Opinion
Volume 15:9, 2022

Journal of Computer Science & Systems Biology

ISSN: 0974-7230 Open Access

Semantic Categorization of Aerial Imagery Using a Transformer Framework and Context Variables

Mohite Weber*

Department of Information Science, University of Cairo, Giza Governorate 12613, Egypt

Introduction

In the field of computer vision, semantic image segmentation, or assigning labels to each pixel, has received a lot of attention. Its primary applications include medical imaging, self-driving cars, satellite image segmentation and point cloud scene segmentation. Because of the rapid development of smart technologies, the use of Unmanned Aerial Vehicles (UAVs), also known as drones, has grown significantly. UAVs can take photos even in remote areas where humans would struggle. Aerial images captured by UAV devices contain a wealth of information that can be used for a variety of tasks, including traffic density estimation, building extraction and flood monitoring.

The dense prediction of objects in street scene images at the pixel level is useful for land use monitoring. Convolution-based architectures are widely used in computer vision for segmentation. The majority of segmentation models use an FCN-based encoder network and various decoder architectures. The FCN-based encoder backbone, on the other hand, produces coarse predictions and the receptive field becomes constant after a certain threshold limit. Convolutional Neural Networks (CNNs) maintain local context properly but fail to capture global contexts due to local receptive fields. Constructed a segmentation dataset and designed a segmentation model for UAV scene segmentation, which yielded satisfactory results for UAV image segmentation. For UAV urban scene segmentation, Long Short-Term Memory (LSTM) or optical flow modules were used.

Description

UAV images typically have very complex backgrounds with many variations in object appearance and scale, which makes semantic segmentation difficult. Even though existing segmentation frameworks, including those specifically designed for UAV scene segmentation, produce satisfactory results, their feature extractor struggles to capture the inherent features of aerial images. Segmenting minority classes, particularly small objects such as humans, which have the fewest pixels in the entire image, becomes difficult. This research aims to achieve precise semantic segmentation of UAV street scene images [1-3]. Inspired by the transformer-based design paradigms for computer vision tasks, an encoder-decoder framework is proposed to address such issues in this work. It includes a self-attention-based encoder network that captures long-distance information in UAV images by maintaining global receptive fields. This enables the network to keep global contextual details.

To model low-level contextual details while taking advantage of CNNs' advantage in capturing local information, a convolution-based element, namely the Token Spatial Information Fusion (TSIF) module, is introduced in the encoder network. Capturing local context information allows the network to maintain the proper shape and size of the segmented objects. The powerful self-attention-based encoder network's output is semantically very rich, with both global and local contextual details. For final pixel-level predictions, a decoder network is proposed that processes the encoder network's output information. Frameworks based on transformers have transformed the entire field of Natural Language Processing (NLP). With its enormous success in NLP, researchers have begun to investigate its application for computer vision tasks. The first work in this direction was Vision Transformer (ViT), which used a fully transformer-based design for image classification. Following that, many researchers worked to improve classification accuracy [4,5]. DeiT implemented a teacher-student framework to facilitate network learning.

Conclusion

For UAV images, we used an encoder-decoder framework that captures the inherent features in the global and local context. Because the self-attention-based encoder network captures long-range information, the global context of two similar objects in the image is well preserved. The convolution-based TSIF module in the network fuses local contextual details. This allows the network to segment neighbouring pixels while maintaining the objects' proper shape and size. For final pixel-level predictions, the decoder network employs semantically rich feature representations from the encoder network. It generates smooth segmentation between two classes with well-preserved boundary information. For the sensitivity test, we conducted a series of ablation studies.

References

1. Cao, Bin, Weizheng Zhang, Xuesong Wang and Jianwei Zhao, et al. "A memetic algorithm based on two Arch2 for multi-depot heterogeneous-vehicle capacitated arc routing problem." *Swarm Evol Comput* 63 (2021): 100864.
2. Shi, Yifei, Xin Xu, Junhua Xi and Xiaochang Hu, et al. "Learning to detect 3D symmetry from single-view RGB-D images with weak supervision." *IEEE Trans Pattern Anal Mach Intell* (2022).
3. Yin, Ming, Yangyang Zhu, Guofu Yin and Guoqiang Fu, et al. "Deep Feature Interaction Network for Point Cloud Registration, With Applications to Optical Measurement of Blade Profiles." *IEEE Trans Ind Inform* (2022).
4. Zhou, Wujie, Ying Lv, Jingsheng Lei and Lu Yu, et al. "Global and local-contrast guides content-aware fusion for RGB-D saliency prediction." *IEEE Trans Syst Man Cybern Syst* 51(2019): 3641-3649.
5. Sheng, Shuran, Peng Chen, Zhimin Chen and Lenan Wu, et al. "Deep reinforcement learning-based task scheduling in iot edge computing." *Sensors* 21 (2021): 1666.

***Address for Correspondence:** Mohite Weber, Department of Information Science, University of Cairo, Giza Governorate 12613, Egypt, E-mail: MohiteWeber2@gmail.com*

Received: 07 September, 2022, Manuscript No. jcsb-22-83695; **Editor assigned:** 09 September, 2022, Pre QC No. P-83695; **Reviewed:** 21 September, 2022, QC No. Q-83695; **Revised:** 26 September, 2022, Manuscript No. R-83695; **Published:** 03 October, 2022, DOI: 10.37421/0974-7230.2022.15.432

How to cite this article: Weber, Mohite. "Semantic Categorization of Aerial Imagery Using a Transformer Framework and Context Variables." J Comput Sci Syst Biol 15 (2022):432.