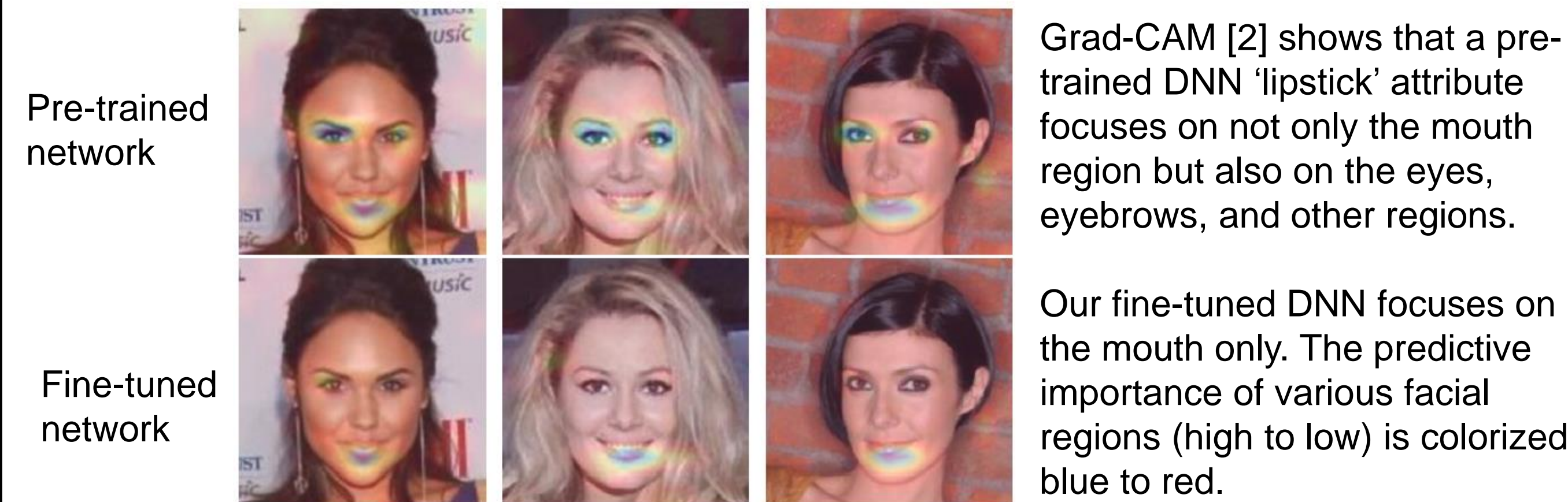


Abstract

Deep neural networks (DNNs) have a high accuracy on image classification tasks. However, DNNs trained by such dataset with co-occurrence bias may rely on wrong features while making decisions for classification. It will greatly affect the transferability of pre-trained DNNs. In this paper, we propose an interactive method to direct classifiers paying attentions to the regions that are manually specified by the users, in order to mitigate the influence of co-occurrence bias. We test on CelebA dataset, the pre-trained AlexNet is fine-tuned to focus on the specific facial attributes based on the results of Grad-CAM.

Introduction



In Large-scale CelebFaces Attributes (CelebA) dataset, for example, the attributes 'Wearing Lipstick' and 'Heavy Makeup' often occur simultaneously with a high probability. Most people in the images, not only applying lipstick but also putting makeup on other facial parts, will be only labelled by the attribute of 'Wearing Lipstick'. Thus, the network recognizes 'Wearing Lipstick' usually relying on the makeup of several parts of a face, such as the eyes, eyebrows, and mouth [1].

Acknowledgement

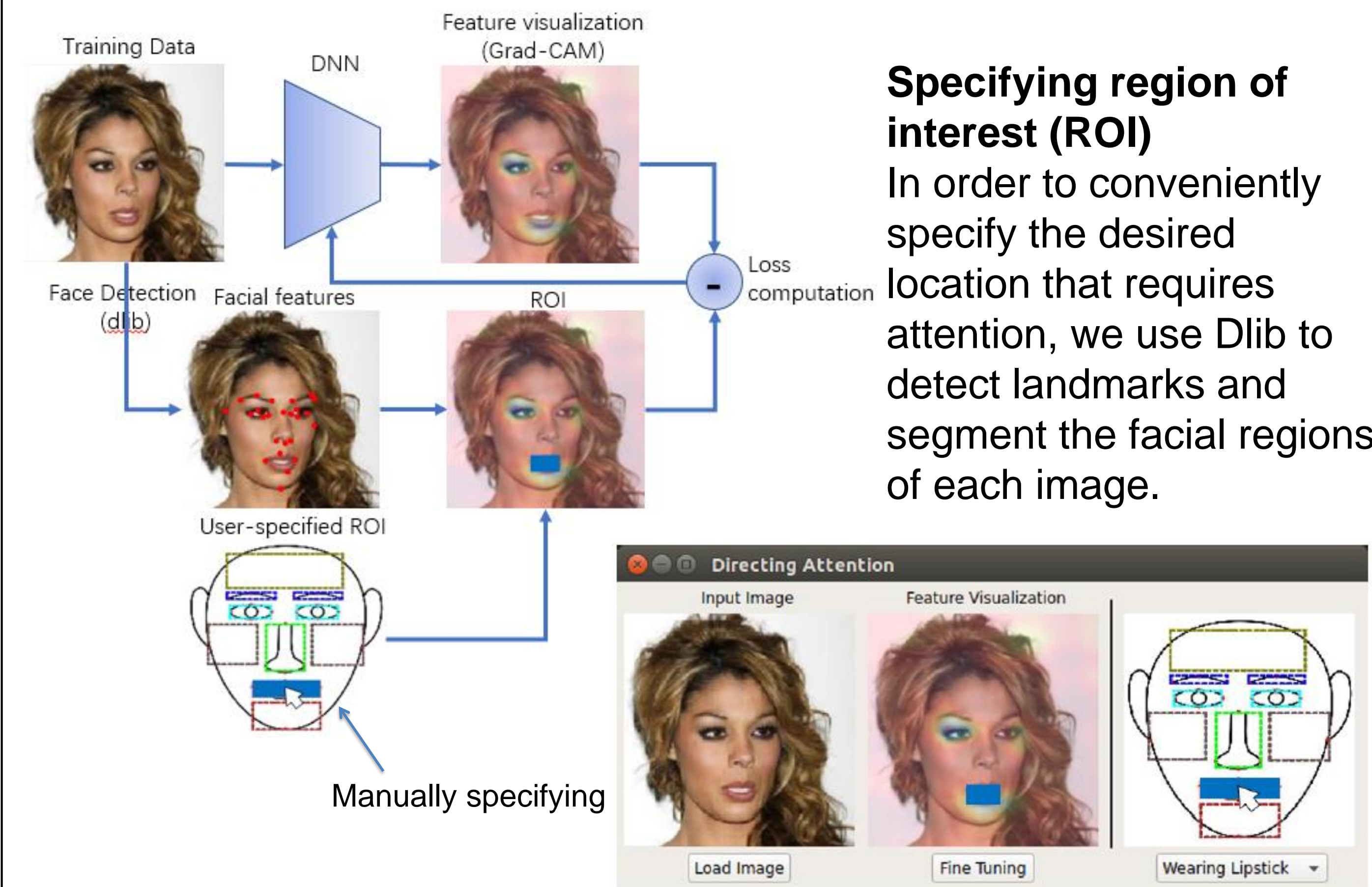
This work was supported by JST CREST Grant NumberJPMJCR17A1, Japan.

References

- [1] Q. Zhang, W. Wang, and S. Zhu. "Examining cnn representations with respect to dataset bias." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
- [2] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In ICCV, 2017.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

Method

Given a pre-trained classification DNN, we visualize the activation in the model to localize the regions where network focuses on for some example images. If the network makes a classification based on biased features, the user manually specifies the correct region on a template. This will fine-tune the pre-trained network to focus on user's defined region and direct the attention of network accordingly.

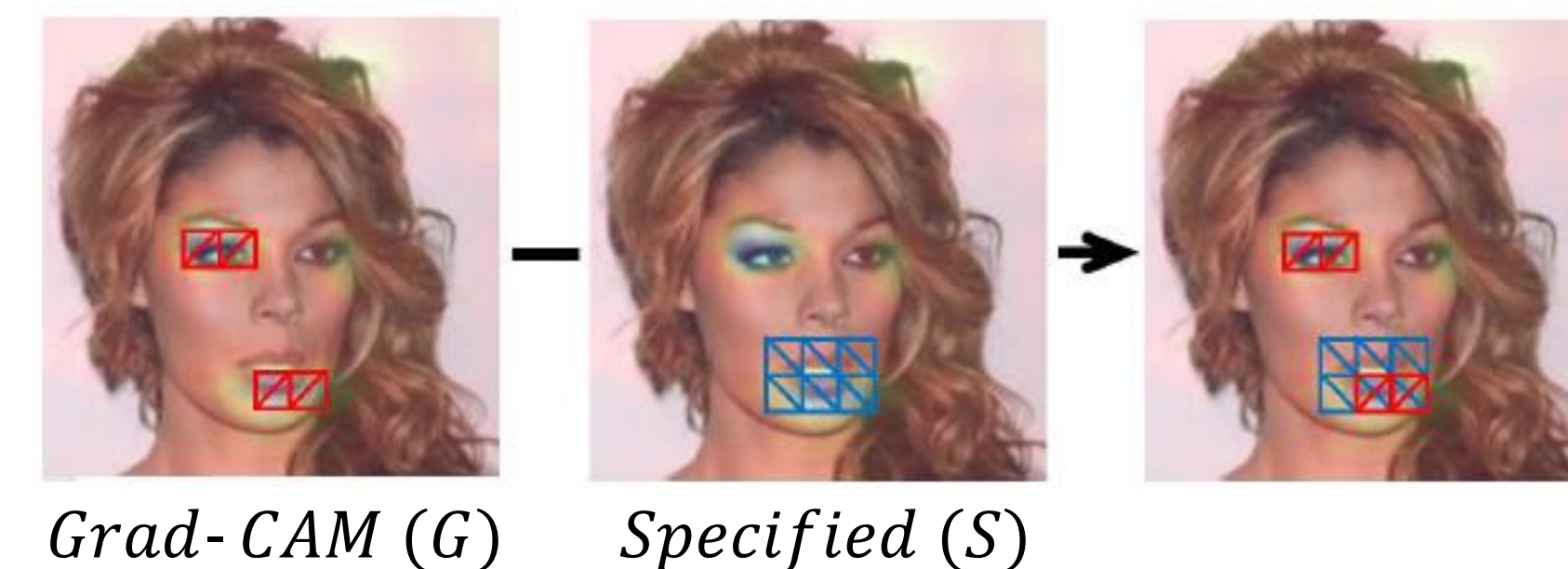


Loss function

The attribute loss $loss_a$ is the difference between the combined binary cross entropy (BCE) of the predicted scores and the labels. The Grad-CAM loss $loss_g$ is computed by comparing the Grad-CAM (G) and the user-specified (S) regions.

$$Loss = w_a \cdot loss_a + w_g \cdot loss_g$$



$$loss_g = -\ln \left(\frac{G \cap S}{G \cup S} \right)$$





Experiments

We tested our method using the CelebA dataset [3], a large-scale facial attributes dataset containing more than 200,000 celebrity images, each featuring 40 attribute annotations. We used AlexNet for facial attribute classification task. The last conv-layer delivers the Grad-CAM results.



“Wearing Lipstick” (has high co-occurrence with ‘Heavy Makeup’)

	Feature visualization	Accuracy
Pre-trained network		Test set ‘W.L.’ but no ‘H.M.’ set 92.9% 82.17% (Lower accuracy due to bias)
Fine-tuned network		93.25% 83.31% (Improved)

“High Cheekbones” (has high co-occurrence with ‘Smiling’)

	Feature visualization	Accuracy
Pre-trained network		Test set ‘H.C.’ but no ‘S.’ set 63.53% 47.41% (Lower accuracy due to bias)
Fine-tuned network		65.56% 61.33% (Improved)

“Double Chin” (has high co-occurrence with ‘Chubby’)

	Feature visualization	Accuracy
Pre-trained network		Test set ‘D.C.’ but no ‘C.’ set 84.18% 81.77% (Lower accuracy due to bias)
Fine-tuned network		84.79% 86.45% (Improved)