SAM: The Sensitivity of Attribution Methods to Hyperparameters

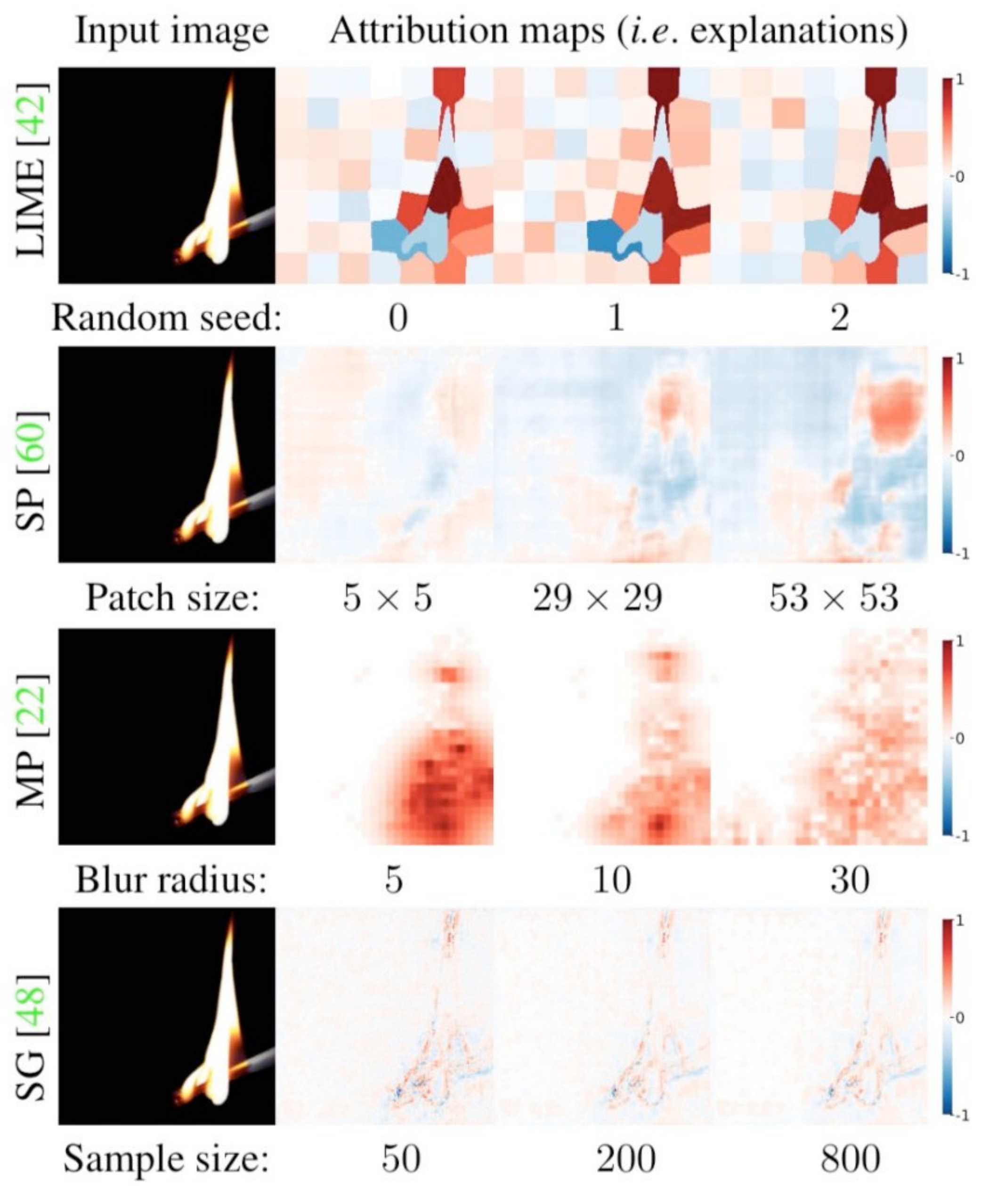
Naman Bansal* (Auburn U), Chirag Agarwal* (U Illinois at Chicago), Anh Nguyen* (Auburn U)

* Equal contribution.   Code and paper: http://anhnguyen.me/project/sam

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

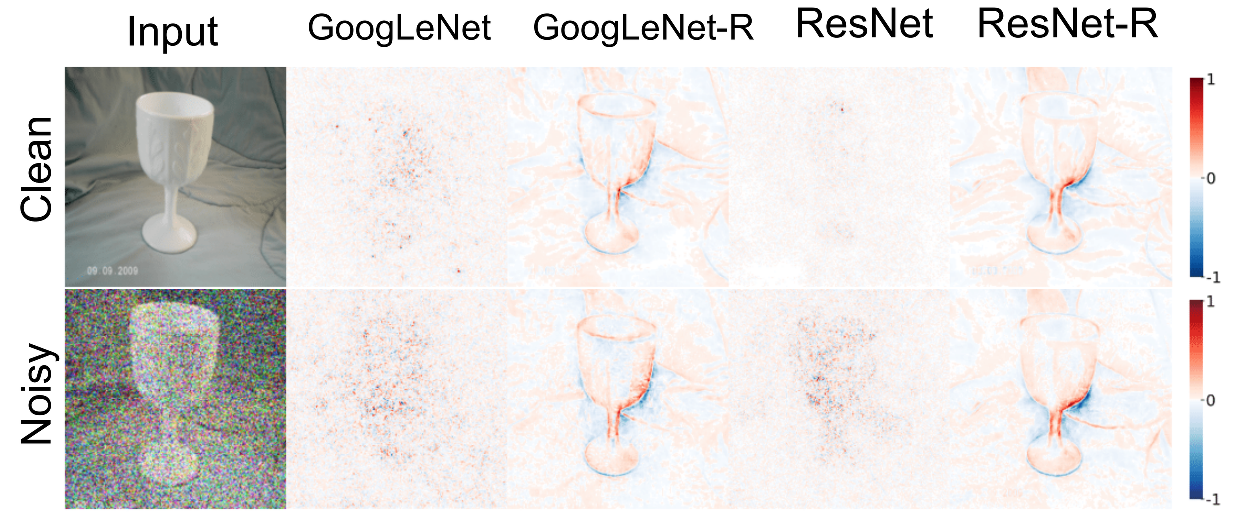
Summary

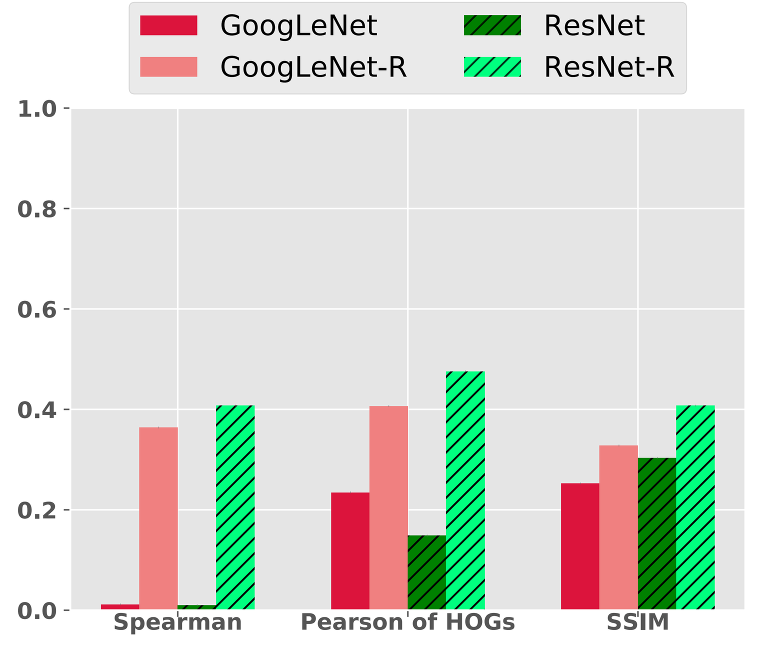
- Many attribution methods are highly sensitive to changes in their common hyperparameters.
- This sensitivity also translates into variation in accuracy scores.
- Compared to regular classifiers, explanations for *robust* classifiers are more invariant to input perturbations and more consistent when hyperparameter changes.
- Vanilla gradient images can exhibit clear visible outlines of objects in the input image.



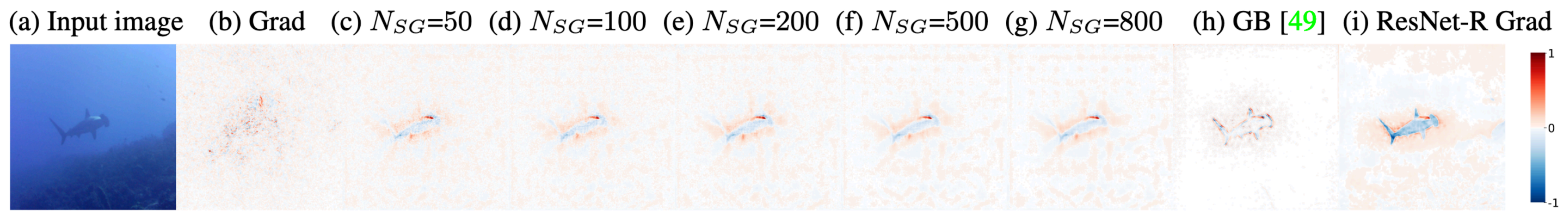
classifier: **matchstick** 0.535

Experiments

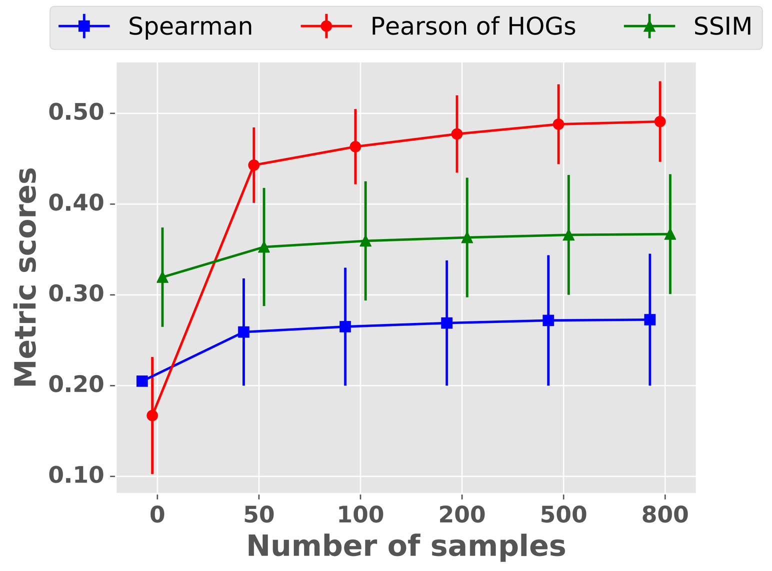




- The vanilla gradients of *robust* classifiers (GoogLeNet-R, ResNet-R) consistently exhibit visible object outlines, which is in stark contrast to the notoriously noisy gradient saliency maps of regular classifiers (GoogLeNet, ResNet).
- The gradient explanations of robust classifiers are significantly more invariant to a large amount of random noise added to the input image.

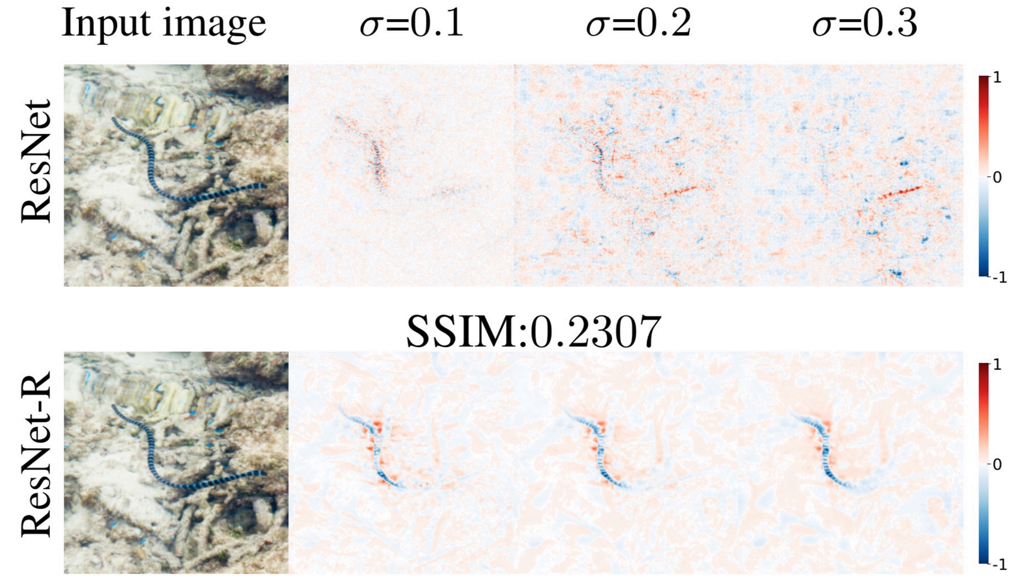


- Smoothing / denoising explanations may mislead our interpretation of explanations.
- E.g. SmoothGrad (SG) heatmaps of a ResNet become increasingly similar to the gradient of a completely different network (ResNet-R) as we increase the smoothing hyperparameter.
- On average, SSIM similarity increases ~1.4 × .

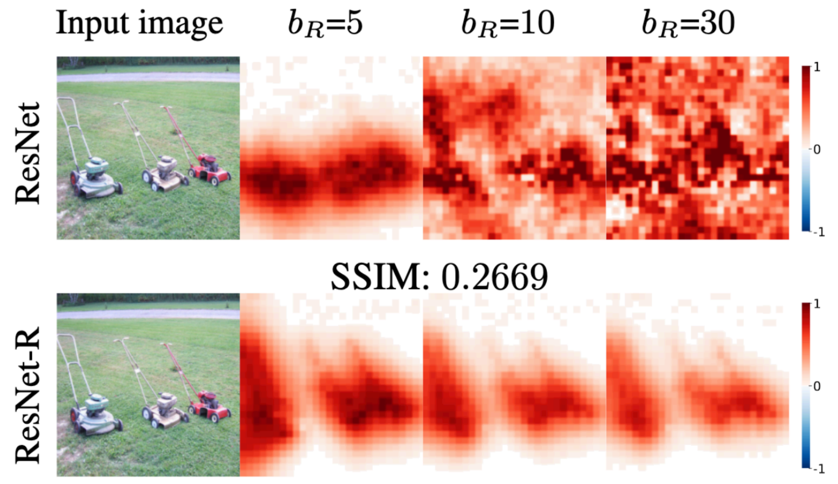


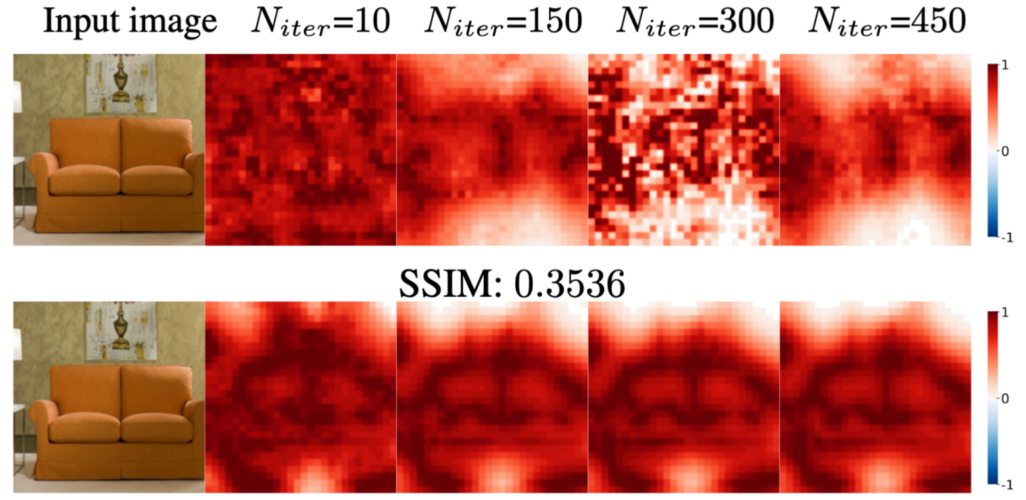
Sensitivity of Attribution Maps

Sliding-Patch (SP) heatmaps are, by design, sensitive to patch sizes.



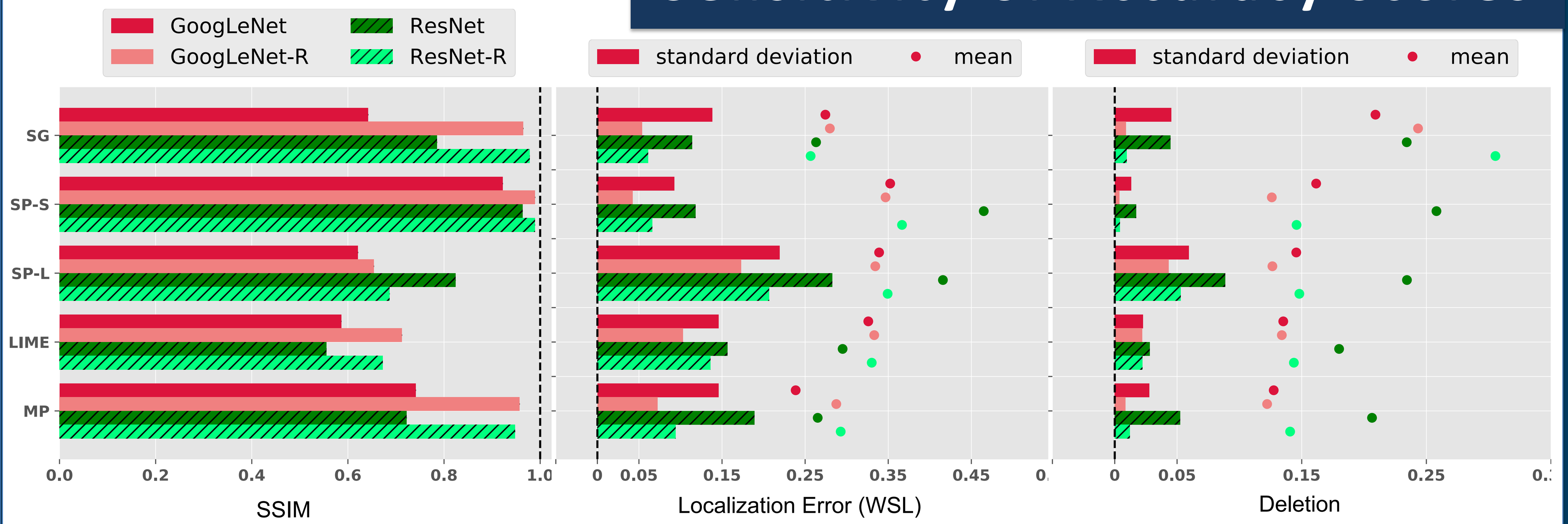
SG heatmaps for ResNet-R are more robust.



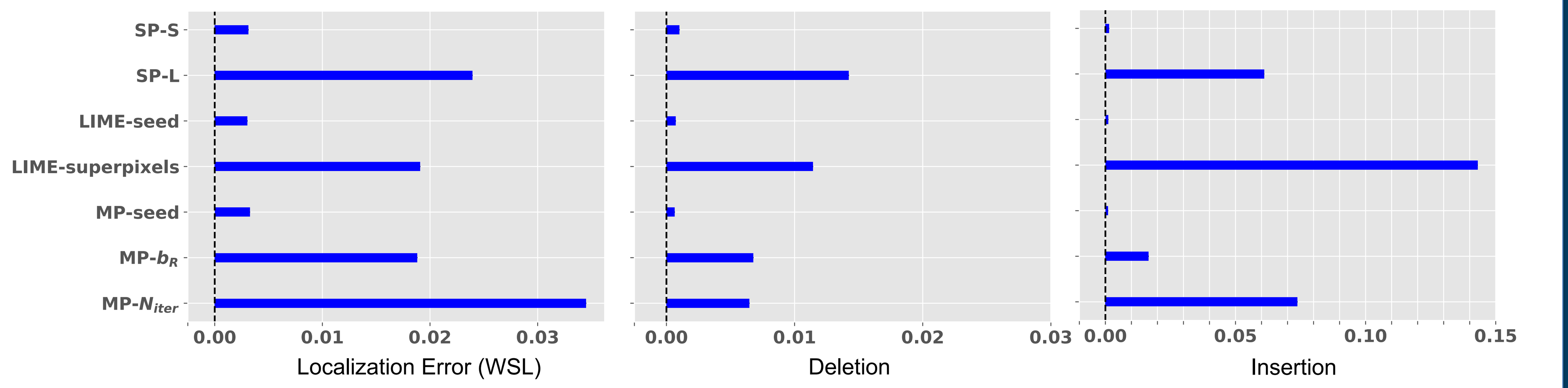


Meaningful Perturbation (MP) heatmaps for ResNet vary dramatically. In contrast, MP heatmaps for robust models (ResNet-R) are ~1.4 × more consistent under SSIM metric and converge faster (10 steps vs. 300 default).

Sensitivity of Accuracy Scores



- Variation in heatmaps (SSIM) also translates into the **variation in the accuracy scores** (WSL and Deletion).
- WSL scores are highly sensitive with average stds being ~0.51 × and ~0.31 × of the mean accuracy scores for both regular and robust models.
- Across all four tested hyperparameters, the correctness of explanations for robust models is on average 2.4 × less variable than regular models
- Even a small pixel-wise variation in explanation (~1 mean SSIM for SP-S) may lead to large variation in accuracy scores (stds are ~10% of mean statistics in SP-S)



- Some hyperparameters leads to higher variation in explanation accuracy scores as opposed to others.
- In LIME, the variation in the number of super-pixels leads to higher sensitivity as compared to the random seed (130.5 × higher std).
- In MP, the std of Insertion scores is 74 × and 16.6 × higher for variation in number of iteration and blur radius respectively as compared to changing the random seed.
- Changing the random seed in LIME vs MP (two different methods) interestingly causes a similar variation in all three-accuracy metrics.