Adaptive Ensemble Prediction for Deep Neural Networks based on Confidence Level - Supplementary Material

Hiroshi Inoue <inouehrs@jp.ibm.com>

February 25, 2019

Here, we show additional results and discussion on mispredictions with high probabilities.

1 How mispredictions with high probability happen

In Section 2.2 of the paper, we discussed that the mispredictions with high probabilities can be caused by the insufficient expressiveness in the used model as one possible reason.

We show a simple example in Figure 1. We built a three-layer perceptron for a simple binary classification problem which maps a 2-D feature into class A or B as shown in Figure 1(a). We label a sample with feature (x, y), where $0 \leq x < 1$ and $0 \leq y < 1$ as

$$label(x, y) = \begin{cases} classA & (x \leq 0.2, y \leq 0.2) \\ classA & (x + y \geq 1.2, x \geq 0.4, y \geq 0.4) \\ classB & (otherwise) \end{cases}$$

We use the three-layer perceptron with 10 or 100 neurons in the hidden layer as the classifier and Sigmoid function as the activation function. We generated 1,000 random samples as the training data. The training is done by using stochastic gradient descent as the optimizer for 10,000 epochs.

Figure 1(b) depicts the classification results with 10 hidden neurons. In this case, the top-left region of class A is not captured by the classifier at all due to the poor expressiveness of the network even though we have enough training samples in this region. As a result, this (weak) classifier misclassifies the samples in this region for the class B with almost 100% probability. Although we repeat the training using different random number seeds, this mispredictions cannot be avoided with 10 hidden neurons. Hence, the ensembling multiple local predictions from this classifier cannot help this type of mispredictions in the top-left region. The decision boundary between class A and B at the bottom-right region is

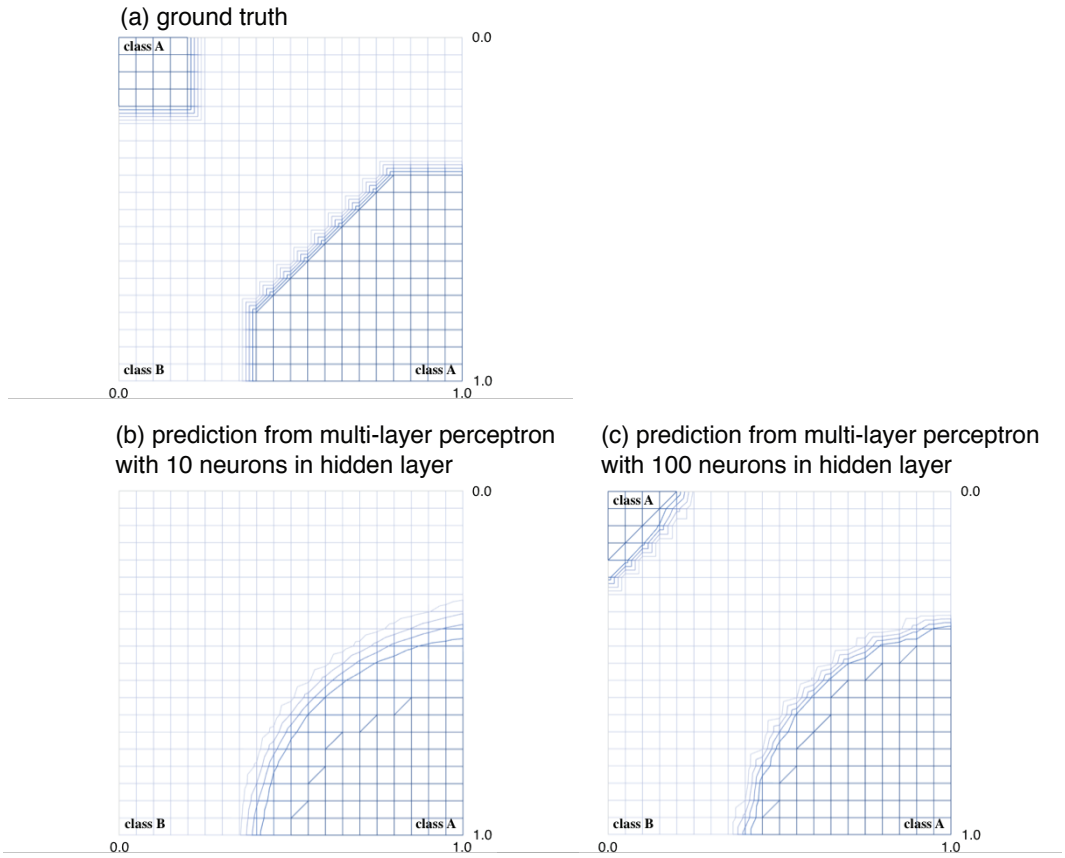

Figure 1: An example of mispredictions with high probabilities due to limited expressiveness of the classifier. A classifier with limited capability (10 hidden neurons) fails to learn the decision boundaries at the top-left region, while a classifier with higher capability (100 hidden neurons) can capture this decision boundary. The weak classifier makes incorrect classifications with almost 100% probabilities in the top-left region.

not sharp and its shape differs run by run due to the random numbers. The ensembling can statistically reduce the mispredictions near the boundary by reducing the effects of the random numbers.

When we increase the number of hidden neurons from 10 to 100, the top-left region of class A is captured as shown in Figure 1(c). So the expressiveness of the used model matters to avoid the mispredictions with high probabilities.

Another type of the mispredictions with high probabilities can happen if we do not have enough training data in small regions, e.g. the top-left region of class A in the above example. In such case, of course, even highly capable classifiers cannot learn the decision boundary around the small region and mispredict the samples in this region with high probability.

Ensembling multiple local predictions from the multiple local classifiers does not help both types of mispredictions and hence stop ensembling for them is effective to avoid wasting computation power without increasing the overall error rate.

2 Ensembling and Probability of Prediction

2.1 Ensemble using different networks

In Section 2 of the paper, we showed that ensembling two predictions from two local classifiers of the same network architecture (Alexnet [Krizhevsky et al., 2012] or GoogLeNety [Szegedy et al., 2015]) can improve the prediction accuracy only for samples that have low probabilities of prediction. Here, we show the results when we mix the predictions from Alexnet and GoogLeNet. Figure 2 shows the result when we use Alexnet in the first prediction and GoogLeNet in the second. The x-axis shows the percentile of the probability of the prediction by Alexnet from high to low as in figures shown in the main paper. The basic characteristics with two different networks are consistent with the cases using two identical networks discussed in the paper, although the improvements from the ensemble is much more significant since the second local classifier (GoogLeNet) is more powerful than the first one (Alexnet). For the leftmost region, i.e. 0- to 20-percentile samples, the ensemble from the two different networks does not improve the accuracy over the results with only the first local classifier. For the rightmost region, the ensemble improves the error rate significantly.

Here, the ensemble improves the accuracy for much wider regions compared to the cases with two identical networks. For the 20- to 40-percentile range, ensembling two local predictions from Alexnet does not improve the accuracy as shown in Figure 1(b) of the main paper while ensembling local predictions from Alexnet and GoogLeNet yields improvements in Figure 2. As discussed above using Figure 1(b) and 1(c), a more powerful classifier can avoid some mispredictions with high probabilities. GoogLeNet, which has higher capability than Alexnet, can correctly classify some samples that are misclassified with high probabilities by Alexnet in this range. However, GoogLeNet cannot do better classification in the 0- to 20-percentile range compared to Alexnet.

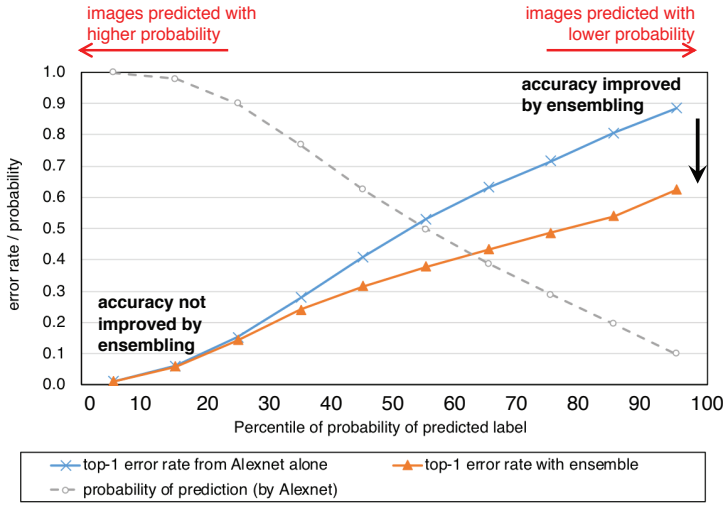

Figure 2: Improvements by ensemble and probabilities of predictions in ILSVRC 2012 validation set using Alexnet and GoogLeNet. X-axis shows percentile of probability of first local predictions from high (left) to low (right). Ensemble reduces error rates for inputs with low probabilities but does not affect inputs with high probabilities.

2.2 Results with ResNet50

In Section 2 of the paper, we showed the relationship between the probability of prediction and the effect of ensembling using GoogLeNet and Alexnet. To show that our observations are still valid with newer networks, the result with ResNet50 [He et al., 2016] is shown in Figure 3. In addition to the random cropping and flipping data augmentation used for experiments with GoogLeNet and Alexnet, we also employ sample pairing data augmentation technique [Inoue, 2018] to achieve further improvements in accuracy. Our observation, ensembling does not help mispredictions for inputs predicted with a high probability, is still valid for a newer network architecture and with more advanced data augmentation technique as we can see by comparing the results for ResNet and GoogLeNet and Alexnet.

References

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[Inoue, 2018] Inoue, H. (2018). Data augmentation by pairing samples for images classification. *arXiv:1801.02929*.

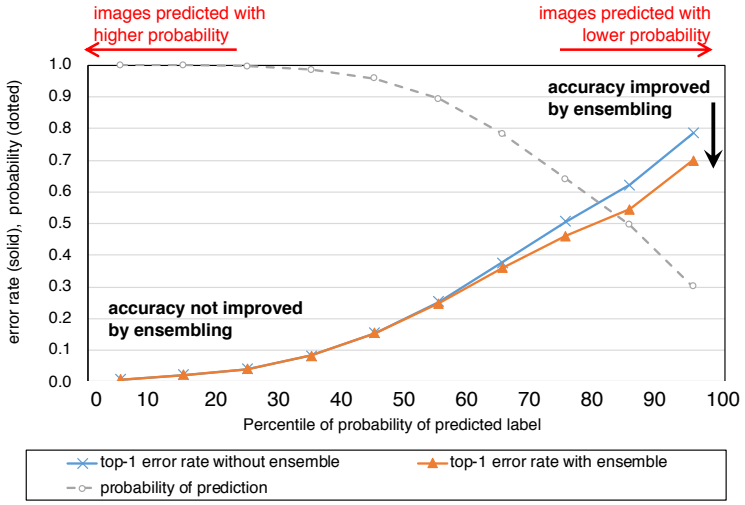

Figure 3: Improvements by ensemble and probabilities of predictions in ILSVRC 2012 validation set using ResNet50. X-axis shows percentile of probability of first local predictions from high (left) to low (right).

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1106–1114.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.