The Importance of Skip Connections in Biomedical Image Segmentation

Michal Drozdzal[1,2(✉)], Eugene Vorontsov[1,2(✉)], Gabriel Chartrand[1,3], Samuel Kadoury[2,4], and Chris Pal[2,5]

[1] Imagia Inc., Montréal, Canada
{michal,eugene,gabriel}@imagia.com
[2] École Polytechnique de Montréal, Montréal, Canada
{samuel.kadoury,christopher.pal}@polymtl.ca
[3] Université de Montréal, Montréal, Canada
[4] CHUM Research Center, Montréal, Canada
[5] Montreal Institute for Learning Algorithms, Montréal, Canada

Abstract. In this paper, we study the influence of both long and short skip connections on Fully Convolutional Networks (FCN) for biomedical image segmentation. In standard FCNs, only long skip connections are used to skip features from the contracting path to the expanding path in order to recover spatial information lost during downsampling. We extend FCNs by adding short skip connections, that are similar to the ones introduced in residual networks, in order to build very deep FCNs (of hundreds of layers). A review of the gradient flow confirms that for a very deep FCN it is beneficial to have both long and short skip connections. Finally, we show that a very deep FCN can achieve near-to-state-of-the-art results on the EM dataset without any further post-processing.

Keywords: Semantic segmentation · FCN · ResNet · Skip connections

1 Introduction

Semantic segmentation is an active area of research in medical image analysis. With the introduction of Convolutional Neural Networks (CNN), significant improvements in performance have been achieved in many standard datasets. For example, for the EM ISBI 2012 dataset [2], BRATS [13] or MS lesions [18], the top entries are built on CNNs [3,4,7,15].

All these methods are based on Fully Convolutional Networks (FCN) [12]. While CNNs are typically realized by a contracting path built from convolutional, pooling and fully connected layers, FCN adds an expanding path built with deconvolutional or unpooling layers. The expanding path recovers spatial information by merging features skipped from the various resolution levels on the contracting path.

M. Drozdzal and E. Vorontsov—Equal contribution.

G. Carneiro et al. (Eds.): LABELS 2016/DLMIA 2016, LNCS 10008, pp. 179–187, 2016.
DOI: 10.1007/978-3-319-46976-8_19

Variants of these skip connections are proposed in the literature. In [12], upsampled feature maps are summed with feature maps skipped from the contractive path while [15] concatenate them and add convolutions and non-linearities between each upsampling step. These skip connections have been shown to help recover the full spatial resolution at the network output, making fully convolutional methods suitable for semantic segmentation. We refer to these skip connections as long skip connections.

Recently, significant network depth has been shown to be helpful for image classification [8,9,14,20]. The recent results suggest that depth can act as a regularizer [8]. However, network depth is limited by the issue of vanishing gradients when backpropagating the signal across many layers. In [20], this problem is addressed with additional levels of supervision, while in [8,9] skip connections are added around non-linearities, thus creating shortcuts through which the gradient can flow uninterrupted allowing parameters to be updated deep in the network. Moreover, [19] have shown that these skip connections allow for faster convergence during training. We refer to these skip connections as short skip connections.

In this paper, we explore deep, fully convolutional networks for semantic segmentation. We expand FCN by adding short skip connections that allow us to build very deep FCNs. With this setup, we perform an analysis of short and long skip connections on a standard biomedical dataset (EM ISBI 2012 challenge data). We observe that short skip connections speed up the convergence of the learning process; moreover, we show that a very deep architecture with a relatively small number of parameters can reach near-state-of-the-art performance on this dataset. Thus, the contributions of the paper can be summarized as follows:

- We extend Residual Networks to fully convolutional networks for semantic image segmentation (see Sect. 2).
- We show that a very deep network without any post-processing achieves performance comparable to the state of the art on EM data (see Sect. 3.1).
- We show that long and short skip connections are beneficial for convergence of very deep networks (see Sect. 3.2)

2 Residual Network for Semantic Image Segmentation

Our approach extends Residual Networks [8] to segmentation tasks by adding an expanding (upsampling) path (Fig. 1(a)). We perform spatial reduction along the contracting path (left) and expansion along the expanding path (right). As in [12,15], spatial information lost along the contracting path is recovered in the expanding path by skipping equal resolution features from the former to the latter. Similarly to the short skip connections in Residual Networks, we choose to sum the features on the expanding path with those skipped over the long skip connections.

We consider three types of blocks, each containing at least one convolution and activation function: bottleneck, basic block, simple block (Fig. 1(b)–(d)). Each block is capable of performing batch normalization on its inputs as well as spatial downsampling at the input (marked blue; used for the contracting path) and

spatial upsampling at the output (marked yellow; for the expanding path). The bottleneck and basic block are based on those introduced in [8] which include short skip connections to skip the block input to its output with minimal modification, encouraging the path through the non-linearities to learn a residual representation of the input data. To minimize the modification of the input, we apply no transformations along the short skip connections, except when the number of filters or the spatial resolution needs to be adjusted to match the block output. We use 1×1 convolutions to adjust the number of filters but for spatial adjustment we rely on simple decimation or simple repetition of rows and columns of the input so as not to increase the number of parameters. We add an optional dropout layer to all blocks along the residual path.

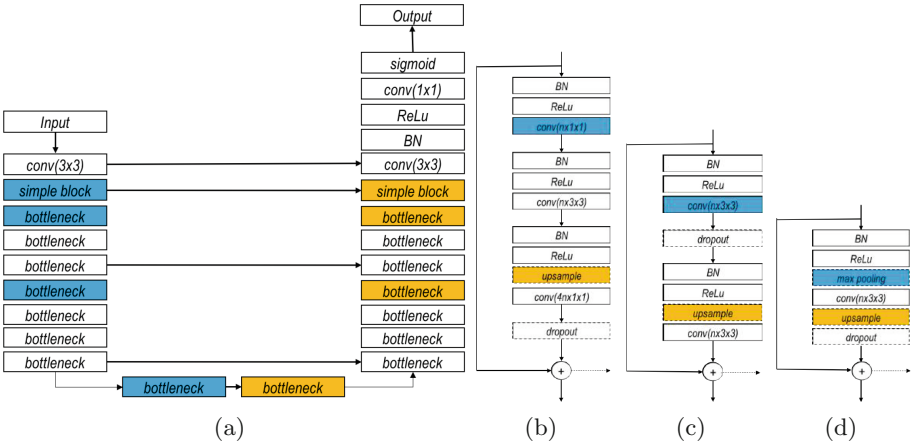


Fig. 1. An example of residual network for image segmentation. (a) Residual Network with long skip connections built from bottleneck blocks, (b) bottleneck block, (c) basic block and (d) simple block. Blue color indicates the blocks where an downsampling is optionally performed, yellow color depicts the (optional) upsampling blocks, dashed arrow in figures (b), (c) and (d) indicates possible long skip connections. Note that all blocks (b), (c) and (d) can have a dropout layer (depicted with dashed line rectangle). (Color figure online)

We experimented with both binary cross-entropy and dice loss functions. Let $o_i \in [0, 1]$ be the i^{th} output of the last network layer passed through a sigmoid non-linearity and let $y_i \in \{0, 1\}$ be the corresponding label. The binary cross-entropy is then defined as follows:

$$L_{bce} = \sum_i y_i \log o_i + (1 - y_i) \log (1 - o_i) \tag{1}$$

The dice loss is:

$$L_{Dice} = -\frac{2 \sum_i o_i y_i}{\sum_i o_i + \sum_i y_i} \tag{2}$$

We implemented the model in Keras [5] using the Theano backend [1] and trained it using RMSprop [21] (learning rate 0.001) with weight decay set to 0.001. We also experimented with various levels of dropout.

3 Experiments

In this section, we test the model on electron microscopy (EM) data [2] (Sect. 3.1) and perform an analysis on the importance of the long and short skip connections (Sect. 3.2).

3.1 Segmenting EM Data

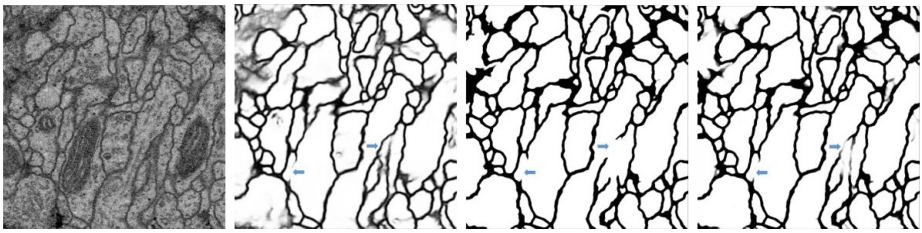
EM training data consist of 30 images (512 × 512 pixels) assembled from serial section transmission electron microscopy of the Drosophila first instar larva ventral nerve cord. The test set is another set of 30 images for which labels are not provided. Throughout the experiments, we used 25 images for training, leaving 5 images for validation.

During training, we augmented the input data using random flipping, sheering, rotations, and spline warping. We used the same spline warping strategy as [15]. We used full resolution (512 × 512) images as input without applying random cropping for data augmentation. For each training run, the model version with the best validation loss was stored and evaluated. The detailed description of the highest performing architecture used in the experiments is shown in Table 1.

Interestingly, we found that while the predictions from models trained with cross-entropy loss were of high quality, those produced by models trained with

Table 1. Detailed model architecture used in the experiments. Repetition number indicates the number of times the block is repeated.

Layer name	Block type	Output resolution	Output width	Repetition number
Down 1	conv 3 × 3	512 × 512	32	1
Down 2	simple block	256 × 256	32	1
Down 3	bottleneck	128 × 128	128	3
Down 4	bottleneck	64 × 64	256	8
Down 5	bottleneck	32 × 32	512	10
Across	bottleneck	32 × 32	1024	3
Up 1	bottleneck	64 × 64	512	10
Up 2	bottleneck	128 × 128	256	8
Up 3	bottleneck	256 × 256	128	3
Up 4	simple block	512 × 512	32	1
Up 5	conv 3 × 3	512 × 512	32	1
Classifier	conv 1 × 1	512 × 512	1	1



(a) (b) (c) (d)

Fig. 2. Qualitative results on the test set. (a) original image, (b) prediction for a model trained with binary cross-entropy, (c) prediction of the model trained with dice loss and (d) model trained with dice loss with 0.2 dropout at the test time.

Table 2. Comparison to published entries for EM dataset. For full ranking of all submitted methods please refer to challenge web page: http://brainiac2.mit.edu/isbi_challenge/leaders-board-new. We note the number of parameter, the use of post-processing, and the use of model averaging only for FCNs.

Method	V_{rand}	V_{info}	FCN	Post-processing	Average over	Parameters (M)
CUMedVision [4]	0.977	0.989	YES	YES	6	8
Unet [15]	0.973	0.987	YES	NO	7	33
IDSIA [6]	0.970	0.985	NO	-	-	-
motif [23]	0.972	0.985	NO	-	-	-
SCI [11]	0.971	0.982	NO	-	-	-
optree-idsia [22]	0.970	0.985	NO	-	-	-
PyraMiD-LSTM [17]	0.968	0.983	NO	-	-	-
Ours (L_{Dice})	0.969	0.986	YES	NO	Dropout	11
Ours (L_{bce})	0.957	0.980	YES	NO	1	11

the Dice loss appeared visually cleaner since they were almost binary; borders that would appear fuzzy in the former (see Fig. 2(b)) would be left as gaps in the latter (Fig. 2(c)). However, we found that the border continuity can be improved for models with the Dice loss by implicit model averaging over output samples drawn at test time, using dropout [10] (Fig. 2(d)). This yields better performance on the validation and test metrics than the output of models trained with binary cross-entropy (see Table 2).

Two metrics used in this dataset are: Maximal foreground-restricted Rand score after thinning (V_{rand}) and maximal foreground-restricted information theoretic score after thinning (V_{info}). For a detailed description of the metrics, please refer to [2].

Our results are comparable to other published results that establish the state of the art for the EM dataset (Table 2). Note that we did not do any post-processing of the resulting segmentations. We match the performance of UNet, for which predictions are averaged over seven rotations of the input images, while using less

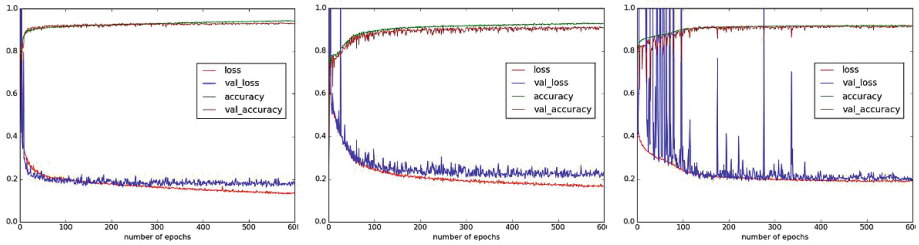


Fig. 3. Training and validation losses and accuracies for different network setups: (a) model 1: long and short skip connections enabled, (b) model 2: only short skip connections enabled and (c) model 3: only long skip connections enabled.

Table 3. Best validation loss and its corresponding training loss for each model.

Method	Training loss	Validation loss
Long and short skip connections	0.163	0.162
Only short skip connections	0.188	0.202
Only long skip connection	0.205	0.188

parameters and without sophisticated class weighting. Note that among other FCN available on the leader board, CUMedVision is using post-processing in order to boost performance.

3.2 On the Importance of Skip Connections

The focus in the paper is to evaluate the utility of long and short skip connections for training fully convolutional networks for image segmentation. In this section, we investigate the learning behavior of the model with short and with long skip connections, paying specific attention to parameter updates at each layer of the network. We first explored variants of our best performing deep architecture (from Table 1), using binary cross-entropy loss. Maintaining the same hyperparameters, we trained (Model 1) with long and short skip connections, (Model 2) with only short skip connections and (Model 3) with only long skip connections. Training curves are presented in Fig. 3 and the final loss and accuracy values on the training and the validation data are presented in Table 3.

We note that for our deep architecture, the variant with both long and short skip connections is not only the one that performs best but also converges faster than without short skip connections. This increase in convergence speed is consistent with the literature [19]. Not surprisingly, the combination of both long and short skip connections performed better than having only one type of skip connection, both in terms of performance and convergence speed. At this depth,

a network could not be trained without any skip connections. Finally, short skip connections appear to stabilize updates (note the smoothness of the validation loss plots in Figs. 3(a) and (b) as compared to Fig. 3(c)).

We expect that layers closer to the center of the model can not be effectively updated due to the vanishing gradient problem which is alleviated by short skip connections. This identity shortcut effectively introduces shorter paths through fewer non-linearities to the deep layers of our models. We validate this empirically on a range of models of varying depth by visualizing the mean model parameter updates at each layer for each epoch (see sample results in Fig. 4). To simplify the analysis and visualization, we used simple blocks instead of bottleneck blocks.

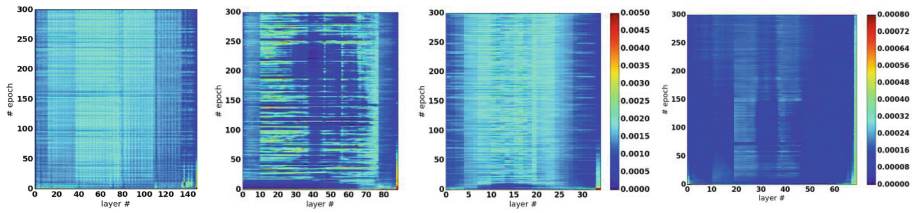


Fig. 4. Weight updates in different network setups: (a) the best performing model with long and short skip connections enabled, (b) only long skip connections enabled with 9 repetitions of simple block, (c) only long skip connections enabled with 3 repetitions of simple block and (d) only long skip connections enabled with 7 repetitions of simple block, without batch normalization. Note that due to a reduction in the learning rate for Figure (d), the scale is different compared to Figures (a), (b) and (c).

Parameter updates appear to be well distributed when short skip connections are present (Fig. 4(a)). When the short skip connections are removed, we find that for deep models, the deep parts of the network (at the center, Fig. 4(b)) get few updates, as expected. When long skip connections are retained, at least the shallow parts of the model can be updated (see both sides of Fig. 4(b)) as these connections provide shortcuts for gradient flow. Interestingly, we observed that model performance actually drops when using short skip connections in those models that are shallow enough for all layers to be well updated (eg. Figure 4(c)). Moreover, batch normalization was observed to increase the maximal updatable depth of the network. Networks without batch normalization had diminishing updates toward the center of the network and with long skip connections were less stable, requiring a lower learning rate (eg. Figure 4(d)).

It is also interesting to observe that the bulk of updates in all tested model variations (also visible in those shown in Fig. 4) were always initially near or at the classification layer. This follows the findings of [16], where it is shown that even randomly initialized weights can confer a surprisingly large portion of a model's performance after training only the classifier.

4 Conclusions

In this paper, we studied the influence of skip connections on FCN for biomedical image segmentation. We showed that a very deep network can achieve results near the state of the art on the EM dataset without any further post-processing. We confirm that although long skip connections provide a shortcut for gradient flow in shallow layers, they do not alleviate the vanishing gradient problem in deep networks. Consequently, we apply short skip connections to FCNs and confirm that this increases convergence speed and allows training of very deep networks.

Acknowledgements. We would like to thank all the developers of Theano and Keras for providing such powerful frameworks. We gratefully acknowledge NVIDIA for GPU donation to our lab at École Polytechnique. The authors would like to thank Lisa di Jorio, Adriana Romero and Nicolas Chapados for insightful discussions. This work was partially funded by Imagia Inc., MITACS (grant number IT05356) and MEDTEQ.

References

1. Al-Rfou, R., Alain, G., Almahairi, A., et al.: Theano: a python framework for fast computation of mathematical expressions. CoRR abs/1605.02688 (2016)
2. Arganda-Carreras, I., Turaga, S.C., Berger, D.R., et al.: Crowdsourcing the creation of image segmentation algorithms for connectomics. Front. Neuroanat. **9**, 142 (2015). doi:10.3389/fnana.2015.00142
3. Brosch, T., Tang, L.Y.W., Yoo, Y., et al.: Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE TMI **35**(5), 1229–1239 (2016)
4. Chen, H., Qi, X., Cheng, J., Heng, P.A.: Deep contextual networks for neuronal structure segmentation. In: Proceedings of the 13th AAAI Conference on Artificial Intelligence, 12–17 February 2016, Phoenix, Arizona, USA, pp. 1167–1173 (2016)
5. Chollet, F.: Keras (2015). https://github.com/fchollet/keras
6. Ciresan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. In: NIPS, vol. 25, pp. 2843–2851. Curran Associates, Inc. (2012)
7. Havaei, M., Davy, A., Warde-Farley, D., et al.: Brain tumor segmentation with deep neural networks. CoRR abs/1505.03540 (2015)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. CoRR abs/1603.05027 (2016)
10. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. CoRR abs/1511.02680 (2015)
11. Liu, T., Jones, C., Seyedhosseini, M., Tasdizen, T.: A modular hierarchical approach to 3D electron microscopy image segmentation. J. Neurosci. Methods **226**, 88–102 (2014)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, November 2015 (to appear)

13. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015). doi:10.1109/TMI.2014.2377694
14. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: FitNets: hints for thin deep nets. CoRR abs/1412.6550 (2014)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. CoRR abs/1505.04597 (2015)
16. Saxe, A., Koh, P.W., Chen, Z., Bhand, M., Suresh, B., Ng, A.Y.: On random weights and unsupervised feature learning. In: Getoor, L., Scheffer, T. (eds.) Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 1089–1096. ACM, New York (2011)
17. Stollenga, M.F., Byeon, W., Liwicki, M., Schmidhuber, J.: Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. CoRR abs/1506.07452 (2015)
18. Styner, M., Lee, J., Chin, B., et al.: 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation, November 2008
19. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR abs/1602.07261 (2016)
20. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CoRR abs/1409.4842 (2014)
21. Tieleman, T., Hinton, G.: Lecture 6.5—RmsProp: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw. Mach. Learn. (2012)
22. Uzunbaş, M.G., Chen, C., Metaxsas, D.: Optree: a learning-based adaptive watershed algorithm for neuron segmentation. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part I. LNCS, vol. 8673, pp. 97–105. Springer International Publishing, Cham (2014)
23. Wu, X.: An iterative convolutional neural network algorithm improves electron microscopy image segmentation. CoRR abs/1506.05849 (2015)