Seeing Bot*

Yingwei Pan
University of Science and Technology of China
Hefei, China
panyw.ustc@gmail.com

Zhaofan Qiu
University of Science and Technology of China
Hefei, China
zhaofanqiu@gmail.com

Ting Yao
Microsoft Research Asia
Beijing, China
tiyao@microsoft.com

Houqiang Li
University of Science and Technology of China
Hefei, China
lihq@ustc.edu.cn

Tao Mei
Microsoft Research Asia
Beijing, China
tmei@microsoft.com

ABSTRACT

We demonstrate a video captioning bot, named Seeing Bot, which can generate a natural language description about what it is seeing in near real time. Specifically, given a live streaming video, Seeing Bot runs two pre-learned and complementary captioning modules in parallel—one for generating image-level caption for each sampled frame, and the other for generating video-level caption for each sampled video clip. In particular, both the image and video captioning modules are boosted by incorporating semantic attributes which can enrich the generated descriptions, leading to human-level caption generation. A visual-semantic embedding model is then exploited to rank and select the final caption from the two parallel modules by considering the semantic relevance between video content and the generated captions. The Seeing Bot finally converts the generated description to speech and sends the speech to an end user via an earphone. Our demonstration is conducted on any videos in the wild and supports live video captioning.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**; **Similarity measures**;

KEYWORDS

Video Captioning; Image Captioning; Multi-view Embedding; Deep Convolutional Neural Networks; Chitchat Bot

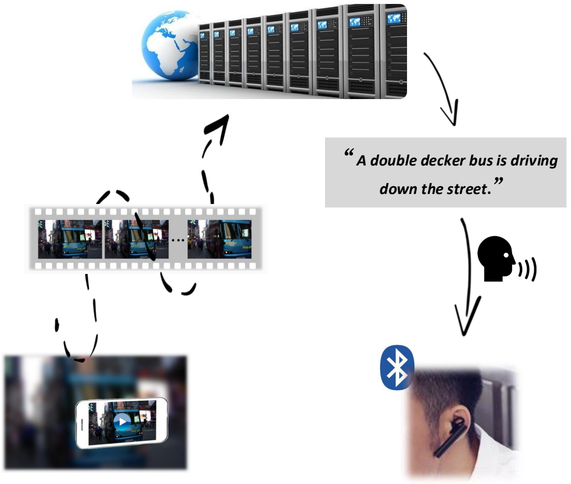

Figure 1: The workflow of Seeing Bot.

1 INTRODUCTION

Recognition of videos has been a fundamental challenge of both computer vision and multimedia communities for decades. Previous research has predominantly focused on recognizing videos with a predefined yet very limited set of individual words. Recently, researchers have strived to automatically describe video content with a complete and natural sentence, which is called video captioning. There is a wide variety of applications based on the generated description, ranging from editing, indexing, search, to practical tools that help blind and visually impaired people.

Although video captioning has been an emerging topic, it is a challenging issue to create a real bot which can convert a live streaming video to a natural language description in real time. We present in this demo such a kind of video captioning bot, named Seeing Bot. Figure 1 shows the workflow of Seeing Bot. Specifically, users are able to easily capture live videos by using mobile devices or wearable camcorders. The recorded video is delivered to the Seeing Bot client, where our video captioning system generates a natural sentence describing the video content. Then, the sentence is converted into speech

*This work was performed when Yingwei Pan and Zhaofan Qiu were visiting Microsoft Research as research interns.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5022-8/17/08...$15.00
https://doi.org/10.1145/3077136.3084144

and sent to an end user via an earphone. Our Seeting Bot provides the capability of real time understanding of real scene and triggering human-bot conversation.

Our Seeing Bot has the following distinct characteristics. First, it runs two parallel image and video captioning modules in real time, which can complement the generated captions to each other. More importantly, both image and video captioning modules are attribute-augmented architectures by integrating semantic attributes into captioning framework for enhancing caption generation. Second, it is equipped with a visual-semantic embedding model to automatically select the best caption from these two modules.

2 TECHNOLOGY

Figure 2 shows an overview of our caption generation framework in Seeing Bot, which is mainly composed of two components: 1) Image captioning model (Img2Cap) and video captioning model (Video2Cap), which are capable of generating natural sentence candidates of sampled frames and video clip. 2) visual-semantic embedding for selecting the most relevant sentence according to current input video stream. Finally, the output sentence is converted into a speech, which is further sent to an end user via an earphone.

In the following, we begin the Section by presenting the motivation and implementation of our image/video captioning modules and visual-semantic embedding model, followed by the introduction of the interface.

2.1 Image and video captioning modules

Given the streaming video from a live camera, our goal is to generate several natural sentence candidates for sampled frames or video clip. Recently, researchers have strived to this target—automatically describing the content of an image or video with a complete and natural sentence, which has a great potential impact for instance on robotic vision or helping visually impaired people. Most of recent attempts on image captioning [3, 13] and video captioning [8, 12] follow the elegant recipe of Convolutional Neural Networks (CNN) plus Recurrent Neural Networks (RNN) architecture, which is to translate directly from image/video representation to language. While encouraging performances are reported in the corresponding in-domain benchmarks, these CNN plus RNN models do not explicitly take more high-level and detailed semantic information from images/videos into account, resulting in limited ability of recognizing rich semantic cues when applied in the wild. Hence, to enrich the generated image/video descriptions in Seeing Bot with more semantic cues, we employ our state-of-the-art semantic attributes based image captioning [15] and video captioning [9] models to generate frame-level caption of each sampled frame and video-level caption of each video clip, respectively.

Semantic attributes based image captioning. Suppose we have an image I to be described by a textual sentence $\mathcal{S}$, where $\mathcal{S} = \{w_1, w_2, ..., w_{N_s}\}$ consisting of N_s words. Let $\mathbf{I}$ and $\mathbf{w}_t$ denote the image representation of the image I and the textual feature of the t-th word w_t in sentence $\mathcal{S}$, respectively. Furthermore, we have feature vector $\mathbf{A_i}$ to represent the probability distribution over the high-level attributes for image I. Specifically, we train the attribute detectors by using the weakly-supervised approach of Multiple Instance Learning (image MIL model) in [4] over image captioning benchmark and treat the final image-level response probabilities of all the attributes as $\mathbf{A_i}$.

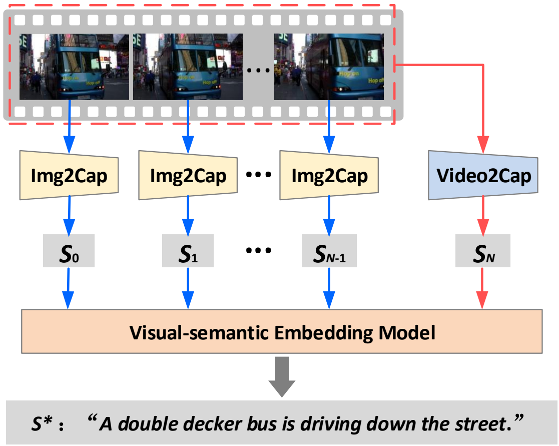
Img2Cap
Img2Cap
Img2Cap
Video2Cap
S_0
S_1
S_{N-1}
S_N
Visual-semantic Embedding Model
S* : "A double decker bus is driving down the street."

Figure 2: An overview of our caption generation framework in Seeing Bot.

Inspired by the recent successes of probabilistic sequence models leveraged in statistical machine translation [11], we formulate this semantic attributes based image captioning model in an end-to-end fashion based on RNN which encodes the given image and its detected attributes into a fixed dimensional vector, and then decodes the vector to the target output sentence. Hence, the sentence generation problem is formulated by minimizing the following loss as

$$E(\mathbf{I}, \mathbf{A_i}, \mathcal{S}) = -\log \Pr(\mathcal{S}|\mathbf{I}, \mathbf{A_i}), \qquad (1)$$

which is the negative log probability of the correct textual sentence given the image representation and detected attributes. Since the CNN plus RNN model produces one word in the sentence at each time step, it is natural to apply chain rule to model the joint probability over the sequential words. Thus, the log probability of the sentence is given by the sum of the log probabilities over the words:

$$\log \Pr(\mathcal{S}|\mathbf{I}, \mathbf{A_i}) = \sum_{t=1}^{N_s} \log \Pr(\mathbf{w}_t|\, \mathbf{I}, \mathbf{A_i}, \mathbf{w}_0, \ldots, \mathbf{w}_{t-1}). \qquad (2)$$

By minimizing this loss, the contextual relationship among the words in the sentence can be guaranteed given the image and its detected semantic attributes. We formulate this task as a variable-length sequence-to-sequence problem and model the parametric distribution $\Pr(\mathbf{w}_t|\, \mathbf{I}, \mathbf{A_i}, \mathbf{w}_0, \ldots, \mathbf{w}_{t-1})$ in Eq.(2) with Long Short-Term Memory (LSTM) network, which is a widely used type of RNN and can capture long-term information in the sequential data by mapping sequences to sequences.

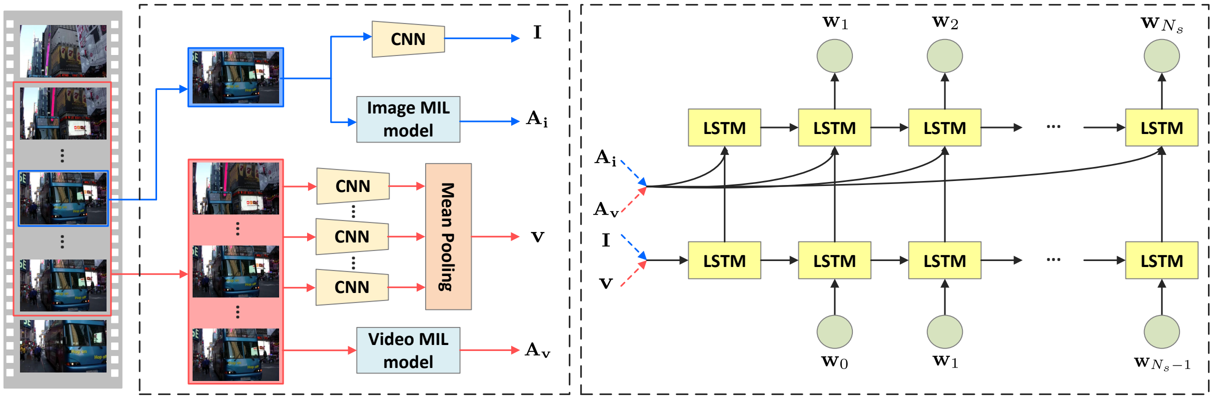

Figure 3: The image and video captioning modules in Seeing Bot.

The basic idea of our utilized semantic attributes based image captioning model is to translate the image representation extracted from CNN to the desired output sentence through LSTM model by additionally injecting the detected high-level semantic attributes, as depicted in Figure 3. In particular, the image captioning model firstly encodes image representation $\mathbf{I}$ at the initial time step to inform the LSTM about the image content and then feeds attributes representation as the additional inputs to the second-layer LSTM unit at each time step to emphasize the semantic information more frequently. Hence, the LSTM decodes each output word based on previous word, previous step's hidden state of LSTM, and the additional input attributes representation. Please note that for the input sentence $\mathbf{W} \equiv [\mathbf{w}_0, \ldots, \mathbf{w}_{N_s}]$, we take $\mathbf{w}_0$ as the start sign word to inform the beginning of sentence and $\mathbf{w}_{N_s}$ as the end sign word which indicates the end of sentence. Both of the special sign words are included in our vocabulary.

After training the whole image captioning model on image captioning benchmark—MSCOCO [7], we uniformly sample N frames from the input video clip and directly apply the learnt image captioning model to generate the corresponding N natural sentence candidates $\{S_0, S_1, ..., S_{N-1}\}$. Specifically, for each sampled frame, we directly choose the word with maximum probability at each time step and set its word representation as LSTM input for next time step until the end sign word is outputted.

Semantic attributes based video captioning. In addition to the image captioning model which is able to recognize scenes and objects within the sampled frames, we also utilize the semantic attributes based video captioning model to capture the temporal dynamics within the input video clip. In particular, given a video V with N sampled frames to be described by a textual sentence $\mathcal{S}$, we first exploit CNN to produce the representation of each sampled frame and then perform "mean pooling" process over all the sampled frames to generate the video representation $\mathbf{v}$. Furthermore, we train the attributes detectors by utilizing the video Multiple Instance Learning (video MIL model) in [9] over video captioning benchmark and achieve the final video-level response probabilities of all the semantic attributes as $\mathbf{A_v}$. The video captioning problem is then formulated by minimizing the negative log probability of the correct textual sentence given the video representation and detected attributes, which is defined as

$$E(\mathbf{v}, \mathbf{A_v}, \mathcal{S}) = -\log \Pr(\mathcal{S}|\mathbf{v}, \mathbf{A_v}), \quad (3)$$

where the log probability of the sentence is given by the sum of the log probabilities over the words:

$$\log \Pr(\mathcal{S}|\mathbf{v}, \mathbf{A_v}) = \sum_{t=1}^{N_s} \log \Pr(\mathbf{w}_t|\, \mathbf{v}, \mathbf{A_v}, \mathbf{w}_0, \ldots, \mathbf{w}_{t-1}). \quad (4)$$

LSTM is again employed to model the corresponding parametric distribution $\Pr(\mathbf{w}_t|\, \mathbf{v}, \mathbf{A_v}, \mathbf{w}_0, \ldots, \mathbf{w}_{t-1})$ in Eq.(4).

The architecture of our utilized semantic attributes based video captioning model is similar to our image captioning model except that the input visual representation is generated through mean pooling over all the sampled frame representations and the the semantic attributes are learnt through video MIL model which is tailored in video domain, as shown in Figure 3. After training the whole video captioning model on video captioning benchmark—Microsoft Research Video Description Corpus (YouTube2Text) [2], we directly leverage the learnt video captioning model to generate the natural sentence S_N for the input video clip.

In sum, with the two complementary image and video captioning models, our Seeing Bot can generate $N+1$ candidate sentences for each input video clip by not only recognizing the visual appearances on the frame level but also exploring temporal dynamics on the video clip level.

2.2 Visual-semantic embedding

Based on the sentence candidates generated by image and video captioning models, we rank the $N+1$ sentence candidates $\{S_0, S_1, ..., S_N\}$ and select the best sentence S^* with the highest relevance score measured by a visual-semantic embedding model. Inspired by [5, 6, 10], we construct a visual-semantic embedding space between textual sentences and their corresponding visual content for sentence ranking. Specifically, to measure the relevance between a video clip and its sentence candidates generated from different sources (sampled frames or video clip), we train the deep visual-semantic embedding model [5] based on the combination of MSCOCO and YouTube2Text datasets.

2.3 Interface

The ultimate target of our Seeing Bot is seeing, captioning, and telling the world to user. To prove this functionality, once fetching the output sentence generated by our Seeing Bot, we not only display the sentence on Seeing Bot client, but also convert the output sentence into a speech and tell the detailed description about the captured video to an end user via an earphone.

3 SYSTEM AND DEMONSTRATION

Demonstration. In the demonstration of Seeing Bot, a camera is carried by a user and keeps capturing the scene. Meanwhile, a Bot client is set up to simultaneously receive the collected video clip from the camera and run the key sentence generation. Once the sentence is generated, Seeing Bot will automatically tell the result sentence to user through the earphone.

Performance. We conduct both objective and subjective evaluations of our Seeing Bot. Regarding the objective evaluations of our image captioning and video captioning models in Seeing Bot, we both achieve the state-of-the-art performances in the sentence generation task: 25.6% and 32.6% in METEOR [1] on MSCOCO c5 testing set and YouTube2Text testing set, respectively. To evaluate the quality of our final generated caption by Seeing Bot subjectively, we collected 1,000 real-life videos recorded by camera from YouTube and invited 10 volunteers to annotate the generated sentences by our Seeing Bot on a three point ordinal scale: 2-Excellent; 1-Good; 0-Bad according to relevance, user friendliness and user experience. The higher score indicates the higher satisfaction. The rate of satisfying result (more than 0 point) for our Seeing Bot is 73.5% and the average score is 1.15.

The video captioning system of Seeing Bot is currently running on a PC with 2.60GHz CPU and 16GB main memory. Given each input one-second video clip, the sentence generation takes about one second in total, which means that our Seeing Bot can caption and tell the description about what the user is seeing in near real time. Figure 4 shows a few video examples with the final generated descriptions by our Seeing Bot.

4 CONCLUSIONS

In this demo, we present our Seeing Bot—a video captioning bot that generates a natural language description about what it is seeing in near real time. Particularly, we exploit the state-of-the-art semantic attribute-augmented image and video captioning models to produce both frame-level and video-level captions, and then harness a visual-semantic embedding model to select the best caption as the output sentence, which is further converted to a speech and delivered to an end user via an earphone. A user study conducted on 1,000 real-life videos validates the user experience of our Seeing Bot.

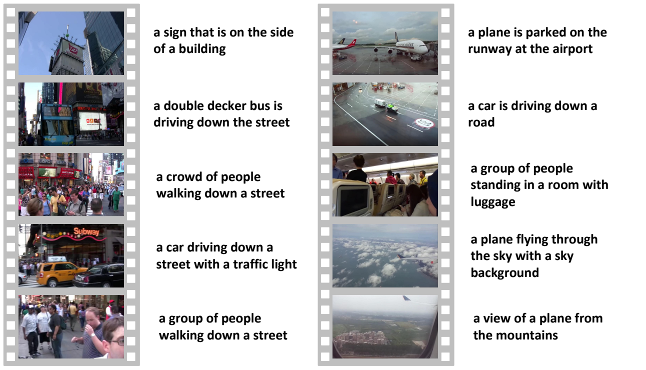
a sign that is on the side of a building
a double decker bus is driving down the street
a crowd of people walking down a street
a car driving down a street with a traffic light
a group of people walking down a street
a plane is parked on the runway at the airport
a car is driving down a road
a group of people standing in a room with luggage
a plane flying through the sky with a sky background
a view of a plane from the mountains

Figure 4: Sentence generation results by Seeing Bot.

Our future works are as follows. First, more semantic attributes will be learnt from large-scale image benchmarks, e.g., YFCC-100M dataset and integrated into image captioning module. Second, similar to [14], generating free-form and open-vocabulary sentences with the semantic attributes is also expected in our Seeing Bot.

Acknowledgments. This work was supported in part by NSFC under contract No. 61325009.

REFERENCES

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL workshop*.

[2] David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*.

[3] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*.

[4] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From Captions to Visual Concepts and Back. In *CVPR*.

[5] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*.

[6] Yehao Li, Ting Yao, Tao Mei, Hongyang Chao, and Yong Rui. 2016. Share-and-Chat: Achieving Human-Level Video Commenting by Search and Multi-View Embedding. In *ACM MM*.

[7] 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 *ECCV*.

[8] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *CVPR*.

[9] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video Captioning with Transferred Semantic Attributes. In *CVPR*.

[10] Yingwei Pan, Ting Yao, Tao Mei, Houqiang Li, Chong-Wah Ngo, and Yong Rui. 2014. Click-through-based cross-view learning for image search. In *SIGIR*.

[11] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

[12] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to Sequence - Video to Text. In *ICCV*.

[13] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *CVPR*.

[14] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects. In *CVPR*.

[15] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2016. Boosting image captioning with attributes. *arXiv preprint arXiv:1611.01646* (2016).