The Role of Visual Attention in Sentiment Prediction

Shaojing Fan
Smart Systems Institute, National University of Singapore
idmfs@nus.edu.sg

Ming Jiang
Department of Computer Science and Engineering, University of Minnesota, mjiang@umn.edu

Zhiqi Shen
Smart Systems Institute, National University of Singapore
idmshenz@nus.edu.sg

Bryan L. Koenig
Department of Psychology, Southern Utah University
bryanleekoenig@gmail.com

Mohan S. Kankanhalli
School of Computing, National University of Singapore
mohan@comp.nus.edu.sg

Qi Zhao
Department of Computer Science and Engineering, University of Minnesota, qzhao@cs.umn.edu

ABSTRACT

Automated assessment of visual sentiment has many applications, such as monitoring social media and facilitating online advertising. In current research on automated visual sentiment assessment, images are mainly input and processed as a whole. However, human attention is biased, and a focal region with high acuity can disproportionately influence visual sentiment. To investigate how attention influences visual sentiment, we conducted experiments that reveal critical insights into human perception. We discover that negative sentiments are elicited by the focal region without a notable influence of contextual information, whereas positive sentiments are influenced by both focal and contextual information. Building on these insights, we create new deep convolutional neural networks for sentiment prediction that have additional channels devoted to encoding focal information. On two benchmark datasets, the proposed models demonstrate superior performance compared with the state-of-the-art methods. Extensive visualizations and statistical analyses indicate that the focal channels are more effective on images with focal objects, especially for images that also elicit negative sentiments.

CCS CONCEPTS

• **Human-centered computing** → **Visualization**; • **Computing methodologies** → *Computer vision*;

KEYWORDS

Visual sentiment; social multimedia; neural network

1 INTRODUCTION

How might you describe an image? Amusing? Pleasant? Scary? The emotions that viewers feel when observing an image are often referred to as the image's *visual sentiment.* Analysis of visual sentiment has become increasingly important due to the huge volume of online visual data generated by the recent explosion of social media. The automatic assessment of visual sentiment has many applications, such as monitoring and predicting crises in social media, facilitating social advertising, and understanding user behavior. However, compared with textual sentiment, visual sentiment is more subjective and implicit [17, 18, 32], which makes it challenging to model computationally.

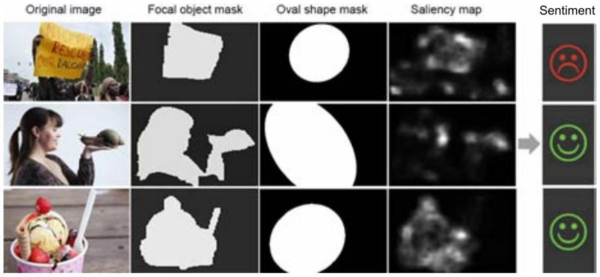
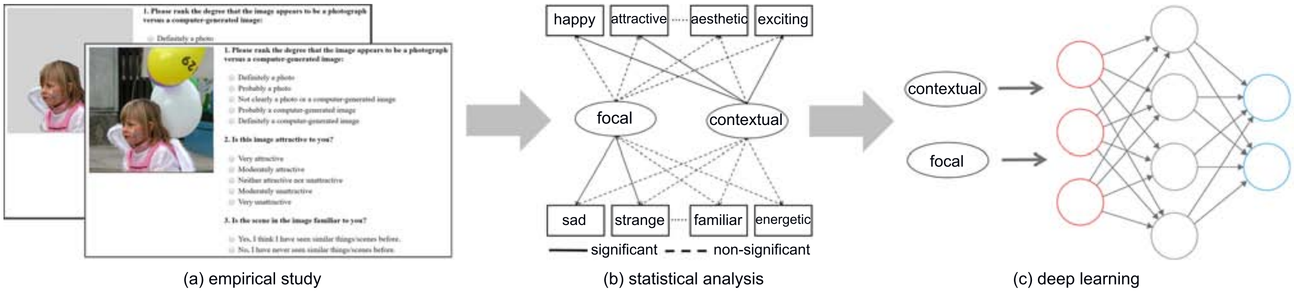


Figure 1: Visual sentiment prediction is distinctive in that human affective responses strongly relate to visual attention. In this study, we demonstrate how visual sentiment prediction is enhanced by incorporating attention information (*i.e.*, focal object masks, saliency maps, as shown in the second to fourth columns in the above image).

Many algorithms have been designed to automatically predict visual sentiment [2, 5, 13, 23, 30]. A common approach is to correlate lower-level image features with higher-level properties, and train a computational model using human ground truth [5, 23]. Recently, Deep Neural Networks (DNNs) have demonstrated superiority in related tasks [10, 12, 50, 51]. DNNs achieve impressive performance, but they provide little insight into why the learned features predict visual sentiment.

Most visual sentiment algorithms process images as a whole. In contrast, human attention mechanisms prioritize regions of relevance [11, 36]. This selective attention interplays with various visual perception tasks [34, 47], particularly tasks related to observer emotion [19]. For example, negative affect leads individuals to focus attention on local details whereas positive affect leads to a broadening of attention [19, 39]. We

MM '17, October 23–27, 2017, Mountain View, CA, USA.
© 2017 ACM. ISBN 978-1-4503-4906-2/17/10. . . $15.00
DOI: https://doi.org/10.1145/3123266.3123445



1. Please rank the degree that the image appears to be a photograph versus a computer-generated image:

○ Definitely a photo
○ Probably a photo
○ Not clearly a photo or a computer-generated image
○ Probably a computer-generated image
○ Definitely a computer-generated image

2. Is this image attractive to you?

○ Very attractive
○ Moderately attractive
○ Neither attractive nor unattractive
○ Moderately unattractive
○ Very unattractive

3. Is the scene in the image familiar to you?

○ Yes, I think I have seen similar things/scenes before.
○ No, I have never seen similar things/scenes before.

Figure 2: (a) Inspired by prior research, we perform studies to experimentally disentangle effects of focal information and contextual information on visual perception. (b) We then statistically analyze the relationships among focal and contextual information on 31 high-level image attributes. (c) Finally, we construct deep neural network models for sentiment prediction that incorporate human perception characteristics by using a devoted focal channel.

therefore thought that visual sentiment prediction might be improved by incorporating selective attention into DNNs.

In this study, we experimentally disentangle effects of focal information and contextual information on human emotional reactions, then we incorporate related insights into computational models. To separate focal and contextual information, we isolate the focal object of each image, defined as the object with the highest saliency score in an automatically computed saliency map [21]. We compare human perception of the isolated focal object with that of the corresponding whole image. This enables us to distinguish how focal and contextual information influence image sentiment. Building on these insights we propose DNNs for sentiment prediction that process focal information through dedicated channels. Two approaches to incorporating focal information into models are evaluated: focal object masks and saliency maps (see Fig. 1). Fig. 2 illustrates our research paradigm.

Our contributions are summarized as follows.

- We discover that negative sentiments are elicited by the focal region without a notable influence of contextual information, whereas positive sentiments are influenced by both focal and contextual information.
- We develop new deep neural networks for predicting visual sentiment that integrate perceptual characteristics of focal vision. Our models demonstrate superior performance in predicting visual sentiment on two benchmark datasets.
- We find that the proposed focal channels are more effective on images with focal objects, especially for images that also elicit negative sentiments.

2 RELATED WORK

Emotion and Attention: Humans have a tremendous ability to direct their gaze rapidly when looking at static or dynamic scenes and selectively process visual information of interest. For example, studies have found that visual attention is attracted to the most informative regions [7], the most surprising regions [22], or regions that maximize task reward [45]. Particularly, studies in psychology and neuroscience found that human attention generally prioritizes emotional content over non-emotional content [6, 46]. For example, smiling people, cute babies, erotic scenes, as well as poisonous snakes and scenes of war attract human attention more than emotionally neutral stimuli [14, 49].

Researchers also find that attention and emotion interact during visual perception [33, 43]. For example, [38] shows that selective attention determines emotional responses to novel visual stimuli. Emotion also enhances the subjective feeling of remembering [19]. Negative emotion leads individuals to focus attention on local details whereas positive emotion leads to a broadening of attention [19, 39]. We extend this line of research by evaluating how selective attention influences visual sentiment.

Predicting Visual Sentiment: In contrast to the abundant research on predicting textual sentiment [1, 18], much less research has been done on the sentiment analysis of visual content. In [30], the authors classify images into eight emotions using hand crafted features. Other researchers [5] propose a bank of visual classifiers (called "SentiBank") with 1,200 linear SVM outputs that use a taxonomy of "adjective-noun pairs". This approach is then extended to object-based visual sentiment [13]. These visual sentiment studies use low-level and mid-level image features (*e.g.*, color, content, composition, GIST [35], SIFT [27]) that are known to predict the visual sentiment of images.

Recently, DNNs have been increasingly used to predict visual sentiment. For example, the SVM-based emotion classifier SentiBank [5] has been extended to DeepSentiBank by using a DNN [12]. Another approach [51] uses progressively trained and domain transferred deep networks for sentiment prediction. The performance of [51] is further boosted by adding augmented data with oversampling [10]. In [50], the DNN is modified to obtain tree-structured recursive neural networks for visual-textual sentiment analysis.

Our approach is distinct from the above methods because we design DNNs based on insights into human perception of visual sentiment. Combining empirical human studies and

Table 1: List of 31 human-annotated attributes in our dataset.

Emotions: Makes you happy? Exciting? Amusing?* Makes you sad? Unusual or strange? Mysterious? Energetic?
Spatial layout: Contain objects of focus? Single focus?* Object centered?* Close or distant view? Neat space? Empty or full space? Common perspective? Clean scene?
Color and illumination: Colorful? Harmonious color? Natural lighting? Natural color?
Aesthetics related attributes: Aesthetic? Image quality? Sharp or blurry; Expert photography? Attractive to you? Appears to be a photograph rather than computer generated?
Semantics-related attributes: People present?* Fine details? Storyline? Natural objects? Natural objects combinations? Familiar to you?

* Attributes designed by the authors. The rest are from [16].

computational modeling, we present an interdisciplinary approach to visual sentiment modeling.

3 PSYCHOLOGY EXPERIMENTS

We perform psychology experiments with human observers to see how focal and contextual information influence visual sentiment. Experiments are conducted on Amazon Mechanical Turk (MTurk) [37].

3.1 Stimuli

Stimuli are 400 images from the Visual Realism Dataset (VRD) [16]. We choose VRD as it includes diverse scenes along with (1) 38 human-annotated attributes on the original images, and (2) extensive object labels annotated on LabelMe [42], which enable focal-object extraction. We are aware of other datasets with salient object labels, such as the PASCAL-S [28], but they do not provide human-annotated attributes.

In order to evaluate the impact of focal information, we extract the focal object of each image through these two steps. First, a global saliency map of each image is computed using the saliency prediction algorithm, SALICON [21]. SALICON is a DNN pre-trained for object classification and fine-tuned on human fixations. It has state-of-the-art performance on saliency prediction [8]. The saliency score of each object is defined as the highest saliency value in the object region annotated via LabelMe. The object with the highest saliency score is extracted as the focal object by removing its background. The average size of the focal object (normalized by the total image size) in the image set is 0.34 ± 0.18. Fig. 3 provides sample images with both versions: the original image, and the manipulated image showing only the focal object.

3.2 Method

Workers on MTurk complete a series of image annotation tasks. Each participant sees only original images or only focal objects. We select a subset of 27 attributes that are explicitly or potential related to emotion from [16]'s attributes list, and add 4 additional attributes relating to emotion and object focus, namely "interesting", "single focused", "centered", "people presence" (see Table 1). The attributes can be classified into four groups: (1) commonly studied human emotions in psychology [15, 32]; (2) low-level image attributes that are known to influence human emotion, such as color and illumination [3]; (3) high-level image attributes potentially related to emotion, such as aesthetics, naturalness, and semantics; (4) spatial layout, which influences human attention, such as whether the image has objects of focus, whether it has a single focus, or whether the focal object is in the center of the scene [34].

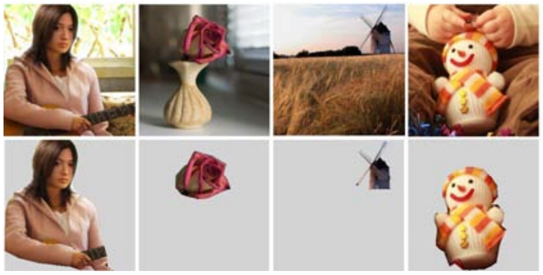


Figure 3: Sample images in our stimulus set. On the bottom of each original image is the version manipulated to show only the focal object.

Each image is rated by 9 MTurk workers for all of the attributes (readers can refer to the supplementary material[1] for the detailed questionnaire). The average response across the 9 ratings for each attribute is normalized between 0 and 1, and stored as the attribute score.

For data reliability analysis, we perform two analyses to assess within and across group consistency in human annotation. First, we use bootstrapping to randomly form two subject groups. That is, we randomly select 9 data points (using sampling with replacement) from all the annotations per image to form an observation of one participant group, and repeat this to create another group. We quantify the degree to which each attribute score for the two sets of participant groups is in agreement using Spearman's rank correlation (ρ). We compute the average ρ over 25 bootstrapping iterations. Overall there is a moderate consistency among all attributes, ρs[2] ≥ 0.39, ps < 0.05 (within-group consistency). We further compute the correlation of our annotations on the original image and the annotations published with VRD which were collected by the authors from [16], to test across-group consistency. Although the two sets of annotations are collected

[1] The supplementary material is available at https://sites.google.com/site/fanshaojing/.

[2] Throughout the paper, ρs and ps represent the plural form of ρ and p, respectively.

during different periods of time with different experimental settings, there is still a statistically meaningful correlation (ρs ≥ 0.15, ps < 0.05), indicating that people are moderately consistent in image perception.

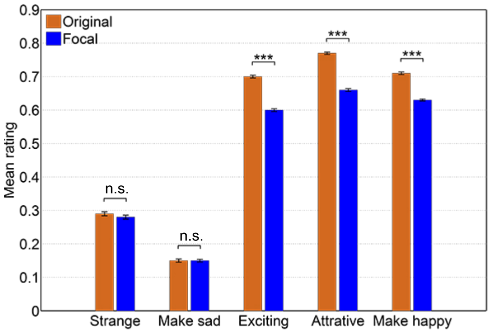


Figure 4: Human ratings on isolated focal objects do not differ significantly from original images on negative sentiments, but are significantly lower on positive sentiments. The asterisks are denoted as following: * $p < 0.001$, ** $p < 0.0002$, *** $p < 0.00002$. $n.s.$ represents non-signiciant results. For all figures in this paper, error bars indicate the standard error of the mean.

3.3 Data Analyses

We use multi-level modeling to analyze the attributes ratings [25]. The ratings for each image are influenced by two levels (lower level: individual human rater; higher level: individual image) and two factors (manipulation—original or isolated focal object; image content). We set the participant-level attribute rating as the dependent variable, image manipulation as a fixed factor and image index (representing different image content) as a random factor. A mixed-model ANOVA is performed on each of the 31 attributes. Due to the large number of comparisons, we use Bonferroni correction with a reduced significant level [3] (α) of 0.001. These analyses are standard in behavioral and other sciences. See, for example, [4] for an introduction to these inferential statistics.

As shown in Fig. 4, negative sentiments like "strange" and "make sad" do not differ significantly between the original images and isolated focal objects, suggesting that information eliciting these reactions is largely present in the focal object. It is reminiscent of the idiom "the rotten apple injures its neighbors"—so long as the most salient object is perceived as negative, the whole image will be affected and perceived the same way. In contrast, the focal object viewed alone has significantly lower scores compared to original images on positive sentiments such as "exciting", "attractive", and "make happy". This indicates that contextual information has an important effect above and beyond that of the focal object for positive visual sentiments. In summary, **negative sentiments are elicited by the focal region without a notable influence of contextual information, whereas positive sentiments are influenced by both focal and contextual information.**

Our findings are reminiscent of the studies from psychology, which report that negative affect is associated with enhanced memory of the focal region whereas positive affect is more related to the memory of contextual details [19, 24, 39, 48].

4 COMPUTATIONAL MODELING OF VISUAL SENTIMENT

In this section, we use our psychology findings to guide the design of DNNs that integrate human attentional bias. Experiments on two benchmark datasets demonstrate the superior performance of the proposed DNN models.

4.1 Proposed DNN Architecture

We base our model on the VGG-19 convolutional neural network architecture [44]. To represent human attention, we introduce a devoted focal channel to the DNNs that uses focal object masks or saliency maps (see the following subsection for detailed designs of the focal channel). We modify the input and the first convolutional layer of the network, to feed the saliency maps or focal object masks together with the input images. The DNN architecture is illustrated in Fig. 5. We resize the images in our datasets to a fixed scale (224×224×3), to be consistent with the input of VGG-19. We further modify the input size of the network to 224×224×4, where three channels contain the RGB colors of the image and the fourth channel contains the saliency map or focal object mask of the image. The first convolutional layer parameters are also modified accordingly for the extra input channel. The model parameters are transferred from pre-trained models on the ImageNet [41] training set. At the first convolutional layer, model parameters for the RGB channels are transferred from the pre-trained models. For the extra focal channel, the model parameters are randomly initialized following a normal distribution. We modify the final fully-connected layer of the network for binary classification, using two classification neurons and a softmax loss. Finally, the network is fine-tuned and evaluated on the targeted sentiment datasets.

4.2 Modeling Human Attention with Focal Channels

Below we describe our methods to model human attention with new focal channels in our DNNs. Two major methods are used to build the focal channels for each image. First, we generate the grayscale mask for the focal object using an automated salient object detection algorithm [29] (model referred to as NUSFocalObj). The steps are similar to those in our psychology study (Sec. 3.1), except here we use automated object segmentation instead of human annotated object labels. Second, we compute the saliency map that predicts human fixations using the SALICON model [21] (model referred to as NUSFocalSal).

[3] The significance level α for a given hypothesis test is a value for which a p-value less than or equal to α is considered statistically significant. The smaller the p-value, the more convincing the evidence is against the null hypothesis of no difference between the means. Typical values for α are 0.1, 0.05, and 0.01.

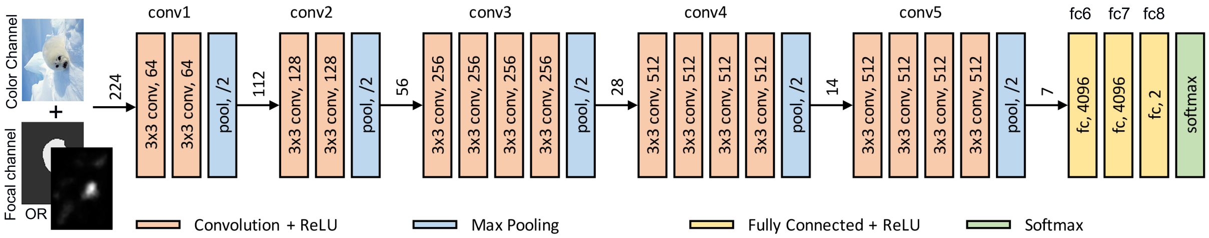


Figure 5: The architecture of the proposed DNNs. It emulates human attentional bias by adding a focal channel (whose input is either focal object mask or saliency map), adapted for visual sentiment prediction.

Table 2: Classification results on Visual Realism Dataset (VRD) and Twitter dataset (DeepSent) by our models and other state-of-the-art methods. The highest performance on each metric is highlighted in bold. The performance of models with focal channels are with gray background. Prec and Acc are short forms of precition, accuracy, respetively.

Model		DeepSent				VRD			
Name	Attention info	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc
NUSNoFocal	None	0.88	0.89	0.88	0.84	0.85	0.77	0.80	0.79
NUSFocalObj	Focal object mask	0.90	0.89	0.89	0.86	**0.86**	**0.79**	**0.82**	**0.81**
NUSFocalOval	Oval shape mask	0.88	**0.91**	**0.90**	0.86	**0.86**	0.78	**0.82**	0.80
NUSFocalSal	Saliency map	**0.91**	0.89	**0.90**	**0.87**	0.85	0.77	0.81	0.80
PCNN [51]	None	0.80	0.88	0.85	0.78	0.81	0.77	0.79	0.77
FTCNN [10]	None	0.80	0.86	0.83	0.76	0.75	0.76	0.76	0.73
DeepSentiBank [12]	None	0.81	0.85	0.83	0.77	0.82	0.73	0.77	0.76

Since a focal object mask encodes object shape information, we further design an elliptical focal channel to evaluate the cause of performance boost, *i.e.,* whether and how much it is from the focal location and from the shape context information. Specifically, we fit the object mask with an ellipse that has the same second moments as the mask region, resulting in an oval shape for each object mask (see Fig. 1). Such oval masks still indicate focal region, but has no shape context information (model referred to as NUSFocalOval).

4.3 Experimental Settings

Datasets: We test our models on two benchmark datasets with human annotations. The first dataset is the Twitter dataset (also called “DeepSent”), collected and released in [51] for visual sentiment prediction. DeepSent contains 1269 images, each labeled for either positive or negative sentiment by five human annotators on MTurk. We use the subset of 882 images that had a consensus across all five annotators [51], for which the number of images with positive and negative labels are 581 and 301, respectively.

The second dataset is Visual Realism Dataset (VRD) [16], from which we select our stimuli for our psychology studies. The VRD includes 2520 images, each with 38 annotations ranging from emotions to semantics, based on 3794 human annotators on MTurk [37]. VRD does not provide human annotated binary sentiment labels. However, based on [17], the sentiment in VRD is strongly correlated with three annotated attributes, namely “make happy”, “attractive”, and “colorful”. We compute the average of these three attributes, resulting in a continuous *sentiment score* ranging from 0 to 1 for each image. Similar to DeepSent, we select a subset of 882 images with the strongest sentiments (top 20% most positive and top 15% most negative along the sentiment score scale), and dichotomize their sentiment scores to get binary sentiment labels (see supplementary material for details). In total, the VRD subset has the same number of images as DeepSent (882), and the number of images with positive and negative labels are 504 and 378, respectively.

DNN parameters: We initialize the training to the pretrained parameters for VGG-19 on ImageNet. The parameters of the DNN are then learned end-to-end on the training images with stochastic gradient descent. We use 32 images for each iteration since we do not get improvement from minibatch. A momentum of 0.9 and a weight decay of 0.0005 are used. The learning rate is fixed at 10^{-4} for the first 10 epochs and decay 1 time for each 10 epochs. Each epoch contains about 1000 iterations. The entire training set is shuffled after each epoch is finished. In each epoch, the network is validated against the validation set of around 200 images to monitor convergence and overfitting. We stop learning when the objective function does not improve on the validation set. We train the network in a single NVIDIA Titan GPU, and it takes approximately 4 hours to finish the training. Each image is horizontally flipped in the training set as augmented data. The data are divided into five different folds to obtain more statistically meaningful results by applying cross-validation.

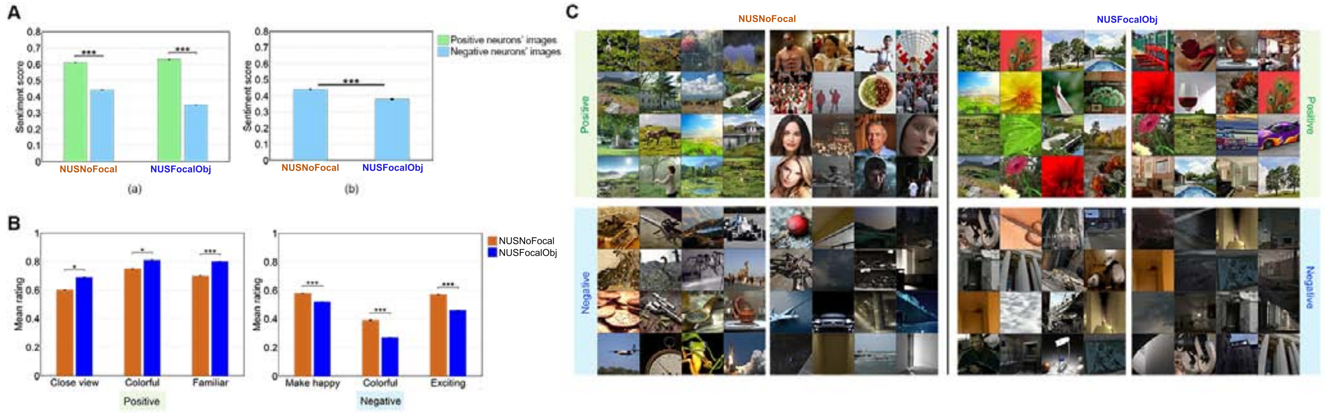


Figure 6: The positive neurons' images of NUSFocalObj are more positive than those of NUSNoFocal, and the negative neurons' images of NUSFocalObj are more negative than those of NUSNoFocal:
(A) The positive neurons' images have significantly higher sentiment score (*i.e.*, more positive) than the negative neurons' images in both NUSNoFocal and NUSFocalObj. Asterisks indicate significant difference between each groups using independent-samples t-test. $* \ p < 0.01$, $** \ p < 0.002$, $* * * \ p < 0.001$.
(B) The positive neurons' images of NUSFocalObj are more colorful, familiar and of closer view (left), and the negative neurons' images of NUSFocalObj are more negative (right), when compared to those of NUSNoFocal.
(C) Visualization of positive neurons' images and negative neurons' images in NUSNoFocal and NUSFocalObj on VRD Dataset. In total 8 neurons are visualized: the top 2 neurons of each of two models for positive sentiment (top row), and the top 2 neurons of each of two models for negative sentiment (bottom row). For each neuron, the 16 images most activate the corresponding neuron are shown.

4.4 Experiment Results

We first compare the performance between our DNN models with focal channels (NUSFocalObj, NUSFocalOval, NUSFocalSal) and our model without a focal channel (NUSNoFocal), to investigate the advantage of focal channels. We then compare our results with three state-of-the-art methods dedicated for visual sentiment prediction: (1) PCNN—a progressively trained and domain transferred DNN for image sentiment analysis [51]; (2) FTCNN—a fine-tuned DNN using AlextNet-style architecture adapted for visual sentiment prediction [10]; (3) DeepSentiBank—a DNN using CaffeNet [26] for sentiments prediction [12]. We download their published models and fine-tune the models on the two datasets to achieve the best performance possible.

The results are shown in Table 2. As shown in the first four rows in Table 2, after adding a focal channel, the performance on both datasets increases compared to the model without a focal channel. NUSFocalOval has a comparable performance as NUSFocalObj, suggesting that the performance boost is mostly from the area of focus rather than object shape context. NUSFocalObj, NUSFocalSal, and NUSFocalOval generate the highest performance among all comparison methods on all metrics, suggesting the advantage of identifying the focal information. Even without a focal channel, NUSNoFocal considerably outperforms the three comparison methods. This advantage may be due to the more complex DNN architecture VGG-19 [44], which provides more parameters to be trained for sentiment classification task.

5 DIVING DEEPER INTO THE DNNS: ANALYSES AND VISUALIZATIONS

In this section, we visualize the neurons that have the highest weights to positive and negative sentiments in our DNNs, to explore the advantage of using a focal channel. We also classify images in terms of their attention patterns (*i.e.*, images with objects of focus, images without obvious focus), and compare the performance of different models on the image groups, to better understand the performance boost.

5.1 Analyses on the VRD Dataset

The images in VRD have extensive human annotated attributes, making it possible to investigate the visualizations quantitatively.

5.1.1 Visualization of the neurons. In this subsection, we compare the visualizations of NUSNoFocal and NUSFocalObj (the model that has the highest performance on VRD), in order to have a deeper understanding on how the additional focal channel contributes to performance. First, to visualize what the neurons have learned, we select 5 neurons in each model (NUSNoFocal and NUSFocalObj) in the layer before the last fully-connected layer (fc7 layer, refer to Fig. 5) with the strongest contribution to the two classification neurons in the last fully-connected layer (*i.e.*, the positive neuron and negative neuron in fc8 layer). Here, contribution is defined as the difference between the weight to the positive and negative neurons on the fc8 layer. For each neuron, we select the 16 images that most activate it. In total, for each model,

Table 3: Classification results on images with and without focal objects in VRD Dataset. The highest performance on each metric is highlighted in bold.

Model	Images have focal objects				Images without focal objects			
	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc
NUSNoFocal	0.89	0.82	0.86	0.80	**0.79**	0.69	0.74	0.78
NUSFocalObj	**0.91**	**0.84**	**0.88**	**0.83**	**0.79**	**0.72**	**0.75**	**0.79**
NUSFocalSal	0.90	**0.84**	0.87	0.82	0.78	0.70	0.74	0.78

we have 16 (images) × 5 (neurons) = 80 images each for positive and negative sentiments. We refer to these images as "positive neurons' images" and "negative neurons' images", respectively. To quantify the the visual difference on these images, we conduct statistical analyses on their sentiment scores and human annotated attributes. Our analyses indicate that **the positive neurons' images of NUSFocalObj are *more positive* than those of NUSNoFocal, and the negative neurons' images of NUSFocalObj are *more negative* than those of NUSNoFocal**. The detailed analyses follow.

First, we perform an independent-samples t-test on the sentiment scores (value ranges between 0 and 1) between the positive neurons' images and negative neurons' images in each of the two models. As shown in Fig. 6 (A (a)), in both NUSNoFocal and NUSFocalSal, the positive neurons' images are more positive than the negative neurons' images, suggesting that both NUSNoFocal and NUSFocalObj have discrimination ability on visual sentiment. More importantly, as illustrated in Fig. 6 (A (b)), t-test shows that the negative neurons' images of NUSFocalObj have lower sentiment scores than those of NUSNoFocal (*i.e.*, more negative), $t(158) = 4.36, p < .001$[4], suggesting that NUSFocalObj has higher discrimination ability on negative sentiment than NUSNoFocal.

To demonstrate the higher discrimination ability of NUSFocalObj, we further perform a series of independent-samples t-tests on the 27 human annotated attributes (provided by [16], the attributes without * in Table 1) between the 80 positive and negative neurons' images of the two models. Bonferroni correction is used due to the large number of comparisons. As shown in Fig. 6 (B), the positive neurons' images of NUSFocalObj are more colorful, familiar and of closer view than those of NUSNoFocal, indicating they are more positive than those of NUSNoFocal (Spearman's rank correlation shows that the attribute "familiar" positively correlates with sentiment score, $\rho = 0.47, p < 0.001$). The negative neurons' images of NUSFocalObj are less colorful, less exciting, and elicit less happiness (*i.e.*, more negative) than those of NUSNoFocal. These observations demonstrate that **NUSFocalObj discriminates visual sentiment better than NUSNoFocal**, suggesting the advantage of the focal channel. Fig. 6 (C) visualizes some of the positive neurons' and negative neurons' images. Due to space limits, only the top two neurons for each sentiment and in each model are shown.

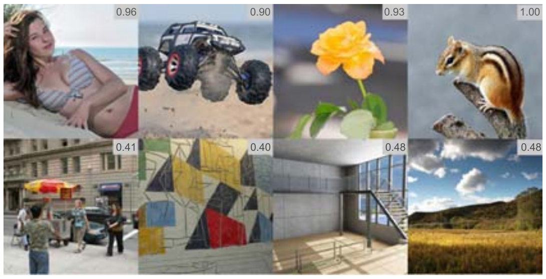


Figure 7: Example images from VRD Dataset that have clear focal objects (top row) and without obvious focal objects (bottom row). The number on the upper right corner indicates the mean ratings on the attribute, "contain objects of focus".

5.1.2 Comparison of Images With and Without Focal Objects. In this subsection, we classify the images of VRD subset by dichotomizing the human annotated attribute "contain objects of focus" with a threshold of 0.75 [20], resulting in two groups of images (the numbers in the parenthesis indicate how many images are in that group): 1) images with clear focal objects (399), and 2) images without obvious focal objects (483). Fig. 7 shows example images from the two groups. As shown in Table 3, NUSFocalSal and NUSFocalObj outperform NUSNoFocal on images with focal objects, whereas for images without obvious focal objects, the performance is more similar. This suggests that focal channel is more effective on images with obvious focal objects. This may be because for images without clear focus, it is difficult for computational algorithms to predict human attention [9], thus diminishing the advantage of the focal channel.

Note that our psychology studies (Sec. 3.3) show that the information evoking negative sentiments is largely present in the focal object. This suggests that our models with a focal channel might be most effective on negative images with focal objects. To test this hypothesis, we investigate the classification accuracy of different models for positive and negative images with and without focal objects. As shown in Fig. 8 (a), NUSFocalObj and NUSFocalSal outperform NUSNoFocal on images with focal objects for negative images but not positive images. For images without obvious focal objects (Fig. 8 (b)), the performances of NUSFocalObj and

[4] We report the results of t-tests as, "t(df) = t value, p = p value". If a p value is smaller than the conventional significance level threshold of .05, we reject the null hypothesis of no difference among the means.

NUSFocalSal are not statistically different from NUSNoFocal. For positive images with and without focal objects, the three models do not significantly differ. These observations suggest that our models with a focal channel are most effective on negative sentiment images with clear focal objects.

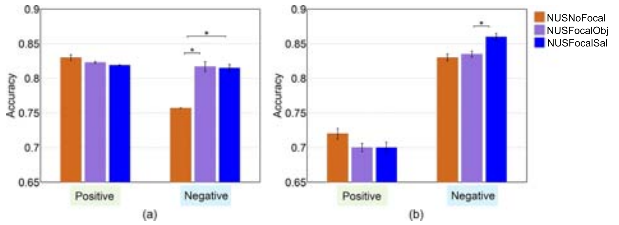


Figure 8: Classification results on (a) images with focal objects and (b) images without obvious focal objects. In both (a) and (b), the left group is the performance of images with positive groundtruth labels, the right group is for images with negative groundtruth labels. The experiments are performed five times for each of the three models on VRD and a series of paired t-test are performed on the five results between different models. Asterisks indicate significant difference between each groups. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Only statistically meaningful differences ($p < 0.05$) are indicated.

5.2 Visualization on the DeepSent Dataset

For the DeepSent Dataset, we select for each model the images with highest confidence of being positive and negative (5 each per model) based on their computational prediction. Figure 9 shows these examples for each model. The labels of top ranked positive images in all models are correctly predicted. For top-ranked negative images, one image from FTCNN and two images from DeepSentiBank are misclassified. Misclassifications suggest that negative images may be more difficult to identify.

We further visualize different neurons in the fc7 layer—the layer before the last fully-connected layer of NUSFocalObj to see what the neurons have learned. The results are reported and discussed in the supplementary material.

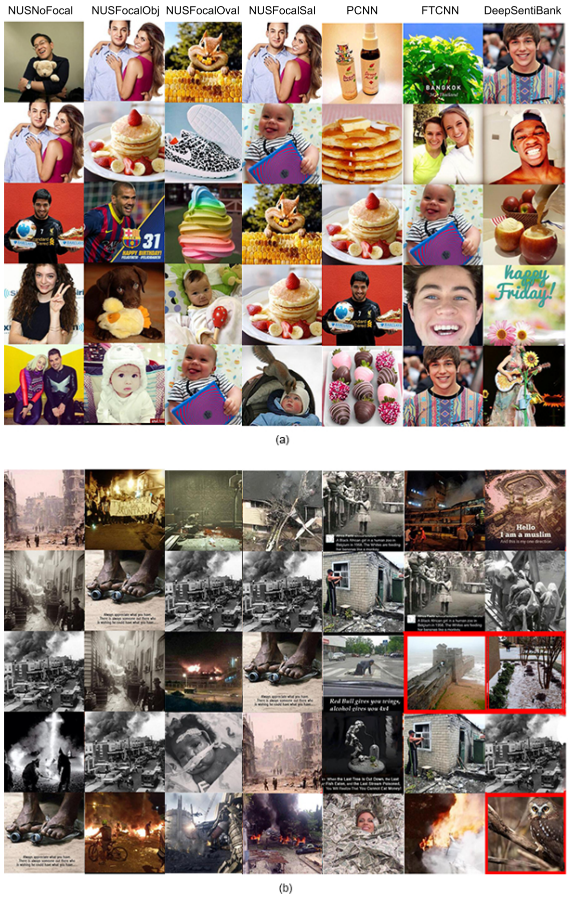


Figure 9: (a) Positive and (b) negative predictions on DeepSent Dataset. Each column shows the images with highest probability for each algorithm. The images are ranked by the prediction score from top to bottom in a decreasing order. Images with red bouding boxes are those misclassified (*i.e.*, the predicted labels disagree with the human labels). The rest are correctly classified on visual sentiment.

6 CONCLUSIONS

In this work we perform psychology studies to empirically evaluate the impact of focal attention on human visual sentiment perception. We discover that negative sentiments are elicited by the focal region without a notable influence of contextual information, whereas positive sentiments are influenced by both focal and contextual information. Based on these findings, we build DNN models for automated assessment of visual sentiment that are augmented with devoted channels for focal information. Our models outperform the state-of-the-art methods on two benchmark datasets. Visualizations of the DNN neurons demonstrate that our models predict visual sentiment better than comparison methods. The analyses on images of different attention patterns echo the findings of the psychology studies by showing that the focal channels are most effective on images with focal objects, especially for images that also elicit negative sentiments.

In the future, we plan to use our understanding of the interplay of attention and emotion to manipulate human affective response by re-targeting human attention [40]. Another interesting application is to apply our models on automated image captioning with sentiments [31].

ACKNOWLEDGMENTS

We would like to thank Dr. Tian-Tsong Ng for helpful discussions. This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative, and a University of Minnesota Department of Computer Science and Engineering Start-up Fund (QZ).

REFERENCES

[1] Charu C Aggarwal and ChengXiang Zhai. Mining text data. In *Springer Science & Business Media, 2012.*

[2] Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Nicu Sebe. Recognizing emotions from abstract paintings using non-linear matrix completion. In *CVPR, 2016.*

[3] Joel Aronoff. How we recognize angry and happy emotion in people, places, and things. In *Cross-cultural research, 2006.*

[4] Rosemary A Bailey. Design of comparative experiments. In *Cambridge University Press, 2008*, Vol. 25.

[5] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM, 2013.*

[6] Tobias Brosch, Gilles Pourtois, and David Sander. The perception and categorisation of emotional stimuli: A review. In *Cognition and Emotion, 2010.*

[7] Neil DB Bruce and John K Tsotsos. Saliency, attention, and visual search: An information theoretic approach. In *Journal of vision, 2009.*

[8] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. *MIT Saliency Benchmark*. MIT, 2017.

[9] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where Should Saliency Models Look Next?. In *ECCV, 2016*. Springer.

[10] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto. From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. In *Image and Vision Computing, 2017.*

[11] Moran Cerf, E Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. In *Journal of vision, 2009.*

[12] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. In *arXiv preprint arXiv:1410.8586, 2014.*

[13] Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In *ACM MM, 2014.*

[14] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. In *Journal of Vision, 2008.*

[15] Paul Ekman. An argument for basic emotions. In *Cognition & emotion, 1992.*

[16] Shaojing Fan, Tian-Tsong Ng, Jonathan S Herberg, Bryan L Koenig, Cheston Y-C Tan, and Rangding Wang. An automated estimator of image visual realism based on human cognition. In *CVPR, 2014.*

[17] Shaojing Fan, Tian-Tsong Ng, Bryan L Koenig, Ming Jiang, and Qi Zhao. A paradigm for building generalized models of human image perception through data fusion. In *CVPR, 2016.*

[18] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. In *ACM Computing Surveys (CSUR), 2016.*

[19] Carlos FA Gomes, Charles J Brainerd, and Lilian M Stein. Effects of emotional valence and arousal on recollective and nonrecollective recall.. In *Journal of Experimental Psychology: Learning, Memory, and Cognition, 2013.*

[20] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *ICCV, 2013.*

[21] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV, 2015.*

[22] Laurent Itti and Pierre F Baldi. Bayesian surprise attracts human attention. In *NIPS, 2005.*

[23] Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. Predicting viewer perceived emotions in animated GIFs. In *ACM MM, 2014.*

[24] Elizabeth A Kensinger. Remembering the details: Effects of emotion. In *Emotion review, 2009.*

[25] Ita GG Kreft, Ita Kreft, and Jan de Leeuw. Introducing multilevel modeling. In *Sage Publication,1998.*

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS, 2012.*

[27] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR, 2006.*

[28] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR, 2014.*

[29] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *IEEE Transactions on Pattern analysis and machine intelligence, 2011.*

[30] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM, 2010.*

[31] Alexander Mathews, Lexing Xie, and Xuming He. SentiCap: generating image descriptions with sentiments. In *arXiv preprint arXiv:1510.01431, 2015.*

[32] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the International Affective Picture System. In *Behavior research methods, 2005.*

[33] Tirin Moore and Marc Zirnsak. Neural mechanisms of selective visual attention. In *Annual Review of Psychology, 2015.*

[34] Ken Nakayama, Julian S Joseph, and R Parasuraman. Attention, pattern recognition and popout in visual search. In *The attentive brain, 1998.*

[35] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. In *Progress in brain research, 2006.*

[36] Stephen E Palmer. 1999. *Vision science: Photons to phenomenology*. Vol. 1. MIT press Cambridge, MA.

[37] Gabriele Paolacci, Jesse Chandler, and Panagiotis Ipeirotis. Running experiments on amazon mechanical turk. In *Judgment and Decision Making, 2010.*

[38] Jane E Raymond, Mark J Fenske, and Nader T Tavassoli. Selective attention determines emotional responses to novel visual stimuli. In *Psychological science, 2013.*

[39] Ulrike Rimmele, Lila Davachi, Radoslav Petrov, Sonya Dougal, and Elizabeth A Phelps. Emotion enhances the subjective feeling of remembering, despite lower accuracy for contextual details.. In *Emotion, 2011.*

[40] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. A comparative study of image retargeting. In *ACM transactions on graphics, 2010.*

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. Imagenet large scale visual recognition challenge. In *International Journal of Computer Vision, 2015.*

[42] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. LabelMe: a database and web-based tool for image annotation. In *International journal of computer vision, 2008.*

[43] Harald T Schupp, Jessica Stockburger, Maurizio Codispoti, Markus Junghöfer, Almut I Weike, and Alfons O Hamm. Selective visual attention to emotion. In *Journal of neuroscience, 2007.*

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

[45] Nathan Sprague and Dana Ballard. Eye movements for reward maximization. In *NIPS, 2003.*

[46] Patrik Vuilleumier. How brains beware: neural mechanisms of emotional attention. In *Trends in cognitive sciences, 2005.*

[47] Patrik Vuilleumier, Jorge L Armony, Jon Driver, and Raymond J Dolan. Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. In *Neuron, 2001.*

[48] Adrian Wells and Gerald Matthews. 2014. *Attention and Emotion (Classic Edition): A Clinical Perspective*. Psychology Press.

[49] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. In *Journal of vision, 2014.*

[50] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks. In *ACM MM, 2016.*

[51] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI, 2015.*