Synthetic Data Generation

saiprasanth paladugula

May 2023

1 Introduction

The objective of the project is to Generate Synthetic data from real data. The kind of data we are dealing with currently and text, and image data. We will be using Generative adversarial networks, and diffusion models to generate fake images and for the text currently, we are using BERT models but in the future, we will expand the scope.

2 Synthetic Data Generation Models

2.1 Image Generation models

2.1.1 StyleGAN3

StyleGAN is a powerful generative adversarial network architecture designed to create high-quality synthetic images. Developed by NVIDIA in 2018, StyleGAN uses a style-based generator that separates the high-level structure of the image from the low-level details, allowing for greater control over image attributes. The generator takes a random noise vector as input, which is transformed by a learned mapping network into a "style code" that controls the image's features. The model progressively adds detail to the image through upsampling layers and includes a noise input at each layer to add variation and randomness to the output. The discriminator is a convolutional neural network that distinguishes between real and generated images and encourages diversity to avoid mode collapse.

StyleGAN3 is an improved version of the original algorithm that introduces several innovations to enhance the model's ability to handle spatially variant features, varying resolutions, and spatially varying styles and object structures. One of these innovations is the Attention Augmentation module which enhances the model's ability to handle spatially variant features, while the Adaptive Pooling module allows the model to better handle images of varying resolutions. The Modulated Convolutions module enhances the model's ability to model spatially varying styles and object structures.

2.1.2 Stable Diffusion

Stable diffusion is a recent development in the field of generative models that aims to improve the stability and convergence properties of diffusion-based generative models. Diffusion-based models are a class of generative models that operate by iteratively applying a series of diffusion processes to a random initial state, resulting in a sequence of intermediate states that converge to the target distribution. Stable diffusion improves upon the original diffusion models by introducing a new regularization term that penalizes the deviation of the intermediate states from the initial state, resulting in more stable and accurate convergence.

To generate synthetic images using stable diffusion, the model first takes a random initial state and applies a series of diffusion processes to generate a sequence of intermediate states. The model then uses a learned mapping function to map the final intermediate state to an output image, which approximates the target distribution. By adjusting the number of diffusion steps and the strength of the regularization term, the model can control the level of detail and quality of the generated images. Stable diffusion has shown promising results in generating high-quality images across a variety of domains, including natural images, medical images, and molecular structures.

2.1.3 StyleSwin

StyleSwin is a deep learning model developed by Microsoft Research Asia for generating synthetic images with high visual quality and diversity. It is an extension of the popular StyleGAN architecture, which separates the control of high-level features (such as pose, lighting, and facial expression) from low-level details (such as texture and color). StyleSwin incorporates a novel "Swin Transformer" module, which is a type of attention mechanism that allows the model to capture global and local features in the image.

To generate synthetic images using StyleSwin, the model takes as input a random noise vector and a set of "style codes" that control the high-level features of the image. The model then applies a series of transformation layers that progressively add detail and complexity to the image. The final output is a high-resolution synthetic image that closely resembles real images of the same category (e.g., faces, animals, landscapes). StyleSwin has been shown to outperform previous state-of-the-art generative models on a variety of image synthesis tasks.

2.1.4 Conditional Variational Autoencoder

Variational autoencoders (VAEs) are a type of neural network that generate synthetic images by learning a low-dimensional representation of the input data called a latent space. The input data is encoded into the latent space by an encoder network and then decoded by a decoder network to produce the synthetic output. During training, the VAE minimizes the difference between the generated output and the original input while also encouraging desirable properties in the latent space.

Conditional variational autoencoders (cVAEs) are a type of VAE that generate synthetic data samples based on given input conditions or labels. They learn to map input images to a latent space and reconstruct the original image from the latent space using the conditions or labels as input. To generate synthetic images using a cVAE, the user provides input conditions or labels to the model, which maps them to the corresponding latent variables in the latent space. A random noise vector is then used to generate a sample from the latent space, and the decoder maps the latent variables and conditions to a synthetic image that corresponds to the given conditions or labels. The resulting synthetic images have similar attributes to the input images but also reflect variations introduced by the random noise vector.

2.1.5 Deep Pix2Pix GAN

DeepPix2Pix GAN is a deep learning-based model used for image-to-image translation tasks. It consists of a generator network and a discriminator network that are trained in an adversarial manner to generate high-quality synthetic images. The generator network takes an input image and generates a corresponding output image that is as similar as possible to the ground truth. The discriminator network is trained to distinguish between real and synthetic images. The generator network is updated to generate more realistic images, while the discriminator network is updated to correctly classify real and synthetic images.

It generates synthetic images from real images using a conditional GAN architecture, which takes both the input image and the desired output image as input to the generator network. The generator network then maps the input image to the output image in a pixel-to-pixel manner, effectively translating the input image into the output domain.

During training, the generator network is trained to minimize the difference between the generated output image and the ground truth output image, while the discriminator network is trained to distinguish between the generated output image and the ground truth output image. The training process results in a generator network that can produce realistic-looking synthetic images that closely resemble the desired output images.

2.1.6 Progressive Growing of GAN's

The "Progressive Growing of GANs" is a deep learning technique for training generative adversarial networks (GANs) to generate high-resolution images. The method involves gradually increasing the resolution of the generated images during training, starting from a low resolution and gradually adding more detail. This is done by progressively adding layers to the generator and discriminator networks, starting from a low-resolution base and adding layers to generate more details. This allows the GAN to generate high-resolution images with a level of detail that is not possible with traditional GANs. The technique has been used to generate high-quality synthetic images of faces, landscapes, and other objects.

The method involves training the GAN on a dataset of real images and using it to generate synthetic images that are similar to the real images. The generator network takes a random noise vector as input and maps it to an image, while the discriminator network tries to distinguish between the generated images and the real images. The GAN is trained in an adversarial manner, with the generator network trying to fool the discriminator network by generating images that are indistinguishable from real images, and the discriminator network trying to correctly classify real and synthetic images. By progressively adding layers to the generator and discriminator networks during training, the GAN is able to generate higher-resolution images with more details.

2.1.7 MediGAN

The MedIGAN is a generative adversarial network (GAN) designed to generate synthetic medical images, such as X-rays and MRIs, for use in medical research and diagnosis. MedIGAN is designed to address the challenges of generating high-quality medical images, such as low data availability, variability in imaging modalities, and the need for anatomical accuracy. MedIGAN utilizes a novel architecture that incorporates both generator and discriminator networks with attention mechanisms, as well as a task-specific loss function to ensure anatomical accuracy.

MedIGAN is trained on a large dataset of real medical images and learns to generate synthetic images that are visually similar to real images. The generator network is trained to produce images that are as similar as possible to the real images, while the discriminator network is trained to distinguish between real and synthetic images. MedIGAN has been shown to generate high-quality synthetic images that can be used for medical research and diagnosis, such as augmenting limited datasets and reducing the need for invasive procedures.

2.2 Text Generation Models

2.2.1 Google T5 Transformer

The Google T5 (Transformers in 5 sizes) is a state-of-the-art natural language processing (NLP) model developed by Google. It is based on the Transformer architecture, which is a neural network architecture that was introduced in a 2017 paper by Vaswani et al. The T5 model is pre-trained on large amounts of text data and can be fine-tuned on a variety of downstream NLP tasks, such as question answering, language translation, and text classification.

One application of the T5 model is to generate fake text from real text using a technique called text generation or text completion. This involves giving the T5 model a prompt or starting sentence and then asking it to generate a continuation of the text. The T5 model can generate fake text that is highly coherent and can resemble human-written text. This is achieved by training the T5 model to learn the statistical patterns and structure of language in the training data, and then using this knowledge to generate new text that is similar in structure and style to the original text. However, it is important to note that the T5 model can generate text that is not factually accurate or ethical, and as such, it should be used with caution.

2.2.2 GPT-3

GPT-3 (Generative Pre-trained Transformer 3) is a natural language processing model developed by OpenAI. It is a state-of-the-art language model that uses deep learning techniques to generate human-like text. GPT-3 is pre-trained on a large corpus of text data and can be fine-tuned for specific natural languages processing tasks, such as language translation, summarization, and question-answering.

GPT-3 generates fake text from the real text by using a deep-learning architecture called a transformer. The transformer model is designed to learn from large amounts of text data and generate coherent and fluent text in response to a given prompt. GPT-3 is trained on a massive dataset of diverse text, which allows it to generate text that is similar in style and tone to human-written text. It uses a technique called "unsupervised learning" to learn the patterns and structures of natural language and generate text that is grammatically correct and semantically meaningful.

2.2.3 BART-large

BART (Bidirectional and Auto-Regressive Transformer) is a pre-trained model developed by Facebook AI that can generate coherent and fluent text. It uses a combination of auto-regressive and denoising auto-encoding techniques to generate text that incorporates the meaning and context of the input text. BART is capable of various natural language processing tasks, such as text summarization, question answering, and text generation. BART Large has more parameters than BART.

To generate fake text from real text, BART can be fine-tuned on a specific task using a large dataset of real text. During fine-tuning, BART is trained to generate text that is similar to the real text in the dataset. Once the model is fine-tuned, it can generate new text that is similar to the real text, but with some variations.

2.2.4 BART-large-cnn

BART-large-CNN (Convolutional Neural Network) is a larger and more powerful version of the BART model. It is a pre-trained sequence-to-sequence model developed by Facebook AI and is based on transformer architecture. It uses a combination of auto-regressive and denoising auto-encoding techniques to generate text, similar to the base BART model. However, it also incorporates a convolutional neural network architecture, which allows it to better capture the local dependencies and structures in the input text.

To generate fake text from real text, BART-large-CNN can be fine-tuned on a specific task using a large dataset of real text. During fine-tuning, the model is trained to generate text that is similar to the real text in the dataset. Once the model is fine-tuned, it can generate new text that is similar to the real text, but with some variations.

2.2.5 Pegasus

PEGASUS is a state-of-the-art text generation model developed by Google. It is a variant of the transformer architecture and is pre-trained on a massive amount of text data. PEGASUS generates text by using a process called "pre-training with reconstruction objectives," which involves reconstructing masked-out parts of a document based on the remaining context.

During pre-training, PEGASUS is trained to generate a summary of a long document given only a few input sentences. This process is called "abstractive summarization," and it involves generating a summary that captures the essential information of the input document in a concise and coherent manner. Once pre-trained, PEGASUS can be fine-tuned on specific tasks, such as fake text generation or summarization, by providing it with a small amount of task-specific data.

2.2.6 Summarizer-clinical-jsl

Summarizer-clinical-jsl is a pre-trained model developed by John Snow Labs that can be used for text summarization tasks in the clinical domain. It is based on transformer architecture and is trained on a large dataset of clinical text to generate coherent and concise summaries. The model takes in a long piece of clinical text as input and outputs a shorter summary that captures the main points of the input text.

2.2.7 Summarizer-clinical-jsl-augmented

Summarizer-Clinical-JSL-Augmented is trained on a larger dataset that includes additional data augmentation techniques to improve performance.

Specifically, Summarizer-Clinical-JSL-Augmented is trained on a dataset of 100,000 clinical notes, which is larger than the dataset used to train Summarizer-

Clinical-JSL. Additionally, Summarizer-Clinical-JSL-Augmented uses data augmentation techniques such as paraphrasing, back-translation, and word replacement to further enhance the quality of the training data and improve the performance of the model.