Review for Object-Contextual Representations for Semantic Segmentation

Irem Arpag

May 24, 2021

As the process of linking each pixel in an image to a class label, semantic segmentation is a fundamental issue in computer vision but especially crucial in autonomous driving since it is important for this discipline to understand the context in the environment in which they are operation, which demonstrate the significance of "contextual aggregation", one of the study field of the domain and subject of this article. The authors represent a vigorous approach called object-contextual representation (OCR) based on the idea that the class label assigned to one pixel is the category of the object that the pixel belongs to. The model is composed of the three stage: Firstly, the contextual pixels are divided into a set of object regions; each of which correspond to a class and learn under the supervision of the ground-truth segmentation; Secondly, object region representation is calculated by aggregating the representation of the pixels in each object region and finally, after augmenting the representation of each pixel with the object contextual representation (OCR), the relation between each pixel and each object region is computed.

The aim of the study that is augmenting the representation of one pixel by using the representation of object region of the corresponding class is verified with various empirical studies that demonstrate high quality segmentation performance of the model when the ground-truth object region is given. The approach, through combining the contextual pixels into object regions, differs from previous models which consider contextual pixels separately, and that brings competitive achievement to the model via evaluating most lead semantic segmentation benchmarks such as with a 84.5 in percentage performance on Cityspace test. Also extending the experiments to Panoptic Segmentation task which joins instance segmentation and semantic segmentation confirm the generalization capacity of the model.

During the experiment, while researchers' final model "HRNet + OCR + SegFix" performs 84.5 in percentage on the Cityspaces leaderboard, a following study [1] by combining their "HRNet + OCR" and a new hierarchical multi-scale attention mechanism achieves a novel state-of-art performance with 85.4 in percentage on Cityspace leaderboard and taking the first rank. By injecting the hierarchical structure knowledge of human parts, another recent study CNIF [2] accomplishes the best performance with 56.93 in percentage rather than OCR's 56.65 in percentage carrying out with simple baseline. It is thought, such hierarchical structure knowledge can provide benefit for the latter model and it can be the interest of future work.

As empirical analysis, the researchers study the influence of two attributes of the model that are object region supervision and pixel-region relations on performance, the superior results confirm the importance of both. For experiments on semantic segmentation, they use different challenging datasets of the domain; Cityspaces, ADE20K, LIP, PASCAL-Context and COCO-Stuff. First, they compare their OCR model with well-known, multi-scale context schemas, by using the same training/testing settings for fairness, OCR performs all well. Comparison of OCR with different relational context schemes confirms good results, in spite of its smaller complexity than the existing models. As another study, they compare efficiency by measuring GPU memory, computation complexity (number of FLOPs) and inference time, OCR is considered to be the best in terms of memory, GFLOP's and running time. The last comparison is performed with SOTA approaches based on baseline as simple and advance on three benchmarks. For example, OCR performs best results of 81.8 in percentage on Cityspace dataset based on simple baseline which is even better than most of the advance baseline methods. Performing Panoptic Segmentation experiments are another challenging motivation confirming the generalization ability, for example, Panoptic-FPN + OCR achieves 44.2 in percentage on COCO val.

By constructing simple but effective structure and evaluating it with comprehensive empirical studies, the research achieves competitive performance.

REFERENCES

[1] Tao, A. Sapra, K. Catanzero, B: Hierarchical Multi-Scale Attention for Semantic Segmentation arXiv: 2005.10821 (2020)

[2]Wang, W, Zhang, Z, Qi, Shen, J. Pony, V. Shao, Learning Compositional Neural Information Fusion for Human Parsing. In: ICCV (2019)