Impact of Sampling and Augmentation on Generalization Accuracy of Microscopy Image Segmentation Methods

Michael Majurski, Petre Manescu, Joe Chalfoun, Peter Bajcsy, and Mary Brady
National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899
POC: peter.bajcsy@nist.gov

1. Motivation

Terascale microscopy imaging requires automated software-based measurements of objects found via segmentation. Convolutional Neural Network (CNN) models have become a popular and successful supervised segmentation method incorporating the domain expert knowledge via annotations. To alleviate the annotation effort, sampling and augmentation methods have been leveraged to generate the large numbers of representative examples required for CNN training. As CNNs frequently report very high accuracies, there is a need to understand the impact of sampling/augmentation methods and their parameters on the generalization accuracy of image segmentation to improve our confidence in the reported accuracy. The confidence problem is illustrated in Figure 1.

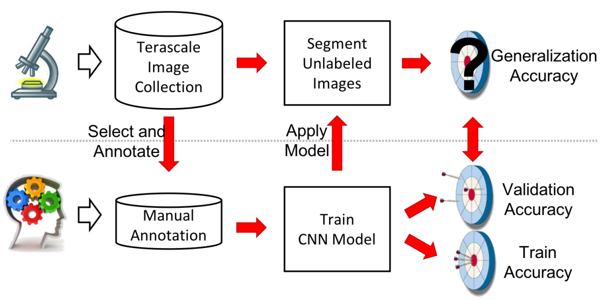

Figure 1 The confidence problem in segmentation generalization accuracy for terascale image collections

We approach the problem of estimating our confidence in CNN-based segmentation accuracy by performing several quantitative evaluations varying sampling and augmentation methods and their parameters over large image collections. The collections were acquired by time-lapse microscopy imaging of cell colonies. The ground truth segmentation was obtained by (a) using a special stain and a fluorescent imaging channel, and (b) segmenting generated high contrast images via thresholding.

2. Methodology

To deliver a sufficiently large number of representative samples for training complex CNN models with millions of parameters, one must analyze (1) the sampling method, (2) the sampling size (count), (3) the augmentation method, and (4) the augmentation parameters. Random sampling was chosen since it is the most frequently used method in the literature. Sample size range was selected based on its statistical relationship to estimation confidence spanning a 95 % confidence interval.

We classified the widely-used augmentation methods based on their types of transformations. Each image augmentation consisted of applying parametrized label-preserving transformations (e.g., affine, reflection, noise) to the annotated examples. To simplify the augmentation parametrization, we used at most one parameter per augmentation transformation, deriving the parameter range from the image data (transformation severity).

We used the Dice metric[1] for evaluating segmentation accuracy and focused primarily on the improvement in our accuracy confidence due to augmentation over the accuracy confidence provided by a selected sampling size.

3. Conclusions

We quantified the impact of sampling and augmentation models and their parametrization on the validation and generalization accuracies of CNN-based segmentation as the generalization error gaps (see the deltas in Figure 2) over 60 configurations of sampling size, augmentation model + parameter, and image object. We observed that training the CNN-based segmentation using rotation, reflection, and jitter lowered the generalization error gap the most (improved our accuracy confidence). We hypothesize that these quantitative results indicate that the augmentation configurations are closely mimicking the imaging variations.

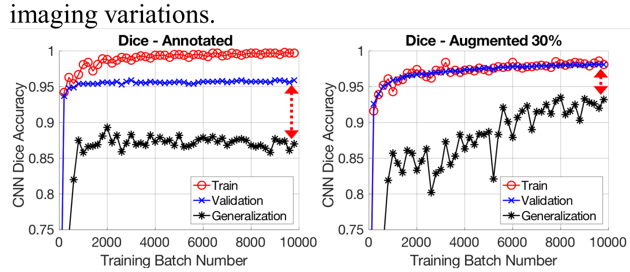

Figure 2: Constant vs. decreasing gap between train, validation, and generalization CNN segmentation accuracy.

[1] Measures spatial overlap between two segmentations. Dice, L. (1945) Measures of the amount of ecologic association between species. Ecology 26, 297–302