

Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning

CHENG WANG, HAOJIN YANG, and CHRISTOPH MEINEL, Hasso Plattner Institute, University of Potsdam

Generating a novel and descriptive caption of an image is drawing increasing interests in computer vision, natural language processing, and multimedia communities. In this work, we propose an end-to-end trainable deep bidirectional LSTM (Bi-LSTM (Long Short-Term Memory)) model to address the problem. By combining a deep convolutional neural network (CNN) and two separate LSTM networks, our model is capable of learning long-term visual-language interactions by making use of history and future context information at high-level semantic space. We also explore deep multimodal bidirectional models, in which we increase the depth of nonlinearity transition in different ways to learn hierarchical visual-language embeddings. Data augmentation techniques such as multi-crop, multi-scale, and vertical mirror are proposed to prevent overfitting in training deep models. To understand how our models “translate” image to sentence, we visualize and qualitatively analyze the evolution of Bi-LSTM internal states over time. The effectiveness and generality of proposed models are evaluated on four benchmark datasets: Flickr8K, Flickr30K, MSCOCO, and Pascal1K datasets. We demonstrate that Bi-LSTM models achieve highly competitive performance on both caption generation and image-sentence retrieval even without integrating an additional mechanism (e.g., object detection, attention model). Our experiments also prove that multi-task learning is beneficial to increase model generality and gain performance. We also demonstrate the performance of transfer learning of the Bi-LSTM model significantly outperforms previous methods on the Pascal1K dataset.

CCS Concepts: • Computing methodologies → Natural language generation; Neural networks; Computer vision representations;

Additional Key Words and Phrases: Deep learning, LSTM, multimodal representations, image captioning, mutli-task learning

40

ACM Reference format:

Cheng Wang, Haojin Yang, and Christoph Meinel. 2018. Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 2s, Article 40 (April 2018), 20 pages.

<https://doi.org/10.1145/3115432>

1 INTRODUCTION

It is challenging to describe an image using sentence-level captions (Karpathy and Li 2015; Karpathy et al. 2014; Kiros et al. 2014b; Kuznetsova et al. 2012, 2014; Mao et al. 2015; Socher et al. 2014; Vinyals et al. 2015), where the task is to map the input image to a sentence output that possesses its own structure. Inspired by the success of machine translation:translate source language to target language, image captioning system tries to “translate” an image to a sentence. It

Authors' addresses: C. Wang, H. Yang, and C. Meinel, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany; emails: {Cheng.Wang, Haojin.Yang, Christoph.Meinel}@hpi.de.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 ACM 1551-6857/2018/04-ART40 \$15.00

<https://doi.org/10.1145/3115432>

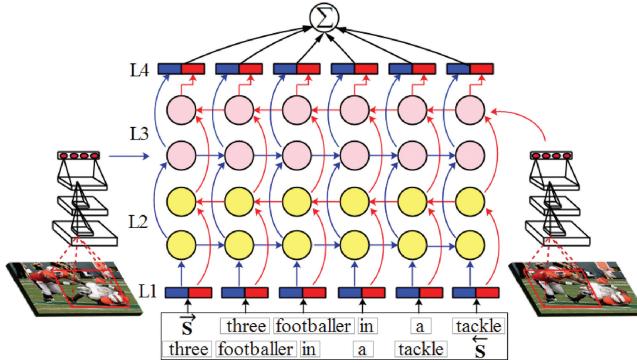


Fig. 1. Deep Multimodal Bidirectional LSTM. L1: sentence embedding layer. L2: Text-LSTM (T-LSTM) layer which receives text only. L3: Multimodal-LSTM (M-LSTM) layer which receives both image and text input. L4: Softmax layer. We feed sentence in both forward (blue arrows) and backward (red arrows) order which allows our model to summarize context information from both the left and right sides for generating a sentence word by word over time. Our model is end-to-end trainable by minimizing a joint loss.

requires not only the recognition of visual objects in an image and the semantic interactions between objects, but the ability to capture visual-language interactions and learn how to “translate” the visual understanding to sensible sentence descriptions. A general approach is to train a visual model using images and train a language model using provided captions. By learning a multimodal joint representation on images and captions, the semantic similarity of images and captions can be measured and thus recommend the most descriptive caption for a given input image. The most important part at the center of this visual-language modeling is to capture the semantic correlations across image and text modalities. While some previous works (Li et al. 2011; Kulkarni et al. 2013; Mitchell et al. 2012; Kuznetsova et al. 2012, 2014) have been proposed to address the problem of image captioning, they mostly use sentence templates, or treat image captioning as a retrieval task through ranking the best matching sentence in the database as the caption. Those approaches usually suffer difficulties in generating variable-length and novel sentences. Recent work (Karpathy and Li 2015; Karpathy et al. 2014; Kiros et al. 2014b; Mao et al. 2015; Socher et al. 2014; Vinyals et al. 2015) indicates that embedding visual and language to common semantic space with relatively shallow recurrent neural network (RNN) yields promising results.

In this work, we propose novel architectures to generate novel image descriptions. The overview of architecture is shown in Figure 1. Different from previous approaches, we learn a visual-language space where sentence embeddings are encoded using bidirectional Long Short-Term Memory (Bi-LSTM) and visual embeddings are encoded with Convolutional Neural Network (CNN). Typically, in unidirectional sentence generation, one general way of predicting next word w_t with visual context I and history textual context $w_{1:t-1}$ is to maximize $\log P(w_t|I, w_{1:t-1})$. While the unidirectional model includes past context, it is still limited to retaining future context $w_{t+1:T}$ that can be used for reasoning previous word w_t by maximizing $\log P(w_t|I, w_{t+1:T})$. The bidirectional model tries to overcome the shortcomings that each unidirectional (forward and backward direction) model suffers on its own and exploits the past and future dependence to give a prediction. As in Figure 2, two example images with bidirectionally generated sentences intuitively support our assumption that bidirectional captions are complementary; combining them can generate more sensible captions. Thus, our Bi-LSTM is able to summarize long-range visual-language interactions from forward and backward directions.



Fig. 2. Illustration of generated captions. Two example images from Flickr8K dataset and their best matching captions that generated in forward order (blue) and backward order (red). Bidirectional models capture different levels of visual-language interactions (more evidence see Section 4.7). The final caption is the sentence with higher probabilities (histogram under sentence). In both examples, backward caption is selected as final caption for corresponding images.

Inspired by the architectural depth of human brain, to learn higher level visual-language embeddings, we also explore the deeper bidirectional LSTM architectures where we increase the non-linearity by adding a hidden-to-hidden transformation layer. All of our proposed models can be trained in an end-to-end way by optimizing a joint loss in forward and backward directions. In addition, we design multi-task learning (Caruana 1998) and transfer learning (Pan and Yang 2010) to increase the generality of the proposed method on different datasets.

The core contributions of this work are fourfold:

- We propose an end-to-end trainable multimodal bidirectional LSTM and its deeper variant models (see Section 3.3) that embed image and sentence into a high-level semantic space by exploiting both long-term history and future context. The code, networks, and examples for this work can be found at our Github repository.¹
- We evaluate the effectiveness of proposed models on three benchmark datasets: Flickr8K, Flickr30K, and MSCOCO. Our experimental results show that bidirectional LSTM models achieve highly competitive performance on caption generation (Section 4.6).
- We explore the generality on multi-task/transfer learning models on Pascal1K (Section 4.5). It demonstrates that transferring a multi-task joint model on Flickr8K, Flickr30K, and MSCOCO to Pascal1K is beneficial and performs significantly better than recent methods (see Section 4.6).
- We visualize the evolution of hidden states of bidirectional LSTM units to qualitatively analyze and understand how to generate a sentence that is conditioned by visual context information over time (see Section 4.7).

The rest of the article is organized as follows. In Section 2, we review the related work on image captioning using deep architectures. In Section 3, we introduce the proposed deep multimodal bidirectional LSTM for image captioning and explore its deeper variant models. Section 4 presents several groups of experiments to illustrate the effectiveness of proposed methods. In Section 4.6, we compare our models with state-of-the-art methods; it shows that Bi-LSTM models achieve very competitive performance. In Section 4.7, we visualize the internal states of LSTM hidden units and show how our methods generalize to new datasets with multi-task/transfer learning; we also provide some illustrative examples. Section 5 summarizes our methods and presents future work.

2 RELATED WORK

This section gives the related knowledge. It starts by introducing Recurrent Neural Network (RNN) which equips neural networks with memories, followed by the review of recently proposed approaches on the image captioning task.

¹https://github.com/deepsemantic/image_captioning.

2.1 RNN

RNN is a powerful network architecture for processing sequential data. It has been widely used in natural language processing (Socher et al. 2011), speech recognition (Graves et al. 2013), and handwriting recognition (Graves et al. 2009) in recent years. In RNN, it allows cyclical connection and reuse of the weights across different instances of neurons; each of them is associated with different timesteps. This idea can explicitly support the network to learn the entire history of previous states and map them to current states. With this property, RNN is able to map an arbitrary length sequence to a fixed length vector.

LSTM (Long short-term memory) (Hochreiter and Schmidhuber 1997) is a particular form of traditional RNN. Compared to traditional RNN, LSTM can learn the long-term dependencies between inputs and outputs; it can also effectively prevent backpropagation errors from vanishing or exploding. LSTM has increasing popularity in the field of machine translation (Cho et al. 2014), speech recognition (Graves et al. 2013), and sequence learning (Sutskever et al. 2014) recently. Another special type of RNN is Gated Recurrent Unit (GRU) (Cho et al. 2014). GRU simplifies LSTM by removing the memory cell and provides a different way to prevent the vanishing gradient problem. GRU has been recently explored in language modeling (Chung et al. 2015), face aging (Wang et al. 2016a), face alignment (Wang et al. 2016b), and speech synthesis (Wu and King 2016). Motivated by those works, in the context of automatic image captioning, our networks build on bidirectional LSTM in order to learn the long-term interaction across image and sentence from both history and future information.

2.2 Image Captioning

Multimodal representation learning (Ngiam et al. 2011; Srivastava and Salakhutdinov 2012; Wang et al. 2016c) has significant value in multimedia understanding and retrieval. The shared concept across modalities plays an important role in bridging the “semantic gap” of multimodal data (Rasiwasia et al. 2007; Yang et al. 2015, 2016). Image captioning falls into this general category of learning multimodal representations.

Recently, several approaches have been proposed for image captioning. We can roughly classify those methods into three categories. The first category is template-based approaches that generate caption templates through detecting objects and discovering attributes in an image. For example, the work Li et al. (2011) was proposed to parse a whole sentence into several phrases, and learn the relationships between phrases and objects in an image. In Kulkarni et al. (2013), conditional random field (CRF) was used to correspond objects, attributes, and prepositions of image content and predict the best label. Other similar methods were presented in Mitchell et al. (2012), Kuznetsova et al. (2012, 2014). These methods are typically hard-designed and rely on a fixed template, which mostly lead to poor performance in generating variable-length sentences. The second category is retrieval-based approaches. This sort of method treats image captioning as a retrieval task by leveraging a distance metric to retrieve similar captioned images, and then modifying and combining retrieved captions to generate a caption (Kuznetsova et al. 2014). But these approaches generally need additional procedures such as modification and generalization process to fit image query.

Inspired by the recent success of CNN (Krizhevsky et al. 2012; Zeiler and Fergus 2014) and RNN (Mikolov et al. 2010, 2011; Bahdanau et al. 2015), the third category emerged as neural network based methods (Vinyals et al. 2015; Xu et al. 2015; Kiros et al. 2014b; Karpathy et al. 2014; Karpathy and Li 2015). Our work also belongs to this category. The work conducted by Kiros et al. (2014a) can be seen as a pioneer work to use neural network for image captioning with a multimodal neural language model. In their follow-up work (Kiros et al. 2014b), Kiros et al. introduced an encoder-decoder pipeline where a sentence was encoded by LSTM and decoded with a structure-content

neural language model (SC-NLM). Socher et al. (2014) presented a DT-RNN (Dependency Tree-Recursive Neural Network) to embed a sentence into a vector space in order to retrieve images. Later on, Mao et al. (2015) proposed m-RNN which replaces the feed-forward neural language model in Kiros et al. (2014b). Similar architectures were introduced in NIC (Vinyals et al. 2015) and LRCN (Donahue et al. 2015); both approaches use LSTM to learn text context. But NIC only feeds visual information at the first timestep while Mao et al. (2015) and LRCN (Donahue et al. 2015) consider image context at each timestep. Another group of neural network based approaches has been introduced in Karpathy et al. (2014) and Karpathy and Li (2015) where object detection with R-CNN (region-CNN) (Girshick et al. 2014) was used for inferring the alignment between image regions and descriptions.

Most recently, Fang et al. (2015) used multi-instance learning and a traditional maximum-entropy language model for image description generation. Chen and Zitnick (2015) proposed to learn visual representation with RNN for generating image captions. Xu et al. (2015) introduced an attention mechanism of human visual system into an encoder-decoder framework (Cho et al. 2015). It is shown that an attention model can visualize what the model “sees” and yields significant improvements on image caption generation. In You et al. (2016), the authors proposed a semantic attention model by combining top-down and bottom-up approaches in the framework of recurrent neural networks. In the bottom-up approach, semantic concepts or attributes are used as candidates. In the top-down approach, visual features are employed to guide where and when attention should be activated.

Unlike those models, our model directly assumes the mapping relationship between visual-semantic is antisymmetric and dynamically learns long-term bidirectional and hierarchical visual-semantic interactions with deep LSTM models. This is proved to be very effective in generation and retrieval tasks as we demonstrate in Section 4.

3 MODEL

In this section, we describe our multimodal Bi-LSTM model and explore its deeper variants. We first briefly introduce LSTM; the LSTM we used is described in Zaremba and Sutskever (2014).

3.1 Long Short-Term Memory

Our model builds on the LSTM cell; as shown in Figure 3, the reading and writing memory cell c is controlled by a group of sigmoid gates. At given timestep t , LSTM receives inputs from different sources: current input x , the previous hidden state of all LSTM units \mathbf{h}_{t-1} , as well as previous memory cell state \mathbf{c}_{t-1} . The updating of those gates at timestep t for given inputs \mathbf{x}_t , \mathbf{h}_{t-1} , and \mathbf{c}_{t-1} is as follows:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i), \quad (1)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f), \quad (2)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o), \quad (3)$$

$$\mathbf{g}_t = \phi(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c), \quad (4)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, \quad (5)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \phi(\mathbf{c}_t), \quad (6)$$

where without considering the optional peephole connections, \mathbf{W} is the weight matrix learned from the network and \mathbf{b} is the bias term. σ is the sigmoid activation function $\sigma(x) = \frac{1}{1+\exp(-x)}$

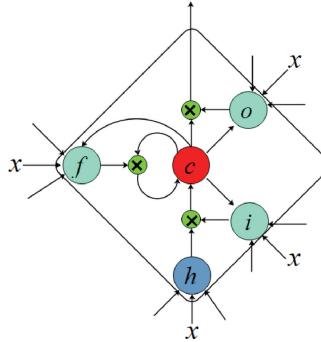


Fig. 3. Long Short-Term Memory (LSTM) cell. It consists of an input gate i , a forget gate f , a memory cell c , and an output gate o . The input gate decides to let an incoming signal go through to the memory cell or block it. The output gate can allow new output or prevent it. The forget gate decides to remember or forget the cell's previous state. Updating cell states is performed by feeding previous cell output to itself by recurrent connections in two consecutive timesteps.

and ϕ presents hyperbolic tangent $\phi(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$. \odot denotes the products with a gate value. The LSTM hidden output $\mathbf{h}_t = \{\mathbf{h}_{tk}\}_{k=0}^K, \mathbf{h}_t \in \mathbb{R}^K$ will be used to predict the next word by Softmax function with parameters \mathbf{W}_s and \mathbf{b}_s :

$$\mathcal{F}(\mathbf{p}_{ti}; \mathbf{W}_s, \mathbf{b}_s) = \frac{\exp(\mathbf{W}_s \mathbf{h}_{ti} + \mathbf{b}_s)}{\sum_{j=1}^K \exp(\mathbf{W}_s \mathbf{h}_{tj} + \mathbf{b}_s)}, \quad (7)$$

where \mathbf{p}_{ti} is the probability distribution for predicted word.

Our key motivation of chosen LSTM is that it can learn long-term temporal activities and avoid quick exploding and vanishing problems that traditional RNN suffers from during backpropagation optimization.

3.2 Bidirectional LSTM

In order to make use of both the past and future context information of a word in sentence prediction, we propose a bidirectional model by feeding a sentence to LSTM from forward and backward order. Figure 1 presents the overview of our model; it is comprised of three modules: a CNN for encoding image inputs, a Text-LSTM (T-LSTM) for encoding sentence inputs, and a Multimodal LSTM (M-LSTM) for embedding visual and textual vectors to a common semantic space and decoding to sentence. The bidirectional LSTM is implemented with two separate LSTM layers for computing forward hidden sequences $\vec{\mathbf{h}}$ and backward hidden sequences $\overleftarrow{\mathbf{h}}$. The forward LSTM starts at time $t = 1$ and the backward LSTM starts at time $t = T$. Formally, our model works as follows: for a given raw image input \tilde{I} , forward order sentence \vec{S} , and backward order sentence \overleftarrow{S} , the encoding performs as

$$\mathbf{I}_t = C(\tilde{I}; \Theta_v), \quad \vec{\mathbf{h}}_t^1 = \mathcal{T}(\vec{\mathbf{E}} \vec{S}; \vec{\Theta}_l), \quad \overleftarrow{\mathbf{h}}_t^1 = \mathcal{T}(\overleftarrow{\mathbf{E}} \overleftarrow{S}; \overleftarrow{\Theta}_l), \quad (8)$$

where C , \mathcal{T} represent CNN, T-LSTM, respectively, and Θ_v , Θ_l are their corresponding weights. Following previous work (Mao et al. 2015; Donahue et al. 2015), \mathbf{I}_t is considered at all timesteps as visual context information. $\vec{\mathbf{E}}$ and $\overleftarrow{\mathbf{E}}$ are bidirectional embedding matrices learned from network. Encoded visual and textual representations are then embedded to multimodal LSTM by

$$\vec{\mathbf{h}}_t^2 = \mathcal{M}(\vec{\mathbf{h}}_t^1, \mathbf{I}_t; \vec{\Theta}_m), \quad \overleftarrow{\mathbf{h}}_t^2 = \mathcal{M}(\overleftarrow{\mathbf{h}}_t^1, \mathbf{I}_t; \overleftarrow{\Theta}_m), \quad (9)$$

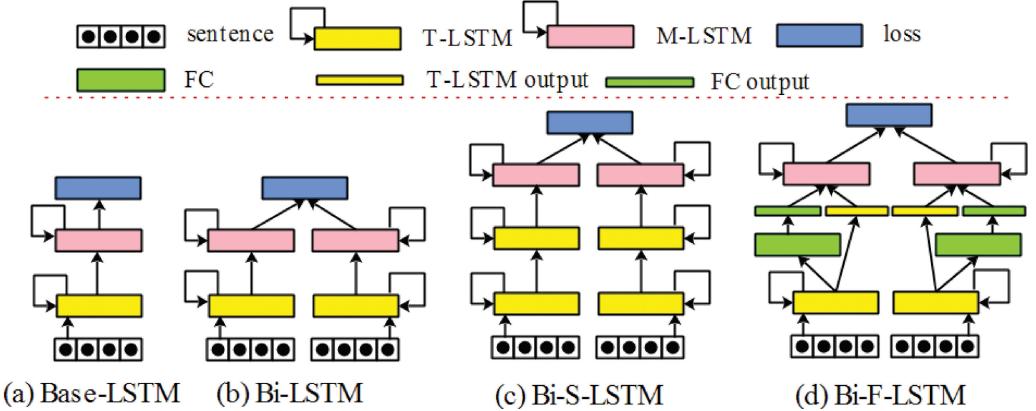


Fig. 4. Illustrations of proposed deep architectures for image captioning. The network in (a) is commonly used in previous work. (b) Our proposed Bidirectional LSTM (Bi-LSTM). (c) Our proposed Bidirectional Stacked LSTM (Bi-S-LSTM). (d) Our proposed Bidirectional LSTM with full connected (FC) transition layer (Bi-F-LSTM). T-LSTM receives text input only and M-LSTM receives both image and text input.

where \mathcal{M} presents M-LSTM and its weight Θ_m . \mathcal{M} aims to capture the correlation of visual context and words at different timesteps. We feed visual vector \mathbf{I}_t to the model at each timestep for capturing strong visual-word correlation. On the top of M-LSTM are Softmax layers with parameters \mathbf{W}_s and \mathbf{b}_s which compute the probability distribution of the next predicted word by

$$\overrightarrow{\mathbf{p}}_{t+1} = \mathcal{F}(\overrightarrow{\mathbf{h}}_t^2; \overrightarrow{\mathbf{W}}_s, \overrightarrow{\mathbf{b}}_s), \quad \overleftarrow{\mathbf{p}}_{t+1} = \mathcal{F}(\overleftarrow{\mathbf{h}}_t^2; \overleftarrow{\mathbf{W}}_s, \overleftarrow{\mathbf{b}}_s), \quad (10)$$

where $\mathbf{p} \in \mathbb{R}^K$ and K is the vocabulary size.

3.3 Deeper LSTM Architecture

The recent success of deep CNN in image classification and object detection (Krizhevsky et al. 2012; Simonyan and Zisserman 2014b) demonstrates that deep, hierarchical models can be more efficient at learning representation than shallower ones. This motivated our work to explore deeper LSTM architectures in the context of learning bidirectional visual-language embeddings. As claimed in Pascanu et al. (2013), if we consider LSTM as a composition of multiple hidden layers that unfolded in time, LSTM is already a deep network. But this is a way of increasing the “horizontal depth” in which network weights W are reused at each timestep and limited to learn more representative features such as increasing the “vertical depth” of the network. To design deep LSTM, one straightforward way is to stack multiple LSTM layers as a hidden-to-hidden transition. Alternatively, instead of stacking multiple LSTM layers, we propose to add multilayer perceptron (MLP) as an intermediate transition between LSTM layers. This can not only increase LSTM network depth, but can also prevent the parameter size from growing dramatically because the number of recurrent connections at a hidden layer can be largely decreased.

Directly stacking multiple LSTMs on top of each other leads to Bi-S-LSTM (Figure 4(c)). In addition, we propose to use a fully connected layer as an intermediate transition layer. Our motivation comes from the finding of Pascanu et al. (2013), in which DT(S)-RNN (deep transition RNN with shortcut) is designed by adding a hidden-to-hidden multilayer perceptron (MLP) transition. It is arguably easier to train such network. Inspired by this, we extend Bi-LSTM (Figure 4(b)) with a fully connected layer that we called Bi-F-LSTM (Figure 4(d)); a shortcut connection between the input and hidden states is introduced to make it easier to train the model. The aim of the extension

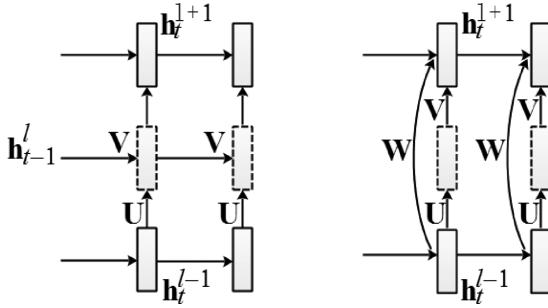


Fig. 5. Transition for Bi-S-LSTM (left) and Bi-F-LSTM (right).

models is to learn an extra hidden transition function F_h . Formally, in Bi-S-LSTM

$$\mathbf{h}_t^{l+1} = F_h(\mathbf{h}_t^{l-1}, \mathbf{h}_{t-1}^l) = \mathbf{U}\mathbf{h}_t^{l-1} + \mathbf{V}\mathbf{h}_{t-1}^l, \quad (11)$$

where \mathbf{h}_t^l presents the hidden states of the l -th layer at time t , and \mathbf{U} and \mathbf{V} are matrices connected to the transition layer (also see Figure 5 (left)). For readability, we consider one direction training and suppress bias terms. Similarly, in Bi-F-LSTM, to learn a hidden transition function F_h by

$$\mathbf{h}_t^{l+1} = F_h(\mathbf{h}_t^{l-1}) = \phi_r(\mathbf{W}\mathbf{h}_t^{l-1} \oplus (\mathbf{V}(\mathbf{U}\mathbf{h}_t^{l-1}))), \quad (12)$$

where \oplus is the operator that concatenates \mathbf{h}_t^{l-1} and its abstractions to a long hidden state (also see Figure 5 (right)). ϕ_r represents the rectified linear unit (Relu) activation function for transition layer, which performs $\phi_r(x) = \max(0, x)$.

3.4 Data Augmentation

One of the most challenging aspects of training deep bidirectional LSTM models is preventing overfitting. Since our largest dataset has only 80K images (Lin et al. 2014) which might cause overfitting easily, we adopted several techniques such as fine-tuning on a pre-trained visual model, weight decay, dropout, and early stopping that were commonly used in previous work. Additionally, it has been proved that data augmentation such as randomly cropping and horizontal mirror (Simonyan and Zisserman 2014a; Lu et al. 2014), adding noise, blur, and rotation (Wang et al. 2015) can effectively alleviate overfitting. Inspired by this, we designed new data augmentation techniques to increase the number of image-sentence pairs. Our implementation performs on a visual model, as follows:

- **Multi-Crop:** Instead of randomly cropping on input image, we crop at the four corners and center region because we found that random cropping tends to select center region and cause overfitting easily. By cropping four corners and center, the variations of network input can be increased to alleviate overfitting.
- **Multi-Scale:** To further increase the number of image-sentence pairs, we rescale input image to multiple scales. For each input image \tilde{I} with size $H \times W$, it is resized to 256×256 , then we randomly select a region with a size of $s * H \times s * W$, where $s \in [1, 0.925, 0.875, 0.85]$ is the scale ratio. $s = 1$ means we do not multi-scale operation on a given image. Finally, we resize it to AlexNet input size 227×227 or VGG-16 input size 224×224 .
- **Vertical Mirror:** Motivated by the effectiveness of the widely used horizontal mirror, it is natural to also consider the vertical mirror of image for the same purpose.

Those augmentation techniques are implemented in a real-time fashion. Each input image is randomly transformed using one of the augmentations to network input for training. In principle, our data augmentation can increase image-sentence training pairs by roughly 40 times ($5 \times 4 \times 2$). We report the evaluation of data augmentation in Section 4.4.

3.5 Multi-Task/Transfer Learning

Although our data augmentation can reduce overfitting in training deep LSTM network, it only helps to a certain extent. Increasing the effective training size with fresh training examples can further enlarge the variations of training data. This can effectively prevent training loss from going down quickly and reduce overfitting. On the other hand, it is also beneficial to increase the model robustness and generality. To address this issue, we propose to combine the training examples from different datasets; for example, in our case, $D_{multi} = D_{flickr8K} \cup D_{flickr30K} \cup D_{mscoco}$. With combined dataset D_{multi} , we train a multi-task joint model \mathcal{M}_{multi} , then we evaluate model performance on validation/test sets of different datasets, respectively.

In order to further test the generality and performance of multi-task joint model \mathcal{M}_{multi} in transferring knowledge learned on D_{multi} to new dataset, we propose to use \mathcal{M}_{multi} to perform image captioning and image-sentence retrieval on target dataset $D_{pascal1K}$. Here, we do not use any images from Pascal1K for training, only for validation. We report the evaluation of multi-task/transfer learning of Bi-LSTM in Section 4.5.

3.6 Training and Inference

Our model is end-to-end trainable by using Stochastic Gradient Descent (SGD). The joint loss function $L = \overrightarrow{L} + \overleftarrow{L}$ is computed by accumulating the Softmax losses of forward and backward directions. Our objective is to minimize L , which is equivalent to maximizing the probabilities of correctly generated sentences. We compute the gradient ∇L with the Back-Propagation Through Time (BPTT) algorithm (Werbos 1990).

The trained model is used to predict a word w_t with given image context I and previous word context $w_{1:t-1}$ by $P(w_t|w_{1:t-1}, I)$ in forward order, or by $P(w_t|w_{t+1:T}, I)$ in backward order. We set $w_1=w_T=0$ at the start point for forward and backward directions, respectively. Ultimately, with generated sentences from two directions, we decide the final sentence for a given image $p(w_{1:T}|I)$ according to the average of word probability within the sentence:

$$p(w_{1:T}|I) = \max \left(\frac{1}{T} \sum_{t=1}^T (\vec{p}(w_t|I)), \frac{1}{T} \sum_{t=1}^T (\overleftarrow{p}(w_t|I)) \right), \quad (13)$$

$$\vec{p}(w_t|I) = \prod_{t=1}^T p(w_t|w_1, w_2, \dots, w_{t-1}, I), \quad (14)$$

$$\overleftarrow{p}(w_t|I) = \prod_{t=1}^T p(w_t|w_{t+1}, w_{t+2}, \dots, w_T, I). \quad (15)$$

Following previous work, we adopted beam search to consider the best k candidate sentences at time t to infer the sentence at next timestep. In our work, we fix $k = 1$ on all experiments, although the average of 2 BLEU (Papineni et al. 2002) points out that better results can be achieved with $k = 20$ compared to $k = 1$ as reported in Vinyals et al. (2015).

4 EXPERIMENTS

In this section, we design several groups of experiments to accomplish the following objectives:

- Measure the benefits and performance of the proposed bidirectional model and its deeper variant models so that we increase their nonlinearity depth in different ways.
- Examine the influences of data augmentation and multi-task/transfer learning on bidirectional LSTM.
- Compare our approach with state-of-the-art methods in terms of sentence generation and image-sentence retrieval tasks on popular benchmark datasets.
- Qualitatively analyze and understand how bidirectional multimodal LSTM learns to generate a sentence conditioned by visual context information over time.

4.1 Datasets

To validate the effectiveness, generality, and robustness of our models, we conduct experiments on four benchmark datasets: Flickr8K (Hodosh et al. 2013), Flickr30K (Young et al. 2014), MSCOCO (Lin et al. 2014), and Pascal1K (Rashtchian et al. 2010) (used only for transfer learning experiment).

Flickr8K. It consists of 8,000 images and each of them has five sentence-level captions. We follow the standard dataset divisions provided by authors; 6,000/1,000/1,000 images for training/validation/testing, respectively.

Flickr30K. An extension version of Flickr8K. It has 31,783 images and each of them has five captions. We follow the publicly accessible² dataset division by Karpathy and Li (2015). In this dataset split, 29,000/1,000/1,000 images are used for training/validation/testing, respectively.

MSCOCO. This is a recent released dataset that covers 82,783 images for training and 40,504 images for validation. Each of the images has five sentence annotations. Since there is a lack of standard splits, we also follow the splits provided by Karpathy and Li (2015). Namely, 80,000 training images and 5,000 images for both validation and testing.

Pascal1K. This dataset is only used for evaluating the generalities of models in our transfer learning experiment. It is a subset of images from the PASCAL VOC challenge. It contains 1,000 images; each of them has five sentence descriptions. We do not use any images from this dataset for training. Following the protocol in Socher et al. (2014), we randomly selected 100 images for validation.

4.2 Implementation Details

Visual feature. We use two visual models for encoding images: Caffe (Jia et al. 2014) reference model which is pre-trained with AlexNet (Krizhevsky et al. 2012) and 16-layer VGG model (Simonyan and Zisserman 2014b). We extract features from the last fully connected layer and feed them to train the visual-language model with LSTM. Previous work (Vinyals et al. 2015; Mao et al. 2015) has demonstrated that more powerful image models such as GoogleNet (Szegedy et al. 2015) and ResNet (He et al. 2016) can achieve promising improvements. To make a fair comparison with recent works, we selected two widely used models for experiments.

Textual feature. We first represent each word w within a sentence as a one-hot vector, $w \in \mathbb{R}^K$, where K is the vocabulary size built on training sentences for a given dataset. By performing basic tokenization and removing the words that occur less than five times in the training set, we have 2,028, 7,400, and 8,801 words for Flickr8K, Flickr30K, and MSCOCO dataset vocabularies, respectively.

Our work uses the LSTM implementation of Donahue et al. (2015) on the Caffe framework. All of our experiments were conducted on Ubuntu 14.04, 16G RAM and single Titan X GPU with 12G memory. Our LSTMs use 1,000 hidden units and weights were initialized uniformly from $[-0.08, 0.08]$. The batch sizes are 150, 100, and 100 for Bi-LSTM, Bi-S-LSTM, and Bi-F-LSTM, respectively,

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

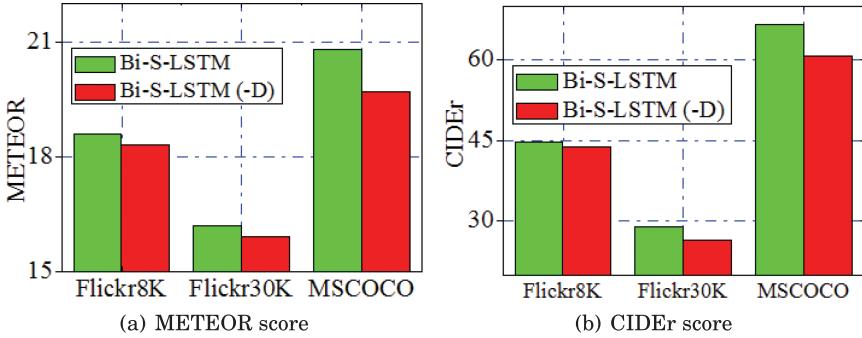


Fig. 6. METEOR/CIDEr scores on data augmentation.

when we use AlexNet as the visual model. When we use VGG as the visual model, the batch size is set to 32. Models are trained with learning rates $\eta = 0.01$ (AlexNet-based training) and $\eta = 0.005$ (VGG-based training), weight decay λ is 0.0005, and we used momentum 0.9. Each model is trained for 18–35 epochs with early stopping.

4.3 Evaluation Metrics

We evaluate our models mainly on caption generation; we follow previous work to use BLEU-N ($N=1,2,3,4$) scores (Papineni et al. 2002):

$$B_N = \min \left(1, e^{1 - \frac{r}{c}} \right) \cdot e^{\frac{1}{N} \sum_{n=1}^N \log p_n}, \quad (16)$$

where r, c represent the length of the reference sentence and the generated sentence, respectively, and p_n is the modified n -gram precisions. We also report the METEOR (Lavie 2014) and CIDEr (Vedantam et al. 2015) scores for further comparison. To evaluate the generality of our models, we conduct a transfer learning experiment using Pascal1K on image-sentence retrieval³ (image query sentence and vice versa). It performs by computing the score of each image-sentence pair, and ranking the scores to obtain the top-K ($K = 1, 5, 10$) retrieved results. We adopt R@K and Mean r as the evaluation metrics. R@K is the recall rate R at top K candidates and Mean r is the mean rank. All mentioned metric scores are computed by the MSCOCO caption evaluation server,⁴ which is commonly used for image captioning challenge.⁵

4.4 Experiments on Data Augmentation

In this subsection, we design a group of experiments to examine the effects of utilized data augmentation techniques. To this end, we use Bi-S-LSTM for experiment, because it has deeper LSTM and we believe that training a deeper LSTM network on limited data is more challenging and helpful to measure the benefits brought by data augmentation. In this experiment, we turn off the introduced augmentation techniques in Section 3.4 and keep other configurations unchanged. The BLEU performance is reported in Table 1 and Table 2; METEOR/CIDEr performance is reported in Figure 6 (shown as $\text{Bi-S-LSTM}^{A,-D}$). It is clear to see that without using data augmentation, the model performance drops significantly on all metrics. Those results also reveal how data augmentation affects datasets at different scales. For example, the model performance on small-scale

³Although this work focuses on image captioning task, we conduct an image-sentence retrieval experiment here to examine the generality of our models across datasets and tasks. The task has been discussed widely in our previous work (Wang et al. 2016d).

⁴<https://github.com/tylin/coco-caption>.

⁵<http://mscoco.org/home/>.

Table 1. BLEU-N Performance Comparison on Flickr8K and Flickr30K (High Score is Good)

Models	Flickr8K				Flickr30K			
	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
NIC (Vinyals et al. 2015) ^{G,‡}	63	41	27.2	-	66.3	42.3	27.7	18.3
X. Chen et al. (Chen and Zitnick 2014)	-	-	-	14.1	-	-	-	12.6
LRCN (Donahue et al. 2015) ^{A,‡}	-	-	-	-	58.8	39.1	25.1	16.5
DeepVS (Karpathy and Li 2015) ^V	57.9	38.3	24.5	16	57.3	36.9	24.0	15.7
m-RNN (Mao et al. 2015) ^{A,‡}	56.5	38.6	25.6	17.0	54	36	23	15
m-RNN (Mao et al. 2015) ^{V,‡}	-	-	-	-	60	41	28	19
Hard-Attention (Xu et al. 2015) ^V	67	45.7	31.4	21.3	66.9	43.9	29.6	19.9
ATT-FCN (You et al. 2016) ^G	-	-	-	-	64.7	46.0	32.4	23.0
C. Wang et al. (Wang et al. 2016d) ^V	65.5	46.8	32.0	21.5	62.1	42.6	28.1	19.3
Bi-LSTM ^A	63.7	44.7	31	20.9	61.0	40.9	27.1	18.1
Bi-S-LSTM ^A	65.1	45.0	29.3	18.4	60.0	40.3	27.1	18.2
Bi-F-LSTM ^A	63.9	44.6	30.2	19.9	60.7	41.0	27.5	18.5
Bi-LSTM ^V	66.7	48.3	33.7	23	63.3	44.1	29.6	20.1
Bi-S-LSTM ^V	66.9	48.8	<u>33.3</u>	<u>22.8</u>	63.6	<u>44.8</u>	<u>30.4</u>	<u>20.5</u>
Bi-F-LSTM ^V	66.5	48.4	32.8	22.4	63.4	44.3	30.1	20.4
Bi-LSTM ^{A,+M}	58.4	42.1	28.6	18.2	61.0	41.4	27.8	18.5
Bi-S-LSTM ^{A,-D}	55.4	38.0	24.6	15.3	58.2	39.0	25.1	16.3

The superscript “A” means the visual model is AlexNet (or similar network), “V” is VGG-16, “G” is GoogleNet, “-D” means without using data augmentations in Section 3.4, “+M” means using multi-task learning in Section 3.5, “-” indicates unknown value, “‡” means different data splits.⁶ The best results are marked in **bold** and the second best results with an underline (the superscripts are also applicable to Tables 2, 3, and 4).

dataset Flickr8K is worse than that on Flickr30K and MSCOCO. This confirms that data augmentation is beneficial in preventing overfitting and particularly helpful on small-scale dataset.

4.5 Experiments on Multi-Task/Transfer Learning

In addition to using data augmentation to increase the variations of training examples and reduce overfitting, another effective way should be multi-task learning. Inspired by Simonyan and Zisserman (2014a) and Donahue et al. (2015) in which datasets were combined to train a joint model, we combine the training set of Flickr8K, Flickr30K, and MSCOCO in order to increase the number of training examples. Then we train a multi-task joint model with combined training sets and evaluate on each validation set to examine its performance and generality. To save training time, we initialize the training of the multi-task joint model with the best-performing pre-trained MSCOCO model. We change the input unit numbers of embedding layers and the output units number of the last fully connected layer; they are the vocabulary size (11,557) of the combined training set.

To compare with baseline models without using multi-task learning, we select the best-performing models⁷ for Flickr8K, Flickr30K, MSCOCO, and Pascal1K datasets, respectively. The comparison with baseline models in terms of BLEU scores is reported in Table 1 and Table 2. The results show that the multi-task joint model (shown as Bi-LSTM^{A,+M}) did not improve the BLEU

⁶On the MSCOCO dataset, NIC uses 4K images for validation and test. LRCN randomly selects 5K images from MSCOCO validation set for validation and test. m-RNN uses 4K images for validation and 1K as test.

⁷The model from 100000th iterations has the best performance on Flickr8K, the model from 90000th iterations performs best on the rest of datasets.

Table 2. BLEU-N, METEOR and CIDEr Performance Comparison on MSCOCO

Models	MSCOCO					
	B-1	B-2	B-3	B-4	METEOR	CIDEr
NIC (Vinyals et al. 2015) ^{G,‡}	66.6	46.1	32.9	24.6	-	-
X. Chen et al. (Chen and Zitnick 2014)	-	-	-	19.0	20.4	-
LRCN (Donahue et al. 2015) ^{A,‡}	62.8	44.2	30.4	-	-	-
DeepVS (Karpathy and Li 2015) ^V	62.5	45	32.1	23	19.5	66.0
m-RNN (Mao et al. 2015) ^{V,‡}	67	49	35	25	-	-
Hard-Attention (Xu et al. 2015) ^V	71.8	50.4	35.7	25	<u>23.0</u>	-
ATT-FCN (You et al. 2016) ^G	<u>70.9</u>	<u>53.7</u>	<u>40.2</u>	<u>30.4</u>	<u>24.3</u>	-
C. Wang et al. (Wang et al. 2016d) ^V	67.2	49.2	35.2	24.4	21.6	71.0
Bi-LSTM ^A	65.1	45.0	29.3	18.4	20.0	64.1
Bi-S-LSTM ^A	64.1	45.4	31.3	21.1	20.7	68.1
Bi-F-LSTM ^A	64.0	45.5	31.5	21.5	20.5	67.5
Bi-LSTM ^V	68.5	50.5	36.0	25.3	22.1	73.0
Bi-S-LSTM ^V	68.7	<u>50.9</u>	<u>36.4</u>	<u>25.8</u>	22.9	73.9
Bi-F-LSTM ^V	68.2	50.6	36.1	25.6	22.6	73.5
Bi-LSTM ^{A,+M}	65.6	47.4	33.3	23.0	21.1	69.5
Bi-S-LSTM ^{A,-D}	62.8	44.4	30.2	20.0	19.7	60.7

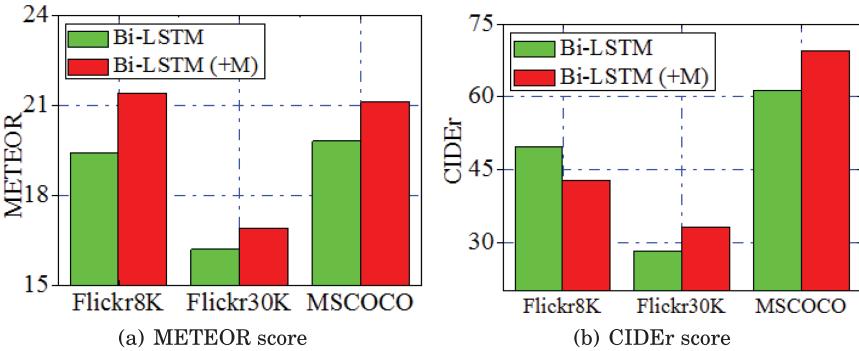


Fig. 7. METEOR/CIDEr scores on multi-task learning.

score on small dataset Flickr8K. We conjecture the reason is, on the one hand, multi-task joint model helps to make diversity of training examples and increase model generality. On the other hand, it enlarges the differences between training and validation data. Those factors lead to worse BLEU performance on Flickr8K even though the generated sentences are highly descriptive and sensible (also see examples in Figure 11). However, the multi-task joint model shows promising improvements on Flickr30K and MSCOCO. In addition, in Table 1 and Table 2 we also found that the multi-task joint model tends to improve B-2, B-3, and B-4 performance (2.4, 4.0, and 4.6 points increased on MSCOCO). In Figure 7, The METEOR/CIDEr performance is improved with multi-task learning except METETOR on Flickr8K.

4.6 Comparison with State-of-The-Art Methods

4.6.1 Model Performance. Now we compare with state-of-the-art methods. Table 1 and Table 2 summarize the comparison results in terms of BLEU-N. Our approach achieves very competitive

Table 3. Image Captioning Performance Comparison on Pascal1K

Methods	BLEU	METEOR
Midge (Mitchell et al. 2012)	2.89	8.80
Baby talk (Kulkarni et al. 2011)	0.49	9.69
RNN (Chen and Zitnick 2014)	2.79	10.08
RNN+IF (Chen and Zitnick 2014)	10.16	16.43
RNN+IF+FT (Chen and Zitnick 2014)	10.18	16.45
X.Chen et al. (Chen and Zitnick 2014)	10.48	16.69
X.Chen et al.+FT (Chen and Zitnick 2014)	10.77	16.87
Bi-LSTM ^A (transfer)	16.4	18.30

performance on evaluated datasets, although with a less powerful visual model—AlexNet. Increasing the depth of LSTM is beneficial on generation task. Deeper variant models mostly obtain better performance compared to Bi-LSTM, but they are inferior to the latter one in B-3 and B-4 on Flickr8K. We believe it should be the reason that Flickr8K is a relatively small dataset which suffers difficulty in training deep models with limited data. One of the interesting facts we found is that stacking multiple LSTM layers is generally superior to LSTM with a fully connected transition layer, although Bi-S-LSTM needs more training time. Replacing AlexNet with VGG-16 results in significant improvement on all BLEU evaluation metrics. We should be aware that a recent interesting work (Xu et al. 2015) achieves the best results on B-1 by integrating an attention mechanism (LeCun et al. 2015; Xu et al. 2015). Semantic attention (You et al. 2016) with GoogleNet achieves the best performance on B-2, B-3, and B-4.

Regarding METEOR and CIDEr performance, our baseline model (Bi-LSTM^A) outperforms DeepVS^V (Karpathy and Li 2015) in a certain margin. It achieves 19.1/51.8 on Flickr8K (compare to 16.7/31.8 of DeepVS^V) and 16.1/29.0 on Flickr30K (15.3/24.7 of DeepVS^V). On MSCOCO, our best results are 22.9/73.9; the METEOR score is slightly inferior to 23.0 in Xu et al. (2015) and 24.3 in You et al. (2016) but exceeds the rest of the methods. Although we believe incorporating an attention mechanism into our framework can make further improvements, note that our current model achieves competitive results while the small gap between our model and the attention-based model (Xu et al. 2015; You et al. 2016) existed.

Comparing to our prior work (Wang et al. 2016d), we use the mean probability in Equation (13), rather than the sum probability of all words when selecting the final caption from the bidirectionally generated captions. This slightly improves our model performance on nearly all metrics by an average 1.7 points on Flickr8K, 1.2 points on Flickr30K, and 1.1 points on MSCOCO.

4.6.2 Model Generality. In order to further evaluate the generality of our model on image captioning, we test our joint model on the Pascal1K validation dataset. Table 3 presents the comparison with related work on BLEU and METEOR. We can see that even with our base model Bi-LSTM^A, the performance on generation task exceeds previous approaches in a certain margin even without using the training images of Pascal1K.

On the same dataset, we also examine the generality of our model on a different task: image-sentence retrieval. The results are reported in Table 4. It shows that without using any training images from Pascal1K, our model substantially outperforms previous work in all metrics. Particularly on R@1, transfer learning achieves more than 20 points on both image-to-sentence and sentence-to-image retrieval tasks.

Those experiments demonstrate that although with less powerful visual model, our simplest network (Bi-LSTM) achieves the best performance on both image captioning and image-sentence retrieval tasks.

Table 4. Image-Sentence Retrieval Performance Comparison on Pascal1K

Methods	Image to Sentence				Sentence to Image			
	R@1	R@5	R@10	M_r	R@1	R@5	R@10	M_r
Random Ranking	4.0	9.0	12.0	71.0	1.6	5.2	10.6	50.0
KCCA (Socher et al. 2014)	21.0	47.0	61.0	18.0	16.4	41.4	58.0	15.9
DeViSE (Frome et al. 2013)	17.0	57.0	68.0	11.9	21.6	54.6	72.4	9.5
SDT-RNN (Socher et al. 2014)	25.0	56.0	70.0	13.4	35.4	65.2	84.4	7.0
DeepFE (Karpathy et al. 2014)	39.0	68.0	79.0	10.5	23.6	65.2	79.8	7.6
RNN+IF (Chen and Zitnick 2014)	31.0	68.0	87.0	6.0	27.2	65.4	79.8	7.0
X. Chen et al. (Chen and Zitnick 2014)	25.0	71.0	86.0	5.4	28.0	65.4	82.2	6.8
X. Chen et al. (T+I) (Chen and Zitnick 2014)	30.0	75.0	87.0	5.0	28.0	67.4	83.4	6.2
Bi-LSTM ^A (transfer)	65.0	90.0	95.0	2.0	52.8	86.0	95.4	2.1

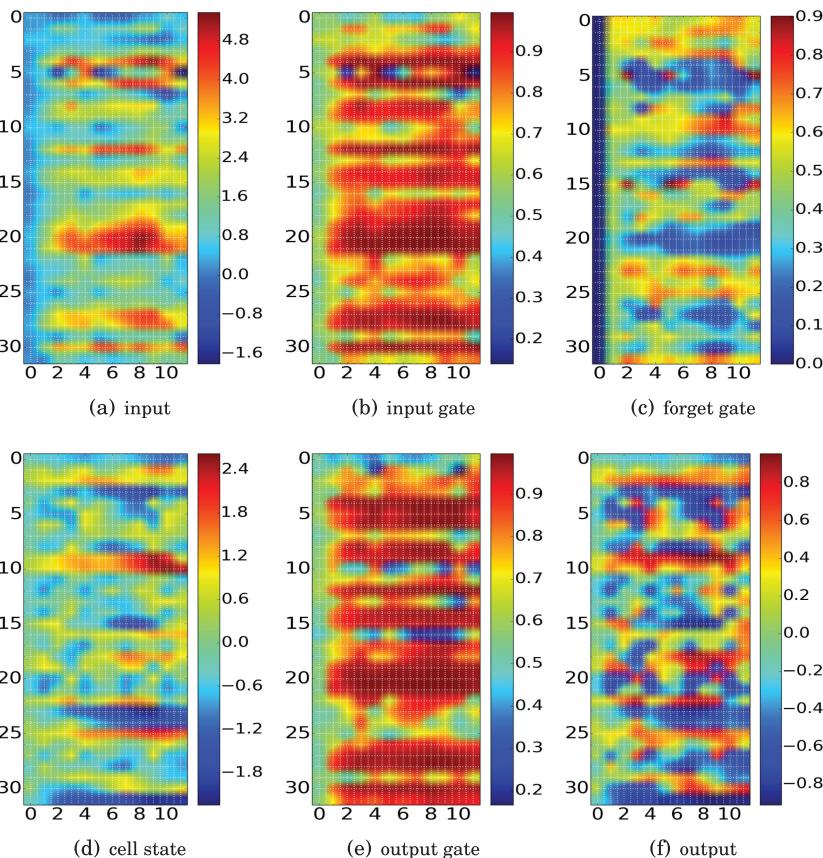


Fig. 8. Visualization of LSTM cell. The horizontal axis corresponds to timesteps. The vertical axis is cell index. Here we visualize the gates and cell states of the first 32 Bi-LSTM units of T-LSTM in forward direction over 11 timesteps.

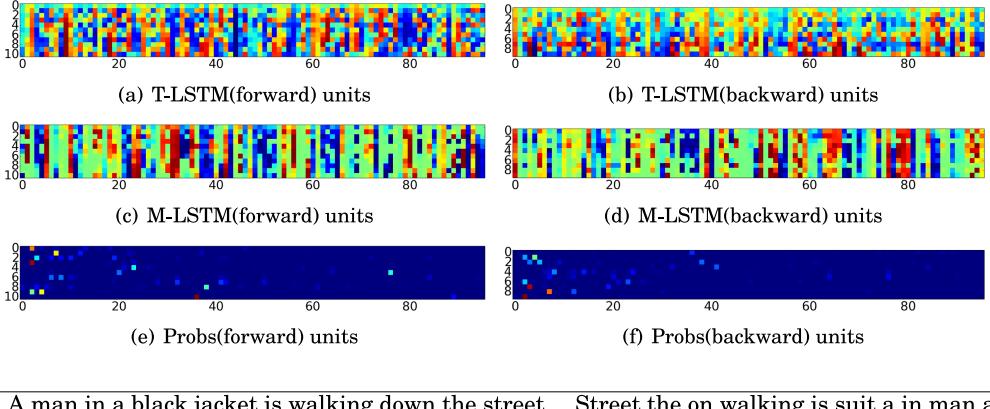


Fig. 9. Pattern of the first 96 hidden units chosen at each layer of Bi-LSTM in both forward and backward directions. The vertical axis presents timesteps. The horizontal axis corresponds to different LSTM units. In this example, we visualize the T-LSTM layer for text only, the M-LSTM layer for both text and image, and the Softmax layer for word prediction. The model was trained on Flickr 30K dataset for generating a sentence word by word at each timestep. In (g), we provide the predicted words at different timesteps and their corresponding index in vocabulary where we can also read from (e) and (f) (the highlight point at each row). Word with highest probability is selected as the predicted word.

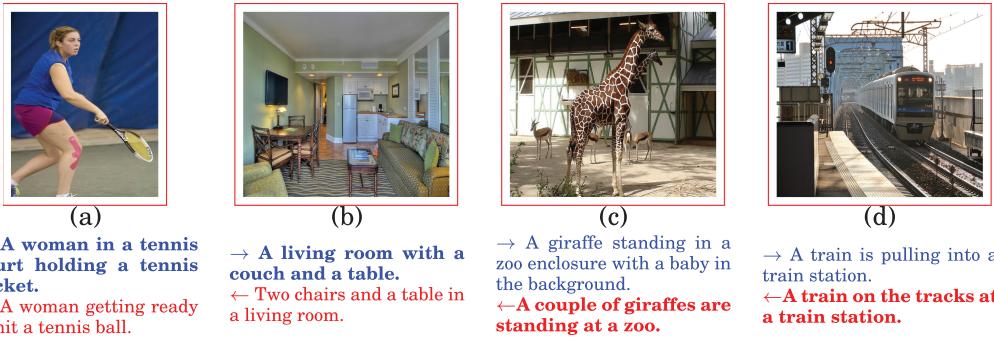


Fig. 10. Examples of generated captions for a given query image on MSCOCO validation set. Blue captions are generated in forward direction and red captions are generated in backward direction. The final caption is selected according to Equation (13) which selects the sentence with the higher mean probability. The final captions are marked in **bold**.

4.7 Visualization and Qualitative Analysis

The aim of this set experiment is to visualize the properties of the proposed bidirectional LSTM model and explain how it works in generating a sentence word by word over time.

First, we examine the temporal evolution of internal gate states and understand how bidirectional LSTM units retain valuable context information and attenuate unimportant information. Figure 8 shows input and output data, the pattern of three sigmoid gates (input, forget, and output), as well as cell states. We can clearly see that dynamic states are periodically distilled to units from timestep $t = 0$ to $t = 11$. At $t = 0$, the input data are sigmoid modulated to input gate $i(t)$

Flickr8K		\rightarrow A man in a blue shirt be ride a bicycle on a street \leftarrow A man be ride on a bicycle \rightarrow^M A group of people riding bikes down a street \leftarrow^M A group of people riding bikes down a street
		\rightarrow A man in a blue shirt be ride a bicycle on a street \leftarrow A man be ride on a bicycle \rightarrow^M A group of people riding bikes down a street \leftarrow^M A group of people riding bikes down a street
Flickr30K		\rightarrow A man in a blue wetsuit is surfing in the ocean \leftarrow A man is surfing on a wave in the water \rightarrow^M A person on a surfboard in the ocean \leftarrow^M A man is surfing in the ocean
		\rightarrow A group of people are playing soccer \leftarrow A group of children are playing in the grass \rightarrow^M A group of young boys playing soccer \leftarrow^M A boy be try to kick a soccer ball
MSCOCO		\rightarrow A fire hydrant sitting on a street next to a street \leftarrow A yellow fire hydrant on the side of a street \rightarrow^M A fire hydrant on a sidewalk next to a street \leftarrow^M A fire hydrant on the side of the street
		\rightarrow A double decker bus driving down a street \leftarrow A double decker bus parked on the side of a street \rightarrow^M A double decker bus is parked in a parking lot \leftarrow^M A double decker bus driving down the street
Pascal1K		\rightarrow^T A black and white cat is looking out of a window \leftarrow^T A close up of a cat looking out of a window
		\rightarrow^T A train is parked in front of a train station \leftarrow^T A man that is standing in front of a train

Fig. 11. Examples of generated captions for given query images on Flickr8K, Flickr30K, MSCOCO, and Pascal1K validation set. Left: input images. Right: \rightarrow and \leftarrow present the generated captions in forward and backward direction, respectively. The superscript *M* or *T* means the captions generated with multi-task or transfer learning. The final captions are marked in **bold**.

where values lie within in [0,1]. At this step, the values of forget gates $f(t)$ of different LSTM units are zeros. Along with the increasing of timestep, forget gate starts to decide which unimportant information should be forgotten, and meanwhile, decide to retain useful information. Then the memory cell states $c(t)$ and output gate $o(t)$ gradually absorb the valuable context information over time and make a rich representation $h(t)$ of the output data.

Next, we examine how visual and textual features are embedded to common semantic space and used to predict words over time. Figure 9 shows the evolution of hidden units at different layers. For the T-LSTM layer where LSTM units are conditioned by textual context from the past

and future, It performs as the encoder of forward and backward sentences. At the M-LSTM layer, LSTM units are conditioned by both visual and textual context. It learns the correlations between input word sequence and visual information that were encoded by CNN. At a given timestep, by removing unimportant information that makes less contribution to correlate input word and visual context, the units tend to appear sparsity pattern and learn more discriminative representations from inputs. At higher layer, embedded multimodal representations are used to compute the probability distribution of next predict word with Softmax. It should be noted, for a given image, the number of words in a generated sentence from forward and backward direction can be different.

Figure 10 presents some example images with generated captions. From generated captions, we found bidirectionally generated captions cover different semantic information; for example, in (b) the forward sentence captures “couch” and “table” while the backward one describes “chairs” and “table.” We also found that a significant proportion (88% by randomly selected 1,000 images on MSCOCO validation set) of generated sentences are novel (do not appear in training set). But generated sentences are highly similar to ground-truth captions; for example, in (d), forward caption is similar to one of the ground-truth captions (“A passenger train that is pulling into a station”) and the backward caption is similar to the ground-truth caption (“a train is in a tunnel by a station”). It illustrates that our model has a strong capability in learning visual-language correlation and generates novel sentences.

More example sentence generations on Flickr8K, Flickr30K, MSCOCO, and Pascal1K can be found in Figure 11. Those examples demonstrate that without using an explicit pre-trained language model on additional corpus, our models generate sentences which are highly descriptive and semantically relevant to corresponding images.

5 CONCLUSIONS

We proposed a bidirectional LSTM model that generates a descriptive sentence for an image by taking both history and future context into account. We further designed deep bidirectional LSTM architectures to embed image and sentence at high semantic space for learning visual-language model. We proved multi-task learning of Bi-LSTM is beneficial to increase model generality and further confirmed by transfer learning experiment. We also qualitatively visualized internal states of the proposed model to understand how multimodal bidirectional LSTM generates words at consecutive timesteps. The effectiveness, generality, and robustness of the proposed models were evaluated with numerous datasets on two different tasks: image captioning and image-sentence retrieval. Our models achieve highly competitive results on both tasks. Our future work will focus on exploring more sophisticated language representation (e.g., word2vec) and incorporating an attention mechanism into our model. It would also be interesting to explore the multilingual caption generation problem. We also plan to apply our models to other sequence learning tasks such as text recognition and video captioning.

REFERENCES

- D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR 2015*.
- Rich Caruana. 1998. Multitask learning. In *Learning to Learn*. Springer, 95–133.
- Xinlei Chen and C. Lawrence Zitnick. 2014. Learning a recurrent visual representation for image caption generation. *arXiv:1411.5654*.
- X. Chen and C. Lawrence Zitnick. 2015. Mind’s eye: A recurrent visual representation for image caption generation. In *CVPR*. 2422–2431.
- Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. 2015. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia* 17, 11 (2015), 1875–1886.
- K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014*.

- Junyoung Chung, Caglar Gülcöhre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *ICML*. 2067–2075.
- J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*. 2625–2634.
- H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, and J. Platt. 2015. From captions to visual concepts and back. In *CVPR*. 1473–1482.
- A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*. 2121–2129.
- Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*.
- Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. 2009. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 5 (2009), 855–868.
- A. Graves, A. Mohamed, and G. E. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *ICASSP*. IEEE, 6645–6649.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.
- Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACMMM*. ACM, 675–678.
- A. Karpathy, A. Joulin, and F.-F. Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*. 1889–1897.
- A. Karpathy and F.-F. Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*. 3128–3137.
- R. Kiros, R. Salakhutdinov, and R. Zemel. 2014a. Multimodal neural language models. In *ICML*. 595–603.
- R. Kiros, R. Salakhutdinov, and R. Zemel. 2014b. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539*.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.
- Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, 1601–1608.
- G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 35, 12 (2013), 2891–2903.
- P. Kuznetsova, V. Ordonez, A. C. Berg, T. Berg, and Y. Choi. 2012. Collective generation of natural image descriptions. In *ACL*, Vol. 1. ACL, 359–368.
- P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi. 2014. TREETALK: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics (TACL)* 2, 10 (2014), 351–362.
- M. Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. *ACL* (2014), 376.
- Y. LeCun, Y. Bengio, and G. E. Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
- S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *CoNLL*. ACL, 220–228.
- T-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer, 740–755.
- Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 457–466.
- J. H. Mao, W. Xu, Y. Yang, J. Wang, Z. H. Huang, and A. Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR* 2015.
- T. Mikolov, M. Karafiat, L. Burget, J. Černocký, and S. Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*. 1045–1048.
- T. Mikolov, S. Kombrink, L. Burget, J. H. Černocký, and S. Khudanpur. 2011. Extensions of recurrent neural network language model. In *ICASSP*. IEEE, 5528–5531.
- M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *ACL*. ACL, 747–756.
- J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. 2011. Multimodal deep learning. In *ICML*. 689–696.

- Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.
- K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*. ACL, 311–318.
- R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. 2013. How to construct deep recurrent neural networks. *arXiv:1312.6026*.
- C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. 2010. Collecting image annotations using Amazon’s Mechanical Turk. In *NAACL HLT Workshop*. Association for Computational Linguistics, 139–147.
- Nikhil Rasiwasia, Pedro J. Moreno, and Nuno Vasconcelos. 2007. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia* 9, 5 (2007), 923–938.
- K. Simonyan and A. Zisserman. 2014a. Two-stream convolutional networks for action recognition in videos. In *NIPS*. 568–576.
- K. Simonyan and A. Zisserman. 2014b. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.
- R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics (TACL)* 2 (2014), 207–218.
- Richard Socher, Cliff C. Lin, Chris Manning, and Andrew Y. Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *28th International Conference on Machine Learning (ICML’11)*. 129–136.
- N. Srivastava and R. Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *NIPS*. 2222–2230.
- I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*. 3104–3112.
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *CVPR*. 1–9.
- R. Vedantam, Z. Lawrence, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*. 4566–4575.
- O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. 3156–3164.
- Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016d. Image captioning with deep bidirectional LSTMs. *arXiv:1604.00790*.
- Cheng Wang, Haojin Yang, and Christoph Meinel. 2016c. A deep semantic framework for multimodal representation learning. *Multimedia Tools and Applications* (2016), 1–22.
- Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, and Nicu Sebe. 2016a. Recurrent face aging. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2378–2386.
- Wei Wang, Sergey Tulyakov, and Nicu Sebe. 2016b. Recurrent convolutional face alignment. In *Asian Conference on Computer Vision*. Springer, 104–120.
- Zhangyang Wang, Jianchao Yang, Hailin Jin, Eli Shechtman, Aseem Agarwala, Jonathan Brandt, and Thomas S. Huang. 2015. DeepFont: Identify your font from an image. In *23rd ACM International Conference on Multimedia (MM’15)*. ACM, New York, 451–459. DOI: <http://dx.doi.org/10.1145/2733373.2806219>
- Paul J. Werbos. 1990. Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE* 78, 10 (1990), 1550–1560.
- Zhizheng Wu and Simon King. 2016. Investigating gated recurrent networks for speech synthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’16)*. IEEE, 5140–5144.
- K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *ICML 2015*.
- Xuyong Yang, Tao Mei, Ying-Qing Xu, Yong Rui, and Shipeng Li. 2016. Automatic generation of visual-textual presentation layout. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 12, 2 (2016), 33.
- Xiaoshan Yang, Tianzhu Zhang, and Changsheng Xu. 2015. Cross-domain feature learning in multimedia. *IEEE Transactions on Multimedia* 17, 1 (2015), 64–78.
- Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*. 4651–4659.
- P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)* 2 (2014), 67–78.
- W. Zaremba and I. Sutskever. 2014. Learning to execute. *arXiv:1410.4615*.
- M. D. Zeiler and R. Fergus. 2014. Visualizing and understanding convolutional networks. In *ECCV*. Springer, 818–833.

Received December 2016; revised March 2017; accepted March 2017