GSOC 2023 Proposal

Project Title - Creating Synthetic MRI Data

First Name	Vara Lakshmi
Last Name	Bayanagari
Timezone	EST (UTC-5)
Github	lb-97

I am a Master's student from NYU, currently working as Graduate Research Assistant at NYU Langone. As part of my Graduate Thesis, I've worked on Sex Classification using Diffusion MRI data based on deep learning approaches. As part of this research, I worked on Contrastive based self-supervised learning methods on CNNs and Masked Modeling on Vision Transformer. I developed a novel training strategy using 2DCNNs that decreased training time by 50% and helped boost the AUC to 0.98, more than that of Transformers counterpart. We interpreted the results by visualizing occlusion analysis on 30 ROI between male and female subjects. In this process, I learnt and refined skills pertaining to writing clean code using PyTorch, data pre-processing, deploying models with GPU, storing experiment results and most importantly interpreting results. I wish to learn and extend my knowledge to this project during the summer.

Introduction

DIPY has been a growing Imaging library for 3D and higher tensors spanning Classical image processing techniques as well as Deep Learning based processing. A collection of datasets are available as part of the library, therefore synthetic conditional MRI images can help enhance the richness of DIPY. Generative AI has grown exponentially in the past few years and has also seen some applications in the field of Medical Image generation. One of the many advantages of synthetic MRI data is the independence from limited availability of human data that contains identifiable sensitive information. Many existing medical datasets require rigorous training and paperwork by the users before availing the data, thereby limiting the sharing capabilities among different research institutions. Also, the volumes created through Diffusion Generation have been proved to be more diverse in structural similarity than by GANs. Past works show that Deep Learning models pretrained on this diverse synthetic data have helped boost Breast segmentation accuracy with higher DICE similarity. A similar trend in images produced through Diffusion Modeling has also been seen in Fundus Images, Ultrasound Images etc.

Abstract

The aim of this project is to contribute to the open source community in synthetic image generation for MRI images. The advantages of Diffusion Models have been witnessed only recently in the field of Medical imaging[1,2,3] leaving scope for more exploration. For GSOC 2023, we intend to design models that can produce images conditioned on damage size, location, type etc. A probable use case for this project would be to perform inpainting to introduce a desired tumor or lesion. This would help in training large-scale data hungry models such as ViT to perform downstream tasks. And also help avoid training on multiple datasets sourced from different machines with varying parameters nor rely on Imagenet pre-trained models that suffer from change in domain distributional shift.

My contribution to DIPY community -

1. #2733[Merged][Bug Fix] - Updated SynRegistrationFlow workflow to allow saving inverse transformed static image when using *dipy_align_syn* command. Earlier, only forward transformed/warped static files were saved even with explicit *–out_inv_static* option. This bug fix takes care of the same.

2. #2769[PR review pending][Documentation] - Updated the documentation for adding cluster information while saving tractogram(*save_tractogram*) using *StatefulTractogram* property *data_per_streamline*. One of the use cases for the same is when creating own clusters of streamlines by using either existing tractogram or from scratch. This documentation will help in having more control over the streamlines by identifying each of these new clusters with a label.

3. #2770[PR review pending][Bug Fix] - *dipy_horizon* command runs HorizonFlow workflow in interactive mode to open a temporary window to interact with 3D diffusion brain map. This command throws an error message when there's an existing file with name 'tmp.png' in the current directory, although unnecessary in non-stealth mode. Figured the root cause for it and updated the workflow with an overriding *manage_output_overwrite* method.

I have started working with DIPY since my introduction to GSOC 2023, so I am fairly new to its usage. Through my contribution to this repository in the past few weeks I have tried to understand its applications and implemented them locally. Two out of three of my contributions are Bug Fixes on workflows which require a good amount of understanding of the codebase. For example, during my first BF issue in DIPY, firstly I went through the doc and tutorial page of the website to understand varying applications such as pre-processing tools, brain manipulation, tractography and more. I then understood that *dipy_align_syn* aligns a moving image to a reference static image using the SyN algorithm that employs diffeomorphic transformation to preserve topology and other properties. As I explored and gained understanding on different registration techniques such as the ones based on sum-of-squares metric, I believe dipy's

implementation of SyN based on similarity metric could be very useful in analyzing complex DiffusionMRI properties. I am thrilled to learn more about the library and continue contributing to it.

Timeline -

Week 1-2:

- Familiarize with the MRI data and the relevant literature on conditional Generative AI models for medical imaging.
- Collect and preprocess the MRI data, including data cleaning, data normalization, and data augmentation.
- Explore and visualize the data to gain insights into its characteristics.
- Define the conditional variable(s) to be used in the generative model, such as age, gender, or diagnosis.

Week 3-4:

- Choose and implement a deep learning framework to build the conditional generative AI model, such as TensorFlow or PyTorch.
- Select an appropriate architecture for the generative model such as conditional Diffusion Latent model.
- Train and evaluate a baseline model on the MRI data, and optimize the hyperparameters.

Week 5-6:

- Experiment with different MRI data - Structural Vs Diffusion MRI data.
- Conduct experiments to investigate the model's sensitivity to different factors, such as the amount of training data, the architecture of the model, or the choice of loss function.

Week 7-8:

- Fine-tune the model on the MRI data and validate the model's performance using a held-out dataset or cross-validation.
- Evaluate the model's ability to generate realistic and useful synthetic MRI data conditional on the specified variable(s).
- Refine the model's architecture and hyperparameters based on the results of the experiments.

Week 9-10:

- Conduct a final round of experiments to validate the model's performance and compare it to existing methods.
- Document the model's performance and limitations, and record the results in a scientific paper or report.

Week 11-12:

- Finalize the documentation and prepare the model for release as open-source software.

Again, this timeline is just a rough estimate, and the actual timeline may vary depending on the complexity of the data and the model, the results obtained and interpretability of the model. Allotted last two weeks to have some buffer time for unexpected issues or delays that may arise during the project.

Limitations-

Current idea is to work on Structural MRI data primarily, observing the generative capability conditioned on one variable - damage type. The outcome of this project depends on training duration, computational requirements, data availability and evaluation schemes. If time permits, experiments can be extended to multiple data modalities(such as FA, MD, RD in Diffusion MRI data) accommodating other conditional variables.

Other commitments-

My semester ends on May 15th and my EAD work visa starts at the end of June. I do not have any commitments during the summer and I would be able to devote 40 hr/week during the time of GSOC project duration.

References-

1. Firas Khader and Gustav Mueller-Franzes and Soroosh Tayebi Arasteh and Tianyu Han and Christoph Haarburger and Maximilian Schulze-Hagen and Philipp Schad and Sandy Engelhardt and Bettina Baessler and Sebastian Foersch and Johannes Stegmaier and Christiane Kuhl and Sven Nebelung and Jakob Nikolas Kather and Daniel Truhn, “Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation”, Image and Video Processing 2023
2. Junde Wu and Rao Fu and Huihui Fang and Yu Zhang and Yehui Yang and Haoyi Xiong and Huiying Liu and Yanwu Xu, “MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model”, CVPR 2023

3. Walter H. L. Pinaya and Petru-Daniel Tudosiu and Jessica Dafflon and Pedro F da Costa and Virginia Fernandez and Parashkev Nachev and Sebastien Ourselin and M. Jorge Cardoso, “Brain Imaging Generation with Latent Diffusion Models”, arXiv, 2022
4. Jonathan Ho and Ajay Jain and Pieter Abbeel, “Denoising Diffusion Probabilistic Models”, 2020
5. William Peebles and Saining Xie, “Scalable Diffusion Models with Transformers}”, CVPR 2023
6. Axel Sauer and Katja Schwarz and Andreas Geiger, “StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets”, 2022
7. Lillian Weng, “What are Diffusion Models?”, https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
8. Wei Li, “Understanding the Denoising Diffusion Probabilistic Model (DDPMs), the Socratic Way
“https://towardsdatascience.com/understanding-the-denoising-diffusion-probabilistic-model-the-socratic-way-445c1bdc5756