The paper "Evaluation of Generative Adversarial Network (GAN) Performance Based on Direct Analysis of Generated Images" introduces a novel metric called the **Creativity-Inheritance-Diversity (CID) index** to evaluate the performance of GANs. Unlike traditional evaluation methods that rely on pretrained classifiers (e.g., Inception Score or Fréchet Inception Distance), the CID index directly assesses generated images based on three core principles:

Key Evaluation Criteria

1. **Creativity**:
    - Measures whether generated images are distinct from the training images.
    - Evaluates overfitting by ensuring generated images are not simple duplicates of real images.
    - Computed by removing near-duplicate images (using Structural Similarity Index, SSIM) and finding the percentage of unique outputs.
2. **Inheritance**:
    - Ensures generated images retain stylistic features of the training set, capturing the "essence" of the input data.
    - Measured using the Gray-Level Co-occurrence Matrix (GLCM) contrast to compare textures between real and generated images.
3. **Diversity**:
    - Assesses variety among generated images.
    - Uses clustering based on SSIM to calculate the entropy of image distributions, with higher entropy indicating more diversity.

Comprehensive Metric

The CID index combines the three individual measures as:

CID=Creativity×Inheritance×Diversity.CID = \text{Creativity} \times \text{Inheritance} \times \text{Diversity}.CID=Creativity×Inheritance×Diversity.

Comparison with Traditional Metrics

- **Inception Score (IS)**: Measures inheritance and diversity using a pretrained network but ignores creativity and is dependent on specific datasets (e.g., ImageNet).
- **Fréchet Inception Distance (FID)**: Measures distributional similarity between real and generated images but assumes Gaussianity of feature vectors.
- **1-Nearest Neighbor Classifier (1NNC)**: Compares distributions using a leave-one-out accuracy but requires equal-sized real and generated datasets.

Experiments and Findings

- Tested on three GAN models: DCGAN, WGAN-GP, and SNGAN, using synthetic datasets with different image types.
- Results:
    - DCGAN often overfits, producing low creativity scores.
    - SNGAN generates diverse outputs but struggles with inheritance (failing to capture input style).
    - WGAN-GP achieves the best balance among creativity, inheritance, and diversity, earning the highest CID index.

Strengths of CID

- Evaluates GANs without relying on external classifiers.
- Captures overfitting issues more effectively than IS and FID.
- Provides interpretable results by isolating performance aspects (creativity, inheritance, diversity).

Applications and Future Work

- The CID index can be extended to other GAN architectures and datasets.
- More robust thresholds for SSIM and advanced measures for inheritance and diversity could refine evaluations.
- The methodology could help in training GANs by identifying specific weaknesses during development.

Certainly! Let's explore every part of the paper **"Evaluation of Generative Adversarial Network (GAN) Performance Based on Direct Analysis of Generated Images"** in detail.

Abstract

- **Objective**:
    - The paper proposes a new evaluation method for GANs that directly analyzes generated images rather than using them as input for classifiers.
    - It introduces the **Creativity-Inheritance-Diversity (CID)** index to measure three key aspects of GAN performance:
        1. **Creativity**: Ensuring images are not duplicates of real images.
        2. **Inheritance**: Ensuring generated images retain the style and key features of real images.
        3. **Diversity**: Ensuring generated images are varied among themselves.
- **Comparison**: The CID index is compared with traditional GAN evaluation methods like **Inception Score (IS)**, **Fréchet Inception Distance (FID)**, and **1-Nearest Neighbor Classifier (1NNC)**.

Introduction

1. **Background**:
    - Generative Adversarial Networks (GANs) were introduced by Goodfellow et al. in 2014 and have become a state-of-the-art technique in deep learning for image generation.
    - Over 500 types of GANs have been proposed, and many applications have been found in image processing (e.g., image-to-image translation, super-resolution, etc.).
2. **Problem**:
    - While significant progress has been made in developing GAN architectures, fewer studies focus on **evaluating their performance effectively**.

- Existing measures often rely on classifiers or statistical methods, such as IS and FID, which may not fully capture the characteristics of ideal GAN outputs.

3. **Proposed Solution**:
    - A new evaluation framework, the CID index, is proposed. It emphasizes:
        - Creativity (avoiding overfitting to training data),
        - Inheritance (maintaining stylistic similarity to input data),
        - Diversity (ensuring a wide range of outputs).

Methods

A. GAN Evaluation Metrics

- **Ideal GAN Output**:
    - Realistic and varied images matching the distribution of training data.
    - Three criteria:
        1. **Non-duplication of real images (Creativity)**.
        2. **Retention of style and features of real images (Inheritance)**.
        3. **Variety among generated images (Diversity)**.
- **Common Problems**:
    - **Overfitting**: The GAN memorizes and replicates training images.
    - **Mode Collapse**: The GAN produces limited variations of outputs.
    - **Mode Dropping**: Some data modes in the training set are ignored.

B. Proposed Measures

1. Creativity:

- Measures whether the generated images are distinct from real ones.
- Method:
    - Use **Structural Similarity Index (SSIM)** to compare real images (RRR) and generated images (GGG).
    - Duplicates are identified if SSIM ≥0.8\geq 0.8≥0.8 (a threshold chosen empirically).
    - Creativity is defined as: Creativity Index=|G′||G|\text{Creativity Index} = \frac{|G'|}{|G|}Creativity Index=|G||G′|

        - Where G’G'G’ is the subset of generated images that are not duplicates.

2. Inheritance:

- Measures the stylistic similarity between generated and real images.
- Method:
    - Use **Gray-Level Co-occurrence Matrix (GLCM)** contrast values to capture texture information.
    - Compute the average contrast for real (RRR) and generated (GGG) images.
    - Normalize the difference between their contrast values:
      Inheritance Index=1−|CR−CG|CR\text{Inheritance Index} = 1 - \frac{|C_R - C_G|}{C_R}Inheritance Index=1−CR|CR−CG|
        - Where CRC_RCR and CGC_GCG are the average GLCM contrast values for RRR and GGG, respectively.

3. Diversity:

- Measures the variation among generated images.
- Method:
    - Cluster generated images using SSIM.
    - Compute the entropy of the cluster distribution:
      Diversity Index=−∑i=1mpilog⁡(pi)\text{Diversity Index} = - \sum_{i=1}^m p_i \log(p_i)Diversity Index=−i=1∑mpilog(pi)
        - Where pip_ipi is the proportion of images in the iii-th cluster, and mmm is the total number of clusters.
    - Higher entropy indicates more diversity.

4. CID Index:

- Combines the three measures into a single score:
  CID=Creativity×Inheritance×Diversity\text{CID} = \text{Creativity} \times \text{Inheritance} \times \text{Diversity}CID=Creativity×Inheritance×Diversity

---

Comparison with Existing Metrics

1. **Inception Score (IS)**:
    - Measures inheritance and diversity using a pretrained Inception network on the ImageNet dataset.
    - Limitations:

        - Does not evaluate creativity (cannot detect overfitting).
        - Dataset-dependent and unsuitable for non-classification tasks.

2. **Fréchet Inception Distance (FID)**:
    - Measures the difference between real and generated image distributions using feature vectors extracted from the Inception network.
    - Limitations:
        - Assumes Gaussian distributions, which may not hold in practice.
        - Relies on the pretrained Inception network.
3. **1-Nearest Neighbor Classifier (1NNC)**:
    - Compares distributions of real and generated images using a leave-one-out accuracy score.
    - Limitations:
        - Requires equal numbers of real and generated images.
        - Sensitive to local distribution properties.

Experiments and Results

Experiment 1: One Image Type with DCGAN

- Trained a DCGAN using a single class of real images (Plastics).
- Evaluated performance with varying numbers of generated images.
- Results:
    - CID, IS, FID, and 1NNC were stable for larger sets of generated images (>1000).
    - CID captured overfitting better than IS or FID.

Experiment 2: Multiple Image Types with Three GANs

- Used four image types (Holes, Small Leaves, Big Leaves, Plastics) to train three GANs: DCGAN, WGAN-GP, and SNGAN.
- Generated 1200 images for each type and computed all metrics.
- Observations:
    - **DCGAN**: High inheritance but low creativity (overfitting).
    - **WGAN-GP**: Balanced performance with the highest CID score.
    - **SNGAN**: High creativity but low diversity (mode collapse).

Key Insights:

- CID reflects GAN performance more comprehensively by penalizing overfitting (low creativity).
- Visual evaluations align better with CID than with IS, FID, or 1NNC.

Discussion

- CID index addresses key limitations of traditional measures by:
    - Evaluating generated images directly without requiring pretrained classifiers.
    - Capturing overfitting through the creativity metric.
    - Being applicable to diverse datasets and GAN architectures.
- Limitations:
    - GLCM is texture-specific and may not generalize to non-texture datasets.
    - The SSIM threshold (0.8) requires further study for optimal values.

Conclusion

- The CID index provides a robust and interpretable method for evaluating GANs.
- It captures the three critical aspects of GAN performance: creativity, inheritance, and diversity.
- This approach has potential applications in GAN training and performance tuning.