CapVis: Toward Better Understanding of Visual-Verbal Saliency Consistency

HAORAN LIANG, Zhejiang University of Technology, China
MING JIANG, University of Minnesota, USA
RONGHUA LIANG, Zhejiang University of Technology, China
QI ZHAO, University of Minnesota, USA

When looking at an image, humans shift their attention toward interesting regions, making sequences of eye fixations. When describing an image, they also come up with simple sentences that highlight the key elements in the scene. What is the correlation between where people look and what they describe in an image? To investigate this problem intuitively, we develop a visual analytics system, CapVis, to look into visual attention and image captioning, two types of subjective annotations that are relatively task-free and natural. Using these annotations, we propose a word-weighting scheme to extract visual and verbal saliency ranks to compare against each other. In our approach, a number of low-level and semantic-level features relevant to visual-verbal saliency consistency are proposed and visualized for a better understanding of image content. Our method also shows the different ways that a human and a computational model look at and describe images, which provides reliable information for a captioning model. Experiment also shows that the visualized feature can be integrated into a computational model to effectively predict the consistency between the two modalities on an image dataset with both types of annotations.

CCS Concepts: • **Human-centered computing** → **Visualization**; **Visualization application domains**; **Visual analytics**;

Additional Key Words and Phrases: Image captioning, visual saliency, visual analytics

ACM Reference format:
Haoran Liang, Ming Jiang, Ronghua Liang, and Qi Zhao. 2018. CapVis: Toward Better Understanding of Visual-Verbal Saliency Consistency. *ACM Trans. Intell. Syst. Technol.* 10, 1, Article 10 (November 2018), 23 pages.
https://doi.org/10.1145/3200767

This work is an extended version of a previously accepted conference paper, H. Liang et al. Visual-Verbal Consistency of Image Saliency, IEEE International Conference on Systems, Man, and Cybernetics, 2017.
This work is supported by the National Science Foundation of China under grant 61702457 and grant 61602409, a University of Minnesota Department of Computer Science and Engineering Start-up Fund (QZ).
This work was done when Haoran Liang was a visiting student in the Zhao Lab.
Authors' addresses: H. Liang and R. Liang (corresponding author), Department of Information Engineering, Zhejiang University of Technology, 288 Liuhe Rd, Xihu District, Hangzhou, 310013, PR China; emails: {haoran, rhliang}@zjut.edu.cn; M. Jiang and Q. Zhao, Department of Computer Science and Engineering, University of Minnesota, MN, 55455, USA; emails: {mjiang, qzhao}@cs.umn.edu.

2157-6904/2018/11-ART10 $15.00
https://doi.org/10.1145/3200767

1 INTRODUCTION

In the field of computer vision and natural language processing, it is challenging to generate properly formed image captions based on an understanding of image contents. As a vital part of artificial intelligence, image captioning strongly relies on the level of semantic perception of a visual scene. Progress in this task can greatly benefit various real-life applications such as traffic navigation [47], robotics [48], and education [15].

Most works that generate image captions mainly focus on the extraction of features from both images and captions, mapping image region fragments with words in a generative model to produce a caption.

To better describe an image, particularly in a cluttered scene, it is essential to capture the key elements in the image instead of describing everything. Previous studies [9, 17, 37] reveal that visual attention is a helpful proxy for perceiving importance in images. Visual attention is a bottleneck mechanism that allows only a small portion of the visual input to reach higher level processing units. It breaks down a scene into a sequence of localized visual analysis problems. We hypothesize that patterns in image captions strongly rely on what people regard as important. On the other hand, annotations of visual attention offer a natural ranking while a human free-views a scene. Thus, it is of interest to understand the consistency between how people view images and how they describe them.

Visual saliency is the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention. We find that the same pattern exists in human descriptions of a visual scene: A visually salient object is often emphasized in descriptions, which we define as verbal saliency. Furthermore, we observe that the level of visual-verbal saliency consistency can vary despite an overall high correlation between the two modalities (see example in Figure 2). Some natural questions are then: What image features make them more or less consistent? And, is consistency predictable with image features?

Until today, the impact of image contents that decide the difference between visual saliency and verbal saliency is virtually unknown. We consider this difference as a useful signal if modeled instead of treated as noise. The biggest motivation of this work is that applications involving images and text can benefit from an understanding of which images are specific (they elicit consistent descriptions from different people) and which ones are ambiguous (descriptions across people vary considerably). For instance, consider text-based image retrieval. If a query description is moderately similar to the caption of an ambiguous image, that query may be considered a decent match to the image. But if the image is very specific or iconic, a moderate similarity between the query and caption may not be sufficient to retrieve the image because the term and order should be accurate enough.

Visual and verbal saliency are two different modalities of data that cannot be directly compared. For images, it is quite clear to see and understand the content. Also, it is easy and quick to know how people observe images through visual saliency maps. However, multiple human-provided descriptions of the same image are more subjective, and this requires necessary processes of reading and concluding. In addition, comparing the difference between image-level and semantic-level description takes even more time.

We want to be able to quickly gain insight into the difference by converting descriptions (semantic-level) into verbal saliency maps (image-level), showing that there is in fact variance in how consistently people look at and describe scenes. What's more, we also want to know exactly what and how features affect this difference. Therefore, a visual analytics system is needed to dig into the problem to find possible explanations and applications for visual-verbal saliency consistency.

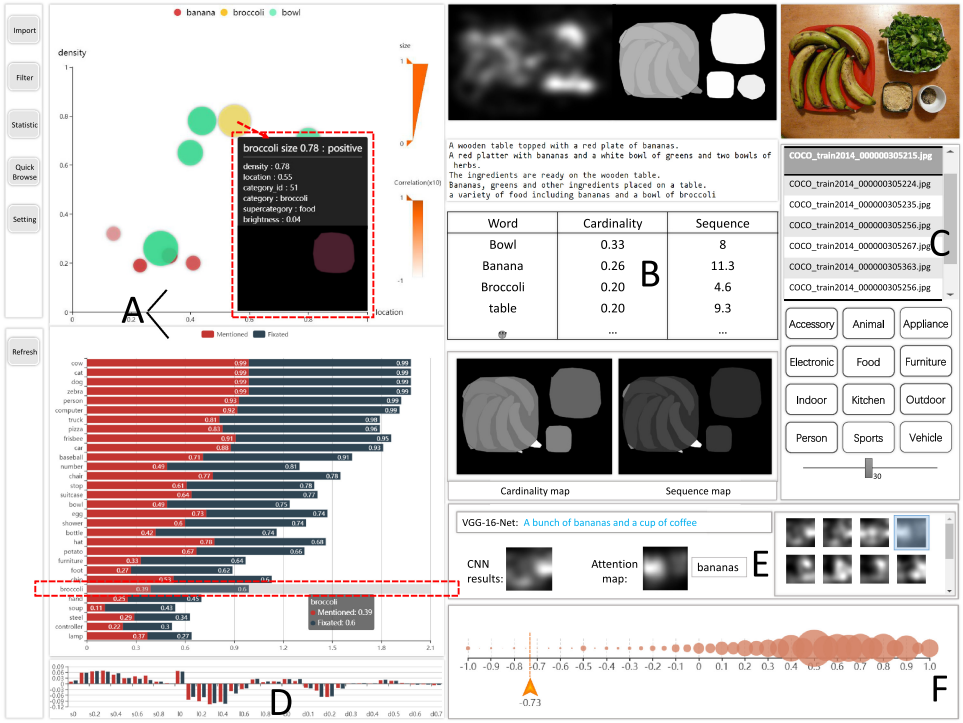

Fig. 1. An overview of CapVis, an analytics system for visual-verbal consistency of image saliency. (A) Graphics module that converts multidimensional data into visible graphs. (B) Saliency map module that shows both visual and verbal saliency maps along with statistics of words in captions. (C) Image selection module that presents images filtered using given options. (D) Histogram module that provides the impact factors of different features. (E) Deep feature visualization module that shows region-to-word mapping from the results obtained from a deep neural network. (F) Graphics module that visualizes the visual-verbal correlation score of the selected image.

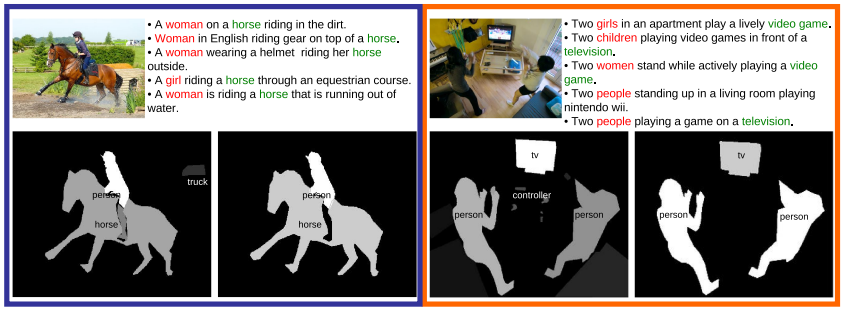

Fig. 2. Examples of high-consistency (blue box) and low-consistency (orange box) patterns between how people look at and describe images. For each case, the grayscale map on the left indicates visual saliency, and the one on the right indicates verbal saliency. The brightness of the object mask refers to the visual/verbal salient value: An object with a brighter mask means that it is more salient than others. In the blue box, the visual saliency map is quite similar to the verbal saliency map, while in the orange box, people tend to fixate on the TV but describe the children first.

In this work, we develop an analytics system, CapVis, to help study the relationship between the words used in image captions and the regions where people look in natural images, which is the main contribution of this work. First, we explicitly define verbal saliency as the importance of the words in a sentence and report the correlation between verbal saliency and visual saliency. Second, we propose a number of features based on image and object information and quantify their effects on the consistency between visual and verbal saliency. By using visualization techniques, we are able to show the correlation quantitatively and qualitatively. We also show that CapVis can be used to diagnose captioning models by comparing attention maps from human subjects and models. Finally, we demonstrate the effectiveness of the visualized features in predicting visual-verbal consistency using a Support Vector Regression (SVR).

2 RELATED WORK

2.1 Image Saliency and Captioning

The computational model of visual attention was first introduced by Itti et al. [20]. They proposed to use a set of feature maps from three complementary channels: intensity, color, and orientation. The normalized feature maps from each channel were then linearly combined to generate the overall saliency map. Based on this, many other researchers have suggested various models that can be categorized based on the algorithms used, such as Bayesian models [57], information theoretic models [4], graphical models [14], spectral analysis models [18], and pattern classification models [24]. In addition to these early features, high-level image semantics have been shown to be much more relevant to image saliency [53]. Therefore, deep neural networks have been proposed to reduce the semantic gap [19, 33] by naturally integrating hierarchical features for saliency prediction.

Recently, also supported by deep neural network models, the methods of image-sentence retrieval [12, 44] and image captioning [11, 26] have been developing dramatically. Many works [23, 25, 35] using a multimodal Recurrent Neural Network (RNN) achieve state-of-the-art performance for the tasks of both image-sentence retrieval and image captioning. Moreover, Xu et al. [54] first correlated a computational captioning model with visual attention by incorporating two mechanisms of attention into the neural networks and demonstrated that the model can selectively focus on certain regions of an image that align with the words in the captions. Ramanishka et al. [39] proposed caption-guided visual saliency to expose the region-to-word mapping in modern encoder-decoder networks. However, their modeling of visual attention was not verified against ground-truth annotations.

2.2 Visual Analytics of Image Importance and Description

The use of a coherent set of keywords for characterizing a particular concept has wide applicability in various document analysis tasks. Many works have focused on word-level content analysis, such as sentiment analysis [36, 46] and topic modeling [6–8, 10, 16, 29, 30, 51]. All these works are completely based on text, so they cannot be used for our system since we also consider a visual reference (i.e., an image).

Image captions and importance are essential parts of an image-based deep learning task, such as image Q&A [13, 27, 34]. Researchers have developed various analytics systems [32, 50, 55] to understand, diagnose, and refine deep networks, but they have not focused on bridging the gap between visual perception and language processing.

Very few works have focused on understanding and visualizing image importance [1] and specificity [21]. Most of them stopped at exploring how a number of factors relate to human perception of importance based only on an image caption. Our work first attempts to use visualization

technique to explore features that contribute to the consistency between how people view and describe an image. We examine this consistency by bridging visual saliency and image description on a large dataset with 15,000 images.

3 REQUIREMENT ANALYSIS

Previous works studied visual saliency and image captioning using common metrics such as Area Under the Curve (AUC) and Bilingual Evaluation Understudy (BLEU) to evaluate the accuracy of a model. Visually analyzing image saliency has been a necessary part of the biologically inspired vision domain, while image captioning is mostly analyzed in a metrics-based way. However, the metric score is obtained from a set of feature values that do not intuitively give us a visual impression about image content or sentence structure. On the other hand, sometimes two different patterns will have very similar feature values that require visual analysis for better understanding.

In this study, we recruited one researcher in data visualization studies and two experts in visual saliency and image captioning studies for a requirements analysis. To visually analyze visual-verbal saliency consistency, we summed up the following requirements according to our discussions with the researcher and experts.

R1 - Showing the consistency/inconsistency between visual and verbal saliency. At the beginning of our analysis, it is necessary to provide easily perceivable information to judge whether the visual and verbal saliency are consistent. The specific requirement is that not only experts, but also general users should be able to quickly see the difference between two types of saliency and, at the same time, grasp brief information (number, size, category) about objects in the image. This basic information is important at the very beginning of analysis for finding problems and setting goals. Unlike the widely used saliency map for visual saliency, currently we do not have a defined "map" to observe when given just a couple of sentences for an image. In this case, we need to keep the verbal saliency map in line with the visual one for the sake of intuitively showing the difference.

R2 - Exploring the potential low- and high-level features that play important roles when people look and describe. The image should be completely examined based on the collected annotations so we can focus on different facets of image content. The annotations allow us to obtain different information about every object in the image, such as size, location, and category. Processing and visualizing these information will help us quickly understand and determine what's important in the image. So, enough elements must be present that can interactively express all the features in our proposed system and let users investigate the difference between visual and verbal saliency. In addition, the visualizations should be intuitive enough to bridge figures and original images. Except for readability, our experts also suggested that the visualized data be quantified properly for machine learning tasks such as recognition and prediction.

R3 - Revealing how features contribute to the consistency/inconsistency of visual-verbal saliency. Previous research has shown several features of concern during the process of image captioning, while our experts believed that pointing out essential elements within images is not enough for the purpose of analyzing the contribution of each feature. For example, we will want to know whether a certain feature is relevant to the consistency/inconsistency of visual-verbal saliency. If it is, we need to know whether the contribution is positive or negative. Therefore, the extent of features' influences should be visualized quantitatively so we can gain a more comprehensive understanding of visual elements in the image. By visualizing the feature contribution to visual-verbal saliency consistency, we are able to select representative image features that can be further applied to improve applications involving images and text.

R4 - Targeting potential pitfalls of the deep captioning model. Previous work [54] proved that visual attention is useful in building a deep network for generating image captions. Deep

captioning models utilize convolutional neural networks to perceive localized image features, followed by a recurrent neural network to generate words based on the learned feature alignment. The feature alignment must correspond well to human intuition to obtain accurate results. The visualization of the verbal saliency map graphically shows the way that people describe an image, which can be considered a valuable reference for building and diagnosing a captioning model. Thus, our experts emphasized the necessity of examining failure cases caused by a mismatch of attention and word.

4 SYSTEM OVERVIEW

In this section, we propose our design of an analytics system (Section 4.1) and introduce the dataset used in this work. (Section 4.2)

4.1 System Components

Motivated by the requirements, we developed CapVis, an analytics system to analyze and evaluate visual-verbal saliency consistency. An overview of the system design is shown in Figure 1. We introduce the modules of CapVis as follow:

- A: A graphics module that converts multidimensional data into visible graphs. The scatter plot and histogram in this module visualize low and semantic levels of the image features to let users view the image content statistically (see details in Section 6.2, Section 7.2). **(R1)**
- B: A saliency map module that shows both visual and verbal saliency maps along with statistics of words in captions. From this panel, the process of generating the verbal saliency map is presented to show the mapping between object and word. The system will automatically calculate the weights of words based on the two methods described in Section 5.2, while the specific value of each weight remains editable in order to allow manual correction when necessary. **(R1, R2)**
- C: An image selection module that presents images filtered using given options. Users can combine several different options using the category button. Once an image inside the list box is selected, all the relevant information and analyses will be presented on other modules (see details in Section 6.3). **(R1)**
- D: A histogram that provides impact factors for different features. It gives information about whether a certain feature contributes positively or negatively to visual-verbal consistency (see details in Section 6.2). **(R3)**
- E: A deep feature visualization module that shows region-to-word mapping from the results obtained from the deep neural network. This module also shows the weight of each word from the captioning solver. (See details in Section 5.3, Section 7.3.) **(R4)**
- F: A graphics module that visualizes the visual-verbal correlation score (see details in Section 5.1) of the selected image. It allows the user to gain an overview of the correlation score distribution of selected category. **(R3)**

With this analytic visualization system, users can view the different patterns people used to describe an image as they looked at it from their different visual references. For example, after selecting an image in the list of module C, users can quickly know how people look at and describe the image in module B by exploring the visual saliency map, the weighs of words in the table, and the obtained verbal saliency maps. Users can get the attention maps and caption obtained from the captioning model in module E to compare the results between the human and computational models. The distribution of the correlation score for visual-verbal consistency for the dataset will be plotted on module F, in which an arrow will indicate the correlation score of the current image.

To identify the factors that lead to the correlation score, users are able to explore the low- and semantic-level features of all objects within the image in module A and D.

4.2 Datasets

Studying the consistency between visual saliency and verbal saliency requires an image dataset with captions, object annotations, and saliency annotations. Popular datasets such as MIT1003 [24] and Toronto [3] that contain images and corresponding eye-tracking data have been used extensively to evaluate saliency models. In terms of captioning, the most frequently used datasets are Flickr8K [40], Flickr32K [56], and MS COCO [31]. However, none of these datasets consists of both saliency annotations and captions. To satisfy this requirement, one can choose to collect eye-tracking data on image captioning datasets or to annotate image captions on saliency datasets. The first approach requires the collection of eye-tracking data on large-scale image sets. With traditional eye-tracking devices, it is difficult to conduct large-scale human experiments because the cost and experimental period are likely to be prohibitive. For the second approach, on the other hand, most saliency datasets contain no more than hundreds of images, which limits making general conclusions on how people describe images.

Recently, a mouse-tracking paradigm has been proposed by Jiang et al. [22] for large-scale saliency annotation. They created SALICON, a large dataset consisting of 10,000 training images and 5,000 validation images from the MS COCO. The flexibility in their experimental setting and the strong correlation with eye-tracking data demonstrate that mouse-tracking data provide a good source of ground truth for visual attention research. Therefore, we investigate the consistency between the verbal importance of words and build the visualization system based on the image captions with saliency annotations in the SALICON dataset.

5 SALIENCY MAP GENERATION

Given an image with object annotations, to measure its visual-verbal consistency, we obtain saliency values from fixations and captions, assign them to the objects, and compare them using Spearman's rank correlation ρ. That is, for image I with N objects, we denote each object as o_i, where $i \in 1, 2, \ldots, N$. Two types of saliency values, visual saliency value $V(o_i)$ and verbal saliency value $W(o_i)$, are computed for o_i from attentional data and image captions, respectively. The visual-verbal consistency is computed as

$$\rho = Spearman(V(I), W(I)), \tag{1}$$

where $V(I) = \{V(o_1), V(o_2), \ldots, V(o_n)\}$ and $W(I) = \{W(o_1), W(o_2), \ldots, W(o_n)\}$, *Spearman* refers to the Spearman's rank correlation coefficient that measures the statistical dependence between the ranking of two variables.

The remainder of this section introduces the methods used to generate a visual saliency map (Section 5.1), verbal saliency map (Section 5.2), and spatial attention map (Section 5.3).

5.1 Visual Saliency Map

The saliency map obtained from attentional data can be directly applied to measure the importance of objects in an image based on the blurred attention map generated using mouse tracking data. We constructed a fixation map of each image by convolving a fovea-sized (i.e., $\theta = 26$ pixels) Gaussian kernel over the successive fixation locations of all subjects and normalizing it to sum 1, which can be considered as a probability density function of eye fixations. Figure 3 shows the process of generating the visual saliency map. In particular, given an input image and its fixation map, the saliency of each object in the scene was computed by taking the maximal value of the fixation

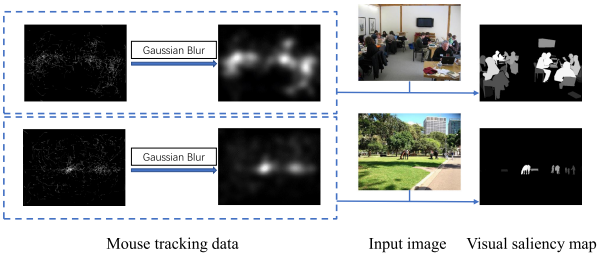

Fig. 3. The process of generating visual saliency maps. We convolve a fovea-sized Gaussian kernel over the fixation locations of all subjects and take the maximal value of the fixation map inside the object mask as the saliency of each object.

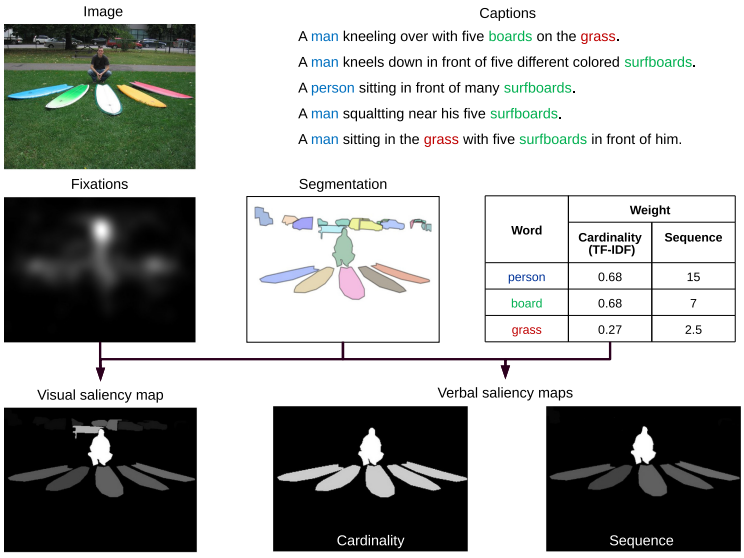

Fig. 4. The generation of verbal saliency maps.

map within the object's segmentation mask; that is, $V(o_i) = \max(P_i)$, P_i denotes the set of saliency values (obtained using mouse tracking in Jiang et al. [22]) inside the ith object mask.

5.2 Verbal Saliency Map

In order to know which objects are described and where they are, we propose a mapping method between words from a caption and objects in an image. Figure 4 demonstrates the approaches of computing verbal saliency from image captions. In particular, we use the leading platform Natural Language Toolkit (NLTK) [2] to parse all the captions into a big vocabulary set that consists of each noun with the help of the Stanford Log-linear Part-of-Speech Tagger [49]. We use a simple Wordnet-based measure of semantic distance [52] to find which of the 80 categories these words belong to. We initialize the salient value for each image using its corresponding 5 captions in two strategies; namely, cardinality-based and sequence-based approaches.

Cardinality-based approaches measures the importance of each word based on the frequency of occurrence. In order to weight the count of each word into floating-point values, the importance (also denoted as the initial saliency value $W_{init}(o_i)$) is computed using Term Frequency-Inverse Document Frequency (TF-IDF) from the scikit-learn software package [38]. Words that are rare

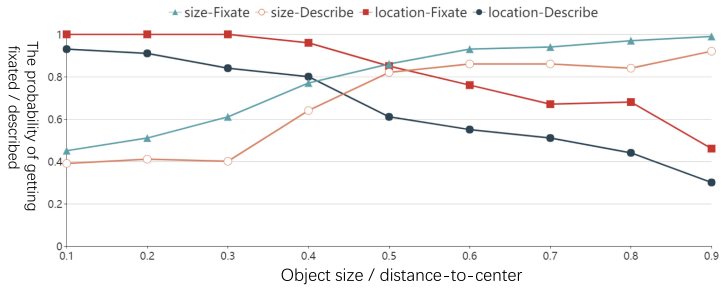

Fig. 5. The effects of location and size on both visual saliency and verbal saliency. The probabilities of getting fixated and described increase significantly when an object gets larger and nearer to the center of image.

in the corpus but occur frequently in a sentence contribute more to the importance. Let J_i be the number of times that the word for object o_i occurs in d sentences, then the term frequency of the object is calculated as

$$W_{init}(o_i) = \frac{J_i}{\sum_{i=1}^{N} J_i}. \tag{2}$$

Sequence-based approaches capture and highlight the order of each word in a sentence. We denote the number of categories in an image as P and the order of a word(object o_i) in an image as q_i, e.g., $q_i = 1$ means that the word comes first in the sentence, so its weight is $\frac{P}{q}$, which is higher if a word comes at the beginning of a sentence and lower otherwise. Therefore, for object o_i in d sentences, the initial saliency value will be

$$W_{init}(o_i) = \sum_{i=1}^{N} \frac{P}{q_i}. \tag{3}$$

These two approaches provide reasonable initializations of the verbal saliency values. Yet ambiguities still exist. For example, with multiple people in a scene, we cannot tell who "the man" in the caption refers to. Furthermore, the mappings from words to objects are not one-to-one. For example, "the vehicle" may represent either a car or a truck that co-exists in an image. To approach this problem, we propose a method based on psychophysical studies. In particular, we borrow the observations from Berg et al. [1] to add to the verbal saliency map by considering the compositional features relating to objects (i.e., size and location). Visual attention is well known to have a center bias and a preference for dominant objects, and image captions also have a similar tendency. Figure 5 displays the effects of location and size on description probability for the training set of SALICON. We can see clearly that small size and a long distance to image center decrease the probability of getting mentioned or fixated and vice versa. With these observations, we adjust the saliency values of objects with an additional term $T(o_i) = S(o_i) \times (1 - L(o_i))$, where $S(o_i)$ is the normalized value of size and $L(o_i)$ is the normalized distance to center, respectively. The finalized value is calculated as $W(o_i) = W_{init}(o_i) \times T(o_i)$, which is used as the ground truth for verbal saliency.

In order to investigate inter-subject consistency in image captioning, for each image in the training set, we use one of the five captions and the rest to generate two verbal saliency maps and calculate the Spearman correlations, resulting in a human consistency of $\rho = 0.865$. By comparing the ground truth between visual saliency and verbal saliency, we obtain $\rho = 0.435$ and $\rho = 0.439$ ($\rho = 0.361$ and $\rho = 0.368$ without adjusting size and location) for cardinality and sequence-based schemes, respectively. The distribution of all images is shownat the bottom of Figure 6. In this single axis scatter plot, the size of node refers to the number of images that have close correlation

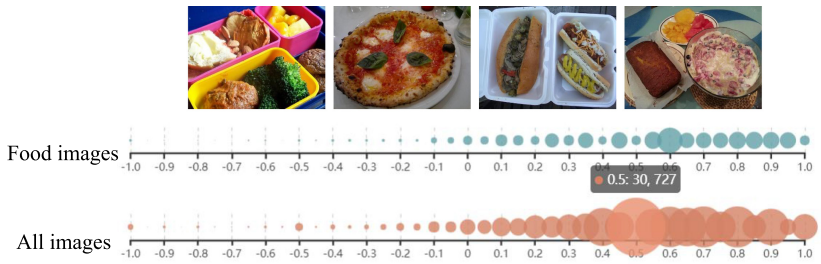

Fig. 6. The distribution of Spearman correlation score for food-related images (top) and all images (bottom). For the tooltip, 0.5 refers to the Spearman correlation score, 30 is the index of this node in the plot, 727 means that 727 images in this dataset have a score of 0.5 and this number controls the size of the node in this module.

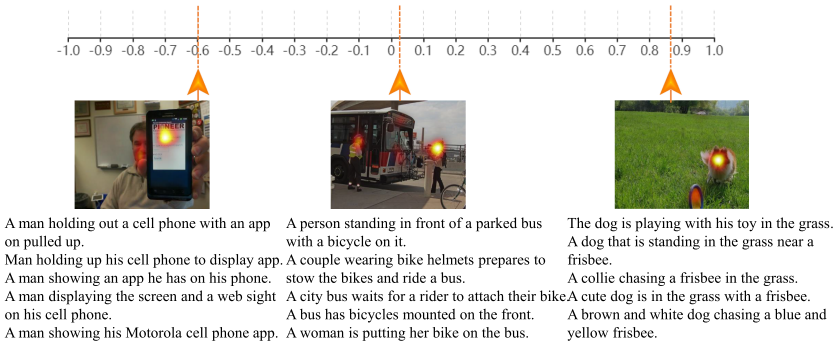

Fig. 7. Example of images with different correlation scores. From left to right, the three images represent "negatively correlated (-0.6)," "uncorrelated (0.02)," and "positively correlated (0.87)," respectively. The highlighted regions indicate the salient part. In the first image, the cellphone attracts attention, while people always describe the man first. In the second image, people seem to have different choices to describe and watch. In the third image, people always describe exactly the most salient object.

scores; that is, the visual saliency in 727 images is positively correlated with the verbal saliency because the correlation scores are around 0.5. Although there are some differences between the whole and part of the dataset, most images are distributed from 0.3 to 0.7.

Figure 7 shows an example of images with different correlation scores. In CapVis, a gauge chart is used to intuitively demonstrate the value of a correlation score.

5.3 Spatial Attention Map from Deep Neural Network

By using a weighted combination of the convolutional feature maps, previous works [41, 45] have achieved attention locating when an image is fed into a deep neural network. In our analytics system (Component F), we adopt the approach of Selvaraju et al. [42] to generate word-specific spatial attention maps. Grad-CAM has achieved state-of-the-art performance for visual element localization in images and shows robustness for visualizing deep convolutional neural networks.

For image I, a target element (word) w, and K feature maps A^k extracted from a certain layer in a trained CNN model. In Grad-CAM, the image I is first propagated forward through the trained CNN model; after that, Grad-CAM generates the spatial attention map $L(I, w)$ using a weighted

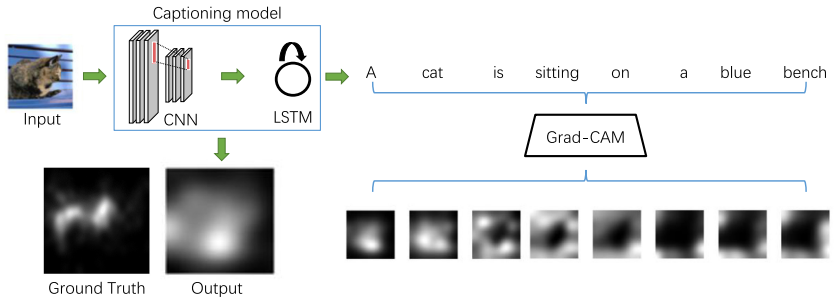

Fig. 8. The design of our system to see the mapping between words and regions (attention) from a captioning model. The input image is first fed into a captioning model that consists of a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) network to produce a caption that describes the image. Then the outputs of the CNN are used to generate spatial attention maps that have one-to-one correspondences with the words through Grad-CAM. The spatial attention maps can then be examined and compared to the visual saliency ground truth by a user to see whether the computational model focuses and extracts features from the correct region.

combination of the convolutional feature maps as follows:

$$L(I, w) = ReLU\left(\sum_k \alpha_k^w A^k\right). \tag{4}$$

The weight α_k^w captures the importance of the kth feature map for the element w and is calculated by backpropagating gradients to the convolutional feature map A^k. Prior to the backpropagation operation, vector quantization is performed on the gradients for the penultimate layer (the layer before softmax) of the CNN model where the dimension of element w is set to 1 and the remaining to 0. The gradients flowing back to A^k are global-average-pooled to obtain α_k^w. More details of Grad-Cam can be found in Selvaraju et al. [42]. Figure 8 demonstrates the design in CapVis to see the mapping between words and regions from a captioning model.

6 FEATURE VISUALIZATION

In this section, we report the statistics of the dataset (Section 6.1). Next, we propose visualization methods (Sections 6.2, 6.3) for a number of potential features for estimating the consistency between visual saliency and verbal saliency, including low-level features (i.e., size, location, density) and semantic-level features (i.e., categories).

6.1 Dataset Composition

We first count the number of each Part-of-Speech (POS) in the image descriptions for the 10,000 training images. Each word is converted to its basic form before the count to make sure the element in our vocabulary list is unique. Figure 9 shows the distribution of POS of all the words in the vocabulary list that people use to describe the training images. Obviously, nouns (NN) stand out as the dominant parts in the descriptions, followed by verbs (VB) and adjectives (JJ). Numerals (DT) and prepositions (PR) may occur quite frequently, but they are less relevant to the image content. Verbs and adjectives apparently relate to the perception of importance and seem worth consideration. However, they are used to describe those attributes that eventually lead us back to the particular objects they derive from. Hence, we only consider nouns in the following analysis. We also show the top 20 frequently used nouns in Figure 9, and we think that the results obtained using frequently used nouns are more convincing than using rare nouns.

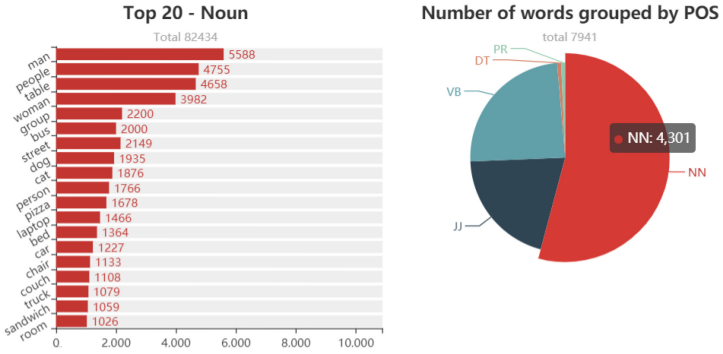

Fig. 9. The dataset composition. The bar chart shows the numbers of top-20 frequently used nouns in this dataset. The pie chart gives information about the number of words grouped by part of speech. Note that each word is converted to its basic form before the count to make sure the element in our vocabulary list is unique.

6.2 Low-Level Features

The average correlation scores of $\rho = 0.435$ and 0.439, respectively, for cardinality and sequence methods show a significant correlation between visual saliency and verbal saliency. To investigate this, we propose three low-level and two semantic-level features extracted from all the objects based on the object annotations of the image. Particularly, given an image I with N objects, let o_i be the ith object. Then, we define low-level features for objects (size, location, and density) as follows:

Size: The size $S(o_i)$ of an object can be measured as the number of pixels in the object mask. We normalize the size by image resolution:

$$S(o_i) = \frac{o_i\_size}{Image\_size}. \tag{5}$$

Thus, each object will have a rounded size value that ranges from 0.1 to 1.0, resulting in a 10-dimensional vector that encodes the numbers of objects of different sizes in an image (e.g., if there are only two objects with size-0.2 in an image, the second variable of the size vector will be 2). In this case, no matter how many objects are in the image, we can always represent the size information using a 10-dimensional vector.

Location: The center coordinate of an object bounding box is used as the exact location for one object. In our feature setting, we use the relative location $L(o_i)$ of an object, which is defined as the distance to the image center:

$$L(o_i) = \frac{dist(o_i, Image\_center) \times 2}{Image\_diagonal}. \tag{6}$$

Similarly, with all the distances ranging from 0.1 to 1.0, the descriptor is a 10-dimensional vector that encodes the numbers of objects in different locations in an image.

Density: Object density consists of the objects' distances to each other. From the segmentation masks of all the objects, the object density can be computed as:

$$D(o_i) = \frac{\sum_j^N dist(o_i, o_j)}{Image\_diagonal \times N}, \tag{7}$$

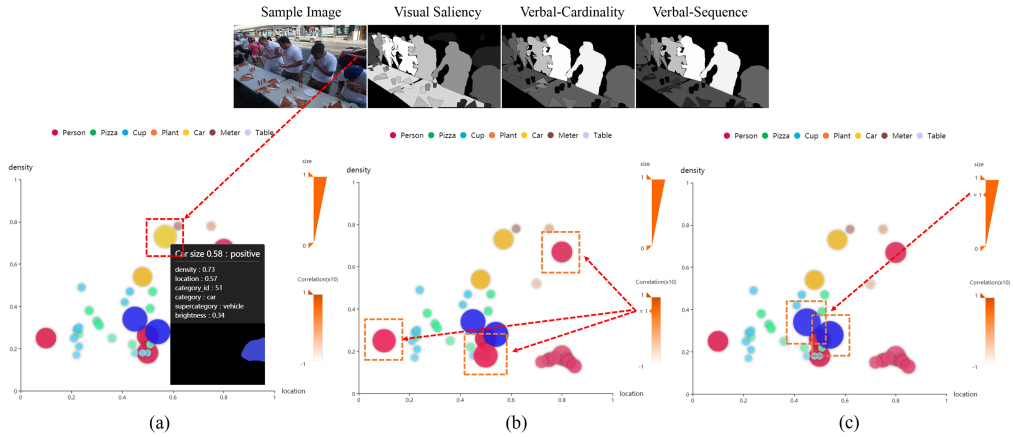

Fig. 10. The scatter plot shows multiple low-level features. One node in the scatter plot represents one object within the image. The X-axis and Y-axis represent the normalized location and density of object, respectively. (a) When the user moves a mouse onto a node, a tool tip will pop up to show the statistics of the object along with a map showing the object mask to help the user locate the object. (b) The brightness of a node indicates the contribution of an object to the visual/verbal correlation. (c) The size of a node represents the normalized size of an object.

where $dist(o_i, o_j)$ denotes the Euclidean distance between the ith and jth objects. After normalization, we use a 20-dimensional vector to encode the numbers of objects with different densities in an image from 0.05 to 1.0 with a step of 0.05.

Next, we independently investigate the contribution of each low-level feature to the visual-verbal consistency. Across all training images, we compute the Pearson's linear correlation coefficients between the feature values and the visual-verbal consistency scores for both the cardinality-based and sequence-based approaches. In this analysis, a positive correlation suggests that the corresponding feature channel contributes positively to the visual-verbal consistency and vice versa. An example of visualization can be found in Figure 1(F), which shows the contribution of each feature obtained from food-related images.

Figure 10 shows the colored scatter plot (module A) we use to visualize low-level features. Specifically, the X-axis refers to the location of the object, Y-axis refers to the density of the object, and the size of node refers to the size of the object. In addition, once selected, the node on the plot will show a pop-up (Figure 10, left) that presents the specific value of each feature along with a segmented image that highlights the referred object. Bars on the right are used to look for objects with close correlation coefficients (Figure 10, middle) and sizes (Figure 10, right).

Across all training images, we compute the Pearson's linear correlation coefficients between the feature values and the visual-verbal consistency scores for both the cardinality-based and sequence-based approaches. In this analysis, a positive correlation score suggests that the corresponding feature channel contributes positively to the visual-verbal consistency and vice versa. The t-test results are shown in Figure 11. This is a static module (D) in CapVis that stands for a reference showing how each feature channel contributes to the visual-verbal saliency consistency when users are exploring the features in module A.

Back to the scatter plot in Figure 10: The brightness of a node demonstrates the average coefficient score shown in Figure 11. For example, an object with size 0.2, location 0.4, density 0.6 (in Figure 11, the corresponding values are 0.065, −0.095, 0.0033, respectively) will have a brightness

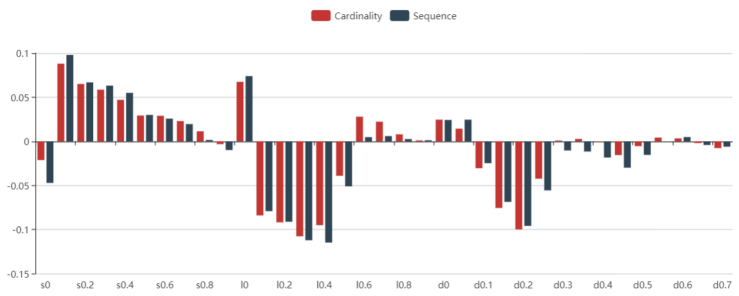

Fig. 11. Correlations of size (s-), location (l-), and density (d-) features for both cardinality-based and sequence-based verbal saliency. Feature channels with all-zero values are skipped. A positive correlation score suggests that the corresponding feature channel contributes positively to the visual-verbal consistency and vice versa.

value of -0.09 $((0.065 - 0.095 + 0.0033)/3 * 10)$. Note that the value of brightness can also be changed to one of the three feature values in the system menu.

6.3 Semantic-Level Features

There is psychophysical evidence [9] that human observers tend to fixate on semantic categories such as people and animals. In this group of features, we wanted to see whether the presence of different categories affects the way people describe an image. We sort the noun list by how many times a noun appears in the image captions in a descending order. We notice that the top 100 nouns are quite commonly used in daily life (i.e., "man," "dog"). For each noun, we look for all the images that contain the described object, along with the corresponding captions. Given a noun mapped to an object, we compute its probability by dividing the number of captions that contain this noun or its synonym by the total number of sentences. Note that summative words are not considered in this case ("a lot of dishes on the table" does not mean an egg (if it exists) is mentioned since we focus more on sibling terms like "boy" for "person," "ship" for "boat"). For the probability of fixations, we simply count the number of fixations inside the object mask and then perform a cut-off at a threshold θ to exclude noises. In our experiments, θ is set at 10 to 30 based on the object size.

In CapVis, the two types of probabilities are plotted in a stacked histogram (module A) as shown in Figure 1; the red bar and blue bar represent the probabilities of being mentioned and fixated, respectively. The values are sorted in ascending order by probability of being fixated. The content of this module is controlled by module C. Once a category in module C is selected, the histogram will show a number of related categories according to the selection. Additionally, the object category in the image that is currently being investigated will always be listed in the histogram.

7 APPLICATION

Two experts (EA and EB) in visual saliency and image captioning studies from two universities were asked to work on this study, identify research problems, and collect design requirements. The system was iteratively improved throughout frequent meetings. The case studies were conducted when the system was ready. The experts provided interesting insights into the research findings.

In this section, we first introduce the workflow of our system (Section 7.1). Next, we present the case studies to demonstrate how CapVis is used to explore the influences of low- and semantic-level features (Section 7.2). Then we show how CapVis helps experts to diagnose captioning models (Section 7.3). We also seek to predict the correlation score between visual and verbal saliency (Section 7.4). Finally, we make our observations (Section 7.5) and conclude the user feedback (Section 7.6).

7.1 System Workflow

Here, we introduce the workflow of our system.

1. After importing the dataset, the image list appears in module C, from which the user can select every single image for investigation. Users can choose different image sets using the category buttons below the image list.
2. Once an image is selected, module B will show the eye-fixation and visual saliency maps on top, the descriptions written by a human in the middle, and the weight of each noun calculated using two methods (cardinality and sequence), along with verbal saliency maps, on the bottom. Module A will show low- and semantic-level features on the top and bottom, respectively. Module E will show the corresponding saliency map obtained by the deep model and the word-based attention maps. Module F will present the correlation score of the image.
3. To investigate low- and semantic-level features of an image, users have to move the mouse on to each node in the scatter plot in module A to see the information provided by a tool tip. By hovering the mouse over the bars on the right, the corresponding nodes will be highlighted. The corresponding category of selected node will also be highlighted in the histogram below.
4. To investigate how the deep captioning model looks and describes, users have to select the attention map on the right of module E or directly click the exact word in blue to see the correspondence between word and attention map.

7.2 Case Study: Influence of Low- and Semantic-Level Features

This case study was a collaboration with expert A (EA). EA is focused on image captioning. He worked with us on the previous work on visual-verbal saliency analysis. We used the obtained visual and verbal saliency to calculate the correlation score and used different levels of features to find those factors that contribute to visual-verbal consistency. The results we obtained were just the image with its correlation score and a set of numbers representing the features. EA thought that the feature values were not intuitive for him to make conclusions, forcing him to look again at the image content, and it took him a lot of time to connect the feature values with objects.

Module A of CapVis is developed to convert feature values into visible nodes and bars. EA wanted to find the most important object in the image after knowing the correlation score. After importing the image, he was able to quickly locate the key object on the scatter plot. As shown in Figure 12(a,b,d), images with positive or negative correlation scores mostly have a lower object number and category. Additionally, salient objects are mostly distributed in a small region. On the contrary, if an image contains a variety of different objects with quite similar low-level features (Figure 12(c)), it will be hard to select the key point in the image, which leads to an uncorrelated correlation score. "Unlike just showing the feature values on object masks directly on the original image, the scatter plot can be seen as a high-level feature filter to let users see the basic structure of an image," EA commented.

To investigate one of the low-level features alone, EA switched the brightness setting of the nodes to only one feature channel instead of three so that he could make observations over different features. The observations are listed next.

First, images with medium-sized objects (s-0.2 to s-0.5) have particularly more consistent visual saliency and verbal saliency. These feature channels contribute significantly to the difference in both cardinality and sequence cases. The downward trend of feature contribution from s-0.2 to s-0.9 demonstrates that visual saliency and verbal saliency become less similar when large objects are presented. It is most likely that large objects are in the background and visually less salient, but are nevertheless important for describing the context of a scene.

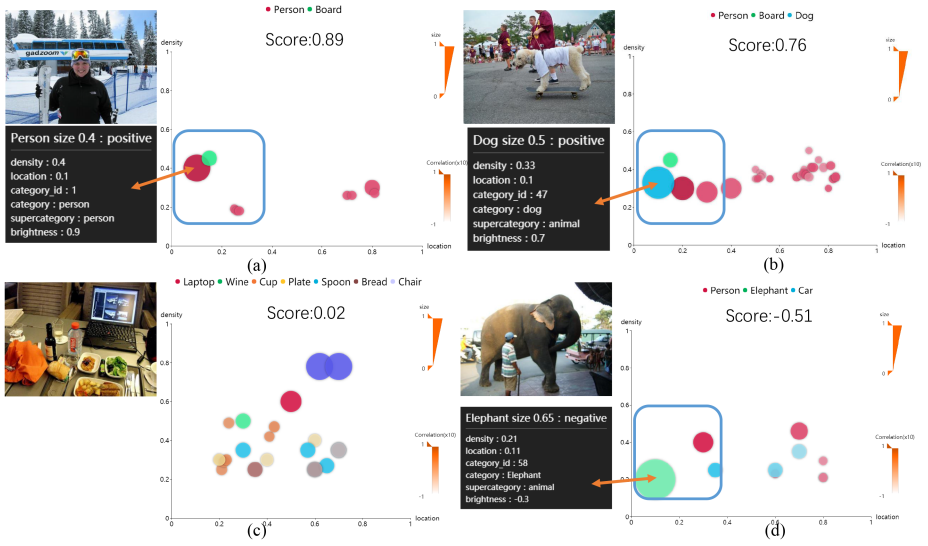

Fig. 12. Examples showing the image in the scatter plot (module A) in CapVis. In most positively or negatively correlated images, we can easily find salient objects inside the blue rectangle in (a) (b) (d). On the other hand, it is hard to point out the key point of image content in some uncorrelated cases, such as (c).

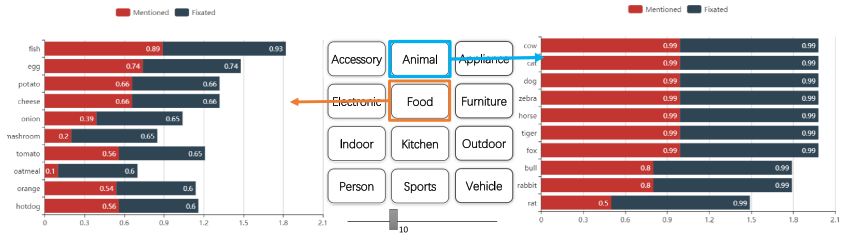

Fig. 13. The probabilities of being fixated (blue bar) and mentioned (red bar) for different categories in module A, selected by the image selection module. The module interface is in the middle; the scroll bar at the bottom controls how many categories to show. Once a button is clicked, the chart will show the fixate/mention probabilities of several related objects.

Second, images with objects close to the center (l-0.1) are significantly more consistent between visual saliency and verbal saliency. As objects get farther from the image center (l-0.2 to l-0.5), they contribute negatively to the visual-verbal consistency. However, when the images contain more objects near the edges (l-0.7 to l-0.9), probably caused by a large dominant object occupying the central region, the saliency ranks become more consistent.

Finally, images with low object density (d-0.35 to d-0.6) are more consistent between visual saliency and verbal saliency, while images with high object density (d-0.15 to d-0.25) are less consistent between visual saliency and verbal saliency. To be specific, sparse contents within images obtain more similar description, while cluttered ones do not.

EA was also curious about the influence of semantic-level features. By using the image selection module C, he randomly chose 10 nouns (see example in Figure 13), each time from the vocabulary set, to investigate the features, from which several observations could be drawn: Fixations cover more contents than image captions because almost all the probabilities from fixations are higher. As shown in previous work [1], humans have quite similar tendencies toward living things both in

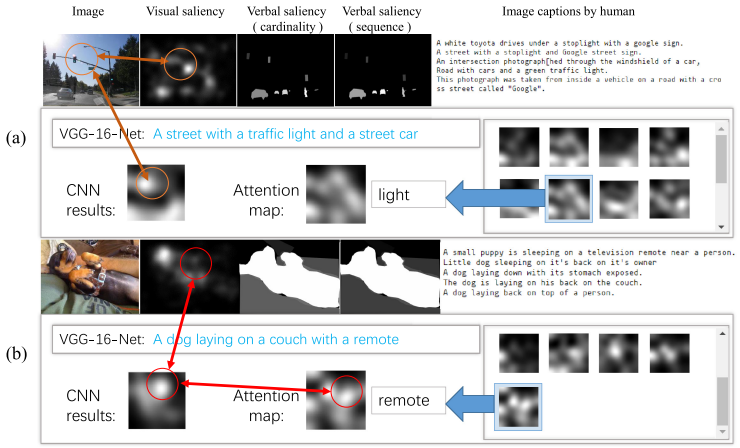

Fig. 14. Example of using CapVis to explore the causes of two failure cases. (a) The captioning model focuses on the traffic light (orange circle) first, as humans do. However, people tend to describe the car first. (b) The captioning model mentions an object that does not exist. The red circle indicates that the model misrecognizes the dog paw.

visual attention and in image captions. In contrast, we see different patterns for inanimate objects. A considerable number of inanimate objects with a certain level of attraction to attention are less likely to be described, such as containers (bottle, bowls) and salient object parts (hand, foot).

7.3 Case Study: Captioning Model Diagnosing

This case study demonstrates how CapVis helps an expert B (EB) diagnose a captioning model. EB is a deep learning researcher who is interested in feature extraction and image understanding. EB tried to build an image captioning model based on CNN and RNN. He faced the problem of finding the reason for failure cases.

Instead of providing a bunch of neuron clusters for network diagnosis, CapVis shows the patterns in which computational models observe and describe (module E). At the same time, they can be compared to patterns from human observation (module B). Although it cannot reveal the problem quantitatively, this method gives a hint about where to improve the model.

"The goal of a captioning model can be (i) to generate a caption that can properly describe the image; (ii) to describe the image as humans do. The second goal is more important if we want to achieve a higher level of artificial intelligence. If the captioning model gives a caption that is quite different from a human's, we cannot stop improving the model even it looks reasonable," EB said. Figure 14 shows two examples that we regard as failure cases and that are frequently met by EB.

In the first example (Figure 14(a)), the model focused on the right region and generated a normal caption that looked appropriate for the image. However, the caption mentioned a traffic light first while people preferred the car. To find the reason, EB turned to module F and found that the CapVis showed a negative visual-verbal consistency score (−0.80) for this image. "The correlation score tells us that the model should not only consider finding and recognizing an object, but also the potential priority of the way people describe it," EB commented. Additionally, he thought that the size, location, and category should be taken into consideration in the early stage of feature extraction. "The traffic light in the middle attracts attention and is well perceived by the model, but in the verbal saliency map we can see the cars are more salient. So I think the consistency score can be a valuable feature before organizing the sentence in captioning model," EB said.

Table 1. Performance (Mean Square Error) of Features at Predicting the Correlation

	Cardinality		Sequence	
Feature type	MSE	STD	MSE	STD
Size	0.2012	0.0540	0.1960	0.0581
Location	0.1963	0.0890	0.1918	0.0896
Density	0.1944	0.0745	0.1902	0.0768
Size + Location + Density	0.1843	0.1102	0.1827	0.1098
Category	0.1756	0.1262	0.1713	0.1273
Super-category	0.1805	0.1066	0.1762	0.1137
Category + super-category	0.1715	0.1366	0.1708	0.1358
All combined	**0.1570**	0.1459	**0.1568**	0.1467
AlexNet [28]	0.1751	0.1393	0.1733	0.1443

In the second example (Figure 14(b)), the model mentioned a "remote" that was not even in the image. "Sometimes we can hardly know what causes the misrecognition when there are many objects to describe," EB said. By looking at the attention map that referred to the word "remote," EB found that the corresponding attention map showed the location of a dog's paw. "This encourages a further training for the recognition part of the captioning model for a better accuracy." The expert also expressed that knowing the causes of failure could save him a lot of time in improving the model in the right direction.

7.4 Predicting the Visual-Verbal Consistency

The impact of the features on visual-verbal consistency motivates us to investigate whether consistency is predictable and how the features contribute to this prediction.

Given an image I, our goal is to automatically decide the consistency between visual saliency and verbal saliency. Based on the preceding observations, we take into consideration size, location, density, and category in our model. The image feature $F(I)$ can be represented by concatenating different feature vectors as follows:

$$F(I) = \{C, C^s, S, L, D\} \in \mathbb{R}^l, \tag{8}$$

where C is an 80-dimensional vector representing object categories; C^s similarly denotes the 12-dimensional vector for super-categories; each element in C and C^s denotes the number of a certain object category in the image; and S, L, and D are 10-, 10-, and 20-dimensional vectors encoding size, distance to center, and density of object, respectively. The preceding features form an l-dimensional ($l = 132$) feature vector for each image. We employ a support vector regressor (SVR, [43]) as our predictive model.

In order to validate the effectiveness of the features used in our experiment, we conduct the same process as described in Section 5 on the validation set of 5,000 images, and we use the Spearman correlation scores as their ground-truth labels. We train SVRs [5] with Gaussian kernels to predict the visual-verbal consistency, using a grid search to select cost and hyper-parameters. We report the results for both standalone and combined features using different sources (cardinality and sequence) as ground truth. We also include a deep convolutional neural network as an additional model for prediction. For this, we employ AlexNet [28] and extract features from the layer just before the final classification layer (often referred to as fc7), resulting in a feature dimension of 4,096. The performance of all models is shown in Table 1.

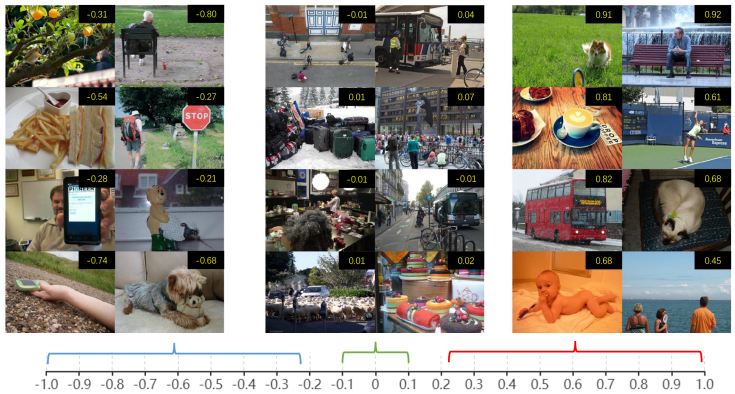

Fig. 15. Example images with low consistency (left), medium consistency (middle), and high consistency (right). Numbers indicate the cardinality-based visual-verbal consistency that our model predicts.

It can be seen that the combined features perform considerably better (MSE = 0.1570 and 0.1568) than the standalone features. The two schemes of verbal saliency generation have shown similar performance, with the difference that the sequence-based method performs slightly better. Moreover, we observe that the object categories, followed by the super-categories, play the most significant roles in the prediction of consistency, whereas low-level features are less useful. This suggests that the verbal description patterns are more related to the semantics of the image.

7.5 Observations and Discussions

What affects visual-verbal consistency? We first look at the image content to unearth possibly consistent patterns that contribute to visual-verbal consistency. Example images predicted by SVR using combined features are shown in Figure 15. The images are ranked by predicted consistency scores in an ascending order.

For the most consistent cases in Figure 15 (right), our model finds relatively clean images with few objects and usual activities. The contents of these images are quite clear at first sight and can be easily understood. The objects in these images are highly likely to be common in daily life and are consistently mentioned in the image captions (e.g., dogs, cars, and people). Also, these objects are quite likely to occur with a dominant size ($0.2 < s < 0.5$) and near the center of the image ($l < 0.2$), thus attracting gaze in the early stage.

Images in Figure 15 (middle) are more complicated scenes containing five or more objects. In these cases, while people still have consistent preferences of visual saliency, the image captions vary in sequence. A typical case can be found in the middle of Figure 7. As the number of objects increases, there are a variety of choices to describe the scene, so the weight of each being mentioned drops. As a result, different things are chosen at the beginning of each sentence.

In most cases, humans and animals are both visually and verbally salient. However, there exist several cases in which inanimate things are more attractive at first sight. Since mostly animate things appear first in captions, the visual saliency in these cases correlates negatively with the verbal saliency. In Figure 15 (left) these rare cases are demonstrated, in which most contain objects more salient than humans and animals. For inconsistent case, more results can be explored using the quick browse function in CapVis (shown in Figure 16), where the overlaid heat maps represent visual saliency. The images in rows 3 and 5 show the tendency of describing people's action first, leading to visual-verbal inconsistency.

Fig. 16. A quick browse of less consistent images between visual saliency and verbal saliency. The examples are arranged as images (1st column), visual saliency maps (2nd column), and verbal saliency for cardinality and sequence maps (3rd and 4th columns, respectively).

Finally, we investigate whether the length of the image captions leads to more variability and hence less visual-verbal consistency. The low correlation between the average length of the captions (measured as the number of words in the sentence) with consistency ($\rho = -0.06$) shows that the length of caption has no obvious impact on visual-verbal consistency. We then check the effect of variation in the length of caption and find that images with high consistency scores ($\rho < -0.4$ or $\rho > 0.4$) usually have less variation in the length of caption.

Overall, visual-verbal consistency is correlated with image content to quite an extent. The proposed features and regressor are able to predict the consistency effectively. Yet there are a number of failure cases that occur because of missing annotations, suggesting further updates to the dataset.

7.6 User Feedback

We provided the whole dataset to two experts who were asked to use CapVis to import a single image from the dataset and see if they could conduct a visual-verbal saliency analysis based on the information and interaction provided by each module. Since they were familiar with all the proposed image features (low- and semantic-level features), we asked several questions and collected feedback after they completely learned how to use each modules in CapVis. The questions involve readability, usability, and limitations of the system. The feedback is summarized as follow:

Visualization Design and Usability: The visual design of our system was received very well by both EA and EB. EA stated that the tool is engaging and easy to understand. He mentioned that showing an image in a statistical way is a novel idea that helps users understand the content of an image. EB was impressed by the visualization view of CapVis, noting that the design "allows me to locate image importance quickly." He also added, "I can easily find all the images with different consistency scores to train a more robust image captioning model in my research."

Limitation and Suggestion: EA stated that we still have to explain how the conclusions of our findings are useful in practice. He said, "After investigation, we do know the difference between how people look and describe, we also come to some conclusions about what affects the visual-verbal saliency. However, the true reason is still not explained nicely." EB suggested that we add details to the deep network module. He said, "For diagnosing the deep model, we still don't know where to refine the model after finding the problem. It is better to find a way to explore the inner

weights of neurons. I would like to see this work extended and shown to be useful for refining deep saliency or captioning models."

8 CONCLUSION

In this work, we developed an analytics system named CapVis to characterize aspects of image recognition and captioning relating to the visual-verbal consistency in how people explore and describe an image. We investigated this question on a large-scale dataset. We extracted the saliency values of objects from eye-tracking data and captions, based on which we proposed several low-level and semantic-level features such as size, location, density, and category and visualized them using a word-based method. Furthermore, using CapVis, we analyzed their relative contributions to visual-verbal consistency both qualitatively and quantitatively. CapVis can also be applied to diagnose deep models for captioning. Finally, we proposed a computational model to predict the consistency between visual saliency and verbal saliency. We envision that understanding the visual-verbal consistency of image saliency could inspire exciting and far-reaching applications in computer vision and answer related questions in human vision.

REFERENCES

[1] Alexander C. Berg, Tamara L. Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. 2012. Understanding and predicting importance in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Providence Rhode Island, 3562–3569.

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

[3] Neil Bruce and John Tsotsos. 2007. Attention based on information maximization. *Journal of Vision* 7, 9 (2007), 950–950.

[4] Neil D. B. Bruce and John K. Tsotsos. 2009. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision* 9, 3 (2009), 5.

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (2011), 27:1–27:27. Software available at https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[6] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 1992–2001.

[7] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, Capri Island, Italy, 74–77.

[8] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. 2011. Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2412–2421.

[9] Robert Desimone and John Duncan. 1995. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18, 1 (1995), 193–222.

[10] Wenwen Dou, Xiaoyu Wang, Remco Chang, and William Ribarsky. 2011. Paralleltopics: A probabilistic approach to exploring document collections. In *IEEE Conference on Visual Analytics Science and Technology (VAST'11)*. IEEE, Providence Rhode Island, 231–240.

[11] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Boston, 1473–1482.

[12] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*. MIT, Lake Tahoe, 2121–2129.

[13] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? Dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*. MIT, Montreal Canada, 2296–2304.

[14] Jonathan Harel, Christof Koch, and Pietro Perona. 2006. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*. MIT, Vancouver, 545–552.

[15] W. Harwin, A. Ginige, and R. Jackson. 1986. A potential application in early education and a possible role for a vision system in a workstation based robotic aid for physically disabled persons. *Interactive Robotic Aids-One Option for Independent Living: An International Perspective, Volume Monograph* 37 (1986), 18–23.

[16] Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. 2002. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (2002), 9–20.

[17] James E. Hoffman and Baskaran Subramaniam. 1995. The role of visual attention in saccadic eye movements. *Attention, Perception, & Psychophysics* 57, 6 (1995), 787–795.

[18] Xiaodi Hou and Liqing Zhang. 2007. Saliency detection: A spectral residual approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Minneapolis, 1–8.

[19] 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*. IEEE, Boston, 262–270.

[20] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259.

[21] Mainak Jas and Devi Parikh. 2015. Image specificity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Boston, 2727–2736.

[22] M. Jiang, S. Huang, J. Duan, and Q. Zhao. 2015. SALICON: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Boston, 1072–1080.

[23] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Las Vegas, 4565–4574.

[24] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, Miami.

[25] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Boston. 3128–3137.

[26] Andrej Karpathy, Armand Joulin, and Feifei Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*. MIT, Montreal, 1889–1897.

[27] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. IEEE, Beijing, 595–603.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. MIT, Lake Tahoe, 1097–1105.

[29] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788.

[30] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. iVisClustering: An interactive visual document clustering via topic modeling. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 1155–1164.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computar Vision*. IEEE, Zurich, 740–755.

[32] Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, and Shixia Liu. 2017. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 91–100.

[33] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. 2015. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Boston, 362–370.

[34] Mateusz Malinowski and Mario Fritz. 2015. Hard to cheat: A Turing test based on answering questions about images. *arXiv Preprint arXiv:1501.03302* (2015).

[35] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv Preprint arXiv:1410.1090* (2014).

[36] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[37] Derrick Parkhurst, Klinton Law, and Ernst Niebur. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Research* 42, 1 (2002), 107–123.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[39] Vasili Ramanishka, Abir Das, Jianming Zhang, and Kate Saenko. 2016. Top-down visual saliency guided by captions. *arXiv Preprint arXiv:1612.07360* (2016).

[40] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 139–147.

[41] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*. MIT, Montreal, 2953–2961.

[42] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Visual explanations from deep networks via gradient-based localization. See https://arxiv.org/abs/1610.02391v3 (2016).

[43] Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14, 3 (2004), 199–222.

[44] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2 (2014), 207–218.

[45] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv Preprint arXiv:1412.6806* (2014).

[46] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37, 2 (2011), 267–307.

[47] Hwang Tae-Hyun, Joo In-Hak, and Cho Seong-Ik. 2006. Detection of traffic lights for vision-based car navigation system. In *Advances in Image and Video Technology*. 682–691.

[48] Charles Thorpe, Martial Hebert, Takeo Kanade, and Steven Shafer. 1989. *Vision and Navigation for the Carnegie Mellon Navlab*. Springer.

[49] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*. 63–70.

[50] F.-Y. Tzeng and K.-L. Ma. 2005. Opening the black box-data driven visualization of neural networks. In *Visualization, 2005. VIS 05. IEEE*. IEEE, Baltimore, 383–390.

[51] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. Tiara: A visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 153–162.

[52] Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*. 133–138.

[53] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. 2014. Predicting human gaze beyond pixels. *Journal of Vision* 14, 1 (2014), 1–20.

[54] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*. IEEE, Lille, 2048–2057.

[55] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv Preprint arXiv:1506.06579* (2015).

[56] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[57] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. 2008. SUN: A bayesian framework for saliency using natural statistics. *Journal of Vision* 8, 7 (2008), 32.

Received August 2017; revised March 2018; accepted March 2018