VISUAL SALIENCY FOR IMAGE CAPTIONING IN NEW MULTIMEDIA SERVICES

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, Rita Cucchiara

Dipartimento di Ingegneria "Enzo Ferrari"
Università degli Studi di Modena e Reggio Emilia
name.surname@unimore.it

ABSTRACT

Image and video captioning are important tasks in visual data analytics, as they concern the capability of describing visual content in natural language. They are the pillars of query answering systems, improve indexing and search and allow a natural form of human-machine interaction. Even though promising deep learning strategies are becoming popular, the heterogeneity of large image archives makes this task still far from being solved. In this paper we explore how visual saliency prediction can support image captioning. Recently, some forms of unsupervised machine attention mechanisms have been spreading, but the role of human attention prediction has never been examined extensively for captioning. We propose a machine attention model driven by saliency prediction to provide captions in images, which can be exploited for many services on cloud and on multimedia data. Experimental evaluations are conducted on the SALICON dataset, which provides groundtruths for both saliency and captioning, and on the large Microsoft COCO dataset, the most widely used for image captioning.

Index Terms— Image Captioning, Visual Saliency, Human Eye Fixations, Attentive Mechanisms, Deep Learning.

1. INTRODUCTION

Replicating the human ability of describing an image in natural language, providing a rich set of details at a first glance, has been one of the primary goals of different research communities in the last years. Captioning models, indeed, should not only be able to solve the challenge of identifying each and every object in the scene, but they should also be capable of expressing their names and relationships in natural language. The enormous variety of visual data makes this task particularly challenging. It is very hard, indeed, to predict a-priori and only driven by data what could be interesting in an image and what should be described. Nevertheless, describing visual data in natural language opens the door to many future applications: the one with the largest potential impact is that of defining new services for search and retrieval in visual data archives, using query-answering tools, working on natural language as well as improving the performance of more traditional keyword-based search engines.

Fig. 1. Saliency prediction and captions generated by our approach on images from the Microsoft COCO Dataset [1].

With the advance of deep neural networks [2] and large annotated datasets [1], recent works have significantly improved the quality of caption generation, bringing the field to a rather mature stage, in which proper captions can be automatically generated for a wide variety of natural images. Most of the existing approaches rely on a combination of Convolutional Neural Networks (CNN), to extract a vectorized representation of an input image, and Recurrent Neural Networks (RNN), as a language model and to generate the corresponding caption [3]. As such, they treat the input image as a whole, neglecting the human tendency to focus on specific parts of the scene when watching an image [4], which is is instead crucial for a convincing human-like description of the scene.

An attempt to emulate such ability in captioning models has been carried out by the machine attention literature [5]: machine attention mechanisms, indeed, focus on different regions of the input image during the generation of the caption, in a fully unsupervised manner, so that regions of focus are chosen only with the objective of generating a better description, without considering the actual human attentive mecha-

978-1-5386-0560-8/17/$31.00 ©2017 IEEE

nisms.

On a different note, the computer vision community has also studied the development of approaches capable of predicting human eye fixations on images [6, 7, 8], by relying on datasets taken with eye-tracking devices. This task, namely saliency prediction, aims at replicating the human selective mechanisms which drive the gaze towards some specific regions of the scene, and has never been incorporated in a captioning architecture, even though, in principle, such supervision could result in better image captioning performance.

In this paper, we present a preliminary investigation on the role of saliency prediction in image captioning architectures. We propose an architecture in which the classical machine attention paradigm is extended in order to take into account salient region as well as the context of the image. Referring to this as a "saliency-guided attention", we perform experiments on the SALICON dataset [9] and on Microsoft COCO [1]. Fig. 1 shows examples of image captions generated by our method on the COCO Dataset [1], along with the corresponding visual saliency predictions. As it can be seen, visual saliency can give valuable information on the objects which should be named in the caption.

In the rest of the paper, after reviewing some of the most relevant related works, we will present our machine attention approach, which integrates saliency prediction. Finally, an experimental evaluation and a use case will follow.

2. RELATED WORK

In this section we briefly review related works in image captioning and visual saliency prediction, and also describe recent studies that incorporate human gaze in image captioning architectures.

2.1. Image and video captioning

Early captioning methods were based on the identification of semantic triplets (with subject, object and verb) using visual classifiers, and captions were generated through a language model which fitted predicted triplets to predefined sentence templates. Of course, this kind of sentences could not satisfy the richness of natural language: for these reasons, research on image and video captioning has soon moved to the use of recurrent networks, which, given a vectored description of a visual content, could naturally deal with sequences of words [3, 10, 11].

Karpathy *et al.* [10] used a ranking loss to align image regions with sentence fragments, while Vinyals *et al.* [3] developed a generative model in which the caption is generated by a LSTM layer, trained to maximize the likelihood of the target description given the input image. Johnson *et al.* [12] addressed the task of dense captioning, which detects and describes dense regions of interest.

Xu *et al.* [5] developed an approach to image captioning which incorporates a form of machine attention in two variants (namely, "soft" and "hard" attention), by which a generative LSTM can focus on different regions of the image while generating the corresponding caption.

2.2. Visual saliency prediction

Inspired by biological studies, traditional saliency prediction methods have defined hand-crafted features that capture low-level cues such as color, contrast and texture and semantic concepts such as faces, people and text [13, 14, 15, 16]. However, these techniques were not able to effectively capture the large variety of factors that contribute to define visual saliency maps. With the advent of deep neural networks, saliency prediction has achieved strong improvements both thanks to specific architectures [6, 7, 8, 17, 18] and to large annotated datasets [9]. In fact, recent deep saliency models have reached significant performances, approaching to those of humans.

Huang *et al.* [7] proposed an architecture that integrates saliency prediction into deep convolutional networks trained with a saliency evaluation metric as loss function. Jetley *et al.* [8] introduced a saliency map model that formulates a map as a generalized Bernoulli distribution and they used these maps to train a deep network trying different loss functions. Kruthiventi *et al.* [19] instead presented an unified framework that is capable of predicting eye fixations and segmenting salient objects on input images. Recently, Cornia *et al.* [18] proposed an attentive mechanism incorporated in a deep saliency architecture to iteratively refine the predicted saliency map and significantly improve prediction results.

2.3. Captioning and saliency

Recent studies have started to investigate the use of visual saliency to automatically describe an input image in natural language. In particular, Sugano *et al.* [20] proposed a machine attentive model that exploits gaze-annotated images: their architecture employs human fixation points to predict image captions for the SALICON dataset [9]. Since this is a subset of the Microsoft COCO dataset [1], it is the only dataset providing both gaze and saliency annotations.

The main drawback of their approach is the need of big amounts of images annotated with human captions and human fixation points. Fixation points, moreover, are needed also in the test phase, thus making this proposal unusable in practice. For this reason, we investigate the use of saliency maps predicted by a state of the art saliency model [18] to improve image captioning performance. Our approach can be potentially trained using any image captioning dataset, and can predict captions on any image.

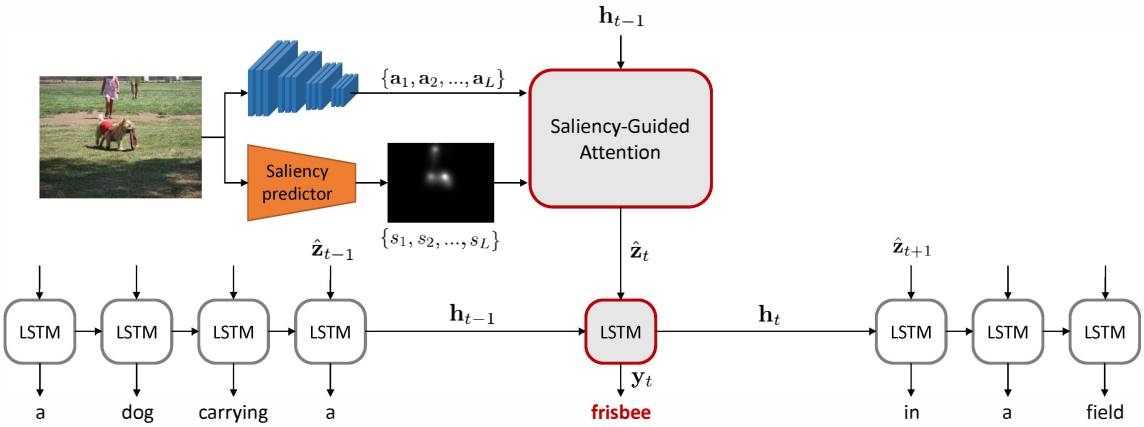

Fig. 2. Overview of our image captioning model. A Saliency-Guided machine attention mechanism drives the generation of the next word in the caption, by taking into account both salient and non-salient regions.

3. SALIENCY-GUIDED CAPTIONING

Machine attention mechanisms [5] are a popular way of obtaining time-varying inputs for recurrent architectures. In image captioning, it is well-known that performances can be improved by providing the generative LSTM with the specific region of the image it needs to generate a word: at each timestep the attention mechanism selects a region of the image, based on the previous LSTM state, and feeds it to the LSTM, so that the generation of a word is conditioned on that specific region, instead of being driven by the entire image.

The most popular attentive mechanism is the so-called "soft-attention" [5]. The input image is encoded as a grid of feature vectors $\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_L\}$, each corresponding to a spatial location of the image. These are usually obtained from the activations of a convolutional or pooling layer of a CNN. At each timestep, the soft-attention mechanism computes a context feature vector $\hat{\mathbf{z}}_t$ representing a specific part of the input image, by combining feature vectors $\{\mathbf{a}_i\}_i$ with weights obtained from a *softmax* operator. Formally, the context vector $\hat{\mathbf{z}}_t$ is obtained as

$$\hat{\mathbf{z}}_t = \sum_{i=1}^{L} \alpha_{ti} \mathbf{a}_i, \tag{1}$$

where α_{ti} are weights representing the current state of the machine attention. These are driven by the original image feature vectors and by the previous hidden state $\mathbf{h}_{t-1}$ of the LSTM:

$$e_{ti} = v_e^T \cdot \phi(W_{ae} \cdot \mathbf{a}_i + W_{he} \cdot \mathbf{h}_{t-1})) \tag{2}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}, \tag{3}$$

where ϕ is the hyperbolic tangent `tanh`, W_{ae}, W_{he} are learned matrix weights and v_e^T is a learned row vector.

To investigate the role of visual saliency in the context of attentive captioning models, we extend this schema by splitting the machine attention into saliency and non saliency regions, and learning different weights for both of them. Given a visual saliency predictor [18] which predicts a saliency map $\{s_1, s_2, ..., s_L\}$, having the same resolution of the feature vector grid $\{\mathbf{a}_i\}_i$, and with $s_i \in [0, 1]$, we propose to modify Eq. 2 as follows:

$$e_{ti}^{sal} = v_{e,sal}^T \cdot \phi(W_{ae} \cdot \mathbf{a}_i + W_{he} \cdot \mathbf{h}_{t-1})) \tag{4}$$

$$e_{ti}^{nosal} = v_{e,nosal}^T \cdot \phi(W_{ae} \cdot \mathbf{a}_i + W_{he} \cdot \mathbf{h}_{t-1})) \tag{5}$$

$$e_{ti} = s_i \cdot e_{ti}^{sal} + (1 - s_i) \cdot e_{ti}^{nosal}. \tag{6}$$

Notice that our model learns different weights for saliency and non-saliency regions ($v_{e,sal}^T$ and $v_{e,nosal}^T$ respectively), and combines them into a final attentive map in which the contributions of salient and non-salient regions are merged together. Similarly to the classical soft-attention approach, the proposed generative LSTM can focus on every region of the image, but the focus on salient region is driven by the output of the saliency predictor.

3.1. Sentence generation

Given an image and sentence $(\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_T)$, encoded with one-hot vectors (1-of-N encoding, where N is the size of the vocabulary), we build a generative LSTM decoder. This is conditioned step by step on the first t words of the caption and on the corresponding context vector, and is trained to produce the next word of the caption. The objective function which we optimize is the log-likelihood of correct words over the sequence

$$\max_{\mathbf{w}} \sum_{t=1}^{T} \log \Pr(\mathbf{y}_t | \hat{\mathbf{z}}_t, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_0) \tag{7}$$

where $\mathbf{w}$ are all the parameters of the model. The probability of a word is modeled via a softmax layer applied on the

output of the decoder. To reduce the dimensionality of the decoder, a linear embedding transformation is used to project one-hot word vectors into the input space of the decoder and, viceversa, to project the output of the decoder to the dictionary space.

$$\Pr(\mathbf{y}_t|\hat{\mathbf{z}}_t, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_0) \propto \exp(\mathbf{y}_t^T W_p \mathbf{p}_t) \quad (8)$$

where W_p is a matrix for transforming the decoder output space to the word space and $\mathbf{h}_t$ is the output of the decoder, computed with a LSTM layer. In particular, we use a LSTM implemented by the following equations

$$\mathbf{i}_t = \sigma(W_{ix}\hat{\mathbf{z}}_t + W_{ih}\mathbf{h}_{t-1} + \mathbf{b}_i) \quad (9)$$
$$\mathbf{f}_t = \sigma(W_{fx}\hat{\mathbf{z}}_t + W_{fh}\mathbf{h}_{t-1} + \mathbf{b}_f) \quad (10)$$
$$\mathbf{g}_t = \phi(W_{gx}\hat{\mathbf{z}}_t + W_{gh}\mathbf{h}_{t-1} + \mathbf{b}_g) \quad (11)$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \quad (12)$$
$$\mathbf{o}_t = \phi(W_{fx}\hat{\mathbf{z}}_t + W_{fh}\mathbf{h}_{t-1} + \mathbf{b}_f) \quad (13)$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \phi(\mathbf{c}_t) \quad (14)$$

where $\odot$ denotes the element-wise Hadamard product, σ is the sigmoid function, ϕ is the hyperbolic tangent `tanh`, W_* are learned weight matrices and $\mathbf{b}_*$ are learned biases vectors. The internal state $\mathbf{h}$ and memory cell $\mathbf{c}$ are initialized to zero.

4. EXPERIMENTAL EVALUATION

4.1. Datasets and metrics

We evaluate the contribution of saliency maps in our image captioning network on two different datasets: SALICON [9] and Microsoft COCO [1].

The Microsoft COCO dataset is composed by more than 120,000 images divided in training and validation sets, where each of them is annotated with five sentences using Amazon Mechanical Turk.

The SALICON dataset is a subset of COCO in which images are provided with their saliency maps. Gaze annotations are collected with a mouse-contingent paradigm which results to be very similar to an eye-tracking system, as demonstrated in [9]. This dataset contains 10,000 training images, 5,000 validation images and 5,000 testing images, all having a size of 480×640.

We employ four popular metrics for evaluation: BLEU [21], ROUGE$_L$ [22], METEOR [23] and CIDEr [24]. BLEU is a modified form of precision between n-grams to compare a candidate translation against multiple reference translations. We evaluate our predictions with BLEU using mono-grams, bi-grams, three-grams and four-grams. ROUGE$_L$ computes an F-measure considering the longest co-occurring in sequence n-grams. METEOR, instead, is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also has several features that are not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching. CIDEr, finally, computes the average cosine similarity between n-grams found in the generated caption and those found in reference sentences, weighting them using TF-IDF. To ensure a fair evaluation, we use the Microsoft COCO evaluation toolkit[1] to compute all scores.

4.2. Implementation details

As mentioned, the input image is encoded as a grid of feature vectors coming from a CNN. In our experiments on the SALICON dataset, we extract image features from the last convolutional layer of two different CNNs: the VGG-16 [25] and the ResNet-50 [26]. On the Microsoft COCO, instead, we train our network using only image features coming from the ResNet-50. Since all images from the SALICON dataset have all the same size of 480×640, we set the image size for this dataset to 480×640 thus obtaining $L = 15 \times 20 = 300$. For the COCO dataset, we set the image size to 480×480 obtaining $L = 15 \times 15 = 225$.

Saliency maps predicted with [18] have the same size of the input images. For this reason, we resize saliency maps to a size of 15×20 for training on the SALICON dataset and to a size of 15×15 for training on the Microsoft COCO dataset.

All other implementation details are kept the same as in Xu *et al.* [5]. In all our experiments, we train our network with the Nestorov Adam optimizer [27].

4.3. Results

Table 1 compares the performances of our approach against the unsupervised machine attention approach in [5], using all the metrics described in Section 4.1. In this case, training is performed on the SALICON training set, and evaluation is carried out on the SALICON validation set. We employ, as the base CNN, the recent ResNet-50 model [26], as well as the more widely used VGG-16 [25].

As it can be seen, our attention model, which incorporates visual saliency, is able to achieve better results on all metrics, except from ROUGE$_L$ in the VGG-16 setting, in which we achieve exactly the same result. For reference, we also report the performance of the architecture when using groundtruth saliency maps, instead of those predicted by [18]: as it can be seen, even though using groundtruth maps provides slightly better results, a proper saliency prediction model can be used without significant loss of performance.

We also perform the same test on the COCO dataset. Being the saliency predictor of [18] trained on SALICON, the experiment is useful to assess the generalization capabilities of the complete model. Results are reported in Table 2: as it can be seen, also in this case, our model can surpass the performance of the soft-attention proposal of [5].

[1] `https://github.com/tylin/coco-caption`

Table 1. Image captioning results on SALICON validation set [9] in terms of BLEU@1-4, METEOR, ROUGE$_L$ and CIDEr. The results are reported using two different CNNs to extract features from input images: the VGG-16 and the ResNet-50.

	CNN	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE$_L$	CIDEr
Soft Attention [5]	VGG-16	0.680	0.501	0.358	0.256	0.222	**0.497**	0.691
Saliency-Guided Attention	VGG-16	**0.682**	**0.505**	**0.361**	**0.258**	**0.223**	**0.497**	**0.694**
Saliency-Guided Att. (with GT saliency maps)	VGG-16	*0.684*	*0.503*	*0.360*	*0.257*	*0.224*	*0.501*	*0.696*
Soft Attention [5]	ResNet-50	0.700	0.523	0.379	0.274	0.235	0.510	0.771
Saliency-Guided Attention	ResNet-50	**0.709**	**0.534**	**0.388**	**0.280**	**0.233**	**0.513**	**0.774**
Saliency-Guided Att. (with GT saliency maps)	ResNet-50	*0.702*	*0.527*	*0.383*	*0.277*	*0.236*	*0.513*	*0.779*

Table 2. Image captioning results on Microsoft COCO validation set [1] in terms of BLEU@1-4, METEOR, ROUGE$_L$ and CIDEr.

	CNN	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE$_L$	CIDEr
Soft Attention [5]	ResNet-50	0.717	0.546	0.402	0.294	0.253	0.529	0.939
Saliency-Guided Attention	ResNet-50	**0.718**	**0.547**	**0.404**	**0.296**	**0.254**	**0.530**	**0.944**

Ours: A man and a woman are playing frisbee on a field.
Soft Attention [5]: A man standing next to a man holding a frisbee.
GT: Two people in Swarthmore College sweatshirts are playing frisbee.

Ours: A group of people sitting on a boat in a lake.
Soft Attention [5]: A group of people sitting on top of a boat.
GT: Family of five people in a green canoe on a lake.

Ours: A large jetliner sitting on top of an airport runway.
Soft Attention [5]: A large air plane on a runway.
GT: A large passenger jet sitting on top of an airport runway.

Fig. 3. Example results on the Microsoft COCO dataset [1].

4.4. A use case in the cloud: NeuralStory

We conclude by presenting an interesting use-case of the proposed architecture. This work is, indeed, part of a large project called *NeuralStory*, which aims at providing new services for annotation, retrieval and re-use of video material in education. The goal of the project is to re-organize video material by extracting its storytelling structure and presenting it with new forms of summarization for quick browsing. Videos are divided into shots and scenes with a deep learning-based approach [28], using images, audio and semantic concepts extracted with a suitable CNN. The resulting annotation is also provided with text, extracted with speech-to-text tools, concepts and possibly user-generated annotations.

The system behind the project works on the cloud and is powered by the eXo Platform ECMS[2]. Videos can be provided by private users or content owners, and the analysis process is carried out automatically on the cloud. A web interface allows students, teachers and any user to browse and create multimodal slides (called *MeSlides*) for re-using visual and textual data enriched with automatic annotations.

Fig. 4 shows some captions automatically generated by our architecture on images taken from an art documentary which is part of NeuralStory. As it can be seen, even though the model has been trained on a different domain, it is still able to generalize and provide appropriate captions. With this work we intend to enrich the annotation and key-frame description on the web interface. Automatically generated captions will be useful for human search, for automatic search by query, and possibly for future query-answering services.

5. CONCLUSION

In this paper, we investigated the role of visual saliency for image captioning. A novel machine attention architecture, which seamlessy incorporates visual saliency prediction, has been proposed and experimentally validated. Finally, a case study involving a video platform has been presented.

[2] `https://www.exoplatform.com`

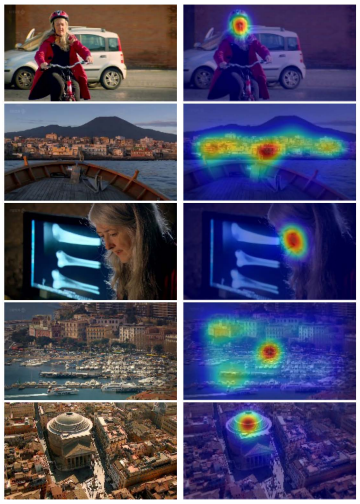

Fig. 4. Saliency maps and captions generated on sample images taken from the *Meet the Romans with Mary Beard* TV series.

6. REFERENCES

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *ANIPS*, 2012.

[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[4] Ronald A. Rensink, "The Dynamic Representation of Scenes," *Visual Cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[5] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[6] Matthias Kümmerer, Lucas Theis, and Matthias Bethge, "DeepGaze I: Boosting saliency prediction with feature maps trained on ImageNet," in *ICLR Workshop*, 2015.

[7] 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.

[8] Saumya Jetley, Naila Murray, and Eleonora Vig, "End-to-End Saliency Mapping via Probability Distribution Prediction," in *CVPR*, 2016.

[9] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao, "SALICON: Saliency in context," in *CVPR*, 2015.

[10] Andrej Karpathy and Li Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *CVPR*, 2015.

[11] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara, "Hierarchical Boundary-Aware Neural Encoder for Video Captioning," in *CVPR*, 2017.

[12] Justin Johnson, Andrej Karpathy, and Li Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *CVPR*, 2016, pp. 4565–4574.

[13] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *ANIPS*, 2006.

[14] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal, "Context-aware saliency detection," *IEEE TPAMI*, vol. 34, no. 10, pp. 1915–1926, 2012.

[15] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, "Learning to predict where humans look," in *ICCV*, 2009.

[16] Jianming Zhang and Stan Sclaroff, "Saliency detection: A boolean map approach," in *ICCV*, 2013.

[17] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "A Deep Multi-Level Network for Saliency Prediction," in *ICPR*, 2016.

[18] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model," *arXiv preprint arXiv:1611.09571*, 2017.

[19] Srinivas SS Kruthiventi, Vennela Gudisa, Jaley H Dholakiya, and R Venkatesh Babu, "Saliency Unified: A Deep Architecture for Simultaneous Eye Fixation Prediction and Salient Object Segmentation," in *CVPR*, 2016.

[20] Yusuke Sugano and Andreas Bulling, "Seeing with humans: Gaze-assisted neural image captioning," *arXiv preprint arXiv:1608.05203*, 2016.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *40th annual meeting on association for computational linguistics*, 2002.

[22] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, 2004.

[23] Satanjeev Banerjee and Alon Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

[24] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.

[25] K Simonyan and A Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[27] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara, "Recognizing and presenting the storytelling video structure with deep multimodal networks," *IEEE TMM*, 2017.