Classifier-Agnostic Saliency Map Extraction

Konrad Żołna,[1, 2] **Krzysztof J. Geras,**[2, 3] **Kyunghyun Cho**[1, 4, 5]
[1]Jagiellonian University*
[2]New York University
[3]NYU School of Medicine
[4]CIFAR Azrieli Global Scholar
[5]Facebook AI Research
konrad.zolna@gmail.com, k.j.geras@nyu.edu, kyunghyun.cho@nyu.edu

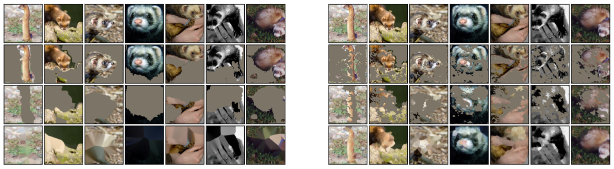
Abstract

Extracting saliency maps, which indicate parts of the image important to classification, requires many tricks to achieve satisfactory performance when using classifier-dependent methods. Instead, we propose classifier-agnostic saliency map extraction. This allows to find all parts of the image that any classifier could use, not just one given in advance. This way we extract much higher quality saliency maps.

Introduction

The success of deep convolutional networks for large-scale object recognition (Simonyan and Zisserman 2014; He et al. 2016) has spurred interest in utilizing them to automatically detect and localize objects in natural images. It has been demonstrated that the gradient of the class-specific score of a given classifier could be used for extracting a saliency map of an image (Simonyan, Vedaldi, and Zisserman 2013). Such classifier-dependent saliency maps can be utilized to analyze inner working of a specific network. However, these are not identifying *all* "evidence" in a given image because only the part of the image that is used by a given model is highlighted.

In this work, we aim at finding saliency maps indicating pixels which aids classification, i.e. we want to find pixels in the input image such that if they were masked, it would confuse *an unknown classifier*. Assuming we were given a classifier, a naive approach would be to train a generative model to output a mask (a saliency map) confusing that classifier. That can be achieved using a simple GAN-like approach (Goodfellow et al. 2014) where the classifier acts as a fixed discriminator. Unfortunately, as we prove experimentally, this solution suffers from the same issues as prior approaches. We argue that the strong dependence on a given classifier lies at the center of the problem. To tackle this directly we propose to train a saliency mapping that is not strongly coupled with any specific classifier.



(a) CASM (Ours)

(b) Baseline

Figure 1: The original images are in the first row. In the following rows masked-in images, masked-out images and inpainted masked-out images are shown, respectively. The proposed approach remove all relevant pixels and hence the inpainted images show the background only.

We qualitatively find that the proposed approach extracts higher quality saliency maps compared to classifier-dependent methods, as can be seen in Fig. 1. Extracted saliency maps show all the evidence without using any symptom-masking methods such as total variation regularization. We also evaluate our method quantitatively by using the extracted saliency maps for object localization. We observe that the proposed approach outperforms the existing weakly-supervised techniques setting new state-of-the art on ImageNet dataset and closely approaches the localization performance of a strongly supervised model.

Our method has many potential applications, in which being classifier-agnostic is of primary importance. For instance, in medical image analysis, where we are interested not only in class prediction but also in indicating which part of the image is important to classification. Importantly, it is criticial to indicate all parts of the image, which can influence diagnosis, not just ones used by a specific classifier.

Classifier-agnostic saliency map extraction

We tackle a problem of extracting a salient region of an image as a problem of extracting a mapping $m : \mathbb{R}^{W \times H \times 3} \rightarrow [0, 1]^{W \times H}$ over an input image $x \in \mathbb{R}^{W \times H \times 3}$. Such a mapping should retain (=1) any pixel of the input image if it aids classification, while it should mask (=0) any other pixel. Earlier work has largely focused on a setting in which a classifier f was given. These approaches can be implemented as finding $m = \arg\max_{m'} S(m', f)$, where S is a score function corresponding to a classification loss, i.e., $S(m, f) = \frac{1}{N} \sum_{n=1}^{N} l(f((1 - m(x_n)) \odot x_n), y_n) + R(m)$, where $\odot$ denotes a masking operation, $R(m)$ is a regularization term and l is a per example classification loss, such as cross-entropy. We are given a training set

*ul. Łojasiewicza 6, 30-348 Kraków, Poland, +48 12 664 66 29

Algorithm 1: Classifier-agnostic saliency map extraction

```
input  : an initial classifier $f^{(0)}$ parameterized by $\theta_f$,
         an initial mapping $m^{(0)}$ parameterized by $\theta_m$,
         dataset $D$, learning rates $\eta_f$ and $\eta_m$,
         number of iterations $K$
output: the final mapping $m^{(K)}$
Initialize a sample set $F^{(0)} = \left\{ f^{(0)} \right\}$.
for $k \leftarrow 1$ to $K$ do
    $\theta_{f^{(k)}} \leftarrow \theta_{f^{(k-1)}} - \eta_f \nabla_{\theta_f} S(m^{(k-1)}, f^{(k-1)})$
    $f' \leftarrow$ Sample($\left\{ f^{(0)}, f^{(1)}, \dots, f^{(k)} \right\}$)
    $\theta_{m^{(k)}} \leftarrow \theta_{m^{(k-1)}} + \eta_m \nabla_{\theta_m} S(m^{(k-1)}, f')$
```

$D = \{(x_1, y_1), \dots, (x_N, y_N)\}$. This optimization procedure could be interpreted as finding a mapping m that maximally confuses the given classifier f. We refer to it as a *classifier-dependent saliency map extraction*.

We propose to consider not only a single fixed classifier but all possible classifiers weighted by their posterior probabilities. That is, now $m = \arg\max_{m'} \mathbb{E}_f \left[S(m', f)\right]$, where the posterior, $p(f|D, m)$, is defined to be proportional to the exponentiated classification loss on masked images. Solving this optimization problem is equivalent to searching over the space of all possible classifiers, and finding a mapping m that works with all of them. We call the proposed approach a *classifier-agnostic saliency map extraction*.

Finding m as defined above is, unfortunately, generally intractable. This arises from the intractable expectation over the posterior distribution. Thus, we solve this problem approximately, as presented in Alg. 1, using the fact that stochastic gradient descent performs approximate Bayesian posterior inference (Welling and Teh 2011; Mandt, Hoffman, and Blei 2017). Note that our algorithm resembles the training procedure of GANs, where mapping m takes the role of a generator and the classifier f (and all its previous iterations) can be understood as a discriminator. Mapping m and the classifier f are trained simultaneously. See the official implementation or full paper for the training procedure details (`https://github.com/kondiz/casme`).

Experiments

Our models were trained on the official ImageNet training set with ground truth class labels. We use ResNet-50 (He et al. 2016) as a classifier f in our experiments. We follow an encoder-decoder architecture for constructing a mapping m. The encoder is implemented also as a ResNet-50.

We use the abbreviation CASM (classifier-agnostic saliency mapping) to denote the final model obtained using the proposed method. Our baseline model (Baseline) is of the same architecture but it is trained with a fixed classifier. We visualize the learned mapping m by inspecting the saliency map of each image in three different ways. We consider the **masked-in image** $m(x) \odot x$, which ideally leaves only the relevant pixels visible, and the **masked-out image** $(1 - m(x)) \odot x$, which highlights pixels irrelevant to classification. We also visualize the **inpainted masked-out image** using an inpainting algorithm (Telea 2004). This allows us to inspect whether the object that should be masked out cannot be easily reconstructed from nearby pixels (see Figure 1).

Model	$\downarrow$
Baseline	53.5
CASM (ours)	**36.1**
ALN (Fan, Zhao, and Ermon 2017)	43.5
Mask (Fong and Vedaldi 2017)	43.1
Grad (Simonyan, Vedaldi, and Zisserman 2013)	41.7
Masking model (Dabkowski and Gal 2017)	36.7
Supervised:	
VGG Net (Simonyan and Zisserman 2014)	**34.3**

Table 1: Localization task results.

We also evaluate our method quantitatively by using the extracted saliency maps for object localization (ILSVRC'14 localization task). We report the performance of CASM, Baseline and prior works in Table 1 using the most widely used metric (Fong and Vedaldi 2017). Most of the existing approaches, except for ALN (Fan, Zhao, and Ermon 2017), assume the knowledge of the target class, unlike the proposed approach. The table clearly shows that CASM performs better than all prior approaches including the classifier-dependent Baseline. The fully supervised approach is the only approach that outperforms CASM.

Acknowledgments

K. Cho thanks support by AdeptMind, eBay, TenCent, NVIDIA and CIFAR. K. Żołna is supported by the National Science Center, Poland (2017/27/N/ST6/00828).

References

Dabkowski, P., and Gal, Y. 2017. Real time image saliency for black box classifiers. In *NIPS*.

Fan, L.; Zhao, S.; and Ermon, S. 2017. Adversarial localization network. In *Learning with limited labeled data: weak supervision and beyond, NIPS Workshop*.

Fong, R. C., and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Mandt, S.; Hoffman, M. D.; and Blei, D. M. 2017. Stochastic gradient descent as approximate bayesian inference. *The JMLR*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint*.

Telea, A. 2004. An image inpainting technique based on the fast marching method. *Journal of graphics tools*.

Welling, M., and Teh, Y. W. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*.