Pooling-Free Fully Convolutional Networks with Dense Skip Connections for Semantic Segmentation, with Application to Brain Tumor Segmentation

Richard McKinley[1(✉)], Alain Jungo[2], Roland Wiest[1], and Mauricio Reyes[2]

[1] Support Centre for Advanced Neuroimaging, Inselspital, University Hospital, University of Bern, Bern, Switzerland
RichardIain.McKinley@insel.ch

[2] Institute for Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland

Abstract. Segmentation of medical images requires multi-scale information, combining local boundary detection with global context. State-of-the-art convolutional neural network (CNN) architectures for semantic segmentation are often composed of a downsampling path which computes features at multiple scales, followed by an upsampling path, required to recover those features at the same scale as the input image. Skip connections allow features discovered in the downward path to be integrated in the upward path. The downsampling mechanism is typically a pooling operation. However, pooling was introduced in CNNs to enable translation invariance, which is not desirable in segmentation tasks. For this reason, we propose an architecture, based on the recently proposed Densenet, for semantic segmentation, in which pooling has been replaced with dilated convolutions. We also present a variant approach, used in the 2017 BRATS challenge, in which a cascade of densely connected nets is used to first exclude non-brain tissue, and then segment tumor structures. We present results on the validation dataset of the Multimodal Brain Tumor Segmentation Challenge 2017.

1 Introduction

We present a network architecture for semantic segmentation, heavily inspired by the recent Densenet architecture for image classification [1], in which pooling layers are replaced by heavy use of dilated convolutions [2]. Densenet employs dense blocks, in which the output of each layer is concatenated with its input before passing to the next layer. A typical Densenet architecture consists of a number of dense blocks separated by transition layers: the transition layers contain a pooling operation, which allows some degree of translation invariance and downsamples the feature maps. A Densenet architecture adapted for semantic segmentation was presented in [3], which adopted the now standard approach of U-net [4]: a downsampling path, followed by an upsampling path, with skip

A. Crimi et al. (Eds.): BrainLes 2017, LNCS 10670, pp. 169–177, 2018.
https://doi.org/10.1007/978-3-319-75238-9_15

connections passing feature maps of the sample spatial dimension from the downsampling path to the upsampling path.

In this paper, we describe an alternative architecture adapting Densenet for semantic segmentation: in this architecture, which we call DeepSCAN, there are no transition layers and no pooling operations. Instead, dilated convolutions are used to increase the receptive field of the classifier. The absence of transition layers means that the whole network can be seen as a single dense block, enabling gradients to pass easily to the deepest layers.

We describe the general architecture of DeepSCAN, plus the particular features of the network as applied to brain tumor segmentation, and report preliminary results on the validation portion of the BRATS 2017 dataset. We then discuss a major source of errors on the BRATS2017 dataset - namely, the imperfect stripping of non-brain tissue. We then introduce a cascade approach in which an initial network strips away remaining non-brain tissue, and a subsequent network to identify the tumor tissues.

2 The DeepSCAN Architecture

2.1 Densely Connected Layers and Densenet

Densenet [1] is a recently introduced architecture for image classification. The fundamental unit of a densenet architecture is the densely connected block, or dense block. In such a block, the output of each layer (where a layer here means some combination of convolutional filters, nonlinearities and perhaps batch normalization) is concatenated to its input before passing to the next layer. The goal behind Densenet is to build an architecture which supports the training of very deep networks: the skip connections implicit in the concatenation of filter maps between layers allows the flow of gradients directly to those layers, providing an implicit deep supervision of those layers.

In the original Densenet architecture, which has state-of-the-art performance on the CIFAR image recognition task, dense blocks are combined with transition blocks: non-densely connected convolutional layers, followed by a maxpooling layer. This helps to control parameter explosion (by limiting the size of the input to each dense block) and limit redundancy between features, but also means that the deep supervision is not direct, at the lowest layers of the network. This Dense-plus-transition architecture was also adopted by Jegou et al. [3], whose whimsically named Tiramisu network is a U-net-style variation of the Densenet architecture designed for semantic segmentation.

In contrast to the standard Densenet architecture, in our approach we dispense with the transition layers: this means, in effect that the whole network (except for the final one by one convolutions) is a single dense block. The layers in our dense blocks have the shape shown in Fig. 1. Depending on its position in the network, the convolution might have kernel size 3 by 3 or 5 by 5, and might or might not be dilated. At deeper levels of the network (where the feature depth is rather high) a "bottleneck" is used, meaning that before the 2D convolution a convolution with 1 by 1 kernels is performed to reduce the number of

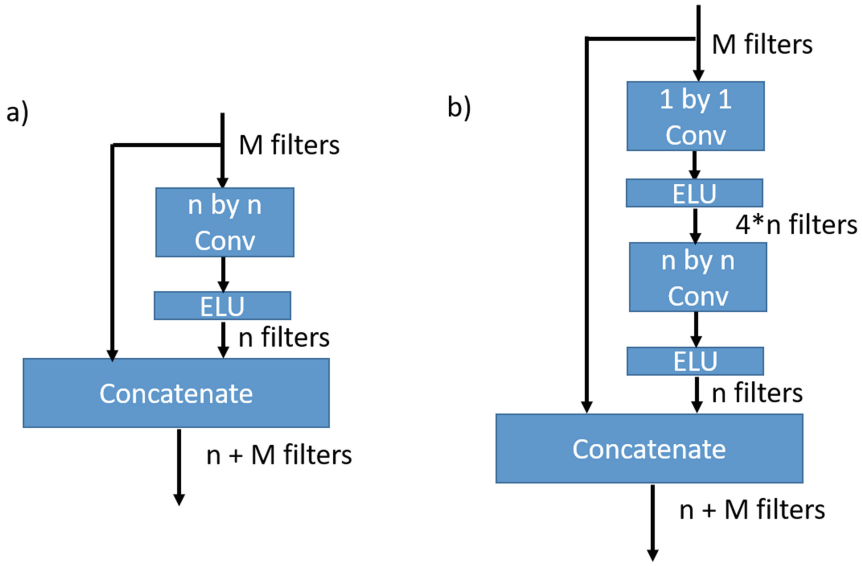

Fig. 1. Dense units, as used in the DeepSCAN architecture (a) a dense unit without bottleneck, and (b) a dense unit with bottleneck

parameters. As a nonlinearity, we use Exponential Linear Units (ELU) [5] rather than the combination of rectified linear unit and Batch Normalization [6] used in the original Densenet paper. There are two reasons for this: the first is that densenets are very memory intensive: removing batch norm layers reduces the overall memory footprint of network. Secondly eliminating batch normalization makes training less sensitive to high levels of variance in batches.

2.2 Dilated Convolutions

The role of pooling layers in CNNs is twofold: to efficiently increase the receptive field and to allow some translation invariance. Translation invariance is of course undesirable in semantic segmentation problems, where what is needed is instead translation equivariance: a translated input corresponding to a translated output. To that end, we use layers with dilated convolutions to aggregate features at multiple scales. Dilated convolutions, sometimes called atrous convolutions, can be best visualised as convolutional layers "with holes": a 3 by 3 convolutional layer with dilation 2 is a 5 by 5 convolution, in which only the centre and corner values of the filter are nonzero, as illustrated in Fig. 3. Dilated convolutions are a simple way to increase the receptive field of a classifier without losing spatial information.

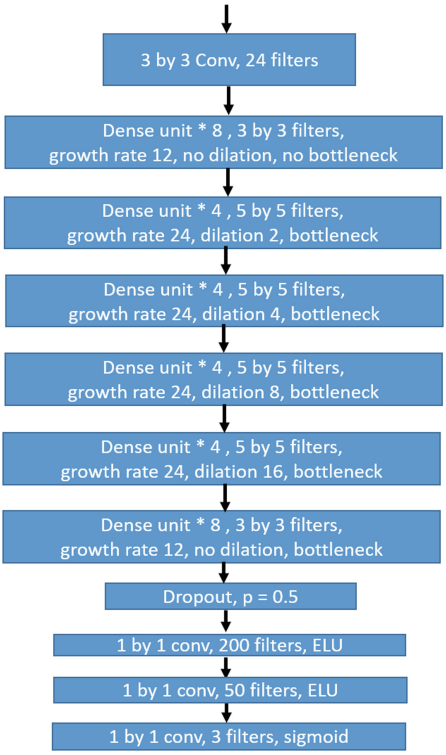

Fig. 2. The DeepSCAN architecture, as applied to brain tumor segmentation

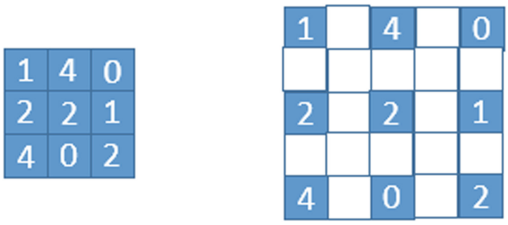

Fig. 3. Left, a 3 by 3 kernel. Right, a 3 by 3 kernel with dilation 2, visualised as a 5 by 5 kernel

2.3 Multi-task Learning and Data Imbalance

Data imbalance in classification in general, and in medical image analysis in particular, is an important theme: in a typical medical image, the background class (healthy appearing tissue) will typically outnumber pathological tissue by a factor of between 10 and 1000 to one. This can lead to the parameter updates arising from the target classes to be strongly diluted by updates from the background class, slowing learning, and can also lead to underidentification of the

target class, if this leads to an overall increase in the accuracy of the classifier. Standard approaches to this imbalance problem include undersampling the background class, oversampling the target class(es), or weighting training examples according to their prevalence in the training set.

For training the DeepSCAN classifiers, we adopted a newer technique from the Bayesian theory of uncertainty in learning [7]. In this framework, one network is used to perform a number of different tasks, each with a loss $\mathcal{L}_i(W)$:, where W are the weights of the network. The *homoscedastic uncertainty* (the noise inherent in the model's output) σ_i of the network is a learned parameter for each task i, and the loss associated with each task is

$$\frac{1}{2\sigma_i^2}\mathcal{L}_i(W) + log(\sigma_i^2) \tag{1}$$

The factor $\frac{1}{2\sigma_i^2}$ provides an adaptive, rather than a fixed, weighting of the loss associated to task i, regularized by the term $log(\sigma_i^2)$. By recasting the segmentation problem in the language of multi-task learning, the appropriate weighting of the different components of the problem is therefore allowed to arise from the data, rather than being imposed.

3 Initial Application to Brain Tumor Segmentation

Brain Tumor segmentation has become a benchmark problem in medical image segmentation, due to the existence since 2012 of a long-running competition, BRATS, together with a large curated dataset [8–10] of annotated images. Both fully-automated and semi-automatic approaches to brain-tumor segmentation are accepted to the challenge, with supervised learning approaches dominating the fully-automated part of the challenge. A good survey of approaches which dominated BRATS up to 2013 can be found here [11]. More recently, CNN-based approaches have dominated the fully-automated approaches to the problem [12–14]

The network used is pictured in Fig. 2. The network was built using Keras [15] and Tensorflow [16], and trained using stochastic gradient descent with momentum for 100 epochs. Rather than using a softmax layer to classify the three labels (edema, enhancing, other tumor) we employ a multi-task approach to hierarchically segment the tumor into the three overlapping targets: whole tumor, tumor core and enhancing: thus the output of the network is three sigmoid units, one for each target. The multi-task uncertainty weighting approach as described above, was applied to each of these tasks.

3.1 Data Preparation and Homogenization

The raw values of MRI sequences cannot be compared across scanners and sequences, and therefore a homogenization is necessary across the training examples. In addition, learning in CNNs proceeds best when the inputs are standardized (i.e. mean zero, and unit variance). To this end, the nonzero intensities

in the training, validation and testing sets were standardised, this being done across individual volumes rather than across the training set. This achieves both standardisation and homogenisation.

3.2 Training and Results

The network segments the volume slice-by slice: the input data is five consecutive slices from all four modalities, Ground truth for such a set of slices is the lesion mask of the central slice.

Slices from all three directions (sagittal, axial, coronal) were fed to the classifier for training, and in testing the results of these three directions were ensembled by averaging. When applied to the BRATS 2017 validation dataset, the mean Dice scores for Whole Tumor, Tumor core and enhancing tumor were 0.87, 0.68 and 0.71 respectively. After some mild postprocessing (removing small connected components), the Dice coefficient for whole tumor increased to 0.88, and for tumor core to 0.70.

4 Cascaded Network for Nonbrain-Tissue Removal

The BRATS dataset was assembled from a large number of datasources, and does not comprise raw imaging data: the volumes are resampled to 1 mm isovoxels, and in addition have been automatically skull-stripped (Fig. 4). Unfortunately, the results of this skull-stripping vary: see Fig. 5 for an example with large amounts of residual skull tissue. Other examples have remnants of the dura or optic nerves. This remaining tissue can confound classification in two ways: it can be misidentified by the classification algorithm (though this is increasingly less likely as classifiers improve) and it can affect the distribution of the

Fig. 4. An axial slice through the FLAIR acquisition for case TCIA_208_1, showing hyperintense nonbrain tissue which has not been removed by the skullstripping algorithm

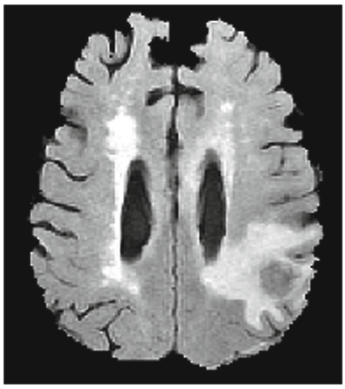

Fig. 5. The same axial slice, after masking with the output of our skullstripping network

intensities in a volume, adversely impacting the global standardisation of voxel values. To combat this effect, we used a cascade of networks to first segment the parenchymia from the poorly skull-stripped images, followed by a second network which identifies the tumor compartments as above. The ground truth for the first network in the cascade was obtained by applying FSL-FAST to the T1 post Gadolinium imaging, as this tended to have the best definition in all three planes. The brain tissue label was assembled by taking the union of tumor, white matter and grey matter labels, and then taking the largest connected component. This network was trained analogously to the tumor classification network above, and the four modalities of the training, validation and testing data were masked with its output. The masked modalities were then used to train a tumor

Label	Dice_ET	Dice_WT	Dice_TC
Mean	0.70985911	0.857922534	0.708739178
StdDev	0.26719754	0.142340453	0.275811784
Median	0.787695	0.90034	0.815625
25quantile	0.653735	0.84384	0.657485
75quantile	0.868275	0.9319575	0.88642

Label	Hausdorff95_ET	Hausdorff95_WT	Hausdorff95_TC
Mean	32.16084212	7.397078973	28.32401603
StdDev	98.1050951	11.26534053	83.88898478
Median	2.23607	4	6.40312
25quantile	1.73205	2.544225	3.4641
75quantile	5.3379075	6.40312	11.8573725

Fig. 6. Results of DeepSCAN on the BRATS 2017 test dataset

segmentation network, as above. For this cascade, the mean Dice scores for Whole Tumor, Tumor core and enhancing tumor were 0.88, 0.76 and 0.71, representing a substantial improvement in classification of the tumor core.

4.1 Challenge Results

The above cascaded network was used to compete the in the BRATS 2017 challenge: results on the testing data of that challenge are shown in Fig. 6.

5 Conclusions and Further Directions

Densely connected networks can provide a convincing tool for semantic segmentation of medical images. The problem of data imbalance in image segmentation can be helped by recasting the problem as a multi-task learning problem and using concepts from Bayesian learning to calibrate the weights on the various component problems. It would be interesting to compare the effects of homoscedastic uncertainty (which is calculated per task) and heteroscedastic uncertainty (which is calculated per datapoint) on quality of segmentation and speed of convergence.

This paper arose in the context of a challenge, where comparisons to other approaches are provided by the challenge leaderboard: while this is useful, it should not take the place of robust comparison against strong benchmarks, and as a result we are preparing further work in which the DeepSCAN architecture is rigorously compared to U-net, Tiramisu, and other existing architectures. We will also investigate the contribution of depth, nonlinearity used, and the contribution of the Bayesian loss function. In addition, we will apply the architecture to additional datasets, to confirm its broad applicability.

References

1. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
2. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: Proceedings of International Conference on Learning Representations (ICLR 2017) (2017)
3. Jégou, S., Drozdzal, M., Vázquez, D., Romero, A., Bengio, Y.: The one hundred layers liramisu: fully convolutional denseNets for semantic segmentation. CoRR (2016)
4. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
5. Clevert, D., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (eLUs). CoRR abs/1511.07289 (2015)

6. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167, pp. 1–11 (2015)
7. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. arXiv:1705.07115 (2017)
8. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Leemput, K.V.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**, 1993–2024 (2015)
9. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Nat. Sci. Data (2017, in press)
10. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. Cancer Imaging Archive (2017)
11. Bauer, S., Wiest, R., Nolte, L.-P., Reyes, M.: A survey of MRI-based medical image analysis for brain tumor studies. Phys. Med. Biol. **58**, R97–R129 (2013)
12. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017)
13. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)
14. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans. Med. Imaging **35**, 1240–1251 (2016)
15. Chollet, F., et al.: Keras (2015)
16. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). http://tensorflow.org/