Can data augmentation really replace explicit regularization?

Critical review on "Data augmentation instead of explicit regularization" by Alex Hernández-García and Peter König

Deep Learning : Statistical perspective
2018-10036 Park Yoonsoo

Abstract

"Data augmentation instead of explicit regularization" by Alex Hernández-García and Peter König(2019) suggested new possibility and future of data augmentation. Normally, it was considered as one of the methods to supplement the lack of training datasets. However, what the paper had proposed was that this can actually replace explicit regularization such as weight decay and dropout. This critical review verifies the validity of the paper's proposals regarding data augmentation methods, neural network designs and datasets. Finally, the result demonstrates that the data augmentation alone cannot always replace the explicit regularization.

1 Introduction

Generally in studies, researchers use both explicit regularizations and data augmentation to decrement overfitting in larger and deeper neural networks. However, it is always difficult to separate the effect of the two since they do similar roles and act similarly. According to Hernández-García and Peter König(2019), data augmentation can perform better than explicit regularization and even more, it can sometimes work better being alone without any regularization. This suggestion seems appealing and reasonable, but they left some to be desired in methodological perspective.

In order to supplement their paper, this paper contributes to:

- Summarize the main thesis of the paper.
- Re-examine the proposal of the paper "Data augmentation instead of explicit regularization" by Alex Hernández-García and Peter König(2019) on newer data augmentation methods.
- Expand the experiment to a deeper neural network.
- Apply the experiment to a larger datasets in order to support the argument of the paper.

2 Explicit and Implicit Regularization

The borderline between regularization and data augmentation is blurred. As regularization can be defined as modification on the learning algorithm in order to prevent overfitting, data augmentation, which increases generalization by incrementing training data counts, can be considered as one of the regularization methods. On the other hand, explicit regularizations like weight decay and dropout add random noises on training datasets, so they are covered by the scope of data augmentation.

So to clarify the concept and the terminology of the paper, this paper borrows the notion of explicit regularization and implicit regularization from the existing paper (Hernández-García & Peter König, 2019).

- Explicit regularization : modification of the effective capacity of neural network models but not the interference in the architecture nor in the structure of neural network. In this paper, explicit regularization usually refers to weight decay and dropout.
- Implicit regularization : modification of the network architecture like training dataset or algorithm. Here, implicit regularization mainly implies data augmentation.

3 Suggestions of the Paper

There are three main arguments of the paper in order to prove why data augmentation can be used instead of explicit regularizations and even more, why explicit regularizations are actually unnecessary.

3.1 Theoretical Perspective

First, in theoretical perspective, data augmentation has an advantage over explicit regularization.

$$L_P(h) - L_S(h) \leq 2\,\mathrm{Rad}(F \circ S) + 4\sqrt{\frac{2\ln(4/\delta)}{m}}$$

As seen above, for every $\delta < 0$, with probability at least $1 - \delta$, for every hypothesis $h \in H$, generalization error is bounded by empirical Rademacher complexity. [1] In this inequality, both data augmentation and explicit regularization have relevance to m, the number of training data counts. Well-designed data augmentation directly increases m, since data augmentation with same distribution as the original can create numerous datasets in principle.

However, general explicit regularizations restrict h, so that this change reduces complexity indirectly. This may be helpful but there is yet small knowledge on how explicit regularizations affect generalization errors.

3.2 Needlessness of Explicit Regularizations

Second, the paper proved empirically that data augmentation exceeds the performance of explicit regularization. The paper chose three neural network models and three datasets for experiments. Based on this experiment, the paper suggests that explicit regularization is unnecessary to decrease overfitting and to increment the performance. Moreover, this paper proposes that data augmentation is able to reach its highest competence without explicit regularization, in other words, when used alone.

3.3 Poor Adaptability of Explicit Regularizations

Lastly, the experiment of the paper shows how explicit and implicit regularizations respond to the changes on the number of training data respectively. According to the result, while models without explicit regularizations naturally adapted to the change, models with explicit regularizations seem to be unstable upon the changes. There are mainly two reasons for this result: first, the parameters of regularized models cannot adapt to the change since they have been tuned to the original data. Second, models with data augmentation originally benefit from the increased number of data, but it seems that constrained hypothesis class prohibits models from deriving benefit.

[1] Shalev-Shwartz, Shai; Ben-David, Shai (2014). Understanding Machine Learning – from Theory to Algorithms. Cambridge University Press.

4 Re-examination of the Experiment

With respectful attitude, there are some possibilities of expanding the experiment from the paper in various ways. In this paper, the expansion will cover three parts: data augmentation methods, neural network designs, and datasets.

4.1 Data Augmentation Methods

The paper of Hernández-García and Peter König(2019) has tested on each model with *Light* augmentation and *Heavier* augmentation. This notation is qualified in sense. According to the paper, *Light* augmentation consists of horizontal flips and translations of 10% of each image in both horizontal and vertical way. On the other hand, *Heavier* augmentation means more various sets of transformations including scaling, rotations and shear mappings, contrast and brightness adjustment.

While its main subject is data augmentation, this paper selected limited methods. In recent years, a variety of new data augmentation methods have emerged. This part raises question of whether these new methods also can replace regularization. Among them, this paper chose two, Cutout(DeVries & Taylor, 2017) and AutoAugment(Cubuk et al., 2019).

4.1.1 Cutout

Figure 1: Cutout applied to images from the CIFAR-10 dataset. From DeVries & Taylor, 2017, p.1.

Cutout was proposed in 2017 and this method crops a random box from the original images and sets that part to zero. It inherited the notion of *Occlusion* which was suggested in 2011(Bengio et al., 2011). Occlusion means hiding some part of original data during training. Actually, Cutout is the notion between regularization and data augmentation. Since Cutout sets random value of original data to zero, the basic principle is kind of similar to Dropout in Neural Network.

4.1.2 AutoAugment

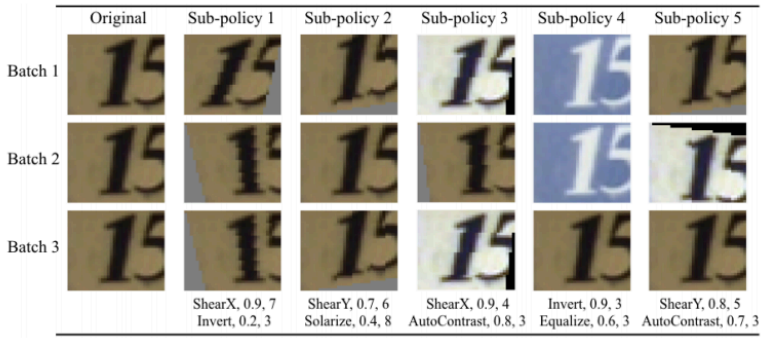

Figure 2. One of the policies found on SVHN. From Cubuk et al., 2019, p.3.

The Second method is AutoAugment. AutoAugment means the reinforce learning process for data augmentation before the actual neural network training. The *Controller* finds the best 5 policies in the search place by reinforce learning. Here, the search place consists of the basic data augmentation operations such as Share, Translate, Rotate, AutoContrast, Invert, Cutout. So it covers almost all the existing methods in the paper.

4.2 Neural Network

Second point is that the algorithm of this paper should be extended to another neural network architecture. This paper used three most famous models, All-CNN, WRN, and DenseNet. However, the experiment should always be consistent with other neural networks.

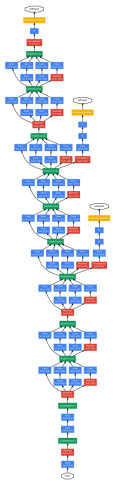

Figure 3: GoogLeNet network with all the bells and whistles. From Szegedy et al., 2014, p.7.

GoogLeNet(Szegedy et al., 2014) is a model that can control the computing resource efficiently while being sparse and dense at the same time. Since dropout is a non-uniform operation, it takes a lot of hardware resources and computing expenses. So they found the way to limit such operations and the number of parameters while increasing the accuracy.

However, with regard to GoogLeNet, there is no clear explanation why there is still a dropout phase in this model even though they have changed the whole model structure in order to replace dropout. Therefore this can be an adequate model to test whether data augmentation can replace dropout and weight decay.

4.3 Datasets

Lastly, the paper took three image datasets, CIFAR-10, CIFAR-100, and ImageNet. The three may be enough but experimental expansion can increase the validity of the suggestion. So as an additional dataset, this paper chose Google OpenImage Datasets, which is composed of almost 9 million different images with annotations. To simplify the test, random 5 classes are selected from the whole dataset.

5 Result

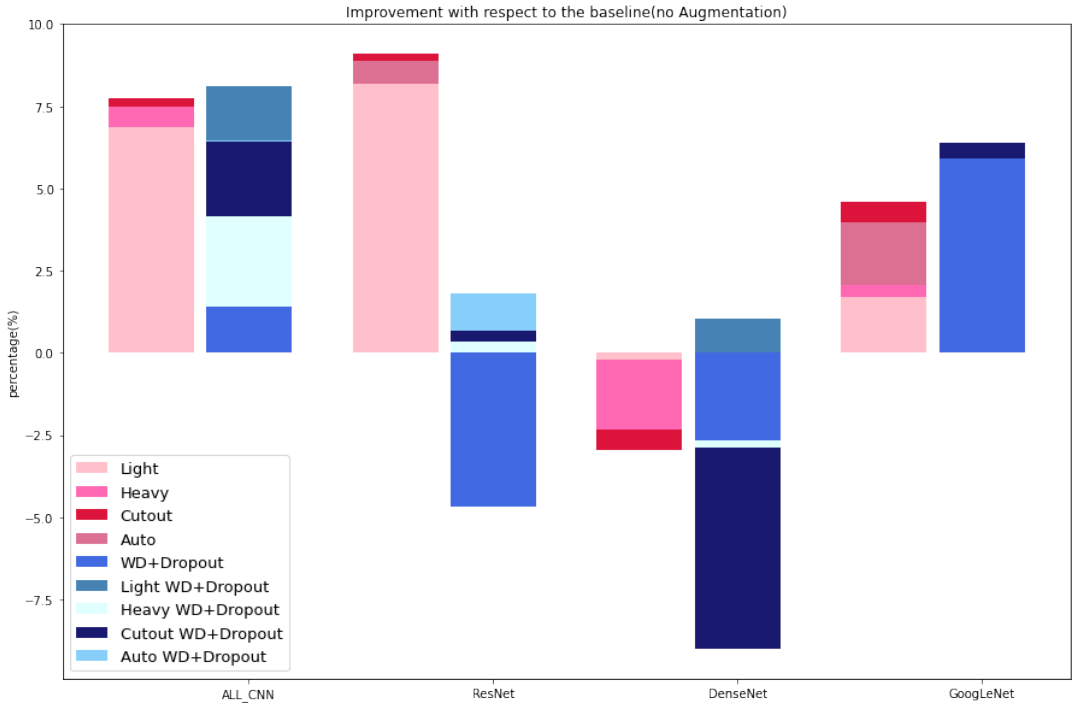

Figure 4: Improvement with respect to the baseline as data augmentation and explicit regularization are added to models.

As seen above, most models without explicit regularizations can perform not only as well as, but also better than those with explicit regularizations. This is consistent with the result of the original paper. Other than that, there are some points that need to be demonstrated regarding the paper.

5.1 Necessity of Explicit Regularization

First of all, among four models, all except only one got the best performing model by applying both data augmentation and explicit regularization. In particular, All-CNN and DenseNet achieve the highest performance by combining Light augmentation with weight decay and dropout. In case of GoogLeNet, the model with Cutout and explicit regularization makes the best result. Therefore, it is hard to say that explicit regularization is totally unnecessary and it cannot be perfectly replaced by data augmentation. It is obvious that explicit regularization itself cannot outperform data augmentation method, but when they are combined together, it can achieves highest accuracy.

Second, unlike the result of the original paper, GoogLeNet shows that data augmentation alone cannot yield better performance than explicit regularization alone. In other words, training with weight decay and dropout surpasses all the data augmentation methods. As a result, explicit regularization has cemented itself as one of main regularization tools.

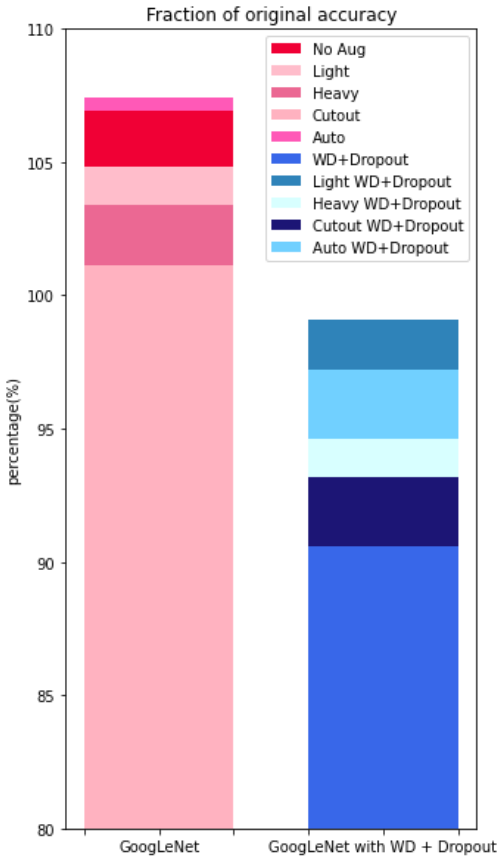

Figure 5: Fraction of original accuracy as only 50% of training data was used

5.2 Reconfirmation of Poor Adaptability of Explicit Regularization by GoogLeNet

As mentioned in the original paper, models with explicit regularization tend to poorly adapt to the limited usage of training dataset. As this experiment has tested the same mechanism on GoogLeNet, the result is similar.

After the GoogLeNet models were trained by using 100% of training data, each model was trained by only 50% of the same data. The figure above shows the performance of the latter models compared to the former ones. While models without explicit regularizations adapt well to the reduction of training images, those with explicit regularizations shows evident decline of the accuracy.

6 Discussion

This paper has shown three different domains that the previous study can be extended. First, recent data augmentation methods can be applied to the original experiment and they have produces consistent result with the original. Second, another deep learning architecture, GoogLeNet is adopted to the new experiment. Lastly, a dataset other than CIFAR-10, CIFAR-100, and ImageNet is used to improve the validity of the paper.

There are mainly two suggestions of this paper. Firstly, explicit regularization is necessary and can improve the performance when collaborating with data augmentation. Especially, GoogLeNet shows that only explcit regularization can achieve better accuracy than data augmentation and even more, GoogLeNet model with both explicit regularization and data augmentation can reach its highest performance. Second, including the new GoogLeNet model, all deep neural network models used in the experiments exhibit a lack of adaptability of explicit regularization.

Some points in this paper can be improved later. One of them is the diversity of datasets. Data augmentation and explicit regularization can work efficiently not only for image classification but also for natural language processing(Wei & Zou, 2017). So the experiment of this paper can be also applied to this kind of data. Also, as mentioned before, data augmentation takes advantage of the prior knowledge of datasets. However, this paper ignores this point to some extent. While adopting multiple data augmentation methods and especially AutoAugment complemented the problem, but following research may compensate the defect.

Reference

Alex Hernández-García, Peter König. Data augmentation instead of explicit regularization. ICLR 2018. arXiv:1806.03852v4, 2019.

Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, Quoc V. Le. AutoAugment: Learning Augmentation Strategies from Data. CVPR 2019. arXiv:1805.09501v3, 2019.

Terrance DeVries, Graham W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. arXiv:1708.04552v2, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 1-9. 2015.

Jason Wei, Kai Zou. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. EMNLP-IJCNLP 2019. arXiv:1901.11196v2, 2019.

Bengio, Y., Bastien, F., Bergeron, A., Boulanger–Lewandowski, N., Breuel, T., Chherawala, Y., Cisse, M., Côté, M., Erhan, D., Eustache, J., Glorot, X., Muller, X., Pannetier Lebeuf, S., Pascanu, R., Rifai, S., Savard, F. & Sicard, G.. Deep Learners Benefit More from Out-of-Distribution Examples. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, in PMLR 15:164-172, 2011.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, Tsuhan Chen. Recent advances in convolutional neural networks. Pattern Recognition 77. p. 354-377. 2018.

Rajput, S., Feng, Z., Charles, Z., Loh, P.-L., and Papailiopoulos, D. Does data augmentation lead to positive margin?. International Conference on Machine Learning. 2019.

Alex Hernandez-Garcia, Peter König. Do deep nets really need weight decay and dropout?. 2018.

David Helmbold, Philip Long. Fundamental differences between Dropout and Weight Decay in Deep Networks. 2016.

Connor Shorten, Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data **6**, Article number: 60. 2019.

Luis Perez, Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. 2017.