# More is Better: Precise and Detailed Image Captioning using Online Positive Recall and Missing Concepts Mining

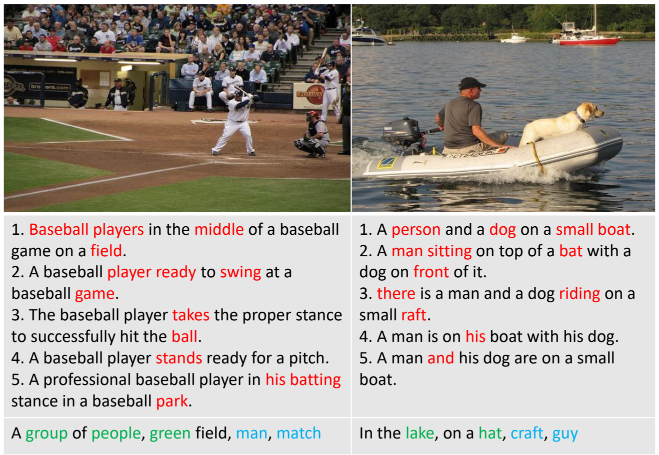
Mingxing Zhang, Yang Yang, Hanwang Zhang, Yanli Ji, Heng Tao Shen, Tat-Seng Chua

***Abstract*—Recently, a great progress in automatic image captioning has been achieved by using semantic concepts detected from the image. However, we argue that existing concepts-to-caption framework, in which the concept detector is trained using the image-caption pairs to minimize the vocabulary discrepancy, suffers from the deficiency of insufficient concepts. The reasons are two-fold: 1) the extreme imbalance between the number of occurrence positive and negative samples of the concept; and 2) the incomplete labelling in training captions caused by the biased annotation and usage of synonyms. In this paper, we propose a method, termed *Online Positive Recall and Missing Concepts Mining* (OPR-MCM), to overcome those problems. Our method adaptively re-weights the loss of different samples according to their predictions for online positive recall and uses a two-stage optimization strategy for missing concepts mining. In this way, more semantic concepts can be detected and a high accuracy will be expected. On the caption generation stage, we explore an element-wise selection process to automatically choose the most suitable concepts at each time step. Thus, our method can generate more precise and detailed caption to describe the image. We conduct extensive experiments on the MSCOCO image captioning dataset and the MSCOCO online test server, which shows that our method achieves superior image captioning performance compared with other competitive methods.**

***Index Terms*—precise and detailed image captioning, semantic concepts, online positive recall, missing concepts mining, element-wise selection.**

## I. Introduction

Image captioning, aiming at machine-generated natural language descriptions for an image, has attracted great interests in the computer vision community. This is because the task is a long way of touching the holy grail in artificial intelligence. How can we say a machine "understands" an image like us? One possible judgment is that the machine should generate a meaningful caption. Generating good captions requires comprehensive content understanding and natural language modeling, both of them are challenging. As the image captioning problem is similar to the machine translation between languages to some extent, many research endeavors have been dedicated to develop an encoder-to-decoder framework: from Convolutional Neural Network (CNN) encoding to Recurrent Neural Network (RNN) decoding [1]–[4]. Thanks to the visual representation ability of CNN and the language modeling power of RNN, those CNN-RNN models have shown promising results. However, the decoder struggles to solve the following two tasks simultaneously: interpreting visual representation and learning language model, which are unnecessarily correlated to each other inherently.



| | |
|---|---|
| 1. Baseball players in the middle of a baseball game on a field.<br>2. A baseball player ready to swing at a baseball game.<br>3. The baseball player takes the proper stance to successfully hit the ball.<br>4. A baseball player stands ready for a pitch.<br>5. A professional baseball player in his batting stance in a baseball park. | 1. A person and a dog on a small boat.<br>2. A man sitting on top of a bat with a dog on front of it.<br>3. there is a man and a dog riding on a small raft.<br>4. A man is on his boat with his dog.<br>5. A man and his dog are on a small boat. |
| A group of people, green field, man, match | In the lake, on a hat, craft, guy |

Fig. 1: Two image samples with their captions and semantic concepts. Five sentences are the annotated training captions for the image. Red words are the semantic concepts contained in the captions. Green words are the missing concepts not described in the captions. Blue words are the synonymous concepts missed in the captions. Best viewed in color.

Recently, a great progress in image captioning has been achieved by using semantic concepts detected from the image, which is very similar to the cognition process of humans [5]–[8]. In order to minimize the vocabulary discrepancy, those methods use the image-caption pairs to train the concept detector. Specifically, those methods consider the most common words in the annotated captions as semantic concepts and leverage multi-label classification methods [9]–[11] or Multiple Instance Learning (MIL) [12], [13] to detect those concepts. Then, a decoder is used to generate the caption according to the detected concepts. Thanks to the decoupling process for visual interpreting and language modeling, the decoder achieves superior captioning performance. Particularly, when directly using the ground-truth of concepts, those models

Mingxing Zhang, Yang Yang, Yanli Ji and Heng Tao Shen are with the Center for Future Media and the School of Computer Science and Engineering, University of Electronic Science and Technology of China. Email: superstar_zhang@hotmail.com, dlyyang@gmail.com, yanliji@uestc.edu.cn, shenhengtao@hotmail.com. Corresponding author: Yang Yang.

Hanwang Zhang is an assistant professor with the School of Computer Science and Engineering, Nanyang Technological University. Email: hanwangzhang@gmail.com

Tat-Seng Chua is the KITHCT Chair Professor with the School of Computing, National University of Singapore. Email: dcscts@nus.edu.sg

can achieve high performance comparable to humans [6].

However, we argue that existing concepts-to-caption framework suffers from the deficiency of insufficient semantics, i.e., detected concepts. The reasons are mainly two-fold. First, in most of the concepts, we observe a severe sample imbalance issue, i.e., for a concept, the number of its negative samples is much larger than that of its positive samples. This makes the concept detector tend to negative prediction and difficult to identify positive concepts. Secondly, because training captions are usually annotated with biased opinions of different annotators, only partial contents in an image are described and many other details are not annotated in the captions. Besides, the same content may be annotated with synonymous concepts due to the variety of language usage and language habits. As illustrated in the left image of Fig. 1, annotators focus on the baseball player and ignore "*a group of people*" who are watching the game. Besides, "*player*" in this image has the same meaning with "*man*", which leads to missing the concept "*man*" in the image. Those missing concepts are not essential for caption generation, however, they inevitably bring noisy labelling information into training data. Consequently, the concept detector cannot precisely characterize their corresponding visual appearance, thereby degrading the performance of concept detection. In brief, both of the above obstacles cause the system lacking sufficient useful semantics, resulting in generating inaccurate and incomplete captions for the images.

In this paper, we propose an effective and efficient approach, termed *Online Positive Recall and Missing Concepts Mining* (OPR-MCM), to overcome the aforementioned problems in image captioning. Specifically, our OPR-MCM method adaptively re-weights the loss of different samples according to their predictions for online positive recall and uses a two-step optimization scheme for missing concepts mining. OPR-MCM not only keeps a higher recall for positive concepts, but also alleviates the bad impact of missing concepts. In this way, the concept detector can well correspond to the visual appearance and preserve sufficient semantics. During the subsequent process of caption generation, we exploit an element-wise selection process to automatically select the most suitable concepts at each time step. Although our detected concepts may contain redundant contents, the caption generator (i.e., RNN) can implicitly model the relations of semantic concepts and choose the most appropriate ones.

We evaluate the proposed method with extensive experiments on the public MSCOCO image captioning dataset and the MSCOCO online test server. Experimental results demonstrate the proposed method achieves superior performance compared with other competitive methods. We visualize the detected concepts and generated captions and the results show that our proposed method can detect more relevant concepts and generate better captions with higher precision and more details. Our contributions are summarized as follows:

- We introduce the deficiency of semantics problem which are neglected in the previous researches for image captioning. Further, we propose the idea that first detecting/providing more semantic concepts to caption generator, and then applying a concepts selection process to automatically choose the most suitable concepts on the caption generation stage. Although our detected concepts may contain some redundant semantics, the generator can automatically choose the appropriate ones. As a result, it can generate more precise and detailed caption to describe the image. This is a novel point of view for the relationship between concept detection and caption generation.
- We propose the OPR-MCM method to overcome the deficiency of semantics caused by the sample imbalance and missing concepts problems for image captioning. We detect sufficient semantic concepts with high accuracy and feed them into the subsequent process for generating more accurate captions. Our proposed OPR-MCM method can be easily extended to many other applications related to the sample imbalance and missing concepts problems.
- We conduct extensive experiments on multiple benchmarks. The results demonstrate our proposed approach outperforms many other competitive approaches across various evaluation metrics. In addition, we show many visualized examples and they as well indicate our proposed method can generate more precise and detailed image captions.

## II. RELATED WORK

### A. Neural Image Captioning

With the success of neural networks for various multimedia applications [14]–[18], methods based on neural networks for image captioning have been proposed. Similar to machine translation [19]–[21], those methods attempt to translate an image into a sentence. Kiros et al. [22] firstly developed a multimodal log-bilinear model using image visual feature, which sets the corner stone for caption generation with neural networks. Kiros et al. [2] furthered their work by simultaneously realizing ranking and generation in a natural fashion. After that, Mao et al. [3] took a next step by employing a recurrent neural language model instead of the feed-forward ones. Vinyals et al. [4] and Donahue et al. [1] took advantage of Long Short Term Network (LSTM) [23] for language modelling. Latterly, attention mechanism is also introduced for image captioning. Rather than encoding the whole image into a static feature vector, attention models [24]–[26] process the image as a set of local regions and dynamically focus on different local regions when generating caption words. Yang et al. [27] proposed a review network for image caption generation. The review network performs a number of review steps using attention mechanism and outputs a set of thought vectors that are used as the input for caption generation. Lu et al. [28] proposed an adaptive attention model which decides whether and where to attend to the image at each time step, in order to extract meaningful information for sequential words generation. Although those methods have achieved promising results, they still make the decoder solve two subtasks of interpreting visual representation and learning language model simultaneously, which diminishes the ability of both subtasks.

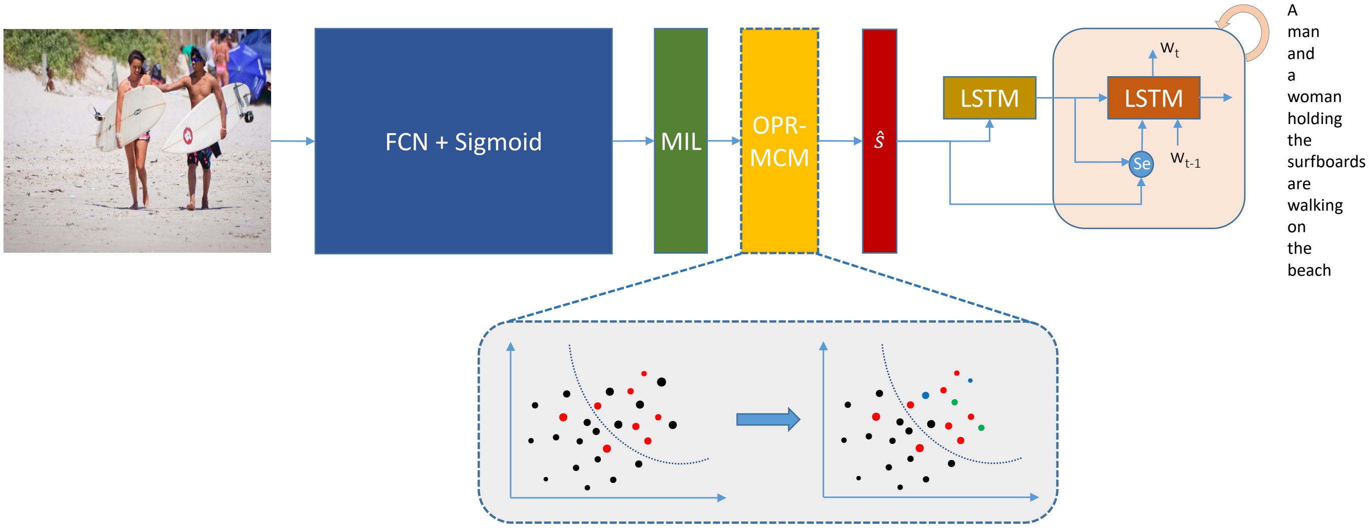

FCN + Sigmoid
MIL
OPR-MCM
$\hat{s}$
LSTM
LSTM
$w_t$
$w_{t-1}$
Se
A man and a woman holding the surfboards are walking on the beach

Fig. 2: The architecture of our proposed deep learning framework for image captioning. Dashed line box is the proposed *Online Positive Recall and Missing Concepts Mining* (OPR-MCM) method which is used in the training stage. FCN + Sigmoid represents the Fully Convolutional Network followed by a Sigmoid activation layer. MIL represents the Multiple Instance Learning. $\hat{s}$ is the prediction score vector for concepts. Red and black points in OPR-MCM represent positive and negative concepts, respectively. Green points are the missing details and blue points are the missing synonymous concepts. Different point sizes represent different weights for the sample loss. “Se” represents the concepts selection process for words generation. Best viewed in color.

### B. Concept Detection for Captioning

Recently, with the development of visual understanding techniques [29]–[36], great progress in image captioning has been achieved by first detecting semantic concepts in an image and then generating the caption for the image [5]–[7], [37], [38]. Because of decoupling the visual interpreting and language modelling processes, the decoder can achieve superior captioning performance. In [5], Fang et al. proposed a model that firstly used the Fully Convolutional Network (FCN) [39], [40] and Multiple Instance Learning (MIL) [12], [13] to detect some keywords (concepts), then those keywords are used to generate the caption. A similar model was proposed in [37] which conducts max pooling on the prediction score vectors from different image sub-regions and then takes leverage of LSTM as the language model for caption generation. Following those methods, You et al. [6] proposed a semantic attention model which dynamically attends to the related words for image captioning. You et al. [7] further jointly utilized the detected high-level attributes (concepts) and image representation to boost image captioning performance. Similar to image captioning, [41] and [42] detected various concepts in the video stream for video captioning. In this paper, our proposed method is also based on the concept detection process. However, our method aims to generate more precise and detailed captions, which is not well explored in the previous researches.

### C. Diverse Captioning

In order to better describe the image, several methods aiming to generate diverse captions have been proposed. In [43], Wang et al. proposed a model called GroupTalk which acts as if a group of people are describing the image with different preferences. In [44], Vijayakumar et al. proposed the Diverse Beam Search (DBS) to replace the original beam search method. The DBS can decode a list of diverse captions by optimizing a diversity-augmented objective. To describe a large number of objects not present in dataset, Venugopalan et al. [45] proposed a method that simultaneously learns from multiple data sources which have auxiliary objectives. Wang et al. [46] explored more diverse image captioning using conditional variational autoencoders (CVAEs) with an additive Gaussian encoding space. Lately, as the development of Generative Adversarial Networks (GAN), adversarial training is also introduced for diverse caption generation [47]–[49]. For instance, Dai et al. [47] introduced a Conditional GAN, which jointly learns a generator to produce descriptions and an evaluator to assess the descriptions, to improve caption diversity. In [48], Shetty et al. employed adversarial training and combined with an approximate Gumbel sampler to generate multiple and diverse captions. Unlike those methods, our method with the purpose of precise and detailed captioning also results in diverse contents in the generated captions.

## III. The Proposed Method

Fig. 2 shows the architecture of the proposed method for image captioning. Our method is based on the Concepts-to-Caption framework. The forward part is the concept detection module and the recurrent part is the captioning module.

### A. Concepts-to-Caption

A Concepts-to-Caption framework consists of two modules: the concept detecting module identifies semantic concepts from the image; and the caption generator produces the caption using the detected concepts. Given an image $I$, a set of semantic concepts is detected, denoted as a fixed length vector $\hat{s} \in \mathbb{R}^n$ which is computed as follows:

$$\hat{s} = f_{enc}(I), \tag{1}$$

where $n$ is the number of concepts, $f_{enc}$ is the detector, each element $\hat{s}_i$ in $\hat{s}$ is the prediction score which represents the probability of the $i$-th concept appearing in the image. Note that $\hat{s}$ is a high-level representation for the image, which contains more informative local and global semantics, such as objects, actions, scenes.

LSTM, a special implementation of RNN, is widely used to model the words dependency and generate the language sentence. In particular, the caption generator takes $\hat{s}$ as a condition and generates a caption $C = \{w_{-1}, w_0, w_1, ..., w_L\}$, where $w_t$ is the word token ($w_{-1}$ and $w_L$ are the *START* token and *END* token, respectively), $L$ is the length of caption. We use the LSTM following the work [50], which takes forward propagation at time step $t$ as follows:

$$\begin{cases} i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \\ f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \\ o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \\ g_t = tanh(W_g[h_{t-1}, x_t] + b_g), \\ c_t = f_t \odot c_{t-1} + i_t \odot g_t, \\ h_t = o_t \odot tanh(c_t), \end{cases} \tag{2}$$

where $c_t$ and $h_t$ are the unit's cell and hidden states, respectively. $i_t, f_t, o_t$ are the activations of input gate, forget gate, and output gate, respectively. $x_t$ is the input content. $W$ and $b$ are the weight and bias parameters, respectively. $\sigma(\cdot)$ is the sigmoid function, and $tanh(\cdot)$ is the output activation function.

At time step $t$, the caption generator takes $x_t$ as input, and computes a distribution over all possible words:

$$\begin{cases} h_t = LSTM_c(x_t, h_{t-1}, c_{t-1}), \\ y_t = softmax(W_g \cdot h_t + b_g), \end{cases} \tag{3}$$

where $h_{t-1}$ is the hidden state of the recurrent units at $t-1$, $W_g, b_g$ are the parameters of the output layer, and $LSTM_c(\cdot)$ is a forward step of LSTM unit. $y_t$ defines a distribution over all possible words, from which the word $w_t$ is generated.

### B. Detecting More Concepts

To detect sufficient concepts from the image, we take leverage of a similar detection process proposed in [5]. As Fig. 2 shows, a Fully Convolutional Network (FCN) followed by a Sigmoid activation layer is performed over the image to generate a spatial score map. The activation in each map grid is the prediction score vector, of which each element represents the probability of a concept appearing in its corresponding image sub-region. This equals to applying a CNN and sigmoid classifier on the image sub-region. The FCN can effectively scan different sub-regions for possible contents. After that, we treat the whole image as a bag of sub-regions and use Multiple Instances Learning (MIL) to pool the prediction score vectors from all sub-regions to obtain the last score vector $\hat{s}$ for the whole image.

*1) Alleviating Sample Imbalance:* Conventionally, we can use the standard cross entropy loss function to optimize the concept detection network which is aforementioned. However, we may suffer from the severe problems of sample imbalance and missing concepts. To deal with the sample imbalance problem, down-sampling of majority and up-sampling of minority are often employed. Nonetheless, those strategies will change the original dataset and easily cause the over-fitting problem. Instead, we propose an online method which adaptively re-weights the loss of different samples according to their predictions on the training stage. Concretely, the final loss of our method is the combination of 1) the ground-truth of training label, and 2) the current prediction score of the concept, which is depicted as follows:

$$L(s_j^{(i)}, \hat{s}_j^{(i)}) = \begin{cases} -\left(p + \alpha \cdot \left(1 - \hat{s}_j^{(i)}\right)\right) \cdot log\hat{s}_j^{(i)}, & if\ s_j^{(i)} = 1; \\ -\left(q + \beta \cdot \hat{s}_j^{(i)}\right) \cdot log\left(1 - \hat{s}_j^{(i)}\right), & if\ s_j^{(i)} = 0, \end{cases} \tag{4}$$

where $s_j^{(i)}$ is the ground truth corresponding to the $i$-th image and the $j$-th concept, and $\hat{s}_j^{(i)}$ is the corresponding concept prediction score. $p$, $q$ and $\alpha$, $\beta$ are the non-negative constant hyper-parameter.

The underlying principle of designing Eq. 4 is to enable flexible control of positive and negative samples in the training process. As seen, if the sample is difficult for classification (i.e., the cross entropy loss tend to be large), we can then impose a large weight to focus on this difficult sample. For instance, if a positive sample is predicted more likely to be negative, we give more penalization to its loss. Subsequently, the gradient with respective to $L(s_j^{(i)},\ \hat{s}_j^{(i)})$ becomes larger, thereby gradually pushing the detector to correctly predict the positive sample. For the negative samples, we observe that their prediction scores are usually quite low, which implies that they can be easily classified. Therefore, we can appropriately ignore a large number of these easy negative samples to alleviate the adverse impact of sample imbalance problem and improve the discriminative ability for the concept detector. It is worth noting that $p$ or $q$ in Eq. 4 is used to setup a base weight for the sample loss, which aims to further differentiate the prediction for positive and negative samples.

*2) Handling Missing Concepts:* By minimizing the value of Eq. 4 to learn the concept detector, we can recall more positive concepts for the image. However, as analyzed before, there exist many missing concepts in the training image data (i.e., noise), which exerts negative effects on learning reliable concept detector for precisely capturing the visual appearance of concepts. To overcome this problem, we propose a two-step optimization scheme based on the consistent prediction. Specifically, we first use the original concept labels and Eq. 4 to pre-train the concept detector. With the pre-training of the detector, the samples with a missing concept will be probably inconsistent with true negative samples in terms of the prediction scores of this concept. This is because the samples with the missing concept have the similar visual appearance of positive samples of this concept, while having a large difference of visual appearance from the true negative samples. To recognize the missing concepts, we first obtain the prediction scores for all the training samples and then compute a probability $p(s_j^{(i)} = 1|\hat{s}_j^{(i)})$, which represents how likely the $i$-th image misses the $j$-th concept. Concretely, we compute the probability according to the original concept label and the prediction scores as follows:

$$p(s_j^{(i)}=1|\hat{s}_j^{(i)})=\frac{\frac{|\{s:s\in S_{+,j}\wedge s\leqslant\hat{s}_j^{(i)}\}|}{N_{+,j}}}{\frac{|\{s:s\in S_{+,j}\wedge s\leqslant\hat{s}_j^{(i)}\}|}{N_{+,j}}+\frac{|\{s:s\in S_{-,j}\wedge s\geqslant\hat{s}_j^{(i)}\}|}{N_{-,j}}}, \tag{5}$$

where $S_{+,j}$ and $S_{-,j}$ are the set of the prediction scores for all the labeled positive and negative samples in the $j$-th concept, respectively. $N_{+,j}$ and $N_{-,j}$ are the numbers of all the labeled positive and negative samples for the $j$-th concept, respectively. $|\cdot|$ is the cardinality of a set. As illustrated in Eq. 5, $\frac{|\{s:s\in S_{+,j}\wedge s\leqslant\hat{s}_j^{(i)}\}|}{N_{+,j}}$ is the proportion of labeled positive samples with a prediction score no larger than $\hat{s}_j^{(i)}$, which can be regarded as the confidence for the sample with the prediction score $\hat{s}_j^{(i)}$ being positive. Similarly, $\frac{|\{s:s\in S_{-,j}\wedge s\geqslant\hat{s}_j^{(i)}\}|}{N_{-,j}}$ is the proportion of labeled negative samples with a prediction score no smaller than $\hat{s}_j^{(i)}$. It can be treated as the confidence for the sample with the prediction score $\hat{s}_j^{(i)}$ being negative.

According to Eq. 5, we recognize the $i$-th concept as the missing concept in the $j$-th negative sample if we have that $p(s_j^{(i)} = 1|\hat{s}_j^{(i)})$ is larger than a threshold $\tau$. Then, we turn these negative samples to be positive ones (i.e., setting $s_j^{(i)} = 1$) and finally re-train the concept detector using Eq. 4 again to create more reliable semantics for images.

In fact, our proposed method for missing concepts mining has the similar effect with the minimum entropy regularization, which was previously studied in [51]. The minimum entropy regularization encourages the model to have a high confidence in label prediction, while our method reducing the label noise also results in label prediction with a high confidence.

### C. Concept Selection

As shown in Fig. 2, the caption generator takes the prediction score vector $\hat{s}$ as a condition to generate the image caption. Specifically, we use the score vector $\hat{s}$ to initialize the LSTM and also feed it to the LSTM at each time step for caption generation. Because we have detected sufficient semantic concepts, we introduce a concept selection process to help the caption generator choose the most suitable concepts when generating each caption word. Rather than applying the standard attention mechanism, we explore an element-wise selection process for the prediction score vector $\hat{s}$ as follows:

$$\begin{cases} x_{-1} = W_s \cdot \hat{s}, \\ (h_{-1}, c_{-1}) = LSTM_i\left\{x_{-1}, \mathbf{0}, \mathbf{0}\right\}, \\ \alpha_t = sigmoid(W_a \cdot h_{t-1} + b_a), \quad t \geq 0, \\ \tilde{s}_t = \alpha_t \odot \hat{s}, \quad t \geq 0, \\ x_t = E \cdot a_{t-1} + W_c \cdot \tilde{s}_t, \quad t \geq 0, \end{cases} \tag{6}$$

where $\mathbf{0}$ is the all-zero vector. $W_s$ and $W_c$ are the weight parameters for transforming concept prediction scores into LSTM input space. $W_a$ and $b_a$ are the selection parameters. $\alpha_t$ is the weights vector of element-wise selection at time step $t$. $a_{t-1}$ is the one-hot coding of generated word $w_{t-1}$. $E$ is the word embedding matrix. $\tilde{s}_t$ is the score vector after the element-wise selection process and $x_t$ is the input of LSTM.

In Eq 6, we use the $sigmoid$ instead of $softmax$ as the activation function. Thus, our element-wise selection has a better adaptability which can either choose several concepts at one time or not choose any semantic concepts when generating some non-meaningful words, such as *"the", "of", "to"*. In this situation, the caption generator generates the next word only relying on the previous words.

We also study the variants of weight relaxation for the initial module. Let $w_c$ and $w_i$ denote the parameters for the $LSTM_c$ and $LSTM_i$, respectively. The first variant follows the common setting in LSTMs, where weights are shared among all the units, i.e., $w_c = w_i$. We also observe that the initial LSTM has different function compared with the LSTM for caption generation. Hence, we also conduct the experiments using the second variant where weights are untied, i.e. $w_c \neq w_i$.

## IV. Experiments

### A. Datasets and Settings

We mainly evaluate our approach on the popular MSCOCO image captioning dataset. The MSCOCO dataset has 123,287 images, most of which contain multiple objects in the context of complex scenes. Each image in the MSCOCO dataset has at least 5 captions which are manually annotated by different people. In order to conduct a fair comparison, we use the publicly available split[1] of training, validating and testing sets.

We follow the work [5] that uses the 1,000 most common words appearing in the training captions as our semantic concepts. Those concepts may belong to any part of a sentence, such as nouns, verbs, and adjectives. In our experiments, we set $p = 1.0$, $q = 0.001$, $\alpha = 3.0$ and $\beta = 1.0$. The threshold $\tau$ is set to 0.99. We implement our concept detector based on Caffe [52], and use the SGD algorithm to optimize the concept detector.

To construct the vocabulary used in the caption generator, we select 9,998 most common words appearing in the training captions, and another 2 words with one representing *START/END* token and the other one representing all other words. In our LSTM caption generator, the dimension of cell and hidden vectors is set to 512. We use the *adam* algorithm [53] to optimize the caption generator. In this work, the caption generator and the concept detector are trained independently.

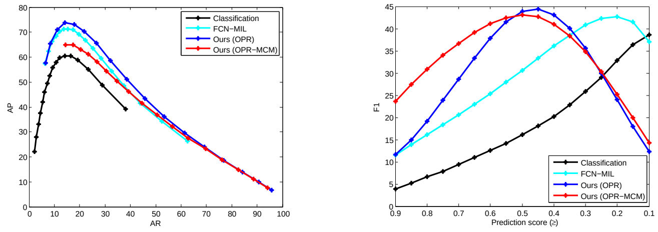
### B. Compared Approaches

To comprehensively verify the performance of our method, we compared our method with the following competitive methods.

- Google NIC [4]: NIC encodes the image into a static representation though CNN, then feeds the image representation into LSTM at the initial time step.

[1] https://github.com/karpathy/neuraltalk

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2855415, IEEE Transactions on Image Processing



(a) AP v.s. AR results

(b) F1 results

Fig. 3: The AP, AR and F1 measure results for different methods. The points from left to right in each curve are the measure values when the prediction score thresholds are $0.9, 0.85, 0.8, \ldots, 0.1$ respectively. All measure values are reported in terms of percentage (%). Best viewed in color.

- LRCN [1]: LRCN inputs both image representation and previous word into LSTM at each time step for caption words generation.
- Visual Hard Attention & Soft Attention [24]: Visual attention is applied on the convolutional feature map of an image when generating caption words through two kinds of mechanisms: 1) stochastic attention mechanism achieved by reinforcement learning (Hard Attention) and 2) deterministic attention mechanism with standard back-propagation (Soft Attention).
- GAN [47]: This method utilizes the Conditional Generative Adversarial Networks (CGAN) to jointly learn a generator and a evaluator to improve the caption naturalness and diversity.
- Review Network [27]: This method performs a number of review steps with attention mechanism on the image sub-regions to output a set of thought vectors, which are more compact, abstractive, and global representations for caption generation.
- GroupTalk [43]: This method learns multiple image caption distributions simultaneously and effectively mimics the diversity of the image captions to generate the appropriate descriptions with both diversity and high quality.
- ATT [6]: This method takes semantic attention on the detected attributes (concepts) at each time step in LSTM for image captioning.
- Sentence-Condition [54]: This method exploits text-conditional semantic attention to generate semantic guidance for sentence generation.
- Att-CNN+LSTM [37]: This method firstly detects the keywords (concepts) in different image regions and then conducts max pooling to obtain the prediction score vector of words, which is used to initialize the LSTM to generate the image caption.
- MSM [7]: This method exploits several ways to jointly leverage the high-level attributes and image visual representation to boost image captioning performance. We only report their best results for comparison.
- SCA-CNN [55]: This method incorporates Spatial and Channel-wise Attentions on multi-layer feature maps in a CNN for image caption generation.
- SCN-LSTM [8]: The method uses the detected concepts and develops a Semantic Compositional Network (SCN) which extends each weight matrix of the LSTM to more effectively assemble the meanings of individual concepts to generate the caption.
- Skeleton-Attribute [56]: This method decomposes the original image description into a skeleton sentence and its attributes, and generates the skeleton sentence and attribute phrases separately.
- Adaptive Attention [28]: At each time step, this method decides whether and where to attend to the image, in order to extract meaningful information for sequential words generation.

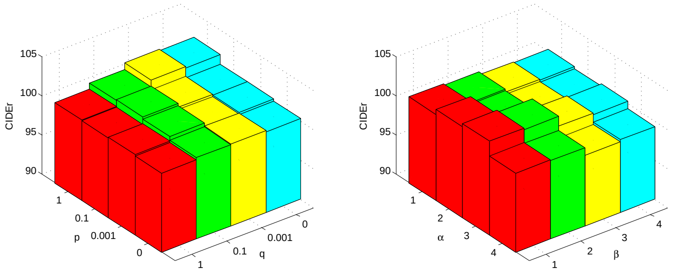
## *C. Performance Evaluation*

*1) Evaluation of Concept Detection:* To gain insight into our method for concept detection, we show the results of the Average Precision (AP), Average Recall (AR) and F1 measures on the MSCOCO testing split. Specifically, the AP is the average of all samples for how many percentages the detected concepts are the correct ones, while the AR is the average of all samples for how many percentages the detected correct concepts account for all concepts in the image. The F1 measure is the harmonic average of AP and AR. We collect the concepts whose prediction scores are no smaller than a threshold as the detected concepts. We ignore 14 functional concepts without any specific meanings such as *"the", "of", "to"*. We compare our method with two baselines. The first one is a whole image classifier which uses the 4096-dimensional feature from the last fully connected layer of VGG16 network [57]. This feature is fine-tuned for the concepts classification using a logistic regression loss. The second is the FCN-MIL model proposed in [5], which lacks our OPR-MCM process.

Fig. 3 shows the results of AP, AR and F1 measures for different methods. Ours (OPR) represents using only Eq. 4 to pre-train the concept detector while Ours (OPR-MCM) represents using the two-step optimization scheme to train the concept detector. The points from left to right in each curve are the measure values when the prediction score thresholds are $0.9, 0.85, 0.8, \ldots, 0.1$. From Fig. 3 (a), we can see that our method with OPR achieves the highest AP-AR scores. This demonstrates that our OPR method can effectively alleviate the negative influence of sample imbalance and improve the detection accuracy. While our OPR-MCM method achieves slightly lower AP-AR performance than the FCN-MIL baseline. This is because our OPR-MCM method can detect more relevant contents which are not annotated in the training data. We also see that our method with OPR-MCM has a much higher recall compared with other methods. For example, when the prediction score threshold is $0.9$, OPR-MCM achieves AR performance of 14.4%, while others are all lower than 7%. Moreover, our OPR-MCM method achieves a much higher AP result when the threshold is high, such as 64.9% for OPR-MCM compared with 57.6% for OPR and 57.5% for FCN-MIL at the threshold point 0.9. The reason is that our OPR-MCM method can reduce the noise in the training label set and make the concept detector more precisely capture the visual appearance of concepts. Therefore, our OPR-MCM method can detect more relevant concepts with a higher confidence.

| Method | OPR-MCM (VGG16) | | | | | FCN-MIL (VGG16) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BLEU-1 | BLEU-4 | METEOR | ROUGE_L | CIDEr | BLEU-1 | BLEU-4 | METEOR | ROUGE_L | CIDEr |
| V1 | 72.6 | 32.5 | 25.9 | 54.0 | 100.8 | 71.8 | 31.7 | 25.0 | 53.1 | 96.3 |
| V1 + WR | 73.0 | 32.7 | 26.1 | 54.4 | 101.6 | 71.9 | 31.5 | 25.1 | 53.2 | 96.6 |
| V2 | 73.1 | 32.7 | 26.0 | 54.3 | 101.5 | 71.9 | 31.2 | 25.0 | 53.1 | 96.4 |
| V2 + WR | 73.3 | 32.9 | 25.9 | 54.3 | 101.2 | 72.4 | 31.4 | 25.2 | 53.3 | 97.5 |
| V3 | 73.7 | 33.3 | 26.3 | 54.6 | 103.6 | 72.3 | 31.9 | 25.3 | 53.3 | 97.9 |
| V3 + WR | **74.2** | **33.9** | **26.2** | **54.8** | **104.3** | **72.4** | **32.2** | **25.3** | **53.5** | **98.6** |
| V3 + WR (OPR) | 73.6 | 33.1 | 25.9 | 54.1 | 102.0 | - | - | - | - | - |

TABLE I: Comparison of image captioning performance between different interfaces and weight relaxation. All values are reported in terms of percentage (%).



(a)

(b)

Fig. 4: CIDEr scores of our method (OPR) using different selection of $p$, $q$, $\alpha$ and $\beta$. Specifically, sub-figure (a) shows the results for different selection of $p$ and $q$ when we set $\alpha = 3$ and $\beta = 1$. Sub-figure (b) shows the results for different selection of $\alpha$ and $\beta$ when we set $p = 1$ and $q = 0.001$. All results are reported in terms of percentage (%).

We also show the F1 measure in Fig. 3 (b). From those results, we can see that the best F1 scores of our OPR and OPR-MCM methods are 44.5% and 43.1%, respectively, which are higher than the best F1 score of the FCN-MIL baseline, namely 42.6%. Moreover, our OPR-MCM method reaches the best F1 score at the threshold point 0.5, and also achieves high F1 scores when the threshold approaches 1. While the FCN-MIL baseline reaches the best F1 score at the threshold point 0.2, and achieves high F1 scores only when the threshold is near to 0. This demonstrates the scores of many concepts predicted by the FCN-MIL baseline are concentrated on a limited range of small values. Thus, the semantic representation $\hat{s}$ produced by our OPR-MCM method is more diverse and discriminative than the representation produced by the FCN-MIL baseline.

*2) Evaluation of interfaces and weight relaxation:* Table I shows the performance of different interfaces between concept detector and caption generator as well as the weight relaxation trick on the MSCOCO dataset. For better comparison, we compare our model with several other variations. The first one (V1) only inputs the prediction score vector $\hat{s}$ into the first LSTM unit to initialize the LSTM sequence, while the second one (V2) also feeds the score vector $\hat{s}$ at each time step but without the concept selection process. V3 represents the model that takes leverage of our concept selection process when generating each caption word. WR is the weight relaxation trick. The columns in OPR-MCM are the results of our OPR-MCM method and the columns in FCN-MIL are the results using the FCN-MIL baseline for concept detection. Note that V3 + WR (OPR) are the results of our OPR method for concept detection, namely only using Eq. 4 to pre-train the concept detector.

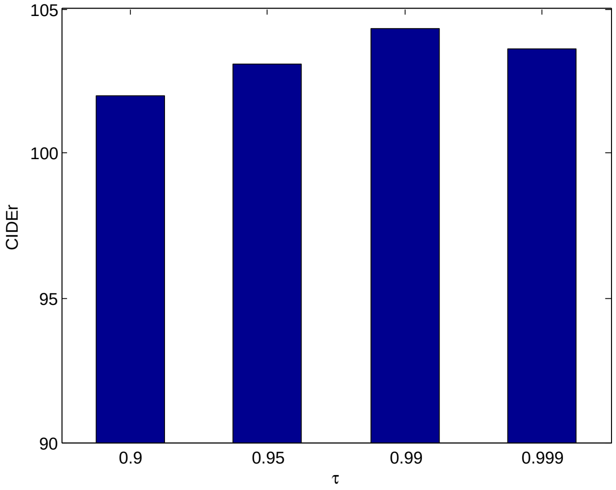


Fig. 5: CIDEr scores of our method (OPR-MCM) using different selection of $\tau$. All results are reported in terms of percentage (%).

From those results, we can see that taking our element-wise selection for the prediction score vector at each step with weight relaxation achieves the best performance. Furthermore, we can also see all variations using the score vector obtained by our OPR-MCM achieve higher performance than those obtained by the FCN-MIL baseline. In the V3 + WR models, the performance of our OPR is lower than our OPR-MCM but better than the FCN-MIL baseline. Those results demonstrate the effectiveness of our methods (OPR and OPR-MCM).

*3) Influence of hyper-parameters:* We conduct additional experiments with different selection of hyper-parameters, including $p$, $q$, $\alpha$, $\beta$ and $\tau$, to show the influence of those hyper-parameters on our method. In all experiments, we use the CIDEr score to evaluate the performance of our method with different parameter selection.

Fig. 4 shows the CIDEr scores of our method (OPR) using different selection of $p$, $q$, $\alpha$, $\beta$. Specifically, Fig. 4 (a) shows the results for different selection of $p$ and $q$ when we set $\alpha = 3$ and $\beta = 1$. From those results, we can see that $p$ with large

| Method | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE_L | CIDEr |
|---|---|---|---|---|---|---|---|
| Google NIC† [4] | 66.3 | 42.3 | 27.7 | 18.3 | - | - | - |
| LRCN [1] | 60.3 | 38.0 | 25.4 | 17.1 | 16.9 | - | - |
| Soft Attention [24] | 66.7 | 43.4 | 28.8 | 19.1 | 18.5 | - | - |
| Hard Attention [24] | 66.9 | 43.9 | 29.6 | 19.9 | 18.5 | - | - |
| GAN [47] | - | - | 30.5 | 20.7 | 22.4 | 47.5 | 79.5 |
| Review Network [27] | - | - | - | 29.0 | 23.7 | - | 88.6 |
| GroupTalk [43] | 68.5 | 46.9 | 32.8 | 23.5 | - | - | - |
| ATT [6] | 70.9 | 53.7 | 40.2 | 30.4 | 24.3 | - | - |
| Sentence-Condition [54] | 71.6 | 54.5 | 40.5 | 30.1 | 24.7 | - | 97.0 |
| Att-CNN+LSTM [37] | 74.0 | 56.0 | 42.0 | 31.0 | 26.0 | - | 94.0 |
| MSM [7] | 73.0 | 56.5 | 42.9 | 32.5 | 25.1 | 53.8 | 98.6 |
| **Ours** (VGG16) | **74.2** | **57.9** | **44.3** | **33.9** | **26.2** | **54.8** | **104.3** |
| SCA-CNN [55] | 71.9 | 54.8 | 41.1 | 31.1 | 25.0 | - | - |
| SCN-LSTM [8] | 74.1 | 57.8 | 44.4 | 34.1 | 26.1 | - | 104.1 |
| Skeleton-Attribute [56] | 74.2 | 57.7 | 44.0 | 33.6 | 26.8 | 55.2 | 107.3 |
| Adaptive Attention [28] | 74.2 | 58.0 | 43.9 | 33.2 | 26.6 | - | 108.5 |
| **Ours** (ResNet101) | **75.8** | **59.6** | **46.0** | **35.6** | **27.3** | **56.0** | **110.5** |

TABLE II: Image captioning performance compared with other methods on the MSCOCO dataset. † indicates a different test data split. All values are reported in terms of percentage (%).

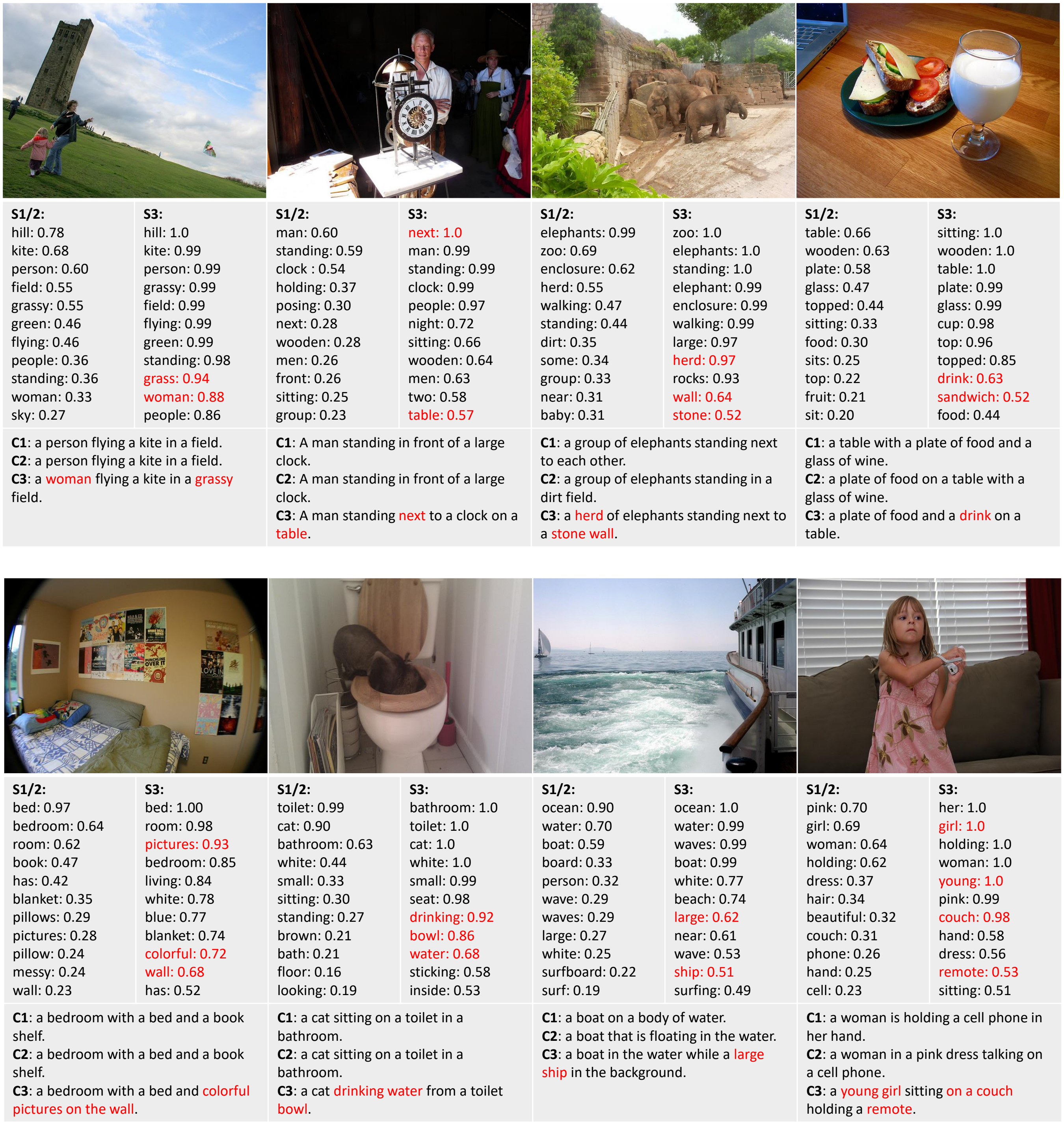
| Method | BLEU-1 | | BLEU-2 | | BLEU-3 | | BLEU-4 | | METEOR | | ROUGE | | CIDEr | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C5 | C40 | C5 | C40 | C5 | C40 | C5 | C40 | C5 | C40 | C5 | C40 | C5 | C40 |
| Google NIC [4] | 71.3 | 89.5 | 54.2 | 80.2 | 40.7 | 69.4 | 30.9 | 58.7 | 25.4 | 34.6 | 53.0 | 68.2 | 94.3 | 94.6 |
| LRCN [1] | 71.8 | 89.5 | 54.8 | 80.4 | 40.9 | 69.5 | 30.6 | 58.5 | 24.7 | 33.5 | 52.8 | 67.8 | 92.1 | 93.4 |
| Hard Attention [24] | 70.5 | 88.1 | 52.8 | 77.9 | 38.3 | 65.8 | 27.7 | 53.7 | 24.1 | 32.2 | 51.6 | 65.4 | 86.5 | 89.3 |
| SCA-CNN [55] | 71.2 | 89.4 | 54.2 | 80.2 | 40.4 | 69.1 | 30.2 | 57.9 | 24.4 | 33.1 | 52.4 | 67.4 | 91.2 | 92.1 |
| ATT [6] | 73.1 | 90.0 | 56.5 | 81.5 | 42.4 | 70.9 | 31.6 | 59.9 | 25.0 | 33.5 | 53.5 | 68.2 | 94.3 | 95.8 |
| MSM [7] | 73.9 | 91.9 | 57.5 | 84.2 | 43.6 | 74.0 | 33.0 | 63.2 | 25.6 | 35.0 | 54.2 | 70.0 | 98.4 | 100.3 |
| SCN-LSTM [8] | 74.0 | 91.7 | 57.5 | 83.9 | 43.6 | 73.9 | 33.1 | 63.1 | 25.7 | 34.8 | 54.3 | 69.6 | 100.3 | 101.3 |
| Adaptive Attention [28] | 74.6 | 91.8 | 58.2 | 84.2 | 44.3 | 74.0 | 33.5 | 63.3 | 26.4 | 35.9 | 55.0 | 70.6 | 103.7 | 105.1 |
| **Ours** (ResNet101) | **74.9** | **92.7** | **58.7** | **85.3** | **45.0** | **75.5** | **34.5** | **65.0** | **26.9** | **36.6** | **55.3** | **71.3** | **105.0** | **105.6** |

TABLE III: The leaderboard for different methods on the MSCOCO online test server. All values are reported in terms of percentage (%).

values is generally better than that with small values. This is because larger $p$ makes positive samples have a larger base weight for their loss. Thus, it makes all the positive samples have relatively enough gradient to obtain higher activations (e.g., pushes the positive samples whose predictions are near to 1 to come more close to 1). This will make the detector recall more concepts for caption generation. We can also see small $q$ can achieve better results. This is due to that small $q$ can make the detector ignore large amount of easily classified negative examples, thus reduce the bad influence of sample imbalance problem. However, too small value of $q$ (e.g., $q = 0$) is not always better. In our paper, we set $p = 1$ and $q = 0.001$ to keep a base weight for the loss of positive and negative samples. Fig. 4 (b) shows the results for different selection of $\alpha$ and $\beta$ when we set $p = 1$ and $q = 0.001$. Because $\alpha$ and $\beta$ control the degree of adaptive re-weighting for positive and negative samples respectively, the ratio between them is very important for the last performance. From those results, we can see that $\alpha = 3$, $\beta = 1$ (ratio = 3) achieves the best performance. Other selections which the ratio is too large or too small perform worse. Besides, when $\alpha$ and $\beta$ keep the same ratio, smaller values of them generally achieves better results. The reason is that $\alpha$ and $\beta$ take a part of deciding the parameter update rate on the optimization stage, smaller value keeps a stable parameter update rate (the learning rate of VGG network for concept detection is 0.001). Thus, it contributes to reach a better optimization result.

Moreover, we can see that CIDEr scores of all the experiments are around 100%. This demonstrates our method is not very sensitive to the selection of those parameters. We introduce these parameters only to add more flexibility to our method and further improve the performance. In fact, we can remove those parameters, which equals to set $p = 0$, $q = 0$, $\alpha = 1$ and $\beta = 1$. In this case, our method also has the ability of adaptive re-weighting. We have conducted the experiment without those parameters. The CIDEr score of this case is 101.2% which is still better than the baseline score 98.6%.

Fig. 5 shows the CIDEr scores of our method (OPR-MCM) using different selection of $\tau$. From those results, we can see $\tau = 0.99$ achieves the best performance. Too small or too large values of $\tau$ decrease the captioning performance. There are mainly two reasons. Firstly, too large values of $\tau$ will make some missing concepts not recognized. Thus, the detector cannot capture the visual features of concepts very well, which results in that some related concepts cannot be detected. Thus, the captioning performance is decreased. Secondly, too small values of $\tau$ will make some true negative concepts misclassified as missing concepts, which will decrease the detection accuracy. Consequently, it also decreases the captioning performance.



| S1/2: | S3: |
|---|---|
| hill: 0.78 | hill: 1.0 |
| kite: 0.68 | kite: 0.99 |
| person: 0.60 | person: 0.99 |
| field: 0.55 | grassy: 0.99 |
| grassy: 0.55 | field: 0.99 |
| green: 0.46 | flying: 0.99 |
| flying: 0.46 | green: 0.99 |
| people: 0.36 | standing: 0.98 |
| standing: 0.36 | grass: 0.94 |
| woman: 0.33 | woman: 0.88 |
| sky: 0.27 | people: 0.86 |

**C1**: a person flying a kite in a field.
**C2**: a person flying a kite in a field.
**C3**: a woman flying a kite in a grassy field.

| S1/2: | S3: |
|---|---|
| man: 0.60 | next: 1.0 |
| standing: 0.59 | man: 0.99 |
| clock : 0.54 | standing: 0.99 |
| holding: 0.37 | clock: 0.99 |
| posing: 0.30 | people: 0.97 |
| next: 0.28 | night: 0.72 |
| wooden: 0.28 | sitting: 0.66 |
| men: 0.26 | wooden: 0.64 |
| front: 0.26 | men: 0.63 |
| sitting: 0.25 | two: 0.58 |
| group: 0.23 | table: 0.57 |

**C1**: A man standing in front of a large clock.
**C2**: A man standing in front of a large clock.
**C3**: A man standing next to a clock on a table.

| S1/2: | S3: |
|---|---|
| elephants: 0.99 | zoo: 1.0 |
| zoo: 0.69 | elephants: 1.0 |
| enclosure: 0.62 | standing: 1.0 |
| herd: 0.55 | elephant: 0.99 |
| walking: 0.47 | enclosure: 0.99 |
| standing: 0.44 | walking: 0.99 |
| dirt: 0.35 | large: 0.97 |
| some: 0.34 | herd: 0.97 |
| group: 0.33 | rocks: 0.93 |
| near: 0.31 | wall: 0.64 |
| baby: 0.31 | stone: 0.52 |

**C1**: a group of elephants standing next to each other.
**C2**: a group of elephants standing in a dirt field.
**C3**: a herd of elephants standing next to a stone wall.

| S1/2: | S3: |
|---|---|
| table: 0.66 | sitting: 1.0 |
| wooden: 0.63 | wooden: 1.0 |
| plate: 0.58 | table: 1.0 |
| glass: 0.47 | plate: 0.99 |
| topped: 0.44 | glass: 0.99 |
| sitting: 0.33 | cup: 0.98 |
| food: 0.30 | top: 0.96 |
| sits: 0.25 | topped: 0.85 |
| top: 0.22 | drink: 0.63 |
| fruit: 0.21 | sandwich: 0.52 |
| sit: 0.20 | food: 0.44 |

**C1**: a table with a plate of food and a glass of wine.
**C2**: a plate of food on a table with a glass of wine.
**C3**: a plate of food and a drink on a table.

| S1/2: | S3: |
|---|---|
| bed: 0.97 | bed: 1.00 |
| bedroom: 0.64 | room: 0.98 |
| room: 0.62 | pictures: 0.93 |
| book: 0.47 | bedroom: 0.85 |
| has: 0.42 | living: 0.84 |
| blanket: 0.35 | white: 0.78 |
| pillows: 0.29 | blue: 0.77 |
| pictures: 0.28 | blanket: 0.74 |
| pillow: 0.24 | colorful: 0.72 |
| messy: 0.24 | wall: 0.68 |
| wall: 0.23 | has: 0.52 |

**C1**: a bedroom with a bed and a book shelf.
**C2**: a bedroom with a bed and a book shelf.
**C3**: a bedroom with a bed and colorful pictures on the wall.

| S1/2: | S3: |
|---|---|
| toilet: 0.99 | bathroom: 1.0 |
| cat: 0.90 | toilet: 1.0 |
| bathroom: 0.63 | cat: 1.0 |
| white: 0.44 | white: 1.0 |
| small: 0.33 | small: 0.99 |
| sitting: 0.30 | seat: 0.98 |
| standing: 0.27 | drinking: 0.92 |
| brown: 0.21 | bowl: 0.86 |
| bath: 0.21 | water: 0.68 |
| floor: 0.16 | sticking: 0.58 |
| looking: 0.19 | inside: 0.53 |

**C1**: a cat sitting on a toilet in a bathroom.
**C2**: a cat sitting on a toilet in a bathroom.
**C3**: a cat drinking water from a toilet bowl.

| S1/2: | S3: |
|---|---|
| ocean: 0.90 | ocean: 1.0 |
| water: 0.70 | water: 0.99 |
| boat: 0.59 | waves: 0.99 |
| board: 0.33 | boat: 0.99 |
| person: 0.32 | white: 0.77 |
| wave: 0.29 | beach: 0.74 |
| waves: 0.29 | large: 0.62 |
| large: 0.27 | near: 0.61 |
| white: 0.25 | wave: 0.53 |
| surfboard: 0.22 | ship: 0.51 |
| surf: 0.19 | surfing: 0.49 |

**C1**: a boat on a body of water.
**C2**: a boat that is floating in the water.
**C3**: a boat in the water while a large ship in the background.

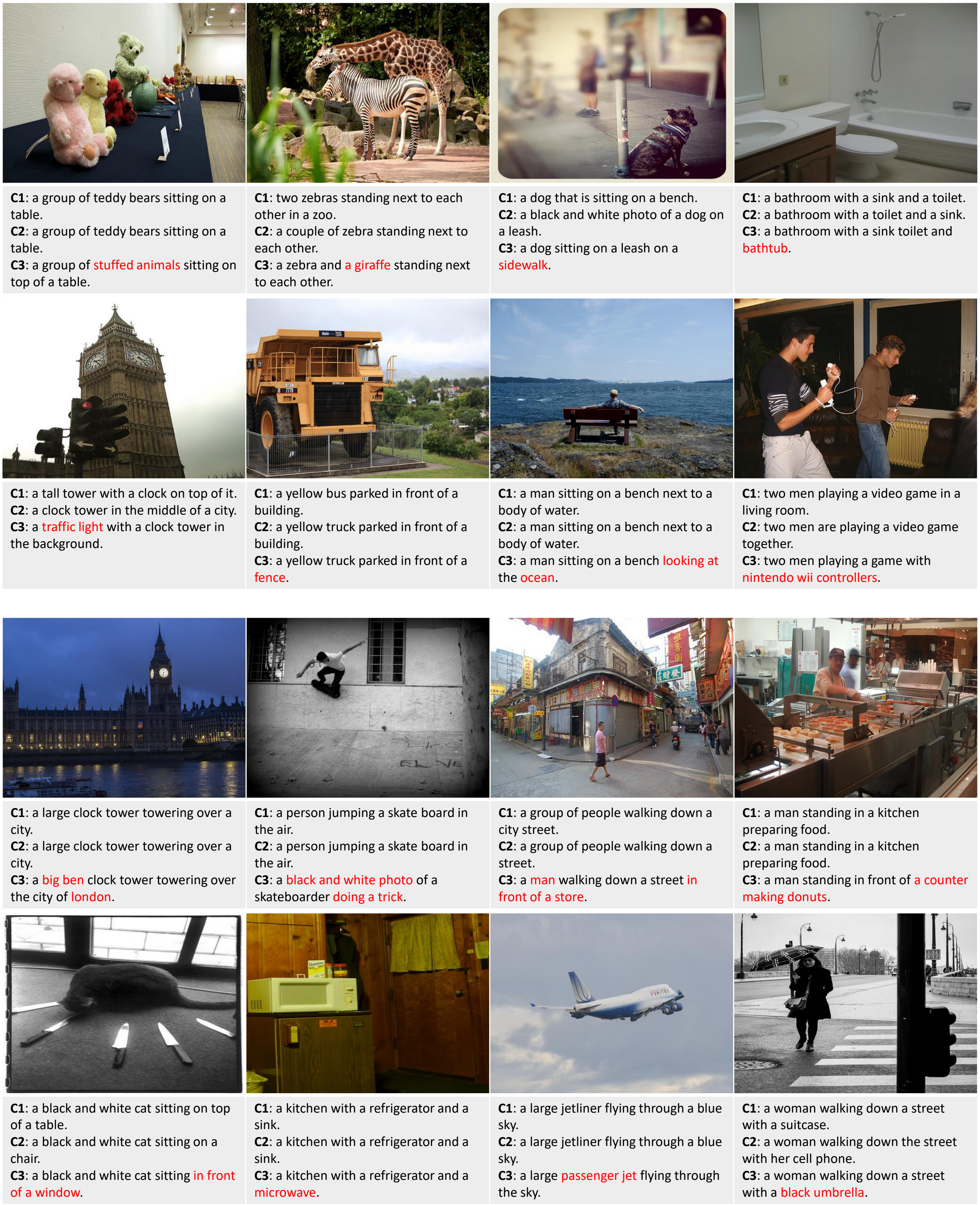
| S1/2: | S3: |
|---|---|
| pink: 0.70 | her: 1.0 |
| girl: 0.69 | girl: 1.0 |
| woman: 0.64 | holding: 1.0 |
| holding: 0.62 | woman: 1.0 |
| dress: 0.37 | young: 1.0 |
| hair: 0.34 | pink: 0.99 |
| beautiful: 0.32 | couch: 0.98 |
| couch: 0.31 | hand: 0.58 |
| phone: 0.26 | dress: 0.56 |
| hand: 0.25 | remote: 0.53 |
| cell: 0.23 | sitting: 0.51 |

**C1**: a woman is holding a cell phone in her hand.
**C2**: a woman in a pink dress talking on a cell phone.
**C3**: a young girl sitting on a couch holding a remote.

Fig. 6: Visualization of the detected concepts, image captions generated by different models. **S1/2** are the concepts and their prediction scores detected by the FCN-MIL baseline, **S3** are the concepts and their prediction scores detected by our OPR-MCM method. Note that words such as *"a", "the", "of", "with", "on"* are not displayed in **S1/2** and **S3**. **C1**, **C2** are the captions generated by V1 + WR and V3 + WR models respectively and both using the concepts from **S1/2**. **C3** is the caption generated by V3 + WR model using the concepts from **S3**. Some words are colored and best viewed in color.

*4) Overall Evaluation of Image Captioning:* Table II shows the image captioning performance of different models on the MSCOCO dataset. Following the conventions of the latest methods, we use Resnet101 [58] as our deep network for the concept detector. In order to compare fairly with the models based on VGG16 network, we also report our results using VGG16 network. We can see that the models utilizing semantic concepts generally perform better than the models using only image visual features. In particular, when using VGG16 network, our model achieves the best performance across

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2855415, IEEE Transactions on Image Processing



**C1**: a group of teddy bears sitting on a table.
**C2**: a group of teddy bears sitting on a table.
**C3**: a group of stuffed animals sitting on top of a table.

**C1**: two zebras standing next to each other in a zoo.
**C2**: a couple of zebra standing next to each other.
**C3**: a zebra and a giraffe standing next to each other.

**C1**: a dog that is sitting on a bench.
**C2**: a black and white photo of a dog on a leash.
**C3**: a dog sitting on a leash on a sidewalk.

**C1**: a bathroom with a sink and a toilet.
**C2**: a bathroom with a toilet and a sink.
**C3**: a bathroom with a sink toilet and bathtub.

**C1**: a tall tower with a clock on top of it.
**C2**: a clock tower in the middle of a city.
**C3**: a traffic light with a clock tower in the background.

**C1**: a yellow bus parked in front of a building.
**C2**: a yellow truck parked in front of a building.
**C3**: a yellow truck parked in front of a fence.

**C1**: a man sitting on a bench next to a body of water.
**C2**: a man sitting on a bench next to a body of water.
**C3**: a man sitting on a bench looking at the ocean.

**C1**: two men playing a video game in a living room.
**C2**: two men are playing a video game together.
**C3**: two men playing a game with nintendo wii controllers.

**C1**: a large clock tower towering over a city.
**C2**: a large clock tower towering over a city.
**C3**: a big ben clock tower towering over the city of london.

**C1**: a person jumping a skate board in the air.
**C2**: a person jumping a skate board in the air.
**C3**: a black and white photo of a skateboarder doing a trick.

**C1**: a group of people walking down a city street.
**C2**: a group of people walking down a street.
**C3**: a man walking down a street in front of a store.

**C1**: a man standing in a kitchen preparing food.
**C2**: a man standing in a kitchen preparing food.
**C3**: a man standing in front of a counter making donuts.

**C1**: a black and white cat sitting on top of a table.
**C2**: a black and white cat sitting on a chair.
**C3**: a black and white cat sitting in front of a window.

**C1**: a kitchen with a refrigerator and a sink.
**C2**: a kitchen with a refrigerator and a sink.
**C3**: a kitchen with a refrigerator and a microwave.

**C1**: a large jetliner flying through a blue sky.
**C2**: a large jetliner flying through a blue sky.
**C3**: a large passenger jet flying through the sky.

**C1**: a woman walking down a street with a suitcase.
**C2**: a woman walking down the street with her cell phone.
**C3**: a woman walking down a street with a black umbrella.

Fig. 7: More image captioning results of different methods. **C1**, **C2** are the captions generated by V1 + WR and V3 + WR models respectively and both using prediction score vector detected by the FCN-MIL baseline. **C3** is the caption generated by V3 + WR model using score vector detected by our OPR-MCM method. Some words are colored and best viewed in color.

all metrics. Note that ATT [6], Att-CNN+LSTM [37] and MSM [7] also use the semantic concepts for image captioning. We can see our method achieves significant improvements in comparison with those methods, which shows the effectiveness of our proposed method. Besides, compared with GroupTalk [43] and GAN [47] which aim to generate diverse captions to describe the image, our model shows much better performance as well. When using Resnet101 as our deep network, our method outperforms the latest models, such as SCA-CNN [55], SCN-LSTM [8], Skeleton-Attribute [56], Adaptive Attention [28]. Specifically, when compared with SCN-LSTM [8], which relies on the original prediction score vector of concepts to generate the caption, our method achieves much better performance. This demonstrates our method achieves much superior concept detection performance.

Furthermore, we evaluate our method on the MSCOCO online test server and compare it with other competitive approaches. In this experiment, we also use Resnet101 in our concept detector. Because our method uses the cross-entropy loss of caption words to train the caption generator, we do not compare the methods using reinforcement learning which directly optimize the quality metrics (e.g., CIDEr) to train the model. The comparison results are shown in Table III. From those results, we can see our method achieves better performance compared with other methods. Note that ATT [6], MSM [7] and SCN-LSTM [8] also make use of semantic concepts to generate image captions.

### D. *Qualitative Analysis*

Fig. 6 shows a few visualized examples of detected concepts and generated captions for different models. **S3** are the concepts and their prediction scores detected by our OPR-MCM method, **C3** is the caption generated by our V3 + WR model using the concepts from **S3**. From these visualized results, we can see that all of compared models can generate somewhat relevant captions, while our proposed method can detect more related concepts with a high accuracy. The captions generated by our model are more precise and contains more detailed contents. For example, in the first image (from left to right, top to bottom), our model detects "*woman*" and "*grassy*" with higher confidence compared with the baseline, so our model generates a caption "*a woman flying a kite in a grassy field*" which describes the image more concretely. In the second one, because our model can detect the word "*table*", the caption generated by ours not only describes "*a man standing next to a clock*" but also contains "*a clock on a table*". In the third one, our model can detect "*stone*" and "*wall*", so it generates the detailed description "*next to a stone wall*". In the fourth one, because "*drink*" is detected by our model, thus our model avoids generating the wrong caption "*a plate of food on a table with a glass of wine*". Other similar results can be found in the following examples.

In Fig. 7, we also show more image captioning results of different methods to better demonstrate the performance of our method. In order to show more examples, we only display the original images and the generated captions. From those results, we can see our method achieves more precise and detailed captioning performance. For instance, in the second image of the first row, our method can correctly recognize "*a zebra and a giraffe*" while other methods mis-recognize them as "*two zebras*" or "*a couple of zebra*". This shows our method can detect various concepts more precisely. In the first image of the third row, our method can describe the clock tower with more detail, namely the "*big ben clock tower*", and also describe the city more specifically with "*the city of london*". This demonstrates our method can detect more semantic concepts for better image captioning.

## V. Conclusion

In this paper, we argued the deficiency of sufficient semantics in the concepts-to-caption framework. We proposed to first feed more semantic concepts to caption generator. Then, we applied a concepts selection process to automatically choose the most suitable ones for better image captioning. Specifically, we proposed a OPR-MCM method to overcome the sample imbalance and missing concepts problems which cause the deficiency of sufficient semantics for image captioning. We also explored an element-wise selection process at the caption generation stage. We conducted extensive experiments on multiple benchmarks and the results show our method achieves superior results compared with many other methods. We visualized the detected concepts and generated captions of different methods, which also show our method can generate more precise and detailed captions. In the future, we intend to explicitly extract the relations between concepts to further improve the image captioning performance.

## VI. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Project 61572108, Project 61632007 and the 111 Project No. B17008.

## References

[1] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[2] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[3] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[5] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt *et al.*, "From captions to visual concepts and back," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482.

[6] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.

[7] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," *arXiv preprint arXiv:1611.01646*, 2016.

[8] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[9] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, 2006.

[10] A. de Carvalho and A. Freitas, "A tutorial on multi-label classification techniques," *Foundations of Computational Intelligence Volume 5*, pp. 177–195, 2009.

[11] M. Hu, Y. Yang, F. Shen, L. Zhang, H. T. Shen, and L. Xuelong, "Robust web image annotation via exploring multi-facet and structural knowledge," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4871–4884, 2017.

[12] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*, 1998, pp. 570–576.

[13] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482.

[14] W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[15] H. Zhu, R. Vial, S. Lu, X. Peng, H. Fu, Y. Tian, and X. Cao, "Yotube: Searching action proposal via recurrent and static regression networks," *IEEE Trans Image Process*, vol. 27, no. 6, pp. 2609–2622, Jun. 2018.

[16] W. Zhang, Q. Chen, W. Zhang, and X. He, "Long-range terrain perception using convolutional neural networks," *Neurocomputing*, vol. 275, pp. 781–787, 2018.

[17] J. T. Zhou, H. Zhao, X. Peng, M. Fang, Z. Qin, and R. S. M. Goh, "Transfer hashing: From shallow to deep," *IEEE Trans Neural Netw. Learn. Syst.*, pp. 1–11, 2018.

[18] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention based bidirectional lstm," *IEEE Transactions on Cybernetics*, 2018, doi:10.1109/TCYB.2018.2831447.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[20] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[21] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[22] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Multimodal neural language models." in *International Conference on Machine Learning*, vol. 14, 2014, pp. 595–603.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," pp. 2048–2057, 2015.

[25] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," *arXiv preprint arXiv:1612.01033*, 2016.

[26] M. Zhang, Y. Yang, H. Zhang, Y. Ji, N. Xie, and H. T. Shen, "Deep semantic indexing using convolutional localization network with region-based visual attention for image database," in *Australasian Database Conference*, 2017, pp. 261–272.

[27] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen, "Review networks for caption generation," in *neural information processing systems*, 2016, pp. 2361–2369.

[28] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[29] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation," *IEEE transactions on cybernetics*, vol. 45, no. 5, pp. 1083–1094, 2015.

[30] Y. Yang, F. Shen, Z. Huang, , H. T. Shen, and X. Li, "Discrete nonnegative spectral clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1834–1845, 2017.

[31] J. Song, H. T. Shen, J. Wang, Z. Huang, N. Sebe, and J. Wang, "A distance-computation-free search scheme for binary code databases," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 484–495, 2016.

[32] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Processing*, vol. 26, no. 5, pp. 2494–2507, 2017.

[33] F. Shen, X. Zhou, Y. Yang, J. Song, H. T. Shen, and D. Tao, "A fast optimization method for general binary code learning," *IEEE Trans. Image Processing*, vol. 25, no. 12, pp. 5610–5621, 2016.

[34] X. Zhu, S. Zhang, R. Hu, Y. Zhu *et al.*, "Local and global structure preservation for robust unsupervised spectral feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 3, pp. 517–529.

[35] X. Zhu, X. Li, S. Zhang, Z. Xu, L. Yu, and C. Wang, "Graph pca hashing for similarity search," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2033–2044, 2017.

[36] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient knn classification with different numbers of nearest neighbors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774–1785, 2018.

[37] Q. Wu, C. Shen, L. Liu, A. Dick, and A. V. Den Hengel, "What value do explicit high level concepts have in vision to language problems," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 203–212.

[38] Y. Bin, Y. Yang, J. Zhou, Z. Huang, and H. T. Shen, "Adaptively attending to visual attributes and linguistic knowledge for captioning," in *ACM on Multimedia Conference*, 2017, pp. 1345–1353.

[39] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[40] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4565–4574.

[41] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[42] Y. Yu, H. Ko, J. Choi, and G. Kim, "End-to-end concept word detection for video captioning, retrieval, and question answering," in *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[43] Z. Wang, F. Wu, W. Lu, J. Xiao, X. Li, Z. Zhang, and Y. Zhuang, "Diverse image captioning via grouptalk," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 2957–2964.

[44] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," *arXiv preprint arXiv:1610.02424*, 2016.

[45] S. Venugopalan, L. Anne Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[46] L. Wang, A. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," in *Advances in Neural Information Processing Systems*, 2017, pp. 5752–5762.

[47] B. Dai, D. Lin, R. Urtasun, and S. Fidler, "Towards diverse and natural image descriptions via a conditional gan," *arXiv preprint arXiv:1703.06029*, 2017.

[48] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," *arXiv preprint arXiv:1703.10476*, 2017.

[49] Y. Yang, J. Zhou, J. Ai, Y. Bin, A. Hanjalic, and H. T. Shen, "Video captioning by adversarial lstm," *IEEE Transactions on Image Processing*, 2018.

[50] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," *international conference on acoustics, speech, and signal processing*, pp. 6645–6649, 2013.

[51] Y. Grandvalet and Y. Bengio, "Entropy regularization," *Semi-Supervised Learning*, 2005.

[52] Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, and Jonathan, "Caffe: Convolutional architecture for fast feature embedding," *Eprint Arxiv*, pp. 675–678, 2014.

[53] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[54] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Image caption generation with text-conditional semantic attention," *arXiv preprint arXiv:1606.04621*, 2016.

[55] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

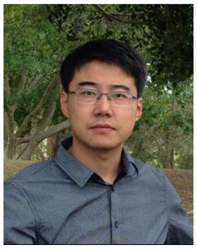
[56] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton key: Image captioning by skeleton-attribute decomposition," in *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2855415, IEEE Transactions on Image Processing

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 770–778.

**Mingxing Zhang** is currently a PhD student at University of Electronic Science and Technology of China. His research interests mainly focus on multimedia content analysis, computer vision and machine learning. He was a Research Assistant under the supervision of Prof. Tat-Seng Chua at National University of Singapore in 2016.



**Yang Yang** is currently with University of Electronic Science and Technology of China. He was a Research Fellow under the supervision of Prof. Tat-Seng Chua in National University of Singapore during 2012-2014. He was conferred his Ph.D. Degree (2012) from The University of Queensland, Australia. During the PhD study, Yang Yang was supervised by Prof. Heng Tao Shen and Prof. Xiaofang Zhou. He obtained Master Degree (2009) and Bachelor Degree (2006) from Peking University and Jilin University, respectively. His research interests include multimedia content analysis, computer vision and social media analytics.

**Hanwang Zhang** is currently an assistant professor in the School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU). In 2017, he was doing a post-doc job at DVMM of Columbia University, whose supervisor is Prof. Shih-Fu Chang. Before this job, he worked with Prof. Tat-Seng Chua, who is also his Ph.D supervisor (2009 - 2014) and Assoc. Prof. Shuicheng Yan (2013 - 2014) at National University of Singapore (NUS). Before this, he was an undergraduate student of Zhejiang University (2005 - 2009). In particular, he was a member of Chu Kochen Honors College.

**Yanli Ji** is currently an Associate Professor in the University of Electronic Science and Technology of China (UESTC). She obtained her Ph.D degree from Department of Advanced Information Technology, Kyushu University, Japan at Sep. 2012. Her research interests include Human Robot Interaction related topics, e.g. human activity recognition, emotion analysis, hand gesture recognition and facial expression recognition.

**Heng Tao Shen** is currently a Professor of National "Thousand Talents Plan", the Dean of School of Computer Science and Engineering, and the Director of Center for Future Media at the University of Electronic Science and Technology of China. He obtained his BSc with 1st class Honours and PhD from Department of Computer Science, National University of Singapore in 2000 and 2004 respectively. He then joined the University of Queensland as a Lecturer, Senior Lecturer, Reader, and became a Professor in late 2011. His research interests mainly include Multimedia Search, Computer Vision, Artificial Intelligence, and Big Data Management. Heng Tao has published 200+ papers, most of which appeared in prestigious publication venues of interests, such as ACM Multimedia, CVPR, AAAI, IJCAI, SIGMOD, VLDB, ICDE, TOIS, TIP, TPAMI, TKDE, VLDB Journal, etc. He has received 7 Best Paper Awards from international conferences, including the Best Paper Award from ACM Multimedia 2017 and Best Paper Award - Honorable Mention from ACM SIGIR 2017. He got the Chris Wallace Award for outstanding Research Contribution in 2010 conferred by Computing Research and Education Association, Australasia, and the Future Fellowship from Australia Research Council in 2012. He has served as a PC Co-Chair for ACM Multimedia 2015 and currently is an Associate Editor of IEEE Transactions on Knowledge and Data Engineering. He is an Honorary Professor at the University of Queensland, and holds a position of Visiting Professor at Nagoya University and National University of Singapore.

**Tat-Seng Chua** is currently the KITHCT Chair Professor with the School of Computing, National University of Singapore (NUS), Singapore. From 1998 to 2000, he was the Acting and Founding Dean of the School of Computing. He joined NUS in 1983, and spent three years as a Research Staff Member in the Institute of Systems Science (now I2R) in the 1980s. He worked on several multimillion-dollar projects interactive media search, local contextual search, and real-time live media search. His research interests include multimedia information retrieval, multimedia question, and the analysis and structuring of user-generated contents.

Dr. Chua has organized and served as a Program Committee Member of numerous international conferences in the areas of computer graphics, multimedia, and text processing. He was the Conference Co-Chair of the ACM Multimedia in 2005, the Conference on Image and Video Retrieval in 2005, and the ACM SIGIR in 2008, and the Technical PC Co-Chair of SIGIR in 2010. He serves on the editorial boards of the ACM Transactions of Information Systems (ACM), Foundation and Trends in Information Retrieval (Now), The Visual Computer (Springer Verlag), and Multimedia Tools and Applications (Kluwer). He is on the Steering Committees of the International Conference on Multimedia Retrieval, the Computer Graphics International, and the Multimedia Modeling Conference Series. He serves as a member of international review panels of two large-scale research projects in Europe. He is the Independent Director of two listed companies in Singapore.