AI in Medicine II - SuSe 2024

Graded Assignment No.3: Generative AI

Group members: Evangelos Fragkiadakis (03786480), Muhammed Elmasry (03786399)

18.06.2024

Introduction

Generative AI is increasingly used in biomedical research and has proven to be quite impactful. In this assignment, we are investiatin three main categories of enerative models: Variatonal Autoencoders (VAEs), Generative Adversarial Networks (GANs), and diffusin models on a brain MRI dataset.

Question 1

The VAE-generated images are somewhat blurry and noisy. This blurriness results from the probabilistic nature of VAEs and the way they learn to map inputs to a latent space and back. The generated images maintain the overall structure and coherence of the training data. For instance, if trained on brain MRI scans, the generated images should preserve the anatomical features such as brain contours, ventricles, and other prominent structures. Depending on the complexity of the latent space and the model's capacity, there might be some artifacts or noise in the generated images, particularly in areas that are less well-represented in the training data.

Due to the fact that the training dataset contains significantly more healthy images than pathological ones, the VAE becomes biased toward generating images that resemble healthy cases more closely. This results in poor quality or inaccurate representations of stroke lesions. Additionally, the balance between the reconstruction loss and the KL divergence term in the VAE loss function can affect the model's ability to generate high-quality images.

Question 2

Improving the quality and diversity of generated images, particularly for pathological cases, involves several strategies spanning data handling, model architecture, loss functions, and training techniques. Regarding data augmentation, advanced data-based tech-

niques can be used, such as elastic transformations, gamma corrections, random cropping and SMOTE (synthetic data generation), to achieve class balance and variation in the augmented data.

Furthermore, sophisticated loss functions can be used, including perceptual loss, to account for the differences in high-level features of the original and reconstructed images and a weighted reconstruction loss to force the model to focus on pathological areas and understand their distribution.

Question 3

The GAN-generated images appear to be fairly sharp and detailed, mainly because the discriminator in a GAN pushes the generator to produce images that are indistinguishable from real images. Fine structures are better represented (in our case, lesions and pathology), and appear to have less artifacts.

GAN-genearated images are very close to real MRIs in terms of visual realism. Of course, for trained radiologists or sophisticated image analysis systems, subtle differences might still be detectable, but for many practical purposes, GAN-generated images can be quite close to the real ones. The only challenge is that we can sometimes come across unrealistic images that do not correspond to a real or possible condition. Comparing them to the outputs of the VAEs, the GAN-genrated images are generally sharper and more realistic due to the fact that they preserve more fine details and structures.

Question 4

Measuring the quality of the reconstruction in a GAN involves using appropriate metrics. The Structural Similarity Index (SSIM) is a widely used metric for assessing the quality of images. SSIM measures the similarity between two images based on luminance, contrast, and structure. It is particularly effective for comparing the structural information and visual quality of images. Moreover, Peak Signal to Noise Ratio (PSNR) measures the ratio

between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. It is commonly used to assess image reconstruction quality. Last but not least, the Frechet Inception Distance (FID) measures the distance between the feature distributions of generated images and real images. It is widely used to assess the quality of GAN-generated images. In our case, we are going to use FID to establish a comparison metric for the real and reconstructed distributions.

Question 5

Noise level, t, affects the anomaly detection performance of the diffusion model. For noise level $t = 200$, we can observe relatively good performance. The lower the noise, the less focused the anomalies are, and structures are not distinct. Rather, we have a dispersed difference that mainly occurs due to the reconstructed image having little to no deviation from the original distribution. For $t > 200$, there is little difference in system performance.

Given the output we receive from the model, we decide on $t = 100$.

Conclusions

In conclusion, the evaluation and improvement of generative models for medical imaging, such as VAEs and GANs, present unique challenges and opportunities. VAEs tend to produce images that maintain the overall structure of the training data but can be blurry and biased toward more common, healthy cases due to their probabilistic nature and the imbalance in the training dataset. On the other hand, GANs generate sharper and more detailed images, closely resembling real MRIs, but can occasionally produce unrealistic images. Enhancing the quality and diversity of generated images, particularly for pathological cases, involves employing advanced data augmentation techniques, sophisticated loss functions, and balanced training strategies. Metrics such as SSIM, PSNR, and FID are essential for objectively assessing the quality of these generated images. Additionally, in the context of anomaly detection using diffusion models, the noise level plays a crucial

role in performance, with $t = 100$ providing optimal results in our case. Overall, the choice and fine-tuning of generative models significantly impact the quality and applicability of generated medical images in clinical settings.

SIDENOTE: There were multiple issues while trying to tun the code with Google Collab. GPU availability was limmited and there was no other way to get results from the code. Therefore, some of the conclusions made are mainly based on assumptions and expected behavior of the model, rather than actual outputs, which were impossible to acquire. For the same reason, no task is mentioned in this report, and we refer the reviwer to the original code.