A Robust Learning Method for Deep Neural Networks Using Unsupervised Competitive Learning and Brain-Like Information Processing

Takashi Shinozaki [1,2]

[1] Center for Information and Neural Networks (CiNet),
National Institute of Information and Communications Technology (NICT),
1-4 Yamadaoka, Suita, Osaka 565-0871, Japan
[2] Graduate School of Information Science and Technology, Osaka University,
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
tshino@nict.go.jp

Abstract

Modern deep neural networks (DNNs) have a wide applicability in many fields; however they are vulnerable mainly because of the use of supervised learning for their representation learning. The features acquired through supervised learning are optimal; hence, they are the minimum requirements that allow the task to be discriminated without wasting resources. This frequently leads to the neglect of common and natural features that are not important for discrimination, making the network vulnerable to adversarial attacks. We recently developed a novel learning method for applying competitive learning, a classical unsupervised learning method, to modern DNNs. To improve this method and make it more robust, we attempted to mimic a real robust system in the wild. In particular, we simulated information processing in the brain. We used a color space similar to human perception, separated the information into color channels and spatial frequencies, and introduced local signal normalization. The top-5 accuracy of the ImageNet discrimination task was 40.73%, achieving the state-of-the-art performance as a DNN without back propagation learning. It is expected that the method will enable robust information processing in the wild using task-independent features.

Introduction

Deep learning is a powerful artificial intelligence (AI) technique that was inspired by brain mechanisms and has achieved groundbreaking results in several fields. However, current deep learning methods rely heavily on back propagation learning, which is considered an obstacle to the realization of flexible and robust information processing like that found in the brain.

As the name implies, back propagation learning proceeds from the output side of the network to the input side, in the opposite direction of information flow. Therefore, learning is most effective at the output layer and diminishes as it approaches the input side. In transfer learning, this characteristic is appropriate for modifying only the features on the output side without significantly changing the features on the input side. However, this implies that the learning and information processing on the input side is weak, leaving it vulnerable to adversarial attacks (Goodfellow, Shlens, and Szegedy 2014; Athalye et al. 2018).

To address this problem, it is natural to look for hints from the brain, which is capable of robust information processing in the wild. Recently, many studies have attempted to develop new physiologically plausible learning methods for deep neural networks (DNNs) based on brain mechanisms (Bartunov et al. 2018; Nøkland and Eidnes 2019; Krotov and Hopfield 2019). In this regard, we have been working on developing classical and physiologically plausible competitive learning (Fukushima 1980; Kohonen 1982) for use with modern DNNs (Shinozaki 2021). In this study, we developed a method to achieve more robust image processing by incorporating the information processing mechanism of the brain in greater detail.

Methods

As a learning method without error back propagation, we used our recently developed simple competitive learning for convolutional neural networks with no inter-unit interaction (Shinozaki 2021). This method learns with only feed-forward signals by performing winner-takes-all (WTA) computing between filters in the convolutional layer at each position of the input. The gradient of competitive learning is represented as follows:

$$\Delta \boldsymbol{w}_{l,i} = \begin{cases} -\rho \boldsymbol{z}_{l-1}, & \text{if } i = \operatorname{argmax}_k u_{l,k} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where ρ is the competitive learning coefficient, which we empirically set to 0.01. This gradient was used for conventional stochastic gradient descent (SGD) to update the weights, with each update being normalization by 2-norm. Moreover, as in our previous study, we introduce a conscience factor (DeSieno 1988) with a coefficient of 5.0 to improve the efficacy of competitive learning.

The ReLU activation function used in conventional DNNs is a type of threshold function, and its threshold is heavily influenced by the weight bias learned through back propagation learning. Therefore, in this study, which does not use back propagation learning, we employed WTA as the activation function rather than ReLU. This WTA activation function, like the WTA in competitive learning, produces a bundle of one-hot-vectors with only the maximum (winner)

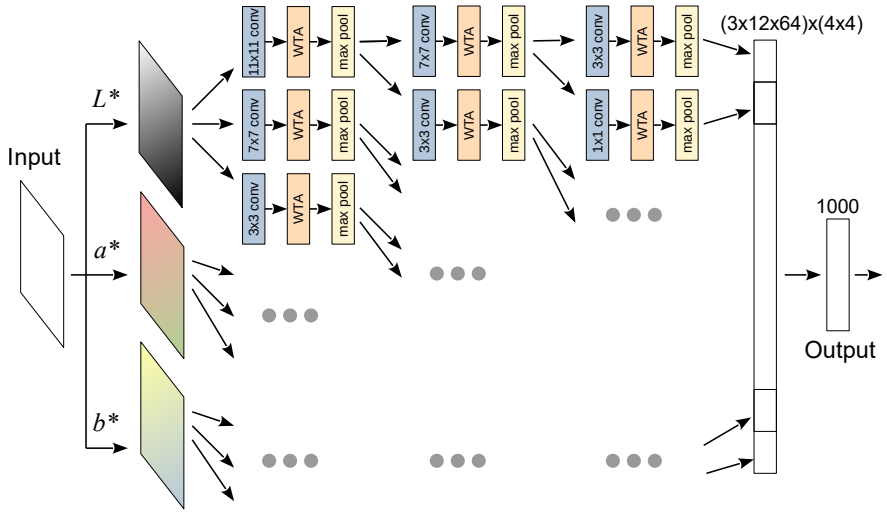

Figure 1: Schematic illustration of the proposed network structure. The RGB color input is converted to Lab color space and then propagated in separate branches.

value of each location set to one. The bundle of one-hot-vectors becomes a bundle of few-hot-vectors by subsequent max pooling and propagates through the network.

This study attempts to incorporate the robust image processing that occurs in the brain using these frameworks of competitive learning in convolutional neural networks. Luminance and color information are processed in separate pathways in our brain during image processing (Nassi and Callaway 2009). This separation allows for stable processing of color information while maintaining shape information; moreover, it is thought to provide robust discrimination against illumination changes. To reproduce such information processing, the current study used the Lab color space, which is a color space reflecting human perception and separated the paths for each channel. The Lab color space consists of three channels, 'L,' 'a,' and 'b.' These channels correspond to the gray scale, red-green axis, and blue-yellow axis, respectively. Figure 1 shows a schematic illustration of the network. The input image is converted from RGB to Lab color space, and each channel is processed separately.

Input signals with different spatial frequencies are processed also separately in the brain (Nassi and Callaway 2009). The spatial frequency processed in the convolutional layer depends on the filter size. Therefore, we applied the separation to the convolutional layer using the filter size method, similar to the inception and network-in-network models to achieve robust discrimination without relying on a specific spatial frequency.

Additionally, to achieve more robust information processing, local signal normalization was used. It has been reported that normalization has a significant effect on the robustness of signals in the brain (Carandini and Heeger 2012). Therefore, we normalized local inputs by z-values for each receptive field of the convolutional layer.

Experiments

We used the ImageNet dataset to perform discrimination tasks to validate the effectiveness of the proposed method (Russakovsky et al. 2015). The network was based on Alexnet (Krizhevsky, Sutskever, and Hinton 2012), with three convolutional layers and one all-combining layer. The network was branched by color space and filter size, resulting in 36 paths in the final convolutional layer.

The convolutional layer had 32, 32, and 64 filter units for each branch, and the filter sizes were 11×11, 7×7, and 3×3 for the first layer, 7×7 and 3×3 for the second layer, and 3×3 and 1×1 for the third layer. The max pooling was 3×3, 3×3, and 4×4 with stride two. As a result, the number of dimensions of the input to the fully connected layer was 36,864, which was significantly higher than AlexNet's 8,192. In large models, the number of filters in the convolutional layer was 64, 128, and 512 units for each branch, and the number of dimensions of the input to the fully connected layer was 294,912. The output dimensionality of the fully connected layer was 1000, which was used for the ImageNet identification.

We used unsupervised competitive learning without back propagation error for the convolutional layer. For the fully connected layer, only traditional error learning was used. The network was trained using conventional SGD, with the mini-batch size set to 32 for unsupervised competitive learning and 16 for error learning. The competitive learning coefficient was set to 0.01, and the number of iterations was set to 75,000. The learning coefficient for error learning was initially set to 0.01, then to 1/10 for each 20,000 iterations. The final number of iterations for error learning was set to 60,000.

For the baseline, we used the L channel of the Lab color space as a gray-scale signal to validate the color information. We also used RGB, XYZ, YUV, and HSV color spaces for comparison. Separated color signals were processed in

independent paths.

All codes were implemented using Python and Chainer deep learning framework (v.4.5.0) (Tokui et al. 2015) with GPU support. All experiments were run on NVIDIA Tesla P100 16 GB with the CUDA (v.9.0) and cuDNN (v.7.1.4) libraries. For color space conversion, we used scikit-image (van der Walt et al. 2014).

Results

Table 1 shows the results of the discrimination task. When the color channels were processed separately, the accuracy was significantly higher than the results for the baseline gray scale and our previous study, indicating the effectiveness of the proposed method. In particular, a top-5 accuracy of 40.73% was achieved when using a large model, which is the-state-of-the-art as a learning method without back propagation learning. However, when processing without separating the color channels, the top-5 accuracy is below the baseline, particularly in the Lab color space. This suggests that color information interferes with the processing of luminance information.

Figure 2 shows the filters of the first convolutional layer acquired by unsupervised competitive learning in the Lab color space condition. The spatial features of various orientations, but of similar spatial frequencies, are acquired in all cases. This is because the features acquired by competitive learning are dependent on their occurrence, and as a result, the features with the highest occurrence and lowest spatial frequency are dominant. On the other hand, the filter sizes vary greatly from 11×11, 7×7, and 3×3, so processing them in parallel can cover a wide range of spatial frequencies. For the color channels, by separating the gray-scale channels with the highest occurrence, the other channels are able to acquire features with color information.

Figure 3 shows the filter of the first convolution layer in the case of RGB color space. The basic trend is the same as in the case of Lab color space.

Figure 4 shows the filters of the second and third convolutional layers acquired by unsupervised competitive learning under Lab color space conditions. The input dimension is divided into three parts, and each dimension is assigned to RGB for pseudo-color representation. Each filter has a distinct spatial structure, indicating that some significant learning representation has been acquired.

Figure 5(a,b) shows the filters of the first convolutional layer when learning without separating the color channels in (a) RGB or (b) Lab color space. In both cases, the color information is mixed up, and proper representation learning has not been performed.

Figure 5(c) shows the filter of AlexNet's first convolutional layer trained by back propagation learning under the same conditions as the proposed method. Grayscale and color are automatically learned separately, and the features of various spatial frequencies are learned in grayscale.

Conclusion

In this study, we attempted to develop a robust DNN in the wild by combining representation learning with unsupervised competitive learning and mimicking the information processing of natural image recognition in the brain. We achieved the state-of-the-art results as a DNN without supervised back propagation learning using the ImageNet dataset.

We used only one fully connected layer, which is responsible for the output, owing to the limitation of not using error back propagation; however, further performance improvement can be expected by increasing the number of fully connected layers. Furthermore, rather than using unsupervised competitive learning to train all convolutional layers, it is possible to use competitive learning only for the early layers and back propagation learning for the subsequent layers.

The proposed method achieves task-independent acquisition of common and natural features through unsupervised competitive learning. This enables the use of features with significantly higher dimensionality than conventional DNNs using back propagation learning and is expected to result

Method	Top-5 Acc
Shinozaki, 2021	26.16
Gray (baseline)	25.69
RGB	33.75
Lab	**33.77**
XYZ	32.32
YUV	33.14
HSV	32.72
RGB combined	25.71
Lab combined	18.66
Lab Large	**40.73**
Alexnet	49.02
FA (Bartunov et al. 2018)	17.46

Table 1: Top-5 accuracies for Imagenet dataset

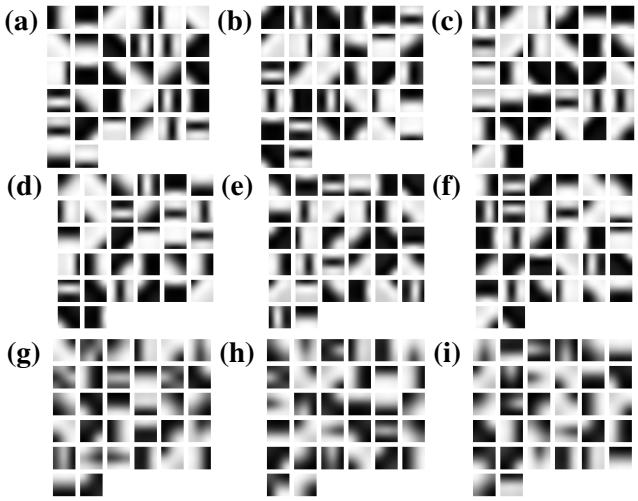

Figure 2: Filters of the first convolutional layer acquired through unsupervised competitive learning in Lab color space. The first, second, and third rows correspond to 11×11, 7×7, and 3×3 filters, respectively. (a,d,g), (b,e,h), and (c,f,i) correspond to L, a, and b channels, respectively.

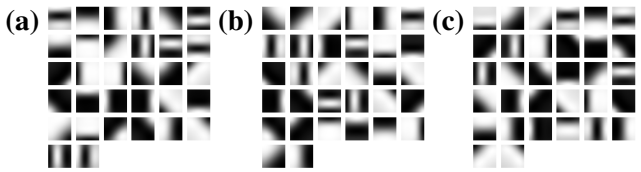

Figure 3: Filters of the first convolutional layer acquired through unsupervised competitive learning in RGB color space. Only 11×11 filters are represented. (a), (b), and (c) correspond to the red, green, and blue channels, respectively.

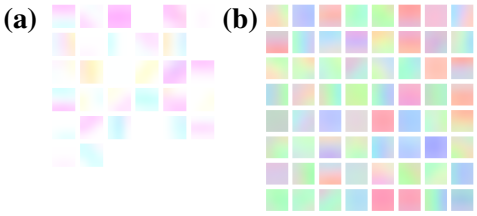

Figure 4: Filters of the (a) second and (b) third convolutional layers acquired through unsupervised competitive learning in Lab color space.

in more robust discrimination in the wild. However, this research is ongoing and we have not been able to evaluate its robustness quantitatively because we have not performed sufficient ablation experiments. We would like to discuss methods for evaluating the robustness against adversarial attacks and in various practical applications in the wild during the workshop.

Acknowledgments

This work was supported by JST ERATO Grant Number JPMJER1801, Japan.

References

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing robust adversarial examples. In *International conference on machine learning*, 284–293. PMLR.

Bartunov, S.; Santoro, A.; Richards, B.; Marris, L.; Hinton, G. E.; and Lillicrap, T. 2018. Assessing the Scalability of Biologically-Motivated Deep Learning Algorithms and Architectures. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Carandini, M.; and Heeger, D. J. 2012. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1): 51–62.

DeSieno, D. 1988. Adding a conscience to competitive learning. In *ICNN*, volume 1.

Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36[4], 193-202.

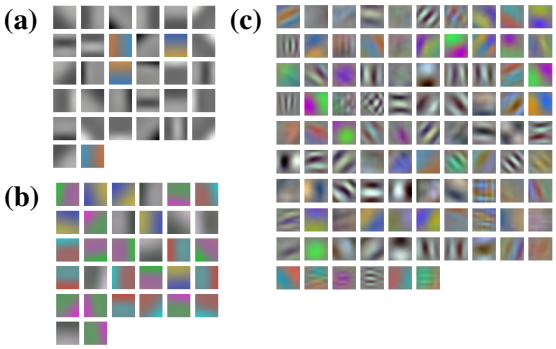

Figure 5: Filters of the first convolutional layer when learning without separating the color channels in (a) RGB or (b) Lab color space, and those of (c) AlexNet trained by back propagation learning under the same conditions as the proposed method.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1): 59–69.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

Krotov, D.; and Hopfield, J. J. 2019. Unsupervised learning by competing hidden units. *Proceedings of the National Academy of Sciences*, 116(16): 7723–7731.

Nassi, J. J.; and Callaway, E. M. 2009. Parallel processing strategies of the primate visual system. *Nature reviews neuroscience*, 10(5): 360–372.

Nøkland, A.; and Eidnes, L. H. 2019. Training neural networks with local error signals. In *International Conference on Machine Learning*, 4839–4850. PMLR.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Shinozaki, T. 2021. Biologically motivated learning method for deep neural networks using hierarchical competitive learning. *Neural Networks*, 144: 271–278.

Tokui, S.; Oono, K.; Hido, S.; and Clayton, J. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, volume 5, 1–6.

van der Walt, S.; Schönberger, J. L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J. D.; Yager, N.; Gouillart, E.; Yu, T.; and the scikit-image contributors. 2014. scikit-image: image processing in Python. *PeerJ*, 2: e453.