

Validation Plan

Intended Use

To measure the volume of a hippocampus from an MRI scan. It is intended for cropped MRI scans of the hippocampus. The model may work for full brain scans, but the performance will be worse since it was trained on cropped scans. It is unlikely to work, or will not work at all, on other types of scans.

Ground Truth

The ground truth can be determined by radiologists manually computing the volume of the hippocampus using distances between points in the sagittal, coronal, and axial planes. This can then be compared to the model's performance in the real world.

Training Data

The training data is from the Medical Decathlon competition "Hippocampus" dataset. It consists of NIFTI files - one per volume and one per corresponding label mask (a volume containing values as follows: 0 for background, 1 for anterior and 2 for posterior portions of the hippocampus). The original images are T2 MRI scans of the full brain but the images were preprocessed into crops of the hippocampus, thereby reducing dataset size leading to good training speeds.

To train the Recursive U-Net, the data was randomly split into 3: 70% consisting of a training set, 20% of a validation set, and 10% of a test set. After each epoch, the validation set was used to measure performance on a held out set to monitor overfitting and underfitting. At the end of training, the test set was used to compute the average Dice and Jaccard scores as well as sensitivity and specificity.

Performance

The algorithm was run on the test data and the following metrics were computed:

- Dice: 0.9, which is good. The highest is 1.
- Jaccard (also known as intersection over union): 0.818. The intersection of the model's output and ground truth is 81.8%.
- Sensitivity: 0.853. Out of those areas (pixels) identified as positively belonging to the hippocampus (either anterior or posterior), 85.3% are.
- Specificity: 0.998. Out of those areas that are not hippocampus, 99.8% are correctly segmented.

The last two taken together say that there's a bit of under segmentation and almost no over segmentation. We can confidently say that the model performs rather well.