accuracy with VGG-face [16], a Bag of Visual Words alone obtaining 79.6%, and a large ensemble achieving 88% accuracy. Arriaga et al.[1] reports 78% and 81% accuracy with a reduced VGG and a mini-Xception network.

Overall most methods for facial emotion recognition use classical neural networks, and Bayesian neural networks are not commonly used, even more recent work that uses ensembles like Siqueira et al. [19] or Surace et al. [20] do not consider the possibility of modeling output uncertainty, despite Lakshminarayanan et al. [13] showing that ensembles are able to produce state of the art uncertainty quantification.

3 Experimental Methodology

For our experiments we use three of the most common CNN model architectures which have shown outstanding performance in image classification competition on ImageNet, namely AlexNet [12], VGG16 [18] and DenseNet121 [8] with some minor modifications. To simplify the models, we reduce the number of neurons in fully connected layers to 256 instead of 4096. For AlexNet, we add batch normalization [9] after each layer. DenseNet-121 is modified to integrate dropout layers into the architecture.

We use the dropout/drop rate of 0.5 for AlexNet and 0.2 for the other two models as suggested by their original implementations. The batch size is set to 32 and we use categorical cross-entropy loss and accuracy metric with Adam optimizer [10]. Note that most implementations of the cross-entropy loss use only a single label per class, even as more labels might be available, so in this work we do not explore the use of soft labels [17]. We decided to only tune the learning rate in range 10^{-1} to 10^{-4}. For VGG16 models we used Stochastic Gradient Descent instead of Adam. For SGD optimizer we tuned the learning rate decay



(a) MC-D (b) MC-DC (c) DE

Fig. 1: Classification error as a function of # of samples/ensembles for all methods in different models on FERPlus dataset.

in range 10^{-1} to 10^{-6}. The actual training after hyper-parameter tuning is done over 80 epochs. We do not perform any kind of data augmentation.

Since full inference in a Bayesian neural network is intractable, we use approximate methods. Due to their scalability and simplicity, we use Monte Carlo Dropout [4], Monte Carlo DropConnect [14], and Deep Ensembles [13]. These methods all have a hyper-parameter in common, the number of stochastic forward passes T for Monte Carlo methods, and the number of ensembles N for Deep Ensembles. Note that while MC Dropout/DropConnect are approximations to a BNN, Deep Ensembles is a non-Bayesian method but it outperforms other methods in uncertainty quantification and out of distribution detection [15].

We evaluate three metrics on the FER+ dataset [2]. We compute classification error, Negative Log Likelihood (NLL), and expected calibration error (ECE) [7], all as a function of number of stochastic forward samples/ensembles, which are varied between 1 to 15. NLL determines whether the network is assigning high confidence to correct classes, and calibration determines if its confidence estimates are compatible with the true likelihood of the data.

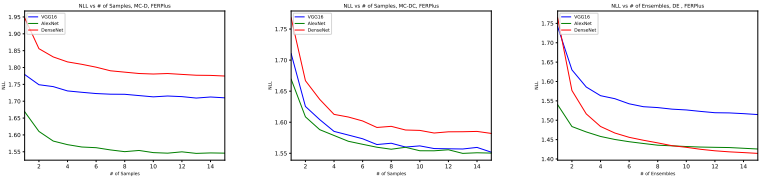
4 Experimental Results and Analysis

Our main results are presented in Figure 1 for classification error, Figure 2 for the negative log-likelihood, and Figure 3 for expected calibration error [7].

The effect on task performance (accuracy/error) is as expected, with overall decreasing error for all uncertainty methods, but this is more pronounced with an ensemble of DenseNets, which is also confirmed with the plots of negative log-likelihood. There are large variations in performance across different models, and MC-Dropout and MC-DropConnect seem to be less stable than ensembles.

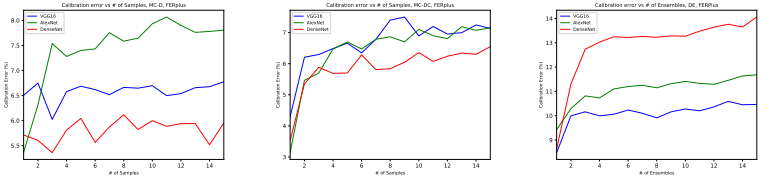
Calibration error shown in Fig. 3 shows an unusual pattern for all models and uncertainty methods, as the calibration error increases with more samples or ensemble members, instead of decreasing as it does with other datasets (like CIFAR10 and SVHN, as shown by Valdenegro-Toro [21]). We interpret these results as that the model's probabilities are closer to represent the true label distribution than the classical network (which can also be seen in Fig. 4).

We believe that our Bayesian neural network models are overall underconfident, which might be undesirable, but this is due to the large aleatoric uncer-



(a) MC-D (b) MC-DC (c) DE

Fig. 2: NLL as a function of # of samples/ensembles for all methods in different models on FERPlus dataset



(a) MC-D (b) MC-DC (c) DE

Fig. 3: Calibration error as a function of # of samples/ensembles for all methods in different models on FERPlus dataset

tainty in the labels and input images (which can be validated by looking at the label entropy), not because the models are producing incorrect predictions. The calibration error considers both the correct class and the prediction confidence of that class, but this considers only one correct class per sample, there are no calibration metrics that consider cases of high aleatoric uncertainty, where some classes are visually similar and should be allowed for the model to be confused.

We visualize the top five most uncertain images as computed using entropy of the output probabilities for the DenseNet model using a Deep Ensemble. This is shown in Figure 4. These results show the uncertainty and visual ambiguity between the classes. An ensemble with a single model is equivalent to a classical neural network, and overall it produces a correct but overconfident result. A Deep Ensemble produces probabilities that are more spread across classes, which make more sense for face images with visual ambiguity in terms of which emotion is actually conveyed.

5 Conclusions and Future Work

Overall we see multiple benefits from using BNNs over a classical neural network for facial emotion recognition in the FER+ dataset. A BNN is able to model aleatoric and epistemic uncertainty, while providing small improvements in accuracy (in the order of 2-4%) compared to a classical network, and providing more realistic probability estimates, specially when considering overconfident point predictions made by classical networks [7].

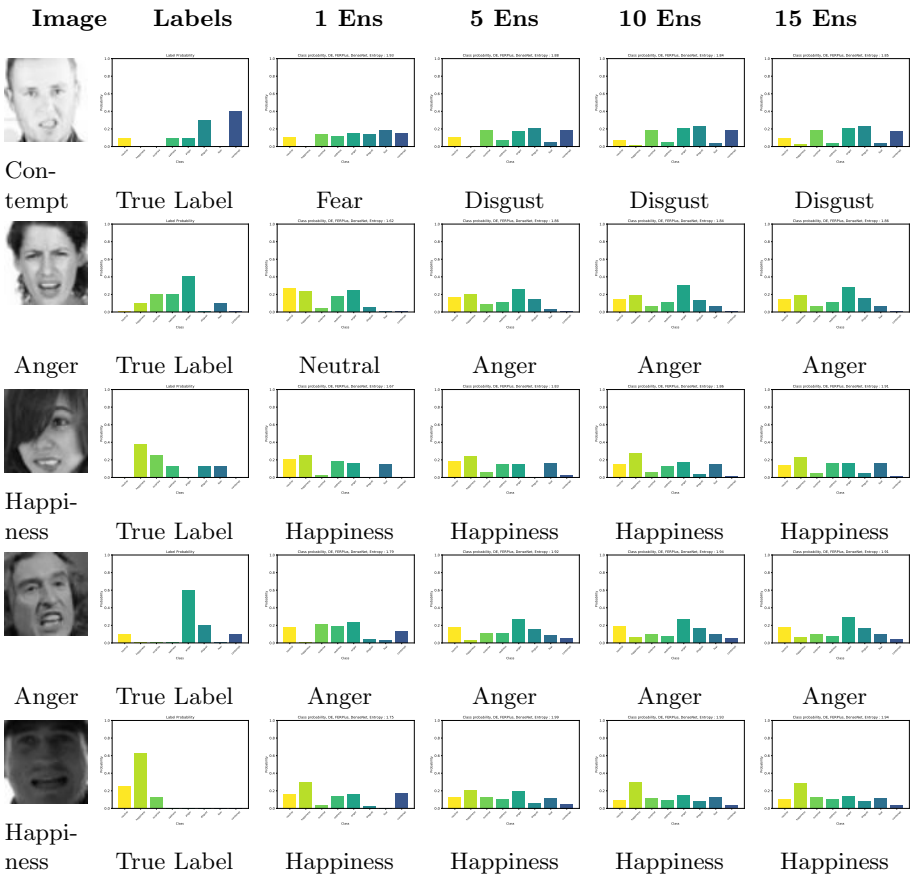


Fig. 4: Five most uncertain images based on DenseNet model and Deep Ensembles with # of ensembles and a plot of predictive probabilities using 1, 5, 10 and 15 ensembles. The first column represents the image, and the second its ground truth label distribution. Under each probability plot, the predicted class is presented.

We also find that usual calibration metrics behave strangely in the presence of high aleatoric uncertainty, with calibration error increasing along with number of samples or ensembles, while in other datasets it generally decreases producing a more calibrated model.

For future work, we wish to evaluate the potential of out of distribution detection based on probability entropy, as a way to detect biases in the model, and prevent wrong predictions to be made in out of distribution settings, which is certainly a concern for skin shades that are far away from the training set [3].

Finally, we wish to explore ways to disentangle aleatoric and epistemic uncertainty, and to train BNNs using other losses that are able to fully utilize the soft labels [17] in the FER+ dataset.

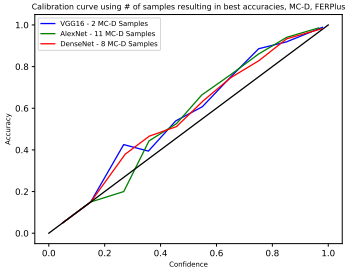
References

1. Arriaga, O., Valdenegro-Toro, M., Plöger, P.G.: Real-time convolutional neural networks for emotion and gender classification. In: European Symposium on Artificial Neural Networks (2019)
2. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM (2016)
3. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91 (2018)
4. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059 (2016)
5. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. arXiv preprint arXiv:1804.10892 (2018)
6. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: International Conference on Neural Information Processing. Springer (2013)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1321–1330. JMLR. org (2017)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
9. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (2015)
10. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Ko, B.C.: A brief review of facial emotion recognition based on visual information. sensors **18**(2), 401 (2018)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
13. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. pp. 6402–6413 (2017)
14. Mobiny, A., Nguyen, H.V., Moulik, S., Garg, N., Wu, C.C.: Dropconnect is effective in modeling uncertainty of bayesian deep networks. arXiv preprint arXiv:1906.04569 (2019)
15. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems. pp. 13991–14002 (2019)
16. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)

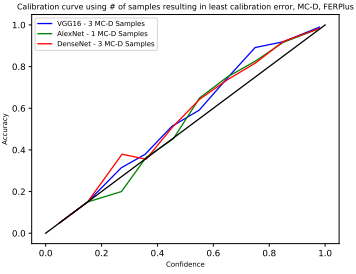
17. Peterson, J.C., Battleday, R.M., Griffiths, T.L., Russakovsky, O.: Human uncertainty makes classification more robust. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9617–9626 (2019)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Siqueira, H., Magg, S., Wermter, S.: Efficient facial feature learning with wide ensemble-based convolutional neural networks. arXiv preprint arXiv:2001.06338 (2020)
20. Surace, L., Patacchiola, M., Battini Sönmez, E., Spataro, W., Cangelosi, A.: Emotion recognition in the wild using deep neural networks and bayesian classifiers. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. pp. 593–597 (2017)
21. Valdenegro-Toro, M.: Deep Sub-ensembles for Fast Uncertainty Estimation in Image Classification. In: NeurIPS Workshop on Bayesian Deep Learning (2019)
22. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6897–6906 (2020)
23. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 435–442 (2015)

A Additional Results on Calibration

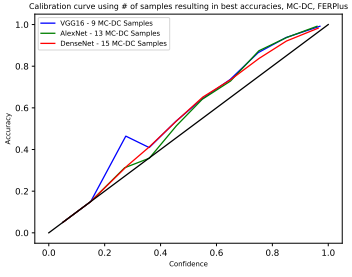
This section presents calibration plots for each model and uncertainty method combination, as they did not fit in the main paper.



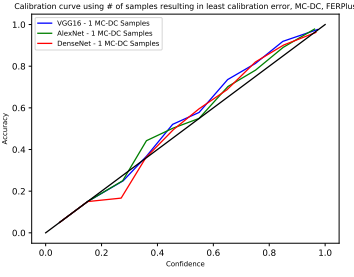
(a) MC-D, # of samples based on best classification accuracy



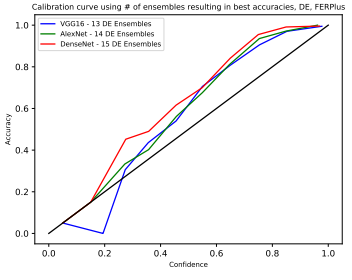
(b) MC-D, # of samples based on best calibration error



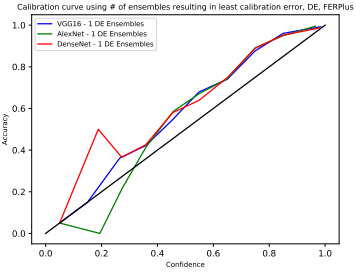
(c) MC-DC, # of samples based on best classification accuracy



(d) MC-DC, # of samples based on best calibration error



(e) DE, # of ensembles based on best classification accuracy



(f) DE, # of ensembles based on best calibration error

Fig. 5: Calibration curve for all methods in different models on FERPlus dataset using # of samples resulting in best classification accuracy (left) or least calibration error (right).

B Additional Results on Most Uncertain Images

This section presents additional probability plots for Deep Ensembles across different models. One important conclusion that can be drawn from these plots is that the number of ensembles has a big influence on the output probabilities, and for task with high uncertainty such as facial emotion recognition, this can lead to very different class predictions as computed by taking the maximum probability.

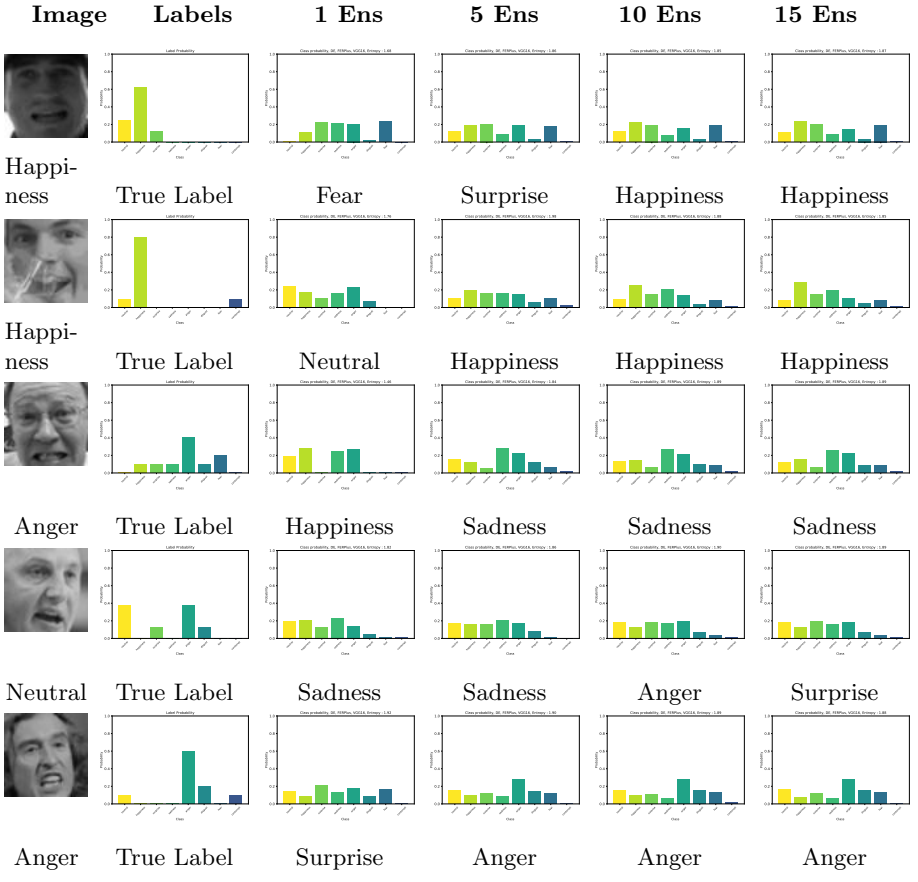


Fig. 6: Five most uncertain images based on VGG model and Deep Ensembles with # of ensembles and a plot of predictive probabilities using 1, 5, 10 and 15 ensembles. The first column represents the image, and the second its ground truth label distribution. Under each probability plot, the predicted class is presented.

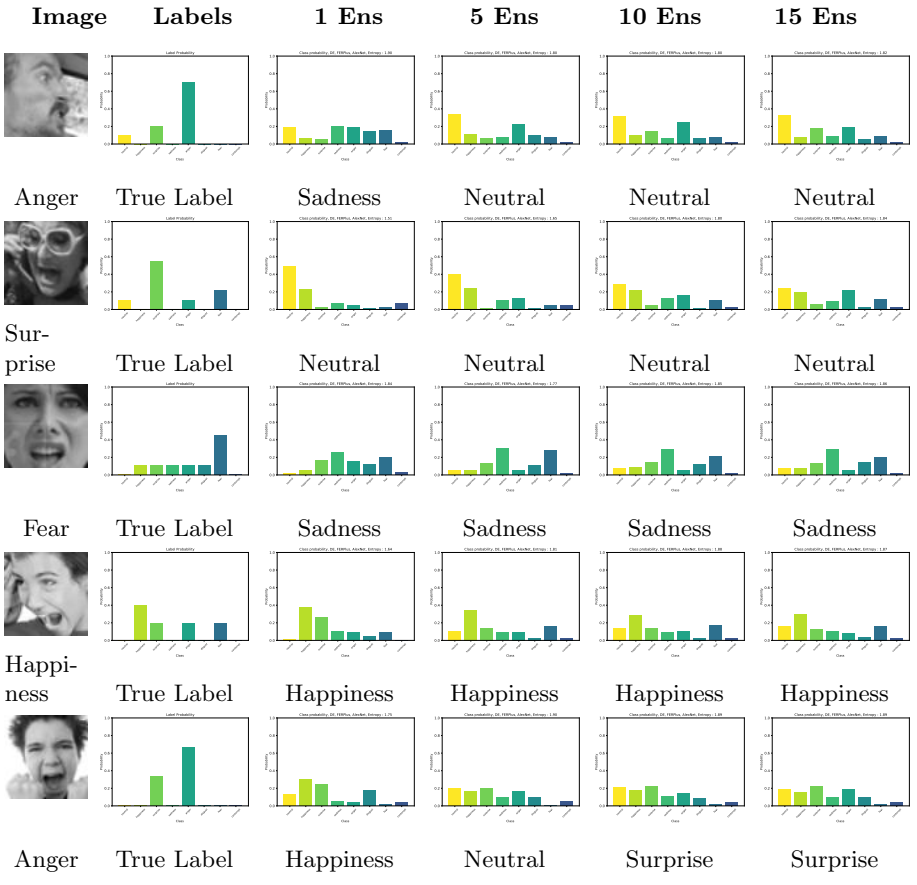


Fig. 7: Five most uncertain images based on AlexNet model and Deep Ensembles with # of ensembles and a plot of predictive probabilities using 1, 5, 10 and 15 ensembles. The first column represents the image, and the second its ground truth label distribution. Under each probability plot, the predicted class is presented.