AN ACCURATE SALIENCY PREDICTION METHOD BASED ON GENERATIVE ADVERSARIAL NETWORKS

Bing Yan[1], Haoqian Wang[1,2], Xingzheng Wang[1,2], Yongbing Zhang[1,2]

[1] Key Lab of Broadband Network and Multimedia, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2] Shenzhen Institute of Future Media Technology, Shenzhen 518071, China

ABSTRACT

In this paper, we propose a saliency prediction algorithm utilizing generative adversarial networks. The proposed system contains two parts: saliency network and adversarial networks. The saliency network is the basis for saliency prediction, which calculates an Euclidean cost function on the grayscale values between the predicted saliency map and the ground truth. In order to improve the accuracy of the algorithm, adversarial networks are subsequently utilized to extract the features of input data by coordinating the learning rates of the two sub-networks contained in the networks. Experimental results validate the high accuracy of the proposed approach compared with the state-of-the-art models on three public datasets, SALICON, MIT1003 and Cerf.

Index Terms— Saliency prediction, accuracy, GAN, adversarial networks, saliency network

1. INTRODUCTION

Saliency prediction has the ability to preferentially allocate finite computational resources for subsequent image processing [1]. It aims to understand human visual attention and predict eye movement by three steps for traditional approaches: feature extraction, feature contrast inference, and contrast fusion. Due to the recent publication of large datasets for this significant task, data-driven methods by training the convolutional neural networks (CNN) could be achieved. As a deep learning task, saliency prediction is a concise process to find the best mapping from the input images to saliency maps. Recently a few approaches [2, 3] have been presented which employ CNN to conduct saliency prediction. However, these methods depend on traditional feature extractions more or less. For instance, [2] employs both top-down and bottom-up visual features via three convolutional networks working in parallel, which is substantially a combination of traditional feature extraction strategies. In fact, CNN models can learn visual features automatically, which is more concise compared to the traditional models. Recently, generative adversarial networks (GAN) [4] have shown the powerful ability in feature extraction and feature fusion by dividing better schemes for coordinating the two sub-networks generative net G and discriminative net D. This framework corresponds to a minimax two-player game, which is considerably difficult. However, with the appearance of deep convolutional generative adversarial networks (DCGAN), the GAN models could be trained more readily and stably.

In this paper, we tackle the saliency prediction task based on generative adversarial networks. In the following, the two parts of our model - saliency network and adversarial networks (generative adversarial networks) will be described separately. In regard to saliency network, two challenges need to be solved. Firstly, since saliency prediction is an end-to-end problem whose output image has the same size with the input one, traditional convolutions that would reduce the size of input feature maps have difficulty in solving this challenge. Thus, we present a feasible solution by means of the transposed convolution [5], which is a reverse operation of a convolution and enlarges the size of the input image. Through the transposed convolution, the size of output image could contain correspondingly with the input. Secondly, the previous CNN based approaches[2, 3] mostly define saliency prediction as a binary classification task. However, binary classification could not describe the continuous pixel values of saliency maps. In this paper, we address the challenge by employing regression.

Next, the innovation that we present to accelerate the balance of adversarial networks will be explained. Generative net is designed to make a compromise so as to reduce the difficulty of achieving equilibrium in a non-cooperative game. In other words, generative net could be defeated by discriminative net, which is different from the previous GAN models. After coordinating generative net and discriminative net, the distribution of training data could be learned efficiently and thoroughly by both of the two networks, which promotes the

This work is partially supported by the NSFC fund (61571259, 61531014, 61471213), National High-tech R&D Program of China (863 Program, 2015AA015901), and Shenzhen Fundamental Research fund (JCYJ20160331185006518).

high accuracy of the final result. Parts of the discriminative net are reused as feature extractor which is denoted as saliency network along with the final transposed convolutional layer.

The remainder of this paper is organized as follows. Section 2 presents the recent works about generative adversarial networks and the proposed algorithm. Section 3 reports the quantitative and qualitative experimental results on three datasets. Concluded remarks are made in Section 4.

2. THE PROPOSED ALGORITHM

2.1. DCGAN

Based on the GAN model presented by Goodfellow et al. [4], a more stable adversarial model, deep convolutional generative adversarial networks (DCGAN) [6], has shown its ability in dealing with unsupervised learning problems and tackled a series of assignments such as semantic image inpainting [7] and semi-supervised classification [8]. DCGAN framework trains two networks, a generative net mapping a random vector z to the image space and a discriminative net mapping an input image to a likelihood. The purpose of generative net is to yield real-like images which could bamboozle the discriminative net. Meanwhile, the discriminative net tries to distinguish between fake images and images sampled from data distribution p_{data}. Each network wants to minimize its own cost function, which is to find a Nash equilibrium between two players in a non-cooperative game. The networks are trained by optimizing the cost function [4]:

$$\min_G \max_D V(G,D) = \mathbb{E}_{h\sim p_{data}(h)}[\log D(h)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z))] \quad (1)$$

where h is the sample from the p_{data} distribution; z is randomly generated and lies in some latent space.

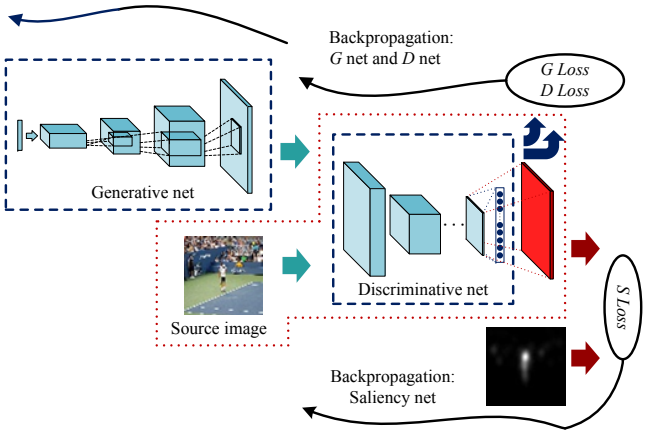


Fig. 1. Our saliency prediction model SGAN.

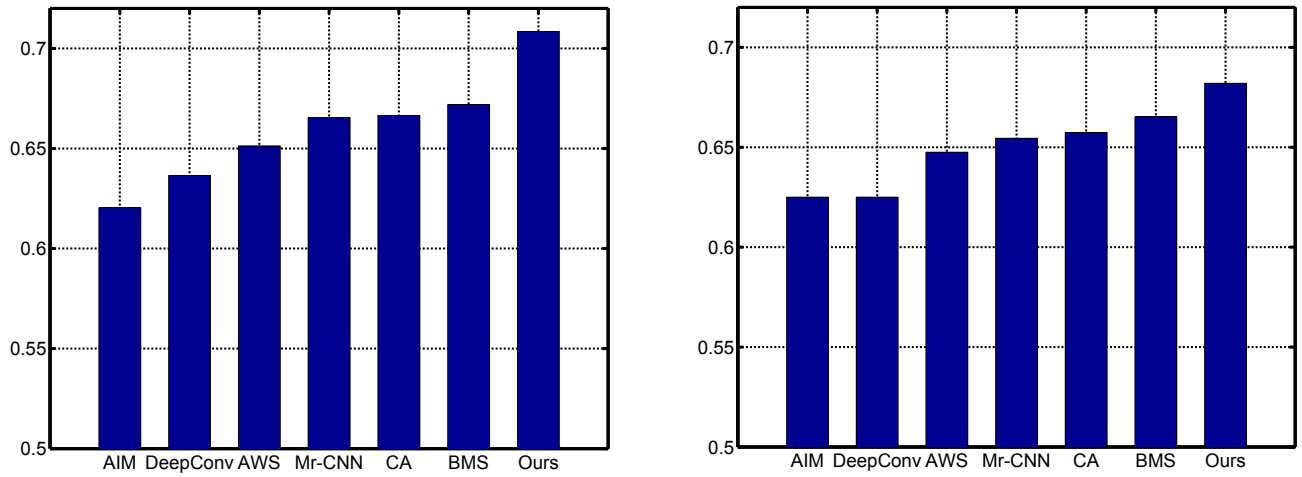
2.2. Saliency prediction based on DCGAN

The diagram of our algorithm is illustrated in Fig.1. To address saliency prediction problem employing GAN model, there are three fundamental issues that need to be tackled.

Achieving end-to-end saliency network. Traditional CNN models with fully connected layers are usually adapted to settle classification problems which has less number of categories; Nevertheless, saliency maps have continuous values between zero and one. It's obvious that the solution of fully connected layers is unsuitable for saliency prediction assignment. We make use of fully convolutional networks where fully connected layers are replaced by transposed convolutional layers [5] for regression. This strategy largely improves the precision of final saliency maps by producing continuous saliency values. Moreover, the transposed convolutional layer is used to connect coarse outputs to dense pixels for the fact that the input and output image have the same scale. The red layer in Fig.1 denotes the predicted saliency map. Since saliency network is an intact model to produce saliency maps. To demonstrate the effectiveness in improving accuracy of adversarial networks, we take advantage of saliency network as noGAN model to obtain saliency maps.

Utilizing the adversarial networks to assist saliency network. GAN model has been applied to image generation and image inpainting which are trained unsupervisedly. However, saliency prediction is a supervised learning task. For saliency network the learning object is naturally the ground truth image while for adversarial networks the learning object is the source image. As shown in Fig.1, the discriminative net contains a series of convolutional layers (they also belong to saliency network) and a fully connected layer. The output is the judgement result - whether the input is a true image (denoted as 1) or not (denoted as 0). The generative net is composed of four transposed convolutions which could learn their own spatial upsampling in order to reconstruct the source image. The saliency network and adversarial networks are connected through the shared convolutional layers between saliency network and discriminative net. During the training process, adversarial networks learn the features of source images thoroughly so that the outcome of feature extraction is excellent, which confirms the high precision of the predicted saliency maps.

Determining the learning rates of both generative net and discriminative net to make training successfully. The learning rates of two sub-networks are the most fundamental parameters in the whole system. By coordinating the two learning rates, the distribution of input data can be learned efficiently, which helps to promote the accuracy. Considering that discriminative net is adapted as feature extraction to yield the final maps, its output - the judgement result is important and should not be fooled by generative net. Meanwhile, the usage of generative net is to analyse and learn the distribution of the source image. It is worth noting that we pay



(a) MIT1003 (b) Cerf

Fig. 2. Comparisons with state-of-the-art in two datasets. We report the shuffled AUC scores of different saliency models under optimal blurring. For each diagram the rightmost blue bar represents our algorithm.

no attention to the generated images from the generative net. Consequently, generative net is designed to be beaten by discriminative net and thus the learning rate of generative net is smaller than that of discriminative net. Experimental details are described in the next section.

In the process of constructing the network architectures, we attempted a series of strategies such as batchnorm and dropout. In general, we found that:

• Batchnorm has little impact on accuracy.

• Dropout regularization increases training time because it slows down the optimization speed of cost function.

• The trends of generative loss and discriminative loss are opposite: one will decrease with the increment in the other.

3. EXPERIMENTS

In this section, the datasets and evaluation metrics are introduced, followed by the implementation details of our model. Then, experimental results are reported for evaluating the proposed approach. We also provide the comparisons with other six competitive saliency models over three datasets.

3.1. Datasets

We conduct evaluation on three eye fixation datasets which capture a broad range of image types. The first dataset is SALICON [9], which is the largest one for saliency prediction. It contains 20000 color images with resolution of 480 $\times$ 640 pixels. Among them 15000 images contain ground truth fixation density maps (GT) which are utilized to train our model. The second dataset, MIT1003 [10], is the most widely used eye fixation dataset containing 1003 images. Before SALICON was released, the dataset is the largest one for model comparison. The third dataset is Cerf [11], which contains 181 images with resolution of 1024 $\times$ 768 pixels.

3.2. Evaluation metrics

Recently the most significant evaluation metric for saliency prediction is shuffled AUC (sAUC) [12] whose false positive rate is approximated by sampling negatives from fixation locations from other images. We utilize sAUC to compare the effectiveness of all approaches. Besides, to further demonstrate the accuracy of adversarial networks, more metrics [12] such as AUC-Judd, CC, and NSS are used. For all the metrics, a higher score means better accuracy of model. We utilize small Gaussian filters with various standard deviation σ to find the optimal blurring of the saliency map for each model. The evaluation scores we report are acquired as the highest scores with blurring.

3.3. Implementation details

In this subsection, we will describe the experimental details including data preprocessing, CNN parameters and settings, and the limitations of our algorithm.

Data preprocessing. We resized the source images to 64 $\times$ 64 and rescaled the pixel values linearly to be in the interval [-1, 1]. Simultaneously, the ground truth fixation density maps are processed to [-1, 1]. We did not subtract the mean pixel value of the source image to zero center them like [13].

CNN parameters and settings. We trained our model on 5,000 labelled images from the SALICON dataset, setting aside 5,000 images for validation and 5,000 for testing. The other two datasets were also tested on the model which was trained on SALICON. In our experiment, when tested using

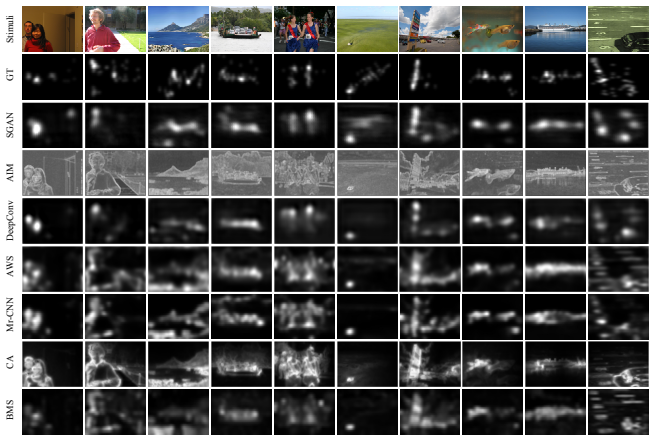


Fig. 3. Qualitative results. We compare our results (SGAN) with six saliency prediction models, namely AIM, DeepConv, AWS, Mr-CNN, CA, BMS. The first row shows the input images from the Cerf (the first 2 columns), MIT1003 (the 3rd to 10th columns). The second row illustrates ground truth maps. The other rows are saliency maps of all the seven models.

MIT1003 dataset, the model trained on SALICON performed even better than directly trained on MIT1003 itself. To optimize the cost function on the grayscale values between the predicted saliency map and ground truth, we utilized the stochastic gradient descent (SGD) with momentum in order to accelerate SGD in the relevant direction and dampen oscillations. The learning rate of saliency network is set as 0.01. When training the adversarial networks, we employed the adaptive moment estimation to optimize both generative and discriminative net loss function. The parameters that are most difficult to select are the two learning rates for adversarial networks. As shown in Section 2.2, the discriminative net is allowed to beat the generative net. Finally, the learning rate for discriminative net is set as 0.0032 and generative net 0.0030. The settings could guarantee the win of discriminative net. Considering the reliability of comparative experiment between our SGAN model and noGAN model, all the parameters of noGAN are the same as SGAN model except for the unavailability of adversarial networks.

Limitations. Considering the stability of the generative net, we set the size of output image to 64×64 as in [6]. In fact, the restriction of size 64×64 appears not only in image generation, but also in videos prediction [14].

3.4. Comparison with state-of-the-art

We evaluated the proposed model by comparison to six competitive models including AIM [15], DeepConv [13], AWS [16], Mr-CNN [2], CA [17], and BMS [18] on three public fixation prediction datasets: SALICON, MIT1003 and Cerf (AIM, AWS, Mr-CNN, CA and BMS model did not publish their saliency maps on SALICON dataset). For the last two datasets, all of the methods participate in comparison. First, we evaluate the sAUC scores over all of the models for quantitative comparison. Fig.2 illustrates that our model outperforms all the methods in MIT1003 and Cerf dataset. BMS achieves the second-best performance. For instance, our model achieves 0.71 sAUC while BMS achieves 0.66 in the popular dataset MIT1003. The excellent results on the two datasets also evince the high generalization of our model since it was trained on SALICON dataset. Second, in Fig.3, we give the qualitative comparison results. For individual cases, our model outstrips any others on images of outdoor scenes, for instance, the five, nine, ten and twelve columns in Fig.3. The CNN based model Mr-CNN could not suppress the background regions smoothly. Consequently, both quantitative and qualitative comparisons show that our algorithm outperforms the other six methods in accuracy.

Table 1. Results in the SALICON dataset. Four different evaluation metrics are used to compare the results of our model, noGAN, and DeepConv.

Evaluation Metric	sAUC	AUC-Judd	CC	NSS
DeepConv	0.60	0.83	0.52	0.41
noGAN	0.61	0.82	0.53	0.50
SGAN	0.64	0.83	0.56	0.51

In order to further demonstrate the accuracy of adversarial networks, we evaluate our model and noGAN by utilizing four metrics on SALICON dataset in Table 1. We provide the evaluation results of DeepConv [13] since it was also trained on SALICON dataset. The evaluation results show that SGAN outperforms noGAN on all the four evaluation metrics especially the sAUC (increased by 3%), which could directly verify the capability of GAN model in promoting accuracy. Moreover, our model defeats DeepConv on four evaluation metrics, which demonstrates the effectiveness of SGAN.

4. CONCLUSIONS

In this paper, we propose an accurate generative adversarial networks based saliency prediction model. Saliency network is an intact model to produce saliency maps. With the help of adversarial networks, feature extraction is more smooth and thorough. Moreover, the fully convolutional networks in saliency network facilitate the continuity and accuracy of pixel values in a saliency map. Compared with the six state-of-the-art methods, the proposed model has achieved highest accuracy. Besides, the performance of our model indicates that adversarial networks could be applied to more than classification. For future work, we will extend the algorithm to semi-supervised saliency prediction since DCGAN is a strong candidate for unsupervised learning.

5. REFERENCES

[1] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[2] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 362–370.

[3] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2798–2805.

[4] 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*, 2014, pp. 2672–2680.

[5] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, 2016.

[7] R. Yeh, C. Chen, T. Lim, M. Hasegawa-Johnson, and M. Do, "Semantic image inpainting with perceptual and contextual losses," in *arXiv preprint arXiv:1607.07539*, 2016.

[8] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2226–2234.

[9] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1072–1080.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 2106–2113.

[11] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in Neural Information Processing Systems*, 2008, pp. 241–248.

[12] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *arXiv preprint arXiv:1604.03605*, 2016.

[13] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. OConnor, "Shallow and deep convolutional networks for saliency prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in Neural Information Processing Systems*, 2016, pp. 613–621.

[15] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, 2005, pp. 155–162.

[16] A. Garcia-Diaz, X.R.Fdez-Vidal, X.M.Pardo, and R.Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012.

[17] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[18] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 153–160.