Project Interim Report: Improving Mixed Example Data Augmentation

Group 13
Sama Samrin [1191609], Nur Alam [1189508], Rahat Hasan [1192154]
Department of Computer Science, Lakehead University

***Index Terms*—data augmentation, image augmentation, mixed example**

I. Introduction

Since the beginning of this century, scientific components like neural networks, machine learning and deep learning algorithms have been playing a vital role in the technological revolutions of numerous industries. For instance, Convolutional Neural Networks (CNNs) [1] and Deep Convolutional Neural Networks (DCNNs) [2] can help humans to diagnose diseases in different organs like brain and lungs through image classification, detection, localization and more. They have also been applied to develop advanced security measures like face recognition [3] on surveillance cameras and signature verification on important documents [4]. Additionally, Long Short Term Memory (LSTM) algorithm is capable of preparing reliable weather forecast for a specific location based on its relevant historical data.

In all of these applications, the neural network models required training over an extensive dataset to detect, follow and memorize the underlying correlations between different variables. Experts and innovators utilize the abundance of such data to train these learning models they designed.

Based on the provided training dataset, the model gets to learn the hidden patterns that are invisible to our human eyes. It can showcase which factors contribute to the peak of certain numerical variables, or which non-numerical feature occurs most frequently in the given dataset. After extracting these observations, the model is then expected to identify similar patterns on the testing data and to successfully answer our queries.

If the model fails to do this part, it is often accredited to the problem of overfitting. This common issue happens when the machine learning model can provide the correct results for training data, but fails to do the same for testing data. Overfitting typically happens when the training dataset is too small. A dataset lacking in size or data points fails to provide enough samples for the model to learn from.

This frequently occurring issue can be resolved with data augmentation - a process that produces artificially generated data to increase the dataset size. It is heavily used in deep learning and image processing tasks to avoid overfitting and train a model that can produce dependable results. Data augmentation can generate the new data points by modifying certain aspects of the original data [5]. In case of image augmentation, one image from the original dataset can lead to many new images through transformation techniques like rotation, crop, shear, translation and more. It becomes possible because the computer perceives an image as a multi-dimensional array where each value of the array represents a pixel. Data augmentation creates new images by modifying these particular values according to the parameters set by researchers. An important aspect of choosing the augmentation technique is to make sure that the resulting change in image is not too drastic, since it should reflect the patterns of the original image dataset.

II. Summary of Original Paper

The chosen paper [6] acknowledges the success of linear transformations in data augmentation and takes the initiative to explore if a generalized version of the non-linear transformations also leads to similarly successful outcome. It evaluates the necessity of linearity in mixed example training for data augmentation and implements a handful of relevant algorithms to test the impact.

The paper runs its tests on three widely used datasets, namely CIFAR-10, CIFAR-100 and CALTECH-256. The results of first two datasets show error percentages of 3.8-6.2% and 19.7-24% respectively, whereas the last dataset produces an accuracy rate of 48.6-59.7% after the discussed image transformation techniques are applied to them. Most mixed-example training methods performed equal or better than ResNet18, which the authors regarded as the state-of-the-art benchmark for mixed example methods at the time.

III. Motivation

Our motivation primarily lies in implementing the methods of the provided paper and reproducing similar results for a field that can benefit mankind. For instance, the medical field can benefit greatly from effective data augmentation in detecting and analyzing potential but unforeseen cases of diseases. On the other hand, climate researchers can also utilize it to understand the progress and implications of climate change on the planet. Using a dataset related to healthcare or climate would be significantly more influential for the greater good as opposed to the datasets used in the original paper.

That is why we have decided to focus on the medical field and we came up with the following research questions -

Primary Research Question:

Can the implemented methods improve crucial stages like disease diagnosis?

Secondary Research Questions:

1. How can the new wave of artificial intelligence change or affect mixed example data augmentation?

2. Does integrating AI into image augmentation make sense for the medical field?

IV. Dataset

Since one of our main motivations was to utilize the mixed example training for beneficial fields like healthcare, we selected a dataset of skin cancer images named HAM10000. Originally published in 2018 [7], it consists of dermatoscopic images portraying common pigmented skin lesions. It was later modified in 2020 to have segmentation masks and is now available on Kaggle to be easily downloaded on notebooks [8]. The original authors collected these images for 20 years from Cliff Rosendahl of Queensland, Australia and Medical University of Vienna, Austria.

This dataset contains 10,015 images which is a small amount compared to the three datasets used in the original paper. We chose a dataset of this particular size since our goal is to produce more training images for the data augmentation process and that will automatically increase its size. Ultimately, it should have as many images as the CIFAR-100 and CIFAR-10 datasets which have 60,000 images each. It has no missing or mismatched values, which should make our work easier.

The images here primarily belong to seven diagnostic categories of pigmented lesions or spots resulting from melanin-producing cells. These categories include dermatofibroma (DF), Bowen's disease (AKIEC), melanoma (MEL), melanocytic nevi (NV), basal cell carcinoma (BCC), vascular lesions and benign keratosis-like lesions. There are eight columns in total where the first column has the image's name as a string and the following seven columns are named after the seven pigmented lesion categories mentioned above.

V. Preliminary Results

We are yet to implement the most significant phases of the project. So far, we have only implemented the methods for a selected pair of images from the given dataset.

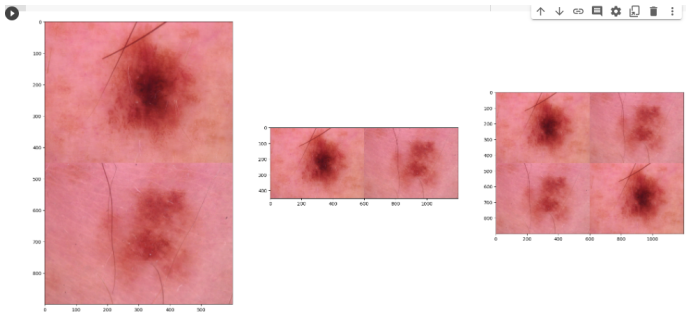

Fig. 1. Three kinds of concatenation (vertical, horizontal, mixed) on a pair of images from the dataset

VI. Challenges

Initially, we struggled to finalize a dataset that belongs to the medical field, contains images that we can transform in a variety of ways and doesn't need to be downloaded manually. It should also not be too small to perform the augmentation on, or too large to operate on through code. Moreover, transforming its images into different orientations should make sense or serve the purpose. For instance, we don't need to rotate images of vehicles since they are hardly upside down in typical circumstances.

The bigger challenge associated with replication of this paper is how to make our version different enough since the code of this paper is already available on GitHub. This is a significant challenge because our group members were not experienced enough with OpenCV or related coding features.

During the initial implementation, we faced a number of issues due to our lack of experience regarding image augmentation codes. For instance, we ended up augmenting the same two images in the same way (mixed concatenate) 4000 times and only realized it while displaying the first few images side by side.

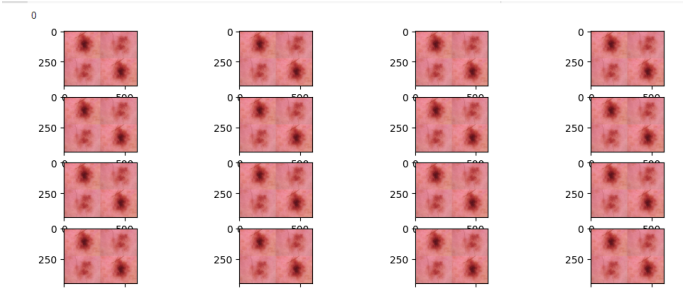

Fig. 2. Same two images concatenated 4000 times by mistake

We need to train ourselves better and fix these issues to obtain the results on time.

VII. Proposed Completion Plan

Our minimum goal is to successfully implement the methods described in paper with a ResNet18 model. For the nominal plan, we will look for ways to enhance the results by reducing the error percentage. Ideally, it should be lower than the majority of error rates showcased in the paper. Our ambitious goal includes to integrate artificial intelligence or something equally modern into this image augmentation project, and observe its effectiveness.

References

[1] D. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," *Evolutionary intelligence*, vol. 15, no. 1, pp. 1–22, 2022.

[2] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big data*, vol. 6, no. 1, pp. 1–18, 2019.

[3] S. Almabdy and L. Elrefaei, "Deep convolutional neural network-based approaches for face recognition," *Applied Sciences*, vol. 9, no. 20, p. 4397, 2019.

[4] E. Alajrami, B. A. M. Ashqar, B. S. Abu-Nasser, A. J. Khalil, M. M. Musleh, A. M. Barhoom, and S. S. Abu-Naser, "Handwritten signature verification using deep learning," *International Journal of Academic Multidisciplinary Research (IJAMR)*, vol. 3, no. 12, pp. 39–44, 2020.

[5] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann, "Fixing data augmentation to improve adversarial robustness," *arXiv preprint arXiv:2103.01946*, 2021.
[6] C. Summers and M. J. Dinneen, "Improved mixed-example data augmentation," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 1262–1270.
[7] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.
[8] Suraj Ghuwalewala, "Skin cancer: Ham10000," 2020, https://www.kaggle.com/datasets/surajghuwalewala/ham1000-segmentation-and-classification.