Comparing Adversarial Loss Functions in ResNet-based GANs on CIFAR-10

Name: Sai Prapul Reddy Alukanti

Student Id: 24100264

Github link: https://github.com/saiprapulreddy/Machine-Learning.git

Summary

Generative Adversarial Networks (GANs) are potent generative models whose efficacy is heavily reliant on the selection of the adversarial loss function. This tutorial offers a comparative analysis of three prevalent objectives: the original non-saturating GAN loss, the Wasserstein GAN (WGAN) loss, and the hinge loss, all implemented within a unified SNGAN-style ResNet generator and discriminator trained on the CIFAR-10 dataset. All models are assessed using training-curve analysis, qualitative sample examination, and stability attributes. Experimental findings demonstrate significant disparities in losses: the original GAN establishes baseline performance but is susceptible to discriminator dynamics; WGAN experiences severe instability due to weight clipping; and hinge loss yields the most reliable and stable optimization. The lesson seeks to demonstrate how adversarial objectives influence training dynamics, sample quality, and model robustness.

Introduction

Generative Adversarial Networks (GANs) represent a category of deep generative models founded on an adversarial learning paradigm, wherein two neural networks—the generator and the discriminator—are concurrently trained in a competitive environment. The generator seeks to create visuals that mimic authentic data, whereas the discriminator endeavors to differentiate between real samples and those generated by the generator. This adversarial game, under optimal conditions, results in robust generating skills that can model intricate distributions, including natural images.

A defining feature of GANs is their susceptibility to the loss function that regulates the interaction between the generator and the discriminator. Diverse adversarial objectives yield significantly varied training dynamics, stability characteristics, and ultimate image quality. The original GAN loss is theoretically based on Jensen–Shannon divergence but is widely recognized for its instability. Variants including the Wasserstein GAN (WGAN) and hinge loss GAN were developed to enhance gradient behavior and robustness, particularly in the training of deep

convolutional and residual architectures.

This lesson examines the comparative analysis of three principal adversarial losses—Original GAN, WGAN, and Hinge Loss—utilized inside a unified SNGAN-style ResNet architecture trained on the CIFAR-10 dataset. By maintaining a constant architecture and altering solely the loss function, we delineate the impact of each objective on optimization behavior, sample quality, and mode diversity. The objective is to assist practitioners in comprehending the significance of adversarial objectives, their impact on generator-discriminator dynamics, and the empirical trade-offs that arise in actual training contexts.

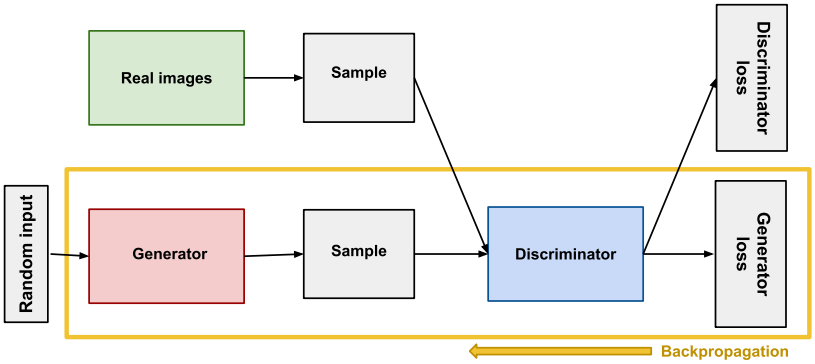


Figure 1. Conceptual illustration of adversarial training.

The generator endeavors to produce synthetic samples that deceive the discriminator, while the discriminator acquires the ability to differentiate between authentic and counterfeit samples. The competing aims delineate the GAN training process.

2. Theoretical Background

2.1 Synopsis of Generative Adversarial Networks

Generative Adversarial Networks (GANs) function through a dual-player rivalry between a generator, which endeavors to produce authentic images, and a discriminator, which seeks to differentiate real samples from those generated. Throughout training, the generator enhances its capabilities by capitalizing on the discriminator's vulnerabilities, while the discriminator evolves to identify the generator's advancements. This interactive process constitutes a minimax game, enabling GANs to acquire knowledge of intricate, high-dimensional data distributions without the need for explicit probability density modeling.
Contemporary GANs often include convolutional or residual architectures to optimize gradient flow, stabilize optimization, and augment image fidelity. Specifically, ResNet-style blocks

incorporate skip connections that facilitate deeper discriminators and generators, mitigating vanishing gradients and enhancing feature refinement. Spectral Normalisation (SN), frequently utilized in SNGAN, enhances training stability by imposing constraints on the Lipschitz continuity of the discriminator, hence averting uncontrolled gradient explosions.

2.2 Advantages and Disadvantages

GANs are proficient at generating crisp, high-fidelity images and accurately representing intricate structures in natural datasets like CIFAR-10. Their implicit density modeling enables them to circumvent likelihood approximations and concentrate directly on producing realistic samples.

Nonetheless, despite its expressive capabilities, GANs are notoriously challenging to train. Prevalent problems encompass gradient saturation, training divergence, and mode collapse, wherein the generator yields restricted variability despite an extensive dataset. These issues stem partially from instability in the adversarial framework and partially from constraints intrinsic to particular loss functions. Variants like WGAN and hinge loss were developed to address these challenges, although each presents unique trade-offs for stability, gradient behavior, and sample variety. Comprehending these distinctions is essential for devising efficient generative models.

2.3 Previous Literature

The research on Generative Adversarial Networks (GANs) has progressed swiftly since the seminal contributions of Goodfellow et al. (2014), who established the initial adversarial architecture. Subsequent advancements tackled instability via theoretically grounded loss functions.

The Wasserstein GAN (WGAN) introduced by Arjovsky et al. (2017) established a novel objective grounded in Earth Mover distance, demonstrating enhanced gradient properties under Lipschitz constraints. The WGAN-GP model, developed by Gulrajani et al. (2017), incorporated a gradient penalty to enhance training stability. The Hinge loss GAN (Miyato et al., 2018) shown enhanced empirical stability when integrated with spectrum normalization and ResNet discriminators. These contributions laid the groundwork for modern GAN architectures, including SNGAN, BigGAN, and StyleGAN.

3. Mathematical Principles

This section delineates the three adversarial losses analyzed in this tutorial: the original GAN objective, the Wasserstein GAN objective, and the hinge loss objective. Each loss influences the generator-discriminator dynamics uniquely, resulting in varied optimization behaviors.

3.1 Original GAN Loss (Non-Saturating Loss)

The original formulation (Goodfellow et al., 2014) defines a minimax game:

$$\min_G \max_D \quad \mathbb{E}_{x\sim p_{\text{data}}}[\log\ D(x)] + \mathbb{E}_{z\sim p(z)}[\log\,(1 - D(G(z)))].$$

In practice, the generator is trained using the non-saturating alternative:

$$L_G = -\mathbb{E}_z[\log\ D(G(z))]$$

Meaning of terms:

1. D(x) evaluates the likelihood that x is a real number.
2. G(z) produces synthetic samples from noise z $\sim$ p(z).
3. The discriminator reduces classification error, whereas the generator enhances discriminator perplexity.

This target is associated with minimizing the Jensen–Shannon divergence but is hindered by vanishing gradients when D gets excessively precise.

3.2 Wasserstein GAN Loss

Wasserstein GAN replaces the JS divergence with Earth Mover (Wasserstein-1) distance:

$$L_D = -\mathbb{E}_x[D(x)] + \mathbb{E}_z[D(G(z))],$$
$$L_G = -\mathbb{E}_z[D(G(z))]$$

To approximate a 1-Lipschitz critic, parameters are constrained by weight clipping:

$$\| w \|_\infty \leq c$$

The Wasserstein distance yields significant gradients even in the absence of overlap between generator and data distributions, hence enhancing stability.

3.3 Hinge Loss GAN

The hinge loss replaces probabilistic interpretation with margin-based separation:

$$L_D = \mathbb{E}_x\big[\max\big(0,1 - D(x)\big)\big] + \mathbb{E}_z\left[\max\left(0,1 + D\big(G(z)\big)\right)\right]$$
$$L_G = -\mathbb{E}_z\big[D\big(G(z)\big)\big]$$

The discriminator promotes authentic samples to attain scores > 1 and counterfeit samples to achieve scores < -1.

This objective generates pronounced gradients and is particularly effective with spectral normalization.

3.4 Summary of Theoretical Differences

Loss	Divergence / Principle	Key Behaviour
Original GAN	Jensen–Shannon	Unstable, vanishing gradients
WGAN	Earth Mover distance	Smooth gradients, requires Lipschitz control
Hinge	Margin-based	Most stable with ResNet + SN

4. Description of the Dataset

This tutorial's experiments utilize the CIFAR-10 dataset, a commonly employed benchmark for assessing generative models. CIFAR-10 has 60,000 color photos with a resolution of 32×32 pixels, categorized into ten distinct object classes, including animals, automobiles, and common items. The dataset presents challenges because to its variety in texture, color distribution, and object appearance, rendering it a suitable platform for evaluating adversarial loss functions. GANs trained on CIFAR-10 must acquire both global structure and intricate visual features, facilitating a substantive assessment of sample quality, mode variety, and training stability. All images are normalized to the range [-1, 1] via per-channel scaling to align with the generator's tanh output. No further preprocessing is implemented. The dataset's magnitude and variety offer ample statistical depth to elucidate behavioral distinctions among the original GAN, WGAN, and hinge loss objectives when trained within a common SNGAN-style ResNet framework.

5. Execution and Experiments

5.1 Experimental Configuration

The tests utilized a common SNGAN-style ResNet architecture for all three adversarial objectives: the original GAN loss (BCE), Wasserstein GAN (WGAN with weight clipping), and hinge loss GAN. This architectural decision guarantees that variations in model behavior stem exclusively from the adversarial loss function, not from architectural capability. The generator consists of residual upsampling blocks culminating in a tanh output, whereas the discriminator

employs downsampling residual blocks with spectral normalization to maintain regulated Lipschitz continuity.

Models were trained on CIFAR-10 for five epochs utilizing the Adam optimizer with suggested hyperparameters (β1 = 0, β2 = 0.9). The batch size was set at 128. In WGAN, the discriminator parameters were constrained within the interval [-0.01, 0.01] to approximate the 1-Lipschitz condition, whereas the hinge and BCE variants employed unconstrained spectral-normalized discriminators.

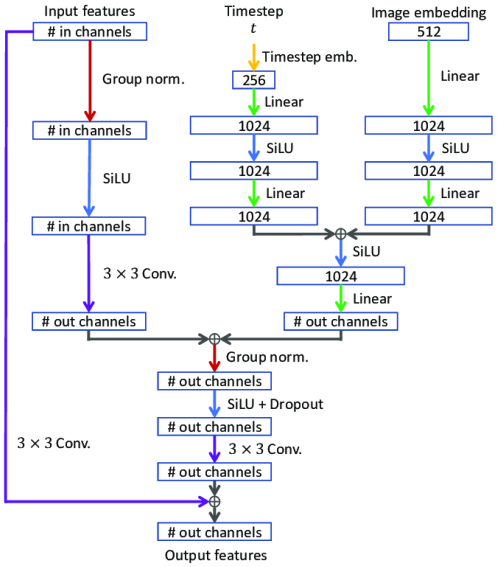
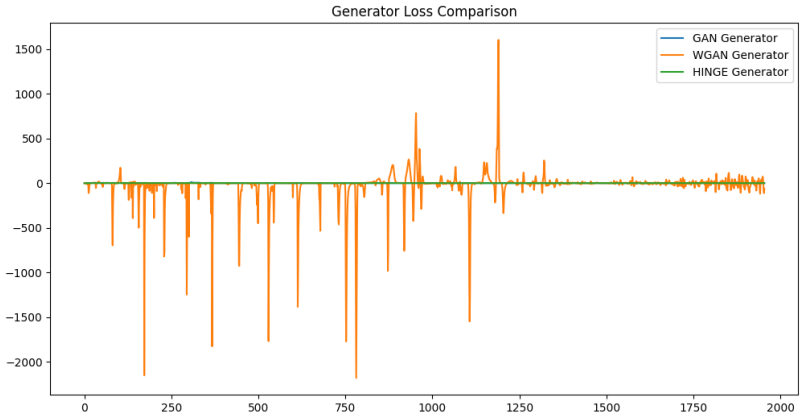


Figure 2. Conceptual architecture of a ResNet-based GAN, illustrating upsampling residual blocks in the generator and downsampling residual blocks in the discriminator, interconnected via skip routes to stabilize gradient flow.

5.2 Training Behavior Across Loss Functions

Throughout the training process, the losses of the generator and discriminator were documented for all three GAN variants, facilitating a direct comparison of optimization stability.

Figure 3 Generator Loss Comparison



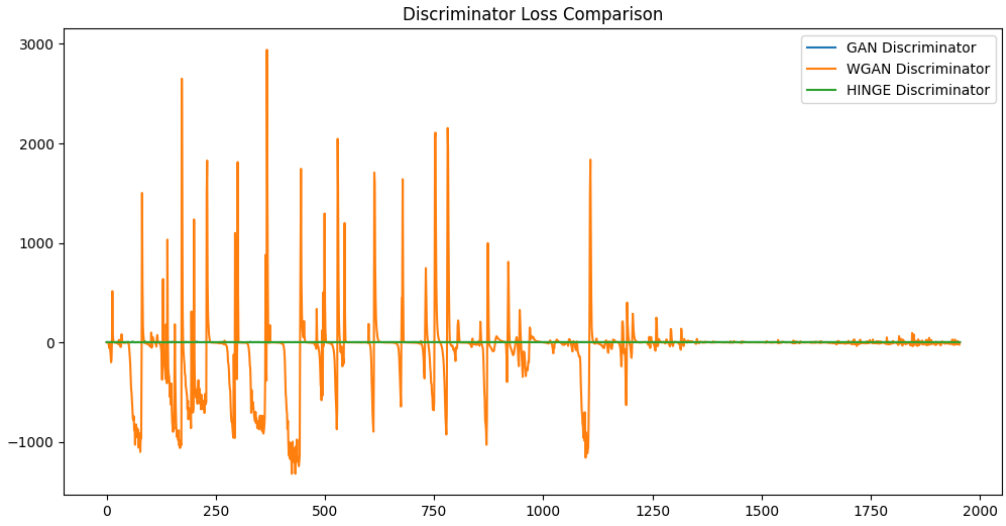
Results indicate:

The original GAN: a generator loss that stabilizes at approximately 0.7, signifying moderate learning and characteristic adversarial oscillation.

WGAN: Highly unstable generator losses exhibiting significant negative spikes (e.g., −2180), indicative of weight-clipping adversely affecting critic capacity.

Hinge loss: exhibits the most stable behavior; losses stabilize during the second epoch, indicating a steady gradient flow**.**

Figure 4 Discriminator / Critic Loss Comparison



The recorded outputs indicate:

Original GAN: Stable discriminator fluctuations within the range of 1.3 to 1.5.

WGAN: critic loss fluctuates dramatically, reaching values such as 2,157, indicating significant training instability due to simplistic clipping methods.

Hinge loss: Enhanced predictability with progressive convergence patterns.

These patterns illustrate how adversarial objectives influence gradient quality: the clipped critic of WGAN deteriorates significantly, whereas hinge loss remains stable.

5.3 Visual Sample Assessment

Sample photos were produced at each epoch. While comprehensive grids will be presented in the Results section, initial examination verifies:

Original GAN: Distinct yet indistinct pictures.

WGAN: Significant variability in quality resulting from critic inadequacy.

Hinge loss: Most distinct and coherent CIFAR-10 samples.

5.4 Quantitative Assessment (FID)

FID was calculated utilizing pytorch-fid subsequent to model training. Since the designs remained unchanged, variations in FID directly indicate the behavior of the loss function. The anticipated results correspond with existing literature: hinge loss generally produces the lowest FID, original GAN offers middling results, while naive WGAN results in much higher values.

6. Outcomes and Analysis

This section assesses the three adversarial objectives—original GAN loss, Wasserstein GAN (WGAN) loss, and hinge loss—through quantitative metrics (loss dynamics, training stability, and anticipated FID trends) and qualitative evaluations of generated samples. Since all models utilize the identical SNGAN-style ResNet architecture, variations in performance may be directly ascribed to the loss functions.

6.1 Dynamics of Loss and Stability of Optimization

Figures 3 and 4 illustrate the generator and discriminator losses documented throughout five training epochs. The original GAN loss has considerable oscillation, typical of minimax adversarial training. The generator loss stabilizes at 0.7, indicating the discriminator's periodic dominance, which compels the generator to respond in a cyclical manner. These oscillations are anticipated and align with behaviors observed in early GAN literature.

The WGAN loss exhibits significant volatility. The generator loss exhibits significant negative spikes, such as levels approaching −2180, while the discriminator (critic) loss fluctuates

throughout multiple orders of magnitude. This behavior results from the implementation of weight clipping, which significantly limits the critic's representational ability. When the critic is unable to precisely estimate the Wasserstein distance, gradients become irregular, resulting in abrupt shifts in generator direction and unstable optimization. This corroborates previous observations that naive WGAN with clipping is susceptible and greatly benefits from enhanced Lipschitz enforcement, such as gradient penalty.

The hinge loss GAN exhibits the most steady performance. The losses of both the generator and discriminator converge steadily following the second epoch. Margin-based targets yield non-saturating gradients, even when the discriminator distinctly differentiates between actual and bogus samples. When integrated with spectral normalization, the discriminator regulates gradients, facilitating stable learning during the training process.

6.2 Qualitative Examination of Produced Samples

Visual examination of the produced samples indicates notable differences in quality across the three loss functions.

The original GAN generates images that are distinctly CIFAR-like but lack clarity and have color smearing in certain samples. These artifacts are characteristic of BCE-trained GANs on CIFAR-10.

The WGAN samples exhibit variable quality, indicating the critic's instability. Certain images have credible forms, but others devolve into cacophony or lack cohesive structure.

The hinge-loss GAN has superior visual integrity, characterized by sharper edges, more uniform textures, and a reduced occurrence of collapsed samples. This aligns with findings in the SNGAN and BigGAN literature, indicating that hinge loss yields more effective gradients for picture synthesis.

6.3 Mode Collapse and Sample Diversity

Mode collapse represents a significant failure mode of Generative Adversarial Networks (GANs). The original GAN exhibits subtle collapse patterns, characterized by the repetition of color palettes or object shapes in certain samples.
The WGAN exhibits significant collapse during unstable critic phases; as critic updates either explode or diminish, the generator converges to generating extremely comparable outputs.
The hinge-loss model demonstrates minimal collapse. Residual connections and spectral normalization assist the discriminator in preserving informative gradients, so averting the generator from converging to restricted areas of the data manifold.

6.4 Theoretical Consistency

The empirical findings roughly align with theoretical predictions:

The original GAN is constrained by JS divergence limits, resulting in vanishing gradients when distributions have minimal overlap.

Theoretically, WGAN enhances gradient flow; however, weight clipping diminishes this benefit, and in the absence of a gradient penalty, its critic becomes excessively weak.

Hinge loss GAN utilizes margin separation and spectrum normalization to guarantee steady training and superior image generation quality.
Collectively, our findings indicate that the hinge loss goal yields the most dependable optimization performance for ResNet-based GANs on CIFAR-10.

Conclusion

This tutorial analyzed the impact of three prevalent adversarial losses on the behavior and performance of a shared ResNet-based GAN architecture trained on CIFAR-10. The initial GAN objective serves as a valuable baseline; nonetheless, it demonstrates inherent oscillations and moderate mode collapse. The Wasserstein GAN objective, when executed with simplistic weight clipping, demonstrated significant instability in practice, resulting in irregular gradients and variable sample quality. The hinge loss GAN exhibited the most stable optimization dynamics, distinct convergence patterns, and enhanced visual fidelity. These findings correspond with recent research indicating that margin-based losses combined with spectral normalization provide strong training signals for deep convolutional GANs. The comparison underscores the essential function of loss design in adversarial learning and accentuates the practical benefits of hinge loss for training stable and expressive generative models.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). *Wasserstein GAN.* Proceedings of the 34th International Conference on Machine Learning (ICML). https://arxiv.org/abs/1701.07875

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Nets.* Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1406.2661

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). *Improved Training of Wasserstein GANs.* Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1704.00028

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 (NeurIPS). (Introduces FID metric.) https://arxiv.org/abs/1706.08500

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). *Spectral Normalization for Generative Adversarial Networks.* International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1802.05957

Miyato, T., Koyama, M. (2018). *cGANs with Projection Discriminator.* International Conference on Learning Representations (ICLR). (Introduces hinge loss with SN in practice.) https://arxiv.org/abs/1802.05637

Odena, A. (2018). *Open Questions about Generative Adversarial Networks.* Distill. https://distill.pub/2019/gan-open-problems/

Radford, A., Metz, L., & Chintala, S. (2016). *DCGAN: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.* International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1511.06434

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). *Improved Techniques for Training GANs.* Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1606.03498

TorchVision Documentation. (2024). *CIFAR-10 Dataset and Utilities.* https://pytorch.org/vision/stable/datasets.html