Brain Tumor Image Analysis using Self-supervised Learning

A proposal by Anirudh Suri and Varsha Srinivasan

Problem Investigation

A significant drawback and issue with working with biological datasets lie in the complexities of labeling the data. Owing to the intricate nature of the data arising from the complex biological phenomenon, it is a task to attain features or labels and, even more, to label the data accurately. As we know, the accuracy/ output of any ML model boils down to the quantity and quality of data upon which it is trained. In most cases (context: only biological data), securing large datasets is not a problem, but ensuring that the data has extremely accurate labels or features can prove to be a task. Additionally, the amount of unlabelled data for biological samples is intrinsically significantly more than labeled data.

In most cases, human intervention is required to assign labels (which can result in faulty features), and in the case of vast datasets close to impossible. Furthermore, annotating biological datasets (image, tabular, etc.) requires only an expert's insight into the field to decipher and distinguish between features. This proves to be a rate-limiting step in computational research and the discovery of Biological problems.

Another issue when trying to deploy ML models for Biological datasets is that the pre-trained models become highly redundant due to the extremely high degree of variation in biological samples. To emphasize this: two healthy individuals won't have the same immune profile yet lack any disease. Thus building one model for either sample won't be effective due to the large amount of generalization required to accommodate both samples.

This has been a problem that has plagued computational biologists for a while. We will explore overcoming this issue by analyzing MRI scans of brain tumor patient samples (brain tumor segmentation - gliomas) coupled with context restoration.

Work

A recent novel trend that is gaining popularity and has been identified as a solution to this issue is self-supervised learning. In a nutshell, the model (CNN) identifies features on its own accord rather than can be used to calculate weights for downstream analysis (classification, segmentation, etc.).

We were inspired by the Multimodal Brain Tumour (BRaTS) Segmentation Challenge conducted by the University of Pennsylvania's Medical School in collaboration with the Computer Science Dept. The outcomes enable 1) Segmentation of intrinsically heterogeneous (shape, histology) MRI scans and 2) prediction of the patient's OS. Most of our referencing and knowledge stems from the contributors of this organization/ initiative, along with other resources such as research and review papers published in this domain.

Data

The data required can be sourced from the Centre of Biomedical Image Computing and Analytics webpage. This process involves authentication and permission grants to access the data. There are several years of backup data. If a smaller dataset is to be used (owing to

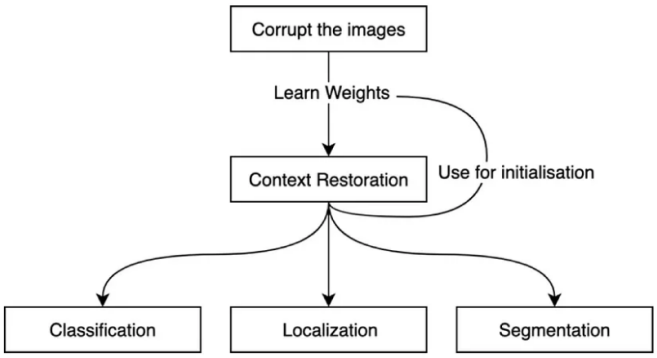
computational utilities not being able to uphold the tasks), a smaller but just effective data set is available from Kaggle. The file format for either is hdf5.

Methodology/ Algorithm

As described in the problem statement, the issue (dealing with large unlabeled data sets) applies to image data too. In this example (MRI images), we have found the implementation of self-supervised learning to be an ideal candidate for overcoming most of the issues,

Self-supervised models can create labels from large unlabeled data sets. It extracts and uses available contexts naturally to estimate the features. In short, to get supervision from the data or image itself. It still falls under the bracket of supervised learning, but the user does not provide the labels as part of the dataset; instead, the model defines the labels independently for further downstream analysis, such as classification and segmentation. Another advantage of this technique is that the model can now predict weights to initialize any downstream CNN-based tasks with the limited labeled data. What does this solve? Dealing with large unlabeled biological datasets and enriching a sample-specific model (trained and initialized over the same data)

We plan on improving this by amalgamating self-supervised learning with context restoration to the image data. That is, corrupting the image with noise or modifications to the original image and restoring it to its original state. The model can decipher the optimum weights required for subsequent tasks during this restoration process.



Source

Evaluation

We aim to compare our method and model against the ones published by the BRaTS consortium and explore how context restoration and self-supervised learning models compare to other models that perform the same task (MRI brain scans and segmentation). This would be done with box plots (as of now).

Contributions

As both team members conceptualized the idea, we aimed to tackle the problem and work together. We believe instead of delegating tasks, it would be cohesive and aid us in gaining valuable skills if we worked on each task together. If the proposed plan does not work, another

project we found equally interesting shall be explored (ML models applied to immune profile data sets of cancer patients to explore prediction and prognosis) and worked upon. If we continue to hit roadblocks, we would switch projects and work on a review paper pertaining to this primary project, as we would already have a fair amount of foundational knowledge concerning the topic.