Graph Neural Blocks on Segmentation

Darwin Saire
Institute of Computing, University of Campinas
Campinas, SP - Brazil
darwin.pilco@ic.unicamp.br

Salvatore Tabbone
LORIA, Université de Lorraine
Nancy, France
antoine.tabbone@univ-lorraine.fr

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 GNNblock). 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 the Cityscape database, achieving close performance with state-of-the-art.

1. Introduction

With the growing increase in devices capable of obtaining photos or videos, the demand to process and obtain more useful information to perform different tasks also increased. The area of computing that works with these images is called Computer Vision. It is in charge of carrying out several applications such as x-ray, agriculture, remote sensing, and autonomous driving.

Currently, computer vision applications require more effective information processing in the intermediate steps inside the pipelines. For instance, a frequently used pipeline step is called semantic segmentation SS. It aims for dense segmentation at the pixel-wise level, i.e., detecting, classifying objects, and adjusting the boundaries of segmented objects. Note, improving the segmentation step has a direct impact on the final result of computer vision applications.

In recent years, Convolutional Neural Networks (CNNs) have led to several improvements in computer vision. Fully convolutional networks (FCN) [18] achieved a significant improvement in the SS in contrast to the traditional SS techniques [30]. However, (i) the low-resolution at the CNNs output and (ii) the loss of spatial precision of objects within the image are still the main problems that affect the segmentation results [4, 17], see Fig. 1.

(a) DUNet (b) OCNet (c) HRNet+OCR

Figure 1: Loss of spatial precision, generally displayed on the segmented objects' boundaries. Red color denotes incorrectly segmented regions.

Different models [9, 15] have tackled these problems and advanced solutions to the output maps' low-resolution problem. For better refinement, previous models [33] post-process the results to enhance them. New architectures [2, 25, 1] propose a global feature extraction with more information and sparse convolution operations [28, 4]. Other methods in SS use hourglass models [24, 20, 3, 13] that add a decoder stage for the reconstruction image (i.e., segmentation) by using deconvolution and unpooling operations. Additionally, to obtain robustness in resolution, models [21, 31, 27] obtain this information as an input. In contrast, models [6, 14] broaden the receptive field of kernels employing Atrous Spatial Pyramid Pooling (ASPP) [4, 5]. Subsequently, to obtain features with information of various resolutions, models [22, 32, 19, 26] modify the size of the features (i.e., features extraction) instead of the convolution kernels [34].

Finally, the self-attention models [8, 11, 29] perform intelligent operations focusing on specific image regions with complex segmentation (i.e., salience). However, the loss of spatial precision, commonly visualized in the contour of segmented objects remains (see Fig. 1). 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 [7, 16] based on GNNs proved to have a greater receptive field by having

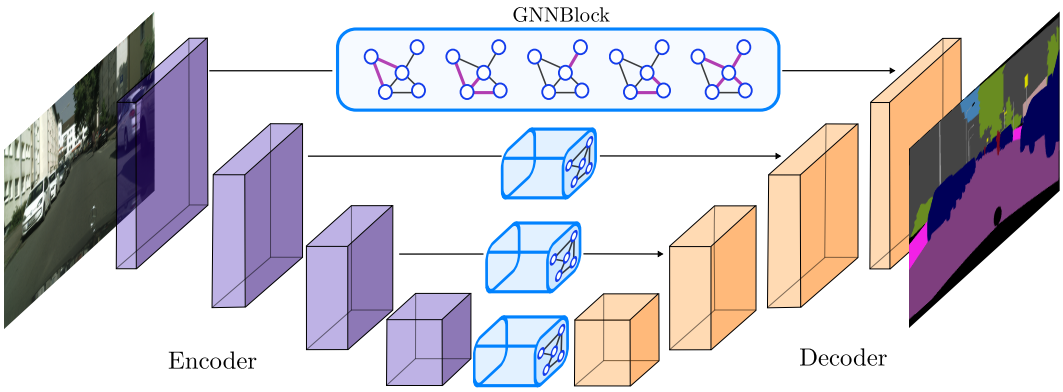

Figure 2: Illustration of our model, which takes an image, extracts the local information in the encoder (left), creates a prediction map, and then in the decoder (right) performs a refinement process using the previous local information with an upsampling and the global information processed by the graph blocks (sky connections).

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. 2). Thus, our neural network aims to produce densely labeled images.

We adopt deconvolution operations, where our input is the multiply of the local features produced at lower levels and the global features extracted from the sky connection blocks (GNNblock). This global information (necessary for slide segmentation) comes at the same level but from the codification stage. In Fig. 2, we can see the process of deconvolution and fusion of local and global information from the right side (i.e., share and merge information). Besides, we show the GNN block (in the sky connections) responsible for processing global information for each level.

The hourglass architecture is similar to previous approaches [12, 23] that use local filters to guide the segmentation. However, they rely on frame-wise feature extraction—which may not provide the best features or consistency thereof. Furthermore, our model integrates global information through blocks embedded within our architecture.

Note, we use the GNN block to provide our model with global contextual information. These blocks are responsible for transforming each level's local information into global ones by creating a function f. This f function is responsible for learning a mapping space to transform the feature

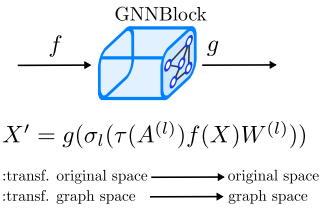

Figure 3: Function mapping, from feature vector into node embedding, g, and vice versa, f.

vector into node embeddings. In other words, the f function converts images feature space into graphs feature space. In this new space (graph space), we find the force of dependence between the regions, i.e., the edges obtain the degree of relationship between regions.

The GNNblock learns to propagate information across all vertices on the graph (node embedding) and through the edges (edge embedding), making it possible to share global information through a set of operations (i.e., graph convolutional operations). Finally, to complete the transfer of information of the convolution state for the deconvolution state performed by the GNNblock, we use another mapping function, g, to convert from the graph space to the original space.

$$X' = g(f(\text{GNNblock}(X))). \quad (1)$$

In this work, the functions f and g are defined by downsampling and upsampling operations (using nearest-neighbor interpolation [10]), respectively. We show this operations in Fig. 3 and Eq. 1.

3. Preliminary Results

Our work is still in progress. Thus, we report our preliminary results, qualitative and quantitative, inside the poster.

References

[1] Naif Alshammari, Samet Akçay, and Toby P Breckon. Multi-task learning for automotive foggy scene understanding via domain adaptation to an illumination-invariant representation. *arXiv*, (arXiv:1909.07697v1), 2019.

[2] Md Amirul Islam, Mrigank Rochan, Neil DB Bruce, and Yang Wang. Gated feedback refinement network for dense image labeling. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3751–3759, 2017.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv*, (arXiv:1706.05587v1), 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[7] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 433–442, 2019.

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3146–3154, 2019.

[9] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. 70:41–65, 2018.

[10] Rafael Gonzalez and Richard Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Upper Saddle River, NJ, USA, 2006.

[11] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S Huang. CCNet: Criss-cross attention for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[12] Varun Jampani, Raghudeep Gadde, and Peter V. Gehler. Video propagation networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[13] Qiangguo Jin, Zhaopeng Meng, Tuan D Pham, Qi Chen, Leyi Wei, and Ran Su. DUNet: A deformable network for retinal vessel segmentation. *Knowledge-Based Systems*, 178:149–162, 2019.

[14] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. CASINet: content-adaptive scale interaction networks for scene parsing. *Neurocomputing*, 419:9–22, 2021.

[15] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019.

[16] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9225–9235, 2018.

[17] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Exploring context with deep structured models for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1352–1366, 2017.

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2016.

[19] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9190–9200, 2019.

[20] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1528, 2015.

[21] Gabriel L Oliveira, Abhinav Valada, Claas Bollen, Wolfram Burgard, and Thomas Brox. Deep learning for human part discovery in images. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1634–1641. IEEE, 2016.

[22] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. ENet: A deep neural network architecture for real-time semantic segmentation. *arXiv*, (arXiv:1606.02147v1), 2016.

[23] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A.Sorkine-Hornung. Learning video object segmentation from static images. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *IEEE International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[25] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal on Computer Vision*, pages 1–47, 2019.

[26] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[27] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1857–1866, 2018.

[28] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *International Conference on Learning Representations (ICLR)*, 2016.

[29] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. OCNet: Object context network for scene parsing. *arXiv*, (arXiv:1809.00916), 2018.

[30] Nida M Zaitoun and Musbah J Aqel. Survey on image segmentation techniques. 65:797–806, 2015.

[31] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. ICNet for real-time semantic segmentation on high-resolution images. In *European Conference on Computer Vision (ECCV)*, pages 405–420, 2018.

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[33] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1529–1537, 2015.

[34] Thomas Ziegler, Manuel Fritsche, Lorenz Kuhn, and Konstantin Donhauser. Efficient smoothing of dilated convolutions for image segmentation. *arXiv*, (arXiv:1903.07992), 2019.