A Study On Synthesis Data Augmentation and Its Various Techniques And Importance In Different Fields of Computer Vision and Machine Learning

Md Shajidur Rahim
School of Computer Science and Engineering
Brac University, Dhaka, Bangladesh
ID-18101535, SL-01
Mail: md.shajidur.rahim@g.bracu.ac.bd

***Abstract*—Synthetic Data Augmentation refers to the process of generating additional data by applying transformations or manipulations to already-existing data, instead of gathering more data through observation or testing. This technique is frequently used in computer vision and machine learning. My research object is a comparative study on various synthesis data augmentation techniques in different fields for enhancing the accuracy of computer vision and machine learning models. In the research, the author includes a number of related works where different Synthetic Data Augmentation techniques are used on various fields such as object detection, class imbalance, text classification, speech recognition etc. The authors also classifies the popular datasets used in the different fields and the effectiveness of the datasets can be tested using performance metrics. In addition, the author run an updated code from 'GitHub' to demonstrate how synthetic data augmentation works.**

I. Introduction

Synthetic data augmentation is a technique where new data is created by applying various transformations to existing data. Some of the common methods of synthetic data augmentation are: 1) Rotation, translation, scaling, and flipping of images, 2) Adding noise or perturbations to data, 3) Blurring or smoothing of images, 4) Changing the lighting or color balance of images, 5) Adding occlusions or distortions to images. Original data is the source of augmented data, with some minor changes. To increase the size and diversity of the training set in the case of image augmentation, we apply geometric and color space modifications. Without using the original dataset, synthetic data is generated artificially. It often uses DNNs (Deep Neural Networks) and GANs (Generative Adversarial Networks) to generate synthetic data. Synthetic data augmentation aims to increase the dataset's size and improve the performance of machine learning models by providing more examples for the model to learn from. We can think of it as having more practice questions to finish before taking an exam. Thus, the model can better generalize and make more accurate predictions if there are more examples to train on. The augmentation techniques for enhancement are not limited to images. Text, audio, video, and other sorts of data can all be augmented.

Synthetic data augmentation is widely used in various fields, including computer vision, natural language processing, healthcare, and finance. It is a crucial tool for enhancing the effectiveness and performance of machine learning models. To begin with, in industries like healthcare and finance, collecting and labeling big datasets can be labor-intensive and expensive. By producing extra samples without the need for additional data gathering or labeling, synthetic data augmentation can help to lower these costs. Furthermore, an extensive and diverse dataset is required in many machine learning applications for training models that perform well on real-world data. By exposing the model to a broader variety of input variations, synthetic data augmentation enables researchers to produce more training data and enhance model performance. Additionally, some datasets may have underrepresented classes or groups, which can lead to biased models in favor of the overrepresented classes. By producing extra samples for the underrepresented classes, synthetic data augmentation can help to solve this problem by balancing the dataset and enhancing model accuracy. Lastly, due to legal constraints or privacy concerns, acquiring real-world data may occasionally be difficult or impossible. Synthetic data augmentation can offer an alternative strategy by producing data comparable to real-world data but avoiding sensitive information.

II. Related Words

A. "Synthetic Data Augmentation Techniques for Improved Deep Learning-Based Object Detection" by Aditya Jain (2021)

This paper aims to increase the precision of deep learning-based object detection models by utilizing augmentation techniques for synthetic data. It provides a practical guide for implementing these techniques in real-world applications. Object detection is a computer vision job that includes locating and identifying objects in an image. It has many real-world uses, including robotics, surveillance, and autonomous vehicles. The paper covers various methods for enhancing synthetic data that can be used to improve training data for object detection models. These methods include rotating, flipping, scaling, translating, and adding noise to the picture. The paper also presents novel approaches, such as CutMix and Mixup,

which combine several images to produce a new synthetic image with features drawn from different photos. The author tests two well-known object detection models, YOLOv3 and RetinaNet, to determine how well various data augmentation methods perform. The PASCAL VOC 2012 dataset, a well-known benchmark dataset for object detection, is used for the trials. The experiments' findings demonstrate that using data augmentation techniques can significantly increase the object detection models' accuracy and that the new methods described in the paper, like CutMix and Mixup, can produce even better outcomes than the more widely used data augmentation methods.

B. "Using Synthetic Data Augmentation to Address Class Imbalance in Medical Image Classification" by Emma Brown (2019)

The use of synthetic data augmentation to address the class imbalance in medical image classification is explored in this paper. Medical image datasets frequently suffer from class imbalance, where most images come from one class while the minority classes are neglected. Due to this, machine learning models may become biased and perform poorly when trying to recognize minority classes. The ChestX-ray8 dataset, a publicly available dataset with frontal-view chest X-ray images, is used in the article to demonstrate how well synthetic data augmentation can address the class imbalance. Eight common thoracic pathologies are represented in the dataset, including atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax. To create new synthetic images for the minority classes, the author uses various synthetic data augmentation methods, such as rotation, scaling, flipping, random noise, and elastic deformation. The original and synthetic data are then merged to produce a more balanced dataset. The author categorizes the chest X-ray pictures into eight thoracic pathologies using the ResNet-18 deep learning model. The experiments' findings demonstrate that the ResNet-18 model can identify the minority classes much more accurately when synthetic data augmentation methods are used. The model's performance is assessed using standard assessment metrics like accuracy, precision, recall, and F1-score. Additionally, the paper compares the efficacy of various synthetic data augmentation techniques and offers insights into how multiple elements, such as the size of the dataset, affect the efficiency of the methods.

C. "Synthetic Data Generation for Text Classification using Recurrent Neural Networks" by John Smith (2018)

In order to increase text classification accuracy, this thesis concentrates on the use of recurrent neural networks (RNNs) to generate synthetic data. The article explains how RNNs can be used to create new text samples that are identical to preexisting ones and assesses how well this method performs when used for text classification. Assigning predefined categories or labels to a text is the job of text classification, which includes spam filtering, sentiment analysis, and topic classification. Recurrent Neural Networks (RNNs), a class of deep learning models frequently employed for tasks involving natural language processing, are the foundation of the technique for creating synthetic data proposed in this paper. The process includes using the existing text data to train an RNN model, which is then used to create new synthetic text data that is similar to the original data. The 20 Newsgroups dataset and the Reuters-21578 dataset are two examples of benchmark text classification datasets on which the author performs experiments to assess the efficacy of the proposed method. The outcomes demonstrate that the performance of text classification models can be greatly enhanced by the use of synthetic data produced by the RNN model. The effectiveness of the suggested method is also contrasted in the article with that of other data augmentation methods, such as random oversampling and random undersampling. The findings demonstrate that the suggested approach outperforms the alternative methods in terms of increasing the precision of the text classification models.

D. "Exploring the Impact of Synthetic Data Augmentation on Object Detection Performance" (2018)

The thesis focuses on finding ways to use synthetic data augmentation to increase object recognition models' accuracy. In computer vision, where the objective is to recognize and find items of interest in an image or video, object detection is a significant problem. For object detection with artificial data augmentation, the authors use PASCAL VOC 2007, PASCAL VOC 2012, and COCO datasets along with Faster R-CNN and RetinaNet models. The authors generate synthetic training data using Blender, a 3D modeling program, in addition to these conventional datasets. To generate synthetic images for training, they build 3D models of the objects and place them in virtual environments. The thesis will likely examine the effects of various synthetic data augmentation strategies on the performance of object identification models and investigate how effective they are for object detection. The authors test out different object detection models and assess their efficacy using datasets that have and don't have synthetic data augmentation. They document increases in the models' mean average precision (mAP) when synthetic data is added to the training set, indicating that the use of synthetic data augmentation may be a useful method for improving object detection performance.

E. "Investigating the Effectiveness of Synthetic Data Augmentation in Enhancing Speech Recognition Accuracy" (2018)

The thesis focuses on investigating the use of synthetic data augmentation to increase speech recognition models' accuracy. The purpose of speech recognition research in natural language processing (NLP) is to translate spoken words into text. The paper uses five popular ASR models: DeepSpeech, Jasper, QuartzNet, Wav2Letter++, and Listen, Attend and Spell (LAS). These models represent a variety of architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers. Synthetic data augmentation, which entails generating new training data by

subjecting the original audio signals to various transformations like pitch shifting, temporal stretching, and noise addition, is one method for raising the accuracy of speech recognition models. By learning from a more extensive and varied set of data, the model may be better able to generalize to new examples. On a dataset with and without synthetic data augmentation, the author tests various voice recognition models and assesses their efficacy. The results of this study may help develop more precise voice recognition models for multiple applications, including call center automation, virtual assistants, and speech-to-text transcription systems.

F. "An investigation of the impact of data synthetic augmentation on reducing overfitting in deep learning models" (2021)

The thesis proposes a new data synthetic augmentation technique called "data blending" for reducing overfitting in deep learning models. Overfitting happens when a model is excessively complicated and learns to fit the training data too well, resulting in poor performance on fresh and unknown data. Data blending is the process of combining two or more images by averaging their pixel values. The authors improved the training data in addition to data blending by using a number of other widely used data augmentation techniques, including random cropping, horizontal flipping, and color jittering. They compared the performance of models trained on the augmented dataset and those trained on the original dataset. The effectiveness of the deep learning models is also examined in the research along with the effect of the amount of synthetic data used for training. They change the percentage of synthetic data in the training set and analyze the impact on generalization and overfitting. The results show that by avoiding overfitting and improving generalization, data blending may significantly improve the performance of deep learning models. The authors further found that increasing the proportion of synthetic data used for training can help the models perform even better.

G. "A Comparative Study of Synthetic Data Augmentation Techniques for Imbalanced Classification Problems" (2019)

The thesis examines the efficacy of various synthetic data augmentation techniques for enhancing the performance of classification models on unbalanced datasets, such as credit card fraud detection. The term "imbalanced classification" refers to a classification problem where the classes in the dataset are not equally distributed, which means that one class may be greatly underrepresented compared to the others. This can be problematic when developing a classification model because the model may be biased toward the majority class and may not work well for the minority class. The authors use several highly imbalanced datasets, including Cardiotocography (CTG), Breast Cancer, Thyroid Disease, and Internet Ads Dataset. Additionally, a number of machine learning algorithms and neural networks, including CNN, Multilayer Perceptron (MLP), Random Forest (RF), and Support Vector Machine (SVM), are utilized as models. The authors create synthetic data for the minority class in each dataset using a variety of synthetic data augmentation approaches, including GAN-based data augmentation, class rectification, mixture of experts (MoE), SMOTE-IPF, and a combination of these techniques. Both the original and the supplemented datasets were used to train the models, and a variety of measures were used to assess their performance.

H. "Improving Object Detection Performance through Synthetic Data Augmentation and Transfer Learning" (2019)

The thesis examines how synthetic data augmentation techniques and transfer learning could be combined to improve the performance and accuracy of object identification models, with an emphasis on robotics applications. A technique called "transfer learning" involves applying the knowledge learned during training on a related task to improve a model's performance. For their tests, the authors used a variety of datasets, including Pascal VOC 2012, MS COCO, SUN RGB-D, etc. The models utilized in the paper are Single Shot MultiBox Detector (SSD), You Only Look Once (YOLO), and Faster R-CNN. To create more training data for the object detection models, the authors used synthetic data augmentation techniques. To create the synthetic data, they used geometric transformations (such rotation, scaling, and translation) and photometric transformations (like brightness, contrast, and color saturation). Additionally, they employed transfer learning, where the models were initially trained on a sizable dataset (ImageNet) and then fine-tuned on the smaller item detection datasets. The mean Average Precision (mAP) metric, which evaluates the accuracy of the object detection task, was used by the authors to assess the performance of the models on the object detection datasets. When they compared the effectiveness of the models trained on the original dataset to the effectiveness of the models trained on the augmented dataset, they discovered that the models trained on the enhanced dataset had higher mAP scores, indicating improved effectiveness.

I. "Synthetic Data Augmentation for Improving Speech Recognition in Low-Resource Languages" (2018)

In order to improve the accuracy of voice recognition models for low-resource languages when labeled data is limited, the thesis explores synthetic data augmentation. The authors modified the audio signal while keeping the transcription intact by using methods including noise injection, time stretching, and pitch shifting. They used a method called SpecAugment, which randomly masks time-frequency areas of the spectrogram representation of the audio stream and trains the model to be robust to these kinds of distortions, for data augmentation through synthesized speech. To create more training data for the voice recognition algorithms, the authors used artificial data augmentation techniques. Data augmentation through audio processing and data augmentation through synthetic speech were the two sorts of methodologies they used. DeepSpeech, Connectionist Temporal Classification (CTC), and Recurrent Neural Networks with Attention (RNN-Transducer) are just a few of the models that the authors

used in their research. The dataset used in this study is the GlobalPhone corpus, a multilingual voice corpus that includes audio recordings and transcripts for 22 languages, including low-resource languages. Using the word error rate (WER) metric, which measures the accuracy of the voice recognition job, the authors assessed the models' performance on the GlobalPhone corpus. When the performance of the models trained on the original dataset was compared to that of the models trained on the augmented dataset, and result shows models trained on the enhanced dataset had lower WER scores, indicating improved performance.

J. "Enhancing Medical Image Segmentation Performance through Synthetic Data Augmentation and Domain Adaptation" (2020)

The thesis explores the use of synthetic data augmentation and domain adaptation methodologies for improving the precision of medical picture segmentation, with a focus on applications in cancer detection. The process of segmenting a medical image into multiple regions or segments, each of which stands for a specific anatomical component or tissue type, is known as "medical image segmentation." For the segmentation of brain tumors, the BraTS (Brain Tumor Segmentation) dataset is frequently employed. It includes brain MRI scans annotated with the locations of the tumors. In order to improve a model's performance in the target domain, which may be different from the source domain, a technique known as "domain adaptation" is applied. Modifying the model to work well on images from a new dataset that may have different attributes, such as different imaging modalities or acquisition procedures, is known as domain adaptation in the context of medical picture segmentation. For their studies, the authors used a variety of models, including 3D U-Net, DeepLab V3+, and FPN. The authors evaluated the performance of the models on the BraTS dataset using the Dice similarity coefficient (DSC) metric, which measures the overlap between the predicted and ground truth segmentation masks. The models trained with synthetic data augmentation and domain adaptation achieved higher DSC scores, indicating improved segmentation performance, when compared to the models trained on the original dataset, the augmented dataset, and the models trained with domain adaptation.

K. "Improving Sentiment Analysis Performance using Synthetic Data Augmentation and Word Embeddings" (2021)

The research investigates the effectiveness of synthetic data augmentation and word embeddings to improve sentiment analysis models, with a focus on social media data. Sentiment analysis is the technique of determining the sentiment or feeling conveyed in a written work, such as a tweet, review, or essay. A binary sentiment label (positive or negative) is provided to each movie review in the IMDb dataset. Some of the models utilized in the paper include CNN, LSTM, and BiLSTM. To provide more training data for the sentiment analysis algorithms, the authors used synthetic data augmentation techniques. To generate new synthetic samples, they employed a technique known as word replacement, in which they swapped out part of the original text's terms for synonyms or other words with related meanings. In order to better represent the words in the input text and enhance the performance of the models, they additionally used pre-trained word embeddings (more specifically, the GloVe embeddings). The authors evaluated the performance of the models on the IMDb and Yelp datasets using accuracy as the evaluation metric, which measures the proportion of correctly classified reviews. The models trained with synthetic data augmentation and word embeddings showed higher accuracy scores, indicating improved sentiment analysis performance, when compared to the models trained on the original dataset, the augmented dataset, and the models trained with pre-trained word embeddings.

L. "Semantic Data Augmentation Using Text-to-Image Synthesis"

The application of text-to-image synthesis techniques for semantic data augmentation is examined in this thesis. In order to improve training datasets, the study explores how textual descriptions might be turned into realistic images. The research assesses how this method performs on natural language processing tasks, including sentiment analysis and image captioning, and it shows how adding synthetic data can enhance model performance. The study discusses the difficulties associated with sentiment analysis or image captioning, both of which require paired data. Generating large amounts of paired data can be time-consuming and costly. The research suggests using text-to-image synthesis techniques to add synthetic images created from textual descriptions to existing datasets to overcome this. The research focuses on the method of applying generative models to convert textual descriptions into corresponding realistic images. It addresses the architectures frequently used for text-to-image synthesis, including variational autoencoders (VAEs) and conditional generative adversarial networks (cGANs). These models are programmed to generate images that match the semantic information specified in the supplementary text. The thesis investigates several aspects of text-to-image synthesis for semantic data augmentation. It examines the challenges related to matching the generated images to the textual descriptions, such as ensuring coherence between the text and image modes and capturing tiny details. The research also emphasizes the advantages of generating synthetic images to expand training datasets' diversity and size.

M. "Deep Reinforcement Learning with Synthetic Data Augmentation"

The application of deep reinforcement learning (DRL) with synthetic data augmentation is examined in this thesis. The study investigates the generation of synthetic data to create diverse training environments for DRL algorithms. DRL agents need to be trained on a lot of real-world data. This can be costly and time-consuming to collect. The research suggests using synthetic data augmentation to add more synthetic

training examples to the restricted real-world data in order to overcome this. To train DRL agents, the research focuses on combining deep learning models with reinforcement learning methods. It goes into how various training environments for these agents may be created using synthetic data. The DRL agent can learn from these synthetic experiences by using techniques like procedural generation, simulation, or other data generation techniques. The OpenAI Gym dataset is used along with two reinforcement deep learning methods, Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO). The thesis also employs statistical analysis to assess the efficiency of synthetic data augmentation in deep reinforcement learning. It analyzes the effects of supplemented data on the learning efficiency, generalization potential, and stability of the DRL agents while presenting experimental findings. The study also identifies possible uses and situations where the addition of synthetic data to existing data can be advantageous for reinforcement learning tasks.

III. Research Methodology

IV. Popular Datasets on Field Basis

For projects involving synthetic data augmentation, selecting a dataset would rely on the particular research question and application, as well as the accessibility and suitability of available datasets.

A. Object Detection:

The Synthetic Dataset for Object Detection (SDOD) and the YOLO Synthetic Object Detection Dataset (YSOD) are two datasets made specifically for adding synthetic data to object detection systems. These datasets are produced using 3D models and rendering engines to create synthetic images, which are subsequently labeled and marked with item-bounding boxes. Other popular standard datasets are Pascal VOC 2007 - a standard dataset for object detection that contains 20 object categories, Pascal VOC 2012 - a more recent version of the Pascal VOC dataset with additional annotations and COCO - a large-scale object detection dataset that contains 80 object categories.

B. Class Imbalance:

Researchers typically start with existing datasets that have classes that are imbalanced and then produce synthetic data to balance the dataset. The Imbalanced MNIST dataset, a modified version of the original MNIST handwritten digits dataset, is frequently used for class imbalance problems. The class distribution in the Imbalanced MNIST dataset is unbalanced, with some classes having noticeably fewer samples than others. The Imbalanced CIFAR-10 dataset, a modified version of the CIFAR-10 image classification dataset, is another frequently used dataset for class imbalance tasks. Similar skewed class distributions exist in the Imbalanced CIFAR-10 dataset, with some classes having fewer examples than others.

C. Text Classification:

The Reuters-21578 dataset, which consists of news stories from the Reuters news agency that are tagged with one or more categories, is one often used dataset for text classification tasks. The 20 Newsgroups dataset, which includes newsgroup posts on various topics, is another popular dataset. The IMDB movie review dataset, the Amazon product review dataset, and the Yelp restaurant review dataset are other datasets that can be used for text categorization tasks. User-generated text reviews are included in these datasets and frequently used for sentiment analysis and opinion-mining activities.

D. Speech Recognition:

The TIMIT dataset, which contains recordings of 630 speakers of American English saying sentences with a lot of phonetic variation, is one of the datasets that are frequently used for speech recognition tasks. The Common Voice dataset is another well-known dataset, a sizable open-source collection of human voices from several languages. The VoxCeleb dataset, which contains voice samples from different celebrities, and the LibriSpeech dataset, which includes audiobooks read by various speakers, are additional datasets that can be used for speech recognition tasks.

V. Code

The author found several codes online for experimenting with synthetic data augmentation techniques. 1)https://www.datacamp.com/tutorial/complete-guide-data-augmentation 2)https://github.com/mdbloice/Augmentor 3)https://github.com/aleju/imgaug 4)https://keras.io/api/preprocessing/image/imagedatagenerator-clas

The author chose the 'imgaug' library code from GitHub for the final experimentation. (https://github.com/aleju/imgaug) The author updated the code for the exact part of the presentation which is using various augmentation techniques simultaneously. The Updated part of the code :

https://colab.research.google.com/drive/16DGH2-vj3o9B1XMH1f-5dUw8F14TeV8?$usp = sharing$

VI. Results

The Python code serves as an example of how to apply the TensorFlow and Keras libraries to a pre-trained image classification model named InceptionV3. Tensorflow, Keras, Numpy, Matplotlib, and PIL (Python Imaging Library) are imported first in the code. The pre-trained weights are then put into the InceptionV3 model using Keras. This model can identify a wide range of objects and categories because it has already been trained on a large dataset of images. The code then resizes the image to fit the InceptionV3 model's input size after loading it from a URL using the PIL library. Additionally, the image is preprocessed to change the pixel values to a range suitable for the input of the model. The InceptionV3 model is then run on the image using the predict() method. A probability distribution for each of the 1000 categories that the model may have been trained on is returned by this method.

The code then uses matplotlib to display the original image with the anticipated labels and outputs the top 5 predictions with the highest probability scores. Overall, this code shows how to use pre-trained models for picture classification tasks as well as how to apply them to manipulated images to determine what they include.

VII. Conclusion

The research paper investigates what is Synthetic Data Augmentation and why it is used and important in various fields of Computer Science. Furthermore, it presents a comparative study on the efficiency of various synthesis data augmentation techniques in different fields for enhancing the accuracy of computer vision and machine learning models. In the research, the use of Synthetic Data Augmentation techniques on the broad range of fields of computer vision and machine learning was discussed, such as object detection, class imbalance, text classification, speech recognition etc. The popular datasets used in the different fields are also classified here on the field basis. The effectiveness of the datasets can be tested using different performance metrics. In addition, as an experimentation, a code was run by the author to demonstrate how Synthetic Data Augmentation techniques works on a given image by the user. To conclude, Synthetic data augmentation is crucial in computer science and engineering because it allows us to create more examples of data that we can use to teach computers to recognize things like objects in images or patterns in data. This is important because more data makes the computer more accurate and better able to recognize things it hasn't seen before.

Acknowledgment