Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation
DeepGs: Alec Tarashansky, Ehsan Dadgar-Kiani, Yuan Xue, Matthew Kim

Background:

Brain tumor segmentation involves separating tumor tissues from normal brain tissues such as gray matter, white matter, and cerebrospinal fluid. Most of the time, abnormal tissues can be detected easily but accurate and reproducible segmentation of lesions remains challenging. Segmentation through magnetic resonance imaging (MRI) data is valuable for the analysis of neuropathologies and are important for planning treatment strategies and monitoring the progression of cancer. Therefore, the authors propose that quantitative imaging could elucidate the characteristics of a disease and its effects on different physiological structures and functionality.

In order to analyze lesions quantitatively, the algorithm must be able to segment lesions in three-dimensional images taken with different imaging modalities. A large variability in location, size, shape, and frequency of lesions make it crucial to track the exact positions of contusions, edema, and hemorrhages in traumatic brain injury (TBI). Although the task is expensive, time-consuming, and not scalable, the gold standard data processing method for accurate quantitative analysis is through manual segmentation by human experts.

The authors present a fully automatic approach to lesion segmentation of brain MRI images that consists of an 11-layers deep, multi-scale, 3D convolutional neural network (CNN). By utilizing dense training on large image segments that adapts to the class-imbalance of the segmentation problem, the authors built a discriminative and efficient, 3D CNN model with parallel convolutional pathways. The results demonstrate that in highly challenging segmentation tasks, their method outperformed the state-of-the-art. The authors also utilized fully-connected 3D Conditional Random Field (CRF) models for post-processing of the 3D CNN's soft segmentation maps.

Data and Model Summary:

Traumatic Brain Injury (TBI): The experiments were conducted using the TBI dataset with 61 multi-channel MRIs. 66 patients with TBI underwent imaging within the first week of injury. Visible lesions were manually annotated on two of the structural MRI sequences with different labels for each type of lesion. The images were randomly split into 46 training and 15 validation images. During training, the authors extracted 10,000 random patches uniformly sampled from the brain region to approximate the true distribution of lesions and healthy tissue in order to monitor the progress of segmentation accuracy. After every five epochs, the authors segment the validation datasets and calculate the mean Dice similarity coefficient (DSC).

The authors then compared their dense training method with two other training schemes on the 5-layers baseline *Shallow* CNN - one which trains on 17^3 patches extracted uniformly from the brain region and the other sampled from patches from both the lesion and background classes. Then, to evaluate the dense training scheme, the authors trained on multiple models of image segments from sizes 19^3 to 29^3.

Based on the shallow model, the authors created a *Deep* CNN model that extended 5-layers to 9, and replaced each convolutional layer of 5^3 kernels with two layers of 3^3 kernels. The authors also present *Shallow+* and *Deep+* models by changing the initialization scheme for weights from the normal distribution (0, 0.01) to one that is derived from modeling the nonlinearity of rectifiers (ReLU/PReLU). This effectively preserved signal in the initial stage of training.

DeepMedic was the final model that the authors proposed. This extends the *Deep+* model with a second convolutional pathway and adds two hidden layers for combining multi-scale features before classification for a total of 11 layers. The network is able to capture context in a large area of the original image through the 17^3 receptive fields.

In the experimental setup, the used the dual-pathway, 11-layers *DeepMedic* CNN network architecture. The network is evaluated with 5-fold cross-validation on the 61 MRIs. The CRF parameters were determined via a configuration experiment using randomly selected subjects. A Random Forest baseline consisting of 50 trees with maximum depth of 30 was used for comparison and trained on data points from both the lesion and background classes.

Brain Tumor Segmentation: The training set consisted of 220 cases of high grade (HG) and 54 cases of low grade (LG) glioma data with reference segmentations from the 2015 Brain Tumor Segmentation Challenge (BRATS). Each subject had MRI images with four classes of segmentations: necrotic core, edema, non-enhancing, and enhancing core. The brain tissue intensities of each image was normalized.

The *DeepMedic* architecture was modified for multi-class problems to have a five-feature map classification layer. Then the network was trained and 5-fold cross validated on the training set image segments extracted from the tumor and healthy tissue.

Ischemic Stroke Lesion Segmentation: The training set consisted of 28 datasets from the 2015 Ischemic Stroke Lesion Segmentation challenge. The images were skull-stripped and their voxel resolutions were resampled. The test set were 36 datasets with hidden annotated segmentation. Besides configuring the network in a slightly different way than that of TBI, the authors evaluated the network in a similar fashion to the other experiments, namely a 5-fold cross validation.

Results/Summary:

Traumatic Brain Injury: According to the two-sided, paired t-test on the DSC metric, the authors' DeepMedic CNN significantly outperformed the Random Forest baseline. In order to clear unbiased error due to randomness from the networks during training, the authors formed an ensemble of these networks and averaged the output. CRF yielded improvements for the single network and ensemble. The network is able to segment cases with very small lesions and the authors showed that network is neither biased towards lesion nor background classes, all of which is promising.

Brain Tumor Segmentation: As in the TBI experiment, results were published with applications of DeepMedic, CRF, and an ensemble of three similar networks. Though CRF and ensemble offers a small improvement, DeepMedic by itself performed better than the state-of-the-art by preserving the hierarchical structure of the tumor. However, the improvements were not unilateral as the authors' methods yielded decreased performance on the BRATS test data most likely from variability in image acquisition from different clinical centers.

Ischemic Stroke Lesion Segmentation: The authors concluded that their system with the structural regularization from the CRF yielded significant improvement to ischemic stroke lesion segmentation as compared to the state-of-the-art. However, some of the improvement could be explained by overfitting.

Weaknesses:

A general problem in medical imaging analysis involves heterogeneity of data. Differences in scanner type and acquisition methods result in test images from clinical centers that are different from the ones of training datasets. The authors did not attempt a generative model for the data acquisition process in order for the CNN to be invariant to the data heterogeneity. The issue of multi-center data is a major bottleneck for large-scale imaging studies and we see a performance drop from overfitting to the training data when the authors' system is applied on BRATS and ISLES test datasets. However, this was true for other teams that worked on multi-center data. In order to account for the innumerable differences that can occur between data acquired from different data collection sessions, domain adaptation techniques have been developed to overcome this difficulty. Perhaps these models might benefit from such methods.

The authors needed to make a lot of decisions in order to improve the performance of their proposed CNN. Citing Choromanska et al. (2015), one choice was to form an ensemble of three similar networks and average outputs on the experiment in order to reduce random unbiased errors. As in other cases, the authors

were very thorough in motivating a particular design choice but were not exactly clear on how they settled on concrete values to use for their parameters. Likewise, the authors were unfortunately vague when describing the enhancements of the *Shallow+*/*Deep+* CNN models compared to the *Shallow*/*Deep* models. Although they described the motivations for using a different initialization [1], specifying some of the parameters for the rectifiers described in the paper they referenced would be helpful for reproducibility. Alas, this is a problem that plagues much of the deep learning literature. The parameter space is so large, especially for these kinds of complex networks, that it would be helpful for readers to know by what design methodology and search process the authors have been able to arrive at their chosen models.

Furthermore, the authors bounded output to the uniform range and thus used the output as probabilities but without a rigorous Bayesian proof. The value is closer to one that is structured to maximize cross entropy on the training samples. It will be interesting to see how the lack of this assumption plays a role in future research with the architecture.

The authors note that each training session for the network takes approximately two to three days. It would be useful to motivate their research if the authors had provided a comparison between the total amount of time saved using their method compared to manual segmentation with respect to their method's accuracy. For example, if in the time it took to train the neural network someone could have already manually identified all brain lesions and the network does not perform as accurately as manual segmentation, will this method really be a useful tool for clinicians?

The most conspicuous improvements can be made on the CRF architecture. The authors specifically mention that finding optimal parameters for each task for tuning the CRF was challenging for multi-class tumor segmentation. Instead of finding a global set of parameters for refinement, the authors opted to apply CRF in a binary fashion to refine only boundaries of the tumor. The next step would be to shed this model and attempt to cast CRF automatically as a neural network with parameters learned from gradient descent.

The authors were primarily concerned about the results of complex segmentation tasks and did not take the time to process information from the neurons of the low-resolution pathway itself. The deepest hidden layers preserve patterns from the activation of certain feature maps, and these patterns may provide valuable spatial bias such as a tendency of TBI to occur towards the front and sides of the brain. Furthermore, the paper could have gone into more analysis on the differing roles of the high- and low- resolution pathways especially since this dual pathway architecture may influence how the network distinguishes detection of smaller lesions from rough localization.

Extensions:

The most obvious application of this technology would be to use it to automatically identify lesions and tumors in different organs. Furthermore, the network could potentially be adapted to analyze primary images from other imaging modalities like ultrasound, computed tomography, nuclear imaging, etc. Additionally, it would be interesting to see if the same methodology can be applied to segment physiological features with completely different morphology. For example, striated muscles have markedly different geometries than brain lesions. Many biomechanical studies require manual segmentation of striated muscles in mammalian limbs in order to automatically generate accurate kinematic models of limb motion that take into account each individual muscle. In the human arm, for example, there are 33 individual muscles all working in concert to enable coordinated movement. Manual segmentation of striated muscles from MRI images is the primary bottleneck in generating models that take all of these muscles into account. By making segmentation easier, the deep learning methodology discussed in the paper may facilitate the development of more accurate biophysical models in a wide variety of fields.

References:

[1] He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026– 1034.

[2] Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., Glocker, B., 2015. Multiscane 3d convolutional neural networks for lesion segmentation in brain mri. in proc of ISLES-MICCAI.