GridMask Based Data Augmentation For Bengali Handwritten Grapheme Classification

Jiayue Yang
u6399075@anu.edu.au
Australian National University
Canberra ACT 0200, Australia

ABSTRACT

In this paper, we describe the deep learning-based Bengali handwritten grapheme classification. Specifically, our recognition approach is based on the convolutional neural networks (CNNs) as deep CNNs have achieved splendid performance on many different visual recognition tasks. Moreover, we employ GridMask-based data augmentation to improve the recognition performance further. We compare the GridMask-based data augmentation with conventional data augmentations (such as flip, rotation, mixup) on three widely-used CNN architectures: ResNet101, DenseNet169 and EfficientNet B0. Extensive experiments demonstrate GridMask can utilize the information removal to improve the robustness of the neural networks, and the boost of hierarchical macro-averaged recall on the validation set suggest that GridMask data augmentation can be efficiently used for the Bengali handwritten grapheme analysis without any prior grapheme segmentation.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Classification and regression trees**.

KEYWORDS

data augmentation, grapheme recognition, deep convolutional neural network

ACM Reference Format:
Jiayue Yang. 2020. GridMask Based Data Augmentation For Bengali Handwritten Grapheme Classification. In *2020 2nd International Conference on Intelligent Medicine and Image Processing (IMIP 2020), April 23–26, 2020, Tianjin, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3399637.3399650

1 INTRODUCTION

As the 5th most spoken language in the world, Bengali presently has hundreds of million of native-speakers. Moreover, as the official language of Bangladesh, Bengali is also the second most spoken language in India. Due to the widely reach of this language, there's increasing interest to devolve a recognition system which that can automatically recognize images of the language handwritten. However, it is widely known that optical character recognition (OCR) is particularly difficult for Bengali. For the Bengali handwritten grapheme classification, there is a large number of unlabeled data. How to fully utilize the data and improve the recognition performance is a challenge. Recently, Kaggle [1] competition platform organized a handwritten grapheme classification competition, with the goal to stimulus the development of Bengali handwritten grapheme recognition, and accelerate the research to improve Bengali language understanding and interpretation. In this paper, we describe our solution for the Bengali handwritten recognition tasks, which employ deep neural networks and data augmentation techniques [11].

Indeed, sustainable efforts has been made using image feature extraction and machine learning methods through previous efforts. Various features were used for images recognition, which include linear discriminant analysis (LDA) [18], support vector machine (SVM) [13], k-nearest neighbors (KNN) algorithm [10]. Unfortunately, the handcrafted features cannot identify all the pattern in the handwritten images. Consequently, there is relatively limitation for practical applications of recognition system.

Recently, deep learning techniques have witnessed a profound success on many different challenge tasks, such as image classification and recognition [8], speech recognition [7], natural language processing. Just a few attempts have been made to improve the Bengali grapheme classification, and text-to-speech task [5, 9]. In this paper, we explore the deployment of deep neural networks for the Bengali grapheme classification task. Specifically, deep convolutional neural network is trained, due to its excellent performance for many visual recognition tasks. However, it is widely known that deep convolutional neural networks are easy to fall into local optimal, which decrease the performance of deep learning. One solution to solve the issue is to increase the labeled data, which is not easy to be obtained in more piratical settings. Another way is to artificially enlarge the dataset, such as data augmentation. To improve the generalization abilities, sustainable efforts have been made for the data augmentation, such as flipping [12], random crop [7] and random rotation. In the paper, we explore the novel data augmentation approach named GridMask [3], to improve the Bengali handwritten grapheme classification performance.

The rest of the paper is organized as follows. The related data augmentation approaches are outlined in Section 2. While the used image datasets and employed data augmentation approach are given in the section 3. Section 4 presents the training procedure, and presents the results. Conclusions, discussions and future perspectives are given in section 5.

IMIP 2020, April 23–26, 2020, Tianjin, China

ACM ISBN 978-1-4503-7779-9/20/04…$15.00
DOI:https://doi.org/10.1145/3399637.3399650

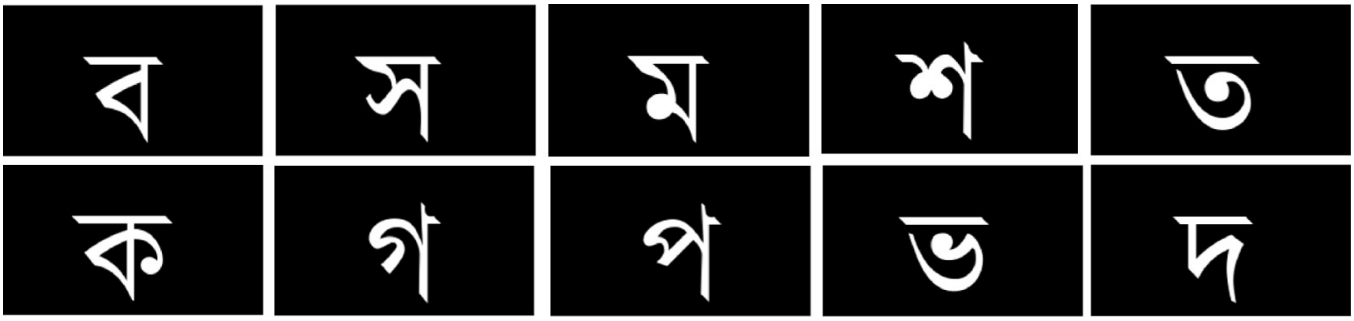

Figure 1: Samples of the Bengali handwritten grapheme.

2 RELATED WORK

Data augmentation is one important regularization approach to improve the generalization ability of neural network, which can greatly prevent over-fitting in the training of networks. Generally speaking, data augmentation aims to increase the training data size by creating artificially labeled samples, as more data can be trained to reduce the performance gap between the train and valid set. Previous attempts for data augmentation are based on transformation of the original image, such as, affine transformation, adding random noise, regions dropout, contract changes, blurring, et al.

Latter, Mixup [19] approach has been proposed to leverage multi-image information to train the model, and achieves better performance with comparison to previous approaches. In more detail, Mixup can create new samples using interpolation between different labeled samples. Unlike previous processing the label using one-hot encoder, Mixup is not within the label-preserving manner, as the new labels for new synthetically-created samples do not belong to two classes [17]. Mixup-based data augmentation approach can be represented by:

$$\begin{aligned} x_{new} &= \lambda * x_i + (1 - \lambda) * x_j \\ y_{new} &= \lambda * y_i + (1 - \lambda) * y_j \end{aligned} \tag{1}$$

where λ is the mixing proportion which follows the $Beta(\alpha, \alpha)$ distribution. x is the input image while y is the label. In our experiments, the α is set as 0.4 for Mixup-based approach.

Similarly, Sample pairing was proposed in [6], which can create a new sample by averaging of two labeled inputs. However, the label is the same as the first sample. Hence, Sample pairing can be represented as:

$$\begin{aligned} x_{new} &= 0.5 * x_i + 0.5 * x_j \\ y_{new} &= y_i \end{aligned} \tag{2}$$

3 METHODOLOGY

3.1 Dataset

The dataset employed in this study consists of the images of individual hand-written Bengali characters. Specifically, the handwritten Bengali characters (graphemes) contains three components: a grapheme_root, vowel_diacritic, and consonant_diacritic. It is worthwhile to notice that: though, in the alphabet Bengali only 11 vowels and 38 consonants, there are also 18 potential diacritics (or accents). This leads to the fact that there are many more grapheme in the practical setting, or the smallest units in a written language. Moreover, it is found that there can be roughly 10,000 possible graphemes which should be classified. As some grapheme occur very few times, this added to more difficulties for the recognition task due to the class-imbalance issues. 200,840 labeled samples are used in our study, in while 20% of them are random selected to evaluate the performance of proposed approach. Samples of the grapheme are depicted in Figure 1.

3.2 Convolutional neural network

As an eminent architectures of deep learning, deep convolutional neural network has demonstrates great potentials on different visual recognition tasks. Generally speaking, CNN can be viewed as the feed-forward neural networks, which is quite similar to other neural network architectures. Recently, CNN has became a vital approach for representation learning.

Since last several years, sustainable efforts have been made [4] to improve the CNN architecture design and numerous variants of CNN architectures are proposed in the literature, such as, VGGNet [16], Inception (GoogLeNet) [14], ResNet, DenseNet huang2017densely, EfficientNet [15], whose has similar components have been proposed over last few years. Specifically, CNN contains the convolution operations and pooling operations (for example: maxpoolings and mean poolings). The goal of pooling operations is to decrease the time consumed for computation, and to build up further spatial the pooling operations for convolutional layer.

For example, ResNet presents micro-architecture modules (a building block is given in Figure 2), which refers to the set of blocks used to construct the network. The success of ResNet also demonstrate that extremely deep neural networks can be trained, using standard stochastic gradient descent through the use of residual modules. However, the basic components of CNN are very similar, which have typically convolutional layers and pooling layers . In this paper, we employ three widely-used CNN architectures for the recognition task: ResNet101, DenseNet169 and EfficientNet B0.

3.3 GridMask data augmentation

As aforementioned, CNNs have millions of parameters to be trained, thus, the training process often require lots of data. Data augmentation can be used to create more useful data from existing samples

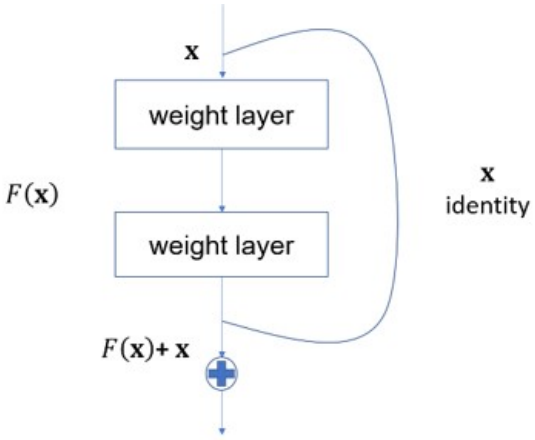

Figure 2: Residual learning: a building block.

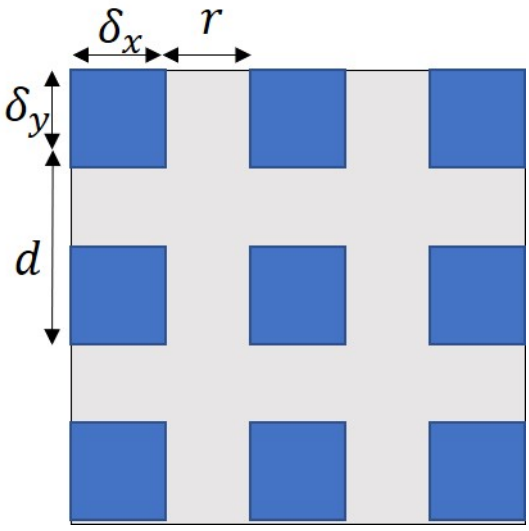

Figure 3: Sample of one unit of the mask (as depicted by the dotted square.)

for training practical sittings and improve the general generality ability of CNNs. In this paper, we explore the GridMask-based data augmentation, which is a simple but effective strategy.

Suppose the input image can be represented as x, and $x \in R_{H\times W\times C}$. where H,W,C is the height, width and channel of the image respectively. GridMask explores to randomly removes some pixels of the image and the augmentation can be represented as follows:

$$\overline{x} = x \times M \tag{3}$$

where the x is the input image with the defined size, and M is the binary mask. Different from previous other methods, GridMask is not aim to remove a continuous region, however, GridMask approach explore to remove a region with disconnected pixel sets (as given in Figure 3.). Specifically, a unique M can be represented by $(r, d, \delta_x, \delta_y)$, and the definition of the hour parameters are depicted in the Figure. r is the ratio in a unit (the shorter gray edge). d is the unit's length, while δ_x) and δ_y) are the distances between first intact unit and the boundary of the image respectively [3].

The samples of the augmented dataset are given in the Figure 4 (with different M settings). And the augmented dataset will be used for the recognition task.

4 EXPERIMENTAL RESULTS

We conduct a series of experiments to validate the performance of GridMask-based augmentation for the Bengali hand written grapheme recognition task, using three different widely used CNN architectures.

4.1 Evaluation metrics

The hierarchical macro-averaged recall is used to evaluate the performance of the trained models. We calculate the standard macro-averaged recall for each component which includes grapheme root, vowel diacritic and consonant diacritic. The final evaluation is weighted average of those three scores, and we give the grapheme root double weight to highlight the importance of it.

4.2 Training schedule

We use Pytorch to implement the network architecture design and the Albumentations python library [2] toolkit package to implement the data augmentation approaches. For all of our experiments, we employ an NVIDIA GeForce GTX 2080Ti GPU for the training and inference. In the experiment setting, Stochastic gradient descent algorithm was selected for the network training, with the start learning rate set as 0.001. We leverage Pytorch CosineAnnealingLR function to decay the learning rate. The mini batch size set as 64 and the number of train epoch is 130. Different data augmentation methods were also adopted in all experiments to avoid overfitting and for quantitative comparison. Moreover, the models loads the parameters of pre-trained on ImageNet dataset in all the experiments.

4.3 Performance comparison

To verify the effectiveness of GridMask-based data augmentation, we use three widely used CNNs architectures for the comparison, while the inputs and learning settings are kept same to provide a fair comparison. To visualize the model's training stage, we show the evolution of loss of ResNet101 during the training process in Figure 5. As can be seen from the Figure, GridMask provides better performance with comparison to Random Rotation, Flip and Mixup-based data augmentation approaches.

Due to the limitation of space, we list macro-averaged recall scores of ResNet101, DenseNet169 and EfficientNet B0 on the Bengali hand written task in Table 1. As can be seen from the Table 1, significant boost of macro-averaged recall scores can be identified for all the three different CNN network architectures. The experimental results indicate that GridMask data augmentation approach can indeed provide a better generalization ability on this grapheme classification task.

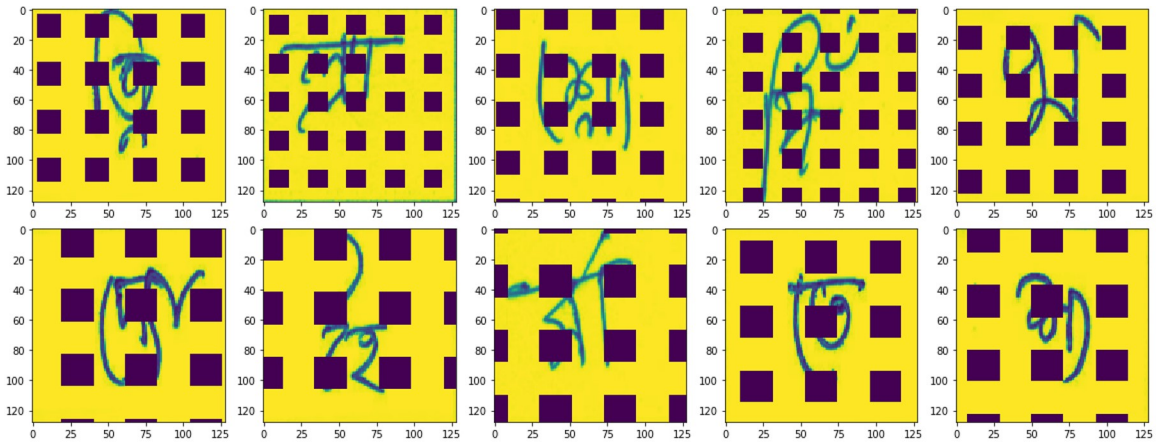

Figure 4: Samples from GridMask-based augmented dataset.

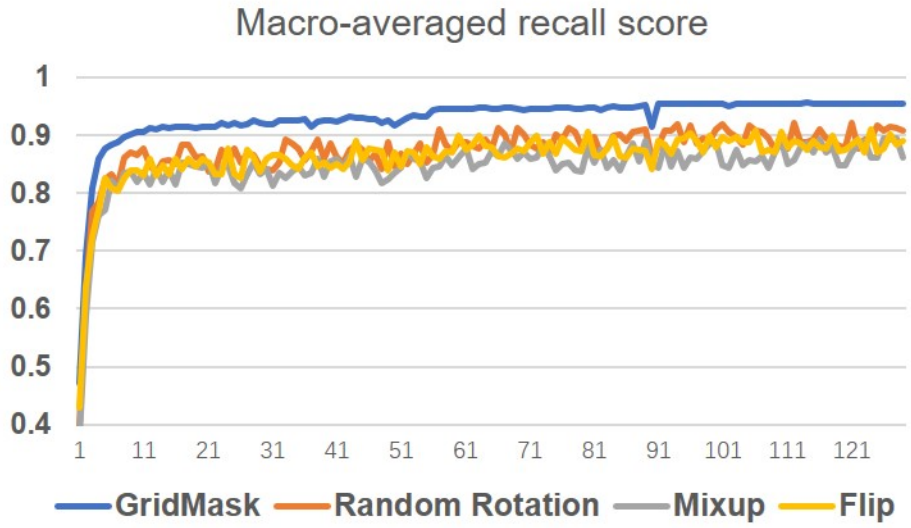

Figure 5: Evolution of macro-averaged recall score of ResNet101 trained with different data augmentation approach.

Table 1: Classification performances (macro-averaged recall score) of ResNet101, DenseNet169 and EfficientNet B0 with hold-out validation on the Bengali hand written grapheme recognition task, respectively. Significant boost of the recall can be identified when the model is trained with GridMask data augmentation.

	ResNet101	DenseNet169	EfficientNet B0
Random Rotation	0.907	0.915	0.930
Flip	0.890	0.935	0.942
Mixup	0.861	0.943	0.947
GridMask	**0.955**	**0.957**	**0.961**

5 CONCLUSION

In this study, we explore the deep learning-based approach for Bengali handwritten grapheme classification task using different data augmentation approach, which demonstrates that GridMask data augmentation can provide better performance for the recognition task. For our further work, we would like to explore the performance of proposed method across large datasets. It would also be interesting to extent our improve the GridMask approach. The current work may also be employed for hand written recognition tasks.

6 ACKNOWLEDGMENT

The author would like thank for the constructive suggestions from the reviewers.

7 REFERENCES

[1] [n.d.]. Kaggle Bengali Competition Description. https://www.kaggle.com/c/bengaliai-cv19/overview. Accessed: 2020-02-20.

[2] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: Fast and Flexible Image Augmentations. *Information* 11, 2 (2020). https://doi.org/10.3390/info11020125

[3] Pengguang Chen. 2020. GridMask Data Augmentation. *arXiv preprint arXiv:2001.04086* (2020).

[4] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. 2017. Dual path networks. In *Advances in Neural Information Processing Systems*. 4467–4475.

[5] Krishnendu Ghosh and K Sreenivasa Rao. 2011. Memory-based data-driven approach for grapheme-to-phoneme conversion in Bengali text-to-speech synthesis system. In *2011 Annual IEEE India Conference*. IEEE, 1–4.

[6] Hiroshi Inoue. 2018. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929* (2018).

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

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[9] NP Narendra, K Sreenivasa Rao, Krishnendu Ghosh, Ramu Reddy Vempada, and Sudhamay Maity. 2011. Development of syllable-based text to speech synthesis system in Bengali. *International journal of speech technology* 14, 3 (2011), 167.

[10] Li-fang Pan and Bing-ru Yang. 2009. Study on KNN arithmetic based on cluster. *Computer Engineering and Design* 30, 18 (2009), 4260–4262.

[11] D Rwrelhart, G Hinton, and R Williams. 1986. Learning representations by back-propagating error. *Nature* 323, 9 (1986), 533–536.

[12] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[13] Johan AK Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural processing letters* 9, 3 (1999), 293–300.

[14] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

[15] Mingxing Tan and Quoc V Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* (2019).

[16] Limin Wang, Sheng Guo, Weilin Huang, and Yu Qiao. 2015. Places205-vggnet models for scene recognition. *arXiv preprint arXiv:1508.01667* (2015).

[17] Shengyun Wei, Kele Xu, Dezhi Wang, Feifan Liao, Huaimin Wang, and Qiuqiang Kong. 2018. Sample mixed-based data augmentation for domestic audio tagging. *arXiv preprint arXiv:1808.03883* (2018).

[18] Jieping Ye, Ravi Janardan, and Qi Li. 2005. Two-dimensional linear discriminant analysis. In *Advances in neural information processing systems*. 1569–1576.

[19] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations* (2018).