Supplementary Material: SESS: Saliency Enhancing with Scaling and Sliding

Osman Tursun, Simon Denman, Sridha Sridharan, and Clinton Fookes

SAIVT Lab, Queensland University of Technology, Australia
{osman.tursun,s.denman,s.sridharan,c.fookes}@qut.edu.au

1 SESS Demo

The shared code of this paper includes a demo of the proposed method SESS. Please run "demo.ipynb" for the demo . This demonstration offers a comparison between SESS and base saliency methods including Grad-CAM [2], Guided Backpropagation [4], Group-CAM [6] and Score-CAM [5].

2 More Qualitative Results

This section provided qualitative results related to the step size and weighted average.

Weighed average In the fusion step, a weighted average is applied to ignore zero saliency values introduced by the calibration step. As Fig. 1 shows, without the weighted average, some parts of the target object will be under activated. For example, near the tale of the snake and cat. The saliency values of those under activated regions are increased with a weighted average.

Step-size In the default implementation of SESS, the step-size is set to 224 for efficiency. However, a smaller step size is beneficial for the generation of an accurate saliency map. As shown in Fig. 2, with a smaller step-size, the boundary of the target object is more accurate.

3 Applications

SESS is also useful for analysing the DNN models and saliency visualisation methods. This can be done through visualising all extracted saliency maps in L' as shown in Fig. 3. This visualisation shows: ResNet50 [1] is more robust to scale and occlusion when compared to VGG-16 [3], and ScoreCAM is more robust to scale variance when compared to Grad-CAM.

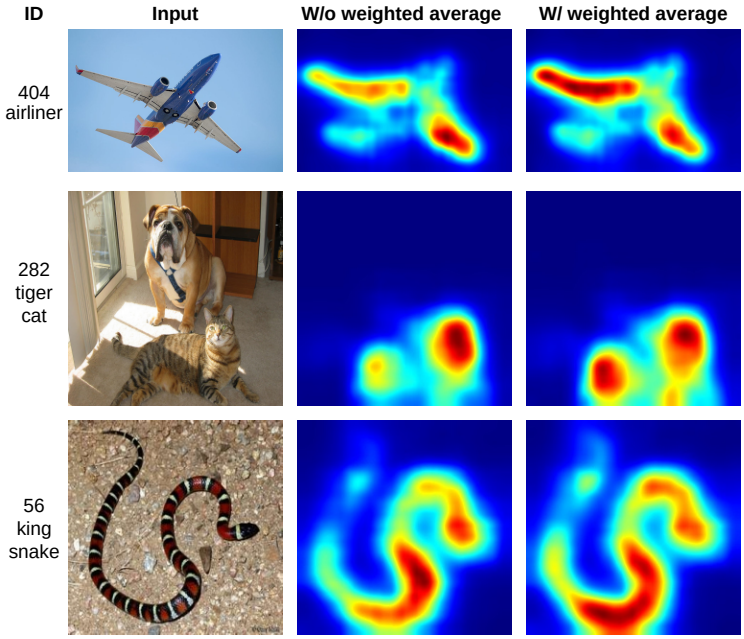

Fig. 1: Impact of weighted average: The weighted average increases the saliency values of under activated regions.

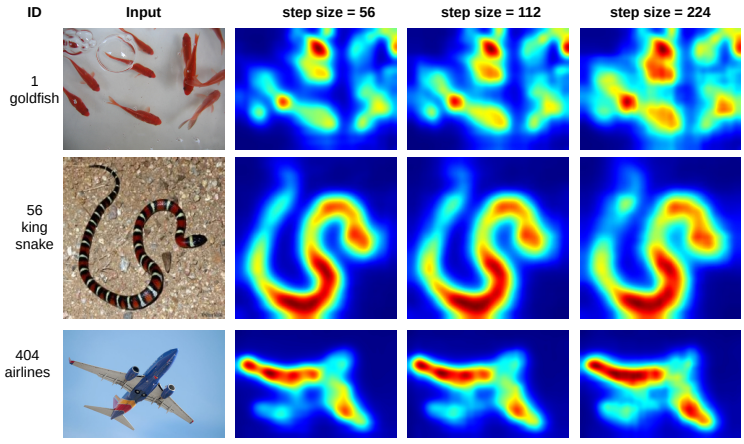

Fig. 2: Impact of step-size: A larger step size reduces the over activated regions near the boundary of the object.

(a) ResNet50 + Grad-CAM

(b) VGG-16 + Grad-CAM

(c) VGG-16 + Score-CAM

Fig. 3: Analysing DNN models and saliency visualisation methods with SESS. In this example the number of scales of SESS is set to 5. The red bounding box denotes the region from which the saliency is extracted. The target class id is 243 (Bull Mastiff).

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
2. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
3. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
4. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track) (2015)
5. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)
6. Zhang, Q., Rao, L., Yang, Y.: Group-cam: Group score-weighted visual explanations for deep convolutional networks. arXiv preprint arXiv:2103.13859 (2021)