Validation plan for HippoVolume.AI

General information

The product is an AI system based on deep neural networks. The architecture used is based on the U-Net network (Ronneberger et al. 2015). The training was done on a GPU device. However, the inference can be done on a CPU device. The product can be integrated into a clinical-grade viewer in order to speed up workflow.

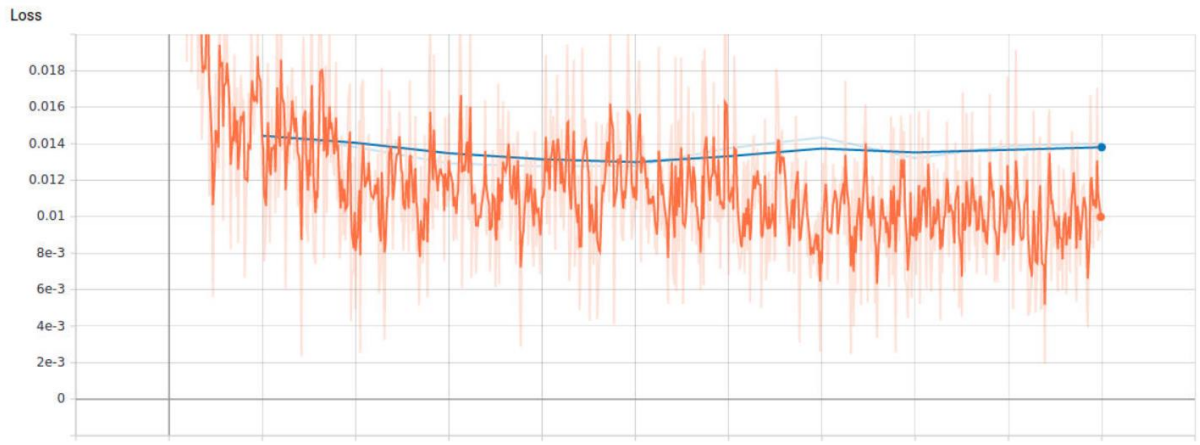
Intended use of the product

The product is used to help clinicians calculate the volume of the hippocampal volume based on cropped MRI scans of the brain. In this fashion, the product can help clinicians detect Alzheimer's disease by monitoring the hippocampal volume and looking for atrophy.

Description of training data

In the training, the "Hippocampus" dataset from the Medical Decathlon competition was used. The data was collected by the Vanderbilt University Medical Center and contains 394 3D volumes. Each volume comes with the corresponding segmentation mask, the annotation being done manually by experts. The original images are T2 MRI sequences of the full brain.

For the purposes of the training, the volumes were cropped by the radiologists so that the network sees only the region around the hippocampus and not the entire image. After cleaning, 260 volumes were viewed by the model, 20% of which were reserved for testing. The other 80% were split 80:20 and used for training and validation respectively. The loss function used was the cross entropy. The following figure shows the evolution of the loss during the training procedure (orange: training loss, blue: validation loss).



Performance on training data

After the training phase, the algorithm was tested on new data it has not seen. The performance during the testing phase was quantified by two metrics: the Jaccard Index, also known as Intersection-Over-Union (IoU), and the Dice coefficient. Both of these metrics are standard ways to evaluate the performance of a segmentation algorithm. The model achieved a Jaccard Index of 0.8160 and a Dice coefficient of 0.8979 on average.

Performance in the real word

To validate the model, we need new data. The data has to be the same modality and have the same statistical properties as the training data. The labels can be obtained by a voting systems based on the opinion of experts.

As the training was done on T2 MRI brain series, the product can only be used on the same type of MRI sequences. For the same reason, the model isn't compatible with other imaging modalities (CT, PET, etc.). The algorithm can only be used to measure the volume of the hippocampus. To measure the volume of others anatomical structures (in the brain or otherwise), the neural network must be retrained.