SalGAN: Visual Saliency Prediction with Adversarial Networks

Junting Pan, Elisa Sayrol and Xavier Giro-i-Nieto
Image Processing Group
Universitat Politecnica de Catalunya
Barcelona, Catalonia/Spain
xavier.giro@upc.edu

Cristian Canton Ferrer
Facebook*
Seattle (WA), USA
ccanton@fb.com

Jordi Torres
Barcelona Supercomputing Center
Barcelona, Catalonia/Spain
jordi.torres@bsc.es

Kevin McGuinness and Noel E. O'Connor
Insight Center for Data Analytics
Dublin City University
Dublin, Ireland
kevin.mcguinness@insight-centre.org

1. Introduction

Visual saliency prediction in computer vision aims at estimating the locations in an image that attract the attention of humans. A saliency map is a heatmap that represents the probability of each corresponding pixel in the image to capture human attention. These saliency maps have been used as soft-attention guides for other computer vision tasks, and also directly for user studies in fields like marketing.

This paper explores adversarial training [2] for visual saliency prediction. The *discriminator* distinguishes between samples from the true data distribution and samples produced by the *generator*. In our case, this data distribution corresponds to pairs of real images and their corresponding visual saliency maps.

We show how adversarial training significantly benefits a wide range of visual saliency metrics, without needing to specify a tailored loss function. Our results achieve state-of-the-art performance with a simple deep convolutional network whose parameters are refined with a discriminator.

2. Architecture

The architecture of the presented SalGAN is based on two deep convolutional neural network (DCNN) modules, namely the generator and discriminator, whose combined efforts aim at predicting a visual saliency map for a given input image.

The generator follows an encoder-decoder architecture, where the encoder part includes max pooling layers that decrease the size of the feature maps, while the decoder part uses upsampling layers followed by convolutional filters to construct an output that is the same resolution as the input.

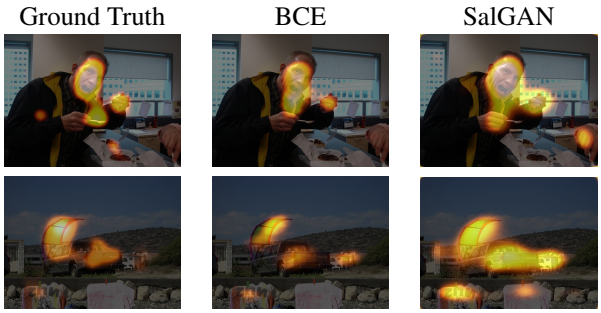

Figure 1: Example of saliency map generation where the proposed system (SalGAN) outperforms a standard binary cross entropy (BCE) prediction model.

The discriminator network is composed of six 3×3 kernel convolutions interspersed with three pooling layers ($\downarrow 2$), and followed by three fully connected layers.

3. Training

The filter weights in SalGAN have been trained over a perceptual loss [5] resulting from combining a content and adversarial loss. The content loss follows a classic approach in which the predicted saliency map is pixel-wise compared with the corresponding one from the ground truth. The adversarial loss depends of the real/synthetic prediction of the discriminator over the generated saliency map.

Generative adversarial networks (GANs) [2] are commonly used to generate images with realistic statistical properties. In our context, the objective is to fit a deterministic function that generates realistic saliency values from images,

*This work was developed while at Microsoft Redmond.

	sAUC ↑	AUC-B ↑	NSS ↑	CC ↑	IG↑
BCE	0.757	0.833	**2.580**	0.772	1.067
GAN	**0.773**	**0.859**	2.560	**0.786**	**1.243**

Table 1: Best results through epochs obtained with non-adversarial (BCE) and adversarial (GAN) training. Saliency maps assessed on SALICON validation.

rather than realistic images from random noise. As such, in our case the input to the generator is not random noise but an image. Second, the knowledge of the image that a saliency map corresponds to is essential for evaluating the quality. We therefore include both the image and the saliency map as inputs to the discriminator. When updating the parameters of the generator function, we found that combining BCE loss and adversarial loss improved the stability and convergence rate of the adversarial training.

The final loss function for the generator during adversarial training can be formulated as:

$$\mathcal{L}_{GAN} = \alpha \cdot \mathcal{L}_{BCE} - \log D(I, \hat{S}), \quad (1)$$

where $D(I, \hat{S})$ is the probability of fooling the discriminator, so that the loss associated to the generator will grow more when the chances of fooling the discriminator are lower. $\mathcal{L}_{BCE}$ is the average of the individual binary cross entropies across all pixels. In our experiments, we used an hyperparameter of $\alpha = 0.05$. During the training of the discriminator, no content loss is available and the sign of the adversarial term is switched.

We train the networks on the SALICON training set. During the adversarial training, we alternate the training of the generator and discriminator after each iteration.

4. Experiments

The presented SalGAN model for visual saliency prediction was assessed and compared from different perspectives. First, the gain of the adversarial training is measured and discussed. Second, the performance of SalGAN is compared to other published works of the current state-of-the-art.

The experiments aimed at finding the best configuration for SalGAN were run using the *train* and *validation* partitions of the SALICON dataset [4], which is the largest dataset available for visual saliency prediction. In addition to SALICON, we also present results on MIT300, the benchmark with the largest amount of submissions.

Table 1 compares validation set accuracy metrics for training with combined GAN and BCE loss versus a BCE alone. The combined GAN/BCE loss shows substantial improvements over BCE for four of five metrics.

SALICON (test)	AUC-B ↑	sAUC ↑	CC ↑	NSS ↑
SalGAN	**0.884**	**0.772**	**0.781**	**2.459**
ML-NET [1]	(0.866)	(0.768)	(0.743)	2.789
SalNet [6]	(0.858)	(0.724)	(0.609)	(1.859)
MIT300	AUC-B ↑	sAUC ↑	CC ↑	NSS ↑
Humans	0.88	0.81	1.0	3.29
SALICON [3]	0.85	0.74	0.74	2.12
SalGAN	**0.81**	**0.72**	**0.73**	**2.04**

Table 2: Comparison of SalGAN with other state-of-the-art solutions on the SALICON (test) and MIT300 benchmarks.

SalGAN is compared in Table 2 to other algorithms from the state-of-the-art on the test partitions of the SALICON and MIT300 benchmarks.

5. Conclusions

In this work we have shown how adversarial training over a deep convolutional neural network can achieve state-of-the-art performance with a simple encoder-decoder architecture. Our experiments showed that adversarial training improved all bar one saliency metric when compared to further training on cross entropy alone. It is worth pointing out that although we use a VGG-16 based encoder-decoder model as the generator in this paper, the proposed GAN training approach is generic and could be applied to improve the performance of other deep saliency models.

Our results can be reproduced with the source code and trained models available at `https://imatge-upc.github.io/saliency-salgan-2017/`.

References

[1] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A deep multi-level network for saliency prediction. In *International Conference on Pattern Recognition (ICPR)*, 2016.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[3] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[4] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

[6] J. Pan, E. Sayrol, X. Giró-i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.