Quality control of BIOP/LNMC Cellpose model

Version Control

Version	Author	Description	Validation	Date
0.1	JP	First version		20220804
0.2	JP	Added pictures, conclusion and remarks		20220825

About This document

This report aims to present and provide quality control for results of a cell detection model ("V2") for Instance segmentation. Results were computed in Jupyter Notebook and benchmarked on 6 annotated slices of sagittal cuts of juvenile mouse brains. The benchmark for quality control is "cyto2", a Cellpose pre-trained model.

Methodology

Three model parameters are compared between 'cyto2' and 'V2', as advised in arXiv:2206.01653:

- **Dice Similarity Coefficient** (DSC) : an overlap based metric (measures overlap between ground truth, annotations, and learned structure). This metric is greatly influenced by the cell size distribution but indifferent to shape. As the DSC approaches 1, the annotation converges to ground truth. Recommended for segmentation problems. A visual representation of the DSC metric is shown in Figure 1.

Rajouter information de quand la metrique est bonne, schémas,

- **Mask Intersection over Union** (IoU) : another overlap based metric. There are multiple different IoU metrics (Mask IoU, Boundary IoU and Intersection over Reference).
  As the IoU approaches 1, the annotation converges to ground truth. Advantages and disadvantages of each metric are list in E.4 (page 54) of [1] . visual representation of the IoU metric is shown in Figure 1.

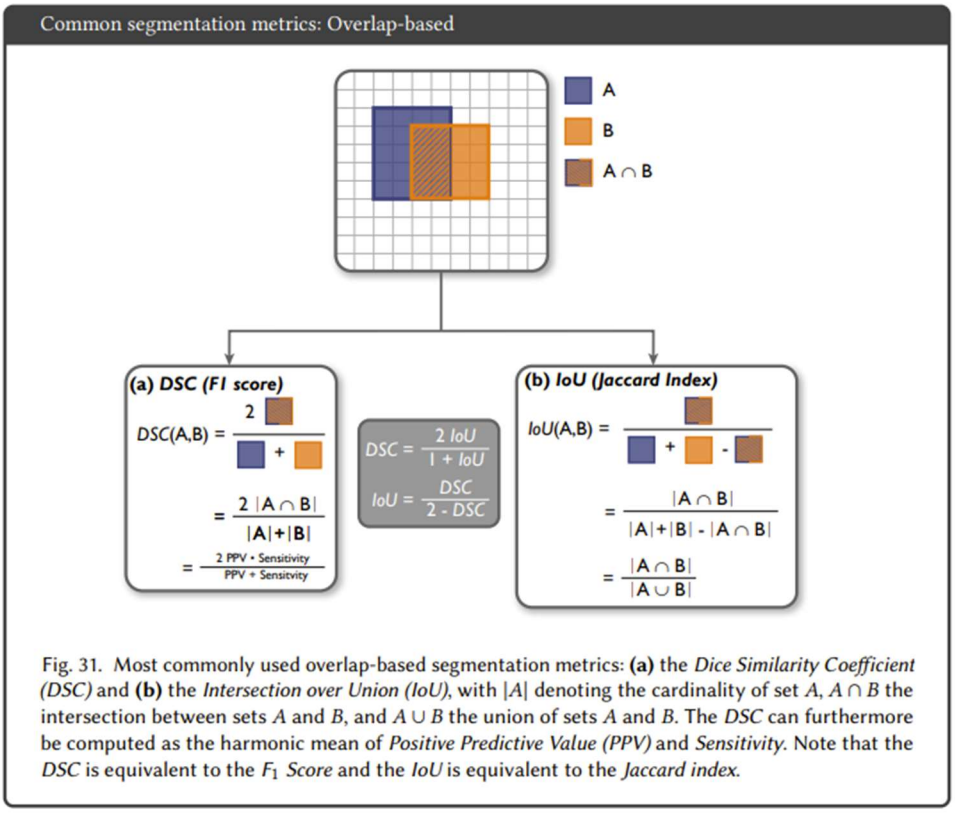


Fig. 31. Most commonly used overlap-based segmentation metrics: **(a)** the *Dice Similarity Coefficient (DSC)* and **(b)** the *Intersection over Union (IoU)*, with $|A|$ denoting the cardinality of set A, $A \cap B$ the intersection between sets A and B, and $A \cup B$ the union of sets A and B. The *DSC* can furthermore be computed as the harmonic mean of *Positive Predictive Value (PPV)* and *Sensitivity*. Note that the *DSC* is equivalent to the F_1 *Score* and the *IoU* is equivalent to the *Jaccard index*.

Figure 1. DSC and IoU representation reproduced from [1]

- **Panoptic quality** (PQ) : a counting metric (based on the count of identified structures). As the metric approaches 1, false negatives and false positives tend to 0 and true positives approach ground truth. This metric evaluates both segmentation (like IoU and DSC) and detection (analysis of false and true positives/negatives).
- **Accuracy :** Metric common used in classification tasks to quantify how much a model is correct (with respect to ground truth) :

$$Accuracy = \frac{True\ positives}{True\ positives + False\ positives + False\ negatives}$$

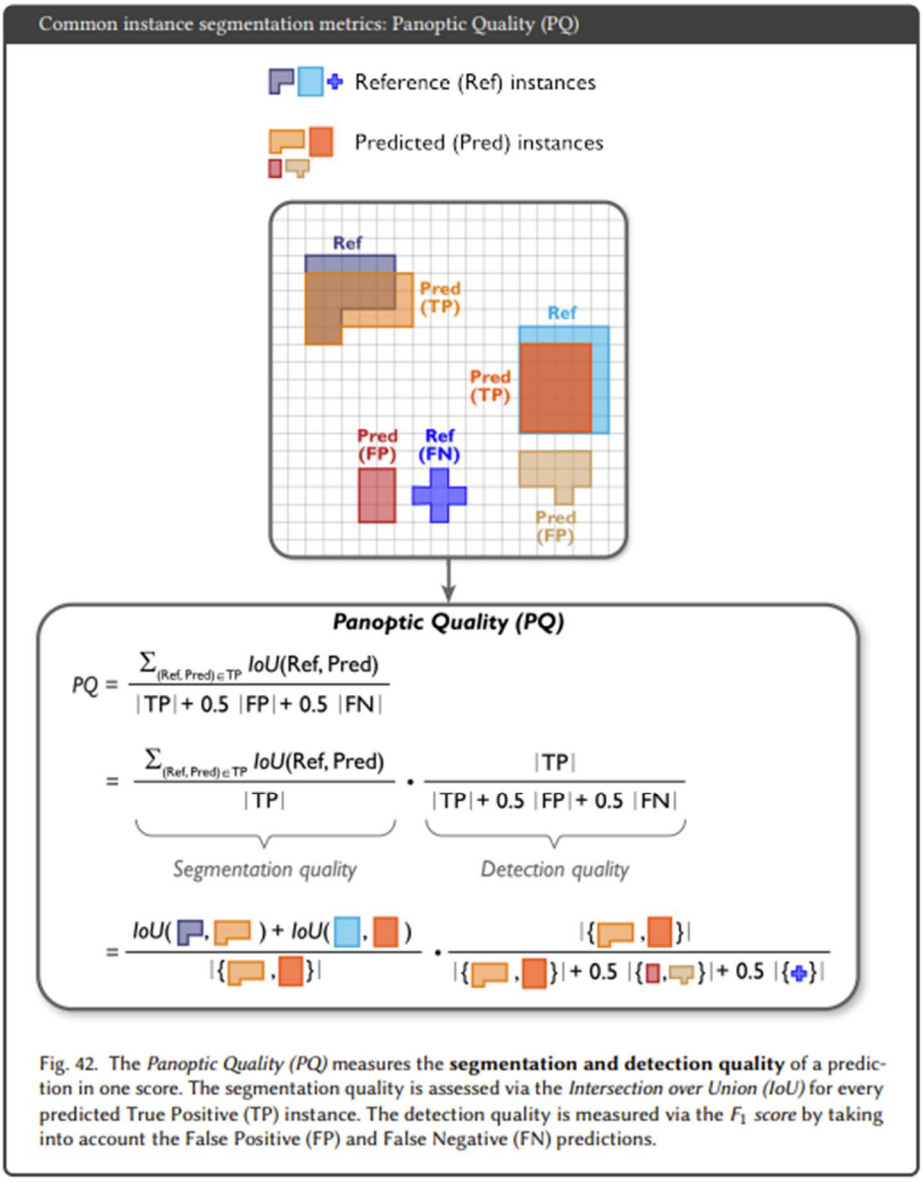


Fig. 42. The *Panoptic Quality (PQ)* measures the **segmentation and detection quality** of a prediction in one score. The segmentation quality is assessed via the *Intersection over Union (IoU)* for every predicted True Positive (TP) instance. The detection quality is measured via the F_1 *score* by taking into account the False Positive (FP) and False Negative (FN) predictions.

Figure 2. PQ representation reproduced from [1]

Unpaired Welch T-tests[1] are applied to) results from ‘cyto2’ and ‘V2’ mean predictions (n=6) using scipy.stats module default configurations. Results are plotted and p-values are reported with asterisk convention (e.g. ‘ * ’ corresponds to p-value <= 0.1, ‘ *** ‘ corresponds to p-value <= 0.001).

Results

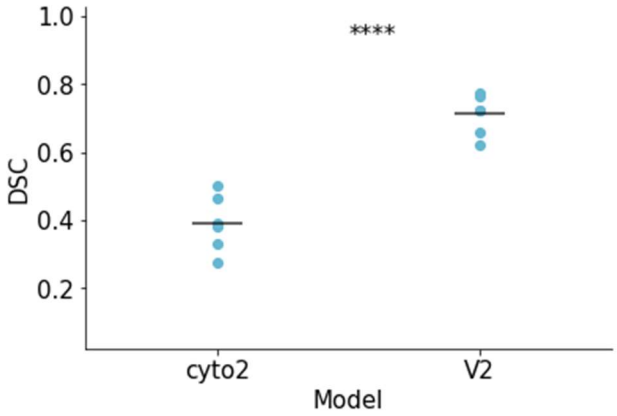
DSC: (p-value = 3.21e-5)

[1] Unpaired Welch tests assume that the means being compared are normally distributed (asymptotically true) and that the variances are not equal (but also normally distributed)

Remarks:

cyto2: 0.2677 ± 0.06167

V2: 0.5513 ± 0.04523

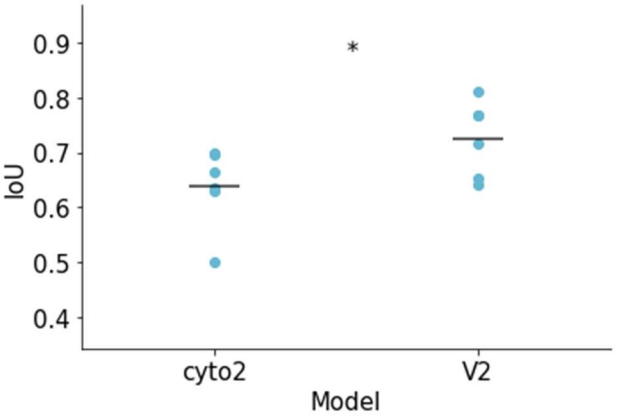


IoU: (p-value =0.059)

Remarks:

cyto2: 0.6383 ± 0.0677

V2: 0.726 ± 0.06262

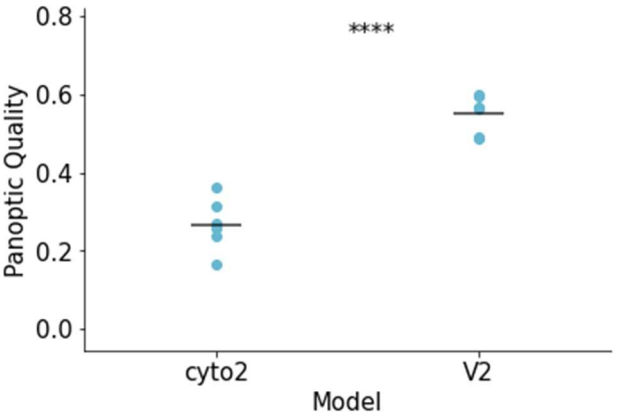


PQ : (p-value =1.47e-5)

Remarks :

cyto2 : 0.3908 ± 0.07681

V2 : 0.7119 ± 0.05417

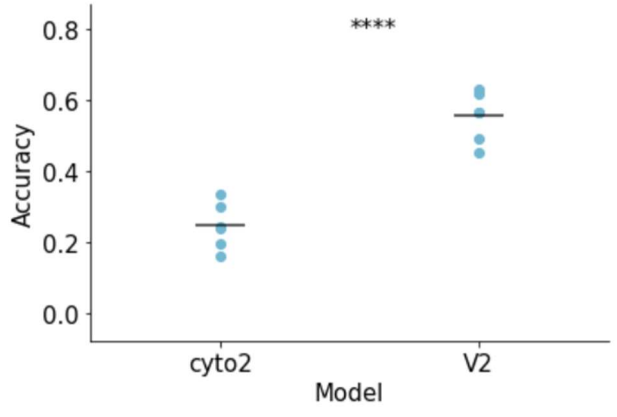


Accuracy: (p-value = 1.36e-05)

Remarks:

cyto2 : 0.2457 ± 0.05961

V2 : 0.5554 ± 0.06426



Conclusion

We observe a significant improvement in DSC and PQ metrics for the V2 classifier in comparison to cyto2. IoU improvement is below a threshold of p = 0.05 and cannot be considered statistically significant. A possible explanation for the generalized increase in all metrics can be an improvement in the segmentation task, as V2 seems to be less biased to annotating bigger cells than cyto2. Nevertheless, statistical analysis with (n = 5) is at the low end limit of applicability and analysis with more samples would be preferred.

Bibliography / Reference Paper

[1] Maier-Hein, L., Reinke, A., Christodoulou, E., Glocker, B., Godau, P., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M., Wiesenfarth, M., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A., Rädsch, T., Tizabi, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Bankhead, P., Benis, A., Cardoso, M., Cheplygina, V., Cimini, B., Collins, G., Farahani, K., Ginneken, B., Hashimoto, D., Hoffman, M., Huisman, M., Jannin, P., Kahn, C., Karargyris, A., Karthikesalingam, A., Kenngott, H., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B., Litjens, G., Madani, A., Maier-Hein, K., Martel, A., Mattson, P., Meijering, E., Menze, B., Moher, D., Moons, K., Müller, H., Nickel, F., Nichyporuk, B., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Gutiérrez, C., Shetty, S., Smeden, M., Sudre, C., Summers, R., Taha, A., Tsaftaris, S., Van Calster, B., Varoquaux, G., & Jäger, P.. (2022). Metrics reloaded: Pitfalls and recommendations for image analysis validation.

https://doi.org/10.48550/arXiv.2206.01653