WMH Segmentation Challenge, MICCAI 2017

Matt Berseth
NLP LOGIX, LLC, matt.berseth@nlplogix.com

INTRODUCTION

The following outlines the training procedure for a system that will automatically identify and segment white matter hyperintensities in T1 and FLAIR MRI sequences. This work was done for the White Matter Hyperintensity Segmentation Challenge coordinated as part of MICCAI 2017.

LESION SEGMENTATION

Overview

The model developed is similar in architecture to [1], with the following key differences:

1. We used receptive fields of size 5x5 in the first convolutional layer, and 3x3 in all other layers.
2. We used fewer filters in all except the first convolutional layer (20, 20, 40, 55 versus 20, 40, 80, 110)
3. We used more neurons in the fully connected layers after fusing (450, 400, 400, 200, 3 versus 900, 200, 2)
4. We did not find value using spatial features

Preprocessing

To prepare the images for the learning algorithm, each training scan was preprocessed using the following sequence of steps:

1. A brain mask was generated for each slice of every training scan using the `bet` tool from the `fsl` library [2]
2. After the brain was removed from the T1 and FLAIR images, the intensity of the remaining pixels was scaled to be between 0.0 and 1.0.

After executing these preprocessing steps, patches were extracted from each of the scans. For each non-brain pixel, three types of patches were extracted: 32x32, 64x64 and 128x128. The larger two patch sizes were scaled down after extraction to 32x32. This process was repeated for both the T1 and FLAIR sequences [1]**.** A patch was labeled as positive if the middle pixel intersected the ground truth mask, otherwise the patch was labeled negative.

For all scans, all positive patches were extracted and a random sample of negative patches were extracted. In total 399,716 positive patches and 550,000 negative patches were extracted and used in the training process. Patches that were labeled as `Other Pathology` by challenge organizers were included as positive patches, but given a different class label when provided to the model.

Negative patches the middle FLAIR pixel had an intensity value over 100 were upsampled so they were twice as likely as other negative examples to be included in the training process.

Model Architecture

A multiscale deep neural network was used to discriminate between positive (WMH), positive (Other Pathology) and negative (Normal Tissue) patches. A stack of 4 convolutional layers was applied to each of the three patch sizes. Each size of T1 and FLAIR patches were stacked to create a channel depth of 2 for the convolution. The output layer from each of the size based stacks were fused and 3 additional fully connected layers were applied to the network before the final softmax output layer. This architecture follows [1] closely, except for the key differences outlined in the overview section.

Rectified linear units were used for all non-linearity's. Dropout of 0.5 was used in all fully connected layers. Pooling was not used. Batch normalization was used between all layers of the network.

Training

The network was trained using the Adam[3] optimization algorithm. The learning rate was initialized to 1e-2 and was annealed throughout the training process. Each cycle of annealing lasted for two epochs, then the learning rate was reset to the initial value. This process repeated for 8 cycles. For each cycle a subset of slices from the validation data set was evaluated and the dice coefficient was computed. The iteration for each cycle that achieved the highest dice coefficient was saved[4]. Stratified 10 fold cross validation was used to train the models, each fold stratified by the site of the scans origination.

Large minibatches of size 768 were used selected for training. Each minibatch of data contained 128 positive and 128 negative examples from each of the three training sites. Of the 128 negative examples, 64 contained patches where the middle FLAIR pixel value exceeded 100.

Postprocessing

To score the validation data set and estimate test time performance, each validation scan had all non-brain pixels scored by the 3 best performing checkpoints for the fold. The probability maps for the 3 different checkpoints were then combined with a simple average. At test time, the probability mass for WHM and Other Pathology were merged. If the combined probability mass of WMH and Other Pathology exceeded 0.975, the voxel was labeled as positive for WMH.

Scoring the test set follows similar logic, except the 3 best checkpoints for all 10 folds are used to score the test scan.

Other Paths Explored

The following is brief overview of the other paths we explored.

Fine tuning model estimates

We evaluated further finetuning the probability maps generated by the network by making a second pass over the voxels and using the estimates, original T1 and FLAIR intensities neighborhoods surrounding the target voxel use this information to finetune the WMH probabilities. We evaluated these using classifiers including Random Forests, Gradient Boosted Trees and Logistic Regression, but did not see enough improvement to warrant the increased complexity.

Incorporating spatial features

Like [1], we evaluated incorporating features encoding the spatial information for the target voxel. The spatial information we evaluated included the relative x,y,z position of the voxel in the scan and the prior of a voxel in this position being positive for WMH. Neither of these types of features appeared to make a difference in the performance of the model and were excluded.

Validation Metrics

Using the evaluation script provided by the challenge organizers, we compute the following metrics for all scans in each of the validation folds and aggregated the metrics so performance could be measured overall and by site.

The tables below summarize the key metrics organized by site. Figure 2 depicts the shape of these additional validation metrics.

Site	Dice Coefficient
GE3T	.798
Singapore	.828
Utrecht	.785

Site	AVD
GE3T	13.61
Singapore	13.13
Utrecht	18.8

Site	Lesion Detection
GE3T	.837
Singapore	.715
Utrecht	.774

Figure 1 shows the original T1 and FLAIR scan, the probability map generated by the mode, the mask computed at the 0.975 cutoff and the ground truth.

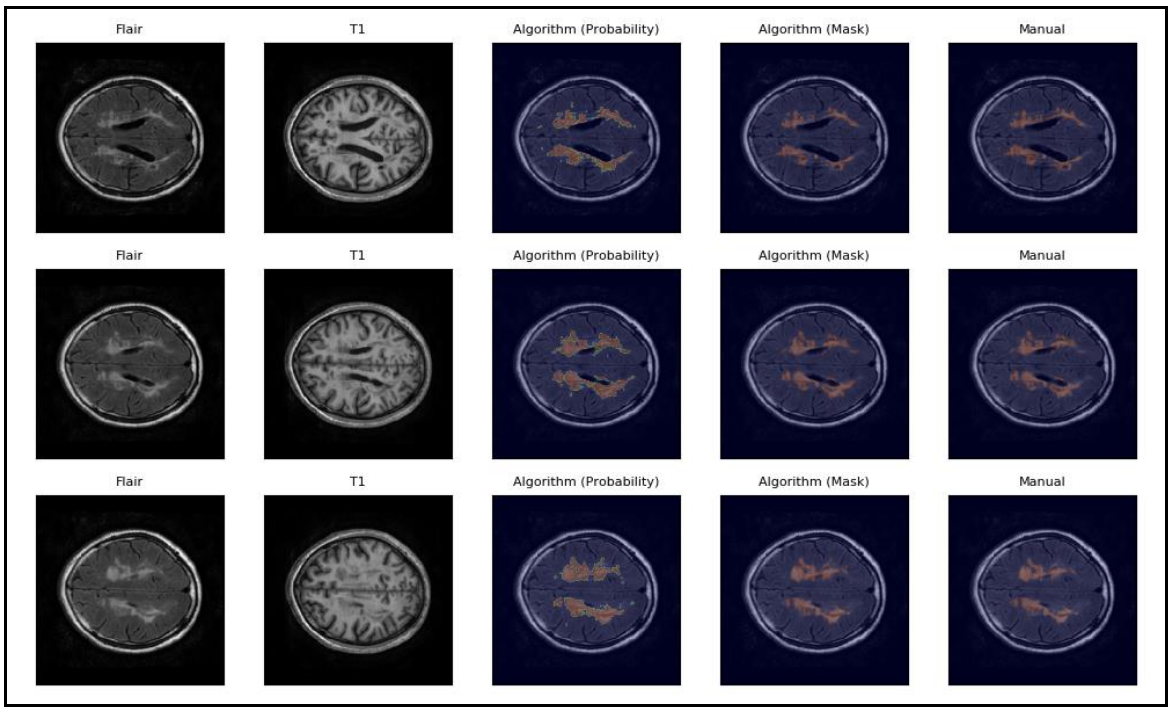

Figure 1 T1, FLAIR Sequences, Model Generated Probability Map and the Corresponding Masks

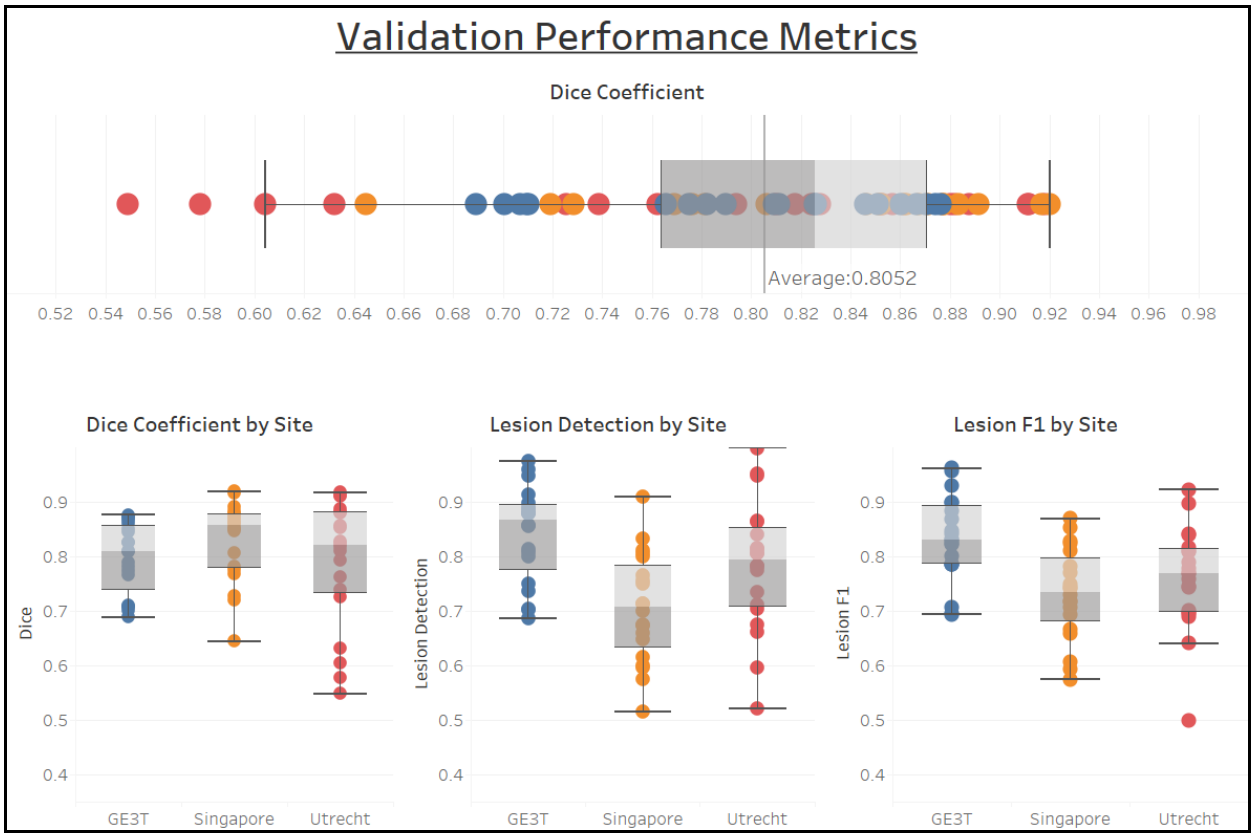

Figure 2 Dice, Lesion Detection and Lesion F1 Metrics

REFERENCES

[1] Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I., Sanchez, C., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., Platel, B., 2016a. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. arXiv:1610.04834.

[2] M.W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, S.M. Smith. Bayesian analysis of neuroimaging data in FSL. NeuroImage, 45:S173-86, 2009

[3] Kingma, Diederik P. and Ba, Jimmy. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG], December 2014.

[4] Huang, Gao, et al. "Snapshot ensembles: Train 1, get m for free." arXiv preprint arXiv:1704.00109 (2017).