Winter in Data Science, 2022

Brain Tumour Segmentation using multi-modal 3D Magnetic Resonance Imaging

Shantanu Welling

January 2023

Contents

1 Introduction

Problem statement is about Brain tumour segmentation & detection from Magnetic Resonance Imaging using multi-modal brain images. This was considered due to various reasons such as availability of dataset, existing research, and my interest to work with a 3D segmentation problem and analyse various methods of solving it, for example, breaking it into 2D problem of multiple slices but that would lead to loss of context and information between the slices along with intensive effort in image registration; breaking it into smaller 3D units (voxels), that would also lead to loss of information but using cross validation and hyperparameter tuning of the sub-unit dimensions, a desirable result could be achieved. I reviewed multiple research papers and shortlisted 3 of them for their techniques. I plan to use the 3D U-Net Architecture.

2 Issues & Measures

Some measures that I took in this project in order to counter the issues.

1. **Class Imbalance**
    - Prevalence of some conditions (frequency of particular classes being significantly more than others), leading to imbalance in the trainset causing those classes to contribute more towards the loss.
    - This leads to the model to train with priority given to these classes, neglecting the other classes.
    - So, the model will only be able to predict conditions related to the majority classes with greater accuracy and other minority classes with very poor accuracy.
    - Resolved with weighted loss i.e. the use of Binary cross-entropy loss assigned with weights.
    - Ideally, we would train our model using an evenly balanced dataset so that the positive and negative training cases would contribute equally to the loss.
    - If we use a normal cross-entropy loss function with a highly unbalanced dataset, then the algorithm will be incentivized to prioritize the majority class, since it contributes more to the loss.
    - Could also be resolved by resampling but it is deprecated as it may not include all datapoints of a majority class and may cause repetition of a single datapoint of minority class.
2. **Dataset Issues**
    - Patient overlap- distinctive features leading to model memorization i.e. model may learn patient-specific features that are not generalizable to other patients.
    - Possible resolution is using set sampling i.e. patient wise sorting and splitting, but that may lead to majority patients with no condition into train set, so the model won't be trained to identify a condition effectively.
    - So, set the train set should include at least X% cases belonging to some medical condition class.
3. **Dataset Augmentation**
    - Rotating, shifting, scaling, intensity, contrast change.
    - Colour noise
    - Flipping & inversion is deprecated as it changes the spatial orientation & labels of the 3D images.
4. **Artifacts**
    - Partial volume effect
    - Motion artifacts (requires image registration)
    - Gibbs Ringing (size of acquisition matrix)
    - Wraparound
    - Bias field noise (intensity in-homogeneities in the RF field)

3 Architecture Reviews

1. **Transformer + UNet (semantic segmentation)**

   - Due to the limitation of convolution kernel size, each convolution kernel only focuses on local information. Therefore, it is difficult for these existing methods to generate any long-distance dependencies when performing image segmentation tasks. The ability to construct global contextual information is crucial for intensive prediction tasks during medical image segmentation.
   - To effectively address the issues on global contextual information, Transformer was proposed to handle the issues in sequence-to-sequence prediction.
   - A one-dimensional sequence is taken as input, so Transformer has a powerful modelling ability, not only in constructing global context information. But it only focuses on building global context information at all stages. Therefore, its ability to obtain local information is weakened, and the lack of detailed location information encoding reduces the distinguishability between background and target.
   - UNet provide a way to extract low-level visual information, which can well compensate for the spatial details of Transformer's local information.
   - TransUNet first used CNNs to extract local features, and then applied Transformer to global context modelling. This architecture not only establishes a self-attention mechanism, but also reduces the loss of local feature resolution brought by Transformer, making it have better image segmentation accuracy.
   - The low-dimensional image texture features mainly include structural features and statistical features.
   - An edge preservation module to enhance low-dimensional edge features, effectively improving the performance of semantic segmentation. Although low-dimensional statistical features play an importance role in grasping global image features, only a small percent of existing solutions try to analyze them.
   - The proposed network can not only utilize the Transformer's ability to construct global contextual information, but can also use the CNN's ability to capture local information. A multi-scale statistical feature extraction module to extract statistical image features to improve segmentation performance.
   - Statistical features as low-dimensional texture features play a key role in improving semantic segmentation performance. Extract the texture information of statistical features by applying Fisher vector layers to enhance features using handcrafting. A texture enhancement module and a pyramid texture extraction module to extract image texture features for the enhancement of semantic segmentation performance.
   - The proposed hybrid network is divided into five stages. Stem is the first stage. CNNs and Transformer alternate in the remaining four stages.
   - At the beginning of each stage, downsampling is applied to decrease feature map size and increase the number of channels. Additionally, the proposed network refers to the residual connection of ResNet.
   - Specifically, stem as the first stage contains two layers of simple 3×3 convolution. CNNs stage is the second stage, because the feature map is too large at this moment and not suitable for using Transformer in global feature extraction. The CNNs stage uses a Depthwise Separable Convolution block (DSConv) to reduce the amount and size of model parameters. There is a 1×1 convolution layer before and after DSConv to change the feature map size and the number of channels. The third stage is the Transformer stage, which extracts global features after CNNs. The proposed network adopts a lightweight multi-head self-attention. In order to reduce the overhead, the proposed network uses a k×k depthwise convolution with a stride of k to reduce the dimensions.

   The U-Net model offers the following advantages:

   - U-Net model can perform efficient segmentation of images using limited number of labelled training images.
   - U-Net architecture combines the location information obtained from the downsampling path and the contextual information obtained from upsampling path to predict a fair segmentation map.

   U-Net models also have few limitations, stated as follows:

   - Input image size is limited to 572 × 572.
   - In the middle layers of deeper UNET models, the learning generally slows down which causes the network to ignore the layers with abstract features.
   - The skip connections of the model impose a restrictive fusion scheme which causes accumulation of the same scale feature maps of the encoder and decoder networks.
   - References: https://doi.org/10.3389/fnins.2022.1009581

2. **V-Net: FCNN for Volumetric Medical Image Segmentation**

   - Refrains from processing the input volumes slice-wise and propose to use volumetric convolutions instead.
   - The left part of the network consists of a compression path, while the right part decompresses the signal until its original size is reached.
   - The left side of the network is divided in different stages that operate at different resolutions. Each stage comprises one to three convolutional layers.
   - The input of each stage is (a) used in the convolutional layers and processed through the non-linearities and (b) added to the output of the last convolutional layer of that stage in order to enable learning a residual function.
   - As the data proceeds through different stages along the compression path, its resolution is reduced. This is performed through convolution with 2 *2 * 2 voxels wide kernels applied with stride 2. Since the second operation extracts features by considering only non-overlapping 2*2*2 volume patches, the size of the resulting feature maps is halved. This serves similar to pooling layers.
   - Since the number of feature channels doubles at each stage of the compression path of the V-Net, and due to the formulation of the model as a residual network, we resort to these convolution operations to double the number of feature maps as we reduce their resolution.
   - Downsampling allows us to reduce the size of the signal presented as input and to increase the receptive field of the features being computed in subsequent network layers. Each of the stages of the left part of the network, computes a number of features which is two times higher than the one of the previous layers.
   - Replacing pooling operations with convolutional osnes results in a smaller memory footprint during training, because no switches mapping the output of pooling layers back to their inputs are needed for back-propagation.
   - The right portion of the network extracts features and expands the spatial support of the lower resolution feature maps in order to gather and assemble the necessary information to output a two-channel volumetric segmentation.

   The conventional FCN model however has the following limitations:

   - It is not fast for real time inference and it does not consider the global context information efficiently.
   - In FCN, the resolution of the feature maps generated at the output is downsampled due to propagation through alternate convolution and pooling layers.
   - This results in low resolution predictions in FCN with fuzziness in object boundaries
   - References: https://arxiv.org/abs/1606.04797

3. **R-CNNs**

   - It generates region proposal network for bounding boxes using selective search process.
   - These region proposals are then warped to standard squares and are forwarded to a CNN so as to generate feature vector map as output.
   - The output dense layer consists of features extracted from the image and these features are then fed to classification algorithm so as to classify the objects lying within the region proposal network.
   - The algorithm also predicts the offset values for increasing the precision level of the region proposal or bounding box.
   - However, a huge amount of time is needed to train network to classify 2000 region proposals per image
   - It cannot be implemented in real time. Selective search algorithm is a fixed algorithm.