doi:10.1017/S1431927621010539
Microsc. Microanal. 27 (Suppl 1), 2021

CEM500K – A large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning.

Ryan Conrad and Kedar Narayan

Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States

Recent advances in volume electron microscopy (vEM) have resulted in the production of massive amounts of cellular EM image data, yet only a minuscule fraction of that data is annotated or segmented. Much work is being done to apply deep learning (DL) approaches to segment vEM data, but a key hurdle remains generalization: models trained on a specific type of data (say mitochondria from 3-D reconstructions of mouse hippocampus) perform poorly when confronted with previously unseen contexts (say mitochondria from the same 3-D reconstruction but within the retina instead) [1][2]. In DL research, a successful paradigm to correct this deficiency includes pre-training a neural network on a large dataset followed by transfer learning to a downstream task with a much smaller dataset. This approach yields better performing models that train more quickly and require fewer labeled examples. Unsupervised algorithms are now able to leverage unlabeled image data for pre-training, thereby overcoming the constraint of image annotation, but now the need for an appropriately relevant, large, heterogenous and information-rich dataset emerges.

To this end, we present an unlabeled dataset of ~500,000 two dimensional Cellular EM images (CEM500K), curated from nearly 600 three dimensional and 10,000 two dimensional images from >100 unrelated imaging projects [3]. We show that a DL model pre-trained on CEM500K not only demonstrates robustness to rotations, blur, noise, pixel spacing, and contrast and brightness adjustments, but also spontaneously learns features that are selective for organelles. We evaluate the effectiveness of transfer learning from the pre-trained model on six publicly available benchmark cellular EM segmentation tasks and report state-of-the-art results on each. Thus pre-training on CEM500K in combination with transfer learning is likely to be a powerful tool for the segmentation of EM datasets irrespective of the actual DL model used. We release the CEM500K dataset and pre-trained models (https://www.ebi.ac.uk/pdbe/emdb/empiar/entry/10592/), as well as code for the curation pipeline (https://github.com/volume-em/cellemnet) for use and expansion by the EM community.

References

[1] J. Buhmann*et al.*, "Automatic Detection of Synaptic Partners in a Whole-Brain Drosophila EM Dataset," *bioRxiv*, p. 2019.12.12.874172, Mar. 2019.

[2] J. W. Lichtman, H. Pfister, and N. Shavit, "The big data challenges of connectomics," *Nat. Neurosci.*, vol. 17, no. 11, pp. 1448–1454, Oct. 2014.
[3] R. Conrad and K. Narayan, "CEM500K – A large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning," *bioRxiv*. bioRxiv, p. 2020.12.11.421792, 11-Dec-2020.

https://doi.org/10.1017/S1431927621010539 Published online by Cambridge University Press