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.

Introduction

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 [1], 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.

Saliency prediction based on DCGAN

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 [1] for regression. 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.

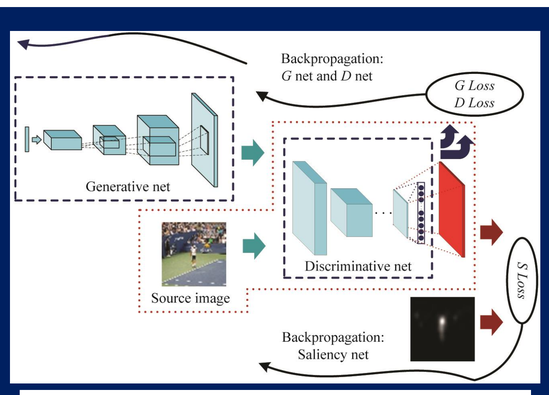

Fig. 1. Our saliency prediction model SGAN.

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 iscriminative 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 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.

Experimental results

Recently the most significant evaluation metric for saliency prediction is shuffled AUC (sAUC) 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 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 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.

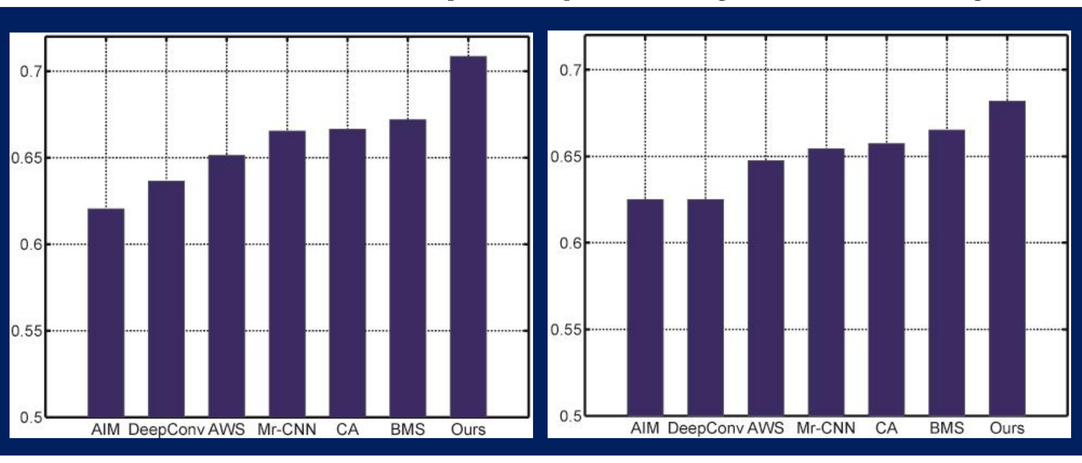

Fig. 2. Comparisons with state-of-the-art in two datasets. We report the shuffled AUC scores of different saliency models under optimal blurring. For each diagram the rightmost blue bar represents our algorithm.

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.

Evaluation Metric	sAUC	AUCJudd	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

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

References

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

[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.

Email:yanbing19930318@126.com