Task configuration impacts annotation quality and model training performance in crowdsourced image segmentation - Supplementary Materials

Benjamin Bauchwitz
Duke University Department of Computer Science
308 Research Dr, Durham, NC 27710
benjamin.bauchwitz@duke.edu

Mary Cummings
George Mason University College of Engineering and Computing
4400 University Dr, Fairfax, VA 22030
cummings@gmu.edu

1. Annotation Quality Distribution

Fig. 1 compares the distribution of mask quality for each set of experimental conditions studied. Mask quality is assessed from the IoU of the crowdsourced mask compared to the curated ground truth mask from the reference dataset.

The upper left plot compares compensation schemes, with annotators in group A paid per image and annotators in group B paid per annotated object in the image, with the same expected payout. Annotations with high IoU compared to the ground truth were more common when annotators were paid per object (light blue bars).

The upper right plot compares time limits, with annotators in group A having no time limit and annotators in group C having a 3-minute time limit. There was no apparent difference in the distribution of IoU scores of annotations produced by the two groups.

The bottom left plot compares task complexity, with annotators in group D1 only labeling a single object class per image and annotators in group D2 labeling at least two different object classes per image. Annotations with high IoU compared to the ground truth were more common when annotators had to label multiple object classes per image (light purple bars).

The bottom right plot compares annotation drawing tool, with annotators in group A having access to all drawing tools, and annotators in groups E, F, and G being restricted to polygon, draggable outline, and paintbrush drawing tools, respectively. Annotations with high IoU compared to the ground truth were most common when annotators were restricted to the paintbrush tool (lightest green bars) and were least common when annotators were restricted to the draggable outline tool (second lightest green bar).

2. Model Training Hyperparameters

We conducted an evaluation of model hyperparameters to determine if different training configurations would influence robustness to different error types. We evaluated batch sizes from 4 to 16, learning rates from 0.01 to 0.1, momentum from 0.85 to 0.99, and weight decays from 0.0001 to 0.005. Comparisons are displayed in Fig. 2.

Response to hyperparameters was generally consistent across annotation sources. A batch size of 16 yielded modest improvements in performance compared to a batch size of 8, though the disparity was minimal beyond 160,000 training iterations. The notable exception was for group A (images where annotators had free choice of drawing tools, no time limit, and were paid per image), where early learning was delayed. In this case, mean IoU on the test set significantly lagged the other dataset versions at 80,000 iterations, though training caught up by 320,000 iterations.

Learning rate, momentum, and weight decay all had consistent effects across annotation sources. We found that a learning rate of 0.02, a momentum of 0.9, and a weight decay of 0.0005 were generally optimal for batch sizes between 8 and 16. The only exceptions were for group F (images annotated with the draggable outline tool only), where increasing the learning rate to 0.05 and decreasing the momentum to 0.85 produced marginal ($< 1\%$) improvements in performance.

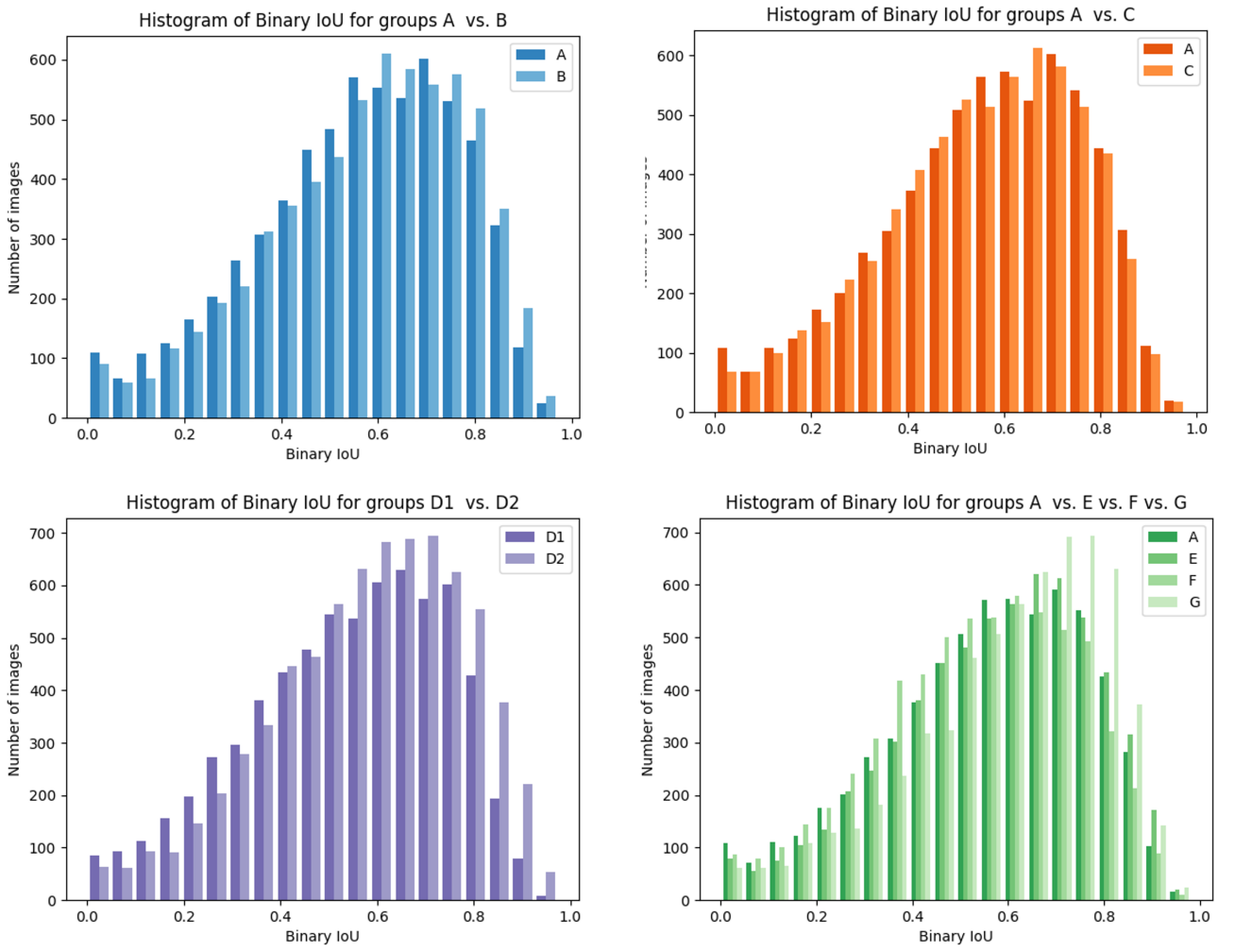
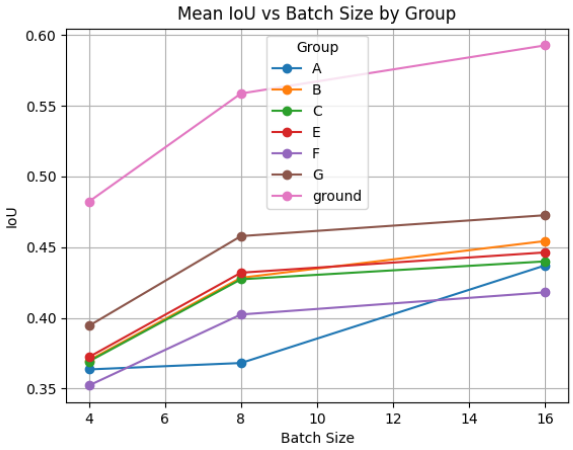
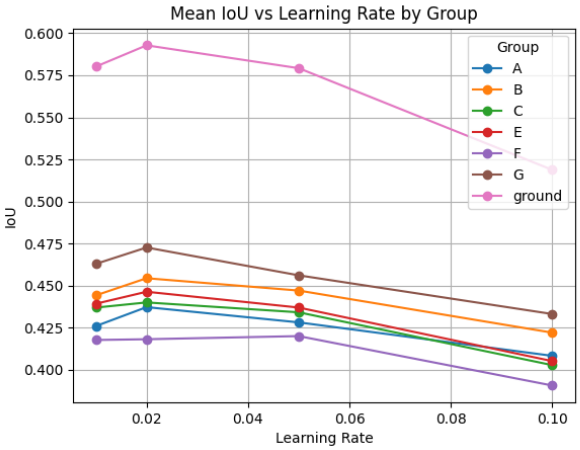


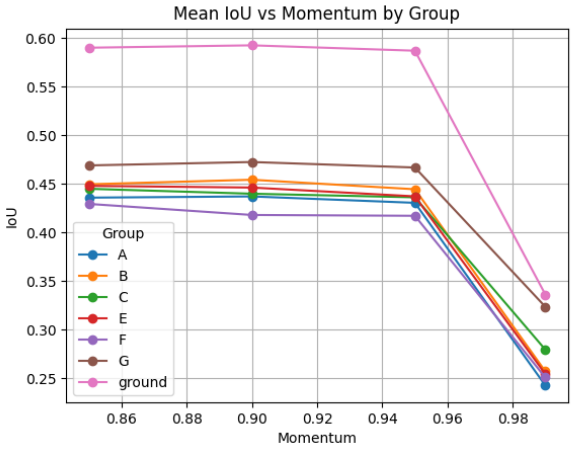
Figure 1. Comparisons of distribution of image-level IoU scores for each task configuration. Top left: annotators paid per image (A) vs. paid per object (B). Top right: annotators given no time limit (A) compared to 3-minute time limit (B). Bottom left: annotators label a single object class at a time (D1) vs. label all potential object classes at once (D2). Bottom right: annotators use any drawing tool (A) vs. polygon tool only (E) vs. draggable outline tool only (F) vs. paintbrush tool only (G).



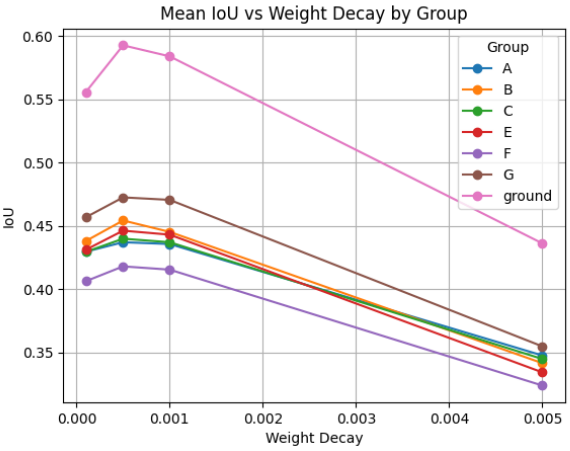
(a) Image 1



(b) Image 2



(c) Image 3



(d) Image 4

Figure 2. Evaluation of hyperparameter space. Top left: Batch size did not systematically influence model performance for any annotation sources except for group A, where learning was delayed at a batch size of 8. Top right: Learning rate had a similar effect on learning across annotation sources, with learning rate of 0.02 leading to optimal or near optimal performance. Bottom left: Momentum values between 0.85 and 0.95 did not significantly affect training, with values of 0.90 producing marginally better learning outcomes. Bottom right: Weight decay of 0.0005 led to superior training performance across annotation sources.