Semantic Segmentation Through Graph Neural Network Blocks

Darwin Saire Pilco [1] **Adín Ramírez Rivera** [1] **Salvatore Tabbone** [2]

Abstract

Semantic segmentation task aims to create a dense classification by labeling pixel-wise each object present in images. Convolutional Neural Network (CNN) approaches have been proved useful by exhibiting the best results in this task. However, some challenges remain, such as the low-resolution of feature maps and the loss of spatial precision, both produced in the CNNs by limited local neighborhoods, i.e., filters with small size. In this work, we propose an encoder-decoder architecture with skip connections based on Graph Neural Network (GNN) (hereafter called GNN-block). This GNN-block proved to have a greater receptive field by having a global vision of objects and their relationships, thus providing additional global information to the model. Finally, we present preliminary results on Cityscape database, achieving close performance with state-of-the-art.

1. Introduction

Humans possess a remarkable ability to parse images simply by looking at it. In a blink of an eye, we are able to fully analyze an image and separate all the components present on it. Furthermore, we can easily generalize from observing a set of objects to recognizing objects that have never been seen before. The human ability to separate components (i.e., join regions) in an image according to some features is called image segmentation (Gonzalez & Woods, 2006). Trying to reproduce this human skill on a computer is not an easy task, and several approaches were proposed to address it (Chouhan et al., 2019). Nevertheless, the segmentation task continues to be challenging, due in large part to variability, i.e., there is a considerable variation in pose, appearance, viewpoint, illumination, and occlusion throughout the image. Thus, a type of segmentation commonly used is semantic segmentation, which is an essential part of the pipeline projects since it extracts and analyzes useful information by classifying the regions into an image. For instance, self-drive vehicles (Zhou et al., 2019), segmentation on X-ray (Bullock et al., 2019), crown detection on dental X-ray (Wang et al., 2016), brain tumor segmentation (Pereira et al., 2019), and remote sensing (Bokhovkin & Burnaev, 2019). Note, by improving the segmentation stage, the final result is also improved. In recent years the Fully Convolutional Networks (FCN) (Long et al., 2016) achieve significant improvement in semantic segmentation task, by converting fully connected layers into convolutional layers and upscale operations. However, with this approach, new problems have been observed, such as (Chen et al., 2017a; Lin et al., 2017b): i) the low-resolution obtained in the output of the CNNs; and ii) the loss of spatial precision of objects within the image. Following, we exhibit different models created to deal with these problems.

FCN were used with post-processing steps. Conditional Random Fields (CRF) (Zheng et al., 2015) or Gaussian CRF (Vemulapalli et al., 2016) are common post-processing steps but are computationally expensive; consequently, embedding it within a network is a viable solution (Chen et al., 2017a). Other authors (Liu et al., 2018; Chen et al., 2018a) proposed to obtain a fine adjustment from the bounding boxes. Instead of making an abrupt prediction of the last layer of CNN, the hourglass approach (Ronneberger et al., 2015; Badrinarayanan et al., 2017; Amirul Islam et al., 2017) created an up-sampling stage in a controlled manner (deconvolutions and unpooling). Moreover, models take into account different scales (Lin et al., 2017a). These models get a full semantic map in low-resolution (coarse prediction map), then refine it with different fusion operations, e.g., fusion cascade (Zhao et al., 2018) and attention blocks (Yu et al., 2018). Contrary to multi-scale models, the approaches that use Atrous Spatial Pyramid Pooling (ASSP) (Chen et al., 2017a;b; Zhao et al., 2017; Chen et al., 2018b; Valada et al., 2019) modify the filters size instead of the size of the images. This modification is achieved using atrous convolution (Chen et al., 2017a), i.e., sparse filters, to generate features with large receptive field without sacrificing spatial resolution. In theory, this should be true, but later experiments showed that there are still insufficiencies to get global

*Equal contribution [1]Institute of Computing, University of Campinas, Campinas, Brazil [2]Université de Lorraine, CNRS, LORIA, Vandœuvre-lès-Nancy, France. Correspondence to: Darwin Saire Pilco <darwin.pilco@ic.unicamp.br>, Salvatore Tabbone <antoine.tabbone@univ-lorraine.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

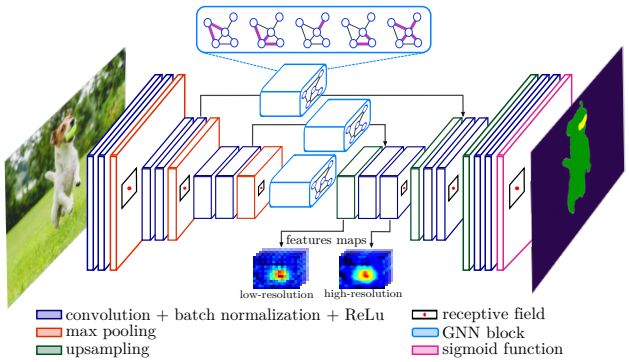

Figure 1. Illustration of our proposed architecture, which works with the local information (encoder and decoder stage) and with the global information processed by the GNN-blocks (skip connections).

features (Wang et al., 2018). Although previous models improve the problem of semantic segmentation, especially the low-resolution of output maps, the issue of spatial precision loss persists. It is produced in the CNNs by limited local neighborhoods, i.e., filters with small size and regular shape. Then, the next stage is dealing with this problem. Thus, current approaches (Chen et al., 2019; Li & Gupta, 2018) based on GNNs proved to have a greater receptive field by having a global vision of objects. Unlike these models that perform graph convolution only in latent space, we create a GNN-block capable of being used anywhere on the network (e.g., between successive CNN layers or into skip connections).

2. Methodology

In this work, we design a new deep learning architecture that is end-to-end trainable to address the semantic segmentation task on images. Our architecture combines local features extraction of CNNs with the global features extraction of GNNs and their irregular connections between pixels through GNN-blocks (light blue squares in Fig. 1). Thus, our neural network aims to produce densely labeled images.

We base our architecture on the well-behaved UNet (Ronneberger et al., 2015) model to work with local and global (poor) information through its encoder and decoder stage. Furthermore, additional global information is embedded in our model through the creation of GNN-blocks. For this, we carry the image features from the original space (Euclidean space) to a graph space. That is, from pixel features to nodes ones. Here, we transform the features of the nodes through convolutions on the graph (Kipf & Welling, 2017). Then, we return to the original space and concatenate it with the features from the down-level (skip connection in Fig. 1).

At a given layer l on the alternating process, we use a function $f(X)$ to bring pixel features $X \in \mathbb{R}^{h\times w\times c}$ to d_l hidden features, $H^{(l)} \in \mathbb{R}^{n\times d_l}$, for each of our n nodes, and the set

Table 1. Results on Cityscape validation set for semantic segmentation task, using 11 classes and with crop size of 384 768. Table taken from (Valada et al., 2019). Our model is denoted by the symbol *. Best performances in bold.

Model	Sky	Building	Road	Sidewalk	Fence	Vegetation	Pole	Car	Sign	Person	Cyclist	mIoU
SegNet	73.74	79.29	92.70	59.88	13.63	81.89	26.18	78.83	31.44	45.03	43.46	52.17
FCN8	76.51	83.97	93.82	67.67	24.91	86.38	31.71	84.80	50.92	59.89	59.11	59.97
FastNet	77.69	86.25	94.97	72.99	31.02	88.06	38.34	88.42	52.34	61.76	61.83	68.52
DeconvNet	89.38	83.08	95.26	68.07	27.58	85.80	34.20	85.01	27.62	45.11	41.11	62.02
DeepLabv2	74.28	81.66	90.86	63.30	26.29	84.33	27.96	86.24	44.79	58.89	60.92	63.59
ParseNet	77.57	86.81	95.27	74.02	33.31	87.37	38.24	88.99	53.34	63.25	63.87	69.28
DeepLabv3	92.82	89.02	96.74	78.13	41.00	90.81	49.74	91.02	64.48	66.52	66.98	75.21
GNN-block*	93.64	88.69	96.42	74.63	41.46	90.97	52.30	89.79	69.40	70.36	68.59	76.02
AdapNet++	**94.18**	**91.49**	**97.93**	**84.40**	**54.98**	**92.09**	**58.85**	**93.86**	**72.61**	**75.52**	**72.90**	**80.80**

of edges encoded into a sparse matrix $A^{(l)} \in [0,1]^{n\times n}$ that represents the graph. Then, to obtain global features, we use convolutional graph operations (Kipf & Welling, 2017)

$$H^{(l+1)} = \sigma_l\left(\tau\left(A^{(l)}\right) H^{(l)} W^{(l)}\right), \tag{1}$$

where $W^{(l)} \in \mathbb{R}^{d_l\times d_{l+1}}$ is the learnable weight matrix for the lth layer, $\sigma_l(\cdot)$ is a non-linear function, and $\tau(\cdot)$ is a symmetric normalization transformation of the sparse matrix, defined by

$$\tau\left(A^{(l)}\right) = \left(\hat{D}^{(l)}\right)^{-\frac{1}{2}} \left(A^{(l)} + I_n\right) \left(\hat{D}^{(l)}\right)^{-\frac{1}{2}}, \tag{2}$$

where $\hat{D}^{(l)}$ is the degree matrix of the graph plus identity, that is,

$$\hat{D}^{(l)} = D^{(l)} + I_n, \tag{3}$$

where $D^{(l)}$ is the degree matrix of $A^{(l)}$, and I_n is the identity matrix of size $n \times n$. Finally, we use the function $g(H^{(l+1)})$ to convert the features from node to pixel (i.e., from graph space to original space). In summary, our GNN block produces a feature vector $X' \in \mathbb{R}^{h\times w\times d_{l+1}}$ defined by,

$$X' = g\left(\sigma_l\left(\tau\left(A^{(l)}\right) f(X) W^{(l)}\right)\right). \tag{4}$$

3. Experiments and Conclusions

We use the Cityscapes (Cordts et al., 2016) dataset, with 2979 images in the training set, and 500 images in the validation set. Each image was resized to 768 $\times$ 384 pixels with labels of 11 semantic classes (Valada et al., 2019); We present quantitative and qualitative results in Table 1 and supplementary material, respectively.

In this work, we show that by adding global information obtained through holistic operations (i.e., graph convolution), we can improve the performance of the semantic segmentation task achieving results proximate to the state-of-the-art. Note that our model only uses unimodal (RGB) information as opposed to multimodal model AdapNet++ (Valada et al., 2019) (RGB and depth maps).

References

Amirul Islam, M., Rochan, M., Bruce, N. D., and Wang, Y. Gated feedback refinement network for dense image labeling. In *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, pp. 3751–3759, 2017.

Badrinarayanan, V., Kendall, A., and Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12): 2481–2495, 2017.

Bokhovkin, A. and Burnaev, E. Boundary loss for remote sensing imagery semantic segmentation. In *Inter. Symp. Neural Netw. (ISNN)*, pp. 388–401. Springer, 2019.

Bullock, J., Cuesta-Lázaro, C., and Quera-Bofarull, A. XNet: a convolutional neural network (CNN) implementation for medical x-ray image segmentation suitable for small datasets. In *Med. Imag. Biomed. Appl. Mol. Struc. Funct. Imag.*, volume 10953. International Society for Optics and Photonics, 2019.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017a.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv*, (arXiv:1706.05587v1), 2017b.

Chen, L.-C., Hermans, A., Papandreou, G., Schroff, F., Wang, P., and Adam, H. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, pp. 4013–4022, 2018a.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conf. Comput. Vis. (ECCV)*, pp. 801–818, 2018b.

Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., and Kalantidis, Y. Graph-based global reasoning networks. In *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, pp. 433–442, 2019.

Chouhan, S. S., Kaul, A., and Singh, U. P. Image segmentation using computational intelligence techniques: Review. *Archives of Computational Methods in Engineering*, 26(3):533–596, Jul 2019. ISSN 1886-1784.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, 2016.

Gonzalez, R. and Woods, R. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Upper Saddle River, NJ, USA, 2006. ISBN 013168728X.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *Inter. Conf. Learn. Represent. (ICLR)*, 2017.

Li, Y. and Gupta, A. Beyond grids: Learning graph representations for visual recognition. In *Adv. Neural Inf. Process. Sys. (NeurIPS)*, pp. 9225–9235, 2018.

Lin, G., Milan, A., Shen, C., and Reid, I. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, pp. 1925–1934, 2017a.

Lin, G., Shen, C., Van Den Hengel, A., and Reid, I. Exploring context with deep structured models for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1352–1366, 2017b.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. Path aggregation network for instance segmentation. In *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, pp. 8759–8768, 2018.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP(99):1–1, 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2572683.

Pereira, S., Pinto, A., Amorim, J., Ribeiro, A., Alves, V., and Silva, C. A. Adaptive feature recombination and recalibration for semantic segmentation with fully convolutional networks. *IEEE Trans. Med. Imag.*, 2019.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *IEEE Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, pp. 234–241. Springer, 2015.

Valada, A., Mohan, R., and Burgard, W. Self-supervised model adaptation for multimodal semantic segmentation. *Inter. J. Comput. Vis.*, pp. 1–47, 2019.

Vemulapalli, R., Tuzel, O., Liu, M.-Y., and Chellapa, R. Gaussian conditional random field network for semantic segmentation. In *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, pp. 3224–3233, 2016.

Wang, C.-W., Huang, C.-T., Lee, J.-H., Li, C.-H., Chang, S.-W., Siao, M.-J., Lai, T.-M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al. A benchmark for comparison of dental radiography analysis algorithms. *Medical image analysis*, 31:63–76, 2016.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., and Cottrell, G. Understanding convolution for semantic segmentation. In *IEEE Wint. Conf. Appl. Comput. Vis. (WACV)*, pp. 1451–1460. IEEE, 2018.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N. Learning a discriminative feature network for semantic segmentation. In *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, pp. 1857–1866, 2018.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. In *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, pp. 2881–2890, 2017.

Zhao, H., Qi, X., Shen, X., Shi, J., and Jia, J. ICNet for real-time semantic segmentation on high-resolution images. In *European Conf. Comput. Vis. (ECCV)*, pp. 405–420, 2018.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. Conditional random fields as recurrent neural networks. In *IEEE Inter. Conf. Comput. Vis. (ICCV)*, pp. 1529–1537, 2015.

Zhou, W., Berrio, J. S., Worrall, S., and Nebot, E. Automated evaluation of semantic segmentation robustness for autonomous driving. *IEEE Trans. Intell. Transp. Syst.*, 2019.