APAC: Augmented PAttern Classification with Neural Networks*

Ikuro SATO Hiroki NISHIMURA Kensuke YOKOI

Deep neural networks have been exhibiting splendid accuracies in many of visual pattern classification problems. Many of the state-of-the-art methods employ a technique known as data augmentation at the training stage. This paper addresses an issue of decision rule for classifiers trained with augmented data. Our method is named as APAC: the Augmented PAttern Classification, which is a way of classification using the optimal decision rule for augmented data learning. Discussion of methods of data augmentation is not our primary focus. We show clear evidences that APAC gives far better generalization performance than the traditional way of class prediction in several experiments. Our convolutional neural network model with APAC achieved a state-of-the-art accuracy on the MNIST dataset among non-ensemble classifiers. Even our multilayer perceptron model beats some of the convolutional models with recently invented stochastic regularization techniques on the CIFAR-10 dataset.

Key words :

Neural Networks, Pattern Classification, Image Classification

1. INTRODUCTION

Many of state-of-the-art methods in visual recognition problems use deep Convolutional Neural Networks (CNNs), trained on augmented datasets (see representative works [1) 2) 6) 10) 11) 12)]). It has been pointed out that CNN models with many layers tend to gain great discriminative power, while on the other hand, theoretical and methodological aspects of data augmentation are not fully revealed. Empirical studies have shown that data augmentation plays an essential role in boosting performance of generic object recognition. Krizhevsky et al. used a few types of image processing, such as random cropping, horizontal reflection, and color processing, to create image patches for the ImageNet training [6)]. More recently, Wu et al. vastly expanded the same dataset with many types of image processing including color casting, vignetting, rotation, aspect ratio change, and lens distortion on top of standard cropping and flipping [12)].

Handwritten character/digit recognition has been important for both industrial applications and algorithm benchmarking [1) 2) 7) 8) 10) 13)]. The problem is relatively simple in a sense that there is no degree of freedom in the background and that stroke can be easily deformed for data augmentation. Elastic

＊第 19 回画像の認識・理解シンポジウム (MIRU2016 *1) への投稿論文を転載（著作権は著者に帰属 *2）。
*1 https://sites.google.com/site/miru2016hamamatsu/home
*2 https://sites.google.com/site/miru2016hamamatsu/submission/instruction

distortion is one such technique that has good properties in giving a large degrees of freedom in the stroke forms, while leaving the topological structure invariant. Indeed, data augmentation by elastic distortion is crucial in boosting classification performance [1) 2) 10)].

Data augmentation can be categorized into two: off-line and on-line. In this work, off-line data augmentation means to increase the number of data points by a fixed factor before the training starts. Every instance is repeatedly processed in the training until convergence [10)]. On-line data augmentation means to increase the number of data points by creating new virtual samples at each iteration in the training [1) 2)]. There, random deformation parameters are sampled at each iteration, hence the classifier always “sees” new samples during the training. Ciresan et al. claims that on-line scheme greatly improves classification performance because learning a very large number of samples likely avoids over-fitting [1) 2)]. Our work is mostly inspired by their work, and is focused on the on-line deformation.

1.1 Contribution

We derive appropriate class decision rule for neural networks trained with augmented data, and show the effectiveness through image classification experiments. Training with the on-line data augmentation minimizes an expectation value of loss function over random deformation parameters. We claim that class decision must be made in a specific way so as to minimize the same expected loss at test time. This requires to process a large number of virtual samples for a given test sample, as discussed in the following sections.

Though we believe that the proposed decision rule is beneficial to broad classification problems, image classification problems are discussed in this paper because we have not conducted experiments in other fields.

2. Augmented data learning

On-line data deformation learning can generate classifiers highly robust to intra-class variations. Such learning generally consumes many iterations to reach a minimum of the objective function. A vast number of training instances are processed because the number of instances increases linearly as the number of iterations increases. In the on-line deformation scheme, the original data themselves are not trained explicitly, they are only trained probabilistically.

In this section we provide a formal definition of augmented data learning, which has been treated rather heuristically so far. Let us first define the data deformation function as $u\colon \mathbb{R}^d \to \mathbb{R}^d$, where d is the dimension of the original data. The function $u(x;\Theta)$ takes a datum $x \in \mathbb{R}^d$ and deformation-controllingparameters $\Theta = \{\theta_1,\cdots,\theta_K\}$, and returns a virtual sample. Each element of the set Θ is defined as a continuous random variable for convenience. Some are responsible for continuous deformation; *e.g.*, θ_1 being a scaling factor. The other are responsible for discrete deformation; *e.g.*, $\theta_2 \in [0,\frac{1}{2})$ meaning side-flipping, and $\theta_2 \in [\frac{1}{2},1]$ meaning no flipping, where $\theta_2 \sim \mathcal{U}(0,1)$. It is assumed that probability density functions of deformation parameters are given at the beginning and held fixed during training and testing. We use the cross entropy as the loss function. The cross entropy requires vector normalization in the output units, where we use the softmax function.

Let $i \in \{1,\cdots,N\}$ denote an index of original training data, $c_i \in \{1,\cdots,N_c\}$ denote the class index of i-th sample, W denote the set of all parameters to be optimized, and $f(\cdot;W)\colon \mathbb{R}^d \rightarrow \mathbb{R}^{N_c}_{>0}$ denote a function realized by a neural network. Let f_c be the c-th component of the output, then $\sum_{c=1}^{N_c} f_c = 1$ and $f_c > 0$, $\forall c \in \{1,\cdots, N_c\}$. Regularization terms are ignored here. Problem of augmented data learning is stated as follows.

Augmented Data Learning:

Given D=$\{(x_i,c_i\}$, $i=1,\cdots, N$, find $W^\star$ such that

$$\mathrm{W}^\star = argmin_W\, J_{\mathcal{D}}(W), \qquad (1)$$

where the objective function $J_D\,(W)$ is defined as

$$J_{\mathcal{D}}(W) = \sum_{i=1}^{N} \mathbb{E}_{\Theta}\left[-\,ln\left(f_{c_i}(u(x_i;\Theta);W)\right)\right]. \qquad (2)$$

The expectation value is computed by marginalizing the cross entropy over deformation parameters that independently obey unconditional probability densities $p_k\,(\theta_k\,), k=1,\cdots,K$. By using appropriate random number generators, one can generate countlessly many virtual samples during training. By sufficiently reducing the objective function, the classifier gains a high level of intra-class invariance with respect to the set of deformations applied, without compromising inter-class distinctiveness.

A truly deformation-robust classifier would be obtained, if the integrals $\mathbb{E}_\Theta\,[\cdot]=\int\cdots\int\prod_k d\theta_k\ \mathrm{p}_k\,(\theta_k)(\cdot)$ were analytically calculated. However, it is hard to integrate out in reality. The integral can generally be converted into a sum of infinitely many terms,

$$\mathbb{E}_\Theta[\cdot] = \lim_{R\to\infty}\frac{1}{R}\sum_{\Theta=\Theta^{(1)},\cdots,\Theta^{(R)}}(\cdot). \qquad (3)$$

Here, $\Theta^{(\ell)} = \left\{\theta_1^{(\ell)},\cdots,\theta_k^{(\ell)}\right\}$ is a set of deformation parameters at ℓ-th sampling, based on the probability densities $\mathrm{p}_k\,(\cdot), k=1,\cdots,K$. With this summation form, the objective function can be approximately minimized by widely-used mini-batch Stochastic Gradient Descent (SGD). Note that a batch optimization algorithm is no longer applicable in a strict sense because the number of terms is infinite. At each iteration in the optimization process, data indices and deformation parameters are randomly sampled to generate a mini-batch. The mini-batch is discarded after a single use.

The total number of trained instances is determined when the training is terminated. Note that the original data samples are not explicitly fed into the network.

3. Decision rule for augmented data learning

In this section we propose a new way of classification, APAC: Augmented PAttern Classification for augmented data learning described in the previous section. It is shown that a single feedforward of a given test sample is no longer a good choice when one minimizes the expected loss at the training stage. Pattern classification problem corresponding to the aforementioned augmented data learning is stated as follows.

APAC (Augmented PAttern Classification):

Given parameters W and data x, find $c^\star$ such that

$$c^\star = argmin_{c\in\{1,\cdots,N_c\}}\, J_{\{(x,c)\}}(W). \qquad (4)$$

It is worth pointing out that class decision making is also an optimization process requiring minimization of the expected loss. The expected loss for a given data sample must be computed at test stage, as it is minimized through training stage. Note that the test sample itself is not explicitly fed into the classifier in the proposed decision rule. In practice, finite-term relaxation must be made at test stage to estimate the expectation value in the objective function,

$$\mathbb{E}_\Theta[\cdot] \simeq \frac{1}{M}\sum_{\Theta=\Theta^{(1)},\cdots,\Theta^{(M)}}(\cdot)\,, M \gg 1. \qquad (5)$$

This means, a large number of sets of deformation parameters must be randomly sampled using the same probability densities used in the training to generate virtual instances are created from test sample *x*. APAC requires to average the logarithms of the softmax outputs of the virtual instances, and then take the maximum argument to give prediction (see **Fig. 1**). We emphasize that 1) taking logarithm of the softmax output is important, otherwise an irrelevant quantity gets minimized at the test stage, and 2) sufficiently many virtual instances must be generated to have a good estimate of the expected loss. APAC is equivalent to picking the maximum argument of the product of the softmax output, which is analogous to selecting the largest joint probability among individual class probabilities of many virtual instances. For a sufficiently trained classifier, it is expected that generalization performance asymptotically reaches the highest as the number of terms, *M*, increases.

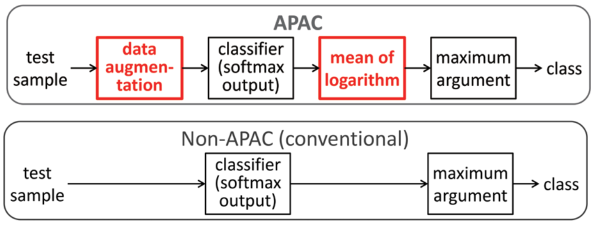


Fig. 1 APAC, the proposed way of classification (above). Non-APAC, conventional way of classification (below).

4. Experiments

Experiments on image classification are carried out to evaluate generalization abilities of APAC.

4.1 Setup

We used MNIST [7], CIFAR-10 [5], and ILSVRC2012 [9]. We evaluated CNNs on all three datasets, and MLPs (multilayer perceptron, meaning fully-connected, layer-by-layer feedforward neural network) on MNIST and CIFAR-10. The MNIST-CNN has 2 convolutional layers (2C, shortly), 2 pooling layers (2P) and 2 fully-connected layers (2F). The CIFAR-10-CNN has (3C, 3P, 2F). The ILSVRC2012-CNN has (10C, 4P, 1F), designed on our own. Each of the MLPs has 3F. On-line data deformation is carried out in each training.

4.2 Classification performance

Table 1 Summary of test error rates. Top-5 validation error rates are shown in the last row.

Trained on		Augmented data		Original data
Tested by		APAC	Non-APAC	Non-APAC
MNIST	CNN	**0.23%**	0.39%	0.69%
	MLP	**0.26%**	0.29%	1.49%
CIFAR-10	CNN	**10.33%**	20.05%	22.63%
	MLP	**14.07%**	23.20%	55.96%
ILSVRC2012	CNN	**15.47%**	20.67%	-

Table 1 summarizes test error rates (For ILSVRC2012 top-5 validation error rates are evaluated) produced by APAC and non-APAC for classifiers trained on augmented data, as well as the results using no data augmentation at all during training and testing. The APAC results shown in **Table 1** are those with M = 16,384 for all the MNIST and CIFAR-10 experiments, and M = 4,096 for the ILSVRC2012 experiment (see Eq. (5)). Two observations can be made. 1) Augmented data learning has positive effect in classification accuracies. 2) APAC consistently gives much better accuracies than non-APAC, prediction made by feedforwarding the original test samples{ albeit they use the same weights trained with augmented data.

4.2.1 Performance on MNIST

Our CNN model achieved 0.23% test error rate. To the best of our knowledge, this is the best when a single model is evaluated. All misclassified test samples are shown in **Fig. 2**. The top-2 prediction error rate is as low as 0.01%; i.e., there is only one misclassified sample out of 10K test samples.

Our single MLP model achieved 0.26% test error rate. To the best of our knowledge, this is the best record among MLP models reported previously. Our MLP model has, again, 0.01% top-2 prediction error rate on the test dataset. Interestingly, the very same test sample (shown at the top-left in **Fig. 2**) is misclassified by our CNN and MLP models.

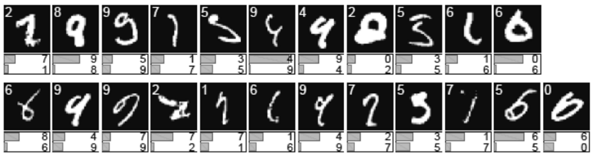


Fig. 2 All MNIST test samples misclassified by our CNN model. In each figure, ground truth is printed at the top-left corner. The bar plot in each figure indicates softmax output of the 1st and 2nd predictions.

4.2.2 Performance on CIFAR-10

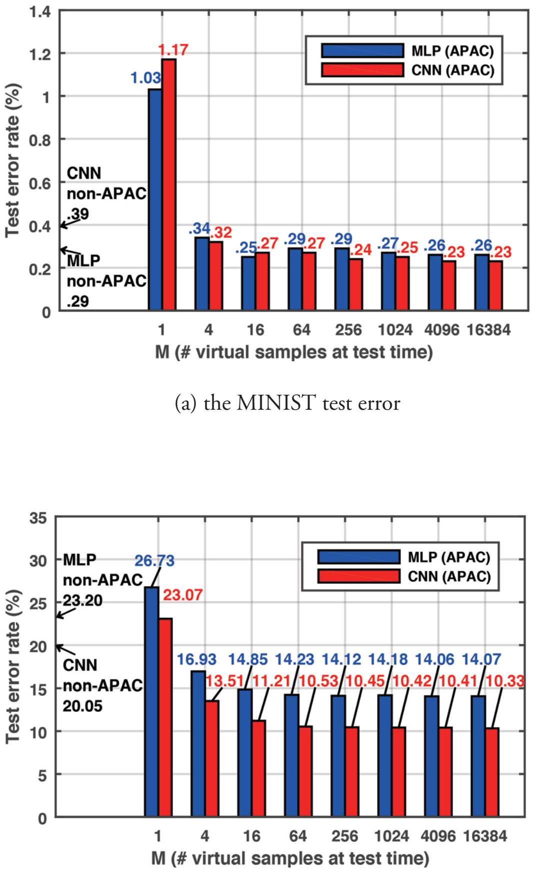
Our single MLP model yields 14.07% test error rate. This is worse than the multi-column CNN (11.21%) [2)], but better than the CNN with stochastic pooling (15.13%) [14)] and the CNN with dropout in final hidden units (15.6%) [4)]. We are aware that MLPs are easy to over-fit when used for image classification tasks. But still, this experiment gives an evidence that a fully-connected network trained with augmented data and tested with APAC can outperform CNNs trained with recently invented regularization techniques and without augmented data [4) 14)].

4.2.3 Performance on ILSVRC2012

Our single CNN model yields 15.47% top-5 validation error rate, which is better than the winning entry of the ILSVRC 2012 competition (16.42% from 5 CNN ensemble [6)]), and worse than the winning entries of more recent competitions (11.74% [15)], 6.67% [11)], 3.57% [3)]). Though our result is not close to those of the state-of-the-art methods, it is still convincing that APAC improves accuracies through this large-scale image classification experiment.

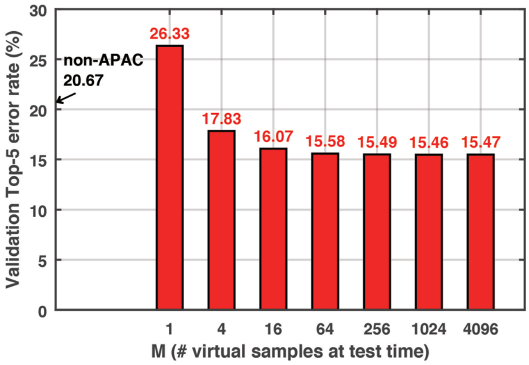
4.3 Asymptotic behavior and computational issues

We evaluate how the classification error rates change as M goes to large values (**Fig. 3**). Non-APAC results are also shown in the figure with texts. The same weights are used for both APAC and non-APAC. The tendency that the classification accuracy raises as M increases is clearly observed in every experiment. This is due to the fact that the expected loss $J(W)$ is better estimated as M gets larger when classifier is sufficiently trained with on-line data augmentation. Though it is computationally demanding to reach extremely large M, in practice M = 16 seems enough to gain high generalization and going beyond M = 16 gives only marginal effects.



(a) the MINIST test error

(b) the CIFAR-10 test error

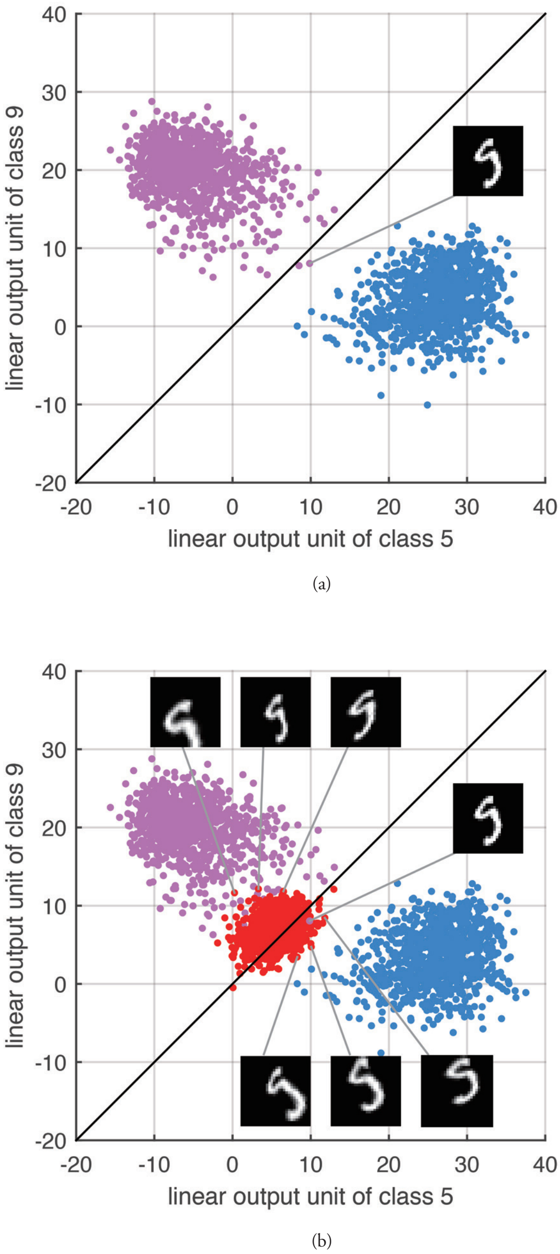


(c) the ILSVRC2012_validation_top-5 error

Fig. 3 Asymptotic behavior of the classification error rates for large *M*.

4.4 Analysis

All the experiments we conducted showed that APAC consistently gives better classification accuracies than non-APAC, when augmented data are learned. Let us illustrate how the class prediction gets altered between the two decision rules in the case of MNIST classification. **Fig. 4 (a)** shows a scatter plot of test data points of class-5 and class-9 in a 2D subspace of the linear output space, with x and y-axis corresponding to class-5 unit and class-9 unit. There, weights are obtained through the on-line deformation learning, and plotted data points do not involve image deformation. A test sample, whose image is superposed in the plot, would be misclassified to class-5 by non-APAC. We deform this test sample in 1,000 different ways, and plot these virtual data points in **Fig. 4 (b)**. The observation is that the majority (661 out of 1,000) of such virtual data points are in favor of the true class ('9'). Indeed, APAC predicts the true class from the 1,000 virtual samples. An important point is that there is a better chance of predicting the correct class by taking the product of softmax output of many virtual samples created from a given test sample, rather than by using the softmax output of the test sample itself.



(a)

(b)

Fig. 4 Illustration of APAC prediction of a class-marginal sample. The violet and light blue points are the class-9 and class-5 test data points, respectively, of MNIST. The red points are the virtual data points created from a particular test sample. See the text for more details.

One might wonder what happens if summation, instead of product, of softmax output of many virtual samples is taken at test stage. We list the results below. Test error rates produced by taking the maximum argument of the softmax sum with M =16,384 are: 0.24% for MNIST-CNN, 0.27% for MNIST-

MLP, 10.42% for CIFAR-10-CNN, and 14.01% for CIFAR-10-MLP. Softmax product gives better performance in all cases except for the CIFAR-10-MLP. We do not have a clear explanation why one out of four experiments exhibits opposite result, but it is safer and more meaningful to use softmax product so as to maximize the joint probability among individual class-probabilities of many virtual instances.

5. Conclusion

This paper addresses an issue of appropriate decision rule for neural networks trained with augmented data created in on-line fashion. Experiments on visual classification tasks revealed that the proposed way of classification, APAC, gives far better generalization than traditional decision rules.

REFERENCES

1) Ciresan, D. C., Meier, U., Gambardella, L. M. and Schmidhuber, J.: Deep, Big, Simple Neural Nets for Handwritten Digit Recognition, Neural Computation, Vol. 22, No. 12, pp. 3207-3220 (2010).

2) Ciresan, D. C., Meier, U. and Schmidhuber, J.: Multicolumn Deep Neural Networks for Image Classification, Computer Vision and Pattern Recognition, pp. 3642-3649 (2012).

3) He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, CoRR, Vol. abs/1512.03385 (2015).

4) Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. R.: Improving neural networks by preventing co-adaptation of feature detectors, ArXiv:1207.0580 (2012).

5) Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images, Master's thesis, Computer Science Department, University of Toronto (2009).

6) Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems, pp. 1097-1105 (2012).

7) Le Cun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient Based Learning Applied to Document Recognition, Proceedings of IEEE, Vol. 86, No. 11, pp. 2278{2324 (1998).

8) LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D.: Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation, Vol. 1(4), pp. 541-551 (1989).

9) Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV), Vol. 115, No. 3, pp. 211-252 (2015).

10) Simard, P. Y., Steinkraus, D. and Platt, J. C.: Best practices for convolutional neural networks applied to visual document analysis, International Conference on Document Analysis and Recognition, pp. 958-963 (2003).

11) Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going Deeper with Convolutions, CVPR 2015 (2015).

12) Wu, R., Yan, S., Shan, Y., Dang, Q. and Sun, G.: DeepImage: Scaling up Image Recognition, ArXiv:1501.02876 (2015).

13) Yaeger, L. S., Lyon, R. F. and Webb, B. J.: Effective Training of a Neural Network Character Classifier for Word Recognition., Advances in Neural Information Processing Systems, pp. 807-816 (1996).

14) Zeiler, M. D. and Fergus, R.: Stochastic Pooling for Regularization of Deep Convolutional Neural Networks, ArXiv:1301.3557 (2013).

15) Zeiler, M. D. and Fergus, R.: Visualizing and Understanding Convolutional Networks, Computer Vision - ECCV 2014 -13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, pp. 818-833 (2014).

著者

佐藤 育郎
さとう いくろう

株式会社デンソーアイティーラボラトリ
博士（理学）
画像認識アルゴリズムの研究開発に従事

西村 裕紀
にしむら ひろき

株式会社デンソー
画像処理・機械学習を応用した走行安全技術の研究開発に従事

横井 健介
よこい けんすけ

株式会社デンソー
ADAS用画像センサの認識仕様開発に従事