DATA AUGMENTATION INSTEAD OF EXPLICIT REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern deep artificial neural networks have achieved impressive results through models with very large capacity—compared to the number of training examples—that control overfitting with the help of different forms of regularization. Regularization can be implicit, as is the case of stochastic gradient descent or parameter sharing in convolutional layers, or explicit. Most common explicit regularization techniques, such as dropout and weight decay, reduce the effective capacity of the model and typically require the use of deeper and wider architectures to compensate for the reduced capacity. Although these techniques have been proven successful in terms of results, they seem to *waste* capacity. In contrast, data augmentation techniques reduce the generalization error by increasing the number of training examples and without reducing the effective capacity. In this paper we systematically analyze the effect of data augmentation on some popular architectures and conclude that data augmentation alone—without any other explicit regularization techniques—can achieve the same performance or higher as regularized models, especially when training with fewer examples.

1 INTRODUCTION

Regularization plays a central role in machine learning. Whereas the goal of learning algorithms is to minimize the in-sample or train error without taking into account the out-of-sample or test error, regularization techniques make the optimizer find a solution that achieves also a small test error, which is the ultimate goal of machine learning. Thus, broadly defined, regularization is any modification applied to a learning algorithm intended to prevent overfitting and improve generalization.

For a simple model with an objective function $J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})$, regularized by a norm penalty on the parameters $\alpha\Omega(\boldsymbol{\theta})$, the regularized objective function would simply be $\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) + \alpha\Omega(\boldsymbol{\theta})$ and it is straightforward to see that the only contribution to the regularization of the model is the second term. In fact, if $\alpha = 0$ the model would not be regularized.

However, in modern deep neural networks the sources of regularization are many and some of them are not explicit, but implicit (Neyshabur et al., 2014). For instance, the popular stochastic gradient descent (SGD) algorithm serves as an implicit regularizer as it tends to converge to solutions with small norm (Zhang et al., 2016). Convolutional layers can also be regarded as implicit regularizers that impose parameter sharing based on prior knowledge. Batch normalization (Ioffe & Szegedy, 2015) implicitly regularizes a network by normalizing the layer responses within each batch, which reduces the covariate shift and improves generalization. Examples of explicit regularizers are weight decay (Hanson & Pratt, 1989), which penalizes large weights; dropout (Srivastava et al., 2014), which randomly removes a fraction of the neural connections during training; or stochastic depth (Huang et al., 2016), which drops whole layers instead.

Most explicit regularizers successfully prevent overfitting by reducing the effective capacity of the models. Ironically, driven by the efficient use and development of GPUs, much research efforts have been devoted to finding ways of training deeper and wider networks of larger capacity (Simonyan & Zisserman, 2014; He et al., 2016; Zagoruyko & Komodakis, 2016). These larger models are typically regularized with weight decay and dropout, for example. However, it is known that with dropout, for instance, the gain in generalization comes at the cost of using larger models and training for longer (Goodfellow et al., 2016). Hence, it seems that with such an approach deep networks

are wasting capacity (Dauphin & Bengio, 2013). As a matter of fact, whereas traditional machine learning models require a reduction of capacity through regularization to prevent overfitting, deep neural networks seem not to need explicit regularizers to generalize well, as recently suggested by Zhang et al. (2016).

Another popular technique to improve generalization is data augmentation. Although it can also be considered a regularization technique, it differs from most explicit regularizers mainly in that it does not reduce the effective capacity of the model. Data augmentation is a very old practice in machine learning (Simard et al., 1992) and it has been identified as a critical component of many models (Ciresan et al., 2010; Krizhevsky et al., 2012; LeCun et al., 2015). However, although some authors have reported the impact of data augmentation on the performance of their models and, in some cases, a comparison of different amount of augmentation (Graham, 2014), to the best of our knowledge, the literature lacks some systematic analysis of the impact of data augmentation on different models and how it performs compared to other regularization techniques.

1.1 Our contributions

In this paper, we systematically analyze the role of data augmentation in deep neural networks for object recognition, compare it to some popular explicit regularization techniques, discuss its relationship with model capacity and test its potential to enhance learning from less training data.

1.1.1 Data augmentation and explicit regularization

Zhang et al. (2016) recently raised the thought-provoking idea that *explicit regularization may improve generalization performance, but is neither necessary nor by itself sufficient for controlling generalization error.* They came to this conclusion from the observation that turning off the explicit regularizers of a model does not prevent the model from generalizing, although the performance becomes degraded. This contrasts with traditional machine learning involving convex optimization, where regularization is necessary to avoid overfitting.

However, in Zhang et al. (2016) data augmentation is considered an explicit form of regularization similar to weight decay and dropout. We instead believe that data augmentation deserves a different classification because of some fundamental properties: Notably, data augmentation does not reduce the effective capacity of the model. Explicit regularizers are often used to counteract overfitting, but as a side effect the architecture needs to be larger and the training process longer (Krizhevsky et al., 2012; Goodfellow et al., 2016). In contrast, data augmentation increases the number of training examples—although not in an independently distributed way—and the robustness against input variability. This has the welcome side-effect of implicitly regularizing the model and improving generalization.

Here, we build upon some of the ideas and procedures from Zhang et al. (2016) and perform some experiments to assess the role of data augmentation in deep neural neural networks and in particular in contrast to explicit regularizers (weight decay and dropout). In our experiments, we consider two levels of augmentation, light and heavier, as well as no augmentation at all. Then, we test them on two popular successful network architectures: the relatively shallow all convolutional network net (Springenberg et al., 2014) and the deeper wide residual network (Zagoruyko & Komodakis, 2016), trained on CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009). Our central conclusion can be summarized as:

In a deep convolutional neural network trained with sufficient level of data augmentation, optimized by SGD, explicit regularizers (weight decay and dropout) might not provide any additional generalization improvement.

1.1.2 Data augmentation and training with fewer examples

Augmented data might be regarded as artificial and very similar to the source examples, therefore with limited contribution for making a network learn more useful representations. However, it has proven to be very useful in extreme cases such as one-shot learning, where only one or few training examples are available (Vinyals et al., 2016).

In order to provide a better insight of the usefulness of data augmentation, we train the networks with only 50 % and 80 % of the available data and test the effect of data augmentation, again in contrast to explicit regularizers. The summary of our findings in this regard can be summarized as:

When a deep neural network is trained with a subset of the training data, the heavier the data augmentation, the smaller the gap with respect to the baseline model, especially if no explicit regularization is used. Thus, data augmentation seems to serve as true data to a great extent.

1.2 Related work

Regularization is a central research topic in machine learning as it is a key component for ensuring good generalization (Girosi et al., 1995; Müller, 2012). In the case of deep learning, where networks tend to have several orders of magnitude more parameters than training examples, statistical learning theory (Vapnik & Chervonenkis, 1971) indicates that regularization becomes even more crucial. Accordingly, a myriad of tools and techniques have been proposed as regularizers: early stopping (Plaut et al., 1986), weight decay (Hanson & Pratt, 1989) and other L^p penalties, dropout (Srivastava et al., 2014) or stochastic depth (Huang et al., 2016), to name a few examples. Besides, other successful techniques have been studied for their regularization effect, despite not being explicitly intended as such. That is the case of unsupervised pre-training (Erhan et al., 2010), multi-task learning (Caruana, 1998), convolutional layers (LeCun et al., 1990), batch normalization (Ioffe & Szegedy, 2015) or adversarial training (Szegedy et al., 2013).

Data augmentation is another almost ubiquitous technique in deep learning, especially in computer vision, which can be regarded as a regularizer because it improves regularization. It was already used in the late 80's and early 90's for handwritten digit recognition (Simard et al., 1992) and it has been identified as a very important element of many modern successful models, like AlexNet (Krizhevsky et al., 2012), All-CNN (Springenberg et al., 2014) or ResNet (He et al., 2016), for instance. In some cases, data augmentation has been applied heavily with successful results (Wu et al., 2015). In domains other than computer vision, data augmentation has also been proven effective, for example in speech recognition (Jaitly & Hinton, 2013), music source separation (Uhlich et al., 2017) or text categorization (Lu et al., 2006).

Bengio et al. (2011) focused on the importance of data augmentation for recognizing handwritten digits (MNIST) through greedy layer-wise unsupervised pre-training (Bengio et al., 2007). The main conclusion of that work was that deeper architectures benefit more from data augmentation than shallow networks. Zhang et al. (2016) included data augmentation in their analysis of the role of regularization in the generalization of deep networks, although it was considered an explicit regularizer similar to weight decay and dropout. A few works have reported the performance of their models when trained with different types of data augmentation levels, as is the case of Graham (2014). Recently, the deep learning community seems to have become more aware of the importance of data augmentation and new techniques, such as cutout (Devries & Taylor, 2017) or augmentation in the feature space (DeVries & Taylor, 2017), have been proposed. Very interestingly, models that automatically learn useful data transformations have also been published very recently (Hauberg et al., 2016; Lemley et al., 2017; Ratner et al., 2017).

2 Experiments and results

This section describes the experimental setup for systematically analyzing the role of data augmentation in modern deep neural networks and presents the most relevant and interesting results.

2.1 Setup

All the experiments are performed on the neural networks API Keras (Chollet et al., 2015) on top of TensorFlow (Abadi et al., 2015) and on a single GPU NVIDIA GeForce GTX 1080 Ti.

2.1.1 Network acrchitectures

We perform our experiments on two popular architectures that have achieved successful results in object recognition tasks: the all convolutional network, All-CNN (Springenberg et al., 2014) and the

wide residual network, WRN (Zagoruyko & Komodakis, 2016). We choose these networks not only because of their effectiveness, but also because they have simple architectures, which is convenient for drawing clearer conclusions. All-CNN has a relatively small number of layers and parameters, whereas WRN is rather deep and has many more parameters.

All convolutional net. All-CNN consists of only convolutional layers with ReLU activations (Glorot et al., 2011), it is relatively shallow (12 layers) and has about 1.3 M parameters. The architecture, identical to the All-CNN-C architecture in the original paper, can be described as follows:

2×96C3(1)–96C3(2)–2×192C3(1)–192C3(2)–192C3(1)–192C1(1)
–*N.Cl.*C1(1)–Gl.Avg.–Softmax

where KC$D(S)$ is a $D \times D$ convolutional layer with K channels and stride S. *N.Cl.* is the number of classes and Gl.Avg. refers to global average pooling. We set the same training parameters as in the original paper in the cases they are reported. Specifically, in all experiments the All-CNN networks are trained using stochastic gradient descent with batch size of 128, during 350 epochs, with fixed momentum 0.9 and learning rate of 0.01 multiplied by 0.1 at epochs 200, 250 and 300. The kernel parameters are initialized according to the Xavier uniform initialization (Glorot & Bengio, 2010).

Wide Residual Network. WRN is a modification of ResNet (He et al., 2016) that achieves better performance with fewer layers, but more units per layer. Although in the original paper several combinations of depth and width are tested, here we choose for our experiments the WRN-28-10 version (28 layers and about 36.5 M parameters), which is reported to achieve the best results on CIFAR. It has the following architecture:

16C3(1)–4×160R–4×320R–4×640R–BN–ReLU–Avg.(8)–FC–Softmax

where KR is a residual block with residual function BN–ReLU–KC3(1)–BN–ReLU–KC3(1). BN is batch normalization, Avg.(8) is spatial average pooling of size 8 and FC is a fully connected layer. The stride of the first convolution within the residual blocks is 1 except in the first block of the series of 4, where it is 2 to subsample the feature maps. As before, we try to replicate the training parameters of the original paper: we use SGD with batch size of 128, during 200 epochs, with fixed Nesterov momentum 0.9 and learning rate of 0.1 multiplied by 0.2 at epochs 60, 120 and 160. The kernel parameters are initialized according to the He normal initialization (He et al., 2015).

2.1.2 DATA

We perform the experiments on the two highly benchmarked data sets CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009), which are labeled according to 10 and 100 object classes respectively. Both data sets consist of 60,000 32 x 32 color images split into 50,000 for training and 10,000 for testing. In all our experiments, the input images are fed into the network with pixel values normalized to the range $[0, 1]$ and with floating precision of 32 bits. So as to analyze the role of data augmentation, we test the network architectures presented above with two different augmentation schemes as well as with no data augmentation at all:

***Light* augmentation.** This scheme is adopted from previous works, for example (Goodfellow et al., 2013; Springenberg et al., 2014), and performs only horizontal flips and horizontal and vertical translations of 10% of the image size.

***Heavier* augmentation.** This scheme performs a larger range of affine transformations, as well as contrast and brightness adjustment. The details of the heavier augmentation scheme are the following:

Affine transformations:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} f_h z_x \cos(\theta) & -z_y \sin(\theta + \phi) & t_x \\ z_x \sin(\theta) & z_y \cos(\theta + \phi) & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

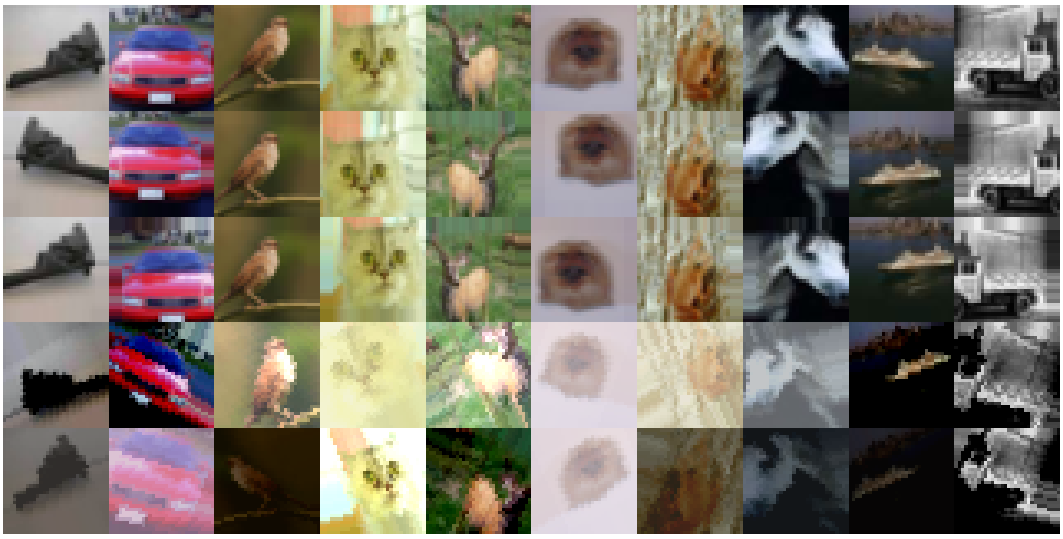

Figure 1: Extreme examples of data augmentation according to the schemes used in our experiments. The first row shows the original images, randomly selected from each CIFAR-10 class. The second and third row are two random combinations of the extreme values that the light data augmentation can take. The fourth and fifth rows are extreme transformations from the heavy data augmentation scheme. Note that these examples are very unlikely to be used during training.

Table 1: Description and range of possible values of the parameters used for the heavier augmentation. $B(p)$ denotes a Bernouilli distribution and $\mathcal{U}(a, b)$ a uniform distribution.

Parameter	**Description**	**Range**
f_h	Horizontal flip	$1 - 2B(0.5)$
t_x	Horizontal translation	$\mathcal{U}(-0.1, 0.1)$
t_y	Vertical translation	$\mathcal{U}(-0.1, 0.1)$
z_x	Horizontal scale	$\mathcal{U}(0.85, 1.15)$
z_y	Vertical scale	$\mathcal{U}(0.85, 1.15)$
θ	Rotation angle	$\mathcal{U}(-\frac{\pi}{180}22.5, \frac{\pi}{180}22.5)$
ϕ	Shear angle	$\mathcal{U}(-0.15, 0.15)$
γ	Contrast	$\mathcal{U}(0.5, 1.5)$
δ	Brightness	$\mathcal{U}(-0.25, 0.25)$

Contrast adjustment:

$$x' = \gamma(x - \overline{x}) + \overline{x}$$

Brightness adjustment:

$$x' = x + \delta$$

The description and range of values of the parameters are specified in Table 1 and examples of transformations on images from CIFAR-10 with extreme values of the parameters are provided in Figure 2.1.2.

2.2 A SUBSTITUTE FOR EXPLICIT REGULARIZATION

To analyze the role of data augmentation and test the hypothesis that it might serve as a substitute of explicit regularization techniques, we first try to replicate the results of All-CNN and WRN provided

Table 2: Test accuracy of the networks All-CNN and WRN on CIFAR-10 and CIFAR-100, comparing the performance with and without explicit regularizers and the different augmentation schemes.

Network	WD	Dropout	BN	Aug. scheme	Test CIFAR-10	Test CIFAR-100
All-CNN	yes	yes	no	no	88.35	60.54
	yes	yes	no	light	91.97	65.57
	yes	yes	no	heavier	92.44	68.62
	no	yes	no	no	87.59	46.89
	no	yes	no	light	92.01	68.01
	no	yes	no	heavier	92.18	68.40
	no	no	no	no	71.98	53.94
	no	no	no	light	90.10	63.00
	no	no	no	heavier	91.48	71.46
	no	no	yes	no	84.53	57.99
	no	no	yes	light	93.26	69.26
	no	no	yes	heavier	93.55	71.25
WRN	yes	yes	yes	no	91.44	71.67
	yes	yes	yes	light	95.01	77.58
	yes	yes	yes	heavier	95.60	76.96
	no	yes	yes	no	91.47	71.31
	no	yes	yes	light	94.76	77.42
	no	yes	yes	heavier	95.58	77.47
	no	no	yes	no	89.56	68.16
	no	no	yes	light	94.71	77.08
	no	no	yes	heavier	95.47	77.30

in the original papers, which were achieved with both weight decay and dropout. Then, we train the models without weight decay and finally without neither weight decay nor dropout. We test all these different models with the three data augmentation schemes presented above: light, heavier and no augmentation. Additionally, in the case of All-CNN we test the effect of adding batch normalization layers before each ReLU activation.

As reported by previous works (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014), when the models are trained with data augmentation, at test time better results are obtained by augmenting as well the test set. Furthermore, we have observed that when the augmentation at training time is heavier, slightly better results are obtained by applying lighter augmentation at test time. Therefore, unless mentioned otherwise, the test accuracy reported here will come from the averaged softmax posteriors over 10 random light augmentations.

The results of the different experiments are presented in Table 2. As expected, all the regularization techniques under test—weight decay, dropout and data augmentation—are successful in reducing the generalization error. However, some relevant observations can be made. Most notably, it seems that data augmentation alone is able to regularize the model as much as in combination with weight decay and dropout and in some cases it clearly achieves better performance, as in the case of All-CNN on CIFAR-100.

Furthermore, the interaction of weight decay and dropout is not always consistent, since in some cases better results are obtained with both explicit regularizers active and in other cases, only dropout achieves better generalization. However, the effect of data augmentation seems to be clearer: just some light augmentation achieves much better results than training only with the original data set and performing heavier augmentation almost always further improves the test accuracy, without the need for explicit regularization.

Not surprisingly, batch normalization also contributes to improve the generalization of All-CNN and it seems to combine well with data augmentation. On the contrary, when combined with explicit regularization the results are much worse than without it (not reported in the table so as not to clutter the results). Finally, it should be pointed out that all experiments have been performed with the same hyperparameters in order to enable fair comparisons, rather than to obtain better performance. However, we observe that when the augmentation is heavier better results are obtained by training

Table 3: Test accuracy of the networks All-CNN and WRN on CIFAR-10 and CIFAR-100 when trained with only 50 % and 80 % of the available training examples. Results in parentheses show the difference with respect to the same model trained with all available data.

Network	Pct. Data	Explicit Reg.	Aug. scheme	Test CIFAR-10	Test CIFAR-100
All-CNN	50 %	yes	no	82.33 (-6.02)	44.94 (-15.60)
		yes	light	87.37 (-4.60)	54.68 (-10.89)
		yes	heavier	88.94 (-3.50)	57.97 (-10.65)
		no	no	78.61 (-5.92)	48.62 (-9.37)
		no	light	90.21 (-3.05)	62.83 (-6.43)
		no	heavier	90.76 (-2.79)	64.41 (-6.84)
	80 %	yes	no	86.61 (-1.74)	52.51 (-8.03)
		yes	light	91.25 (-0.72)	63.24 (-2.33)
		yes	heavier	91.42 (-1.02)	65.89 (-2.73)
		no	no	83.04 (-1.49)	55.78 (-2.21)
		no	light	92.25 (-1.01)	69.05 (-0.21)
		no	heavier	92.80 (-0.75)	69.40 (-1.85)
WRN	50 %	yes	no	86.96 (-4.48)	63.60 (-8.07)
		yes	light	92.65 (-2.36)	70.83 (-7.27)
		yes	heavier	92.86 (-2.74)	70.33 (-6.63)
		no	no	85.56 (-4.00)	60.64 (-7.52)
		no	light	91.87 (-2.84)	69.97 (-7.33)
		no	heavier	92.77 (-2.70)	70.72 (-6.36)
	80 %	yes	no	90.27 (-1.17)	70.41 (-1.26)
		yes	light	94.07 (-0.94)	75.66 (-2.44)
		yes	heavier	94.57 (-1.03)	75.51 (-1.45)
		no	no	88.98 (-0.58)	66.10 (-2.06)
		no	light	93.97 (-0.74)	75.07 (-2.23)
		no	heavier	94.84 (-0.63)	75.38 (-1.70)

with a higher learning rate. For example, with a learning rate of 0.035, All-CNN, trained with the heavier augmentation and without explicit regularization on CIFAR-10, achieves 93.05 % (+1.57), even without batch normalization.

2.3 Fewer available training examples

We extend the analysis of the data augmentation role by training the same networks with fewer training examples—50 % and 80 % of the available data—and comparing the performance between different augmentation schemes. Additionally, we also analyze the combination of data augmentation and explicit regularization (both weight decay and dropout) in this case. All models are trained with the same random subset of data and tested in the same test set as the previous experiments in order to enable fairer comparisons. The results are shown in Table 3.

As expected, the performance decays with the number of available training examples. However, as the level of data augmentation increases, the difference with respect to the baseline performance (by training with all examples) significantly decreases. This indicates that data augmentation serves, to a great extent, as true data. Therefore, this confirms the effectiveness of this technique when not many training examples are available.

Furthermore, in the experiments with only 50 % and 80 % of the available data, the observations presented above become even more clear. It seems that if explicit regularization is removed, data augmentation alone better resists the lack of data. This can probably be explained by the fact that explicit regularization reduces the effective capacity, making it more difficult to take advantage of the augmented data. As a matter of fact, we observe that some of the models trained with explicit regularization underfit the data.

3 DISCUSSION AND CONCLUSION

In this work, we have presented a systematic analysis of the role of data augmentation in deep neural networks for object recognition, focusing on the comparison with popular techniques of explicit regularization. We have built upon the work by Zhang et al. (2016), in which the authors concluded that explicit regularization is not necessary, although it improves generalization performance. Here, we have shown that it is not only unnecessary, but also that the generalization gain provided by explicit regularization can be achieved by data augmentation alone.

The importance of these results lies in the fact that explicit regularization is the standard tool to enable the generalization of most machine learning methods. However, according to Zhang et al. (2016), explicit regularization plays a different role in deep learning, not explained by statistical learning theory (Vapnik & Chervonenkis, 1971). We argue instead that the theory still holds in deep learning, but one has to properly consider the very important role of implicit regularization. Explicit regularization is no longer necessary because its contribution is already provided by the many elements that implicitly regularize deep neural networks: stochastic gradient descent, convolutional layers or data augmentation, among others.

Whereas explicit regularizers, such as weight decay and dropout, succeed in mitigating overfitting by blindly reducing the effective capacity of a model, implicit regularization operates more effectively at capturing reality (Neyshabur et al., 2014). For instance, convolutional layers successfully reduce the capacity of a model by imposing a parameter sharing strategy that incorporates some essential prior domain knowledge, as well as data augmentation by transforming the training examples in a meaningful and plausible way.

In this regard it is worth highlighting some of the advantages of data augmentation, which is the main concern of this work. Not only does data set augmentation not reduce the effective capacity of the model, but it increases the number of training examples, which, according to statistical learning theories, reduces the generalization error. Furthermore, if the transformations are such that they reflect the variations of the real world, it increases the robustness of the model and it can be regarded as a data-dependent prior, similarly to unsupervised pre-training (Erhan et al., 2010). Besides, unlike explicit regularization techniques, data augmentation does not increase the computational complexity because it can be performed in parallel to the gradient updates on the CPU, making it a computationally free operation.

Deep neural networks can especially benefit from data augmentation because they do not rely on pre-computed features and because the large number of parameters allows them to shatter the augmented training set. Actually, if data augmentation is included for training, we might have to reconsider whether deep learning operates in an *overparameterization* regime, since the model capacity should take into account the amount of training data, which is exponentially increased by augmentation.

One may think that despite the advantages of data augmentation, it is a limited approach because it depends on some prior expert knowledge and it cannot be applied to all domains. However, we argue instead that expert knowledge should not be disregarded but exploited. Besides, some recent works show that it is possible to learn the data augmentation strategies (Lemley et al., 2017; Ratner et al., 2017) and future research will probably yield even better results.

Finally, it is important to note that, due to computational limitations, we have performed a systematic analysis only on CIFAR-10 and CIFAR-100, which consist of very small images. These data sets do not allow performing more agressive data augmentation since the low resolution images can easily show distortions that hinder the recognition of the object. However, some previous works (Graham, 2014; Springenberg et al., 2014) have shown impressive results by performing heavier data augmentation on higher resolution versions on CIFAR-10. We plan to extend this analysis to higher resolution data sets such as ImageNet and one could expect even more benefits from data augmentation compared to explicit regularization techniques.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 641805.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`. Software available from tensorflow.org.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pp. 153–160, 2007.

Yoshua Bengio, Arnaud Bergeron, Nicolas Boulanger-Lewandowski, Thomas Breuel, Youssouf Chherawala, Moustapha Cisse, Dumitru Erhan, Jeremy Eustache, Xavier Glorot, Xavier Muller, et al. Deep learners benefit more from out-of-distribution examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 164–172, 2011.

Rich Caruana. Multitask learning. In *Learning to learn*, pp. 95–133. Springer, 1998.

François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

Dan Claudiu Ciresan, Ueli Meier, Luca Maria Gambardella, and Jrgen Schmidhuber. Deep big simple neural nets excel on handwritten digit recognition. *CoRR*, abs/1003.0358, 2010. URL `http://dblp.uni-trier.de/db/journals/corr/corr1003.html#abs-1003-0358`.

Yann N Dauphin and Yoshua Bengio. Big neural networks waste capacity. *arXiv preprint arXiv:1301.3583*, 2013.

Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. URL `http://arxiv.org/abs/1708.04552`.

Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *JMLR W&CP: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, volume 9, pp. 249–256, may 2010.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

Benjamin Graham. Fractional max-pooling. *arXiv preprint arXiv:1412.6071*, 2014.

Stephen José Hanson and Lorien Y Pratt. Comparing biases for minimal network construction with back-propagation. In *Advances in neural information processing systems*, pp. 177–185, 1989.

Søren Hauberg, Oren Freifeld, Anders Boesen Lindbo Larsen, John Fisher, and Lars Hansen. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Artificial Intelligence and Statistics*, pp. 342–350, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, pp. 646–661. Springer, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

Navdeep Jaitly and Geoffrey E Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, pp. 625–660, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pp. 396–404, 1990.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Joseph Lemley, Shabab Bazrafkan, and Peter Corcoran. Smart augmentation-learning an optimal data augmentation strategy. *IEEE Access*, 2017.

Xinghua Lu, Bin Zheng, Atulya Velivelli, and ChengXiang Zhai. Enhancing text categorization with semantic-enriched representation and training data augmentation. *Journal of the American Medical Informatics Association*, 13(5):526–535, 2006.

Klaus-Robert Müller. Regularization techniques to improve generalization. In *Neural Networks: Tricks of the Trade*, pp. 49–51. Springer, 2012.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *CoRR*, abs/1412.6614, 2014. URL `http://arxiv.org/abs/1412.6614`.

D.C. Plaut, S.J. Nowlan, and G.E Hinton. Experiments on learning by back propagation. 1986.

Alexander J Ratner, Henry R Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. *arXiv preprint arXiv:1709.01643*, 2017.

Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop-a formalism for specifying selected invariances in an adaptive network. In *Advances in neural information processing systems*, pp. 895–903, 1992.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji. Improving music source separation based on deep neural networks through data augmentation and network blending. *Submitted to ICASSP*, 2017.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probab. and its Applications*, 16(2):264–280, 1971.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3630–3638. Curran Associates, Inc., 2016. URL `http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning.pdf`.

Ren Wu, Shengen Yan, Yi Shan, Qingqing Dang, and Gang Sun. Deep image: Scaling up image recognition. *arXiv preprint arXiv:1501.02876*, 7(8), 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.