APPENDIX

Saurabh Desai [2] Harish G. Ramaswamy[1,2]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Madras
[2]Robert Bosch Centre for Data Science and AI (RBC-DSAI),
Indian Institute of Technology, Madras
saketd403@gmail.com, hariguru@cse.iitm.ac.in

1. Introduction

In this supplementary document, we provide

- Section 2: Qualitative results showing Ablation-CAM and Grad-CAM visualizations for PASCAL VOC and Imagenet datasets.
- Section 3: Empirical comparison between Ablation-CAM and Grad-CAM for varying number of pixels.
- Section 4: Empirical comparison between Ablation-CAM and Grad-CAM++ for varying number of pixels.
- Section 5: Qualitative results showing Ablation-CAM visualizations for image captioning task.
- Section 6: Robustness of Ablation-CAM for adversarial images.

2. Qualitative results for Ablation CAM and Grad CAM visualizations.

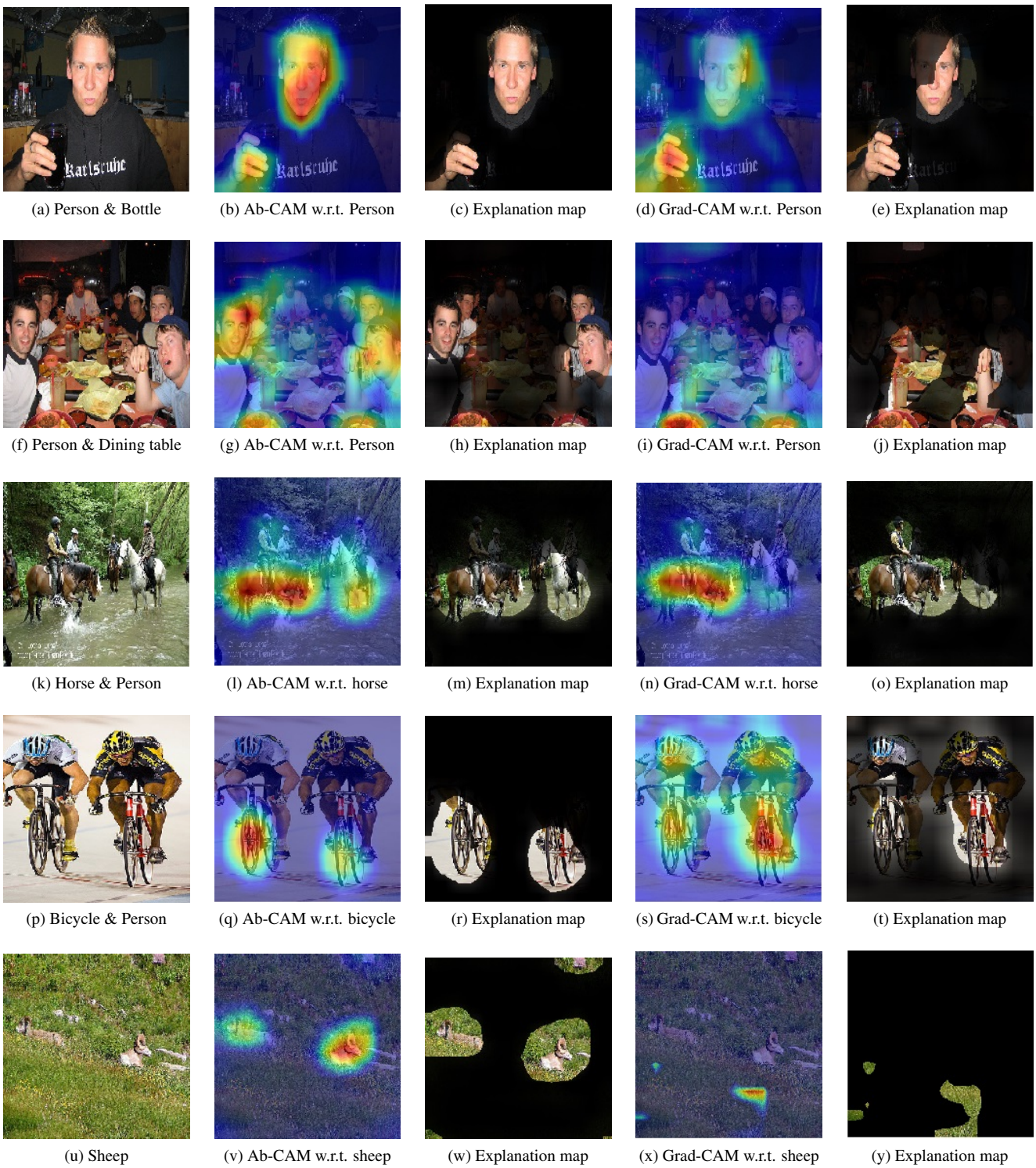
Fig 1 shows images from PASCAL VOC validation dataset which cause decision nodes of VGG16 model to saturate. The top four images contain two class categories each. A good visualization w.r.t a class node will be able to localize only the regions responsible for predicting that class. But it can be seen that Grad-CAM fails to do so. The discrimination capability of Grad-CAM is reduced which is evident as the visualization highlights regions of other classes as well. At the same time, the quality of visualizations is also reduced when the decision nodes are saturated (which can be seen from last example).
Figure 2 shows more examples from Imagenet 2012 validation dataset where Ablation-CAM covers more relevant regions of the object than Grad-CAM. Also, Abation-CAM highlights more instances of an object compared to Ablation-CAM (last image).

3. Comparative analysis of performance of Ablation-CAM and Grad-CAM for varying number of pixels.

In this section, we extend the empirical comparison of Ablation-CAM with Grad-CAM for varying number of pixels. In section 6.1, we limited ourselves to top 20 percent pixels for VGG-16 and Inception-v3 models. Table 2 shows results for VGG-16 for top 30 % and top 50 % pixels. We see that with increase in number of pixels, the activation drop decreases. Further, we have also carried out similar experiments on Alexnet model(table 3) pretrained on Imagenet data for top 20, 30 and 50 percent pixels. Interestingly, we see that for Alexnet the activation/confidence drop does not decrease by much as we one would expect when more pixels form part of the explanation map. Also, for same number of pixels the drop in activation and confidence scores is much greater than VGG-16. This observation stems from the fact that Alexnet does not generalize that well as compared to VGG16 and has lower performance on Imagenet data.
We have also showed performance of Ablation-CAM and Grad-CAM for Resnet-50 model (table 1) for top 20 percent pixels. As noted in section 6.1, since Resnet-50 has no fully-connected layers, Ablation-CAM and Grad-CAM will show almost similar performance. We found similar results for 30 % and 50 % pixels for Inception-v3 and Resnet-50 models.

4. Comparison with GradCAM++.

This section compares Ablation-CAM with Grad-CAM++. GradCAM++ is another gradient-based visualization technique which is improvement over Grad-CAM. We compare Ablation-CAM to GradCAM++ for VGG-16 (table 4) and for Alexnet (table 5) models. For VGG-16, we see that GradCAM++ performs much poorly when



(a) Person & Bottle (b) Ab-CAM w.r.t. Person (c) Explanation map (d) Grad-CAM w.r.t. Person (e) Explanation map

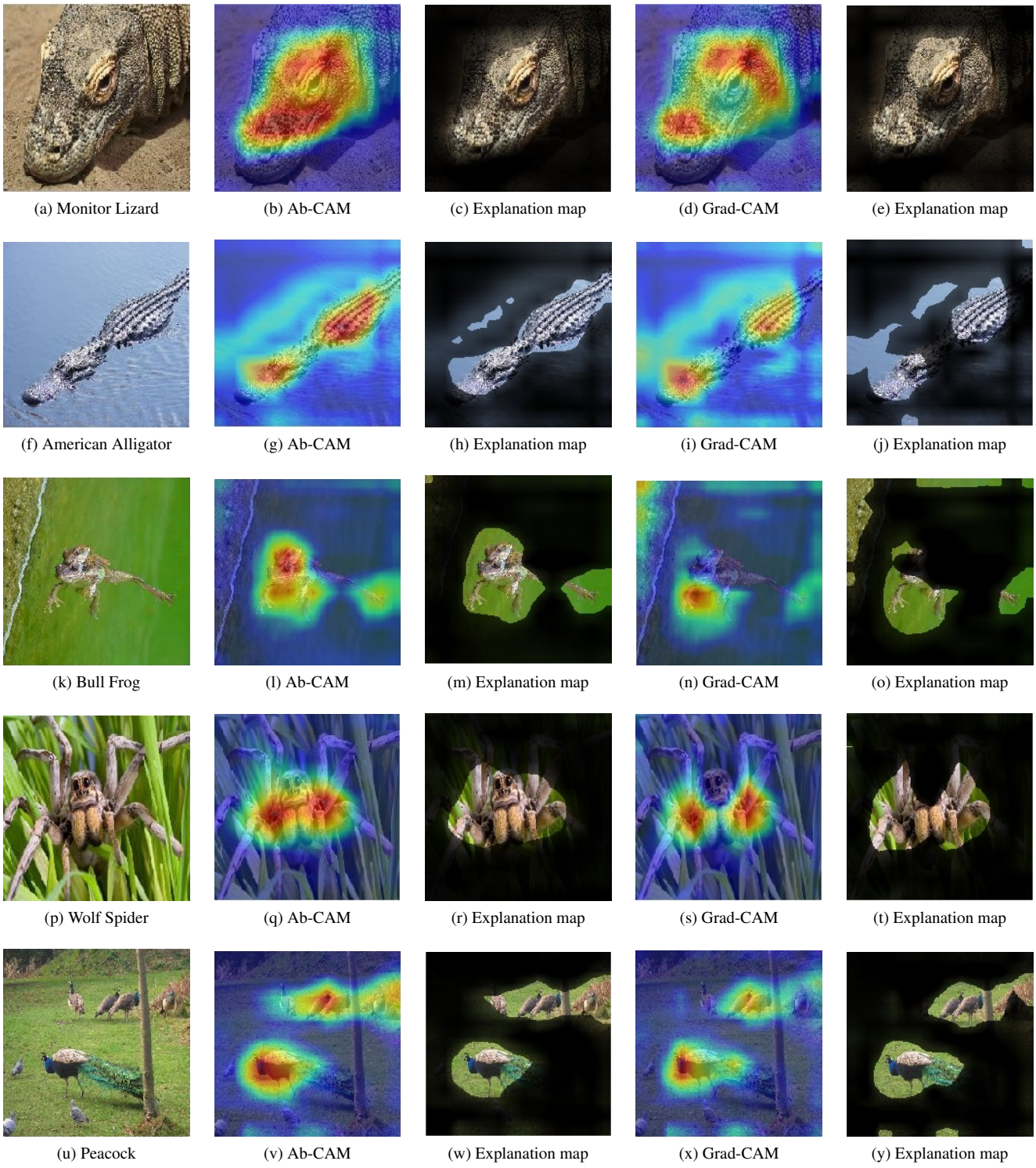
(f) Person & Dining table (g) Ab-CAM w.r.t. Person (h) Explanation map (i) Grad-CAM w.r.t. Person (j) Explanation map

(k) Horse & Person (l) Ab-CAM w.r.t. horse (m) Explanation map (n) Grad-CAM w.r.t. horse (o) Explanation map

(p) Bicycle & Person (q) Ab-CAM w.r.t. bicycle (r) Explanation map (s) Grad-CAM w.r.t. bicycle (t) Explanation map

(u) Sheep (v) Ab-CAM w.r.t. sheep (w) Explanation map (x) Grad-CAM w.r.t. sheep (y) Explanation map

Figure 1: The above are some examples from PASCAL dataset which cause the output class node to saturate. We have generated Ablation-CAM and Grad-CAM visualizations w.r.t. this saturated node. Each visualization is followed by corresponding explanation map containing top 20 percent pixels.



(a) Monitor Lizard (b) Ab-CAM (c) Explanation map (d) Grad-CAM (e) Explanation map

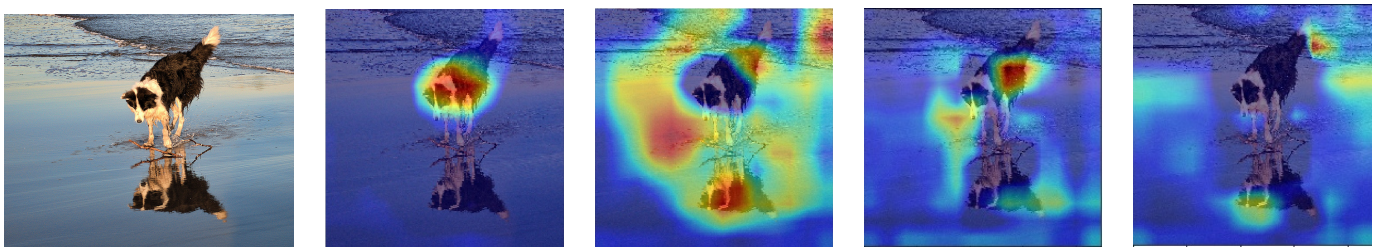
(f) American Alligator (g) Ab-CAM (h) Explanation map (i) Grad-CAM (j) Explanation map

(k) Bull Frog (l) Ab-CAM (m) Explanation map (n) Grad-CAM (o) Explanation map

(p) Wolf Spider (q) Ab-CAM (r) Explanation map (s) Grad-CAM (t) Explanation map

(u) Peacock (v) Ab-CAM (w) Explanation map (x) Grad-CAM (y) Explanation map

Figure 2: Above are some examples from Imagenet dataset. Ablation-CAM and Grad-CAM visualizations along with corresponding explanation map are shown following the original image. Explanation maps contain top 20% pixels.



(a) Dog is running across the beach (b) Ablation-CAM w.r.t. dog (c) Ablation-CAM w.r.t. beach (d) Grad-CAM w.r.t. dog (e) Ablation-CAM w.r.t. beach

Figure 3: The caption generated for original image (a) is "Dog is running across the beach". Ablation-CAM and Grad-CAM visualizations are generated for words "dog" and "beach".

(a) Original image (b) Adversarial image (c) Ablation-CAM for original (d) Ablation-CAM for adversarial

Figure 4: (a) shows original image for which VGG model predicts category "Boxer" with 42 % confidence whereas (b) is corresponding adversarial image for which confidence drops to 0.02 %.(c) and (d) are Ablation-CAM visualizations with respect to Boxer class node for original and adversarial images.

Metric	Ablation CAM	Grad CAM
Average % drop in confidence (lower is better)	**31.01**	31.70
Average % drop in activation (lower is better)	**19.96**	21.41
Percent increase in confidence (higher is better)	29.70	**29.93**
Percent increase in activation (higher is better)	**21.67**	21.27
Win % in confidence (higher is better)	**40.06**	34.75
Win % in activation (higher is better)	**46.78**	36.05

Table 1: Comparing Ablation-CAM with Grad-CAM for Resnet50 on Imagenet 2012 validation data.

compared to Ablation-CAM and even Grad-CAM. This is in contrary to findings of GradCAM++. In GradCAM++, explanation maps did not had same number of pixels and hence one with larger area produced better results. In our experiment, we made sure that both the explanation maps contain same number of pixels for fair comparison. For Alexnet, we see that GradCAM++ performs better than GradCAM. But Ablation-CAM still outperforms GradCAM++.

By the definition of GradCAM++, it is equivalent to GradCAM for networks without fully connected layers such as Inception and Resnet. Hence, we did not show results comparing Ablation-CAM with GradCAM++ for these models.

5. Explanations for image captioning task.

Similar to Grad-CAM, we tried our Ablation-CAM on image captioning task. For this, we considered image captioning model with architecture similar to popular "Show and Tell" model (Vinyals *et al*. [1]). This architecture includes a CNN to encode the image followed by an LSTM

Top % pixels	30%		50%	
Metric	Ablation CAM	Grad CAM	Ablation CAM	Grad CAM
Average % drop in confidence (lower is better)	**38.85**	43.36	**34.45**	40.28
Average % drop in activation (lower is better)	**26.55**	32.34	**22.55**	28.67
Percent increase in confidence (higher is better)	**25.87**	24.42	**28.79**	26.20
Percent increase in activation (higher is better)	**16.73**	14.05	**20.14**	16.78
Win % in confidence (higher is better)	**48.91**	32.79	**50.59**	29.62
Win % in activation (higher is better)	**58.54**	31.22	**58.99**	28.64

Table 2: Comparing Ablation-CAM with Grad-CAM for VGG-16 on Imagenet 2012 validation data.

Top % pixels	20%		30%		50%	
Metric	Ablation CAM	Grad CAM	Ablation CAM	Grad CAM	Ablation CAM	Grad CAM
Average % drop in confidence (lower is better)	**76.38**	80.65	**77.19**	81.32	**76.62**	80.42
Average % drop in activation (lower is better)	**58.48**	66.07	**58.51**	66.47	**55.52**	63.23
Percent increase in confidence (higher is better)	**11.88**	9.73	**11.52**	9.55	**11.87**	10.11
Percent increase in activation (higher is better)	**7.61**	5.45	**7.85**	5.95	**9.30**	8
Win % in confidence (higher is better)	**56.03**	34.58	**56.67**	34.16	**57.18**	33.23
Win % in activation (higher is better)	**63.93**	29.86	**62.89**	30.60	**59.85**	32.16

Table 3: Comparing Ablation-CAM with Grad-CAM for Alexnet on Imagenet 2012 validation data.

to generate the captions. The data used is Flickr30K dataset. Figure 3 shows an image for which caption generated is "Dog is running across the beach". We generated visualizations w.r.t. specific output nodes which represented specific words of caption. The benefit of this experiment is that it helps to understand portion of image responsible for generation of a particular word in caption. We have plotted visualizations w.r.t. words "dog" and "beach". Clearly, Ablation-CAM visualizations are much more interpretable than Grad-CAM's.

6. Robustness to adversarial images

Adversarial examples are slight imperceptible perturbations of input images to fool the network into misclassifying them. We generate adversarial images for the ImageNet trained VGG-16 model such that it assigns a very low probability to categories that are present. We then compute Ablation-CAM visualizations for the categories that are present in image. We can see from figure 4 that inspite of the network being completely certain about the absence of these categories, Ablation-CAM visualizations can correctly localize the categories. This shows the robustness of Ablation-CAM to adversarial noise.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the*

Top % pixels	20 %		30 %		50 %	
Metric	Ablation CAM	Grad CAM++	Ablation CAM	Grad CAM++	Ablation CAM	Grad CAM++
Average % drop in confidence (lower is better)	**41.52**	61.33	**38.85**	60.01	**34.45**	53.29
Average % drop in activation (lower is better)	**29.23**	47.47	**26.55**	45.68	**22.55**	38.75
Percent increase in confidence (higher is better)	**24**	12.67	**25.87**	13.26	**28.79**	15.72
Percent increase in activation (higher is better)	**14**	5.17	**16.73**	5.64	**20.14**	7.05
Win % in confidence (higher is better)	**64.38**	27.02	**56.61**	39.72	**65.18**	23.65
Win % in activation (higher is better)	**74.04**	22.61	**76.50**	19.85	**74.59**	20.53

Table 4: Comparing Ablation-CAM with Grad-CAM++ for VGG16 on Imagenet 2012 validation data.

Top % pixels	20 %		30 %		50 %	
Metric	Ablation CAM	Grad CAM++	Ablation CAM	Grad CAM++	Ablation CAM	Grad CAM++
Average % drop in confidence (lower is better)	**76.38**	78.16	**77.19**	79.34	**76.62**	79.36
Average % drop in activation (lower is better)	**58.48**	61.75	**58.51**	62.28	**55.52**	60.05
Percent increase in confidence (higher is better)	**11.88**	10.60	**11.52**	10.12	**11.87**	10.18
Percent increase in activation (higher is better)	**7.61**	5.88	**7.85**	5.98	**9.30**	7.12
Win % in confidence (higher is better)	**46.95**	43.99	**48.68**	42.62	**51.40**	39.57
Win % in activation (higher is better)	**53.50**	40.74	**54.44**	39.63	**55.44**	37.48

Table 5: Comparing Ablation-CAM with Grad-CAM++ for Alexnet on Imagenet 2012 validation data.

IEEE Conference on Computer Vision and Pattern Recognition, page 31563164, 2015.