International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 3, May-June 2024
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Improve Deep Learning Model Performance with Data Augmentation

Hai Hoang Thanh*, Nguyet Mong Thi**
*(Thai Nguyen University of Economics and Business Administration, Viet Nam
Email: hoangthanhhai03091988@gmail.com)
** (Thai Nguyen High School, Viet Nam
Email: nguyetmt@tnue.edu.vn)

----------**********----------

Abstract:
Data augmentation is crucial for building an efficient deep learning model. The performance of a deep learning model is considerably influenced by the size and the diversity of input data. By generating new data from existing data, we can expand data size as well as data diversity, from which model performance could be enhanced. In this paper, we examine the impact of data augmentation on model performance with different types of input data. Experiments show model accuracy is significantly influenced by augmented training data.

Keywords **—data augmentation, deep learning, model performance.**

----------**********----------

I. INTRODUCTION

Insufficient data in machine learning can lead to overfitting, making the model unable to analyze new data properly [1]. In general, acquiring additional data is challenging, costly, and sometimes impractical. For example, considering the problem of predicting skin cancer based on skin images, it is true that collecting new skin cancer photos and labeling them are expensive and time-consuming. Also, in many cases, copyright laws or private policies prevent researchers from obtaining new data [2].

Diversity is another important property of data that could influence on model performance. Data diversification can provide samples with enough information to train machine learning models, resulting in better performance [3].

Data augmentation is a technique of artificially increasing the training set by creating modified copies of a dataset using existing data. It includes making minor changes to the data samples or using deep learning to generate new data points [4]. Data augmentation can be used when the initial training set is small, to prevent models from overfitting, or to improve the model performance.

Commonly used types of data are images, text, audio, video and tabular. Each type of data has its suitable data augmentation techniques.

In this research, we focus on two types of data that commonly used, which are images and tabular.

In this research, we focus on two types of data commonly used, which are images and tabular. For images, we have a few popular augmentation techniques, such as random flip, crop, rotation, or color transformations. For tabular data, transforming techniques can be applied to modify existing rows or columns to create new ones.

Our main objectives in this work are:

- To assess the influence of data augmentation techniques on model performance with image data;
- To assess the influence of data augmentation techniques on model performance with tabular data.

II. METHODS

2.1 Image Data

We consider the image classification problem. Different pre-trained models are finetuned on a specific dataset and model performance is measured by accuracy. Model quality is assessed based on models trainned on different training datasets, namely, original training sets (without data augmentation technique), and augmented training sets (with one augmentation technique or more).

2.2. Tabular Data

We use a specific model architecture for the clasification problem using a tabular dataset. The model has two separate layers to learn two kinds of features (categorical and continuous ones) before being concatenated into several shared linear layers (Figure 1)

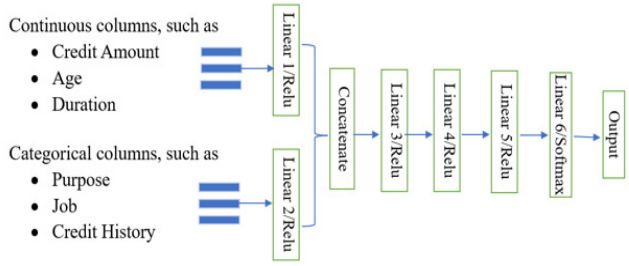

Fig. 1 Used model architecture for tabular data

We examine model accuracy using original training data (without data augmentation) and augmented training data. We add minor changes to continuous features and keep other categorical ones unchanged.

Let $D_{train} = \{(\boldsymbol{x}_i, y_i), i = 1,..,N\}$ be the original training dataset, where

$$\boldsymbol{x}_i = (x_{i1}, \dots, x_{im}, z_{i1}, \dots, z_{in})$$

are the features, y_i is the credit status ($y_i = 1$ or $y_i = 0$) of the i[th] customer, $(x_{i1}, \dots, x_{im})$ are continuous and $(z_{i1}, \dots, z_{in})$ are categorical. Set

$$a_j = \max_{i=1,\dots,N} x_{ij}, j = 1, \dots, m.$$

The new continuous feature values are determined by

$$x'_{ij} = x_{ij} + r.u_{ij}, i = 1, \dots, N, j = 1, .., m$$

where u_{ij} is sampled from uniform distributions $U(-a_j, a_j)$ and r is a small parameter. The new samples corresponding to $(\boldsymbol{x}_i, y_i)$ have the following form

$$(x'_{i1}, \dots, x'_{im}, z_{i1}, \dots, z_{in}).$$

By sampling multiple times from the above uniform distributions, we have a larger training dataset including new sampled instances and the original ones.

III. EXPERIMENTS AND RESULTS

3.1 Data

With images, we use the CIFAR-10 dataset for all experiments. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. This data is publicly available in various libraries, like Torchvision.

With tabular data, we use the German Credit Data, which is available on UC Ivrine Machine Learning [5]. There are 1000 samples of 20 features (7 – numerical, 13 – categorical) including the information of customers like status of checking account, duration in month, or credit history. The outcome is binary (Good/Bad).

3.2 Models

CIFAR – 10 Dataset: We use different pre-trained models, namely, ResNet18, ResNet34, VGG16, and VGG19 tuned on the CIFAR-10 training set. Model accuracy is measured by the test set. We train models using original training data and compare them with models using augmented data. All experiments are randomly run 5 times, the final performance is averaged.

German Credit Dataset: The original dataset (1000 samples) is divided into 70% training set and 30% test set.

Augmented data are sampled 4 times from uniform distributions with $r = 0.01$. Augmented data are concatenated with original training set to createa new training set. Model parameters are shown in Table 1.

Table 1. Parameters and Architecture of the model

Layer	Parameters and Architecture
Input Categorical Columns	Input shape: 13
Continuous Columns	Input shape: 7

Linear 1	in_features = 7, out_features = 100
Linear 2	in_features = 13, out_features = 100
Linear 3	in_features = 200, out_features = 200
Linear 4	in_features = 200, out_features = 100
Linear 5	in_features = 100, out_features = 50
Linear 6	in_features = 50, out_features = 2

The model is randomly trainedfive times with 200 epochs each run using the Pytorch deep learning library. We use the Adam method for optimizing.

3.3 Results

CIFAR – 10 Dataset:

Table 2 shows the accuracy of different models trained on the CIFAR-10 dataset. Using augmented data significantly improves model performance. This improvement is higher when using two augmentation techniques (Flip and Rotation) compared to using only one technique (Flip). Standard deviations of estimated accuracy tend to lower with generated data compared to original training data.

Table 2. Accuracy of models trained on CIFAR-10 (mean ± sd)

Model	Original Training Set (without augmentation)	Augmented Data (one method)	Augmented Data (two methods)
ResNet18	73.32 ± 0.39	77.15 ± 0.23	79.75 ± 0.21
ResNet34	74.77 ± 0.30	78.12 ± 0.32	80.53 ± 0.27
VGG16	85.14 ± 0.64	85.97 ± 0.26	86.92 ± 0.18
VGG19	86.23 ± 0.45	87.42 ± 0.28	87.45± 0.32

German Credit Dataset:

Table 3 shows the model performance of the deep learning model trained on the GERMAN dataset. Using augmented data gives higher accuracy with smaller deviation.

Table 3. Accuracy of models trained on GERMAN dataset (mean ± sd)

Original Training Set (without augmentation)	Augmented Data
81.27± 0.72	81.93± 0.60

IV. CONCLUSIONS

In this work, we examine the influence of augmented data on deep learning model performance. Experiment results show that using data augmentation do improve model accuracy with both image data and tabular data. Estimated accuracy is more stable when training models using augmented data. We aim to do more experiments about impact of different data augmentation techniques on other types of data like video or audio in the future.

REFERENCES

[1] D. Haba, *Data Augmentaion with Python*, Pack Publishing, 2023.
[2] S. Shaked, *Overcome Data Shortages for Machine Learning Model Training with Synthetic Data*, [Online]. Available: https://tdwi.org/articles/2021/06/14/adv-all-overcome-data-shortages-for-ml-model-training-with-synthetic-data.aspx
[3] Z. Gong, P. Zhong and W. Hu, *Diversity in Machine Learning*, *IEEE Access*, vol. 7, pp. 64323-64350, 2019.
[4] A. Awan, *A comlete Guide to Data Augmentation,* [Online]. Available: https://www.datacamp.com/tutorial/complete-guide-data-augmentation
[5] German Credit Data, UC Irvine Machine Learning Repository, [Online]. Available: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

