Image Saliency Prediction

Kay Strama
Department of Audio-visual technology
Technical University Ilmenau, Germany
kay.strama@tu-ilmenau.de

Omer Zafeer Ali
Faculty of Electrical Engineering and Information Technology
Technical University Ilmenau, Germany
omer-zafeer.ali@tu-ilmenau.de

Huzaifa Ahmad
Department of Media and Technology
Technical University Ilmenau, Germany
huzaifa.ahmad@tu-ilmenau.de

Pooja Surjuse
Department of Media and Technology
Technical University Ilmenau, Germany
pooja.surjuse@tu-ilmenau.de

Momal Khan
Department of Media and Technology
Technical University Ilmenau, Germany
momal.khan@tu-ilmenau.de

***Abstract*—This paper focuses on image saliency prediction. In the area of computational modelling, so-called visual attention models (also called saliency algorithms) can predict the gaze locations of human viewers. The general goal of this paper is to perform a comparison between multiple saliency models. As for the image database there are no fixation maps from a subjective test with people available, saliency maps of the images from the database are computed using a top performing DeepGaze II saliency model [1] and are defined as a ground truth for the further steps of the paper. At first, 100 images are selected for all further analysis described in this paper. The selected images do represent a broad range of visual complexity. The complexity is assessed with a simple measurement of how many white contents the respective ground truth saliency map is containing. In the next step, three saliency models are applied on the selected images. For this purpose, the models SAM, SalGAN and ML-Net are used, which are available as open source. After computing the respective saliency maps, the results of the used models are compared to the ground truth saliency maps computed with the DeepGaze II algorithm. For doing so, several popular saliency map performance metrics like Kullback-Leibler Divergence, AUC-Judd and SIM are applied to the computed saliency maps.**

***Index Terms*—saliency prediction, neural network, machine learning**

I. Introduction

Saliency prediction of an image means predicting the spatial location of the image contents that generate the highest attention in the human viewers. Saliency prediction is essential for several machine-based tasks, for instance recognition of objects. Traditionally data for visual saliency was gathered by eye-tracker and now by using webcams and mouse clicks [2].

Saliency predictions are generated with neural networks. The result is a saliency prediction map. A heat map represented as a grey-scale image is generated as a result that shows the probability of every pixel in an image to get human focus. Higher salient areas are shown as white and lesser salient areas in black [1].

Usually, saliency prediction methods used features like color, contrast, texture or faces and other semantic features as cues. But those features are not sufficient to define all factors for the prediction of the visual saliency maps. The introduction of Convoluted Neural Networks (CNN) for the use of saliency prediction algorithm, allows advancements in architectures and the use of large data sets [3]. There are several saliency prediction models. Every model has its own characteristics. Here, we will apply a few already available saliency models on a data set of 100 images selected from 200 images and compare their results with the ground truth saliency maps computed by the DeepGaze II algorithm. For this purpose, metrics typically used for saliency map evaluation are used.

This paper gives a general overview about saliency map prediction algorithms and its performances against the DeepGaze II algorithm. To achieve that, this paper first explores convoluted neural networks (CNN) and its use in the prediction of saliency maps. Afterwards, three different saliency map prediction algorithms are introduced, and their functionality are explained generally. For comparison, the paper first presents the image selection criteria for the 100 images and the chosen saliency comparison metrics, then the results are analyzed and conclusion about the findings is presented.

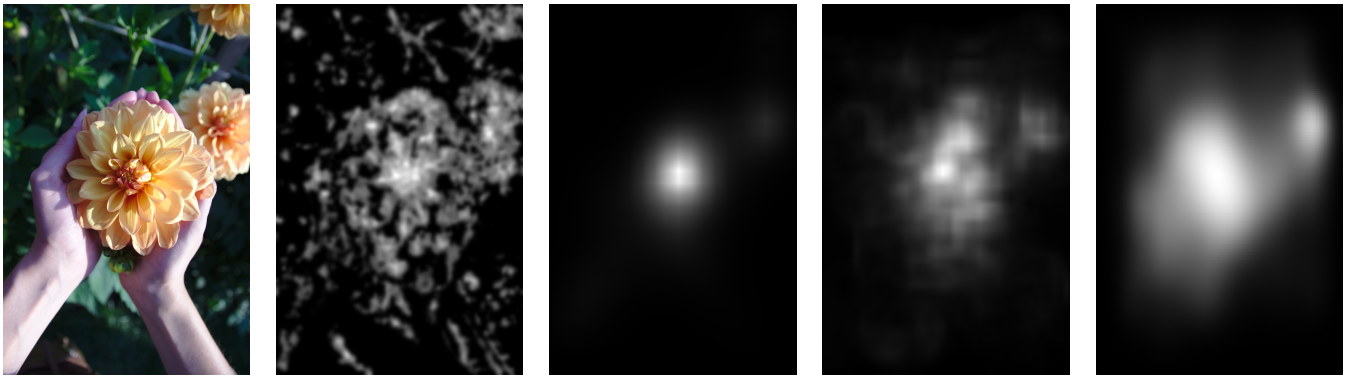
II. Saliency map algorithms

There are several open algorithms available to predict the salient features of an image. Neural network algorithms need to be trained to the images. Algorithms use set of training and test data. Since SALICON [4] has established itself as a standard training data set in the scientific community, the prediction algorithm also use them for training. Training exceeds the scope of this paper, the algorithms were used with pre-trained weights.

Here, in this section we will talk about three algorithms that we used to generate the saliency prediction maps. A brief discussion about performance and architecture forms the main part of the following chapters. In this paper used algorithms are SalGAN, Saliency Attentive Method, and Multi-level Network for Saliency Prediction.

A. SalGAN

SalGAN architecture evaluates the saliency map of an input image using a deep convolution neural network (DCNN) method. Its architecture comprises two CNN's, a generator, SalGAN (to produce saliency maps), and a discriminator to



(a) (b) (c) (d) (e)

Fig. 1: (a) The original image. (b) The ground truth map from DeepGaze II. (c) SAM-ResNet saliency map. (d) ML-Net saliency map. (e) SalGAN saliency map.

identify the differences between the real saliency maps and the synthesized saliency maps.

The generator is made of a convoluted encoder and a decoder model. The Encoder has a convolution layer associated with a max-pool layer that helps in reducing the size of the maps. Same is the case with the decoder layer, it has an up sampled layer followed by a convolution layer to generate the saliency map back to the size of the input image. Furthermore, the discriminator is made of 6 convolution layers linked with a kernel of size 3 x 3 that identifies the difference between the predicted maps. The SalGAN architecture utilizes adversarial and content loss. Content loss is used to calculate the similarity between the synthesized saliency map and the ground truth saliency maps based on images per pixel. The main function of adversarial loss is to compute how precisely the discriminator has identified that a synthesized saliency map is fake or real [2].

For this project, the SalGAN model is trained with adversarial examples. The Initial step includes a generator model in which weights are learned by back propagation. Then it calculates with binary cross entropy (BCE) the down sampled variants of the saliency maps. The model is then refined with a discriminator network, that is trained to sort a binary classification task between saliency maps designed by the SalGAN model and the ground truth saliency maps generated by DeepGaze II. Results of our experiments explain the impact of adversarial training on its performance across different metrics when linked with commonly used loss function like BCE [2].

B. Saliency Attentive Model

The Saliency Attentive Model (SAM) uses an Attentive Convolutional Long Short-Term Memory network (Attentive ConvLSTM). Machine attention is a method to sequentially compute different parts of an input. The algorithm can use two different CNNs, SAM-VGG and SAM-ResNet. Images are scanned horizontally and vertically and use global and scene context to generate image saliency [3]. Following, some parts of the SAM model architecture are explained.

One neuron has a weight and an activation function, which ensures that the output will be between 0 and 1, also for high weight values. To train neural networks, back propagation is a widely used method.

Before training the weights are initialized with random values. Then training starts an iteration with the following steps: Training (an image in our case) is used as input for the network. Next, the output is then compared to the real values and the weights are adjusted to minimize the error. This step is repeated, until the deviation between output and input is lower than the minimum deviation defined before as termination criterion.

When using multi layered networks and neurons with their own activation functions, the weight correction is multiplied with the error for every layer. This means, for outputs between 0 and 1, bigger values disappear during training over longer chains. To prevent this loss of information, neurons from Long Short-Term Memory (LSTM) networks use so-called gates, to let values trough or forget them.

The SAM algorithm further modified the Convolutional LSTM model to refine saliency features over an image instead of a temporal sequence. This is achieved by substituting dot products with convolutional operations. Furthermore, SAM uses the sequential nature of the LSTM model to process saliency features in an iterative way [3].

When observing images, humans mostly focus on the center of the image. This has to do with the fact, that photographers put the interesting object in the center or near the center. Also, if there are no interesting sections of an image, humans tend to focus to the center of the image. Therefore, SAM gives more priority to the center of the image. The priors can also be learned by the model. To simplify the calculation, a 2D Gaussian function is used [3].

C. Multi-Level Network for Saliency Prediction

The Multi-Level Net (ML-Net) saliency map prediction algorithm consists of three main parts. First a CNN for the feature extraction, second an encoding network to weight low- and high-level features and third a prior learning network [5].

The first part is the feature extraction network. The fully convolutional network with 13 layers is used to produce feature maps for the encoding network. It is based on the VGG-16 layers model. The disadvantage of a standard CNN architecture is reducing of the size of the feature maps at higher levels. The VGG-16 model would result in an output saliency map rescaled by a factor of 32 compared to the input image. The modified network removes the last pooling stage and decreases the stride. So, the output saliency map is only rescaled by a factor of 8 [5].

The second part of ML-Net is the encoding network. Feature maps are extracted at convolutional layers 3, 4 and 5. They are combined to a tensor and fed into a convolutional layer with a 3 x 3 kernel and learns 64 saliency specific features [5].

The third part is the prior learning component. Here, instead of using pre-defined priors, the network learns its own custom prior [5].

III. Selection of images

To reduce the time and complexity for generating the saliency maps we selected 100 images from the given data set of 199 images. The images were chosen based on their complexity. To determine the complexity of an image, several methods exist. Following are some methods that can be used to calculate complexity.

- Number of zeroes in ground truth saliency maps. A lower number of zeroes in the saliency map corresponds to large white areas and small black areas. This means there are more salient areas and high image complexity [6].
- Spatial information (SI) with Sobel filter responses. From the distribution of the SI of an image, the Mean, Square-Root-Mean and Standard deviation can be calculated [7].
- Histogram distribution. An image showing wide range of distribution in histogram, i.e. most pixels carrying value other than zero and one corresponds to an image with high complexity. Having peaks in distribution only at zero or one shows a simple image.
- JPEG Compression. It uses discrete cosine transform (DCT) to remove redundancies in an image information and reduce its size. A complex image does not repeat linear structures as much and DCT coefficients not redundant, hence less information to remove. So, comparing bitmap and JPEG file sizes for a complex image show little difference while for simple image it shows noticeable change in file sizes.

Though histogram distribution and JPEG compression serve the purpose, but are laborious tasks in terms of image selection. For instance, plotting 200 histograms and comparing them for complexity or, applying JPEG compression to 200 images and comparing the resulted file sizes to original bitmap sizes of the images is a complex and more time consuming. Hence, for this paper, we used relatively an easier method i.e., the number of non zero pixels in the ground truth saliency map as a simple selection criterion to choose the images for the analysis.

IV. Analysis of the different algorithms

For this paper the saliency maps for selected images of section III were computed. The used algorithms from section II are SalGAN, Sam-ResNet and ML-Net. Then the saliency maps are compared to the ground truth saliency maps computed with the DeepGaze II algorithm. In figure 1 is one original image, the ground truth saliency map and its associated predicted saliency maps.

There are several methods to compare saliency maps from different algorithms. In the following, we will present a short overview of the three methods chosen for this paper.

A. Kullback-Leibler Divergence

The Kullback-Leibler Divergence (KL) measures the difference between two probability distributions. For saliency maps this means, the saliency prediction and the ground truth are interpreted as distributions. KL is a dissimilarity metric, so with a lower score the prediction is closer to the ground truth and with a higher score the prediction is more dissimilar to the ground truth [8].

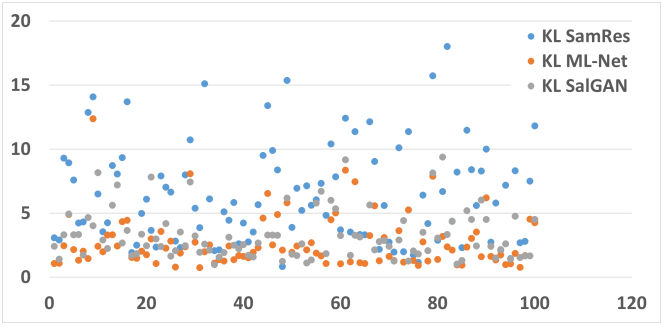


Fig. 2: KL Divergence of SAM-ResNet and ML-Net

The KL Divergence for the three saliency prediction algorithm can be seen in figure 2. There can be seen, that the divergences for the with ML-Net and SalGAN predicted maps are generally lower than for the Sam-ResNet predicted maps.

B. Similarity

The Similarity (SIM) measures the similarity between two distributions, viewed as histograms. It calculates the sum of the minimum values at each pixel. A SIM score of 1 means the prediction and ground truth are the same. A score of 0 means, there is no overlap between the two saliency maps [8].

The comparison with SIM gives similar results as with the KL Divergence. The results for the images can be seen in figure 3. The saliency maps generated with Ml-Net and SalGAN are closer to the ground truth than the saliency maps from the SAM-ResNet algorithm.

C. Area under ROC Curve

Area under ROC Curve (AUC) measures if a pixel is fixated or not by measuring true and false positives under a level threshold. Because of the nature of images, that important parts are mostly shown in the center, most saliency prediction algorithms with a center bias can determine fixations independent

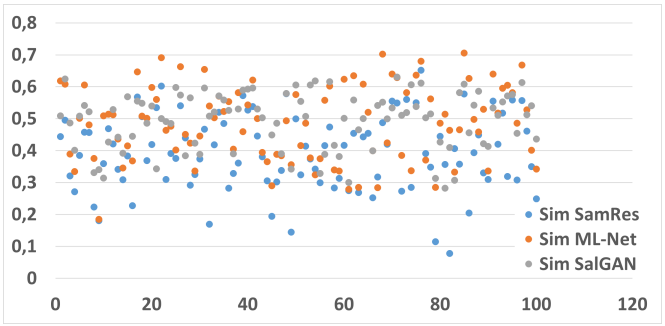


Fig. 3: Similarity of SAM-ResNet and ML-Net

of the image content. AUC has the effect of sampling negatives predominantly from the center of an image. This means, that a saliency prediction model with a center bias achieves an AUC score of 0.5. AUC is effectively penalizing models with center bias [8].

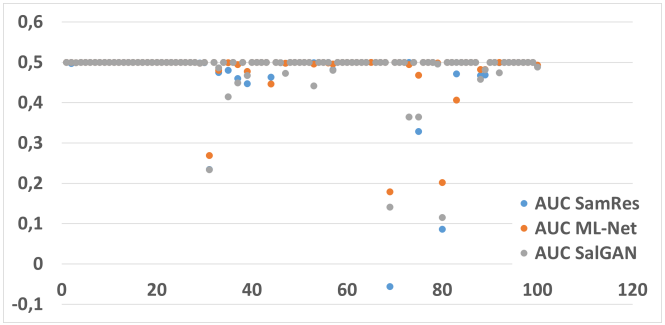


Fig. 4: AUC Judd of SAM-ResNet and ML-Net

The AUC version we used in this paper is AUC-Judd. The results for the three saliency prediction algorithms can be seen in figure 4. AUC gives most of the prediction maps a value of 0.5. That is to be expected, because all the used prediction algorithms use a center bias.

Also, most of the sample images are center focused. So, when the prediction algorithm successfully predicts the main salient area in the center of the image, AUC gives them a score of 0.5. But, in the case of a couple out liners in figure 4, the most salient areas are outside the center. For these images, the algorithm achieves a low AUC score.

That all algorithms use a center bias, comes from the fact, that photographers mostly put the object of interest in the center or near the center of an image. So, saliency prediction models are trained towards a center bias.

V. Conclusion

The findings from the comparison with KL and SIM are shown in table I. In the table the mean and standard deviation of the prediction maps is shown. As mentioned in section IV-A the KL Divergence of SAM-ResNet is on average worse than the other two algorithms. While ML-Net and SalGAN compare quite close together.

On the comparison with SIM all three algorithms are close together. Where SalGAN has a lower standard deviation than Sam-ResNet and ML-Net.

TABLE I: Comparison of saliency prediction algorithms

Comparison	*Sam-ResNet*	*ML-Net*	*SalGAN*
Mean KL	6,42	2,69	3,27
STD KL	3,81	1,97	1,90
Mean Sim	0,39	0,48	0,49
STD Sim	0,12	0,12	0,09

For an evaluation of saliency prediction algorithms, the area of application must be considered. For example, a gaze prediction, where only the general area of the interesting object is relevant, the saliency maps generated by Sam-ResNet, ML-Net and SalGAN would be sufficient.

Another implementation would be the use of saliency maps for the encoding of images. Because low salient areas are generally of less importance to the human viewer, a higher compression can be used on them. For this application, the three used algorithms would not be as useful as the state-of-the-art DeepGaze II would be because of their divergence and difference relative to ground truth maps, also shown in table I.

Also, for the evaluation of saliency maps the used metrics are important. KL divergence and SIM are generally less useful in determining the use of saliency maps for compression. Whereas AUC is more useful because of its effectively penalizing prediction algorithms with a center bias.

References

[1] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Deepgaze ii: Reading fixations from deep features trained on object recognition," 2016.

[2] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. G. i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," 2018.

[3] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.

[4] C. Q. Zhao, "Salicon image dataset," online. [Online]. Available: http://salicon.net/

[5] 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.

[6] M. Perreira Da Silva, V. Courboulay, and P. Estraillier, "IMAGE COMPLEXITY MEASURE BASED ON VISUAL ATTENTION," in *IEEE International Conference on Image Processing - ICIP*, Bruxelles, Belgium, Sep. 2011, pp. 3281–3284. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00617725

[7] *Subjective video quality assessment methods for multimedia applications*, ITU-T Recommendation P.910, nov 2021.

[8] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" 2017.