Review: Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation

Shaimaa Bakr, Daniel Fernandez, Cici Chen, Rahul Palamuttam

December 5, 2016

Summary

The paper is about a model that performs brain lesion segmentations in registered 3D MRI volumes of the brain. The authors develop an architecture, train and test their model separately on 3 disease types:(Traumatic Brain Injury (TBI), Brain Tumors, Ischemic Stroke. The paper proposes several techniques to address the main challenges in applying deep learning to medical image segmentation problems.

Introduction/Overview

Segmentation is an important problem in medical imaging. Most of the information currently extracted from medical images are restricted to what can visually be observed by the radiologist. Few basic quantitative metrics are applied in practice to obtain diagnostic and prognostic value from medical images. Accurate segmentation can provide a more reliable procedure for tracking disease progression, characterizing lesions, and extracting predictive features from medical images. Currently, medical image segmentation is either performed manually by radiologists requiring significant time and effort or using semi-automated methods that still involve expert radiologists. Deep learning promises to solve this problem. However, applying deep learning to medical images many new challenges arise. These include the volumetric nature of medical scans, which would directly translate to exponentially more complex model when conventional 2D architectures are extended to 3D input. This leads to over-fitting, prohibitive computational and memory costs. Another issue with medical imaging data is the common large class imbalance of data, e.g. in a segmentation task, most of the image volume is not a tumor. Furthermore, in a classification task, most of the scans represent healthy data points and very few are of disease-positive individuals. This paper proposes an architecture that attempts to overcome the complexity of 3D input and also tries to deal with class imbalance:

The main contributions of the paper, as applied to medical images are:
1) Application of dense training (saves computation and adapts to class imbalance)
2) Application of a deeper architecture of smaller kernels compared to previous work (extracts higher order feature while reducing computation time)
3) Use of dual path processing: 1) full-resolution image path and 2) subsampled version of the image volume to extract global information as well as high-resolution features in the image.

In addition to these contributions the authors provide a qualitative interpretation of the model and an extension of CRF's to 3D volumes that are used to obtain hard segmentations from soft probability maps that are the output of the CNN.

Method

Deepmedic uses a dual-path convent with one path processing high-resolution image segments and the second path processing a down-sampled version of the whole image volume to preserve global features. The high-resolution path and low-resolution path join and two hidden layers are added resulting in a CNN with 11 layers. The low-resolution input is subsampled 3 times giving a receptive field increase with the same size model. The output of this network produces a soft segmentation in the form of probability maps for each class. This is followed by a 3D CRF which produces hard segmentation labels that essentially clean up stray false positives and refines segmentation boundaries. The full-resolution input is the same size as the subsampled input, which is an image segment rather than a full volume or a patch volume. This reduces the complexity of the model and computational cost in full-volume input and avoids redundant calculation in patch input model. Here the authors describe the baseline model which consists of 4 layers with $5\hat{3}$ kernels leading to a receptive field of $17\hat{3}$ and a fifth fully convolutional classification layer with $1\hat{3}$ kernels.

3D CNN for Dense segmentation

Dense inference occurs when an input of a size larger than the receptive field of the CNN is applied. (Sermanet et al. (2013)). This saves computational cost, since voxels which would repeat in overlapping patches would produce repeated computation in a batch method. Dense input reduces the number of these computation. This is a compromise between using patches and applying the whole 3D volume as input, which would be constrained by GPU memory (Long et al. (2015)).

Dense training and class balance

Training is done by selecting image segments larger than the receptive field. Segments are selected with 50%/50% probability that the center voxel is foreground/background. This is the authors proposal to deal with class balance, since they force the 50-50 selection but at the same time other voxels in the segment, which are also trained on, naturally have the actual class distribution of the data set.

Deeper network

The authors combine a deeper architecture with a smaller kernel and less weights to train following (Simonyan and Zisserman et. al (2014). Deeper networks have higher-level and more abstract features. They used ReLu activation, normal distribution weight initialization with zero mean, and took the variance of the square root of 2 divided by the number of weights through which a neuron is connected to its input.

Multi-scale

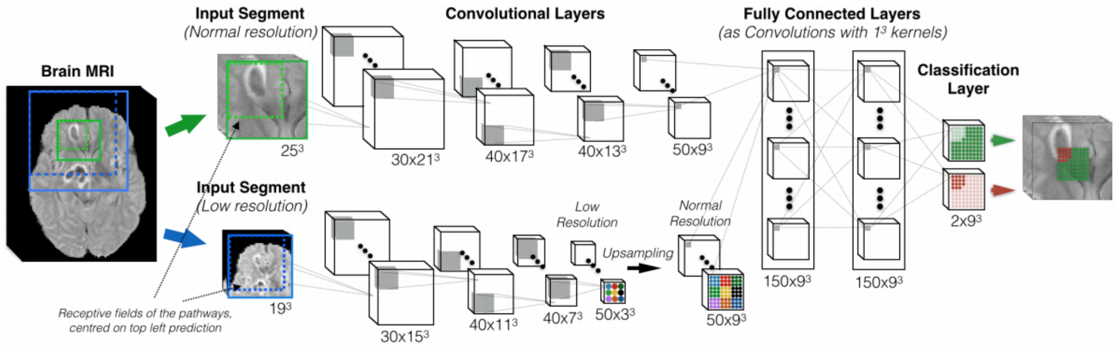
Multi-scale pathways (Long et al. (2015); Ronneberger et al. (2015) allow extraction of local features from higher-resolution image segments combined with global features that can be learned from the whole volume. Subsampling allows combining the two inputs for final classification and maintains a practical computational cost.

3D CRF

CRF's are used to clean-up holes in segmentation maps and spurious isolated regions in the network output, due to noise and local-minima. CRF's utilize neighborhood dependencies to give hard-segmentation output. The CRF used is a 3D extension from the work by (Krahenbruhl and Koltun (2012)).

Analysis of Network Architecture

In this section the authors look at each of the features of their attribute and analyze its contribution on the increase of the performance.



Experimental setting

The experiments for these analysis use the private TBI data set which consists only of 61 patients. They use 46 images for training and the rest for validation. They perform full validation of the data set every 5 epochs using accuracy, specificity, sensitivity and DSC score.

Effect of dense training on image segments

To evaluate the effect of dense training they train this approach on the baseline CNN using various image segment cube sizes of $19 - 29$. The segments are chosen uniformly centered on background or foreground. This is compared to use of cube patches of size 17 extracted 1) uniformly, 2) equally from brain background and lesion regions. These two approaches perform poorly on either specificity or sensitivity, since uniform extraction will lead to under-segmentation and conversely artificial balancing will lead to over-segmentation. The dense training performs better than both but the authors note they adjust memory use and training times to match that of the uniform and equal patch approaches. This suggests that the improvement is partly due to the increase in effective batch size and the authors note that.

Effect of deeper networks

The authors analyze the effect of a deeper architecture and ways to successfully train a deeper mode while avoiding the vanishing gradient problem. They use 33 kernels instead of 53 and obtain an architecture of 9 convolutional layers. They use batch normalization and initialization by (He et al) to deal with this problem and gives superior results to the shallow baseline model. Without batch normalization for training, the deeper model is unsuccessful compared to the baseline 5-layer mode.

Effect of multi-scale

To evaluate the effect of multi-scale the authors perform an experiment where the model contains twice as many feature map as the single path model but with no global sub-sampled input. They show that the performance improvement occurs only in the multi-scale ruling out the increased capacity of the network as the reason for higher performance. An identical second convolutional path is added in parallel. Training this architecture results in overfitting, compared to the improved performance with the dual path using subsampled input.

Evaluation

The authors used 3 data sets corresponding to 3 diseases to evaluate their method:

1 TBI

Data set

61 TBI cases with isotropic MPRAGE, FLARI, PD, T2 and GE sequences, registered and resampled. The abnormalities are merged into one class. The experiment applies sagittal reflected augmentations plus original data to the DeepMedic architecture. Dropout of 2% is applied to convolutional layers and 50% to FC layers. 5-fold cross-validation is used and training is done in 2 days. CRF configuration is done using random search on 15 subjects and input from from a preliminary CNN model. This is evaluated using 5-fold CV.

Results

The baseline model used is random forest with 50 trees and 30 max depth two hundred randomized cross-channel features are used as input to the RF. In addition, authors compare to results from an ensemble to enhance performance of the model. They show that use of an ensemble and CRF improves results. It is worth noting the low DSC score obtained on the data.

2 BRATS

BRATS is a well-known Brain Tumor Segmentation Challenge (BRATS) data sets containing 220 High grade and 54 low grade gliomas for training and 110 cases for testing. Segmentation is done for 4 classes (necrotic core, edema, enhancing and non-enhancing core) and background. FLAIR, T1, T1-contrast and T2 sequences are included registered to a common pace and resampled. For the experiment, sagittal reflections are used for augmentation. There is no use of intensity perturbation or dropout on this data set. Segments are extracted with equal probability centered around tumor and healthy voxels.5-fold validation is performed. Training is completed in 2 days. For CRF configuration, probability maps are the four classes are merged into one tumor map and configuration is done similar to TBI. On BRATS data, DeepMedic performs best on the training data of the challenge, outperforming (Pereira et al) and (Bakas et al) but the performance comparison with other methods on test data is not available. Compared to the ensemble, DeepMedic does worse on the test data, which implies problems in generalization.

3 ISLES

Data set

The authors participated in Ischemic Stroke Lesion Segmentation (ISLES) http://www.isles-challenge.org/ which provided 28 data sets for training and 36 data sets for testing. Each volume included T1, T1-contrast, FLAIR, DWI sequences. Pre-processing consisted of resampling and normalization of mean and variance.

Experiment

The specifics of the model included less feature maps on the global (lower resolution) path, as the accuracy of the model would not be affected for small lesions such as those in sub-acute ischemic stroke lesion segmentation.

They use data augmentation as sagittal reflection.
They use 5-fold cross validation to evaluate their data set.

Results

The authors also point to a qualitative analysis by looking at feature values for deep and early layers of their network and observing that the network is learning to discriminate between different brain tissue types CSF, gray matter etc. They also note that feature maps of the low-resolution path contain spatial information on certain lesion types and suggest that including tissue priors could further improve the performance of their model. We note that an extension for this remark is warranted; most clinical models include other patient information besides medical images e.g. patient history, family history demographics. This means that aside from spatial cues other image features could correlate with these types of data and contribute to the performance of the model. In addition, some types of patient information are independent of image data and can be predictive of pathological behavior. Including this data can enhance the performance of the model.

Criticism

- Although the BRATS data set is reasonable in size, the TBI and ISLES data set is relativel small. This raises the questions of how well the model generalizes and how it will validate on other data sets.
- Analysis of the architecture is applied only to one of the small data sets, the TBI. TBI is also an internal data set, so it is not easy to reproduce the analysis of the paper.
- An important challenge in deep learning in general and in application to medical images in particular is training time. Specifically, for this paper, many of the contributions rely on achieving results while keeping training times practical. However, training-time savings are not reported in the paper. We would have liked to see a much more extensive performance evaluation for the net.
- While the paper introduces the novelty of good results on various brain MRI data sets, most of the techniques are borrowed from applications involving 2D architectures or adaptations thereof.
- When evaluating dense training compared to patches, the dense training performs better than both patch variations but the authors note they adjust memory use and training times to match that of the uniform and equal patch approaches. This suggests that the improvement is partly due to the increase in effective batch size and the authors note that. It is not clear whether this has bigger contribution or the altered distribution of training samples. In this sense, the comparison is not objective. This experiment optimized the hyper-parameter S, dimension of image segment to a cube of 25.

References

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Sur- passing human-level performance on imagenet classification. In: Proceed- ings of the IEEE International Conference on Computer Vision. pp. 1026 1034.

Bakas, S., Zeng, K., Sotiras, A., Rathore, S., Akbari, H., Gaonkar, B., Rozy- cki, M., Pati, S., Davatzikos, C., 2015. Segmentation of gliomas in multi- modal magnetic resonance imaging volumes based on a hybrid generative- discriminative framework. in proc of BRATS-MICCAI

Pereira, S., Adriano, P., Alves, V., Silva, C., 2015. Deep convolutional neural networks for the segmentation of gliomas in multi-sequence mri. in proc of BRATS-MICCAI.

Feng, C., Zhao, D., Huang, M., 2015. Segmentation of stroke lesions in multi- spectral mr images using bias correction embedded fcm and three phase level set. in proc of ISLES-MICCAI.

Halme, H., Korvenoja, A., Salli, E., 2015. Segmentation of stroke lesions using spatial normalization, random forest classification and contextual clustering. in proc of ISLES-MICCAI.

Krahenbruhl, P., Koltun, V., 2012. Efficient inference in fully connected crfs with gaussian edge potentials. arXiv preprint arXiv:1210.5644.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431-3440.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Springer, pp. 234-241.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.