There and Back Again: Revisiting Backpropagation Saliency Methods Supplementary Material

A. Pointing Game details

We use [6]'s Pointing Game for quantitative evaluation of saliency methods. Saliency maps are computed with respect to every object class present in each image. If the maximally salient point for each class lands on the ground truth annotation for that object (within a threshold of 15 pixels), then a "hit" is recorded; otherwise, a "miss" is recorded. Pointing game accuracy is computed as the mean over per-class accuracies given by the following: $\frac{|\text{hits}|}{|\text{hits+misses}|}$.

We evaluate VGG16 and ResNet50 networks that have been trained on ImageNet and fine-tuned on PASCAL VOC and COCO. We evaluate on the PASCAL VOC 2007 test split ($N = 4952$ images) and COCO 2014 val split ($N \approx 50k$). We also show performance on the difficult subsets of the data provided by [6]; these are images for which the total area of the annotations (bounding boxes for PASCAL VOC and segmentation masks for COCO) for the given object class is less than 25% of the image size and for which there is at least one other object class present. We use [2]'s TorchRay library for evaluation; see [6] for more details.

B. Virtual identity trick correlations

We computed the correlation between saliency maps generated both with and without the virtual identity trick (paper section 4.1). The high correlation shown in Table 1, as well as the minimal difference in pointing game performance ($0.53\% \pm 0.62\%$) demonstrates that the identity trick closely approximates the behaviour of calculating the spatial contributions for the original convolutional layers.

Architecture	Layer	Correlation
ResNet50	layer1.0.conv1	99.48 ± 0.55
ResNet50	layer2.0.conv1	99.32 ± 0.35
ResNet50	layer3.0.conv1	99.16 ± 0.42
ResNet50	layer4.0.conv1	98.35 ± 1.05
VGG16	features.2	97.77 ± 1.34
VGG16	features.7	99.15 ± 0.42
VGG16	features.14	98.99 ± 0.57
VGG16	features.21	97.79 ± 1.42
VGG16	features.28	94.33 ± 3.09

Table 1. **Correlations between saliency maps generated with and without the virtual identity trick.** For both VGG16 and ResNet50, the high values of correlation across the network (min of 94.33) justify the use of the identity trick.

	ResNet50		VGG16	
	All	Difficult	All	Difficult
CEB	layer3	layer4	feat.29	feat.29
EB	layer4	layer4	feat.29	feat.29
GC	layer4	layer4	feat.29	feat.29
Gd	layer2	layer2	feat.29	feat.22
Gds	layer2	layer2	feat.8	feat.8
Gui	layer3	layer3	input	input
LA	layer4	layer4	feat.29	feat.29
NG	layer3	layer3	feat.22	feat.22
sNG	layer4	layer4	feat.29	feat.29

Table 2. **Best individual layer for the Pointing game on VOC07.** (C)EB: (Contrastive) Excitation Backprop, GC: GradCAM, Gd(s): Gradient (sum), Gui: guided backprop, LA: linear approximation, (s)NG: (selective) NormGrad.

C. Performance of combining saliency maps

As noted in paper section 4.2, feature spread and classification accuracy are both interpretable as measures of feature sensitivity with respect to the class of the input image. Figure 1 shows that both increase with network depth, with the exception of feature spread for VGG16, which decreases in the last two layers despite simultaneously increasing classification accuracy. We note that classification accuracy is a more reliable metric than feature spread as it directly codes for the separability of features with respect to class, whereas feature spread is susceptible to non-material differences in the absolute scale of activation values across layers. In the case of VGG16, this means the features in the last two layers differ less across images (and hence, classes) in absolute terms, but are nonetheless highly discriminative of class.

The performance gains of our weighting schemes over using the best individual layer are given in table 3. The best individual layers for computing maps are given in table 2, which is notable as in no case does the practice of computing saliency at the earliest layer produce the best performance.

D. Meta-saliency analysis

We hypothesized that meta vs. non-meta saliency maps should be less correlated for validation images than for training images. This is because the inner gradient step of meta-saliency should not be as impactful for a seen training image as for an unseen validation image.

For both NormGrad and selective NormGrad, we evaluated over time the average correlation between the impor-

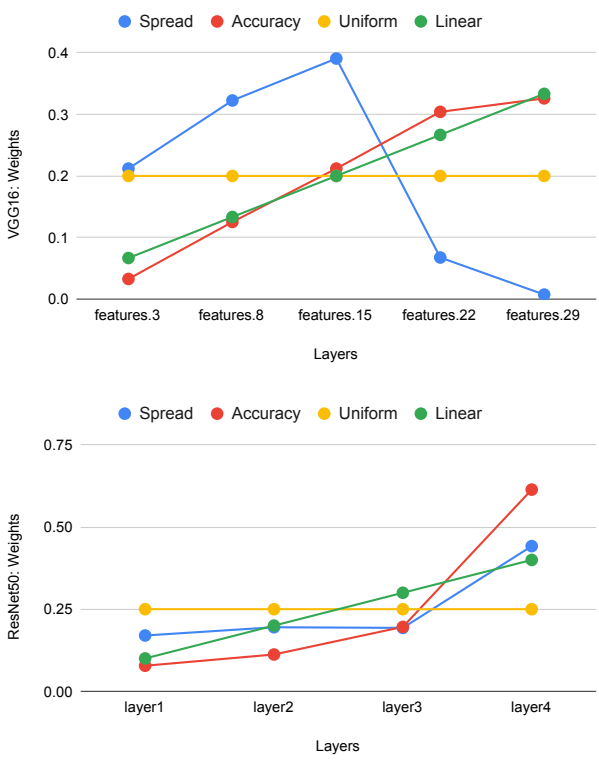

Figure 1. **Weights for the different weighting methods.** VGG16 on top and ResNet50 on the second row.

			ResNet50		VGG16	
			All	Difficult	All	Difficult
LA	+	accuracy	0.57	1.47	0.22	2.09
LA	+	spread	0.81	1.66	-3.01	-7.18
LA	+	linear	0.68	0.74	0.36	2.22
LA	+	uniform	0.58	1.29	-0.61	-0.20
LA	×	accuracy	0.87	**1.84**	0.42	**2.91**
LA	×	spread	**1.01**	1.78	-2.92	-6.80
LA	×	linear	0.52	0.72	**0.49**	2.87
LA	×	uniform	0.32	0.83	-0.25	0.52
sNG	+	accuracy	1.08	1.37	0.48	0.85
sNG	+	spread	0.94	1.40	-3.66	-6.43
sNG	+	linear	0.94	1.67	0.62	0.85
sNG	+	uniform	0.25	0.54	-0.93	-2.55
sNG	×	accuracy	**1.26**	1.65	**0.83**	**1.88**
sNG	×	spread	1.12	**2.09**	-3.13	-6.18
sNG	×	linear	0.63	0.95	0.72	0.62
sNG	×	uniform	0.89	1.79	-0.27	-1.76

Table 3. **Score gains compared to best individual layer performance for weighting methods with Linear Approximation (LA) and selective NormGrad (sNG) on the Pointing game on VOC07.** Paper showed that LA and sNG benefit from weighting methods. Here we can observe that using the weights as exponents in a product (lines with ×) is the most effective solution for both LA and sNG on ResNet50 and VGG16. Both the weightings using the features spread and the layer accuracy perform the best on ResNet50 but only the layer accuracy perform consistently across datasets and saliency methods.

tance maps with and without meta-saliency on the training and testing sets respectively. Figure 2 demonstrates that the correlation scores are indeed decreased for validation images.

E. Model weights sensitivity

In fig. 5, we show the effects of cascading randomization on linear approximation and selective NormGrad methods by using the same images as [1] ("junco", "corn" and "Irish terrier") to illustrate the model weights sensitivity.

Qualitatively, both methods - with and without meta-saliency - demonstrate the desired sensitivity to model weights as the saliency maps progressively lose focus from the object target as layer depth decreases. We also observe that using meta-saliency on top of the chosen base saliency method (shown on every other line of fig. 5) delays the degradation in saliency maps to earlier layers.

F. Image captioning

We show in this section a few examples of our proposed selective NormGrad method applied to the image captioning setting. We use a variant [4] of the original neuraltalk2 [3] with a ResNet101 as backbone network followed by the LSTM caption model. As in [5] we do a backward pass using the log probability of the generated caption as objective function. We apply selective NormGrad before the final global average pooling and at the ReLU layers just after the downsampling shortcuts of the third and fourth macro blocks. We also compare these saliency maps with the product combination map of these layers using a linear weighting scheme (see section 4.2 in the paper). We notice in fig. 6 that using a combination of layers produces a sharper saliency map than for individual layers.

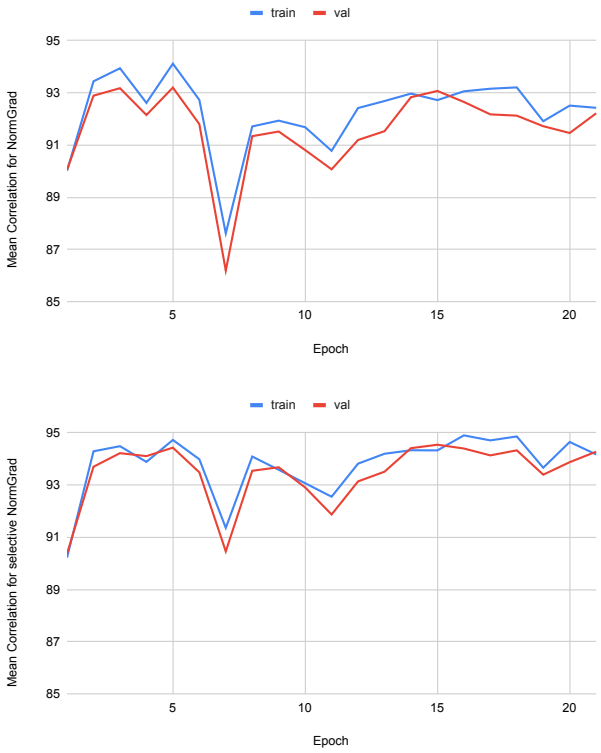

Figure 2. **Correlation between meta and non-meta saliency maps over time for train and val splits.** For both NormGrad (top) and selective NormGrad (bottom), the correlation is lower on the validation split.

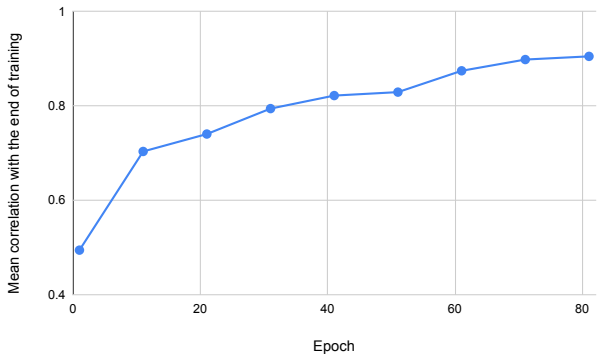

Figure 3. **Mean correlation on the val split between the meta-saliency maps of selective NormGrad at epoch t and the end of training.** The saliency maps stabilize at the end of the training.

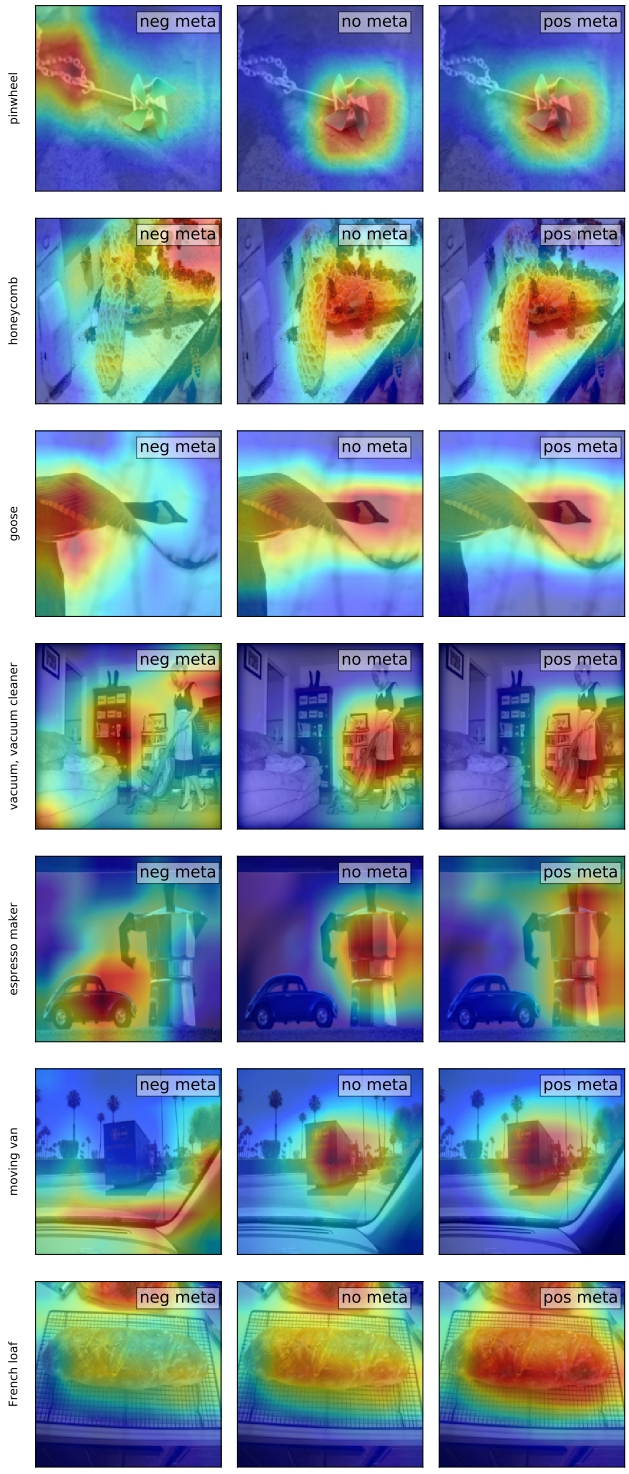

Figure 4. **Examples of positive and negative meta-saliency for selective NormGrad**. Negative meta-saliency (left images) corresponds to a gradient ascent inner step whereas a gradient descent step is used for the positive meta-saliency (images on the right). The center images do not use meta-saliency.

Figure 5. **Model weights sensitivity on VGG16.** Linear Approximation and selective NormGrad with and without meta-saliency.

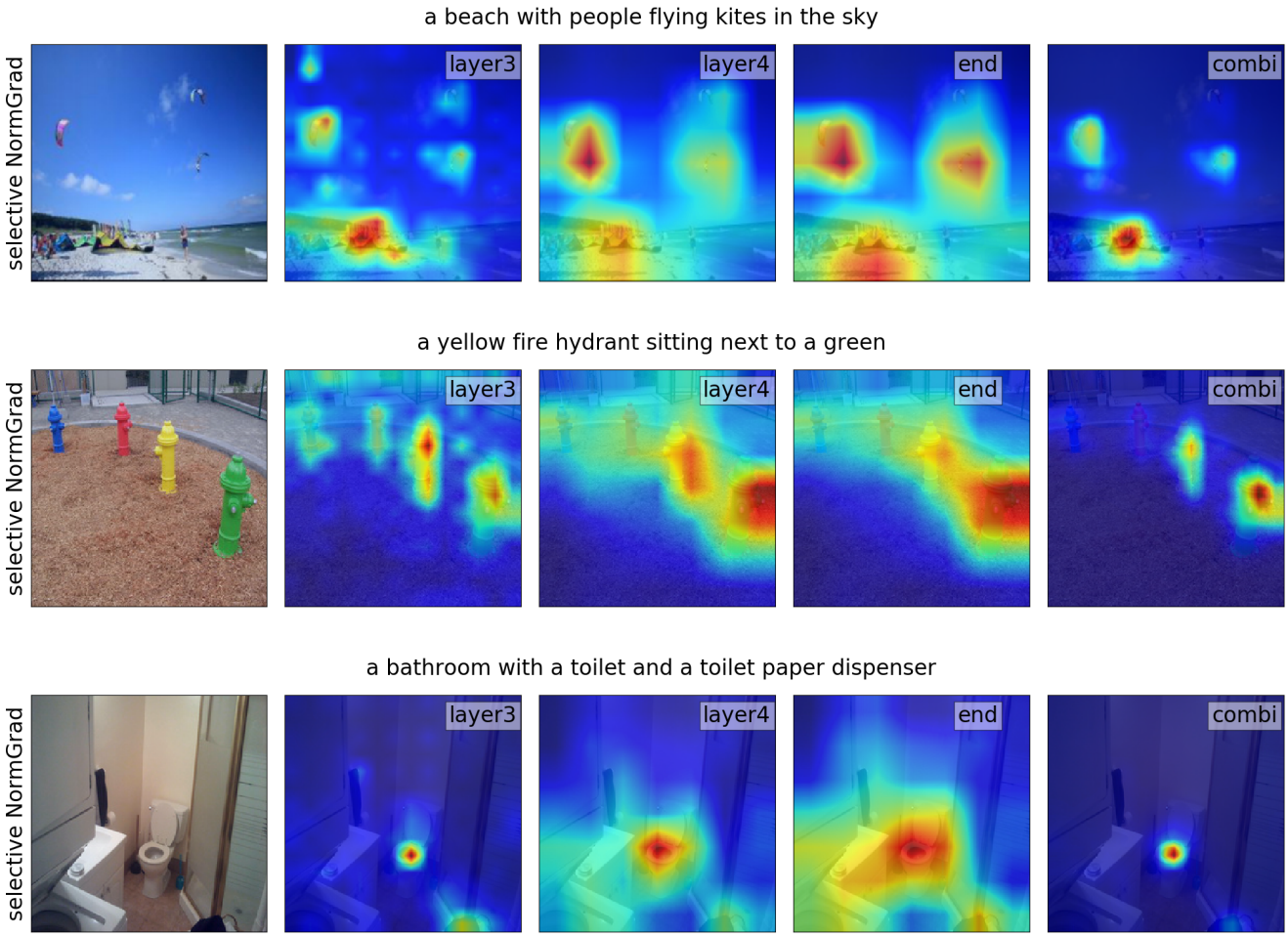

Figure 6. **Image captioning explanations.** Selective NormGrad is applied at different layers or combination of layers (last column) of a ResNet101. We observe that saliency maps at individual layers highlight a big part of the images. Using a combination of layers allows a clearer focus on the important parts of the images.

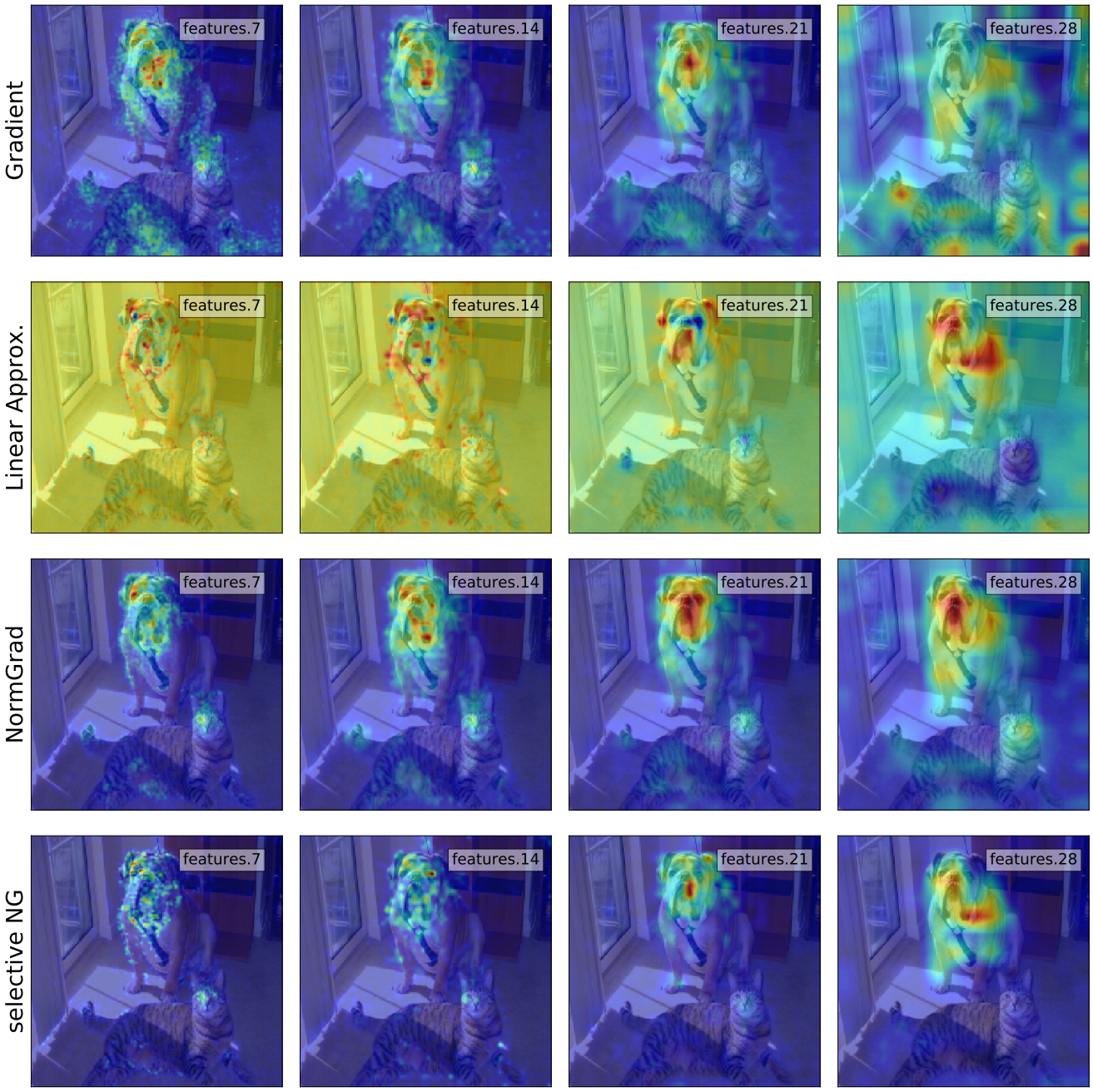

Figure 7. **Visualizations for some methods of the Extract & Aggregate framework at different layers of VGG16.** The gradient backprogation method (first row) works well at all stages except at the end of the network. Selective NormGrad (last row), NormGrad (third row) and Linear Approximation (second row) perform well across the network. Finally we can observe that selective NormGrad and Linear Approximation are more class selective than NormGrad as the non targeted "tiger cat" appears more in the third row.

References

[1] Julius Adebayo, Justin Gilmer, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proc. NeurIPS*, 2018. 2

[2] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proc. ICCV*, 2019. 1

[3] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, 2015. 2

[4] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. *arXiv*, 2018. 2

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. ICCV*, 2017. 2

[6] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *Proc. ECCV*, 2016. 1