Advances, Applications, and Challenges in Enhancing Machine Learning Models with GAN-Based Data Augmentation

Sandeep Reddy

**Abstract:**

Data Augmentation using Generative Adversarial Networks has emerged as crucial technique in enhancing the performance of Machine Learning models. When dealing with datasets, GANs generate realistic synthetic data that can improve model generalization, outperforming traditional augmented methods like cropping and noise addition. Study has been especially effective in complex domains such as object detection, human activity recognition, signal processing, ESP fault diagnosis, chatbot training. GAN based data augmentation offer advantages such as improved data diversity and reduced need for costly data collection. Challenges, including training instability, mode collapse, potential introduction of poor-quality data. Overcoming of these challenges, advanced GAN architecture like condition GANs(cGANs), Wasserstein GANs(WGANs) and Regularization technique has been developed. Architecture has enhanced training stability and data quality by incorporating technique such as 1D convolution and adaptive discriminator augmentation. Application of GAN-augmentation has been extended in various fields, including automative, healthcare, education, E-commerce, finance, telecommunication. GAN-Based Data augmentation remains a powerful tool for enhancing AI driven solution in data-scare environment.

**Keywords:**

*Deep learning, Generative Adversarial Networks (GANs), Data Augmentation,1D Convolution, Conditional GANs (cGANs), Wasserstein GANs (WGANs).*

**Introduction**

Applications and development in the field of HAR have been significant as researchers gained interest in this activity through the last few decades; its applications are so various, ranging from monitoring one's health to that in smart homes and then through fitness tracking. More crucially, the real world depends on the need to accurately recognize human activities by using sensor data with accurate outcomes in developing intelligent systems, which can assist them during their daily lives. Amongst the major problems of HAR is the paucity of labeled data that affects machine learning models from performing very well.

As a potential solution to overcome this limitation, researchers have opted for generating synthetic data. Among such tools that emerged and shown their power to generate the most realistic data to complement existing datasets are the GANs. This paper uses cGANs and their enhanced version, CWGANs, for synthesizing accelerometry data for HAR tasks. Motivation: It tries to study whether these GAN architectures are effective in generating high-quality synthetic data for improving the performance of the HAR models.

The research questions motivating this study include identifying appropriate GAN architectures for synthetic acceleration data generation, assessing the closeness of the generated data to reality, and determining whether changes in the amounts of available synthetic data affect classifier performance. The exploration of these questions contributes to a more general understanding regarding the potential of synthetic data to alleviate some of the shortcomings associated with the limited availability of real data in HAR.

The following sections will describe the experimental methodology used to measure the performance of cGANs and CWGANs, followed by discussions regarding the results obtained in experiments. Ultimately, the main contribution of this research would lie in its contribution towards these efforts in enhancing the effectiveness of HAR systems with synthetic data generation techniques.

**Literature Review:**

To minimize intraclass imbalance, these models used GAN models with WGAN and SNGAN to ensure stability and diversity of samples. Local Outlier Factor and affine transformations also helped in improving the diversity of sparse samples [1].

Small object detection still lags behind larger objects despite the progress achieved. A new GAN architecture named DS-GAN introduced data augmentation by generating realistic small objects from larger ones and reduced the need for annotations [2].

Due to limited fault samples, fault diagnosis in an ESP suffers from the ill effect of data imbalance. MCGAN-VSG generates virtual samples with the support of multi-distribution mega trend diffusion, and its classification accuracy enhances the faults considerably [3].

Using a novel data augmentation algorithm, this work addresses the limited spot datasets for AI-enabled Acquisition, Tracking, and Pointing (ATP) systems in FSO communication. The algorithm integrates wavelet transform into F2GAN to enhance generated spot image diversity, evaluated with PSNR, SSIM, and LPIPS metrics [4].

Generative models have been recently used to address the issues of data scarcity, imbalance, and privacy through the notion of synthetic data. This paper presents an important GAN-based synthetic training model of GAN-ST, which can be used for generating data to train lightweight CNNs that, in turn, enhance the classification accuracy on MNIST and CIFAR datasets [5].

Generative Adversarial Networks (GANs) for generating GPR data to address the data scarcity problem for underground explosive detection. It produces high-quality synthetic GPR defect data by using techniques like Double Normalization, Adaptive Discriminator Augmentation, and a modified self-attention module toward improved detection performance [6].

GANs are used to augment the radar data to enhance the accuracy of human motion recognition. A semisupervised Triple-GAN model is proposed, taking data-label pairs as input and involving three players to improve classification with minimal labeled data and abundant unlabeled data [7].

The works considered previously are imbalanced data, GAN transfer learning, and GANs in the synthesis of skin lesion images. This work proposes a two-stage self-transfer GAN that synthesizes diverse skin lesion images, enhancing greatly the classification performance on imbalanced datasets [8].

Challenges in obtaining labeled data for fault diagnosis under abnormal conditions limit deep learning effectiveness. A hybrid data augmentation mechanism (HDAM) combines a multicategory generative adversarial network (MCGAN) with similarity-based selection to generate high-quality data, significantly enhancing recognition accuracy in few-shot scenarios [9].

Data scarcity in speech emotion recognition tends to overfit when deep models are applied to train a model. To address this, we here propose the Adversarial Data Augmentation Network as an architecture that comprises of GAN and autoencoder in an integrated manner [10].

Data augmentation can be categorized based on synthesis strategies. However, all the techniques fail to handle imbalanced data properly for industrial applications. EID-GAN is proposed to excel at synthesizing samples superior to the state-of-art models and tested using classification tasks in CNN and clustering K-Means algorithm [11].

For surface defect recognition in intelligent manufacturing, which is vision-based, insufficiently labeled samples become an obstacle for deep learning models. Con-GAN addresses the issue of synthesizing high-quality defect images from just 10 samples with the aid of a shared data augmentation module to combat overfitting and mode collapse [12].

Generative models' evolution, especially GANs, is considered to indicate their applications in image generation and data augmentation. Proposing Leaf GAN can generate high-quality images for grape leaf diseases and thereby address the challenge of the limited training images. In addition, the generator model degressive channels for effective creation of images [13].

The challenge of insufficient real training data for deep learning in remote sensing image object detection is addressed. A framework is developed to generate synthetic images of aircraft with ground-truth annotations, enhancing detection performance [14].

Exploring applications of Generative Adversarial Networks to medical imaging, GANs are used for lung-nodule classification and segmentation from chest X-rays to improve COVID-19 detection by generating synthetic chest X-ray images using an Auxiliary Classifier GAN named CovidGAN [15].

**Methodology:**

1. **Frame Work:**

The central problem involves generating high-quality, well-balanced synthetic data for HAR, which addresses problems of intraclass imbalance, the unavailability of samples in the required quantity, and fidelity in GAN-generated data. The priorities lie in enhancing model stability, refining conditional data generation, augmentation of limited samples, and setting up robust evaluation metrics. The advancements in these areas will further optimize GAN applications across diverse domains toward developing more accurate and adaptive HAR systems.

1. **Data Collection and Preprocessing**:

The accelerometry data from 395 students undertaking different activities. This dataset was specifically prepared to design activities and conditions within a human activity recognition (HAR) scenario.

The ability to manipulate the latent space in autoencoders helps synthesize the data through addition of noise and interpolations into compressed representations. SMOTE improves model performance because it's used on datasets with class imbalances. DDPMs use Gaussian noise for generating synthetic accelerometry data. GANs are considered some of the most powerful solutions when it comes to unsupervised time series data generation, whereas VAE has a promising role within a lot of synthetic data generation tasks. Together, these methods push the limits of data synthesis and augmentation.

1. **Model Implementation:**

Two main architectures were studied the conditional Generative Adversarial Networks (cGANs) and the conditional Wasserstein GANs (CWGANs).Quantum-Augmented Model Architecture.

The CWGAN was made with a generator and a discriminator. Generator generates synthetic data depending on random noise and activity ID. Discriminator evaluates the authenticity of that data. The architecture comprises six one-dimensional convolutional layers along with the Rectified Linear Unit (RELU) activation functions and batch normalization layers to improve training speed and stability.

**Conditional Generative Adversarial Networks (cGANs):**

cGANs can be viewed as a generalization of the standard GAN architecture that facilitates conditioned generation of synthetic data given relevant input information. In this work, conditioning information consisting of an activity ID along with random noise was fed into the generator so as to condition its synthesis of synthetic accelerometry data related to a specific activity. Conditioning in this way facilitates more meaningful and realistic data generation for HAR tasks.

The objective function of the CGAN is much the same as that of the original GAN except that it includes the additional information y that is the specific activity for which synthetic data is being generated. The output of the generator will now depend on the noise, as well as the activity ID, and can be represented as G(z, y) 1.

 

**Training Process of cGANs:**

At the training period, the generator generates synthetic data and attempts to develop data which cannot be distinguished from the real data by the discriminator. The discriminator scores the data it is given - whether the data received is real or synthetic and belongs to the correct class of activities. Dual evaluation improves quality data generated

The experiments revealed the positive ability of cGANs to produce synthetic samples that enhance the accuracy of HAR models much better than traditional signal processing techniques

**Wasserstein GAN:**

This paper discusses the WGAN, a variant of GAN with a new loss function, using the Wasserstein distance rather than the Kullback-Leibler divergence for added stability on the training process. For this study, the WGAN was conditionally modified, sometimes referred to as CWGAN, in order to generate synthetic data while keeping the advantages achieved with the conventional WGAN architecture, like higher convergence and stability from training.

Just like cGANs, the architecture of CWGAN allows the generator to provide synthetic samples that not only look realistic but also fall in specific classes. In contrast to the generator, the discriminator of CWGAN needs to be trained with additional steps further aiding in the refinement of generated data quality



**Comparison and Findings:**

The paper compared cGANs and CWGANs in terms of performance. It was concluded that CWGANs outperformed cGANs while generating accelerometry data. Superiority could be revealed in model performance as well as in the generated signals' similarity with real data.

The study highlighted that even if the both architectures are successful, stability and avoiding model collapse in CWGANs make it a more apt architecture for synthetic data generation for the context of HAR

CGANs utilize conditional information such as class labels and apply a traditional loss function for GAN. In practice, they are popular for generating very high-quality but less diverse outputs. Such models face instability issues in the training process and mode collapse as well. On the other hand, CWGANs also make use of conditional information but replace this traditional loss function with the Wasserstein distance. Therefore, it reduces the problem of gradient flow and comes with reduced training instability. They produce much more diverse samples of very high quality; thus, easier and much more reliable to train than that of CGANs.

 **Results**

The metrics of CGAN and CWGAN are definitely different and have clear trade-offs between both architectures. CGAN is significant in terms of classification-based metrics, such as 1-NN Accuracy, TPR, and TNR, since it outperforms CWGAN considerably in these areas. This further justifies that the cGAN is more efficient and capable of synthetic data generation, which is differentiated from real data in classification tasks. The score attained by CGAN is considerably high, especially 0.9453 for 1-NN Accuracy, 0.924 for TPR, and 0.9667 for TNR.

On the other hand, there is significant improvement in Inception Score (IS-Accuracy) by CWGAN, which scored 0.3146 whereas CGAN had very low scores 0.0569, indicating better quality data. Although CWGAN was performing comparable to CGAN on Kernel MMD-a measure of distribution similarity between real and generated data-CWGAN falls behind in the performance of a classification-based approach.

In short, CWGAN shines bright at high-quality, realistic synthetic data generation; in contrast, CGAN remains better at classifying real and synthetic data on clear distinctions. This would come under a basic trade-off: CGAN is more effective in places where accurate classification is of utmost importance, whereas CWGAN will be more suited to applications where high-quality generation of synthetic data is of great concern.

 ****

**Conclusion:**

Summarising it, this term paper emphasizes its innovative ability to produce synthetic data using conditional Wasserstein GAN for Human Activity Recognition. It reveals the potential of the synthetic data in benefiting the performance of the learning models in proportion to the size of datasets. The performance of small datasets improves with added synthetic samples. It emphasizes the importance of dataset size and also suggests that targeted data augmentation may be effective rather than uniform data augmentation across all classes. More importantly, CWGAN's ability to produce high-quality private-preserving data is a critical attribute for sensitive applications. Future work could integrate this with Digital Twining to produce realistic virtual agents, enhancing HAR and synthetic data applications further.

**References:**

1.H. Ding, N. Huang, Y. Wu and X. Cui, "LEGAN: Addressing Intraclass Imbalance in GAN-Based Medical Image Augmentation for Improved Imbalanced Data Classification," in IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1-14, 2024, Art no. 2517914, doi: 10.1109/TIM.2024.3396853.

2.Bosquet, B., Cores, D., Seidenari, L., Brea, V. M., Mucientes, M., & Del Bimbo, “A. (2023). A full data augmentation pipeline for small object detection based on generative adversarial networks”. Pattern Recognition, 133, 108998

3.Gao, X., Zhang, Y., Fu, J., & Li, S. (2024). “Data augmentation using improved conditional GAN under extremely limited fault samples and its application in fault diagnosis of electric submersible pump”. Journal of the Franklin Institute, 361(4), 106629.

4.Liu, Y., Liu, Y., Song, S., Chen, K., & Guo, L. (2024).” GAN-Based Data Augmentation for AI-Enabled ATP in Free Space Optical Communication.” IEEE Communications Letters.

5.Rather, I. H., & Kumar, S. (2024). “Generative adversarial network based synthetic data training model for lightweight convolutional neural networks. Multimedia Tools and Applications”, 83(2), 6249-6271.

6.Xiong, H., Li, J., Li, Z., & Zhang, Z. (2023). “GPR-GAN: A ground-penetrating radar data generative adversarial network”. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-14.

7.Liu, L., Wang, S., Song, C., Xu, H., Li, J., & Wang, B. (2023). “Radar-based Human Motion Recognition Using Semi-supervised Triple-GAN”. IEEE Sensors Journal.

8.Q. Su, H. N. A. Hamed, M. A. Isa, X. Hao and X. Dai, "A GAN-Based Data Augmentation Method for Imbalanced Multi-Class Skin Lesion Classification," in IEEE Access, vol. 12, pp. 16498-16513, 2024, doi: 10.1109/ACCESS.2024.3360215.

9.Y. Quan, C. Liu, Z. Yuan and B. Yan, "Hybrid Data Augmentation Combining Screening-Based MCGAN and Manual Transformation for Few-Shot Tool Wear State Recognition," in IEEE Sensors Journal, vol. 24, no. 8, pp. 12186-12196, 15 April15, 2024, doi: 10.1109/JSEN.2024.3372438

10.L. Yi and M. -W. Mak, "Improving Speech Emotion Recognition With Adversarial Data Augmentation Network," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 1, pp. 172-184, Jan. 2022, doi: 10.1109/TNNLS.2020.3027600.

12.Z. Du, L. Gao and X. Li, "A New Contrastive GAN With Data Augmentation for Surface Defect Recognition Under Limited Data," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-13, 2023, Art no. 3502713, doi: 10.1109/TIM.2022.3232649

13.B. Liu, C. Tan, S. Li, J. He and H. Wang, "A Data Augmentation Method Based on Generative Adversarial Networks for Grape Leaf Disease Identification," in IEEE Access, vol. 8, pp. 102188-102198, 2020, doi: 10.1109/ACCESS.2020.2998839.

14.W. Liu, B. Luo and J. Liu, "Synthetic Data Augmentation Using Multiscale Attention CycleGAN for Aircraft Detection in Remote Sensing Images," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 4009205,doi:10.1109/LGRS.2021.3052017.

15.A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman and P. R. Pinheiro, "CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection," in IEEE Access, vol. 8, pp. 91916-91923, 2020, doi: 10.1109/ACCESS.2020.2994762.