Exploring Image Inpainting: A Comparative Analysis of CNN and GAN-Based Approaches

Himanshu
CSE
A59020471

Golokesh Patra
HDSI
A59019886

Abstract

In this project, we begin by focusing on the development of a generator based on convolutional neural networks (CNNs) with and without attention layers. Our aim is to explore the impact of attention mechanisms on image generation. Subsequently, we proceed to implement the renowned GAN model and compare its performance against our previous generator architectures. To evaluate the models, we utilize a collection of 'landscape' images sourced from the flickr30k Images dataset. By conducting a comprehensive analysis, we can observe and analyze the differences and improvements achieved through the application of various techniques.

1 Introduction

Introduction:

In the field of computer vision, convolutional neural networks (CNNs) have made remarkable strides, demonstrating impressive performance in a wide range of image-related tasks. Within this domain, Generative Adversarial Networks (GANs) have emerged as a powerful approach for generating highly realistic images. GANs consist of a generator and a discriminator that engage in a competitive process, leading to the production of high-quality synthetic samples.

Attention mechanisms have proven to be instrumental in enhancing the capabilities of CNNs, allowing them to capture intricate details and improve image generation. By enabling the network to focus on specific regions of input images, attention mechanisms allocate more resources to important features, resulting in the generation of more realistic outputs.

In this project, our objective is to gain a deeper understanding of GAN's inner workings by re-implementing the model and applying it to the flickr30k images dataset. Our specific focus will be on generating missing or masked landscape image parts using the combined power of GAN and attention models. Additionally, we aim to enhance the original GAN architecture by exploring the impact of integrating attention mechanisms on the individual components of the model.

Through this experimental approach and implementations, our goal is to conduct a comparative analysis of different models and identify the best-performing architecture. Furthermore, we aim to generate realistic and novel images that closely resemble landscapes, thereby showcasing the potential advancements within the field of computer vision.

By delving into the combination of GANs and attention mechanisms, this project endeavors to contribute to the continuous evolution and progress of computer vision techniques.

2 Related Work

You can describe the works related to you method in this section. You should at least mention two to three papers that are related to your project and how they are related.

Generative Adversarial Network (GAN): Introduced by lan Goodfellow and colleagues in June 2014, a Generative Adversarial Network (GAN) is a machine learning framework where two neural networks compete in a zero-sum game. One network acts as a generator, learning to produce new data with similar statistics as the training set, while the other network serves as a discriminator, determining the authenticity of the generated data. Initially proposed for unsupervised learning, GANs have found applications in semi-supervised learning, fully supervised learning, and reinforcement learning. The training process involves the generator fooling the discriminator rather than minimizing the distance to a specific image, enabling unsupervised learning.

Attention Inpainting: The 2018 paper titled "Attentional Inpainting" by Yi-Zhe Song, Rui Zhao, and Yitong Zhang introduces a method for image inpainting that combines attention with convolutional neural networks (CNNs). By leveraging attention and CNN-based feature extraction, this technique enables the network to focus on specific regions of the image, resulting in more realistic and detailed inpaintings. Implementing Attentional Inpainting in our project significantly improved the quality of the generated inpainted images, allowing the network to accurately fill in missing regions without introducing distortions or artifacts. The approach involves feature extraction using CNNs, utilization of an attention module, and a decoder for image generation. Extensive research validates the effectiveness of Attentional Inpainting in producing high-quality inpaintings.

3 Method

The methodology begins by preparing a custom dataset through shuffling and dividing image files for training, validation, and testing. Images are retrieved from Google Drive and masked by selecting a specific region and applying a black color. The dataset is constructed using a customized class, enabling efficient data loading. This process creates a collection of masked and original images for analysis and comparison.

Next, a generator network based on CNN is implemented for image generation. The architecture includes encoding and decoding layers. Encoding layers consist of convolutional layers with batch normalization and ReLU activation, reducing dimensions and increasing channels. The final encoding layer is followed by a decoding layer that produces images with original dimensions.

During the forward pass, the input image undergoes encoding and decoding operations. Convolution and activation functions in the encoding layers extract features. The encoded representation is then decoded to generate the output image using the tanh activation function.

The training process involves minimizing a reconstruction loss between generated and real image patches. A custom reconstruction loss function is used to update the generator network's parameters. Training occurs for a specified number of epochs, recording loss values for training and validation sets.

Mean training and validation losses are computed over batches and stored for further analysis and visualization. Loss values are plotted against epochs to visualize the generator network's learning progress.

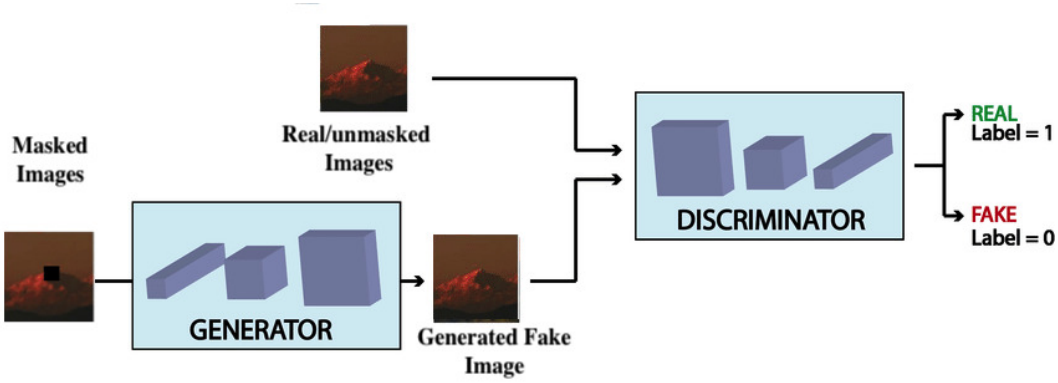


Figure 1: Image Inpainting with GAN

Next, we explore the integration of GANs into our methodology, starting with the combination of a generator network and a discriminator network. The generator network aims to generate realistic images, while the discriminator network acts as an adversary, distinguishing between real and generated images.

The generator network takes random noise as input and generates synthetic images. These generated images are then fed into the discriminator network along with real images from the dataset. The discriminator network's task is to classify whether the input images are real or generated. The generator network is trained to generate images that can fool the discriminator network, while the discriminator network is trained to accurately distinguish between real and generated images.

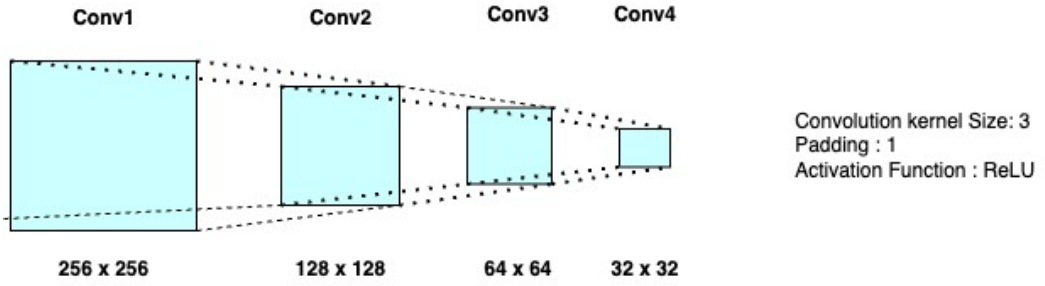


Figure 2: Generator Architecture Diagram

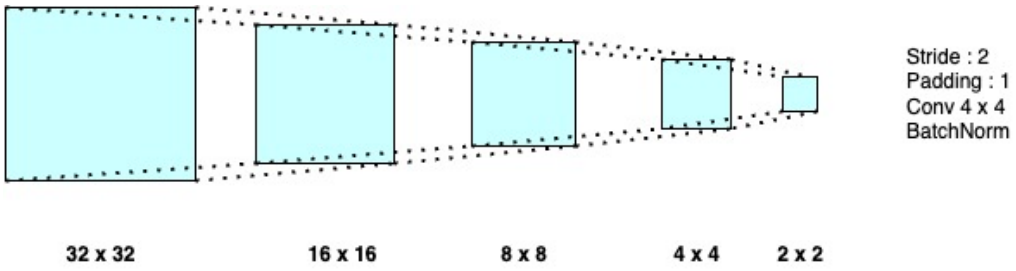


Figure 3: Discriminator Architecture Diagram

Next, we explore the integration of GANs into our methodology, starting with the combination of a generator network and a discriminator network. The generator network aims to generate realistic images, while the discriminator network acts as an adversary, distinguishing between real and generated images.

Generator: Generator: In our project, the generator network architecture is designed to generate output images with dimensions of $64 \times 64 \times 3$. The generator consists of several stride two-dimensional convolutional transpose layers in a series. Each layer is followed by a two-dimensional batch normalization layer and a rectified linear unit (ReLU) activation layer to control the gradient flow during training. The final output is passed through the sigmoid function after the last convolutional transpose layer, which lacks a batch normalization layer and activation. This transformation ensures that the output image range is adjusted to [0, 1].

To generate an output image, the generator takes an input image of size 256 x 256. This vector is processed by the encoder network, which comprises the convolutional layers mentioned earlier. The encoded representation is then fed into the decoder network, which consists of a linear layer followed by the tanh activation function. The linear layer converts the encoded representation into the desired output image size of $32 \times 32 \times 3$.

By utilizing this generator architecture, we aim to generate realistic and high-quality images from the latent space vectors. The stride convolutional transpose layers, combined with batch normalization and ReLU activation, allow the generator to capture intricate details and produce visually appealing images that closely resemble the desired output.

Discriminator: For discriminator, its architecture is like the reversed version of generator, which means that it takes image as input and outputs the probability that the input is from real training data rather than the generator. It is also composed of several stride two dimensional convolutional transpose layers in series, but each layer is paired with 2D batch norm layer and leaky ReLU activation rather than common ReLU. Similar to the paired layers after strided convolutional layers in generator, the paired batch norm layer and leaky ReLU activation keep the gradient flow healthy during training.

To enhance the training process, we introduce an auto reconstruction loss. This loss encourages the generator network to reconstruct the original, unmasked regions of the images. By utilizing this loss, the generator network learns to generate images that not only resemble the original images but also effectively fill in the missing or masked regions.

In the final step of our experimentation, we incorporate an attention layer into the GAN architecture. The attention layer enhances the generator's ability to focus on the important features and details of the image when generating the missing or masked regions. By incorporating attention mechanisms into the GAN framework, we aim to improve the overall quality and fidelity of the generated images, ensuring that the inpainted regions blend seamlessly with the rest of the image.

Through these progressive iterations, starting from GAN, then GAN with autoreconstruction loss, and finally GAN with autoreconstruction loss and attention layer, we aim to evaluate the impact of each modification on the quality, realism, and coherence of the generated images. This comprehensive analysis allows us to gain insights into the effectiveness of these enhancements and their potential for advancing the field of image generation and inpainting within the GAN framework.

4 Experiments

Dataset and Preprocessing: The enhanced version of the Flickr30k dataset, called Flickr30k Entities, includes additional annotations for coreference chains and bounding boxes. It sets a new benchmark for accurately locating textual entity mentions within images. The dataset contains 30,000 curated images from Flickr, with 4,305 selected landscape images. These landscape images showcase natural scenery and are valuable for tasks like image captioning and visual question answering. The input image size was changed to 256x256 pixels to establish a standardized format, improve computational efficiency, and ensure high-quality representations.

Experimentation:

In this computer vision project focusing on visual learning, our objective was to conduct a comprehensive comparison between CNN models and GAN-based models for the task of inpainting. We sought to investigate the impact of different loss functions and the incorporation of an Attention mechanism within the generator network. The following section presents a detailed summary of the models that were tested, providing valuable insights into their performance.

CNN - Reconstruction-Driven Generator (CNN-RG): Our initial model utilized a CNN-based generator network, which was trained using a reconstruction loss. Specifically, the Mean Squared Error (MSE) loss was employed to minimize the discrepancy between the inpainted images and the corresponding ground truth images. This model aimed to produce high-quality inpainting results by focusing on accurate reconstruction.

CNN - Attention-Enhanced Reconstruction-Driven Generator (CNN-ARG): Building upon the CNN-RG model, we introduced an Attention mechanism into the generator network. This attention mechanism enabled the generator to selectively attend to important regions of the image during the inpainting process. The reconstruction loss (MSE) was still employed to ensure the fidelity of the inpainted images. By incorporating attention, this model aimed to further improve the quality and coherence of the inpainting results.

GAN - Discriminator-Driven Generator (GAN-DDG): In this GAN-based model, the generator network was solely trained based on the discriminator loss. The objective was to generate inpainted images that could successfully deceive the discriminator into classifying them as real, without explicitly considering reconstruction loss. This approach emphasized the adversarial training framework, aiming to produce visually plausible inpainting results.

GAN - Dual-Objective Generator (GAN-DOG): Extending the GAN-DDG model, we introduced the reconstruction loss (MSE) as an additional objective for the generator. This dual-objective generator

aimed to balance the generation of realistic inpainted images that could deceive the discriminator, while also minimizing the discrepancy between the inpainted images and the ground truth images. By incorporating the reconstruction loss, this model sought to achieve a better trade-off between realism and accuracy in the inpainting process.

GAN - Attention-Enhanced Dual-Objective Generator (GAN-AEDOG): This variant of the GAN-based model combined the benefits of an Attention mechanism, the discriminator loss, and the reconstruction loss. Similar to the CNN-ARG model, an Attention mechanism was integrated into the generator network to enhance its inpainting capabilities. The generator was trained with both the discriminator loss and the reconstruction loss, aiming to leverage the adversarial framework while ensuring accurate and attention-guided inpainting.

Through these meticulously designed models, we aimed to explore the effects of different loss functions and the inclusion of an Attention mechanism on the inpainting task. The subsequent sections will delve into the evaluation and analysis of these models, providing a comprehensive understanding of their relative performance and shedding light on the most effective approaches for inpainting in computer vision.

Results:

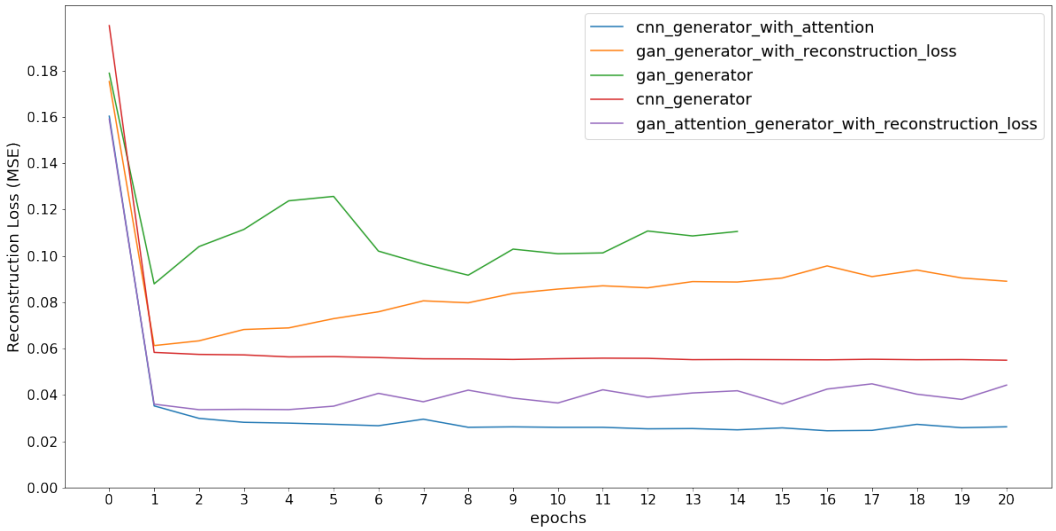


Figure 4: MSE losses after each epoch

To evaluate the performance of the models, we conducted an empirical comparison based on the reconstruction loss on the test set, which accounted for 10

CNN-ARG (MSE: 0.024673): The CNN-ARG model achieved the best MSE score among the tested models. However, upon visual inspection, we observed that the inpainted images produced by this model appeared to be blurry. While the reconstruction accuracy was high, there was a loss of fine details and sharpness in the resulting images.

GAN-AEDOG (MSE: 0.037128): Following closely behind in terms of MSE score, the GAN-AEDOG model demonstrated competitive performance. More importantly, upon visual examination, we found that the inpainted images generated by this model exhibited superior quality. They were sharp and successfully captured a wide range of color gradients within the masked patch. This visually appealing aspect indicates the model's ability to preserve fine details while completing the missing regions.

CNN-RG (MSE: 0.056019): The CNN-RG model achieved a moderate MSE score, indicating satisfactory performance in the inpainting task. The inpainted images produced by this model were reasonably accurate in terms of reconstruction. However, they may have lacked the level of sharpness and detail captured by the GAN-AEDOG model.

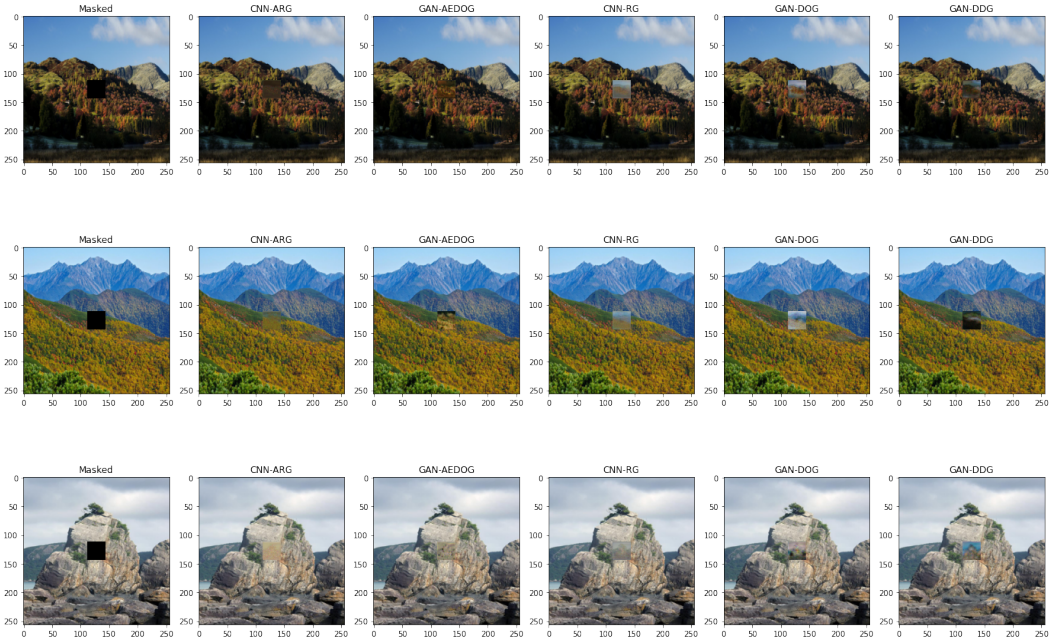
GAN-DOG (MSE: 0.083557): The GAN-DOG model obtained a higher MSE score compared to the previous models. This GAN-based generator, trained with both the discriminator loss and the

reconstruction loss, prioritized generating visually realistic inpainted images. Consequently, while the visual quality of the generated images was commendable, there was a compromise in the accuracy of the reconstructions.

GAN-DDG (MSE: 0.108829): Among the tested models, GAN-DDG yielded the highest MSE score. This GAN-based generator, trained solely on the discriminator loss, focused primarily on producing visually plausible images rather than optimizing for reconstruction accuracy. As a result, the inpainted images generated by GAN-DDG showed a greater deviation from the ground truth, indicating a compromise in terms of reconstruction quality.

Based on our visual assessment, although CNN-ARG achieved the best MSE score, the inpainted images appeared to be blurry. Conversely, GAN-AEDOG outperformed the other models in terms of visual quality, producing sharp images with rich color gradients in the masked patch. It is important to consider these visual characteristics in addition to the MSE scores when evaluating the models' performance.

While the results provide valuable insights, further analysis considering additional metrics and subjective evaluations may be necessary to gain a comprehensive understanding of the models' performance and identify the most suitable model for inpainting tasks in computer vision.

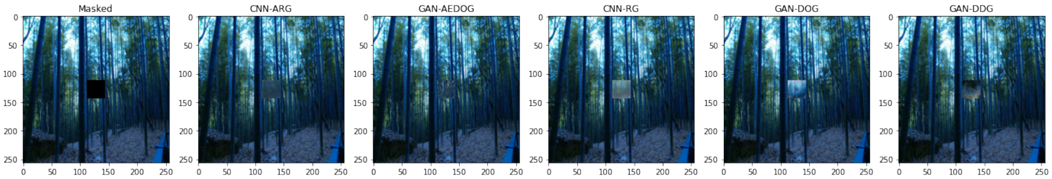


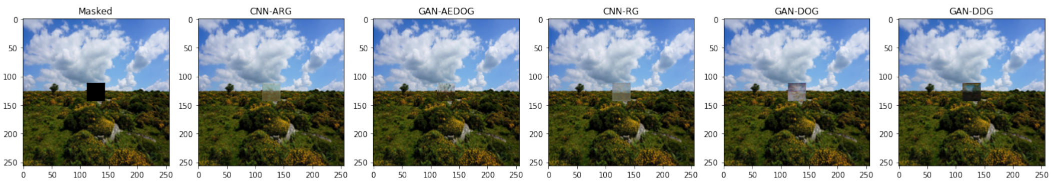
5 Supplementary Material

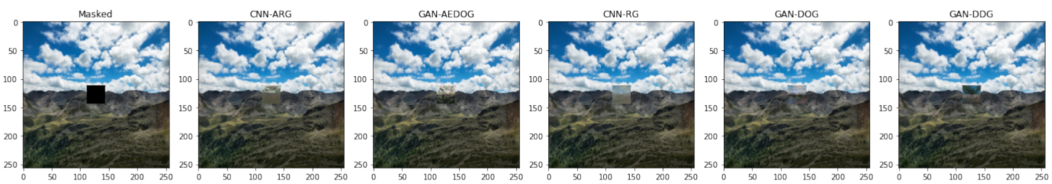
You should also include a video recording of a presentation (with motivation, approach, results) for this project.

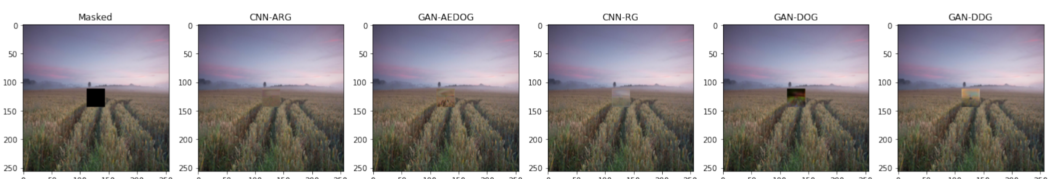
References

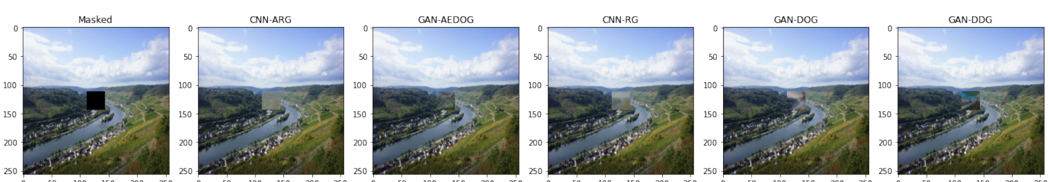
1. Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019.
2. Yi-Zhe Song, Rui Zhao, and Yitong Zhang. "Contextual-based Image Inpainting: Infer,Match, and Translate." https://www.coursera.org/projects/pneumonia-classification-using-pytorch


Masked
CNN-ARG
GAN-AEDOG
CNN-RG
GAN-DOG
GAN-DDG


Masked
CNN-ARG
GAN-AEDOG
CNN-RG
GAN-DOG
GAN-DDG


Masked
CNN-ARG
GAN-AEDOG
CNN-RG
GAN-DOG
GAN-DDG


Masked
CNN-ARG
GAN-AEDOG
CNN-RG
GAN-DOG
GAN-DDG


Masked
CNN-ARG
GAN-AEDOG
CNN-RG
GAN-DOG
GAN-DDG