Semantic Segmentation on Image Using Multi-task Hourglass Networks

Anonymous Author(s)
Affiliation
Address
email

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 last convolution layer of the CNNs. In this work, we propose an hourglass model based on the multi-task approach. Consequently, we combine the tasks of edge detection, semantic segmentation, and distance transform. The refinement of the tasks (getting specific information of each task) is obtained in the last layers of the decodification stage. All the tasks share the rest of the information, that is, shared weights. Thus our model is efficient with respect to the number of tasks and memory used. We obtained encouraging preliminary results still in images using Cityspace and Kitti datasets.

1 Introduction and Related Works

Humans possess a remarkable ability to parse images and videos 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. Even we can perform several tasks at the same time by analyzing an image, e.g., semantic segmentation (SS), and instance segmentation(IS). Addressing the SS and IS tasks are not a trivial problem due to the variability, i.e., considerable variations in pose, appearance, viewpoint, illumination, and occlusion throughout the image. Note by improving segmentation task this directly influences several applications such as self-drive vehicles [1, 2], segmentation on X-ray [3], detect crown on dental X-ray [4], brain tumor segmentation [5, 6], and remote sensing [7, 8, 9].

In recent years the fully convolutional networks (FCN) achieve significant improvement, in SS task, by converting fully connected layers into convolutional layers and upscale operations [10]. However, with this approach, new problems have been observed, such as [11, 12]: i) the low-resolution obtained in the output of the CNNs; and ii) the loss of spatial precision of objects within the image. Then, the next stage is dealing with these problems.

Thus, FCN has used with post-processing steps. Conditional Random Fields (CRF) [13] or Gaussian CRF [14] are common post-processing steps but are computationally expensive; consequently, embedding it within a network is a viable solution [11]. Others researchers proposed to obtain a fine adjustment from the bounding boxes [15, 16, 17]. Instead of making an abrupt prediction of the last layer of CNN, the hourglass approach [18, 19, 20, 21] created an up-sampling stage in a controlled manner (deconvolutions and unpooling). Moreover, to arose models that take into account different scales [22]. These models get a full semantic map in low-resolution (coarse prediction map), then refine it with different fusion operations, e.g., fusion cascade [23] and attention blocks [24]. Contrary to multi-scale models, the approaches that use Atrous Spatial Pyramid Pooling (ASSP) [11, 25, 26, 27] modify the filters size instead of the size of the images. This modification is achieved using atrous

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

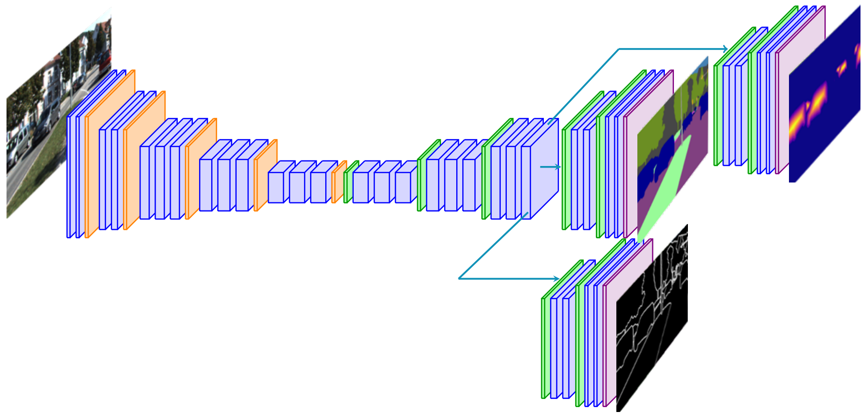

Figure 1: Illustration our multi-task hourglass model, for tasks of edge detection, semantic segmentation, and distance transform (applied in the task of instance segmentation). The blocks blue, orange, and green are convolution, pooling and unpooling operations respectively. Note, the model share weight in the first layers, and the specifics feature for each task are obtained in the last layers.

convolution [11], 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 fit contour segmentation [28].

Although the previous models improved the SS task compared to the traditional works, still needs a greater transfer of information between its different layers. In other words, we need models that take into account more information, i.e., more specific features by using multi-task learning.

2 Multi-task Hourglass Model

The idea of using CNNs as feature extractor is not new, and it has been used widely, achieving better results against traditional methods [29, 30]. Nevertheless, using CNN for SS task also brings new and challenging problems, such as the low-resolution of the feature maps and the loss of spatial precision.

In this work, we focus our model for use multi-task learning, with the target of learning of one task can improve the learning of other tasks. [31, 32, 33]. Hence, task relationships facilitate the transfer of shared knowledge from relevant tasks. For this reason, multi-tasks models only need to learn features for specific tasks [34].

Designing and building a multi-task model for SS is not a trivial task; to achieve this, we need: i) identify which are similar tasks that improve SS; ii) procure independent tasks, unlike multi-task cascade [35]; and iii) merge the semantic information geometric information (distance transform) for the IS task; To carry out this approach, we need tasks that reinforce each other, so we select the tasks of edge detection, SS, and distance transform, which were chosen empirically.

Thus, our multi-task architecture is inspired by SegNet [20] hourglass model due to well-behaving of the prediction maps (better up-sampling domain compared to interpolation) in the codification and decodification stage. Hence, we use convolution and pooling operations in the codification stage in order to extract common features for several tasks in the same image. Then, we produce dense prediction maps at different levels (scale). For the decodification stage, we propose to use deconvolution and unpooling operations where our input is the merge of the features produced at lower levels, and sky connection with information from the same level but from the codification stage. Consequently, we intend to share the information at each level by merge layers. Note, the features necessary to distinguish each task are shared throughout the encoder stage and half of the decoder stage. Thus, the last four layers of the decoder are responsible for learning specific features for each task. Our preliminary results, still on images, showed encouraging results using crops of 300×500 size on the Cityscape [36] and Kitti [37] datasets. This work is still in development. We can see qualitative results in the supplementary material.

References

[1] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, 2016, pp. 2147–2156.

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

[3] J. Bullock, C. Cuesta-Lázaro, and A. Quera-Bofarull, "XNet: a convolutional neural network (CNN) implementation for medical x-ray image segmentation suitable for small datasets," vol. 10953. International Society for Optics and Photonics, 2019, p. 109531Z.

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

[5] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

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

[7] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv*, no. arXiv:1606.02585v1, 2016.

[8] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, 2017.

[9] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation." Springer, 2019, pp. 388–401.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, pp. 1–1, 2016.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[12] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Exploring context with deep structured models for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1352–1366, 2017.

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

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

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE Inter. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2961–2969.

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

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

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation." Springer, 2015, pp. 234–241.

[19] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *IEEE Inter. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1520–1528.

[20] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

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

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

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

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

[25] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv*, no. arXiv:1706.05587v1, 2017.

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

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

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

[29] M. Thoma, “A survey of semantic segmentation,” *arXiv*, no. arXiv:1602.06541v2, 2016.

[30] Y. He, W. Chiu, M. Keuper, M. Fritz, and S. I. Campus, “STD2P: RGBD semantic segmentation using spatio-temporal data driven pooling,” in *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, 2017.

[31] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *arXiv*, no. 1707.08114v2, 2017.

[32] S. Ruder, “An overview of multi-task learning in deep neural networks,” *arXiv*, no. 1706.05098v1, 2017.

[33] K.-H. Thung and C.-Y. Wee, “A brief review on multi-task learning,” *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29 705–29 725, Nov 2018.

[34] M. Long, Z. Cao, J. Wang, and S. Y. Philip, “Learning multiple tasks with multilinear relationship networks,” in *Adv. Neural Inf. Process. Sys. (NeurIPS)*, 2017, pp. 1594–1603.

[35] J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” in *IEEE Inter. Conf. Comput. Vis., Pattern Recog. (CVPR)*, 2016, pp. 3150–3158.

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

[37] H. Alhaija, S. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, “Augmented reality meets computer vision: Efficient data generation for urban driving scenes,” *Inter. J. Comput. Vis.*, 2018.