# Reducing Overfitting and Improving Generalization in Training Convolutional Neural Network (CNN) under Limited Sample Sizes in Image Recognition

Panissara Thanapol *, Kittichai Lavangnananda †, Pascal Bouvry *, Frédéric Pinel * and Franck Leprévost *

* Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

† School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

Email: panissara.thanapol@uni.lu; Kitt@sit.kmutt.ac.th; pascal.bouvry@uni.lu; frederic.pinel@uni.lu; franck.leprevost@uni.lu

*Abstract*— **In deep learning, application of Convolutional Neural Network (CNN) is prolific in image recognition. CNN assumes that large amount of samples are available in the dataset in order to implement an effective CNN model. However, this assumption may not be practical or possible in some real world applications. It is commonly known that training a CNN model under limited samples available often leads to overfitting and inability to generalize. Data augmentation, batch normalization and dropout techniques have been suggested to mitigate such problems. This work studies the effect of overfitting and generalization in image recognition of intentionally contracted CIFAR-10 dataset. Application of these techniques and their combination are considered as well as injection of data augmentation at different epochs. The result of this work reveals that utilizing injection at 30 epoch in the application of width and height shift data augmentation together with dropout yields the best performance and can overcome the overfitting effect best.**

***Keywords—CIFAR-10 Dataset, Convolution Neural Network (CNN), Generalization, Image Recognition, Overfitting***

## I. Introduction

Deep Learning, especially Convolutional Neural Network (CNN), has become one of the most popular tool image classification and recognition. Nevertheless, training CNN usually requires large number of samples for satisfactory outcome [1]. In practice, however, obtaining sufficient number of samples may not be possible or even undesirable. For example, having sufficiently high number of images of defected components in implementing a recognition system implies that the quality control is unacceptable. This is also problematic particularly in some medical fields, for example classification of tumour [2], [3]. Therefore, implementing image recognition using CNN under relatively limited samples available receives a lot of attention recently.

Data Augmentation, a well known technique in deep learning, has been invented to compensate the training under small size dataset by generating artificial images from existing images [4]. These artificial images are created in several ways, such as shifting, altering of width and height, etc. of original images. Another drawback of training under limited samples available is the effect of what is commonly known as overfitting [5].

While the data augmentation technique increases the quantity of dataset and also reduces the overfitting, it unintentionally introduces another drawback that the CNN implemented is unable to generalize the recognition process and hence is unable to recognize unencountered images well. In order to overcome this, techniques known as dropout and batch normalization have been introduce to improve the generalization ability of CNN [6].

This work is concerned with the studies of utilizing data augmentation, dropout and batch normalization techniques in training CNN in image recognition of a well known dataset available in a public domain website [7]. The number of images are intentionally reduced in order to create sitauation where number of samples is imited and insufficient. The work introduces an approach to mitigate such situation by suggesting suitable combination of data augmentation and dropout techniques as well as suggests a suitable epoch to inject data augmentation.

## II. Literature Review

Convolutional Neural Network (CNN) is widely used for image recognition and classification due to its ability to automatically extract useful and particular features. The CNN architecture known as ResNet-34 [8] was implemented in classification of old polish cars with the highest accuracy of 99.18%. CNN has also been applied in medical field, an example of which includes the classification of breast cancer [9] with 97% accuracy.

In attempt to overcome the limited samples available, several approaches had been applied. The work in [10] simply duplicated samples in the limited category to increase the sample size. A more technical approach is the Generative Adversarial Networks (GAN) where more samples are increased by having a generative model which is capable of learning from input samples [11]. GAN-Based data augmentation was further implemented by considering bias of a fake data. This allowed both distributions of the existing data and the generating data to be similar [12].

Application of data augmentation in training CNN is plentiful due to the importance of overcoming both limited samples available as well as overfitting. In recognition of hand writing digits and images [13], the differential data augmentation techniques implemented managed to improve accuracy with different degrees of success in each type of augmentation. Similar work which adopted the random eraser technique was carried out on the image dataset mentioned [14]. Data augmentation was also applied in other procedure, the work in [15] proposed the way to improve accuracy by learning from random images in mini batch and then apply the data augmentation technique with the rest, this resulted in an error rate reduction of 0.6%.

978-1-7281-6694-0/20/$31.00 ©2020 IEEE

Dropout [16] is another popular technique to regularize the model by random omitting some hidden unit within each layer during the training. The objective of the dropout is train the network so that the output is not dominated by some hidden units, an hence alleviates the overfitting. Several dropout approaches also exist. A popular dropout technique is the regularization based approach [17], the approach dropped the hidden unit ruled by finding the correlation of each other hidden unit. It was proven superior to the traditional dropout. [18] and [19], have advanced the studies of dropout technique and were able to regularize the model, which in turn, reduced the overfitting. However, there have been studies which suggested that dropout ought not be used together with data augmentation [20].

Batch Normalization [20] is a technique to conduct the covariate shift in neural network due to the change of parameter during the training. The objective is to ensure a stable distribution of the layer by normalizing the set of activation in mini-batch. It also has another effect in speeding up the training to propel a convergence. Batch normalization was proven to control stability in the neural network training process better than the conventional weight normalization [21]. In classification using CNN, apart from the accuracy, another important aspect is the generalization ability (i.e. the ability to correctly identify unseen images) [22]. This ability was emphasized and has been used in assessing deep learning model in [23]. This was also be demonstrated by means of the learning curve in [24].

In most previous applications, CNNs were implemented with specific objectives, while several researches have also come up with new and novel techniques to overcome drawbacks of CNN from different perspectives. This work is an attempt to studies existing techniques that intended to overcome drawbacks originated from insufficient number of samples available in training CNN to recognize images, by utilizing existing techniques together.

## III. Methodology

In this study, the probably most popular dataset available in the public domain known as CIFAR-10 dataset [7] is selected. This dataset is widely used in experimentation in image processing, not just in deep learning and CNN. One factor that attributes to its popularity is the large quantity of samples within the dataset. CIFAR-10 is a collection of images, where each image comprises 32 by 32 pixels. The dataset contains 10 classes; airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. There are 6,000 images in each class (5,000 are set aside for training set and another 1,000 for testing set). Samples in CIFAR-10 were also thoroughly processed where duplicated images and incomplete images (i.e. containing some white pixels) are eliminated.

As the objective of this study is to overcome the insufficient samples available for training the CNN. Five datasets are created, they are contracted version of the original CIFAR-10. Each new dataset contains 10% of samples in the original. Hence, 600 samples of each class are randomly selected (500 from training set and 100 from test set) to form a 10-category 6,000 sample dataset.

In order to investigate the effect of different data augmentation, dropout and batch normalization techniques to the classification improvement, a CNN model is implemented and trained under limited number of samples as stated earlier. It comprises 5 layers. Input layer consists of $x_1$ to $x_{3072}$ (32 by 32 (number of pixels in an image)) by 3 (for channels red, green and blue colours)). Output layer consists of $y_1$ to $y_{10}$ representing probabilities of the input image belonging to each category. There are 3 hidden layers, the first 2 layers are a combination of convolution and max. pooling for extracting useful features in an image. The last hidden layer is a fully connected layer with 256 nodes. Fig. 1 depicts the model of the CNN architecture in this work.

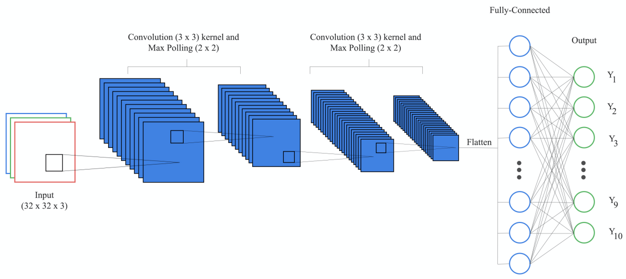
Fully-Connected
Convolution (3 x 3) kernel and Max Polling (2 x 2)
Convolution (3 x 3) kernel and Max Polling (2 x 2)
Output
$Y_1$
$Y_2$
$Y_3$
Flatten
$Y_9$
$Y_{10}$
Input (32 x 32 x 3)

Fig. 1. The architecture of the implemented CNN model

For validity of the work, the five datasets created are used for training and testing the CNN model, each is carried out for 150 epochs. As expected, the classification accuracy obtained is rather low with the average test accuracy of 49.8%. Overfitting is also apparent, this is described and depicted in Section IV. The following Sub-Sections describe the utilization of different techniques and their combination in this work.

### *A. Applying a Combination of Data Augmentation Techniques to CNN*

Numerous types of data augmentation techniques exist. This work follows the four types of data augmentation as suggested in [14]. These are summarized as follows:

- Rotation in range of 25 degrees
- Width and height shift in range of 0.225 degrees
- Shear in range of 0.2 degrees
- Random eraser at 0.5 degrees of each image

Once the CNN model is identified, applications of a single data augmentation and all possible combinations are then studied. The application of width and height shift augmentation yields the best result while the application of random erase yields the poorest. Their training, validation and test accuracies (average values obtained from using the 5 datasets) are shown in Fig. 2 in Section IV where results from all studies are included for ease of comparison.

### *B. Utilizing a Combination of Data Augmentation Techniques together with Dropout in training the CNN*

While the range of omitting rate (i.e. the probability for omitting the hidden unit) is 0 to 1 in theory, the values too close to 0 or 1 are seldom used. Popular range for omitting rate is between 0.5 to 0.8 [16]. This work had investigated several omitting rates within this range. The results reveal an interesting fact. The omitting rate of 0.5 achieves the best generalization to the CNN model, while the omitting rate of 0.7 is the best value for reducing the overfitting effect. Fig. 3 in Section IV illustrates the training, validation and test accuracies (average values obtained from using the 5 datasets) in this study.

### *C. Utilizing a Combination of Data Augmentation Techniques together with Batch Normalization in training the CNN*

The application of batch normalization is to normalize the activation in the hidden layer as well as to regularize the model. The technique enables the mean value of the input close to zero and variance close to one. This process corresponds with two steps that is to subtract the output of a previous activation layer by batch mean value and divide by the batch standard deviation. In this work, the normalizing layer is adopted before non-linear activation function. This technique is applied in the training of the CNN model in order to improve the generalization. Fig. 4 in Section IV illustrates the training, validation and test accuracies (average values obtained from using the 5 datasets) in this study.

It is worth mentioning here that the work in [20] is quite significant to this work, it concludes that batch normalization and dropout ought not be applied simultaneously in training a CNN as their results were detrimental rather than beneficial to the improvement of the classification accuracy. Therefore, this work avoids application of the combination of these two techniques accordingly.

### *D. Utilizing a Combination of Data Augmentation Techniques together with Batch Normalization and Data Augmentation Techniques together with Dropout by Injecting During Training of the CNN*

It had been suggested at application of data augmentation techniques are a lot more effective if they are injected during the training instead of at the initial stage. This work follows the work in [14] where injections were investigated at 30, 60, and 90 epoch. The above studies (Sections III *C* and III *D*) are adapted so data augmentations are injected at different epochs as stated above. The results reconfirm the work in [14] that injecting data augmentation at 30 epoch during the training is the most effective. Figs. 5 and 6 in Section IV illustrate the training, validation and test accuracies (average values obtained from using the 5 datasets) in this two part of studies.

## IV. Result

This Section shows all results in the studies. All accuracy values are the average obtained from using the 5 datasets. From CNN perspective, the overfitting effect can be observed from the discrepancy between the percentage of training loss and the percentage of validation loss. Examples of overfitting and the discussion are in the next Section. The most commonly used for performance verification of a CNN is the test accuracy.

With respect to the application of data augmentation techniques and their combination, this study has carried out every possible combinations of the three data augmentations. The training, validation and test accuracies are shown in Fig 2.

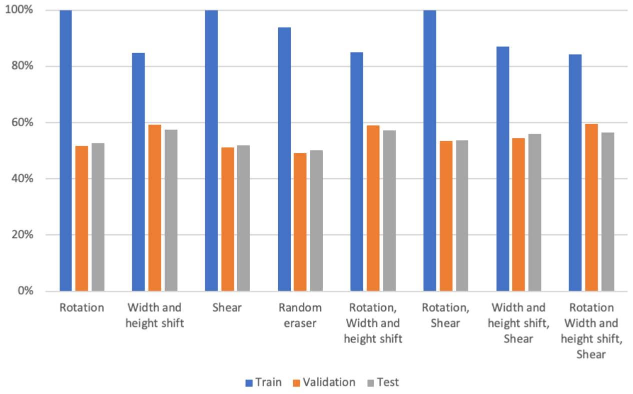

Fig. 2. Comparison of Different Combination of Data Augmentation Techniques with CNN

Referring to Fig. 2, the application of width and height shift yields the best improvement with the test accuracy of 57.5%. while random erase yields the lowest. Replacing several pixels in the original image with non informative colour rather than their actual in colour random erase augmentation may lead to significant loss of important feature(s). The fact that limited samples are available, this may amplify this further. The application of this augmentation has, nevertheless, improved the test accuracy marginally. As stated earlier, dropout and batch normalization are applied together with all four types of data augmentation. However, result of any combination with random eraser yields rather poor performance. For simplicity and clarity in comparison, result of any combination with random eraser is omitted.

The application of data augmentation together with dropout results in better overall performance. The best performance is the combination of width and height shift and shear augmentations with the test accuracy of 58.8%. This indicates the advantage in utilizing dropout. Results of the application of this combination are shown in Fig. 3.

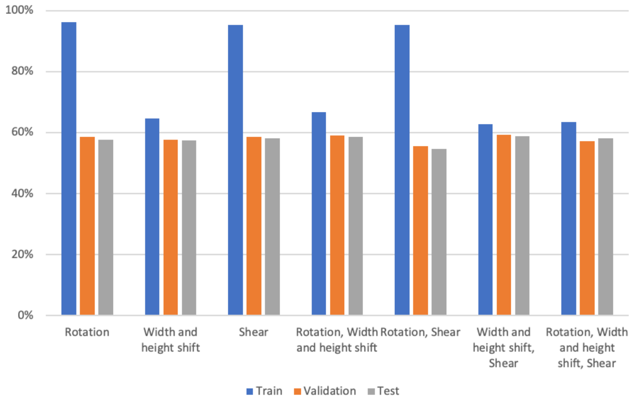

Fig. 3. Comparison of Different Combination of Data Augmentation Techniques with CNN together with Dropout

The application of data augmentation together with batch normalization shows marginal improvement in some data augmentation. The best performance is the combination of width and height shift and rotation augmentations with the test accuracy of 58.3%. Results of the application of this combination are shown in Fig. 4.

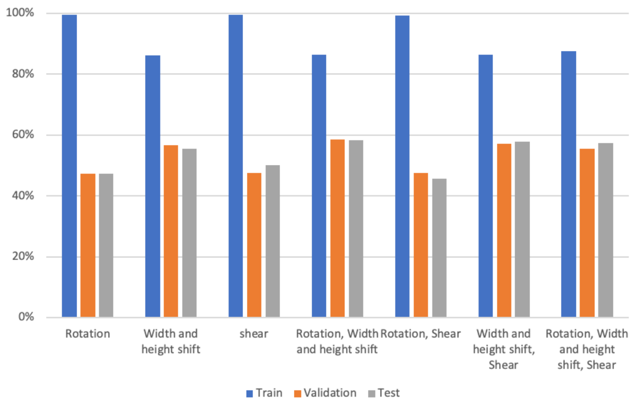

Fig. 4. Comparison of Different Combination of Data Augmentation Techniques with CNN together with Batch Normalization

As mentioned in Section III *D*, the best results were obtained with injection at 30 epoch, hence results and further discussion are focused to the performance achieved from injecting at 30 epoch due to limited space available and also avoiding confusion from unsubstantial results.

Utilization of injecting at 30 epoch with the application of data augmentation together with batch normalization does not improve the overall performance. This is, by no means, a contrast to the advantage of batch normalization as reported in [20] and [21], as previous work were not under limited number of samples and the objective of the works were not about overcoming the overfitting. Results of utilization of injecting at 30 epoch and the application of this combination are shown in Fig. 5.

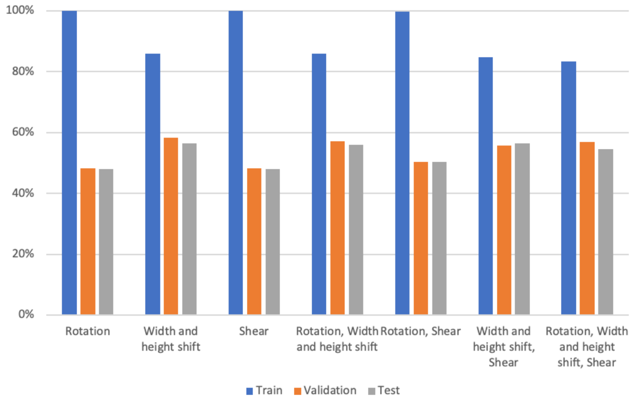

Fig. 5. Comparison of Different Combination of Data Augmentation Techniques by Injecting at 30 Epoch to CNN together with Batch Normalization

Utilization of injecting at 30 epoch with the application of data augmentation together with dropout yields the best overall performance improvement, when compared with results in Fig. 2. The best performance in the utilization of injecting happens with the application of width and height shift augmentation alone with the test accuracy of 61.5%. The results suggest that injecting at 30 epoch is beneficial, especially in the application of width and height shift augmentation together with dropout. Results of utilization of injecting at 30 epoch and the application of this combination are shown in Fig. 6.

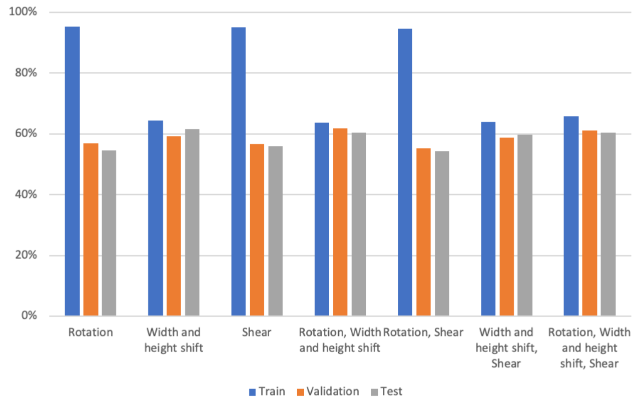

Fig. 6. Comparison of Different Combination of Data Augmentation Techniques by Injecting at 30 Epoch to CNN together with Dropout

## V. Reduction of Overfitting and Improvement of Generalization

In deep learning using CNN, there usually exists a ‘loss function’ in order to updated the weight for the further training. The term loss, refers to a summation of an error is determined by the difference between the predicted value from the model and the actual value. Hence training loss and validation loss represent summation of errors using training set and validation set respectively. A low discrepancy between satisfactory training accuracy and test accuracy is a good indication of a good generalization. As stated earlier, overfitting can then be observed by noting the discrepancy between the training loss and validation loss. This discrepancy also reflects the generalization ability too.

It is impractical to display these two effects of each combination investigated in this work. Apart from occupying too much space, the main message of the work may not be apparent too. Therefore, the best result of the data augmentation technique (i.e. application of width and height shift) is selected to for the discussion. Figs. 7 to 12 depict overfitting and generalization in the six studies in this work. They represent overfitting and level of generalization in application of CNN alone, application of data augmentation with CNN, application of data augmentation with CNN together with batch normalization, application of data augmentation with CNN together with dropout, utilization of injecting at 30 epoch and the application of data augmentation with CNN together with batch normalization and utilization of injecting at 30 epoch and the application of data augmentation with CNN together with dropout respectively.

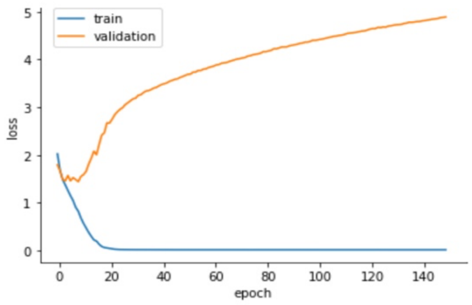

Fig. 7. Overfitting and Low Generalization in CNN Alone

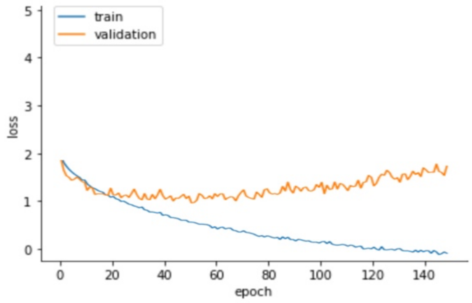

Fig. 8. Reduction of Overfitting and Improvement of Generalization in application of **width and height shift** with CNN

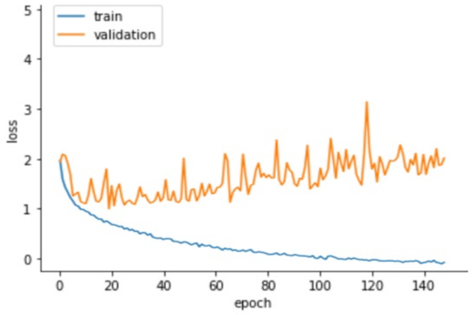

Fig. 9. Reduction of Overfitting and Improvement of Generalization in application of **width and height shift** with CNN together with **batch normalization**

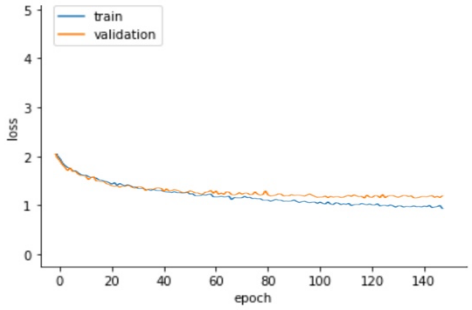

Fig. 10. Reduction of Overfitting and Improvement of Generalization in application of **width and height shift** with CNN together with **dropout**

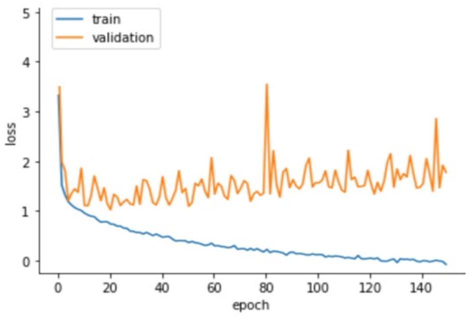

Fig. 11. Reduction of Overfitting and Improvement of Generalization in utilization of **injecting at 30 epoch** and the application of **width and height shift** with CNN together with **batch normalization**

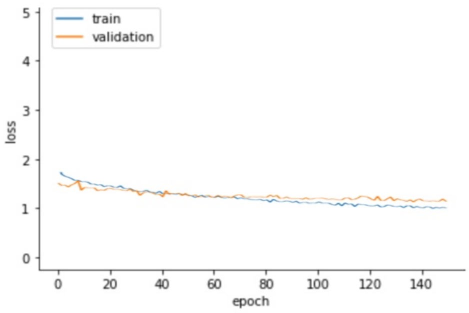

Fig. 12. Reduction of Overfitting and Improvement of Generalization in utilization of **injecting at 30 epoch** and the application of **width and height shift** with CNN together with **dropout**

The heights of y-axis in Figs. 7 to 12 are kept the same for ease of comparison. As can be expected, the effect of overfitting is very apparent in CNN alone. The application of data augmentation manages to reduce the overfitting quite significantly. Introduction of batch normalization seems to introduce more instability while dropout is beneficial in this application. The best performance is the utilization of injecting at 30 epoch and the application of data augmentation with CNN together with dropout with least overfitting effect as shown in Fig. 12. This also reaffirms the work in [14] which recommended injecting at 30 epoch.

## VI. Conclusion

If accuracy is the metric of this work, then it may seem less satisfactory as the best test accuracy is about 61.5%. However, it must be re-emphasized that the main objective of this work is to investigate the existing techniques and identify a strategy to overcome overfitting effect in training CNN model in image recognition under insufficient samples available. This work suggests that using width and height shift data augmentation by injecting at 30 epoch together with dropout is a promising approach which mitigates the overfitting and generalization. While it is arguable that suitable model configuration and characteristic of the dataset have important role in dealing with CNN training under limited samples. This finding in this work ought to be applicable in similar image recognition problems. Nevertheless, it ought not be seen as a panacea to this unique problem. In a situation where too few samples are available, implementing a CNN model with any level of satisfactory performance may not be possible at all.

## VII. Future Work

Future work can be carried out in several directions. A good candidate is to apply this approach to other image datasets available in public websites. Other domain such as signal processing or natural language understanding deserves a similar study. The test function to determine errors in a CNN may be fine tuned for an optimal CNN performance.

The structure of a suitable CNN is also vital to satisfactory performance. This is an area where this work can further explore. A very recent study [25] may pave way to implementing of an even more efficient CNN. Also Recurrent Neural Network should not be overlooked for in this type of problem too.

## Acknowledgment

The authors wish to acknowledge The University of Luxembourg for providing full Scholarship to Miss. P Thanapol. They are grateful for the computing facility support from The School of Information Technology (SIT), King Mongkut's University of Technology Thonburi (KMUTT), Thailand throughout this research.

## References

[1] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?," *arXiv preprint arXiv:1511.06348*. 2015.

[2] C. N. Vasconcelos and B. N. Vasconcelos, "Convolutional Neural Network Committees for Melanoma Classification with Classical And Expert Knowledge Based Image Transforms Data Augmentation," *arXiv preprint arXiv:1702.07025*. 2017.

[3] T. Shaikhina and N. A. Khovanova, "Handling limited datasets with neural networks in medical applications: A small-data approach," *Artif. Intell. Med.*, vol. 75, pp. 51–63, 2017.

[4] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019.

[5] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *arXiv preprint arXiv:1712.04621*. 2017.

[6] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for Deep Learning: A Taxonomy," *arXiv preprint arXiv:1710.10686*. 2017.

[7] Krizhevsky, Alex, and G. Hinton, "Learning Multiple Layers of Features from Tiny Images," 2009.

[8] E. Zawadzka-Gosk, K. Wołk, and W. Czarnowski, "Deep Learning in State-of-the-Art Image Classification Exceeding 99% Accuracy," in *World Conference on Information Systems and Technologies*, 2019, pp. 946–957.

[9] B. Wei, Z. Han, X. He, and Y. Yin, "Deep learning model based breast cancer histopathological image classification," in *2017 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2017*, 2017, pp. 348–353.

[10] H. H. Alam, M. M. Rahoman and M. K. A. Azad, "Sentiment Analysis for Bangla Sentences using Convolutional Neural Network", in *2017 International Conference of Computer and Information Technology, ICCIT 2017*, 2017.

[11] A. Antoniou, A. Storkey, and H. Edwards, "Data Augmentation Generative Adversarial Networks," *arXiv preprint arXiv:1711.04340*. 2017.

[12] H. Mengxiao and J. Li, "Exploring Bias in GAN-based Data Augmentation for Small Samples," *arXiv preprint arXiv:1905.08495*. 2019.

[13] C. Lei, B. Hu, D. Wang, S. Zhang, and Z. Chen, "A preliminary study on data augmentation of deep learning for image classification," in *Proceedings of the 11th Asia-Pacific Symposium on Internetware*, 2019, pp. 1–6.

[14] S. O'Gara and K. McGuinness, "Comparing Data Augmentation Strategies for Deep Image Classification," in *Irish Machine Vision and Image Processing Conference (IMVIP)*, 2019.

[15] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 113–123.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] J. Ba and B. Frey, "Adaptive dropout for training deep neural networks," in *Advances in Neural Information Processing Systems*, 2013, pp. 3084–3092.

[18] H. Wu and X. Gu, "Towards dropout training for convolutional neural networks," *Neural Networks*, vol. 71, pp. 1–10, 2015.

[19] S. Park and N. Kwak, "Analysis on the Dropout Effect in Convolutional Neural Networks," *ACCV 2016 Comput. Vis. – ACCV 2016 pp*, pp. 189–204, 2016.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, pp. 448–456.

[21] I. Gitman and B. Ginsburg, "Comparison of Batch Normalization and Weight Normalization Algorithms for the Large-scale Image Classification," *arXiv preprint arXiv:1709.08145*. 2017.

[22] R. Roelofs, "Measuring Generalization and Overfitting in Machine Learning," *Doctoral dissertation, UC Berkeley*. 2019.

[23] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning," *Adv. Neural Inf. Process. Syst.*, pp. 5947–5956, 2017.

[24] M. J. Anzanello and F. S. Fogliatto, "Learning curve models and applications: Literature review and research directions," *Int. J. Ind. Ergon.*, vol. 41, no. 5, pp. 573–583, 2011.

[25] F. Pinel *et al.*, "Evolving a deep neural network training time estimator," in *Proceedings of the 2020 Int. Conference on Optimization and Learning (OLA'20)*, 2020, pp. 13–24.