Guided Proofreading of Automatic Segmentations for Connectomics

Supplemental Material

Anonymous CVPR submission

Paper ID 0947

1. Classifier

1.1. Architecture

We explored different architectures for the convolutional neural network (CNN) for split error detection. In table 1 we compare traditional CNN architectures versus residual networks [1]. The traditional architecture with dropout regularization generalized better than residual networks on unseen testing data.

	Traditional Network		**Residual Network**	
Conv. Layers	2	4	5	13
Dropout Reg.	y	y	y	n
Cost [m]	27.5	383	5080	1094
Test. Acc.	0.925	0.94	0.93	0.90
Prec./Recall	0.93/0.93	0.94/0.94	0.7/0.53	0.74/0.66
F1 Score	0.93	0.94 *	0.39	0.64

Table 1: Traditional CNN Architecture versus Residual Network Architecture [1]. All configurations are compared using the same parameters. Our final choice (indicated by *) trains relatively fast and performs better.

1.2. Training Parameters

We performed a limited brute force parameter search to tune the split error classifier (Tab. 2). This resulted in 3240 different CNN configurations which were evaluated on 10% of our training data. Learning rate and momentum ranges are defined linearly across 2000 epochs.

1.3. Automatic Method Threshold p_t

For automatic selection, we observed a threshold $p_t = 0.95$ as stable when evaluating on previously unseen testing data (Mouse S1 AC3 Open Connectome Project dataset). This means, that automatic selection stops once all borders with $p_t \geq 0.95$ were proofread. Fig. **??** shows split error classification on three randomly selected subvolumes ($700 \times 700 \times 2$ voxels) of AC3. In all cases, the threshold $p_t = 0.95$ reduces VI.

Parameter	**Search Space**
Filter size:	**3x3**, 5x5, 9x9, 13x13
No. Filters 1:	32, 48, **64**
No. Filters 2-4:	32, **48**, 64
Dense units:	256, **512**
Learning rate:	0.00001, 0.0001, 0.001, 0.01, **0.03-0.00001**
Momentum:	0.9, 0.95, **0.9-0.999**
Mini-Batchsize:	10, 100, **128**

Table 2: Brute force parameter search for the split error classifier. The final parameters are highlighted.

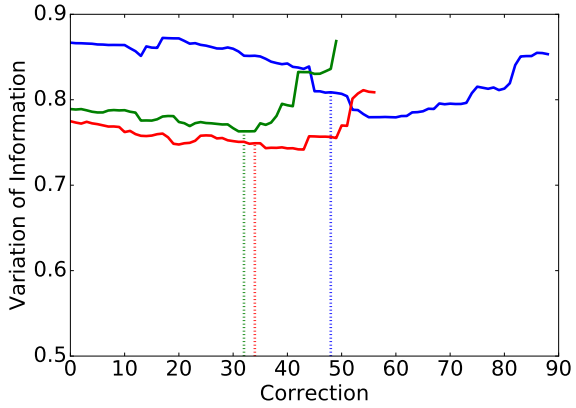

Figure 1: Observations of probability thresholds p_t during automatic selection on three different subvolumes of previously unseen testing data. The dashed lines show when $p_t = 0.95$ is reached.

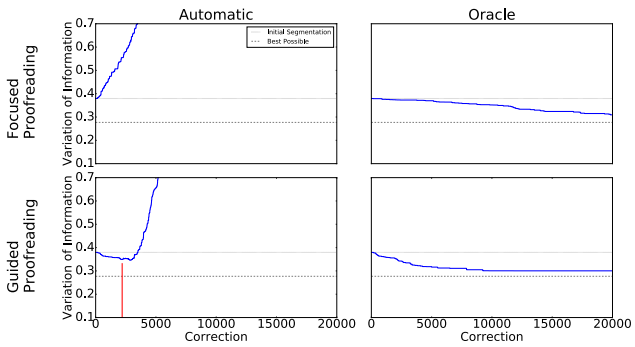

Figure 2: Performance comparison of Plaza's focused proofreading and our guided proofreading on the L. Cylinder dataset. All measurements are reported as median VI, the lower the better. We compare automatic selection with threshold and the selection oracle for accepting or rejecting corrections using each method. In all cases, guided proofreading yields better results with fewer corrections.

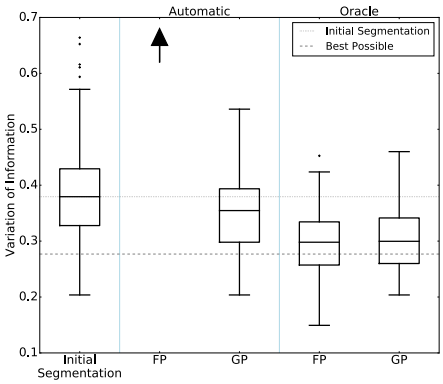

Figure 3: VI distributions of guided proofreading (GP) and focused proofreading (FP) output across slices of the L. Cylinder dataset, with different error correction approaches. The variation resulting from performance of FP with automatic selection is $7.8\times$ higher than GP (as indicated by the arrow), with median VI of 2.75 and $SD = 0.789$.

1.4. Limitations

Guided proofreading works on 2D image sections. This enables error correction without a computationally expensive alignment process. However, the output requires an additional (block-)merging step prior to 3D analysis.

2. L. Cylinder Results

3. CREMI Results

4. Forced Choice User Experiment

4.1. Recruitment and Participation

4.2. Example Classifications

Where did users make a mistake?

4.3. Subjective Responses

NASA TLX ANOVA analysis

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1