Synthetic Dataset Generation Using Generative Adversarial Networks (GANs)

DA312 Advance Machine Learning Lab: Course Project Proposal by

*Soumya Savarn (220150031)

I. Introduction

In modern machine learning, the availability of high-quality datasets is a critical factor for model performance. However, collecting, annotating, and preprocessing real-world data can be expensive, time-consuming, or impractical due to privacy or accessibility concerns. Generative Adversarial Networks (GANs) offer a solution by creating synthetic datasets that closely resemble real data distributions. This project aims to leverage GANs for synthetic dataset generation to address the challenges associated with real dataset limitations.

II. Objectives

- Develop and train GAN models for synthetic dataset generation.
- Assess the quality of synthetic data using metrics like Frechet Inception Distance (FID) and Inception Score (IS).
- Assess usability of synthetic data in ML tasks (e.g., classification).

III. Methodology

- **Dataset Selection:** Identifying a suitable real-world dataset as a reference.
- **GAN Development:** Implementing and training baseline GAN architectures (e.g., DCGAN, WGAN). Exploring advanced variants to improve performance.
- **Evaluation:** Analyzing the quality of generated data through quantitative metrics and visual inspections.
- **Applications:** Validate the utility of synthetic datasets in training machine learning models.

IV. Tennative Timeline

1) **Weeks 1-2:** Conduct a literature review and select a suitable dataset.
2) **Weeks 3-4:** Implement a baseline GAN and evaluate preliminary results.
3) **Weeks 5-6:** Experiment with advanced GAN architectures for enhanced performance.
4) **Week 7:** Assess the quality of generated datasets and compile results.
5) **Week 8:** Finalize the project report and prepare for submission.

V. Deliverables

- A trained GAN capable of generating synthetic datasets.
- Quantitative and qualitative evaluation results demonstrating data quality.
- A comprehensive project report detailing methodology, findings, and future directions.

VI. References

1) Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672-2680.
2) Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 6629-6640.
3) Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X. (2016). Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 29, 2234-2242.
4) Radford, A., Metz, L., Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

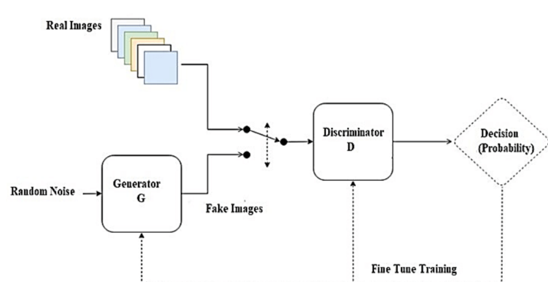

Fig. 1: A conceptual overview of Generative Adversarial Networks (GANs). The generator produces synthetic data to mimic real data, while the discriminator attempts to distinguish between real and synthetic data. Both networks are trained adversarially to improve their performance.