BRIEFnet: Deep Pancreas Segmentation using Binary Sparse Convolutions

Mattias P. Heinrich[1] and Ozan Oktay[2]

[1] Institute of Medical Informatics, University of Lübeck, Germany
[2] Biomedical Image Analysis Group, Imperial College London, UK
heinrich@imi.uni-luebeck.de, www.mpheinrich.de

Abstract. Dense prediction using deep convolutional neural networks (CNNs) has recently advanced the field of segmentation in computer vision and medical imaging. In contrast to patch-based classification, it requires only a single path through a deep network to segment every voxel in an image. However, it is difficult to incorporate contextual information without using contracting (pooling) layers, which would reduce the spatial accuracy for thinner structures. Consequently, huge receptive fields are required which might lead to disproportionate computational demand. Here, we propose to use binary sparse convolutions in the first layer as a particularly effective approach to reduce complexity while achieving high accuracy. The concept is inspired by the successful BRIEF descriptors and complemented with 1×1 convolutions (cf. network in network) to further reduce the number of trainable parameters. Sparsity is in particular important for small datasets often found in medical imaging. Our experimental validation demonstrates accuracies for pancreas segmentation in CT that are comparable with state-of-the-art deep learning approaches and registration based multi-atlas segmentation with label fusion. The whole network, which also includes a classic CNN path to improve local details, can be trained in 10 minutes. Segmenting a new scan takes 3 seconds even without using a GPU.

Keywords: context features, dilated convolutions, dense prediction

1 Introduction

The automatic segmentation of medical volumes relies on methods that are able to delineate objects boundaries on a local detail level, but also to avoid over-segmentation of similar neighbouring structures within the field-of-view. A robust method should therefore capture a large regional context. The segmentation of the pancreas in computer tomography (CT) is very important for computer assisted diagnosis of inflammation (pancreatitis) or cancer. However, this task is challenging due to the highly variable shape, a relatively poor contrast and similar neighouring abdominal structures.

In recent years, convolutional neural networks (CNN) have shown immense progress in image recognition [1], by aggregating activations of object occurrences across the whole image using deep architectures and contracting pooling

layers. Employing classic deep networks for segmentation using sliding patches results in many redundant computations and long inference times. *Dense prediction* (i.e. a parallel voxelwise segmentation) of whole images using the classical contracting architecture was extended by so-called upconvolutional layers in [2], which result in the loss of some detail information during spatial downsampling. Additional links transferring higher resolution information have therefore been proposed [3]. Fully-convolutional networks (FCN) have also been adapted for medical image segmentation [4]. The ease of adapting existing deep network for segmentation is inviting, however, may result in overly complex architectures with many trainable weights that are potentially reliant on pre-training [5] and increase the computational time for training and inference.

While fully-convolutional networks can reach impressive segmentation accuracy when designed properly and trained with enough data, it is interesting to explore whether a completely new approach for including context into dense prediction could be considered. A multiresolution deep network has been proposed in [6] for brain lesion segmentation, which uses two parallel paths for both high-resolution and low-resolution inputs. However, this approach still has to perform a series of convolutions to increase the capture range of the receptive field and might yield a higher correlation of weights across paths. In order to segment smaller abdominal organs, in particular the pancreas that has poor gray value contrast, we believe that it is of great importance to efficiently encode information about its surroundings within the CT scan. The use of sparse long-range features [7], such as local binary patterns and BRIEF [8], has shown great success within classic machine learning approaches. However, robustly and discriminately representing vastly different anatomical shapes across subjects is challenging and pancreas segmentation has not been successfully addressed using random forests or ferns. We therefore conclude that while long-range comparisons are powerful in practice, learning their optimal combination within the context of deep learning can further enhance their usefulness.

In [9], a new concept to aggregate context was introduced by using dilated convolutions with the advantage that very wide kernels with fewer trainable weights can be realised and has been used for MR segmentation in [10]. We propose to extend this idea to *binary sparse convolutions* and thereby model sparse long-range contextual relations with DCNNs to obtain simple and efficient, yet very powerful models for medical image segmentation. Our approach, which we call BRIEFnet, starts with a sparse convolution filter with a huge receptive field, so that each neuron in the following layer has only two non-zero input weights (which are restricted to be $\in \pm 1$). In order to learn relations across these binary comparisons, we succeed this layer by a 1×1 convolution, originally presented as network in network [11]. We will show that this sparse sampling enables the rapid training of expressive networks with very few parameters that outperform many recent alternative ideas for dense prediction. Reducing model complexity using binary weights is a new promising concept [12] that can substantially reduce training and inference times. Similar to [6, 9], we design a network for dense prediction without a need for contracting pooling layers. BRIEFnet can

be seen as a complimentary solution to fully-convolutional architectures with multi-resolution paths [4, 6] or holistically nested networks [5, 13]. The additional sparsity constraint in our method is in particular useful when few training scans are available. We will show in our experiments that our network, which also includes a classic CNN path for improved delineation of local details reaches comparable performance to the much more complex DCNN technique of [5] and the best multi-atlas registration with label fusion on a public abdominal CT dataset [14].

2 Method

The input to the proposed network will be a region of interest around the pancreas. While our approach is fully convolutional and the output therefore invariant to translational offsets (except for boundary effects), a bounding box initialisation helps to obtain roughly comparable organ sizes across scans and reduces the computational complexity. Since this detection is not the focus of this work, we use a manual box that is enlarged by about 300% (in volume) around the pancreas. Several accurate algorithms exist for automatic bounding box and organ localisation, e.g.[15, 16], which could be adapted for this task. The BRIEFnet is designed to use a stack of 3D slices as input and output a dense label prediction for every pixel within a stack of 2D slices. Thus a 3D volume is generated by applying the network to all slices within the region of interest. Finally, an edge-preserving smoothing of the predicted probability maps is performed, which are then thresholded to yield binary segmentations.

The key idea behind BRIEFnet is to use binary sparse convolutions in order to realise larger receptive fields while keeping the model complexity low. The overview of our complete network is given in Fig. 1. We use 3D stacks of slices of the CT scan within the bounding box as input to our framework. The images are padded to keep the same dimensions for all layers using the lowest intensity value (-1000 HU). In total, 2×1536 nonzero weights are determined, one ±1 pair per kernel, by sampling the receptive field using a uniformly random distribution (with a stride of three voxels). This random sparsity of connections in the first layer is inspired by the irregular k-space sampling found in compressed sensing. Note that it is important to increase the throughput of the central pixel as discussed in [9], which is also similar in spirit to residual learning [1]. We increase the probability of drawing the centre voxel within the receptive field so that every third weight pair contains it. Subsequently, the 1536 channels (for each voxel in the grid) are passed through tanh activation. To avoid a complete saturation of the nonlinearity, we divide the inputs by a constant (here 100). Optimising binary weights can be challenging [12] and would in this scenario lead to a very many degrees of freedom, motivating the random selection at model construction.

Why is such a sparse sampling sufficient to gather all necessary image data? The key lies in having the following 1×1 convolution that combines the output of multiple weight pairs. This locally fully-connected layer (with shared weights across all spatial position as in [11]) is able to find the optimal combination of

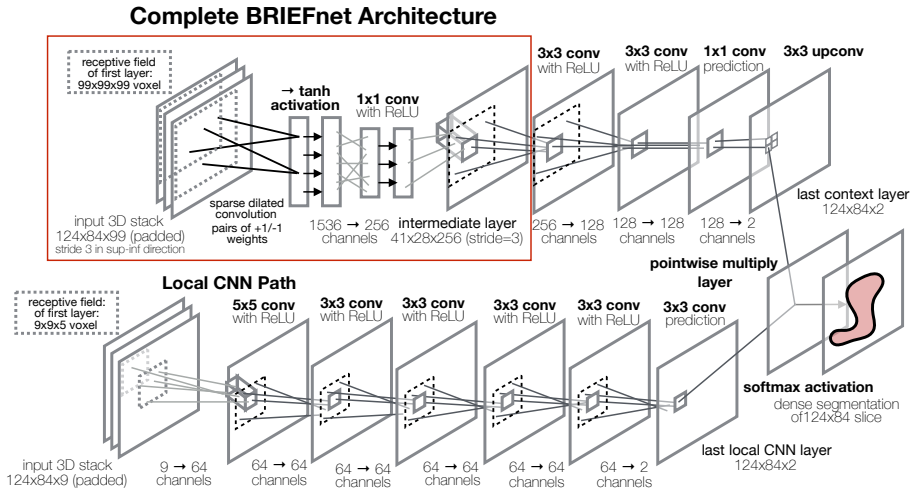

Fig. 1: Architecture of BRIEFnet (the novel part is highlighted with red box). A huge receptive field of 99×99×99 is realised by our sparse sampling (here of 1536 pixel pairs). The weights of this first layer are not learned, restricted to +1 or -1 and followed by a tanh activation. Importantly, a following 1×1 convolution combines information across channels. To reduce computation time a stride of 3 is used for the first contextual layer, requiring an up-convolution at the end. The additional local CNN path uses six traditional small convolutions with 64 channels, which are merged with the context information using a pointwise multiplication layer. Given an input 3D stack, a dense probabilistic segmentation of the centre slice is obtained. Note that each ReLU is preceded by batch-normalisation. In total there are only ≈1 million trainable weights.

sparse pixel pair activations and enables a meaningful dimensionality reduction. When multiplying the trained weights of both layers, one notices patterns that are similar to classic large convolution filters, but our framework removes redundancy and achieves a much lower complexity than most other approaches. The FCN network for semantic segmentation in [2] has over 100 million parameters and the holistically nested network of [13] over ten million. We only require one million, most of them for the 1536x256 matrix multiplication, which is particularly efficient to compute. In order to compensate local errors (e.g. due to the stride of 3 voxels), we additionally include a classic local CNN path for dense prediction, which is combined with the contextual layers using a pointwise multiplication (as done in [17]). While the input to our network is 3D, all following spatial convolutions are 2D (due to memory contraints), we therefore predict the segmentation of each 2D slice individually and stack them together, as also done for pancreas segmentation in [5]. The inference for all slices of one 3D image takes only about 3 seconds on a CPU (<1 sec. on GPU) making our approach in particular suitable for time-sensitive applications. A standard cross entropy loss layer may be unsuitable for semantic segmentation when the object class occurs less frequently. Following [5], we thus use a loss that weighs pancreas voxels more

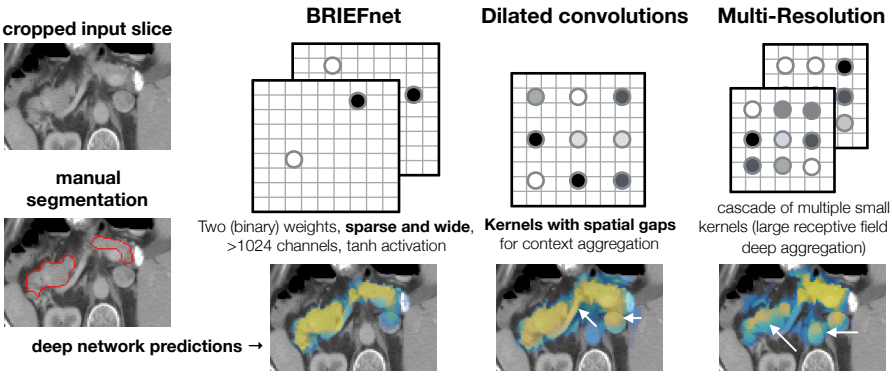

Fig. 2: Visual comparison of BRIEFnet to other recent methods for contextual aggregation. Circles represent nonzero convolutional weights (+1 is white and -1 black). The sparse layer of BRIEFnet is followed by a 1×1 kernel for cross-channel pooling as shown in Fig. 1. Employing standard dilated convolutions results in a regular and non-sparse sampling, while the multi-resolution scheme aggregates context information by successive application of classic convolution layers. The colour overlay of the final prediction of all models clearly demonstrates the ability of BRIEFnet to accurately segment the pancreas even in areas of low contrast. When using dilated convolutions neighbouring structures may be included (such as the vena cava), similarly for the multi-resolution framework, which cannot follow the narrow shape as accurately (see white arrows).

strongly. Since some slices contain no foreground at all, we adapt the weighting factors in Equation (1) of [5] to $\beta = \frac{|B|}{|F|+|B|} + \epsilon$ and $1-\beta = \frac{|F|}{|F|+|B|} + \epsilon$, where ϵ is set to 0.05. We implemented our method in MatConvNet [18] using a directed acyclic graph structure to allow for joint end-to-end training of both paths.

3 Experiments and Results

We apply our new framework to pancreas segmentation in CT, because it presents a challenging scenario where capturing the contextual information is of great value. We use the 30 subjects of the training data of the MICCAI 2015 challenge on multi-atlas segmentation of the abdomen [14][3] to evaluate and compare our approach. In addition, we include 18 scans of the VISCERAL dataset [19] only for the learning stage. We perform six-fold cross-validation with 43 scans for training and five scans for test (we do not evaluate accuracy on the VISCERAL data). All weights, expect for the binary sparse convolution layer as discussed before, are initialised using the Xavier method. We use a mini-batch size of six and stochastic gradient descent with momentum of 0.9 and logarithmically decaying learning rates from 0.1 to 0.01 over seven epochs (≈1400 iterations). To enable a reproduction of our results and further research, our processing pipeline and code are publicly available at `https://github.com/mattiaspaul/BRIEFnet`.

[3] `https://www.synapse.org/#!Synapse:syn3193805/wiki/89480`

Segmentation Results and Model Comparison: We compare the BRIEFnet to two alternative architectures for context aggregation: dilated convolutions as recently proposed in computer vision by [9] and a multi-resolution dual path network similar to [6]. All architectures are designed to have similarly large receptive fields when considering all layers and a comparable number of parameters. The dilated convolution network has a convolution kernel of size $17\times17\times17$ with a dilation of six voxels (in all 3 dimensions) and 256 channels in the first layer, but is otherwise identical to BRIEFnet. A multi-resolution dual path network is designed with a $9\times9\times33$ convolution filter (with stride 3) followed by six 5×5 convolutions with 128 channels each. All architectures share the same local CNN path. Visual examples of the predicted organ probabilities together with a schematic explanation of the structural network differences are shown in Fig. 2.

We use the same 3D images for all models, obtained using the enlarged bounding boxes, which contain on average 2-3% pancreas voxels. For the contextual path in each network, the images are smoothed by a Gaussian kernel with $\sigma = 1.5$ voxels. The intensity range is mapped from the interval of $(-160, 240)$ HU to $(-1, +1)$ as in [5]. An edge-preserving smoothing of the foreground probabilities as proposed in [20] is performed as only post-processing step. A global binary segmentation threshold that represents the best tradeoff across all scans is selected for each tested model separately for a fair comparison. We obtained Dice scores of 64.5% for our proposed BRIEFnet, 59.6% for dilated convolutions and 47.6% for the multi-resolution network. Demonstrating a clear advantage of our new concept. When removing the local path the accuracy drops by $\approx 5.6\%$.

While the use of manual bounding boxes might yield overly optimistic results, it provides a good comparison among the tested models. In many other areas, e.g. for face landmark detection and segmentation, specific bounding boxes are commonly shared with public datasets to exclude the dependency of the following steps on this initialisation. Automatic bounding box estimation is nevertheless of great importance for future work. Encouraging progress has recently been made in computer vision. Another approach to further boost accuracy is to predict organ probabilities for multiple region proposals and fuse the results [5]. Since no data augmentation was performed in our experiments, we can conclude that our compact model together with the sparsity constraint generalises well even for small datasets and further improvements are expected for larger scale experiments.

Comparison to State-of-the-Art: In general Dice scores for pancreas segmentation are relatively low. In [14] overlap scores of 40% and 49% have been reported for two different multi-atlas techniques. The standardised nonlinear registration, which uses NiftyReg (the parameterisation of which is described in [21]) in a leave-one-out validation, followed by a majority vote achieves a Dice score of only 26%, highlighting the difficulty of this dataset. Employing advanced label fusion can increase the accuracy to 64%. Holistically nested networks together with superpixel based region proposals and random forest spatial aggregation as proposed in [5] (trained on a much larger dataset of 82 cancer patients) achieved impressive 62% (or 66% when employing the fusion of multiple models).

Computational complexity: One of the main advantages of our framework is its low model complexity. The currently employed deep segmentation models that are originally based on large scale image classification require training times of several hours (to days) [2]. Our network has only one million free parameters and we found that by employing batch normalisation the training converges after just seven epochs, yielding a training time of around ten minutes on a GTX 950. Even more compelling is the inference time of three seconds on a CPU for the segmentation of a new 3D image, since most deep networks employed for medical scans require several minutes on a GPU [5, 6]. This opens up new possibilities for using DCNNs for time-sensitive applications such as image-guided interventions.

4 Discussion and Conclusion

We have presented a new deep convolutional architecture called BRIEFnet that enables a very efficient modelling of contextual information for semantic segmentation. It realises a very large receptive field by using binary sparse convolutions with only two nonzero weights each - similar to BRIEF features [8] - followed by a nonlinear activation. In addition, a subsequent 1×1 convolution layer enables a suitable combination of these elementary feature activations and also serves as a dimensionality reduction. In contrast to fully convolutional networks (FCN) [2, 3] no contracting pathway is required, reducing the number of trainable weights to one million ($100\times$ less than FCN) and realising training times of only ten minutes. The segmentation of a new image is highly efficient due to the dense prediction of whole image slices and the large matrix multiplications (in particular in the 1×1 convolution layer) leading to test times of three seconds on a CPU. Exemplary results for CT pancreas segmentation demonstrate high accuracies, comparable to more complex models, and with substantial improvements over alternative ways of incorporating nonlocal information into semantic segmentation. Making no domain specific assumptions, our model would be directly applicable to other anatomical regions. In future work, it would be of interest to include data augmentation and/or model fusion to further exploit the fast training and inference time of BRIEFnet and enhance its generalisability e.g. by learning offsets [22]. So far only the first layer captures true 3D information, therefore further improvements could be gained by employing only 3D convolutions throughout the network.

Acknowledgements: We thank Maurice Sambale, whose work during his B.Sc. thesis gave inspiration to some of the new ideas in this paper.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 3431–3440

3. 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, Springer (2015) 234–241
4. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision, IEEE (2016) 565–571
5. Roth, H.R., Lu, L., Farag, A., Sohn, A., Summers, R.M.: Spatial aggregation of holistically-nested networks for automated pancreas segmentation. In Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W., eds.: MICCAI 2016 LNCS, Springer (2016) 451–459
6. 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 Imag Anal **36** (2017) 61–78
7. Criminisi, A., Shotton, J., Bucciarelli, S.: Decision forests with long-range spatial context for organ localization in CT volumes. In: MICCAI PMMIA. (2009) 69–80
8. Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P.: BRIEF: Computing a local binary descriptor very fast. PAMI **34**(7) (2012) 1281–1298
9. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
10. Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease. In: MICCAI RAMBO. (2016) 95–102
11. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv:1312.4400 (2013)
12. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: XNOR-Net: Imagenet classification using binary convolutional neural networks. In Leibe, B., Matas, J., Sebe, N., Welling, M., eds.: ECCV, Springer (2016) 525–542
13. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV. (2015) 1395–1403
14. Xu, Z., Burke, R.P., Lee, C.P., Baucom, R.B., Poulose, B.K., Abramson, R.G., Landman, B.A.: Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning. Med Imag Anal **24**(1) (2015) 18–27
15. de Vos, B., Wolterink, J., de Jong, P., Leiner, T., Viergever, M., Isgum, I.: ConvNet-Based localization of anatomical structures in 3D medical images. IEEE Transactions on Medical Imaging (2017)
16. Xu, Z., Panjwani, S.A., Lee, C.P., Burke, R.P., Baucom, R.B., Poulose, B.K., Abramson, R.G., Landman, B.A.: Evaluation of body-wise and organ-wise registrations for abdominal organs. In: SPIE Medical Imaging. (2016) 97841
17. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using CNNs. In Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W., eds.: MICCAI 2016 LNCS, Springer (2016) 230–238
18. Vedaldi, A., Lenc, K.: Matconvnet: Convolutional neural networks for matlab. In: Proc ACM. (2015) 689–692
19. Jiménez-del Toro, O., et al.: Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. IEEE Trans Med Imag **35**(11) (2016) 2459–2475
20. Heinrich, M.P., Blendowski, M.: Multi-organ segmentation using vantage point forests and binary context features. In Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W., eds.: MICCAI 2016 LNCS, Springer (2016) 598–606
21. Xu, Z., Lee, C., Heinrich, M., Modat, M., Rueckert, D., Ourselin, S., Abramson, R., Landman, B.: Evaluation of six registration methods for the human abdomen on clinically acquired CT. IEEE Trans Biomed Eng (2016) 1–10
22. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. arXiv preprint arXiv:1703.06211 (2017)