Towards Mitigating Spurious Correlations in Image Classifiers with Simple Yes-no Feedback

Seongmin Lee Ali Payani Polo Chau

Georgia Tech CISCO

seongmin@gatech.edu

CRAYON is a practical approach using **yes-no feedback** to mitigate spurious correlations in deep learning models (e.g., a smile classifier attending to forehead)

Paper

Existing methods

Collect ground truth attention map

☹ Costly ☹ Time-intensive

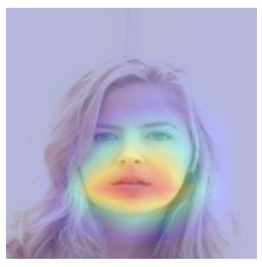
CRAYON

Collect **Yes-No Feedback**

☺ Simple ☺ Scalable ☺ Practical

CRAYON-ATTENTION

1. For each image, collect yes-no feedback on the **relevance of saliency maps**
2. Finetune model using RRR loss so that...



feedback: yes

For image with **relevant** map

Maintain similar saliency maps after refinement

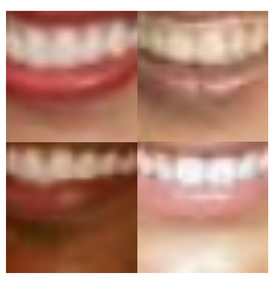
feedback: no

For image with **irrelevant** map

Attend the regions that are originally not highlighted

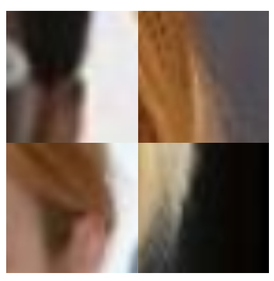
CRAYON-PRUNING

1. For each neuron in the penultimate layer, collect yes-no feedback on the **relevance of visual concepts that activate the neuron**
2. Prune the neurons activated by irrelevant concepts and finetune



feedback: yes

Neuron #609 with **relevant** concepts

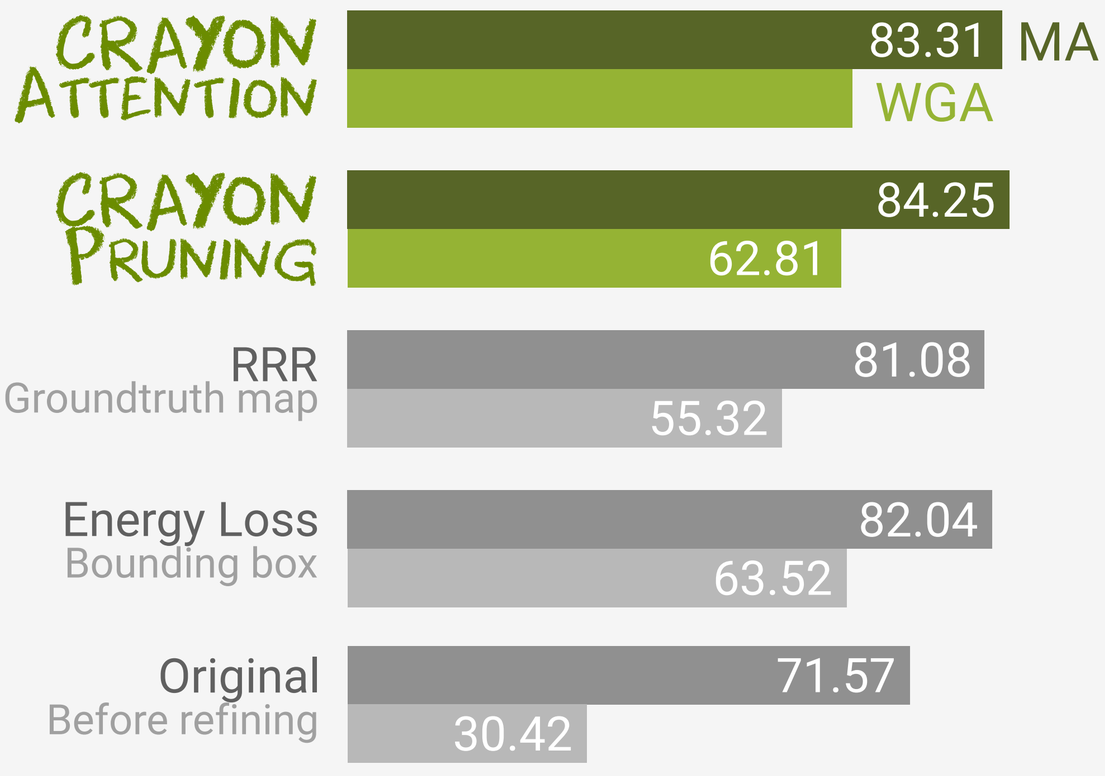


feedback: no

Neuron #0 with **irrelevant** concepts (pruned)

Results

Mean Acc. (MA) & **Worst Group Acc. (WGA)** of Smile Classifiers



Original attention maps of a smile classifier

Attention maps after CRAYON-ATTENTION refinement