

# Deep Hierarchical Encoder-Decoder Network for Image Captioning

Xinyu Xiao, Lingfeng Wang, Kun Ding, Shiming Xiang, and Chunhong Pan

**Abstract**—Encoder-decoder models have been widely used in image captioning, and most of them are designed via single Long Short Term Memory (LSTM). The capacity of single-layer network, whose encoder and decoder are integrated together, is limited for such a complex task of image captioning. Moreover, how to effectively increase the “vertical depth” of encoder-decoder remains to be solved. To deal with these problems, a novel Deep Hierarchical Encoder-Decoder Network (DHEDN) is proposed for image captioning, where a deep hierarchical structure is explored to separate the functions of encoder and decoder. This model is capable of efficiently exerting the representation capacity of deep networks to fuse high level semantics of vision and language in generating captions. Specifically, visual representations in top levels of abstraction are simultaneously considered, and each of these levels is associated to one LSTM. The bottom-most LSTM is applied as the encoder of textual inputs. The application of the middle layer in encoder-decoder is to enhance the decoding ability of top-most LSTM. Furthermore, depending on the introduction of semantic enhancement module of image feature and distribution combine module of text feature, variants of architectures of our model are constructed to explore the impacts and mutual interactions among the visual representation, textual representations and the output of the middle LSTM layer. Particularly, the framework is training under a reinforcement learning method to address the exposure bias problem between the training and the testing by the policy gradient optimization. Qualitative analyses indicate the process that our model “translates” image to sentence and further visualization presents the evolution of the hidden states from different hierarchical LSTMs over time. Extensive experiments demonstrate that our model outperforms current state-of-the-art models on three benchmark datasets: Flickr8K, Flickr30K and MSCOCO. On both image captioning and retrieval tasks, our method achieves the best results. On MSCOCO captioning Leaderboard, our method also achieves superior performance.

**Index Terms**—deep hierarchical structure, encoder-decoder, LSTM, image captioning, retrieval, vision-sentence.

## I. INTRODUCTION

AS one of the vision-language problems, image captioning is a challenging problem in computer vision and machine learning, which has attracted increasing attention of researchers [1]–[6]. The objective of image captioning is to generate a natural language description of a given image, and it essentially applies translation between two disparate modals of information. Compared with conventional computer vision

tasks, image captioning is more difficult as it requires not only capturing the information contained in an image, but also extracting the semantic correlation of the captured visual information to the relevant language expressions.

The encoder-decoder framework, which generally learns a transformation from image to sentence, has been proved to be a successful technique for image captioning. The encoder-decoder framework for image captioning is mainly inspired by the successful applying of Recurrent Neural Network (RNN) to sequence-to-sequence learning in machine translation [7]–[9]. Most of recent image captioning techniques inherit this thought, which adopt a Convolutional Neural Network (CNN) as the encoder to generate high-level abstract of image. A RNN, typically implemented with Long Short Term Memory (LSTM) [10], is employed to decode the generated image representation to a caption. Based on this structure, for better performance, most recent variants focus on adding extra inputs like region attention [6], [11] or attributes [12], [13] into the encoder-decoder. Region attention mechanism, derived from the intuition of visual attention, produces a spatial map highlighting image regions over time according to the textual context [14]. The use of attributes is to import the high-level image attribute vector as complementary knowledge to the encoder-decoder [15].

The aforementioned methods have achieved significant performances. However, the methods with complementary information still have several drawbacks. First, most variants of the classical encoder-decoder model pay attention to input reinforcement sources, and continue to use single LSTM layer. The outstanding performance of deep hierarchical CNN in object-detection [16], image-classification [17], object-recognition [18] and segmentation [19] demonstrates that deep and hierarchical structure of natural networks present more efficient learning and representation capacity than shallower models. For one LSTM, the inner of LSTM can be regarded as multi-stacked hidden layers unfolded in time, but the weights of the “horizontal depth” layers are shared over all time steps. The weights-sharing property limits the learning or feature representation ability of classical encoder-decoder based models. Similar to the latest machine translation frameworks [20]–[23], increasing the “vertical depth” of the encoder-decoder can be taken into account for improving the performance of image caption generation.

Second, for the majority of recent works, the high-level image attributes and visual attention representations are only used as extra knowledge with the image and sentence features into the LSTM units [12], [24]. The research results from Ting et al. [25] demonstrate that the mutual interactions between

Xinyu Xiao, Lingfeng Wang, Kun Ding, Shiming Xiang, and Chunhong Pan are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CAS), Beijing 100190, China (e-mail: {xinyu.xiao, lfwang, smxiang, chpan}@nlpr.ia.ac.cn; kding1225@gmail.com). Lingfeng Wang is the corresponding author.

Xinyu Xiao and Shiming Xiang are also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing, China.

different inputs can be exploited to refactor the architecture of encoder-decoder. Middle layers of stacked multiple LSTMs imply abundant cross-semantic information from the visual and textual representations. Inspired by this, we conduct investigations about the architectures of multiple LSTM layers by exploiting the mutual interactions between different LSTMs to enhance the capacity of caption generation.

The Maximum Likelihood Estimation (MLE) is the conventional training method in image captioning, which is used to learn parameters through minimizing the negative log-likelihood of the ground-truth at each time step. However, this method will lead to a problem called exposure bias. This problem has the result of a discrepancy between training and testing. Research [26] indicates that the reinforcement learning (RL) can provide a solution to solve this problem. The policy gradient method is proposed by [26] in the training procedure. Rennie et al. [27] adopted the idea but replaced the obtained reward by the he current model against the baseline of the inference algorithm.

In this paper, we propose a Deep Hierarchical Encoder-Decoder Network (DHEDN) which is constructed on three LSTM layers. Unlike single-layer model which couples the function of encoder and decoder in a single LSTM module, DHEDN separates the encoder and decoder layers. The middle LSTM which named Vision-Sentence Embedding LSTM (VSE-LSTM) can fuse visual and textual context semantics to enhance the decoding ability of top-most LSTM. More importantly, depending on the introduction of semantic enhancement module and distribution combine module in the top-most LSTM which named Semantic Fusion LSTM (SF-LSTM), three variants of encoder-decoder are devised to better investigate and demonstrate the crucial impact of the deep hierarchical architecture. Moreover, we apply the policy gradient method to directly optimize our model to further prove the validity of our method.

The main contributions of this paper are in three aspects:

- (1) A deep hierarchical framework is developed for image captioning, which separates the language model into multi-layer LSTMs, and receives the image representations with different CNN depths. This provides a mechanism that can increase the vertical depth of the encoder-decoder. With this treatment, the multi-level semantics of vision and language can be fused together for caption generation, and thus the representation capacity of the proposed network is remarkably enhanced.
- (2) A semantic enhancement module of image feature and a distribution combine module of text feature are introduced to the SF-LSTM. Depending on these two modules the decoder capacity of SF-LSTM can be improved.
- (3) Qualitative and visualized analyses of our DHEDN model are performed to understand how the hierarchical LSTMs militate over time. The extensive experiments on both image captioning and image-text cross-modal retrieval demonstrate the effectiveness of our model.

## II. RELATED WORK

Early works in image captioning pay attention to the template-based methods and retrieval-based approaches.

Template-based methods [28]–[33] specify templates for sentence generation which split sentence into fragments (subject, verb, object, etc.) and map them with target image content. For example, Yang et al. [30] employed the Hidden Markov Model (HMM) to filter the highest log-likelihood quadruplet which consists of objects, scenes, verbs, and prepositions to generate relevant image descriptions. Similar in Kulkarni et al. [28], based on the visual detection, Conditional Random Field (CRF) model is employed to predict corresponding detected objects, attributes, and prepositions which would be converted to generate sentence by predefined templates. Retrieval-based approaches [34]–[36] treat image captioning as retrieval task, which copy sentences from other similar images in the training set to the target image. These approaches cannot generate novel descriptions and robust infeasibility to unseen images. For instance, relying on neural network generated image feature, Devlin et al. [36] utilized a simple k-nearest neighbor retrieval model to select a consensus caption to image.

Recent public methods based on deep neural networks in image captioning achieve great success compared to early work. The differences among these various methods mainly lie in the language model.

Attention-based methods and semantic-based approaches are regarded as two main types of complementary knowledge to the language model. Attention-based mechanism [3], [6], [37], [38] has attracted wide interests recently, which learn to confirm where to focus in the image over time according to the text context. Xu et al. [6] proposed an image captioning model that combines visual attention with the hidden state of single LSTM layer. Integrating this spirit, Liu et al. [37] attempted to increase the correctness of visual attention to improve sentence generation performance. Except visual attention, Zhou et al. [39] proposed a textual attention model which defines a time-dependent textual attention in architecture. Lu et al. [40] proposed a “visual sentinel” which can adaptively look at the image to decide if input it to the language model when generate the next word. Semantic-based approaches [12], [25], [41] apply a semantic-concept-detection process to high-level image attributes before generating sentences. Wu et al. [12] and You et al. [41] employed the visual attributes as an extra input with the image and sentence features to the LSTM. Ting et al. [25] explored multi-way to composite the image representation and attributes to the single LSTM. Gan et al. proposed a Semantic Compositional Network (SCN), which attempt to effectively compose the semantic concepts for image captioning. Although these methods achieve state-of-the-art performance, the achievements depend heavily on the quality of the extra complementary knowledge.

Encoder-decoder based methods have received more and more attentions recently. Wang et al. [42] proposed a Bidirectional LSTMs based model which uses bidirectional Long Short Term Memory (Bi-LSTM) to encode the sentence to increase the depth of the encoder-decoder. Yang et al. [43] introduced a review network which extends existing (attentive) encoder-decoder models which performs multiple review steps with attention on the encoder hidden states to improve the performance of attention mechanism. Gan et al. [44] proposed a Semantic Compositional Network that exploits semantic

concepts encoded in an image and effectively composes the semantic concepts in the LSTM with the hidden states and the image CNN representation to improve the caption generation performance.

According to [26], the methods of Reinforcement learning (RL) have been introduced into image captioning, which calculate the reward by the Monte-Carlo sampling method and back propagation by the policy gradient technique in training. A self-critical training method was proposed by Rennie et al. [27] to use the policy gradient method with a normalized reward of metric from the current model to against the baseline under the test-time inference algorithm.

In short, our work belongs to the RL-based encoder-decoder methods, but we do not utilize extra complementary knowledge like attributes or attentions. Through the deep hierarchical structure and vision-sentence embedded semantic fusion, we construct a novel encoder-decoder for image captioning.

### III. METHOD

In this section, we describe the Deep Hierarchical Encoder-Decoder Network (DHEDN) used for image caption generation. First, the widely used LSTM in image captioning field is briefly introduced in Section III-A, then the generic encoder-decoder framework for image captioning is introduced in Section III-B. We elaborate the details of the proposed deep hierarchical mechanism based image captioning model and explore the variants of our deep architecture in Section III-C. At last, the learning methods of our model are summarized in Section III-D.

#### A. Long Short Term Memory

The LSTM cell (see Fig. 1) was proposed in [10] and has been widely used in sequence generation [45] and machine translation [8], [9]. The memory cell  $c$  indicated by red solid circle is the core of LSTM, which can make LSTM control information by a group of sigmoid and hyperbolic tangent gates at every time step. When an LSTM receives multi-source inputs: the current input  $x_t$ , the previous step memory cell state  $c_{t-1}$  and the previous step hidden state  $h_{t-1}$  at the given time step  $t$ , and the group of gates are being modulated to update the inputs as follows:

$$\begin{aligned} i_t &= \sigma(\mathbf{W}_{xi}x_t + \mathbf{W}_{hi}h_{t-1} + b_i), \\ f_t &= \sigma(\mathbf{W}_{xf}x_t + \mathbf{W}_{hf}h_{t-1} + b_f), \\ o_t &= \sigma(\mathbf{W}_{xo}x_t + \mathbf{W}_{ho}h_{t-1} + b_o), \\ g_t &= \phi(\mathbf{W}_{xc}x_t + \mathbf{W}_{hc}h_{t-1} + b_c), \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\ h_t &= o_t \odot \phi(c_t), \end{aligned} \quad (1)$$

where  $\mathbf{W}_*$  denotes the weight matrix and  $b_*$  is the bias vector;  $\odot$  means the elementwise product;  $\sigma$  is the sigmoid nonlinearity activation function which is defined as  $\sigma(x) = (1 + \exp(-x))^{-1}$  and  $\phi$  denotes hyperbolic tangent nonlinearity function  $\phi(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$ ;  $c_t, h_t \in \mathbb{R}^N$  denote the memory cell state and hidden state with  $N$  dimension; the LSTM unit sets  $i_t, f_t, o_t, g_t \in \mathbb{R}^N$  are the input gate, forget

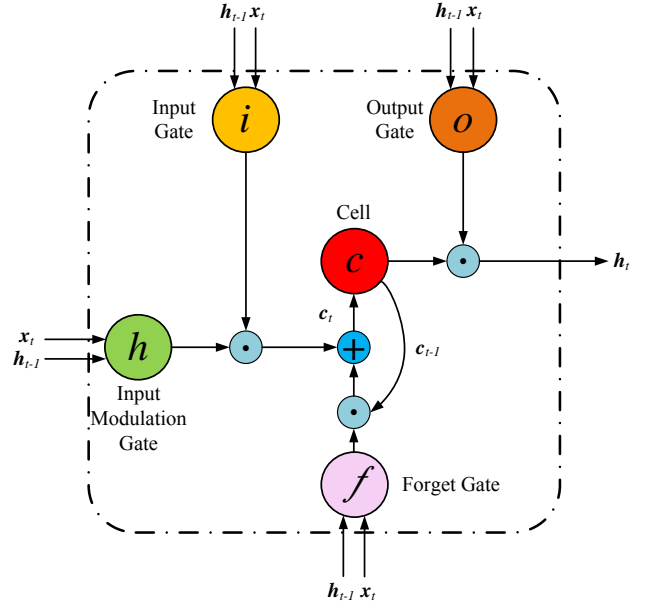


Fig. 1: Long Short Term Memory (LSTM) cell is controlled by three gates: the input gate  $i$ , forget gate  $f$  and output gate  $o$ . The inputs contain the current input  $x_t$  and the previous step hidden state  $h_{t-1}$ . The function of input gate  $i$  is to determine the input or output of the incoming signal. The previous state of cell is decided by the forget gate  $f$ . The output gate  $o$  decides to go through or prevent the current state of cell.

gate, output gate and input modulation gate, respectively. The input gate  $i_t$  and forget gate  $f_t$  have this capacity that makes the LSTM learn to selectively measure the current input and forget the previous memory. As a switch, the output gate  $o_t$  can learn the amount of memory cell and transfer it to the current hidden state.

#### B. Encoder-Decoder Framework

The encoding-decoding process, which generates a natural language description for an image, is to learn a transformation from image to text sequence. In general, the encoder-decoder is composed of an encoder which encodes image or sentence to fixed-dimensional vectors and a decoder that decodes the context vectors to the desired sentence.

**Encoder.** In conventional models [2], [6], there are two types of encoders, i.e., CNN encoder and RNN encoder. For an input image  $I$ , the deep CNN encoder extracts the last fully connected (FC) layer as a global image feature  $v = FC(I)$ .

For a sentence containing  $T$  words  $S = \{s_1, \dots, s_T\}$ , where  $s_t$  denotes the one-hot vector of the  $t$ -th word, the words embedded by a word embedding matrix  $\mathbf{W}_e \in \mathbb{R}^{N \times V}$  before sending the sentence to an RNN, where  $V$  denotes the vocabulary size. At time step  $t$ , the RNN encoder updates its hidden state by  $h_t = f(\mathbf{W}_e s_t, h_{t-1})$ , where  $f$  represents the RNN cell. RNN has been proven difficult to train for long-term sequences, partly because of gradients vanishing and exploding [10] which might result from propagating the gradients back through many recurrent layers. Since LSTM

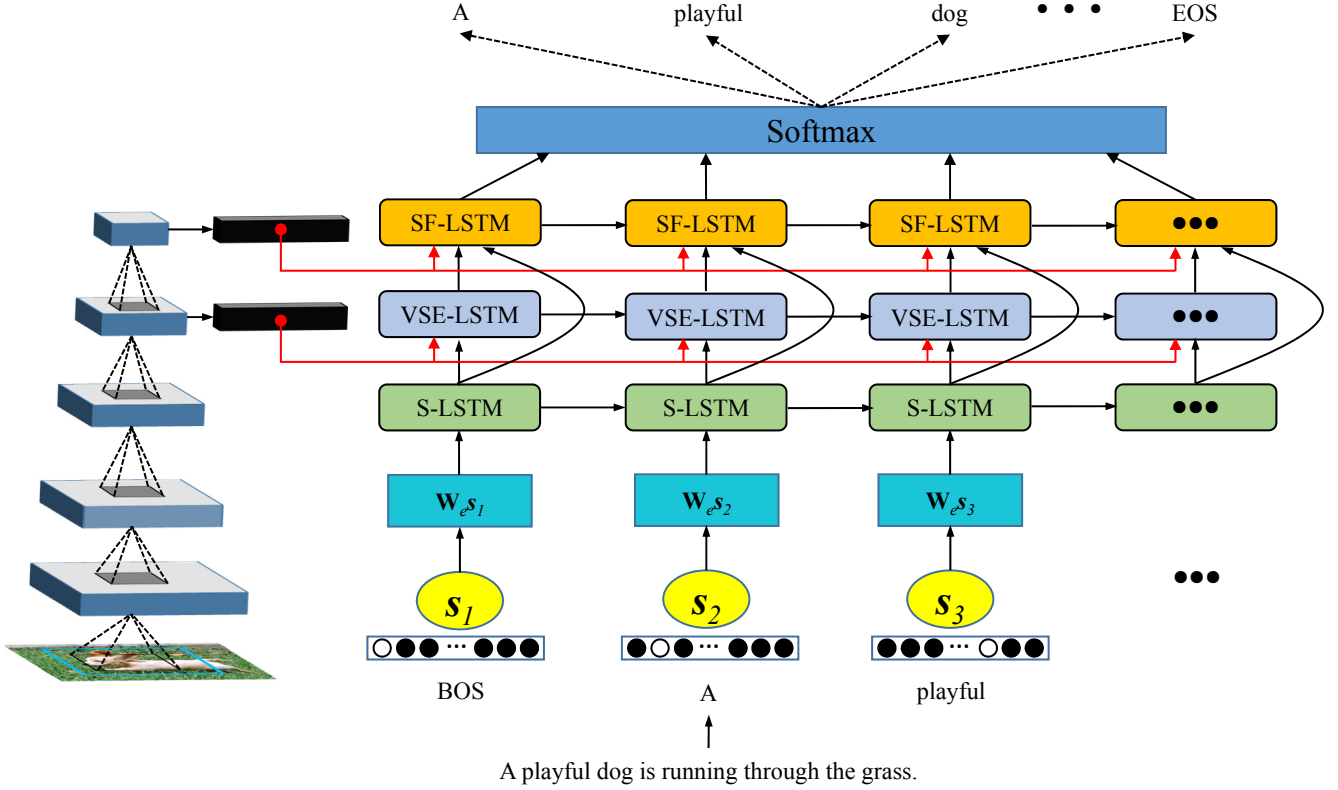


Fig. 2: Model architecture of Deep Hierarchical Encoder-Decoder Network (DHEDN). S-LSTM: Sentence-LSTM encoder. VSE-LSTM: Vision-Sentence Embedding layer. SF-LSTM: Semantic Fusion LSTM decoder.  $s_t$  indicates the one-hot vector of the  $t$ -th time word in a sentence.  $\mathbf{W}_e$  is word embedding matrix. BOS and EOS denote the start-of-sentence and end-of-sentence tokens, respectively. See Section III-C for more details.

cell incorporates memory units which endow network the power to judge if it needs to forget previous time step hidden state or to update hidden state when given new information. The implementation with LSTM to realize the function of RNN in encoder-decoder which can deal with the problem of vanishing and exploding gradients. Likewise, the LSTM encoder in the encoding process of source sentence at the  $t$ -th time step can be denoted as  $\mathbf{h}_t = LSTM(\mathbf{W}_e \mathbf{s}_t, \mathbf{h}_{t-1})$ , where  $LSTM$  implies the LSTM cell, which can be seen in Eqn. (1). The initialization of the memory cell state  $\mathbf{c}_0$  and hidden state  $\mathbf{h}_0$  to the LSTM unit is zero.

**Decoder.** The function of the image captioning decoder is to transform the encoded information, including image feature and source sentence vectors, to the target sentence. In generic architectures, the decoder of image captioning model embedded with the sentence sequence encoder to one shared LSTM layer in general. In other words, adding the image encoded feature to the LSTM sentence encoder in some way, and then outputting the target sentence in sequence could be the typical decoder mode in common. Accordingly, the LSTM updating procedure is as  $\mathbf{h}_t = LSTM(\mathbf{x}_t, \mathbf{h}_{t-1})$ , where  $\mathbf{x}_t$  is the current input.

Then, the hidden state  $\mathbf{h}_t$  is transformed to the distribute space of vocabulary by a fully connected layer  $\mathbf{d}_t = \mathbf{W}_d \mathbf{h}_t + \mathbf{b}_d$ , where  $\mathbf{W}_d$  and  $\mathbf{b}_d$  is the weight matrix and bias vector,

respectively. The distribution representation  $\mathbf{d}_t$  over the  $t$ -th time step to predicted word  $w_t$  by the softmax function:

$$w_t \sim softmax(\mathbf{d}_t). \quad (2)$$

Traditionally the encoder-decoder models for image captioning are learned through maximizing the likelihood of the target sentence in the training process. Given the ground truth sentence  $w_1^*, \dots, w_T^*$ , the objective loss function is defined as:

$$L = - \sum_{t=1}^T \log(p(w_t^* | w_{1:t-1}^*)), \quad (3)$$

where  $p(w_t | w_{1:t-1})$  is generated by the model in Eqn. (2).

### C. Deep Hierarchical Encoder-Decoder Network

As described in the previous Section III-B, the generic frameworks for image captioning normally embeds the textual encoder and decoder in one LSTM layer. Unlike previous works, we separate the encoder and decoder to different LSTMs, so that we can construct a deep hierarchical encoder-decoder to fuse the visual and textual semantics before decoding. The crucial point of our framework is to explore the effective connection patterns between different LSTMs, which can exert a beneficial impact on the feature transformation,



The vision-sentence embedded vector set  $\mathcal{H}^v$  is extracted from the joint space of vision and text. To time step  $t$ , the vision-sentence embedded vector  $\mathbf{h}_t^v$  owns the image feature  $\mathbf{v}_s$  and word representation  $\mathbf{h}_t^s$  semantics, simultaneously.

Visual and textual semantics are different. Text is transformed into a continuous distribution space through an embedding layer. It ensures that each word has a continuous, distributed representation in vocabulary. The vision-sentence embedded vector set  $\mathcal{H}^v$  and S-LSTM word encoded vector set  $\mathcal{H}^s$  combined to a common distribution space can make the decoding process more robust and efficient. The CNN visual feature removes the nonsignificant content and reserves the semantic of significant objects which is close to related visual concepts. The vision-sentence embedded vector  $\mathbf{h}_t^v$  combined with global image feature appropriately, can contribute to visual semantic augmentation at the  $t$ -th time's decoding process for relevant word  $\mathbf{w}_t$  in target sentence.

As shown in Fig. 3, for SF-LSTM, we define two modules  $S^{sf}$  and  $U^{sf}$ , where  $S^{sf}$  is the semantic augmentation module of the deep image feature  $\mathbf{v}^d$  and vision-sentence embedded vector set  $\mathcal{H}^v$ ;  $U^{sf}$  is the distribution combine module of the encoded textual sequence  $\mathcal{H}^s$  and  $\mathcal{H}^v$ . Inspired by the characters of the inputs and their interactions, we devise three variants of SF-LSTM to fuse the semantics of inputs.

**SF-LSTM<sub>1</sub>.** In SF-LSTM<sub>1</sub>, only the textual semantic property of vision-sentence embedded vector need to be utilized, and the design of architecture based on treating the vision-sentence embedded vector set  $\mathcal{H}^v$  and S-LSTM word encoded vector set  $\mathcal{H}^s$  as equivalent atoms in textual distribution space. Specifically, at each time step, the vision-sentence embedded vector  $\mathbf{h}_t^v$  is concatenated with the word encoded vector  $\mathbf{h}_t^s$  as  $\mathbf{u}_t^{sf}$  before fed into LSTM unit. Then the deep image representation  $\mathbf{v}_d$  and the textual concatenated vector  $\mathbf{u}_t^{sf} \in \mathbb{R}^{2N}$  are transformed into LSTM as enhanced semantics at time step  $t$ , respectively. For  $*$  =  $i, f, o, g$ , the SF-LSTM<sub>1</sub> computes the procedure as follows:

$$\begin{aligned} \mathbf{u}_t^{sf} &= U^{sf}(\mathbf{h}_t^v, \mathbf{h}_t^s), \\ \mathbf{s}_t^{sf} &= \mathbf{v}_d, \\ \mathbf{x}_{*,t}^{sf} &= \mathbf{W}_{*,s}^{sf} \mathbf{s}_t^{sf} + \mathbf{W}_{*,u}^{sf} \mathbf{u}_t^{sf}, \\ \mathbf{h}_t^{sf} &= LSTM^{sf}(\mathbf{x}_t^{sf}, \mathbf{h}_{t-1}^{sf}; \Theta^{sf}), t \in \{1, \dots, T\}, \end{aligned} \quad (8)$$

where  $U^{sf}$  is a concatenate module;  $\mathbf{s}_t^{sf} \in \mathbb{R}^N$  is the no-operation semantically enhancement vector which equals to the deep CNN image feature  $\mathbf{v}_d$ ;  $\mathbf{W}_{*,s}^{sf}$  and  $\mathbf{W}_{*,u}^{sf}$  are the transformation matrixes for the textual concatenated vector and semantically enhancement vector, respectively; and  $LSTM^{sf}$  is denoted as the SF-LSTM unit;  $\mathbf{x}_t^{sf}$  and  $\mathbf{h}_t^{sf}$  are the current input and hidden state output of SF-LSTM at time step  $t$ ;  $\Theta^{sf}$  is the parameters of the SF-LSTM units.

**SF-LSTM<sub>2</sub>.** Different from the former SF-LSTM<sub>1</sub>, The SF-LSTM<sub>2</sub> applies both visual and textual semantic properties of vision-sentence embedded vector set  $\mathcal{H}^v$  to benefit the fusion process of alien modals. Similarly, the vision-sentence embedded vector  $\mathbf{h}_t^v$  and word encoded vector  $\mathbf{h}_t^s$  are concatenated together at time  $t$ . In order to extract and strengthen the significant objects semantic of the deep image feature  $\mathbf{v}_d$ , which attends to the according time step target word,

the vision-sentence embedded vector  $\mathbf{h}_t^v$  is added to the deep image feature  $\mathbf{v}_d$  as  $\mathbf{s}_t^{sf}$  at time step  $t$ . And in decoding stage, the concatenated vector  $\mathbf{u}_t^{sf}$  and the mixed visual feature  $\mathbf{s}_t^{sf}$  are transformed to LSTM unit at each time step. Accordingly, the LSTM updating procedure in SF-LSTM<sub>2</sub> is:

$$\begin{aligned} \mathbf{u}_t^{sf} &= U^{sf}(\mathbf{h}_t^v, \mathbf{h}_t^s), \\ \mathbf{s}_t^{sf} &= S^{sf}(\mathbf{v}_d, \mathbf{h}_t^v), \\ \mathbf{x}_{*,t}^{sf} &= \mathbf{W}_{*,s}^{sf} \mathbf{s}_t^{sf} + \mathbf{W}_{*,u}^{sf} \mathbf{u}_t^{sf}, \\ \mathbf{h}_t^{sf} &= LSTM^{sf}(\mathbf{x}_t^{sf}, \mathbf{h}_{t-1}^{sf}; \Theta^{sf}), t \in \{1, \dots, T\}, \end{aligned} \quad (9)$$

where  $S^{sf}$  is a sum module for the deep CNN image feature  $\mathbf{v}_d$  and vision-sentence embedded vector  $\mathbf{h}_t^v$  at time step  $t$ .

**SF-LSTM<sub>3</sub>.** The last design SF-LSTM<sub>3</sub> is similar to SF-LSTM<sub>2</sub> except the semantically enhancement module  $S^{sf}$ . The vision-sentence embedded vector  $\mathbf{h}_t^v$  through a sigmoid layer at time  $t$  and then dot product with the deep image feature  $\mathbf{v}_d$  as the  $\mathbf{s}_t^{sf}$ . Hence, the LSTM updating procedure in SF-LSTM<sub>3</sub> is designed as:

$$\begin{aligned} \mathbf{u}_t^{sf} &= U^{sf}(\mathbf{h}_t^v, \mathbf{h}_t^s), \\ \mathbf{s}_t^{sf} &= S_p^{sf}(\mathbf{v}_d, \sigma^{sf}(\mathbf{h}_t^v)), \\ \mathbf{x}_{*,t}^{sf} &= \mathbf{W}_{*,s}^{sf} \mathbf{s}_t^{sf} + \mathbf{W}_{*,u}^{sf} \mathbf{u}_t^{sf}, \\ \mathbf{h}_t^{sf} &= LSTM^{sf}(\mathbf{x}_t^{sf}, \mathbf{h}_{t-1}^{sf}; \Theta^{sf}), t \in \{1, \dots, T\}, \end{aligned} \quad (10)$$

where  $\sigma^{sf}$  is a sigmoid nonlinearity activation layer and  $S_p^{sf}$  is a dot product module for the deep image feature  $\mathbf{v}_d$  and vision-sentence embedded vector  $\mathbf{h}_t^v$  at time step  $t$ .

4) *Composition of Deep Architecture:* Contrasts to previous image captioning frameworks [2], [46], the architecture of our model is deeper, which is inspired by the stacked LSTM layers architecture in machine translation models [21], [22]. But simply stacking some LSTM layers in image captioning model could make the network hard to train, which is likely because of gradients exploding and vanishing problem [47]. Motivated by the differentiated semantic properties of vision and text, the visual feature concentrates on the latent significant semantic objects but the textual feature project to a continuous distribution semantic space. Different to the design of stacked LSTM layers with residual connections in Google's neural machine translation system [20], the outputs of intermediate LSTM layers can be regarded as vision-sentence embedded vectors which can contribute to enhancing visual object concepts and steadying the textual distribution space.

Summarizing above conclusions, the four modules of our method are composited to a deeper structure, and devised to three variants compound modes. Throughout the entire network structure, the VSE-LSTM layer is utilized to embedding the shallow visual feature  $\mathbf{v}_s$  and S-LSTM word encoded vector set  $\mathcal{H}^s$  into a vision-sentence embedded space. The SF-LSTM decoder is through two processing module  $S^{sf}$  and  $U^{sf}$ . The  $S^{sf}$  processes semantic enhancement of the deep image feature  $\mathbf{v}_d$ , and the  $U^{sf}$  makes a distribution merge with the S-LSTM encoded word sequence  $\mathcal{H}^s$  by vision-sentence embedded set  $\mathcal{H}^v$ . Based on these, the efficiency of SF-LSTM units are lifting to generate the caption.



#### D. Model Learning

1) *Maximum Likelihood Estimation:* Given the image  $\mathbf{I}$  and associated sentence  $\mathcal{W} = \{w_1, \dots, w_T\}$ , the word conditional probability on preceding words can be written as:

$$p(w_t | w_1, \dots, w_{t-1}, \mathbf{v}_d, \mathbf{v}_s) = \frac{\exp[p(w_1, \dots, w_{t-1}, w_t, \mathbf{v}_d, \mathbf{v}_s)]}{\sum_{\bar{w} \in \mathcal{V}} \exp[p(\bar{w}, w_1, \dots, w_{t-1}, \mathbf{v}_d, \mathbf{v}_s)]}, \quad (11)$$

where  $\mathbf{v}_s, \mathbf{v}_d$  are defined in Eqn. (4) and Eqn. (5), respectively;  $\mathcal{V}$  denotes the vocabulary of the full training examples. In convention, the learning method of image captioning models is always applied the Maximum Likelihood Estimation (MLE). Its learning objective is to learn parameters through minimizing the negative log-likelihood of the target sentence with the ground-truth. Specifically, suppose there are  $K$  training examples, the loss function  $L$  of our model is defined as:

$$\begin{aligned} L &= -\frac{1}{K} \sum_{k=1}^K \log(p(\mathcal{W}^k | \mathbf{v}_d^k, \mathbf{v}_s^k, \Theta)) \\ &= -\frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \log(p(w_t^k | w_1^k, \dots, w_{t-1}^k, \mathbf{v}_d^k, \mathbf{v}_s^k, \Theta)), \end{aligned} \quad (12)$$

where  $\Theta$  is the total parameters of the whole network. But because of the limitations of the MLE in the gradient vanishing and overfitting, the discrepancy is existing between training and testing. To solve this problem, the policy network which applies the evaluation metrics as the optimizing objects is taken into consideration.

2) *Policy Gradient Optimization:* Similar to the other reinforcement learning techniques [48], [49], the policy gradient (PG) technique is interacting with the “environment” by an “agent” (e.g. LSTM). The “agent” following the policy  $p_\theta$  of the parameters  $\theta$  of the network to select an “action”, to predict the next word. The reward which is set as one of the evaluation metrics is observed by the “agent”. Following the implementation in [27], the objective in learning is to minimize the negative expected rewards of the complete sampled sentence  $\mathcal{W}^s = \{w_1^s, \dots, w_T^s\}$ :

$$L_\theta = -\frac{1}{K} \sum_{i=1}^K E_{\mathcal{W}^s \sim p_\theta} [r(\mathcal{W}^s)], \quad (13)$$

where the  $r(\mathcal{W}^s)$  is calculated by comparing sampled caption with the reference caption in the specified evaluation metric. We calculate the expected gradient by applying a single Monte-Carlo sample as:

$$\nabla_\theta L_\theta \approx -\frac{1}{K} \sum_{i=1}^K \Delta r(\mathcal{W}^s) \nabla_\theta \log[p_\theta(\mathcal{W}^s)], \quad (14)$$

where  $\Delta r(\mathcal{W}^s)$  is the relative reward, which is computed by relating to a baseline reward  $b$ ,  $b$  is obtained by performing greedy decoding:

$$\begin{aligned} b &= r(\hat{\mathcal{W}}), \quad \hat{\mathcal{W}} = \arg \max p(w_t | h_t^d), \\ \nabla_\theta L_\theta &\approx -\frac{1}{K} \sum_{i=1}^K (r(\mathcal{W}^s) - r(\hat{\mathcal{W}})) \nabla_\theta \log[p_\theta(\mathcal{W}^s)]. \end{aligned} \quad (15)$$

Through the choice the baseline has no impact to the expected gradient but can drastically reduce its variance.

#### IV. EXPERIMENTS

To validate the effectiveness of our model, extensive experiments have been conducted under the guidance of the following objectives:

- (1) Visualizing and analyzing how the proposed DHEDN learns to generate captions.
- (2) Qualitatively measuring the performances and benefits of our model and its variant forms, and then comparing with state-of-the-art methods of image captioning and retrieval task on three benchmark datasets.
- (3) Further mining the performance of our model through the presentation of retrieval task and visual captioning results.

##### A. Datasets

Our experiments are conducted on three datasets: Flickr8K [52], Flickr30K [53] and MSCOCO [54].

**Flickr8K** consists of 8,091 images from the flickr website<sup>1</sup>. Each image is given five captions. We follow the publicly splits<sup>2</sup> which divides 6,000, 1,000 and 1,000 images for training, validation and testing, respectively.

**Flickr30K** is an expansion of Flickr8K. It consists of 31,783 images and each of them is paired with five captions. According to the publicly splits<sup>2</sup>, we divide 29,014, 1,000 and 1,000 images for training, validation and testing, respectively.

**MSCOCO** is the largest dataset of image captioning at present, which consists of 82,783 images for training and the other 40,504 images for validation. Each image is annotated with at least five captions. In addition, this dataset contains 40,775 images for online testing. Because of its larger size and more complex scene images contained, MSCOCO is more challenging. In offline evaluation, we follow the prior work data split as [24], sampling 5,000 images for validation and 5,000 images for testing from the validation set, then extending the remaining 30,504 validation images to the training set for training. In online evaluation, we apply the trained models in the offline to the online test to against the state-of-the-art.

##### B. Evaluation Metrics

We evaluate our model by MSCOCO caption evaluation tool<sup>3</sup> on two tasks: caption generation and image-caption retrieval. To caption generation, the server can report the following metrics: BLEU-N (N=1,2,3,4) [55], Meteor [56], Rouge-L [57], CIDEr [58]. To image-caption retrieval, we follow previous work [59] to use Med r and Recall@K

<sup>1</sup><https://www.flickr.com/>

<sup>2</sup><http://cs.stanford.edu/people/karpathy/deepimagesent/>

<sup>3</sup><https://github.com/tylin/coco-caption>

TABLE I: Performance of our proposed model and the state-of-the-art methods on the Flickr8K and Flickr30K datasets, where B-N refers to BLEU-N score. The superscript “V” means the VggNet and “G” represents the GoogleNet, (-) indicates an unknown metric. The best results are highlighted and the second best results are underlined.

Model	Flickr8k					Flickr30k				
	B-1	B-2	B-3	B-4	METEOR	B-1	B-2	B-3	B-4	METEOR
Google NIC <sup>G</sup> [2]	63	41	27.2	—	—	66.3	42.3	27.7	18.3	—
m-RNN <sup>V</sup> [1]	—	—	—	—	—	60	41	28	19	—
Log Bilinear <sup>V</sup> [50]	<u>65.6</u>	42.4	27.7	17.7	17.3	60.0	38	25.4	17.1	16.9
LRCN <sup>V</sup> [46]	—	—	—	—	—	58.8	39.1	25.1	16.5	—
Soft-Attention <sup>V</sup> [6]	<b>67</b>	44.8	29.9	19.5	18.9	<u>66.7</u>	43.4	28.8	19.1	18.5
Hard-Attention <sup>V</sup> [6]	<b>67</b>	45.7	31.4	21.3	<u>20.3</u>	<b>66.9</b>	43.9	29.6	19.9	18.5
ATT <sup>V</sup> [41]	—	—	—	—	—	64.7	46.0	32.4	23.0	18.9
Bi-LSTM <sup>V</sup> [42]	65.5	<u>46.8</u>	32.0	21.5	—	62.1	42.6	28.1	19.3	—
phi-LSTM <sup>V</sup> [51]	63.6	<u>43.6</u>	27.6	16.6	—	66.6	45.8	28.2	17.0	—
Vgg16-LSTM	62.3	43.6	29.2	19.1	19.7	63.7	45.0	31.1	21.4	18.7
Vgg16+3-LSTM	62.4	44.2	29.8	19.6	19.3	63.8	44.8	31.1	21.5	17.9
Vgg16-LSTM-H	62.3	44.1	29.8	19.9	19.2	64.7	45.5	31.1	21.0	18.5
Vgg16-H	63.1	44.7	30.8	20.9	20.1	64.1	45.8	32.2	22.3	18.8
DHEDN <sup>V</sup>	65.1	<b>47.0</b>	<b>32.6</b>	<b>22.0</b>	20.1	65.3	<b>46.8</b>	<b>32.9</b>	<u>22.9</u>	19.0
DHEDN <sub>2</sub> <sup>V</sup>	64.3	46.2	32.0	21.5	20.0	65.4	46.7	<u>32.5</u>	22.4	<b>19.3</b>
DHEDN <sub>3</sub> <sup>V</sup>	64.8	46.7	<u>32.3</u>	<u>21.8</u>	<b>20.5</b>	65.3	<u>46.7</u>	<b>32.9</b>	<b>23.1</b>	<u>19.2</u>

(K=1,5,10) as the evaluation metrics. Med r is the median rank of the first retrieved ground truth image or caption and Recall@K is the rate R which a right image or caption is retrieved within the top K candidates.

### C. Implementation Details

**CNN Encoder.** We use two kinds of CNN image encoders: 16-layer VggNet [60] with 13 convolution layers and 3 fully connected layers and Inception-Resnet-v2 model [61]. The utilized VggNet model is pre-trained on the 1.2M image ILSVRC-2012 [62] classification training subset of the ImageNet [63] in our experiments. Specifically, to obtain different depth CNN image features, a new group of convolution layers *conv6* is added behind the fifth group of convolutional layers *conv5*, the structure of these convolution layers is the same with other convolution layers of the VggNet. The output of *conv6* is processed by a max-pooling layer and then connect to another 3 same type of fully connected layers as VggNet. We define this hierarchical CNN as Vgg16+3, in which 3 indicates to the number of additional convolutional layers, and the original *fc8* output and deeper *fc8* output are denoted as VggNet shallow and deep representations, respectively. About the deep CNN, we adopt the Inception-Resnet-v2 model as the image encoder, which is initialized by the public source<sup>4</sup>. Specifically, to get the shallow and deep CNN representations, a new *conv6\_1*×1+*fc* is added to the back of inception\_resnet\_v2\_c5 as the shallow feature and the preliminary output is regarded as the deep representation. The dimension of all the last *fc* outputs is  $N = 1,000$ .

**Captioning Vocabulary.** Similar to previous works [2], [42], each word in the sentences is represented with one-hot vector  $s_t$  which is embedded by using word embedding matrix  $W_e \in \mathbb{R}^{N \times V}$ , where  $V$  is the vocabulary dimension which depends on the training sentences of dataset and differs with the size of different datasets. We follow the public tokenization [3] and reserve the words that appear at least five times in the

training caption set, then we get 2,546, 7,448, 8,801 words for Flickr8K, Flickr30K and MSCOCO datasets, respectively.

**Training and Testing.** In the training and testing process, the hidden state dimension of all types of LSTM units is 1,000.

In training, each word is embedded by using word embedding matrix  $W_e$ , and the parameters of the language model are randomly initialized. To VggNet based model, the size of each image is reshaped to  $256 \times 256$  and the crop size of VggNet is  $224 \times 224$ . The initial learning rate  $\eta$  for CNN model is  $1e-3$  and for language model is  $1e-2$ . The gamma  $\gamma = 0.5$ , weight decay  $\lambda$  is 0 and the momentum we used is 0.9. For each dataset we train for 17 ~ 30 epochs with early stop. To Inception-Resnet-v2 based model, each image is resized to  $328 \times 328$  and the crop size is  $299 \times 299$ . We first train this model with MLE. After that, the RL method is applied to optimize the just MLE trained model with the CIDEr metric. At each epoch, the validation set is used to evaluate the training model, and the best CIDEr score achieved model is selected for the final testing. The online testing results are obtained from the Inception-Resnet-v2 based model in MLE and RL training methods, respectively. All of our experiments are conducted on a Titan X GPU with 12G memory.

In testing, the sentence is generated by two available strategies. One is at each time step choosing the maximum probability word as the current output and feeding it into next time calculation as an input until the EOS word appear or the maximum length of sentence is reached. The other is beam search, and at each time step, it searches the top- $k$  optimal fractional sentences, then regards them as the candidates to generate new top- $k$  best sentences at the next time step. According to previous work [25], we adopt the second strategy and set the beam search size  $k$  as 1 or 3 in different tasks.

### D. Quantitative Results and Analysis

On Flickr8k, Flickr30k and MSCOCO datasets, we evaluate our model with previous works, and assess different combinations of our model. The results are presented in Table I and Table II. The models we denote are as follows:

<sup>4</sup><https://github.com/twtygqyy/Inception-resnet-v2>



TABLE II: Performance comparison of BLEU-1,2,3,4, METEOR, ROUGE-L, CIDEr-D compared with other state-of-the-art methods on MSCOCO dataset. The superscript “V” is the VggNet and “G” means the GoogleNet, (-) indicates an unknown metric. We apply the VggNet as our CNN encoder. The best results are in bold face and the second best results with underline.

Model	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr-D
Google NIC <sup>G</sup> [2]	66.6	46.1	32.9	24.6	—	—	—
m-RNN <sup>V</sup> [1]	67	49	35	25	—	—	—
Log Bilinear <sup>V</sup> [50]	70.8	48.9	34.4	24.3	20.0	—	—
LRCN <sup>V</sup> [46]	70.8	53.6	39.9	29.8	24.3	52.1	88.8
Bi-LSTM <sup>V</sup> [42]	67.2	49.2	35.2	24.4	—	—	—
Soft-Attention <sup>V</sup> [6]	70.7	49.2	34.4	24.3	23.9	—	—
Hard-Attention <sup>V</sup> [6]	71.8	50.4	35.7	25.0	23.0	—	—
ATT <sup>V</sup> [41]	70.9	53.7	40.2	30.4	24.3	—	—
Att-CNN+LSTM <sup>V</sup> [12]	<b>74</b>	56	42	31	<b>26</b>	—	94
Areas-Attention <sup>V</sup> [24]	72.1	—	—	31.1	25.0	—	95.6
SC-Tanh <sup>V</sup> [39]	71.6	54.5	40.5	30.1	24.7	—	97.0
RIC <sup>V</sup> [64]	72.1	52.1	36.4	27.3	23.8	—	—
MSM <sup>G</sup> [25]	73	<b>56.5</b>	<b>42.9</b>	<b>32.5</b>	25.1	<b>53.8</b>	98.6
Vgg16-LSTM	71.9	55.0	41.4	31.1	25.0	53.2	95.5
Vgg16+3-LSTM	72.1	55.2	41.4	31.1	25.0	53.3	96.2
Vgg16-LSTM-H	72.2	55.4	41.7	31.2	25.0	53.2	95.3
Vgg16-H	72.3	55.4	41.7	31.4	25.5	53.3	98.1
DHEDN <sub>1</sub>	72.8	56.0	42.3	32.1	25.5	<u>53.7</u>	<b>100.1</b>
DHEDN <sub>2</sub>	72.6	55.9	42.5	<b>32.5</b>	25.7	<b>53.8</b>	99.5
DHEDN <sub>3</sub>	<u>73.1</u>	<u>56.3</u>	<u>42.6</u>	<u>32.3</u>	25.6	<u>53.7</u>	99.3

TABLE III: Comparisons of the image captioning performance of the existing methods, which is using deep CNN or RL, on MSCOCO dataset. The Inception-Resnet-v2 is used in our model here. (-) indicates an unknown metric. The best results are in bold face and the second best results with underline.

Model	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr-D
SCN [44]	72.8	56.6	43.3	33.0	25.7	—	101.2
Adaptive [40]	74.2	58.0	43.9	33.2	26.6	—	108.5
Stack-Cap (MLE) [65]	76.2	60.4	46.4	35.2	26.5	—	109.1
Contrastive [66]	75.5	59.8	46.0	35.3	<u>27.1</u>	55.9	114.2
PG-BCMR [67]	75.4	59.1	44.5	33.2	25.7	55.0	101.3
SCST:Att2in [27]	—	—	—	33.3	26.3	55.3	111.4
TD-FC [68]	75.9	59.5	44.6	33.1	26.0	54.9	109.8
DHEDN <sub>1</sub> (MLE)	75.1	59.0	45.6	35.2	27.1	55.8	108.8
DHEDN <sub>2</sub> (MLE)	74.9	58.8	45.5	35.2	27.0	55.6	108.9
DHEDN <sub>3</sub> (MLE)	75.5	59.5	46.1	35.6	27.1	56.0	109.7
DHEDN <sub>1</sub> (PG)	<u>80.6</u>	<u>63.3</u>	48.3	36.2	<b>27.2</b>	56.9	<u>115.7</u>
DHEDN <sub>2</sub> (PG)	<u>80.4</u>	<u>63.3</u>	48.4	36.4	27.1	57.0	<u>115.5</u>
DHEDN <sub>3</sub> (PG)	<b>80.8</b>	<b>63.7</b>	<b>48.8</b>	<b>36.7</b>	<b>27.2</b>	<b>57.2</b>	<b>117.0</b>

- **Vgg16-LSTM.** Compared to the full model, this one inputs the Vgg16 extracted image feature, omits the top-most SF-LSTM and imports the middle VSE-LSTM output to the prediction layers directly. “Vgg16” means the 16-layer VggNet.
- **Vgg16+3-LSTM.** Its structure is the same as Vgg16-LSTM, but the image feature is extracted by the 16+3 VggNet, which outputs the deep CNN visual representation. “Vgg16+3” represents the defined 16+3 VggNet. In particular, the language model of Vgg16-LSTM and Vgg16+3-LSTM is motivated by [46].
- **Vgg16-LSTM-H.** “LSTM-H” denotes a deep hierarchical LSTM structure. This model implements a 3 layers LSTM, but the used LSTMs are the generic LSTM.
- **Vgg16-H.** “H” denotes our proposed deep hierarchical structure in the language model. In here, this model resembles the DHEDN<sub>1</sub> but without using the deep visual feature.
- **DHEDN<sub>1,2,3</sub>.** These different structures in the disparate SF-LSTMs of our model are presented in Section III-C.

**Performance on Flickr8k and Flickr30k.** The first perfor-

mance evaluation of our model is on Flickr8k and Flickr30k datasets for image captioning which is summarized in Table I. Because of the limitations of database size limitations, the results are sensitive to extra information. Our results reported by VggNet are compared with the others which adopt similar image encoders such as GoogleNet [69] or VggNet. The beam size in testing is 3. The results show that our model performs best on most indicators.

**Performance on MSCOCO.** Results of image captioning on MSCOCO are presented in Table II and Table III. In Table II, we present the result of variable variants of our model in VggNet, and our model is compared with the state-of-the-art methods whether using attention mechanism or other extra information by similar image encoders to the VggNet. In Table III, our model uses the Inception-Resnet-v2 as the CNN model, is trained by the Maximum Likelihood Estimation (MLE) and Policy Gradient (PG), respectively. In this table, our method is comparing with other deep CNN or RL based approaches. The beam size is 3 in testing. It is not surprising that DHEDN model gets the highest performance on most metrics.

**Quantitative Analysis.** The results on Flickr8k, Flickr30k

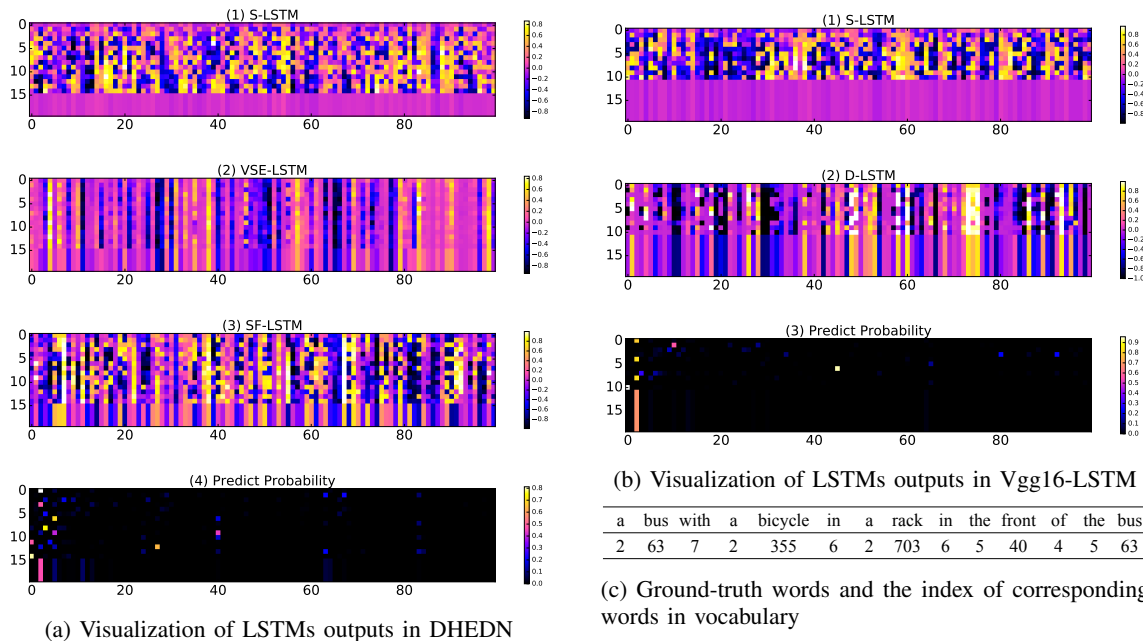


Fig. 4: Illustration of the top 100 units output from each layer of the language model in DHEDN<sub>3</sub> and Vgg16-LSTM. The horizontal axis indicates to different LSTM units and the final probability units. The vertical axis reveals 20 time steps. In the third subfigure, we provide the ground-truth words in chronological order and their corresponding index in vocabulary.

and MSCOCO datasets from the Table I, Table II and Table III which indicate our proposed DHEDNs exhibit better performance than other state-of-the-art methods. In Table I and Table II, on complicated and advanced metrics like Meteor, Rouge-L and CIDEr-D, our VggNet based model achieves the highest performance on all datasets. Besides, on MSCOCO, the CIDEr-D exceed 100%, which is the best report so far when extracting image representation by VggNet. Moreover, we present variable variants. From the results of Vgg16-LSTM and Vgg16+3-LSTM, it shows that the deep image feature has better representation ability to the shallow. The comparison between Vgg16-LSTM-H and Vgg16-H confirms the effectiveness of the DHEDN structure in language model.

In Table III, we introduce the Inception-Resnet-v2 model to replace the VggNet. And compared with the state-of-the-art deep CNN or RL based methods, our method achieves the highest performance by significant margins across most metrics. Compared with the MLE trained methods, our MLE trained results have significant advantage. Furthermore, our PG optimized model widening the advantages. It indicates that the RL is applicable to our model. All the results basically indicate that our deep hierarchical structure makes the dividing of encoder and decoder in image captioning model, which effectively improves the performance of results. DHEDN<sub>1</sub> as the primary composition which regard the vision-sentence embedded vector as one of the inputs embodied remarkable performance boost. DHEDN<sub>2</sub> and DHEDN<sub>3</sub> implement visual semantic strengthening in different ways by the vision-sentence embedded vector and achieve certain effects on the Flickr30k and MSCOCO datasets to DHEDN<sub>1</sub>. In contrast, DHEDN<sub>3</sub> behaves more excellent than DHEDN<sub>2</sub>. It indicates that letting the vision-sentence embedded vector as an aux-

iliary information to image representation is beneficial for caption generation.

**Efficiency Analysis.** In addition to presenting the performance of our model, our model has high computational efficiency as well. In Table IV, on the MSCOCO dataset, we compute the average time of the training for each iteration. And we compute the average time of all the testing image captions generation for various models, respectively. We set the batch size as 50 for training and select the beam search as the testing strategy, the beam size is 3. It can be shown that the deep and complex models cost slightly higher time consumption but achieve significant improvements. The DHEDN can strike the balance between performance and efficiency.

TABLE IV: Average time costs for training and testing.

	Vgg16-LSTM	Vgg16-LSTM-H	DHEDN
training	1.10s	1.14s	1.18s
testing	0.187s	0.243s	0.266s

### E. Visualized Analysis

To indicate how the Deep Hierarchical Encoder-Decoder Network (DHEDN) works in the process of caption generated over time, visualized analysis is conducted to the properties of the model.

First, we visualize and compare the different hierarchy outputs of LSTMs and finally prediction output between the DHEDN<sub>3</sub> and Vgg16-LSTM, which we can see in Fig. 4, to print the properties of our hierarchical structure and contributes to the relevant word prediction over time. The S-LSTMs in Fig. 4a (1) and Fig. 4b (1) are the encoders to get the textual distribution of the vocabulary, which changed

TABLE V: Comparison to leaderboard of the published state-of-the-art image captioning models on the online MSCOCO testing server, where B-N is short for BLEU-N score. Results using 5 references and 40 references captions are shown simultaneously. Our submissions are the ensemble of 3 models with different initialization. The best results are highlighted.

Model	B-1		B-2		B-3		B-4		METEOR		ROUGE-L1		CIDEr-D	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
ATT [41]	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
Att-CNN+LSTM [12]	72.5	89.2	55.6	80.3	41.4	69.4	30.6	58.2	24.6	32.9	52.8	67.2	91.1	92.4
Review Net [43]	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9
MSM [25]	73.9	91.9	57.5	84.2	43.6	74	33	63.2	25.6	35	54.2	70	98.4	100.3
SCA-CNN [70]	71.2	89.4	54.2	80.2	40.4	79.1	30.2	57.9	24.4	33.1	52.4	67.4	91.2	92.1
SCN [44]	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 [40]	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
PG-BCMR [67]	75.1	91.6	59.1	84.2	44.5	73.8	33.1	62.4	25.5	33.9	55.1	69.4	104.2	105.9
SCST:Att2in [27]	78.1	93.1	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Contrastive [66]	74.2	91.0	57.7	83.1	43.6	72.8	32.6	61.7	26.0	35.0	54.4	69.5	101.0	102.9
Stack-Cap (RL) [65]	77.8	93.2	61.6	86.1	46.8	76.0	34.9	64.6	27.0	35.6	56.2	70.6	<b>114.8</b>	<b>118.3</b>
TD-ATT [68]	75.7	91.3	59.1	83.6	44.1	72.6	32.4	60.9	25.9	34.2	54.7	68.9	105.9	109.0
Human	66.3	88.0	46.9	74.4	32.1	60.3	21.7	47.1	25.2	33.5	48.4	62.6	85.4	91.0
DHEDN (MLE)	74.8	92.2	58.5	84.7	44.9	74.9	34.5	64.6	26.9	<b>36.5</b>	55.4	71.2	105.0	104.9
DHEDN (PG)	<b>80.6</b>	<b>94.2</b>	<b>63.5</b>	<b>86.8</b>	<b>48.3</b>	<b>76.7</b>	<b>36.0</b>	<b>65.4</b>	<b>27.0</b>	35.5	<b>56.8</b>	<b>71.3</b>	112.5	114.4

TABLE VI: Performance of retrieval task compare with state-of-the-art methods where R@K according to Recall@K (high is good) and Med r means to median recall (low is good). The superscript “V” means the VggNet and “G” represent the GoogleNet, “I” indicates the Inception-Resnet-v2, (-) points an unknown metric. The 1K test images strategy means the result of 1,000 samples of the test set, the 5K test images strategy indicates the result is obtained on the entire test set. The best results are marked in bold.

Datasets	Model	Image to Caption				Caption to Image			
		R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Flickr8K	m-RNN <sup>V</sup> [1]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
	DeepVS <sup>V</sup> [3]	16.5	40.6	54.2	7.6	11.8	32.1	44.7	12.4
	Google NIC <sup>G</sup> [2]	20	—	60	6	19	—	<b>64</b>	5
	Bi-LSTM <sup>V</sup> [42]	<b>29.3</b>	<b>58.2</b>	<b>69.6</b>	<b>3</b>	<b>19.7</b>	<b>47.0</b>	60.6	<b>5</b>
	DHEDN	23.3	51.5	63.0	5	17.6	43.3	57.3	8
Flickr30K	m-RNN <sup>V</sup> [1]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
	DeepVS <sup>V</sup> [3]	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
	LRCN <sup>V</sup> [46]	23.6	46.6	58.3	7	17.5	40.3	50.8	9
	Google NIC <sup>G</sup> [2]	17	—	56	7	17	—	57	8
	Bi-LSTM <sup>V</sup> [42]	28.1	53.1	64.2	4	19.6	43.8	55.8	7
	DHEDN	<b>41.0</b>	<b>67.5</b>	<b>76.2</b>	<b>2</b>	<b>28.9</b>	<b>56.5</b>	<b>68.6</b>	<b>4</b>
MSCOCO	1K test images								
	DeepVS <sup>V</sup> [3]	16.5	39.2	52.0	9	10.7	29.6	42.2	14.0
	LRCN <sup>V</sup> [46]	53.3	84.3	91.9	1	39.3	74.7	85.9	2
	Bi-LSTM <sup>V</sup> [42]	16.6	39.4	52.4	9	11.6	30.9	43.4	13
	Convnets <sup>V</sup> [71]	16.9	39.8	53.1	8	12.4	31.5	44.0	12
	2WayNet <sup>V</sup> [72]	55.8	75.2	—	—	39.7	63.3	—	—
	DHEDN <sup>V</sup>	59.1	85.7	92.0	1	43.0	77.5	88.5	2
	DHEDN <sup>I</sup>	<b>73.3</b>	<b>92.3</b>	<b>96.2</b>	<b>1</b>	<b>54.5</b>	<b>85.2</b>	<b>92.6</b>	<b>1</b>
	5K test images								
	FA [73]	17.3	39.0	50.2	10	10.8	28.3	40.1	17
	DeepVS <sup>V</sup> [3]	16.5	39.2	52.0	9	10.7	29.6	42.2	14
	Embeddings <sup>V</sup> [74]	23.3	—	65.0	5	18.0	—	57.6	7
	DHEDN <sup>V</sup>	32.5	60.3	73.5	3	21.9	49.4	63.2	6
	DHEDN <sup>I</sup>	<b>50.4</b>	<b>78.0</b>	<b>86.4</b>	<b>1</b>	<b>33.1</b>	<b>63.0</b>	<b>75.0</b>	<b>3</b>

over time until the EOS signal appear or reach the maximum length. The SF-LSTM and D-LSTM in Fig. 4a (3) and Fig. 4b (2) are the decoders to extract correlative information between the visual and textual context to generate the representation of the predicted word. No matter in Fig. 4a or in Fig. 4b, we can find that the distributions between the outputs of the encoder or decoder LSTM are very discrepancy. Its difficult to directly transform the encoder output by the decoder LSTM. But the introduction of the VSE-LSTM, which are conditioned by the hidden state of S-LSTM units and shallow CNN visual representation jointly, can narrow the discrepancy. From Fig.

4a (2), it can be seen that the distribution of the VSE-LSTM output is an intersection, which constructs a vision-sentence embedded space, and relates the processes of the encoder and decoder. It should be indicated that the vision-sentence embedded space which is composited by external image and textual features can contribute to the caption generation.

Second, we present our model with the Vgg16 and Vgg16+3 non-hierarchical models by some representative captioned images (see Fig. 5). Some interesting phenomenons can be found as follows: 1. **Express Accurately and Informatively**. Our model has powerful information extraction and organization






 <p>(a)</p>	<p><b>Generated Sentences:</b></p> <p>DHEDN<sub>3</sub>: The big ben clock tower towering over the city of london.</p> <p>Vgg16-LSTM: A tall clock tower with a sky background.</p> <p>Vgg16+3-LSTM: A very tall clock tower towering over a city.</p> <p>Vgg16-H: The big ben clock tower towering over the city of london.</p>	<p><b>Ground-truth:</b></p> <p>① : A large tower that has a big clock at top.</p> <p>② : A clock tower on top of a tall historic building.</p> <p>③ : Clock tower next to a large building in a city.</p>
 <p>(b)</p>	<p><b>Generated Sentences:</b></p> <p>DHEDN<sub>3</sub>: A polar bear playing with a ball in the water.</p> <p>Vgg16-LSTM: A polar bear swimming in a pool of water.</p> <p>Vgg16+3-LSTM: A polar bear swimming in a pool of water.</p> <p>Vgg16-H: A polar bear swimming in a pool of water.</p>	<p><b>Ground-truth:</b></p> <p>① : A polar bear playing with an orange ball in a cage.</p> <p>② : A wet polar bear half submerged in water playing with a ball.</p> <p>③ : A polar bear in water near an orange ball in a cage.</p>
 <p>(c)</p>	<p><b>Generated Sentences:</b></p> <p>DHEDN<sub>3</sub>: A man flying through the air while riding a skateboard.</p> <p>Vgg16-LSTM: A man riding a skateboard down the side of a ramp.</p> <p>Vgg16+3-LSTM: A man riding a skateboard down the side of a ramp.</p> <p>Vgg16-H: A man flying through the air while riding a skateboard.</p>	<p><b>Ground-truth:</b></p> <p>① : A person doing skateboard ticks at a skate park.</p> <p>② : A guy on a skate board high in the air.</p> <p>③ : A person on a skate board is doing a trick.</p>
 <p>(d)</p>	<p><b>Generated Sentences:</b></p> <p>DHEDN<sub>3</sub>: A toilet that is sitting on the ground.</p> <p>Vgg16-LSTM: A white toilet sitting in a bathroom next to a brick wall.</p> <p>Vgg16+3-LSTM: A white toilet sitting in the middle of a bathroom.</p> <p>Vgg16-H: A white toilet sitting in a bathroom next to a wall.</p>	<p><b>Ground-truth:</b></p> <p>① : A laptop sitting on top of a toilet in a room.</p> <p>② : An open laptop computer sits on a closed toilet.</p> <p>③ : Silver laptop sitting on top of a public toilet.</p>
 <p>(e)</p>	<p><b>Generated Sentences:</b></p> <p>DHEDN<sub>3</sub>: A subway train with its doors open at a station.</p> <p>Vgg16-LSTM: A subway train stopped at a train station.</p> <p>Vgg16+3-LSTM: A subway train stopped at a train station.</p> <p>Vgg16-H: A subway train that is parked at a station.</p>	<p><b>Ground-truth:</b></p> <p>① : A train with open doors at a terminal.</p> <p>② : The London Subway is empty with both doors open.</p> <p>③ : The doors of an underground train are open.</p>

Fig. 5: Examples comparison of generated captions. The middle column sentences are generated by different models: 1) black-colored captions are generated by DHEDN<sub>3</sub>, 2) blue-colored captions are generated by Vgg16-LSTM, 3) red-colored captions are generated by Vgg16+3-LSTM, 4) purple-colored captions are generated by Vgg16-H. The right column presents three randomly selected ground truth sentences.

ability in sampled examples. For example, to (e), our model not only discovers the “train” and “station”, but also finds “its doors open” and properly organized them to a sentence as well. 2. **Interactivity**. In our model, there are sufficient interactions between different objects, to (a), our model more than describes the image as “a very tall clock tower towering over a city”, extracts the interactional of the “clock” and “city”, and regards them as “The big ben clock” and “london”, respectively. 3. **Conform to Human Speech Mode**. The captions generated by our model conform similar to human speech mode. For example, in (d), the caption uses subject clause to describe image and in (c), the application of temporal adverbial clause makes the extracted information properly composited in a sentence, even like a human-being.

In addition, some sampled captions illustrate in Fig. 6. After statistics, 46.3% of our generated sentences on MSCOCO testing set by VggNet exists in its training set. We use beam size 1 to compute the vocabulary words usage rate. On the VggNet, we find that the usage rate of top 1,000 to our model is around 88%, but other two non-hierarchical models exceed 89%. Like “hummingbird”, but the other non-hierarchical models describe to “bird” or “colorful bird”. Furthermore, the nouns used times of hierarchical models (DHEDN<sub>3</sub> and Vgg16-H) from top 500 to 1,000 of vocabulary in testing set are 882 and 1,031 which are much higher than the non-hierarchical models (681 and 739, respectively). We continue count the results of Inception-Resnet-v2 based model, and the usage rate of top 1,000 to this model is around 88% as well.

The nouns used time of this model from top 500 to 1,000 of vocabulary is 1,206.

In a word, the visualized analyses illustrate the strong capability of our model to generating novel captions.

#### F. Online Evaluation on MSCOCO Test Server

To further evaluate our model, we compare the ensemble DHEDN<sub>3</sub> models on the online MSCOCO Test Server with state-of-the-art systems summarized in Table V. We use the Inception-Resnet-v2 model as the image encoder and upload the results of our MLE and PG based 3 ensemble model to the online MSCOCO test server. Our model approach the best performance on most metrics among the published works. Compared to the level of human expression of images, our model achieves significant progress.

#### G. Image-Caption Retrieval

Retrieval is also a challenging task [75], [76]. We further explore the retrieval performance of our DHEDN<sub>3</sub> model on the all Flickr8K and Flickr30K testing data sets. On the MSCOCO, we follow the operation of [59] to choose subset of 1K images and all the 5K images from the MSCOCO testing set to implement the retrieval task. All of the models are trained by MLE. The evaluation results are showed in Table VI. On the smallest dataset Flickr8K, despite the performance of our model is outstanding, it cannot exceed the results of Bi-LSTM<sup>V</sup> which employs bidirectional LSTM as the language





Fig. 6: Examples of sentences generation results by our model. Images are chosen from the testing set. We used beam search with beam size of 3, and display the highest likelihood result above.



A public transit bus on a city street.

A bus that is sitting in the street.

A bus sitting on the side of a road.

A double decker bus driving down a street.

A group of stuffed animals are arranged together



Fig. 7: Two examples of caption to image (top) and image to caption (bottom) by our proposed model. Results are arranged in order from left to right.

model in image captioning. While on the larger datasets, our model presents absolute dominance in all metrics. On the Flickr30K, our model improves the  $R@1$  and  $Medr$  of caption-image retrieval over 50% compare to the second best model. On the MSCOCO, we present two results are based on VggNet and Inception-Resnet-v2, respectively. The VggNet based model is competitive compare with other models, and the Inception-Resnet-v2 based model is absolute dominance in all metrics. These suggest that complicated models may not achieve better performance for some tasks than simple models on small scale datasets. To the DHEDN<sub>3</sub> model, with the size of the dataset growing larger, the effect of retrieval results is better. It indicates that our deep hierarchical structure can better suit larger datasets which can provide more data for training the mass of parameters of complicate model. Besides, with the increasing of training samples, our model can significantly improve the capacity to capture the vision-language correlation for image-caption retrieving. In general cases, our method can guarantee sensible the corresponding

captions or images be retrieved, and two examples are showed in Fig. 7.

## V. CONCLUSION

In this paper, we have presented Deep Hierarchical Encoder-Decoder Network (DHEDN), a novel three-layers LSTM based encoder-decoder to effectively fuse visual and textual semantics to generate appropriate captions. We explore the usages of output from the middle layer of LSTMs to enhance the decoding ability of top-most LSTM. Depending on the introduction of semantic enhancement module and distribution combine module, the DHEDN has devised variants of architectures of the decoder LSTM to better demonstrate the impact of the deep hierarchical architecture. In addition to the MLE, we apply a policy gradient optimized method to improve the performance of our model. Experiments conducted on three datasets and different tasks validate the superiority of our proposal. Significant improvement on performances demonstrate the effectiveness of our model when compared with the

other captioning methods. The results of our DHEDN model outperform state-of-the-art techniques both on multitask and on the MSCOCO image captioning leaderboard.

There are several places of our current work we desire to explore. First, we believe that it makes sense to continue to increase the “vertical depth” of encoder-decoder, but how to stack different layers and set those numerous parameters remain difficult to be solved. Second, importing extra inputs like visual attention or attributes into the deep hierarchical encoder-decoder. The complementary knowledge has been demonstrated beneficial for the performance improvement of image descriptions generation. How to integrate complementary knowledge into the deep hierarchical encoder-decoder structure is worth trying and seems very attractive.

## VI. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grants 91646207, 61773377, and 61573352, and the Beijing Natural Science Foundation under Grant L172053.

## REFERENCES

- [1] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn),” in *International Conference on Learning Representations (ICLR)*, 2015.
- [2] 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 (CVPR)*, pp. 3156–3164, 2015.
- [3] A. Karpathy and F. Li, “Deep visual-semantic alignments for generating image descriptions,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137, 2015.
- [4] R. Kiros, R. Salakhutdinov, and R. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” *Computing Research Repository (CoRR)*, vol. abs/1411.2539, 2014.
- [5] P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi, “TREETALK: composition and compression of trees for image descriptions,” *IEEE Transactions of the Association for Computational Linguistics (TACL)*, vol. 2, pp. 351–362, 2014.
- [6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *IEEE International Conference on Machine Learning (ICML)*, pp. 2048–2057, 2015.
- [7] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *International Conference on Learning Representations (ICLR)*, 2015.
- [8] I. Sutskever, O. Vinyals, and Q. Le, “Sequence to sequence learning with neural networks,” in *Conference on Neural Information Processing Systems (NIPS)*, pp. 3104–3112, 2014.
- [9] K. Cho, B. Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.
- [10] J. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *IEEE Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [11] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. Shen, “Video captioning with attention-based LSTM and semantic consistency,” *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.
- [12] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Hengel, “What value do explicit high level concepts have in vision to language problems?,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 203–212, 2016.
- [13] X. Yang, T. Zhang, C. Xu, S. Yan, M. Hossain, and A. Ghoneim, “Deep relative attributes,” *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1832–1842, 2016.
- [14] M. Spratling and M. Johnson, “A feedback model of visual attention,” *Journal of cognitive neuroscience*, vol. 16, no. 2, pp. 219–237, 2004.
- [15] D. Parikh and K. Grauman, “Relative attributes,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 503–510, 2011.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [17] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Conference on Neural Information Processing Systems (NIPS)*, pp. 1106–1114, 2012.
- [18] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, “A deep neural network-driven feature learning method for multi-view facial expression recognition,” *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.
- [19] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [20] Y. Wu, M. Schuster, Z. Chen, Q. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *Computing Research Repository (CoRR)*, vol. abs/1609.08144, 2016.
- [21] T. Luong, H. Pham, and C. Manning, “Effective approaches to attention-based neural machine translation,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1412–1421, 2015.
- [22] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba, “Addressing the rare word problem in neural machine translation,” in *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 11–19, 2015.
- [23] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, “Deep recurrent models with fast-forward connections for neural machine translation,” *IEEE Transactions of the Association for Computational Linguistics*, vol. 4, pp. 371–383, 2016.
- [24] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, “Areas of attention for image captioning,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1251–1259, 2017.
- [25] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, “Boosting image captioning with attributes,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 4904–4912, 2017.
- [26] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, “Sequence level training with recurrent neural networks,” in *International Conference on Learning Representations (ICLR)*, 2016.
- [27] S. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179–1195, 2017.
- [28] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg, “Babytalk: Understanding and generating simple image descriptions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.
- [29] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, “Composing simple image descriptions using web-scale n-grams,” in *Conference on Computational Natural Language Learning (CoNLL)*, pp. 220–228, 2011.
- [30] Y. Yang, C. Teo, H. III, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 444–454, 2011.
- [31] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. III, “Midge: Generating image descriptions from computer vision detections,” in *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 747–756, 2012.
- [32] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, “Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2712–2719, 2013.
- [33] D. Elliott and F. Keller, “Image description using visual dependency representations,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1292–1302, 2013.
- [34] A. Farhadi, S. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every picture tells a story: Generating sentences from images,” in *European Conference on Computer Vision (ECCV)*, pp. 15–29, 2010.



- [35] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Meeting of the Association for Computational Linguistics, Proceedings (ACL)*, pp. 359–368, 2012.
- [36] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language models for image captioning: The quirks and what works," in *Meeting of the Association for Computational Linguistics (ACL)*, pp. 100–105, 2015.
- [37] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in *Conference on Artificial Intelligence (AAAI)*, pp. 4176–4182, 2017.
- [38] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6077–6086, 2018.
- [39] L. Zhou, C. Xu, P. Koch, and J. Corso, "Watch what you just said: Image captioning with text-conditional attention," in *ACM Multimedia*, pp. 305–313, 2017.
- [40] 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 (CVPR)*, pp. 3242–3250, 2017.
- [41] 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 (CVPR)*, pp. 4651–4659, 2016.
- [42] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional lstms," in *Conference on Multimedia Conference (MM)*, pp. 988–997, 2016.
- [43] Z. Yang, Y. Yuan, Y. Wu, W. Cohen, and R. Salakhutdinov, "Review networks for caption generation," in *Conference on Neural Information Processing Systems (NerulPS)*, pp. 2361–2369, 2016.
- [44] 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 (CVPR)*, pp. 1141–1150, 2017.
- [45] A. Graves, "Generating sequences with recurrent neural networks," *Computing Research Repository (CoRR)*, vol. abs/1308.0850, 2013.
- [46] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.
- [47] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," *Computing Research Repository (CoRR)*, vol. abs/1211.5063, 2012.
- [48] R. Sutton and A. Barto, "Reinforcement learning: An introduction," *IEEE Transactions Neural Networks*, vol. 9, no. 5, pp. 1054–1054, 1998.
- [49] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, vol. 550, 2017.
- [50] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on Machine Learning (ICML)*, pp. 595–603, 2014.
- [51] Y. Tan and C. Chan, "phi-lstm: A phrase-based hierarchical LSTM model for image captioning," in *Asian Conference on Computer Vision (ACCV)*, pp. 101–117, 2016.
- [52] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *NAACL HLT Workshop*, pp. 139–147, Association for Computational Linguistics, 2010.
- [53] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *IEEE Transactions of the Association for Computational Linguistics (TACL)*, vol. 2, pp. 67–78, 2014.
- [54] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft COCO: common objects in context," in *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.
- [55] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.
- [56] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, vol. 29, pp. 65–72, 2005.
- [57] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, vol. 8, 2004.
- [58] R. Vedantam, C. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015.
- [59] A. Karpathy, A. Joulin, and F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Conference on Neural Information Processing Systems (NerulPS)*, pp. 1889–1897, 2014.
- [60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.
- [61] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Conference on Artificial Intelligence (AAAI)*, pp. 4278–4284, 2017.
- [62] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [63] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.
- [64] H. Liu, Y. Yang, F. Shen, L. Duan, and H. Shen, "Recurrent image captioner: Describing images with spatial-invariant transformation and attention filtering," *Computing Research Repository (CoRR)*, vol. abs/1612.04949, 2016.
- [65] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *Conference on Artificial Intelligence (AAAI)*, pp. 6837–6844, 2018.
- [66] B. Dai and D. Lin, "Contrastive learning for image captioning," in *Conference on Neural Information Processing Systems (NerulPS)*, pp. 898–907, 2017.
- [67] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 873–881, 2017.
- [68] H. Chen, G. Ding, S. Zhao, and J. Han, "Temporal-difference learning with sampling baseline for image captioning," in *Conference on Artificial Intelligence (AAAI)*, 2018.
- [69] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.
- [70] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6298–6306, 2017.
- [71] Y. Liu, Y. Guo, and M. Lew, "What convnets make for image captioning?," in *International Conference on MultiMedia Modeling (MMM)*, pp. 416–428, 2017.
- [72] A. Eisenschlat and L. Wolf, "Linking image and text with 2-way nets," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1855–1865, 2017.
- [73] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4437–4446, 2015.
- [74] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *International Conference on Learning Representations (ICLR)*, 2016.
- [75] I. González-Díaz, M. Birinci, F. Díaz-de-María, and E. Delp, "Neighborhood matching for image retrieval," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 544–558, 2017.
- [76] L. Dong, Y. Liang, G. Kong, Q. Zhang, X. Cao, and E. Izquierdo, "Holons visual representation for image retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 714–725, 2016.