Exploring Language Prior for Mode-Sensitive Visual Attention Modeling

Xiaoshuai Sun[1], Xuying Zhang[1], Liujuan Cao[1*], Yongjian Wu[2], Feiyue Huang[2], Rongrong Ji[1]
[1]Media Analytics and Computing Lab, School of Informatics, Xiamen University, China, [2]Youtu Lab, Tencent
xssun@xmu.edu.cn,zhangxuying@stu.xmu.edu.cn,caoliujuan@xmu.edu.cn
{littlekenwu,garyhuang}@tencent.com,rrji@xmu.edu.cn

ABSTRACT

Modeling human visual attention mechanism is a fundamental problem for the understanding of human vision, which has also been demonstrated as an important module for various multimedia applications such as image captioning and visual question answering. In this paper, we propose a new probabilistic framework for attention, and introduce the concept of *mode* to model the flexibility and adaptability of attention modulation in complex environments. We characterize the correlations between the visual input, the activated mode, the saliency and the spatial allocation of attention via a graphical model representation, based on which we explore the lingual guidance from captioning data for the implementation of a mode-sensitive attention (MSA) model. The proposed framework explicitly justifies the usage of center bias for fixation prediction and can convert an arbitrary learning-based backbone attention model to a more robust multi-mode version. Experimental results on the York120, MIT1003 and PASCAL datasets demonstrate the effectiveness of the proposed method.

CCS CONCEPTS

• **Computing methodologies → Probabilistic reasoning**.

KEYWORDS

Language Prior; Caption Semantics; Multi-Mode Attention

ACM Reference Format:
Xiaoshuai Sun[1], Xuying Zhang[1], Liujuan Cao[1*], Yongjian Wu[2], Feiyue Huang[2], Rongrong Ji[1]. 2020. Exploring Language Prior for Mode-Sensitive Visual Attention Modeling. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3414008

1 INTRODUCTION

Computational models of attention can benefit not only the low-level computer vision tasks like object detection [24] and segmentation [31] but also the more advanced AI applications such as image captioning [32] and visual question answering [1].

As one of the most representative attention-driven applications, image captioning aims at generating a natural language description for a given image. Most state-of-art captioning models rely on a CNN-based visual backbone (*e.g.*VggNet [25]) and an RNN-based language module (*e.g.*LSTM [9]) with attention mechanism to generate captions sequentially [32]. Different from the early attention-based captioning models that learn attention implicitly by optimizing the captioning objectives [1, 32], Chen *et al.*[7] proposed a boosted attention model to integrate two types of attention emphasizing both the task-specific top-down semantic guidance and the bottom-up saliency of the visual stimuli. By introducing the supervision of stimuli-driven attention, the boosted model can effectively localize the most salient target in the image and prevent itself from focusing on the incorrect regions of interest while generating caption words. At a microscopic scale, Vaswani *et al.*[28] proposed an effective learning structure called multi-head attention which outperforms traditional RNN in many applications due to its ability of characterizing complex correlations in sequential data.

The success of attention-based approaches (either with implicit or explicit attention modeling) well validates that human attention plays an important role in lingual caption generation, and the integration of attention mechanism did improve the captioning performance remarkably. Given such promising results, we cannot help thinking about an interesting reverse problem:

Could the human-labeled image captions (language prior) be used to benefit the training of visual attention models?

Obviously, it's impossible to directly use the lingual captions to train or fine-tune a classic visual attention model because the data labels, *i.e.*textual sentences and 2D fixation maps, are totally heterogeneous, as shown in Figure 1. However, we do have available several good datasets, *e.g.*, SALICON [12], in which all images have both human fixations and caption labels. In this paper we explore an alternative solution by extracting the usefully priors from the language domain and integrating such prior knowledge into the learning process of deep visual attention models in the vision domain. Specifically, we first build a new probabilistic framework of attention to which priors from other domains can be seamlessly integrated. We then analyzed SALICON dataset, where the images are also in the well known captioning dataset MSCOCO. Two stylish viewing modes are discovered and utilized for the implementation of a dynamic Mode-Sensitive Attention model. The contribution of this paper is threefold:

- We proposed a probabilistic framework with the concept of latent mode for dynamic attention inference. It enables the integration of semantic priors from various domains and can be used as a generic framework to boost the performance of single-mode models.

*Corresponding Author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414008

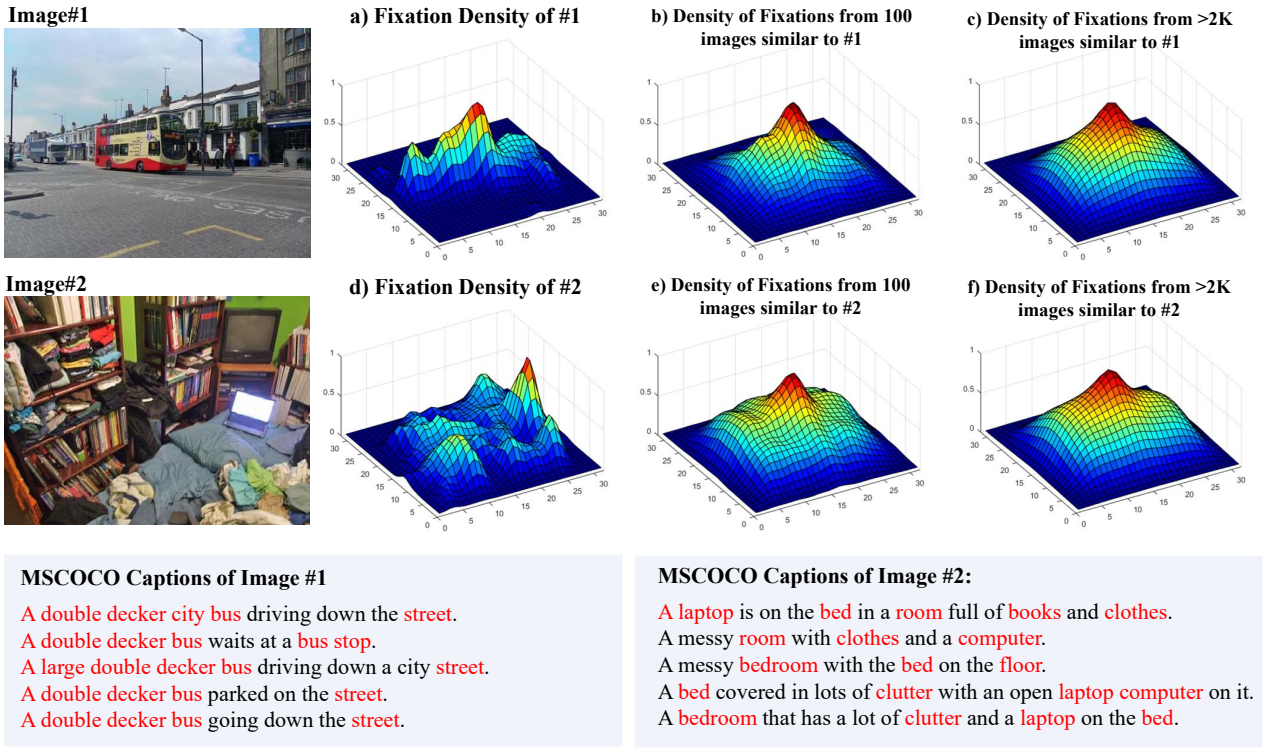

Figure 1: Illustration of the correlation between captions (Language) and eye-Fixations (Vision). Natural images have different styles which strongly affect human perception of the semantic context. For instance, images with a clear salient foreground (#1) lead to consistent viewer attention (eye-fixations, (a)) and descriptions (captions). While images with complex layout (#2) usually result in inconsistent viewer attention (b) and diverse image captions. Such phenomenon can be observed within a small group of similar samples (*e.g.*, (b) and (e)) but vanishes when number of samples increases to a dataset scale as shown in (c) and (f). Unlike traditional works, this paper explicitly models the above phenomenon as attention modes and explores priors from the language domain for robust attention prediction in the vision domain.

- We investigated the correlation between human eye-fixations and image captions, based on which two language-inspired attention modes are defined and a Mode-Sensitive Attention (MSA) model is proposed following our probabilistic framework.
- We conducted extensive experiments on popular benchmarking datasets including YORK120, MIT1003 and PASCAL, and demonstrated that the proposed framework can achieve significant performance gains on top of state-of-the-art backbone models for all metrics.

The rest of this paper is organized as follows. Section 2 introduces recent studies related to our work. Section 3 presents the details of the proposed framework. Experimental settings and results are discussed in Section 4. Finally, we conclude this paper in Section 5.

2 RELATED WORK

2.1 Fixation Prediction in Deep Learning Era

Computational models for fixation prediction aims to predict where human look given an input image or a video clip [11]. Most of the traditional fixation prediction (attention) models are stimuli-driven, which are mostly built upon heuristics and hypotheses from psychology and mathematics, *e.g.*, ITTI [11], AIM [4] and BMS [35]. After the release of the SALICON dataset [12][1], saliency models based on deep neural networks (*e.g.*DeepFix [15], Salicon [10], DPN [22] and DeepGazeII [17]) have emerged and rapidly dominated the leaderborad of saliency benchmarks (*e.g.*MIT300 benchmark [5]). Despite the promising performance, deep models still suffer from poor explainability and extendibility. More details can be found in two recent surveys [2][3] for both classic and deep-learning-based fixation prediction models, respectively.

In this paper, we implement the model of [27] with PyTorch as the backbone network and validate our new probabilistic framework of attention, *i.e.*, MSA, in comparison with the traditional single-step pipelines. As a unified framework, the proposed MSA is more plausible and explainable for the modeling of attention in complex real-world scenarios, which also provides a theoretical interpretation for the integration of statistical priors for attention such as the well known *center bias*.

[1]SALICON contains 15K images with mouse-click labels, which is currently the largest dataset for training fixation prediction models.

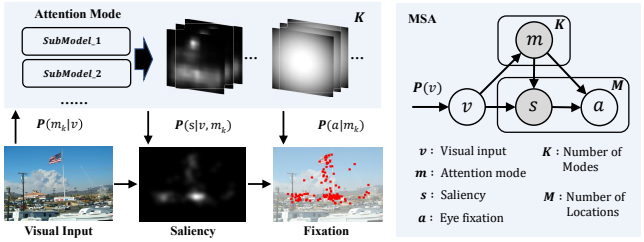

Figure 2: A graphical model representation of the proposed probabilistic framework of attention. Different from all previous works, saliency is estimated based on a dynamic combination of modes instead of a fixed model. This framework enables the integration of prior knowledge from heterogeneous domains and provides a theoretic interpretation for *Center Bias* (the simplest form of mode prior in Eq. 4).

2.2 Attention and Captioning

The attention mechanism has been extensively explored in both image captioning and visual question answering research [1, 30, 32, 33]. However, most of these methods implemented the attention module as an add-on to the existing backbone networks, which is implicitly learned from the semantic structure of sentences. Recent studies [7, 8] explored an alternative architecture by integrating an independent visual attention module specifically designed/trained for stimuli-driven saliency, which has shown significant improvements comparing to traditional frameworks. Efforts have also been made to understand how people allocate their attention while generating image captions. Kiwon *et al.*[34] studied two cross-domain datasets containing both captions and human eye movements, and demonstrate a strong relationship between the shifts of human attention (gaze movements) and the human-generated captions. In a more recent work, Tavakoli *et al.*[26] found that humans mention more salient objects earlier than less salient ones in their descriptions, the better a captioning model performs, the more attention agreement it has with human descriptions. Inspired by these discoveries, He *et al.*[8] constructed the currently largest dataset with synchronously recorded eye-fixations and scene descriptions for the joint analysis of attention and captioning, based on which they further boosted the captioning performance by integrating image saliency with soft-attention. Different from the previous studies that focused on utilizing the attention mechanism to boost captioning performance, our work investigates an interesting reverse problem emphasizing how the large-scale captioning data can be used to benefit the study of human visual attention. To the best of our knowledge, our work is the first attempt to transfer the knowledge of lingual captions to the fine-grained training of fixation prediction networks.

3 THE PROPOSED METHOD

We first proposed a new probabilistic framework of attention in Section 3.1, which introduces the concept of latent mode for attention modeling. Then, in Section 3.2, we show how human-labeled eye-fixations and image captions are correlated to each other, and explicitly define two attention modes according to the semantic consistency of image captions. Based on the above configuration, we further describe our mode-sensitive attention model based on a mode controlling network and a backbone saliency network.

3.1 A Probabilistic Framework of Attention

Assume we have available a set of visual stimuli (*e.g.*images), $\mathcal{V} = \{v_1, \ldots, v_N\}$ which cause the computation of saliency $\mathcal{S} = \{s_1, \ldots, s_M\}$ and the focus of attention (eye fixations) $\mathcal{A} = \{a_1, \ldots, a_M\}$ given M spatial locations. When v_i is presented, $P(s_j|v_i)$ infers the probability of the j-th location being salient and $P(a_l|v_i)$ the probability of the l-th location being attended to by people. Previous works try to compute these conditional probabilities using a single model that treats all input equally. While in this paper, we argue that the traditional framework is inflexible and very much insufficient to handle complex real-world environments, where attention might be modulated in different modes. To address this issue, we introduce an additional factor, termed Latent Attention Mode $\mathcal{M} = \{m_1, \ldots, m_K\}$, and transform the original fixation prediction task into a probabilistic inference problem.

The dependencies between the four factors are illustrated via a graphical model representation in Figure 2. Given a specific input v_i, saliency prediction is achieved by inferring the following posterior probability:

$$P\left(s_j|v_i\right) = \sum_{k=1}^{K} \underbrace{P\left(m_k|v_i\right)}_{\text{mode control}} \underbrace{P\left(s_j|m_k, v_i\right)}_{\text{saliency prediction}}, \tag{1}$$

where saliency is estimated based on a dynamic combination of modes determined by v_i instead of using a single fixed model. Eq. 1 can also be regarded as a weighted summation of the prediction results from different sub-models.

We further define the posterior probability of $P(a_l|v_i)$ as:

$$P\left(a_l|v_i\right) = \sum_{j=1}^{M}\sum_{k=1}^{K} P\left(m_k|v_i\right) P\left(s_j|m_k, v_i\right) P\left(a_l|s_j, m_k\right), \tag{2}$$

where the fixations are inherently determined by saliency and the modes. We assume an extremely simplified relationship between visual saliency at location j, human eye fixation at location l, and the working mode k:

$$P(a_l|s_j, m_k) = \begin{cases} P(a_l|m_k) & l = j, \\ 0 & l \neq j. \end{cases} \tag{3}$$

Then, the posterior $P(a_l|v_i)$ for fixation prediction can be reformulated as:

$$\begin{aligned} P\left(a_l|v_i\right) &= \sum_{j=1}^{M}\sum_{k=1}^{K} P\left(m_k|v_i\right) P\left(s_j|m_k, v_i\right) P\left(a_l|s_j, m_k\right) \\ &= \sum_{k=1}^{K} P\left(m_k|v_i\right) P\left(s_l|m_k, v_i\right) \underbrace{P\left(a_l|m_k\right)}_{\text{mode prior}}. \end{aligned} \tag{4}$$

Note that, when $K = 1$, the term $P(a_l|m_k)$ (mode prior) in Eq. 4 becomes the classic *Center Bias*, which was frequently used as a post-processing term to enhance the prediction performance of an attention model. Based on the proposed probabilistic framework, we describe the mode configuration in Section 3.2 and definition of each term for Eq. 1 and Eq. 4 in Section 3.3 respectively.

3.2 Language-Inspired Mode Configuration

3.2.1 Joint Analysis of Fixations and Captions. Images in the SALICON fixation dataset are all sampled from the MSCOCO dataset [21]. Thus, in addition to fixations, we can also analyze other semantic labels such as object categories and image captions to explore possible latent modes for our probabilistic framework. As shown in Figure 3, images from SALICON can be classified into two main categories: one receives attention with high viewer consistency (all subjects look at the same salient target in the image) and the other with low viewer consistency where people tend to pay their attention to different targets. Interestingly, we found a strong correlation between the consistency of viewer's attention and the diversity of human-labeled captions. People tend to use similar sentences to describe images with clear topics and high viewer consistency, and use diverse sentences to describe images with complex layouts and low viewer consistency.

The above observation inspires us to use the *Consistency of Captions* as a clue to define the latent modes of attention. Intuitively, we present a quantitative metric, termed CapSim, to measure the semantic consistency of a group of captions, based on which we further define two latent modes, *i.e.*, HCC (High Caption Consistency) and HCD (High Caption Diversity). Specifically, given a group of captions $C = \{c_t | t = 1, ..., T\}$, the self-similarity score CapSim(.) is defined as:

$$\text{CapSim}(C) = \frac{1}{T^2} \sum_{t=1}^{T} \sum_{q=1}^{T} \text{CIDEr}(c_t, c_q). \tag{5}$$

where $c_t, c_q \in C$, CIDEr [29] is a classic similarity metric specially designed for the evaluation of captioning models. In Figure 3, we show the distribution of CapSim scores for all images in MSCOCO, from which we can find about 40K highly distinguishable samples.

3.2.2 Automatic Mode Classification. We investigate the possibility of using a CNN model to classify an unlabeled image into HCC or HCD mode. Specifically, we construct a sub-dataset named **COCO-DIV** by extracting 40K samples from the 80K MSCOCO images according to the CapSim score: the top 20K are labeled as **HCC** and the bottom 20K labeled as **HCD**. For quantitative analysis, we randomly took 30K images for training, 5K for validation and 5K for testing. We fine-tuned a pretrained VGG-16 [25] network, where the last classification layer is modified to output a 2D binary prediction. Based on our split, the accuracy of the above VggNet reaches 91.16% on the validation set and 90.18% on the testing set, which we believe is good enough for practical usage.

3.3 Mode-Sensitive Attention Model

Based on the above mode configuration, we build our Mode-Sensitive Attention (MSA) model by integrating a mode controlling network as the implementation for $P(m_k|v_i)$ and two independent saliency prediction networks for $P(s_j|m_k, v_i)$.

3.3.1 Mode Controlling Network. This network determines the working mode of the entire MSA model according to the global appearance of the visual stimuli. Typically, it takes an image as input and outputs the probability (dynamic weight) for each pre-defined mode.

Following the mode classification experiment in Section 3.2, we utilize a pretrained VGG16 model as the backbone of the controlling network and modify the last layer to output a 2D vector, where each dimension represents the corresponding mode (HCC vs. HCD). We fine-tune the mode controlling network on the proposed COCO-DIV dataset, and use the probabilistic output as the implementation for $P(m_k|v_i)$ in Eq. 1 and Eq. 4 during inference. Note that, we also use the binary output of the mode controlling network to determine the training and validation split of the SALICON dataset for fine-tuning the two follow-up saliency prediction networks. From a systematic point of view, the mode controlling network is used not only as a weighting module for result fusion, but also as a data sampling module for mode-sensitive training control.

3.3.2 Saliency Prediction Networks. For each attention mode, we assign a saliency prediction network for the inference of $P(s_j|m_k, v_i)$. We implemented a powerful deep fixation prediction network, *i.e.*, Salicon [10][2], as our backbone for all the following experiments. We follow the codes of **OpenSalicon**[3] for the implementation of this backbone.

We make a few modifications in our implementation to improve the training stability (eliminating *NaN* cases) and reduce the memory cost (enabling larger batch size). Specifically, we reconfigure the input to be a fixed size RGB image of 448×448 pixels and use the corresponding fixation density map as the ground-truth output. We adopt Mean Square Error (MSE) loss with Adam [14] for optimization. The mode controlling network is used to split the original SALICON dataset into two sub-sets (HCC vs. HCD) for the training and validation of the saliency prediction networks. For HCC, we have 3,833 images for training and 1,634 for validation, while for HCD we have 6,167 for training and 3,366 for validation. We first train a backbone model on the full SALICON with 10 epochs to initialize the parameters, and then fine-tune two independent models on the HCC and HCD split with another 10 epochs to finalize the prediction networks for individual modes.

3.3.3 Mode Prior. We learn the mode prior $P(a_l|m_k)$ directly from the mode-based split of SALICON (*i.e.*, HCC and HCD). Let D_k denote the set of training images for mode m_k, then we can compute the mode prior by:

$$P(a_l|m_k) = \mathcal{N}\left(\log \sum_{d \in D_k} P_d(a_l)\right), \tag{6}$$

where $P_d(.)$ is the fixation density function (also know as the fixation density map) estimated using the ground-truth fixations of the d-th image in D_k, and $\mathcal{N}$ is a normalization operator which ensures $\sum_{l=1}^{M} P(a_l|m_k) = 1$.

3.3.4 Post Processing. As demonstrated in many recent studies, blurring the predicted results is an effective and reasonable technique to further boost the overall performance. For our MSA, we also follow this setting to blur the result with a 31×31 Gaussian filter as the final output.

[2]Salicon model is still one of the best in the MIT Saliency Benchmark: http://saliency.mit.edu/

[3]https://github.com/CLT29/OpenSALICON

Figure 3: Examples of images with High Caption Consistency (HCC) and High Caption Diversity (HCD) in MSCOCO. We use the Self-Similarity of Captions (CapSim, Eq. 5) to measure the consistency of captions, according to which we create a new dataset called COCO-DIV by labeling the top 20K images of MSCOCO as HCC and the bottom 20K as HCD. A VGG-based network can achieve over 90% classification accuracy on COCO-DIV, which was used to define and detect the latent modes solely based on the visual input under our probabilistic framework of attention.

4 EXPERIMENTS

We evaluated the effectiveness of our mode-sensitive attention model on popular benchmark datasets in comparison to the top models on the MIT Saliency Benchmark.

4.1 Dataset and Evaluation Metrics

4.1.1 Fixation Datasets. We used the SALICON dataset [12] for the training and fine-tuning of the tested models. This dataset is currently the largest human labeled fixation dataset with 10K images for training and 5K for validation, which is mostly used for training deep attention models. Different from traditional fixation sets that adopt eye-tracking device to capture human fixations. SALICON utilized crowdsourcing to collect mouse clicks from human labeler as the substitute of true fixations. For evaluation, we used three public available large-scale eye fixation datasets constructed by real-world eye-tracking experiments including YORK120 [4], MIT1003 [13] and PASCAL [20]. The above three datasets consist of fixations for 120,1003 and 850 images respectively, which are frequently used for performance evaluation of attention models. Since our model explores cross-modality supervision, where caption labels are used to supervise the training of the prediction network, it would be unfair to directly compare our results with the data posted on the MIT/Tuebingen benchmark. Thus, we mainly compared the results generated by the public available models and our MSA alternatives based on the three offline datasets.

4.1.2 Performance Metrics. There are 8 major performance metrics for the evaluation of fixation prediction models, which measure the quality of a model generated saliency map from different perspectives [6]. As demonstrated in [16], no single prediction can perform best across all metrics. Following [17], in this study we mainly consider the most frequently used metrics, *i.e.*, Area Under ROC curve (AUC[5]), Shuffled AUC (sAUC [36]), Normalized Scanpath Saliency (NSS [23]), and Information Gain (IG [18]).

Details about the definition of **AUC**, **sAUC** and **NSS** can be found in [16] and the implementation codes are available on the MIT Saliency Benchmark. Information Gain (**IG**) measures what one model knows about the data beyond a given baseline model, which has been used by many recent papers and benchmarks for both model evaluation and comparison. Since the IG metric can be used for *model vs. model* comparison, we introduce two variants:

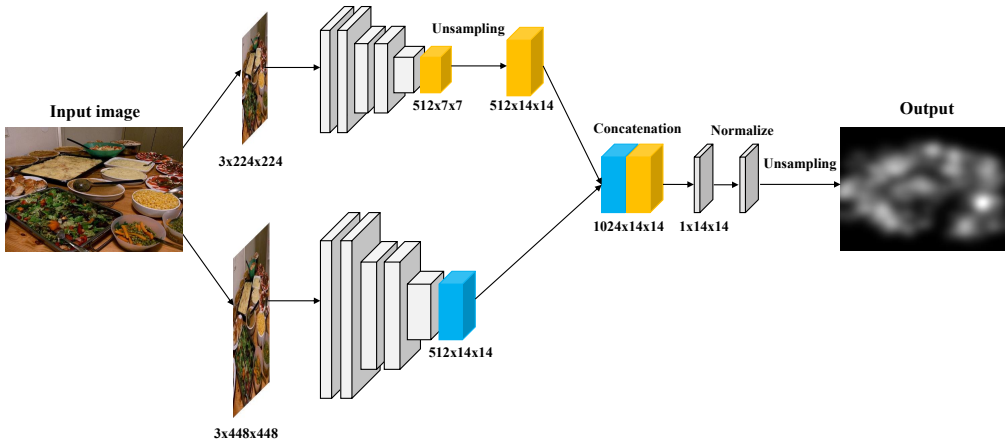

Figure 4: Illustration of the backbone prediction network. The main architecture is based on a state of art saliency network named OpenSalicon, a typical two stream network that fuses visual saliency at both coarse and fine scales.

the original $\mathbf{IG}_{cen}$ metric against center bias and $\mathbf{IG}_{dg2}$ which takes Deep Gaze II [17] as the baseline model.

4.2 Baselines and Implementation Details

All implementations, including the mode controlling network, the mode-sensitive model training, the pre-trained fixation prediction network, as well as the datasets we constructed for this work, can be downloaded at our GitHub project page[4].

We compared our method with several representative open-source fixation prediction models selected from the MIT Saliency Benchmark. Details about all the tested models are given as follows.

- **CEN** This is a reference model under the assumption that the center of the image is most salient. The result is computed by stretching a symmetric Gaussian to fit the aspect ratio of a given image.
- **AWS** AWS [19] is one of the most powerful classic (Non-Deep) bottom-up saliency model which is developed using signal processing techniques such as dynamic whitening.
- **SAL** This is an implementation of the Salicon network [10] which is a typical deep attention model combining the output of two pre-trained CNNs on a different scale (both coarse and fine). We use SAL (100 epoch on SALICON) as a baseline for comparison and its 10 epoch version as the backbone for our MSA.
- **DGII** DeepGazeII [17] utilizes a readout network to decode saliency from a pretrained VGG Net [25]. Untill now, **SAL** and **DGII** are still the best performers on the MIT Saliency Benchmark.
- **MSA-S** and **MSA-A** Two implementations of our proposed Mode-Sensitive Attention model. We took the parameters of a pre-trained **SAL** model (trained on full SALICON data with 10 epoch) for initialization, and achieved two sub-models for our proposed **MSA** by fine-tuning on the new data splits generated using the mode controlling network. MSA-S corresponds to Eq. 1 and MSA-A coresponds to Eq. 4 respectively. MSA-A explicitly takes the mode prior (similar to center bias) into consideration.

[4]https://github.com/zhangxuying1004/MSA

4.3 Results and Analysis

4.3.1 Overview. The main quantitative results are shown in Table 1 and Table 2. We also show the visual comparisons of the predicted saliency maps in Figure 5. The proposed method (MSA-S and MSA-A) achieved significant performance gains compared to its backbone (SAL) on all three datasets. Specially, the IG_{dg2} scores of MSA-A on three datasets are all positive which indicates that it gains more information than the DeepGaze II model[5]. Visual comparisons show that our proposed MSA-A model does perform better for images with dense objects and complex layout.

4.3.2 Ablation Study. We also show the performance of the prediction networks from each mode (HCC and HCD) in comparison to the full MSA model in Table 3. By combining the prediction results from each individual mode in a dynamic manner, MSA-S (Eq. 1) achieved performance gains on all metrics which well demonstrates the effectiveness of the proposed probabilistic framework. By integrating the mode prior, MSA-A (Eq. 2) performs better than MSA-S for most metrics except for sAUC, which is quite reasonable because sAUC was specially designed to eliminate the center bias (a special case of mode prior used in MSA-A) in saliency evaluation.

4.3.3 The Essence. The proposed MSA model by nature is a dynamic ensemble approach. Similar to our MSA, DeepGaze II is also an ensemble model consisting of several decoding networks that share the same feature encoding module. However, the output of DGII is an average fusion of the predictions from each sub-model, while our framework utilizes a mode controlling network to dynamically modulate the offline training and online fusion of the sub-models. The dynamic nature of our framework offers better generalization ability and flexibility for real-world applications.

[5]Note that the public available pre-trained DGII model is an ensemble of 10 models and are fine-tuned on MIT1003. Thus it is unfair to directly compare DGII with other competitors on MIT1003.

Table 1: Information Gain over CEN (IG_{cen}) and DeepGazeII (IG_{dg2}) on YORK120, MIT1003 and PASCAL. The top 2 scores for each metric are highlighted in bold. "↑"means higher is better. The IG_{dg2} scores of our MSA-A are all positive indicating that it gains more information than the DeepGaze II model

	YORK120		MIT1003		PASCAL	
Model	IG_{cen}↑	IG_{dg2}↑	IG_{cen}↑	IG_{dg2}↑	IG_{cen}↑	IG_{dg2}↑
CEN [5]	0.000	-0.660	0.000	-0.939	0.000	-1.255
AWS [19]	0.358	-0.302	0.261	-0.678	0.242	-1.012
SAL [10]	0.927	0.267	1.050	0.111	1.230	-0.025
DGII [17]	0.660	0.000	0.939	0.000	**1.254**	**0.000**
MSA-S (Ours)	**1.015**	**0.355**	**1.061**	**0.122**	1.218	-0.036
MSA-A (Ours)	**1.025**	**0.365**	**1.083**	**0.144**	**1.257**	**0.003**

Table 2: Experimental results for AUC, sAUC and NSS on YORK120, MIT1003 and PASCAL. The top 2 scores for each metric are highlighted in bold. "↑"means higher is better. Our method performs better for AUC & NSS than SAL yet DeepGazeII and SAL both show superior performance in sAUC

	YORK120			MIT1003			PASCAL		
Model	AUC↑	sAUC ↑	NSS↑	AUC↑	sAUC↑	NSS↑	AUC↑	sAUC↑	NSS↑
CEN [5]	0.784	0.500	0.959	0.787	0.513	0.968	0.840	0.516	1.159
AWS [19]	0.760	**0.717**	1.236	0.736	0.686	1.075	0.747	0.657	1.122
SAL [10]	0.819	0.707	1.904	0.830	0.719	1.976	0.854	**0.717**	2.210
DGII [17]	0.783	**0.717**	**2.280**	0.770	**0.749**	**2.681**	0.825	0.700	**2.546**
MSA-S (Ours)	**0.828**	0.708	1.928	**0.839**	**0.723**	1.978	**0.857**	0.712	2.205
MSA-A (Ours)	**0.828**	0.702	**1.951**	**0.839**	0.717	**1.998**	**0.858**	**0.713**	**2.235**

4.3.4 Extensions. The proposed probabilistic framework is designed for the integration of priors which include but not limited to captioning consistency (diversity). We also test the possibility of utilizing the semantic topics of captioning data for mode configuration. Specifically, we regard the MSCOCO dataset as a document dataset and define attention modes according to the topics (document classes) of the corpus. For each image, we group its 5 human labelled captions to form a single document and extract the TF-IDF feature as its basic representation. We then adopt the k-means algorithm to cluster the corpus into N classes(topics). Following Sec.3.2, we fine-tune a VGGNet as the mode controlling network based on the MSCOCO images and their corresponding topic labels. The new mode controlling network can then be used to prepare the training data for the subsequent MSA modules.

In Table 4, we show the results of 4 Topic-Driven MSA alternatives (MSA-S and MSA-A with 5 and 10 topics respectively) in comparison to the Diversity-driven MSA implementations (2 modes only). Based on empirical observations, we found that 10 is the optimal mode number for most evaluation metrics on the evaluation datasets. As the number increases, we can observe a clear performance degeneration for most metrics, which we believe is caused by the unbalanced distribution of modes (some have thousands of samples, while the rest only have a few hundreds). Although improvements have been achieved across all three datasets, the model size also increases dramatically which makes the new alternatives unsuitable for popular lightweight platforms such as mobile phones. However, by such extensions, we show the potential of our probabilistic framework in integrating different prior knowledge to the fine-grained training process of the prediction networks.

Table 3: Ablation Experimental Results on MIT1003. Our framework can effectively fuse the two modes and gain stable improvements for all metrics especially for IG_{dg2}

Model	AUC	sAUC	NSS	IG_{cen}	IG_{dg2}
MSA_{HCC}	0.839	0.722	1.976	1.057	0.118
MSA_{HCD}	0.837	**0.723**	1.975	1.060	0.121
MSA-S $_{HCC+HCD}$	**0.839**	**0.723**	**1.978**	**1.061**	**0.122**
MSA-A $_{HCC+HCD}$	**0.839**	0.717	**1.998**	**1.083**	**0.144**

Table 4: Results of different MSA alternatives. We re-implement MSA by configuring the modes as caption topics (classes). The new Topic-Driven alternatives show better performance as well as larger computational cost

MSA	**YORK120**		**MIT1003**		**PASCAL**		Mode
Alternatives	IG_{cen}	IG_{dg2}	IG_{cen}	IG_{dg2}	IG_{cen}	IG_{dg2}	Number
MSA-S $_{HCC+HCD}$	1.015	0.355	1.061	0.122	1.218	-0.036	2
MSA-S $_{Topics,\ 5}$	1.031	0.370	1.060	0.121	1.239	-0.015	5
MSA-S $_{Topics,\ 10}$	1.036	0.376	1.066	0.127	1.220	-0.034	10
MSA-A $_{HCC+HCD}$	1.025	0.365	1.083	0.144	1.257	0.003	2
MSA-A $_{Topics,\ 5}$	1.032	0.372	1.083	0.144	1.273	0.018	5
MSA-A $_{Topics,\ 10}$	1.049	0.389	1.086	0.147	1.248	-0.006	10

Despite exploring language labels to set up the latent modes, other alternative cues, *e.g.*, object and scene labels as well as photo

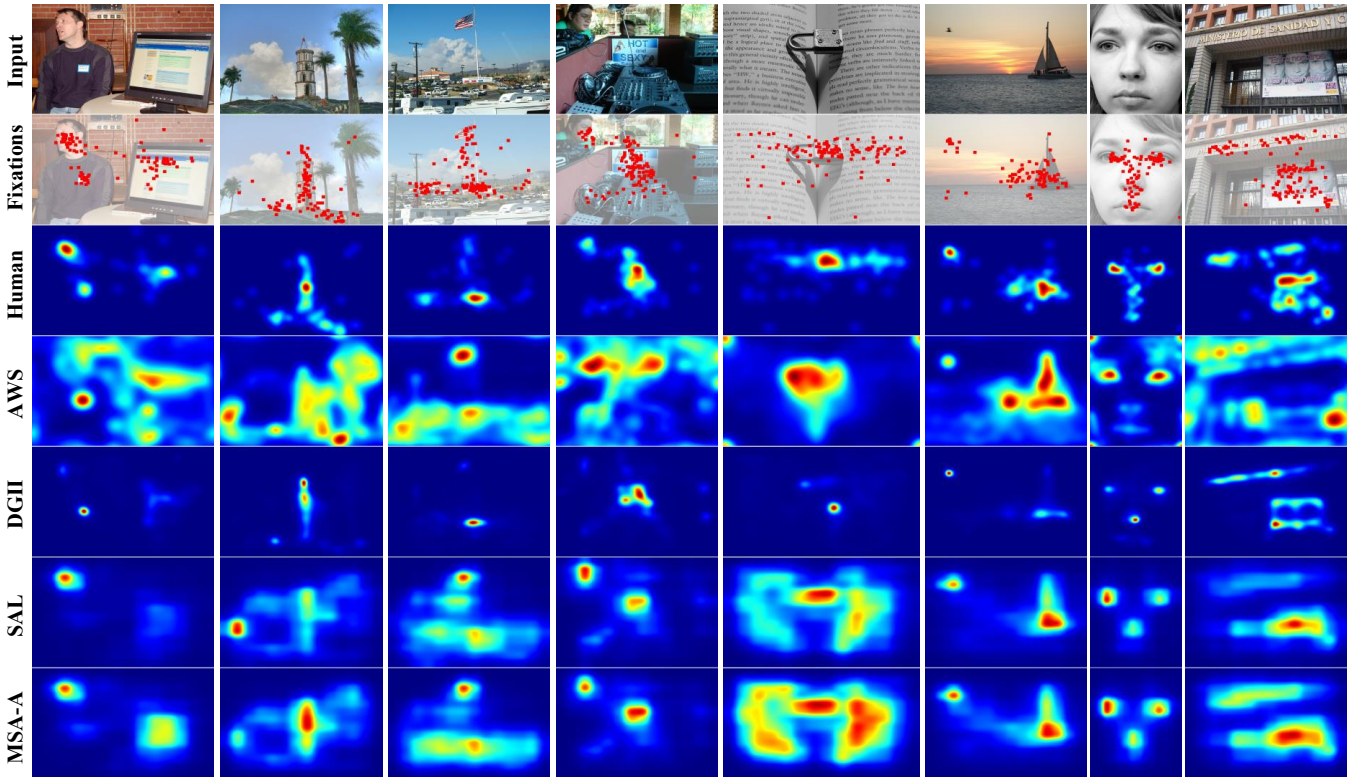

Figure 5: Visual Comparisons of the ground truth fixations (Row#2 and #3) and the predicted results of tested models. The classic non-deep AWS method is strongly affected by regions with high contrast textures or edges which actually attract less attention than meaningful objects (*e.g.*, human faces and texts). All the deep learning-based methods, *e.g.*, DGII, SAL, and our MSA-A can localize the most salient target in easy images with clear foreground objects. The proposed MSA method performs better for images with multiple objects and complex layout.

styles, can also be explored for more complex mode configurations. We expect more novel implementations based on our framework in the follow-up works.

4.3.5 Limitations. One key limitation to the proposed framework is that the size of the model will continually increase when more modes are used, which linearly lower the inference speed. A solution to ensure efficiency is to modify the control module to turn soft model fusion into hard model selection where only one or a few sub-models are activated for online inferences. The acceleration of our proposed framework is beyond the scope of this paper, but we will take it as an important direction in our future work.

5 CONCLUSION

In this paper, we proposed a new probabilistic framework to characterize human viewing behavior which emphasizes the effect of latent mode during attention inference. In this framework, visual input changes the model's belief on the attention modes and consequently determines how saliency and eye-fixations are predicted. We also defined two language-inspired attention modes and implemented a Mode-Sensitive Attention (MSA) model following the presented probabilistic framework. Experimental results on popular datasets have demonstrated the effectiveness of the proposed approach on major evaluation metrics.

In future studies, we will apply our method to enhance other single-mode models by converting them into the more robust multi-mode versions, and explore more semantic cues from different domains for the configuration of latent modes, *e.g.*, biological hypotheses from psychology and semantic graph from knowledge base. Besides, we will continue to conduct more in-depth investigations into the relationship between language generation and human attention mechanism, and try to gain more insights for cross-domain intelligent tasks such as image/video captioning, referring expression comprehension, language-based retrieval and visual question answering.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.U1705262, No.61772443, No.61572410, No.61802324, and No.61702136), National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Key R&D Program of Jiangxi Province (No. 20171ACH80022) and Natural Science Foundation of Guangdong Province in China (No.2019B1515120049).

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*. 6077–6086.

[2] Ali Borji. 2019. Saliency Prediction in the Deep Learning Era: Successes and Limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1. https://doi.org/10.1109/tpami.2019.2935715

[3] Ali Borji and Laurent Itti. 2013. State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35, 1 (2013), 185–207.

[4] N. Bruce and J. Tsotsos. 2006. Saliency Based on Information Maximization. In *Advances in Neural Information Processing Systems,(NIPS)*. 155–162.

[5] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. [n.d.]. MIT Saliency Benchmark.

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2016. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605* (2016).

[7] Shi Chen and Qi Zhao. 2018. Boosted Attention: Leveraging Human Attention for Image Captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 68–84.

[8] Sen He, Hamed R. Tavakoli, Ali Borji, and Nicolas Pugeault. 2019. Human Attention in Image Captioning: Dataset and Analysis. In *The IEEE International Conference on Computer Vision (ICCV)*.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[10] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 262–270.

[11] L. Itti and C. Koch. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience* 2, 3 (2001), 194–203.

[12] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. SALICON: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1072–1080.

[13] T. Judd, K. Ehinger, F. Durand, and A. Torralba. 2009. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2106–2113.

[14] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[15] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. 2015. Deepfix: A fully convolutional neural network for predicting human eye fixations. *arXiv preprint arXiv:1510.02927* (2015).

[16] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. 2018. Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics. In *ECCV*.

[17] Matthias Kümmerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. 2017. Understanding Low- and High-Level Contributions to Fixation Prediction. In *The IEEE International Conference on Computer Vision (ICCV)*.

[18] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. 2015. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences* 112, 52 (dec 2015), 16054–16059. https://doi.org/10.1073/pnas.1510393112

[19] Víctor Leborán, Anton Garcia-Diaz, Xosé R Fdez-Vidal, and Xosé M Pardo. 2017. Dynamic whitening saliency. *IEEE transactions on pattern analysis and machine intelligence* 39, 5 (2017), 893–907.

[20] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. 2014. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 280–287.

[21] Tsungyi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*. 740–755.

[22] Junting Pan, Kevin McGuinness, Elisa Sayrol, Noel O'Connor, and Xavier Giro-i Nieto. 2016. Shallow and Deep Convolutional Networks for Saliency Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 598–606.

[23] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. 2005. Components of bottom-up gaze allocation in natural images. *Vision Research* 45, 18 (aug 2005), 2397–2416. https://doi.org/10.1016/j.visres.2005.03.019

[24] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015), 1137–1149.

[25] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[26] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying Attention to Descriptions Generated by Image Captioning Models. In *The IEEE International Conference on Computer Vision (ICCV)*.

[27] Christopher Lee Thomas. 2016. *OpenSalicon: An Open Source Implementation of the Salicon Saliency Model.* Technical Report TR-2016-02. University of Pittsburgh.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.

[29] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and L. M. Shao. 2019. An Iterative and Cooperative Top-Down and Bottom-Up Inference Network for Salient Object Detection. In *CVPR*.

[32] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*. 2048–2057.

[33] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention. In *CVPR*. 4651–4659.

[34] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. 2013. Studying Relationships Between Human Gaze, Description, and Computer Vision. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Computer Society Conference on*. IEEE.

[35] Jianming Zhang and Stan Sclaroff. 2013. Saliency detection: a boolean map approach. In *ICCV*.

[36] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell. 2008. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision* 8, 7 (2008), 1–20.