UNDERSTANDING SALIENCY PREDICTION WITH DEEP CONVOLUTIONAL NEURAL NETWORKS AND PSYCHOPHYSICAL MODELS

Qiang Li
Image Processing Lab, University of Valencia

ABSTRACT

Convolutional neural networks (CNNs) have achieved great success in natural image saliency prediction. The primary goal of this study is to investigate the performance of saliency prediction in CNN with psychophysical synthetic images. Is it still as decent as natural images in terms of performance? In the meantime, it can be used to investigate the relationship between CNNs and human vision, mainly low-level vision functions. Are CNNs, on the other hand, exact replicas of human visual function? This study used CNNs and Fourier and Spectral models inspired by low-level vision systems to investigate saliency prediction on psychophysical synthetic images rather than natural images. According to our findings, saliency prediction models inspired by Fourier and Spectral theory outperformed current pre-trained deep neural networks on psychophysical images. However, psychophysical models were more unstable in noise than pre-trained deep neural networks. Meanwhile, we suggested that investigating CNNs with psychophysical methods could benefit visual neuroscience and artificial neural network studies.

Index Terms— Convolutional Neural Networks, Vision Neuroscience, Saliency Prediction, Psychophysical

1. INTRODUCTION

The attention function was first introduced into artificial neural networks, and it has achieved great success in computer vision tasks, such as image classification [1], few-show learning [2], and so on. There is no denying CNN’s victory in both academics and industry. Moreover, Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train [1]. The vital message transmitted from all of these is Attention is All you Need. Although there are some arguments about attention in artificial neural networks, it’s beyond the scope of this paper.

The natural world includes heavy redundant information, and the human visual system can’t process all this information under the limited information bottleneck of the vision system. However, related research about attention with physiology methods has a long history, with a specific focus on the vision system [3]. Based on three-stage vision information processing models [4], the input information encoded in the retina with 20MB/second [5], and transmitted information into the selection phase. Here is where attention happens, and it can remove irrelevant messages. After the selection step, information is compressed into 1MB/second then conveyed left information into the primary visual cortex (V1) and leaving only 40 bits/second of information in V1 [6, 4]. A more critical and classical principle that connects to the visual attention mechanism is Efficient Coding Principle [7]. After that, a larger body of psychophysical experiments investigated vision attention. Meanwhile, different specifically designed computational attention models were proposed based on physiology experiments, such as the earliest Itti models for visual saliency prediction [8].

Natural image saliency prediction has been a massive success in the deep learning age [9], such as ML_Net [10], DeepGazeII [11]. However, there are fewer studies that examine saliency prediction on psychophysical images using CNNs and custom-designed psychophysical models and the relationship between CNNs and psychophysical saliency prediction models. This study is limited to predicting saliency for psychophysically oriented images rather than natural images. Investigating artificial neural networks and human vision systems, specifically low-level vision functions, is central to this study. The connections between CNNs and low-level vision functions in humans are not fully understood at the moment. We find it puzzling that CNNs, on the other hand, rely entirely on low-level human visual processes. Additionally, we want to comprehend and test CNNs using psychophysical approaches and then improve their general and transferable learning abilities across various tasks.

2. RELATED WORK

The general saliency prediction on natural images has demonstrated tremendous success in the deep learning era [9]. Predicting visual saliency has become a hot topic in machine vision and visual neuroscience. One way of understanding black-box CNNs is visualization, and attention maps become a way of understanding the features of information process-
ing in CNNs. The classical pre-trained with different depths of general CNNs were employed to investigate their attention maps, such as Alexnet [12], GoogLeNet [13], Inceptionv3 [14], Densenet201 [15], respectively. There are several top-ranking CNNs dedicated to natural image saliency prediction, for example, DeepGazeII based on the MIT/Tuebingen Saliency Benchmark evaluation. Furthermore, three of the highest-ranked psychophysical saliency prediction models were introduced in this study, and their design is entirely based on the functions of the human low-level vision system [16]. The first one is the HTF model, which is a bottom-up paradigm for visual saliency prediction based on frequency domain [17]. The second model is Incremental Coding Length (ICL), which is designed based on information-theoretic principles [18]. The model assumes that attention regions in images may elicit entropy gain in the perception state and then attract high energy. The last one is the spectral-oriented saliency prediction model, DCTS, which predicts image saliency based on image signature [19]. The following section will examine saliency prediction using all of the aforementioned models on psychophysical images.

3. METHODS

3.1. Dataset: Psychophysical images

A SID4VAM dataset was used in this study, composed of 230 synthetic images with known salient regions (see Fig. 1). Images were generated with 15 distinct low-level features (e.g., orientation, brightness, color, size) with a target-distractor popout type of synthetic pattern. We have used free-viewing and visual search task instructions and 7 feature contrasts for each feature category [20, 21].

3.2. Pre-trained CNNs and psychophysical models on psychophysical images

As we mentioned before, here we investigated pre-trained neural networks, such as AlexNet, GoogLeNet, Inceptionv3, Densenet201, and DeepGazeII to visualize saliency maps with psychophysical synthetic images. Meanwhile, we also used some classical Fourier and Spectral saliency prediction models in which inspired by human low-level vision functions, such as HFT, ICL, and DCTS. In pre-trained neural networks, saliency map estimated with Gradient-weighted Class Activation Mapping (Grad-CAM) that is a simple technique for understanding which parts of an image are most important for a deep network’s recognition [22]. Use Grad-CAM to gain a high-level understanding of what image features a network uses to make a particular classification or other tasks. The Grad-CAM map for a convolutional layer with k feature maps (channels), \( A^k_{i,j} \), could be estimated with

\[
S = \text{ReLU} \left( \sum_k \frac{1}{N} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{i,j}} A^k_{i,j} \right),
\]

where \( y^c \) is output, representing the score for class \( c \), \( i, j \) indexes the pixels, \( N \) is the total number of pixels in the feature map. The rectified linear unit (ReLU) activation ensures you get only the features that have a positive contribution to the class of interest. The output is therefore a heatmap for the specified class. The HFT, ICL and DCTS were designed based on Fourier-Spectral and information-theoretic principles, respectively. The fundamental information about models refer to Table. 1.

| Network     | Depth | Parameters (Millions) | Image Input Size |
|-------------|-------|-----------------------|-----------------|
| Alexnet     | 8     | 61.0                  | 227-by-227      |
| Googlenet   | 22    | 7.0                   | 224-by-224      |
| Inceptionv3 | 48    | 23.9                  | 299-by-299      |
| Densenet201 | 201   | 20.0                  | 224-by-224      |
| DeepGazeII  | 19+   | -                     | -               |
| HFT         | -     | -                     | -               |
| ICL         | -     | -                     | -               |
| DCTS        | -     | -                     | -               |

**Table 1: Models information.** The pretrained networks and psychophysical models and some of their properties (e.g., depth, parameters and image input size).

3.3. Pre-trained CNNs and psychophysical models on degraded psychophysical images

Here, we want to investigate the stability of various saliency prediction models under noise interference (See Fig. 2). We
added salt-and-pepper noise with a density of 0.3 to each psychophysical image. After that, we fed degraded images into various models to check how noise affects CNNs and psychophysical models on saliency prediction accuracy.

Fig. 2: Selected part of degraded psychophysical images
The top row shows degraded psychophysical stimuli images, and the second row shows saliency prediction maps from humans.

3.4. Evaluation metrics
Accuracy (ACC), sensitivity (SEN), specificity (SPE), precision (PRE), and F1 score were adopted as statistical measures of the saliency prediction performance. ACC is defined as

\[ \text{ACC} = \frac{TP + TN}{P + N} \]  

where TP, TN, P, and N refer to true positive, true negative, condition positive, and condition negative, respectively. In this study, TP is the number of positive samples correctly predicted as positive, TN the number of correctly predicted negative samples, P the total number of positive samples, and N the total number of negative samples. SEN is defined as\[ \text{SEN} = \frac{TP}{P} \], SPE is defined as\[ \text{SPE} = \frac{TN}{N} \], PRE is calculated as\[ \text{PRE} = \frac{TP}{TP + FP} \], and F1 score is the harmonic mean of precision and sensitivity and determined as\[ F_1 = \frac{2TP}{2TP + FP + FN} \], where FP is the number of negative samples incorrectly predicted as positive, and FN the number of positive samples incorrectly predicted as negative.

4. RESULTS
The visualization of saliency prediction with the above models is present in Fig. 3. Here we plotted a saliency prediction map with different models. We observed that DCTS and HFT have prediction saliency that closely matches human levels (baseline, second row). However, pre-trained CNNs have a worse prediction compared with psychophysical models. To statistically analyze the performance of saliency prediction with different models, we measured the receiver operator characteristic (ROC) curve, referring to Fig. 5. The statistics result also shows psychophysical models have better prediction accuracy than pre-trained CNNs. The reasons will be explained in the next section. Even more interesting, we explored how saliency prediction interferes with different models under noise. In Fig. 4, we found that pre-trained (Alexnet, Googlenet, Inceptionv3, and Densenet201) and specifically designed (DeepGazeII) neural networks did not have a large effect when adding noise. However, psychophysical models (HFT, ICL, and DCTS) have a large effect by adding noise. The above result makes sense because pre-trained neural networks already have fixed parameters and weights with natural images. Still, psychophysical models need to feed degraded images into models without learning to produce saliency maps. However, we can’t evaluate degraded images’ saliency performance because we don’t have human information on degraded images. Therefore, we only report the clean image saliency prediction metric in Fig. 5 and Table 2. Therefore, future work could collect human saliency prediction data on the degraded image, and it would be fascinating.

Fig. 3: Predicting saliency using pre-trained convolutional neural networks and biologically inspired saliency prediction models on selected images. The top row shows psychophysical stimuli images, followed by the second row, which shows the saliency prediction map from humans. The remaining rows show saliency prediction maps from different models.
Saliency predictions were made on selected degraded images using pre-trained convolutional neural networks and psychophysical models. The top four rows show degraded saliency predictions from pre-trained CNNs. The remaining rows show saliency prediction maps from psychophysical models.

Table 2: AUC values

| Models      | Alexnet | Googlenet | Inceptionv3 | Densenet201 | DeepGazeII | ICL | DCTS | HFT |
|-------------|---------|-----------|-------------|-------------|------------|-----|------|-----|
| AUC         | 0.5216  | 0.5493    | 0.5283      | 0.5281      | 0.6153     | 0.7569 | 0.7862 | 0.8297 |

Fig. 5: AUC - ROC curve. The performance of saliency prediction with various depths of pre-trained convolution neural networks and biologically inspired saliency prediction models.

As mentioned above, some difficulties arise when psychophysical images are applied to CNNs. On the one hand, there are not enough psychophysical images used to train deep CNNs. On the other hand, CNN architecture models high-level human visual systems, but do they model low-level visual functions, such as color, orientation, brightness reduction, and adaptation? Finally, saliency prediction on degradation distract tasks could be extended with more diversity degradation on both clean natural images and psychophysical synthetic images to check the stability of saliency prediction. As we aforementioned, we also need to collect human saliency prediction information on degraded images as a baseline to evaluate the model’s performance. It would also be an exciting direction to take in the future.

Last but not least, we should consider psychophysical methods when examining the properties of artificial neural networks. Numerous psychophysical studies on visual attention mechanisms have been conducted by visual neuroscientists, and they have significantly increased our knowledge in those fields. It can improve the generality and stability of artificial neural networks, but there is insufficient research in this area. Nonetheless, it will be more interesting to incorporate additional psychophysical studies into the artificial neural network because they will deepen our understanding of the human visual system and artificial neural networks mechanisms.
6. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[2] Yuxin Peng, Xiangteng He, and Junjie Zhao, “Object-part attention model for fine-grained image classification,” IEEE Trans. Image Process., vol. 27, no. 3, pp. 1487–1500, 2018.

[3] Marisa Carrasco, “Visual attention: The past 25 years,” Vision Research, vol. 51, no. 13, pp. 1484–1525, 2011, Vision Research 50th Anniversary Issue: Part 2.

[4] Zhaoping Li, “Theoretical understanding of the early visual processes by data compression and data selection,” Network: Computation in Neural Systems, vol. 17, no. 4, pp. 301–334, 2006, PMID: 17283516.

[5] D. Kelly, “Information capacity of a single retinal channel,” IRE Transactions on Information Theory, vol. 8, no. 3, pp. 221–226, 1962.

[6] G. Sziklai, “Some studies in the speed of visual perception,” IRE Transactions on Information Theory, vol. 2, no. 3, pp. 125–128, 1956.

[7] Horace Barlow, “Possible principles underlying the transformations of sensory messages,” Sensory Communication, vol. 1, 01 1961.

[8] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] Ali Borji, “Saliency prediction in the deep learning era: Successes and limitations,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, pp. 679–700, 2021.

[10] Marcello Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, “A deep multi-level network for saliency prediction,” in 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 3488–3493.

[11] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge, “Deepgaze ii: Reading fixations from deep features trained on object recognition,” ArXiv, vol. abs/1610.01563, 2016.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, “Going deeper with convolutions,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, 06 2015.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, “Rethinking the inception architecture for computer vision,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), abs/1512.00567, 2015.

[15] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger, “Densely connected convolutional networks,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269, 2017.

[16] Qiang Li, “A psychophysical oriented saliency map prediction model,” arXiv preprint arXiv:2011.04076, 2020.

[17] Jian Li, Martin Levine, Xiangjing An, Xin Xu, and Hangen He, “Visual saliency based on scale-space analysis in the frequency domain,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pp. 996–1010, 11 2012.

[18] Xiaodi Hou and Liqing Zhang, “Dynamic visual attention: Searching for coding length increments,” in Adv. Neural Inf. Process. Syst, 01 2008, vol. 21, pp. 681–688.

[19] Xiaodi Hou, Jonathan Harel, and Christof Koch, “Image signature: Highlighting sparse salient regions,” IEEE transactions on pattern analysis and machine intelligence, vol. 34, 07 2011.

[20] David Berga, Xosé Ramón Fdez-Vidal, Xavier Otazu, and Xose Manuel Pardo, “Sid4vam: A benchmark dataset with synthetic images for visual attention modeling,” IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8788–8797, 2019.

[21] David Berga, Xosé R. Fdez-Vidal, Xavier Otazu, Víctor Leborán, and Xosé M. Pardo, “Psychophysical evaluation of individual low-level feature influences on visual attention,” Vision Research, vol. 154, pp. 60–79, 2019.

[22] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in ICCV. 2017, pp. 618–626, IEEE Computer Society.