CLIP-SalGAN
A Text-guided Image Saliency Model

Dexuan He, *Student, SEIEEE,*

***Abstract*—In recent years, with the rapid development of computer vision techniques, research on human visual attention has also made progress. In this paper, we propose a novel text-guided saliency prediction model, CLIP-SalGAN, by combining the pre-trained CLIP [1] model and the state-of-the-art saliency prediction model salGAN [2]. We evaluate the quality of the model output using benchmark metrics and conduct experiments to compare and analyze the impact of different text guidance on saliency maps. Our results show that CLIP-SalGAN can generate saliency maps that are both accurate and informative. We believe that our work can help to shed light on the principles of human visual attention.**

***Index Terms*—text-guided saliency prediction, CLIP, salGAN, human visual attention.**

I. Introduction

In the analysis and prediction of visual attention have long been crucial tasks in the fields of computer vision and image processing. The images are generally accompanied by various text descriptions; however, few studies have explored the influence of text descriptions on visual attention, let alone developed visual saliency prediction models considering text guidance. The reference paper conducts a comprehensive study on text-guided image saliency (TIS) from both subjective and objective perspectives. The first TIS database named SJTU-TIS [3], which includes 1200 text-image pairs and the corresponding collected eye-tracking data. Based on these data, the researchers analyze the influence of various text descriptions on visual attention. Then, to facilitate the development of saliency prediction models considering text influence, benchmark experiments evaluate the performance of state-of-the-art saliency models on the dataset.

CLIP (Contrastive Language-Image Pre-Training) [1] is a neural network trained on a variety of (image, text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 and 3. It has been found that CLIP matches the performance of the original ResNet50 on ImageNet "zero-shot" without using any of the original 1.28M labeled examples, overcoming several major challenges in computer vision.

SalGAN, is a deep convolutional neural network for visual saliency prediction trained with adversarial examples. The first stage of the network consists of a generator model whose weights are learned by back-propagation computed from a binary cross entropy (BCE) loss over downsampled versions of the saliency maps. The resulting prediction is processed by a discriminator network trained to solve a binary classification task between the saliency maps generated by the generative

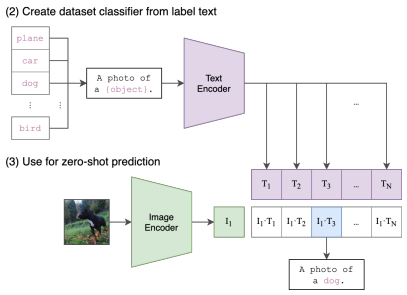

Fig. 1. CLIP Usage1

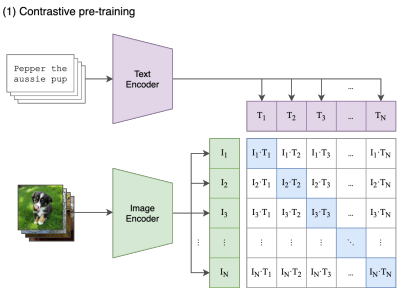

Fig. 2. CLIP Usage2

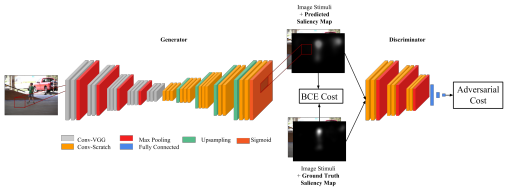

Fig. 3. SalGAN

stage and the ground truth ones. This model show SOTA performance across different metrics.

Based on the SJTU-TIS dataset, we attempt to design a novel text-guided saliency prediction model on the basis of previous image-based saliency prediction models. This model can predict the change of the saliency map under the text bias by comparing the text and the real saliency map. We evaluate the quality of the generated saliency map by metrics such as AUC, sAUC, CC, and NSS. After several attempts, we choose the lightweight CLIP model to encode the text information. We then fuse the features generated by the SalGAN encoder with the text representation. Finally, the saliency map is generated through decoding and adversarial training. We name this model CLIP-SalGAN.We also find that there are researchers on GitHub who have done similar work [4]. We refer to their code to some extent in the specific implementation [5].

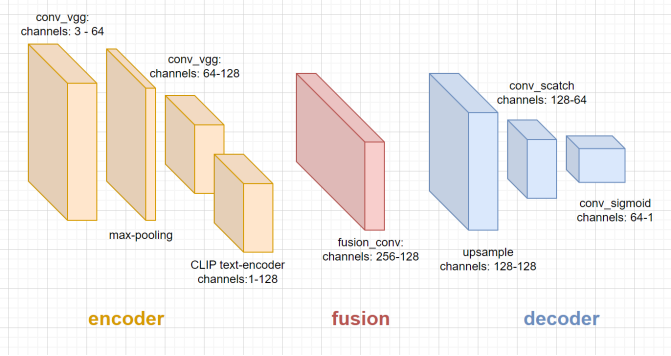

Fig. 4. generator

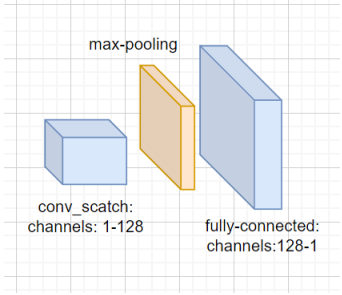

Fig. 5. discriminator

II. Method

In this method section, we present the architecture and implementation details of the CLIP-SalGAN model. The model consists of two main components: a generator and a discriminator. The generator consists of three modules: an encoder (which includes two encoders for images and text), a feature fusion module, and a decoder. The discriminator consists of a single module. The generator and discriminator form a generative adversarial network (GAN), which is optimized using the binary cross-entropy loss function. (In fact, there are two loss functions for the generator and discriminator. For more details, please refer to the SalGAN paper.)

A. Generator

1) Image Encoder: In our work, we use the two-layer VGG network to extract low-level features from the image. These features are then used by the generator to create a saliency map.

$$feat_i = VGG_2 \circ MaxPooling \circ VGG_1(img) \quad (1)$$

2) Text Encoder: For the text, we use the pre-trained CLIP-ViT-B/32 encoder to extract features. In fact, we only use the text encoder of CLIP and ignore the image encoder. This is because the image features are already extracted by a dedicated feature extractor, and re-encoding the image would add unnecessary parameters to the feature fusion layer.

$$feat_t = CLIP_t(text) \quad (2)$$

3) Fusion: In the feature fusion layer, in order to fuse the image features and text features with different shapes, we first transform the text features with shape 512 into shape 2 * 2 * 128, then extract them into shape (128) using a convolutional layer. After that, we concatenate them with the image features, which are also 128 channels. Finally, we extract the fused features from the 256-channel concatenated features using a convolution layer.

$$feat_f = Conv_2(Cat(feat_i, Conv_1(feat_t))) \quad (3)$$

4) Decoder: The decoder decodes the extracted mixed features into the predicted saliency map. Specifically, it passes through an upsampling layer, a convolution layer, and an activation layer.

$$salmap = Sigmoid \circ Conv \circ Upsample(feat_f) \quad (4)$$

B. Discriminator

The discriminator is relatively simple in structure, consisting of only convolutional layers, max-pooling layers, and fully connected layers. It is a binary classifier that plays a supervisory role in the learning of the generator. Here we only give the brief principle of GAN but not to mention all the details.

III. Experiment

A. Dataset

In this experiment, the SJTU-TIS dataset was used, which consists of 600 images. Among them, 300 images contain four types of salient object annotations: no text guidance, text guidance of the overall image, text guidance of salient objects in the image, and text guidance of non-salient objects in the image. The other 300 images contain only two types of annotations: no text guidance and text guidance of the overall image. Therefore, the total number of labeled salient objects is 4 * 300 + 2 * 300 = 1800. Based on this, we combined the data of no text guidance and the other three types of annotations to obtain three datasets, and then added the total dataset to train four types of models: total: 1800, general: 1200, sal: 600, non_sal: 600.

B. Training

Here we plot the loss of the models in the training process and the models converge, especially for the non-sal model.

C. A Test on Typical Example

To investigate the influence of different text guidance on the saliency of the model, we used the total model to predict saliency maps for the same image with text input as sal and non-sal, respectively. The results are shown in the figure. It can be seen that the guidance of text has a significant impact on human attention.

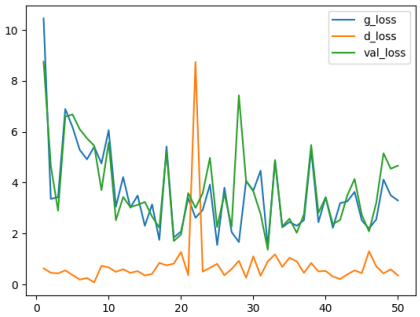

Fig. 6. Total model

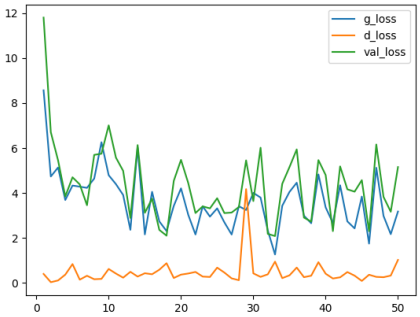

Fig. 7. General model

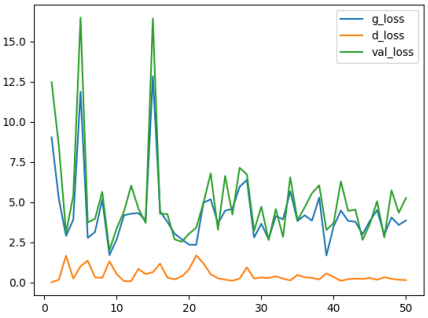

Fig. 8. Sal model

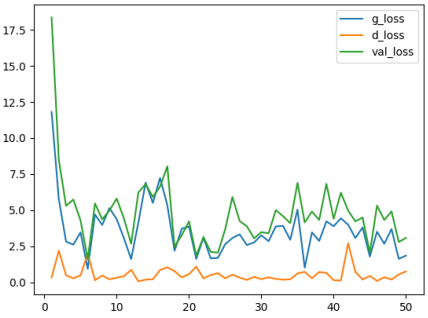

Fig. 9. Non-sal model

Fig. 10. A dog is sitting in the window looking at a car

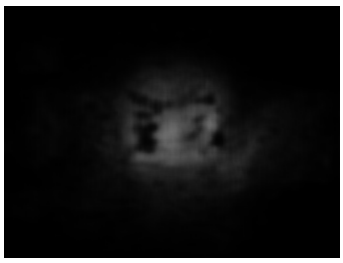

Fig. 11. A dog stares out the window

Fig. 12. Gray cars were parked in the courtyard

D. Metrics scoring

We evaluated the quality of the saliency maps generated by the model using classical saliency map quality metrics. The results are shown below. It can be seen that our model achieves or surpasses the benchmark on the sAUC and NSS metrics, while the AUC and CC are also close to the benchmark.

	Total	General	Sal	Non_sal
AUC	0.684	0.716	0.714	0.682
sAUC	0.597	0.608	0.607	0.591
CC	0.408	0.458	0.423	0.377
NSS	0.224	0.191	0.180	0.186

TABLE I
METRICS

IV. Conclusion

In this experiment, we combined the CLIP model from OpenAI and the state-of-the-art saliency map prediction model SalGAN to build a text-guided saliency map prediction model, CLIP-SalGAN. We measured the prediction quality of CLIP-SalGAN on standard metrics, including sAUC, NSS, AUC, and CC. The results showed that CLIP-SalGAN achieves or surpasses the state-of-the-art performance on all four metrics.

We then used CLIP-SalGAN to investigate the influence of different text guidance on human attention. We found that text guidance can significantly influence the saliency of images.

Overall, our work demonstrates that CLIP-SalGAN is a promising model for text-guided saliency map prediction. It achieves state-of-the-art performance on standard metrics and can be used to investigate the influence of text guidance on human attention.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[2] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. a. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," in *arXiv*, January 2017.

[3] Y. Sun, X. Min, H. Duan, and G. Zhai, "The influence of text-guidance on visual attention," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023, pp. 1–5.

[4] Q. Siyuan, "Clip-salgan: A text-guided saliency model combining gan and clip metrics," 2023. [Online]. Available: https://github.com/gumusserv/CLIP-SalGan

[5] A. Botsa, "Salgan: Visual saliency prediction with generative adversarial networks," 2023. [Online]. Available: https://github.com/batsa003/salgan