Streak2O: Data Augmentation for Handwritten Text Recognition in Neural Networks

Author: Eduardo J. Beltran Feliciano
Advisor: Dr. Marvi Teixeira
Department of Electrical & Computer Engineering

Abstract

Streak2O is a machine learning data augmentation algorithm based on the combination of two other independent algorithms: Streak and Droplet. These three augmentations are implemented as non-trainable TensorFlow custom Keras layers optimized for eager execution in a GPU based environment. They generate configurable random artifacts that imitate real life handwritten historical document or manuscript water damage and document mishandling. Testing these augmentation algorithms with small subsets of the Nationat Institute of Standards and Technology Special Database 19 (NIST-SD19) on a convolutional neural network architecture shows that this new augmentations can help reduce neural network overfitting falling partially into the category of synthetic data generation.

Introduction

One of the most widely studied problems in the field of pattern recognition and computer vision is optical character recognition(OCR)[1]. Handwriting Text Recognition(HTR) is a sub-field of OCR that relates to detecting and classifying non-mechanized characters, those written with ink, graphite or other substances over a physical media. HTR imposes its own challenges including segmentation, style variation by writer, irregular spacing and orientation, usage of non-standard symbols, and noise caused by degradation and mishandling[2][3].

Background

In image classification, augmentation algorithms are routinely utilized to enrich image data sets. Augmentation has two main purposes, generating synthetic data to enrich small data sets and to reduce overfitting over the training data. TensorFlow and MATLAB offer built-in tools for implementing common image augmentations such as rotation, horizontal or vertical reflection, scaling, translation and shearing. Literature offers additional augmentation methods useful for image classification such as CutOut, SamplePairing and CopyPairing[4][5].
In general the modified samples used for training are generally considered a type of synthetic data. For OCR and HTR, it is possible to generate synthetic data that is not based on a currently available data. For example, previous researches generated training samples by using different computer font typefaces that are then process by multiple random augmentations[6][7].

Problem

Previous augmentation algorithms are based on simple transformations, partial elimination of a training sample or creating a chimeric sample based on two existing samples. Non of these methods generate artifacts similar to manuscript degradation and mishandling, and cannot mitigate detection and classification errors caused by these types of artifacts.

Methodology

Streak2O, and its two brother algorithms Droplet and Streak, generate randomized, pseudo-iterative artifacts that imitate mishandling and water damage that should reduce overfitting while allowing the neural network to prepare for real world non-categorizing artifacts.
The International Conference on Document Analysis and Recognition 2017 manuscript dataset (ICDAR 2017)[8] was used to test the realism of the effects on historical images. The artifacts generated by the algorithm can be observed on both historical and synthetic data in samples 1, 2 and 3.

Sample 1:
ICDAR 2017
Streak Augmentation

Sample 2:
ICDAR 2017
Droplet Augmentation

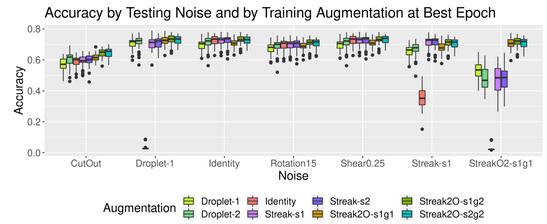
Sample 3:
Streak Augmentation
Synthetic Sample

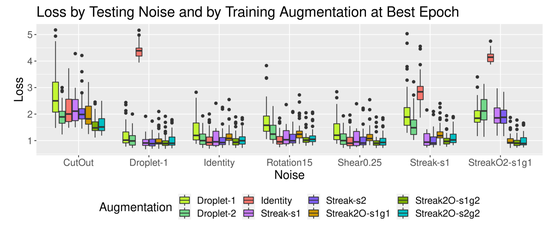
Background noise removal was performed by clipping the values of the sample before running the algorithm. This base noise is reintegrated after executing the algorithm preserving any original artifacts. This allowed closed results to a K-means masking used during prototyping but allowed faster execution during TensorFlow eager execution.

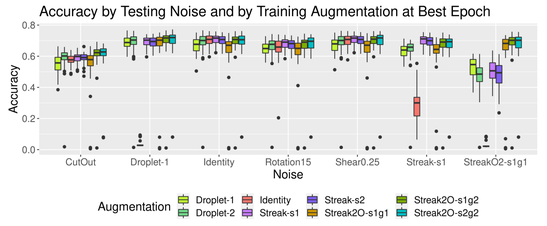
Neural Network architecture was based on the work of Simoyan and Zisserman[?]. Training was perform in subsets of the NIST-SD19[9]. A total of 288 neural networks where trained under four different randomization seeds, nine sub-sample training sizes and eight augmentation configurations: Identity (control group), two Droplet, two Streak, and three Streak2O configurations. Each neural network was then evaluated against 28 different test configurations, 4 seeds and 7 noises. The Identity noise refers to no noise applied during evaluation. Shear, Rotation and CutOut noises are external augmentations applied to the test set which was chosen from the NIST-SD19 recommendation. Non of the test images were used for training. The Droplet, Streak and Streak2O noises match one of the configurations used for training under different seeds generating different random input.
The accuracy and loss, percent and mean squared error respectively, can then be compared at the best epoch and the last epoch after training for a maximum of 50 epochs with an early stop rule based on accuracy with a patience of 10 epochs. A default 20% dropout rate was used, default Keras configuration.

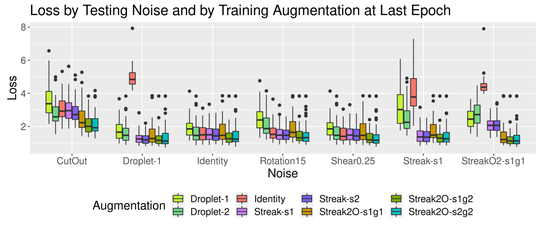
Results and Discussion

The evaluation of the neural networks against multiple types of noise is represented in the graphs bellow. They display the accuracy and loss achieved by the neural networks by testing noise and training augmentation both at their best epoch and their last epoch.









The control group(Identity augmentation) showed not significant difference against any of the augmentations under shear and no-noise testing groups. The control group could not identify testing data under Droplet or Streak2O noise, and had less than ideal performance under Streak noise.
The Droplet augmentation showed lower performance than other augmentations under transformation noise: Rotation and Shear. However, Droplet proofed partially more accurate than the control group when degradation and mishandling artifacts where present: Droplet, Streak and Streak2O. Particularly the Droplet-1 configuration, also used in Streak2O-s1g1, showed lower performance against the rotation noise. The Streak augmentation showed significantly better performance against the control group and Droplet augmentation group for Droplet, Streak and Streak2O noises.
The Streak2O algorithm had the best performance under the CutOut noise and had superior performance against Streak2O noise than just training under Streak or Droplet augmentation.

Conclusions

The Streak2O augmentation shows significant benefit in reducing overfitting in small manuscript datasets. It also shows benefits for synthetic data augmentation as it should help the neural network focus on relevant features and ignore degradation and mishandling artifacts in documents.
The scheduler for the augmentations developed into the callback functions of the library should help introduce these algorithms into incremental learning frameworks.
Although, the Streak algorithm similar performance to the Streak2O in Droplet and Streak noises, the Streak2O augmentation showed better performance in the Streak2O noise. We can conclude that the two artifacts generated are in fact different and combined during training offer better noise handling performance than just using the Streak or the Droplet algorithms independently.

Future Work

The tests against the NIST-SD19 are closer to synthetic data due to the clean background of the images. The augmentation algorithms should be tested using more robust manuscript datasets that already bring their own artifacts.
The effective execution area of the algorithm is generally smaller than the image size. With a quick boundary box calculation the algorithm performance could be increase for whole page augmentation by utilizing Tensor masks. The images used in this paper were small and this optimization was not considered.
After identifying previously trained networks for HTR, the algorithm could be applied to test incremental training. Particularly imitating the synthetic data frameworks previously cited[6][7] could be interesting future work.

Acknowledgements

The author would like to thank the faculty and administration of the School of Electrical and Computer Engineering and the office of Graduate Studies of the PUPR for their continued support during a national emergency. In particular, he would like to thank Dr. Marvi Teixeira for his feedback and patience as mentor, Dr. Alfredo Cruz and Dr. Jeffrey Duffany for the opportunity of participating in the NSF-SFS CyberCorps program, and his family and partner, Irmarie Fraticelli-Rodríguez, for their continued love and support.

References

[1] M. Namysl and I. Konya, “Efficient, Lexicon-Free OCR using DeepLearning,” 2019.
[2] V. Rouchon, M. Desroches, V. Duplat, M. Letouzey, and J. Stordiau-Pallot, “Methods of aqueous treatments: the last resort for badly damaged iron gall ink manuscripts,” Journal of Paper Conservation: IADA reports = Mitteilungen der IADA, vol. 13, no. 3, pp. 7–13, 2012.
[3] C. Cars ̧ote, P. Budrugeac, R. Decheva, N. S. Haralampiev, L. Miu,and E. Badea, “Characterization of a byzantine manuscript by infrared spectroscopy and thermal analysis,” Revue Roumaine de Chimie, vol. 59,no. 6-7, pp. 429–436, 2014.
[4] T. DeVries and G. W. Taylor, “Improved Regularization of Convolutional Neural Networks with Cutout,” 2017.
[5] P. May, “Improved Image Augmentation for Convolutional Neural Networks by Copyout and CopyPairing,” pp. 1–8, 2019.
[6] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Synthetic Data and Artificial Neural Networks forNatural Scene Text Recognition,” pp. 1–10, 2014.
[7] I. Ahmad and G. A. Fink, “Training an Arabic handwriting recognizer without a handwritten training data set,” Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, vol. 2015-Novem, pp. 476–480, 2015.
[8] A. Fornes, V. Romero, A. Baro, J. I. Toledo, J. A. Sanchez, E. Vidal,and J. Llados, “ICDAR2017 Competition on Information Extraction in Historical Handwritten Records,” Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, vol. 1, pp.1389–1394, 2017.
[9] P. J Grother and K. K. Hanaoka, “NIST Special Database19 Hand printed Forms and Characters Database,” pp.1–30, 2016.