Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei
Microsoft Research, Beijing, China
University of Science and Technology of China, Hefei, China
Sun Yat-Sen University, Guangzhou, China
{tiyao, tmei}@microsoft.com, {panyw.ustc, yehaoli.sysu}@gmail.com

Abstract

Image captioning often requires a large set of training image-sentence pairs. In practice, however, acquiring sufficient training pairs is always expensive, making the recent captioning models limited in their ability to describe objects outside of training corpora (i.e., novel objects). In this paper, we present Long Short-Term Memory with Copying Mechanism (LSTM-C) — a new architecture that incorporates copying into the Convolutional Neural Networks (CNN) plus Recurrent Neural Networks (RNN) image captioning framework, for describing novel objects in captions. Specifically, freely available object recognition datasets are leveraged to develop classifiers for novel objects. Our LSTM-C then nicely integrates the standard word-by-word sentence generation by a decoder RNN with copying mechanism which may instead select words from novel objects at proper places in the output sentence. Extensive experiments are conducted on both MSCOCO image captioning and ImageNet datasets, demonstrating the ability of our proposed LSTM-C architecture to describe novel objects. Furthermore, superior results are reported when compared to state-of-the-art deep models.

1. Introduction

Automatically describing the content of an image with a complete and natural sentence, a problem known as image captioning, has great potential impact for instance on robotic vision or helping visually impaired people. Intensive research interests from both computer vision and natural language processing communities have been paid for this emerging topic. Most of recent attempts on this problem [4, 23, 26, 29] are Convolutional Neural Networks (CNN) plus Recurrent Neural Networks (RNN) based sequence learning methods, which are mainly inspired from the advances by using RNN in machine translation [21]. The basic idea is an encoder-decoder mechanism for translation. Specifically, a CNN is employed to encode image content and then a decoder RNN is exploited to generate a natural sentence. While encouraging performances are reported, the sequence learning methods learn directly from image and sentence pairs, which fail in their ability to describe the objects out of the training data, i.e., novel objects. Take the image in Figure 1 as an example, the output sentence generated by a popular image captioning method Long-term Recurrent Convolutional Networks (LRCN) [4] is unable to describe "suitcase" as this object is non-existent in the training corpora. More importantly, manually labeling a large-scale image captioning dataset is an intellectually expensive and time-consuming process.

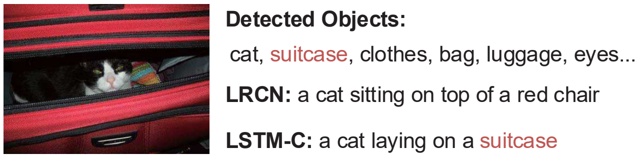

Detected Objects:
cat, suitcase, clothes, bag, luggage, eyes...
LRCN: a cat sitting on top of a red chair
LSTM-C: a cat laying on a suitcase

Figure 1. An example of object recognition and image captioning. The input is an image, while the output is the detected objects and a natural sentence, respectively. (upper row: the detected objects in the image; middle row: the sentence generated by LRCN [4] image captioning approach; bottom row: the sentence generated by our LSTM-C model.)

We demonstrate in this paper that the above limitations could be mitigated by incorporating the knowledge from external visual recognition datasets, which are freely available for developing object detectors. Specifically, we present a novel Long Short-Term Memory with Copying Mechanism (LSTM-C) framework to generate words by integrating "copying mechanism." Copying mechanism is originated from human language communication and refers to the mechanism that locates a certain segment of the input sequence and directly puts the segment in the output sequence [7]. The spirit behind is the rote memorization in language processing of human being, which needs to refer to sub-sequences of the input. We extend the copying idea here to select novel objects learnt from external sources and put them at proper places in the generated sentence. The overview of LSTM-C framework is illustrated in Figure 2. Given an image, a CNN is utilized to extract visual features, which will be fed into LSTM at the initial time step for sentence generation. Meanwhile, the objects of the input image are also predicted by object detectors pre-trained on recognition dataset. A copying layer is devised at the top of the whole architecture to accommodate the generative model of LSTM and copying mechanism from the detected objects. By integrating copying mechanism into image captioning, the word "suitcase" is copied from detected objects and output in the sentence generated by our LSTM-C as shown in Figure 1. The whole architecture is trained end-to-end.

The main contribution of this work is the proposal of LSTM-C framework by incorporating the knowledge from external sources to address the issue of predicting novel objects in image captioning task. This issue also leads to an elegant view of how to accommodate both generative model and copying mechanism from detected objects for sentence generation, which is a problem not yet fully understood.

2. Related Work

We briefly group the related works into two categories: image captioning and novel object captioning. The first category reviews the research in sentence generation for images, while the second investigates a variety of recent models which attempt to describe novel objects in context.

2.1. Image Captioning

The research on image captioning has proceeded along three different dimensions: template-based methods [11, 14, 27], search-based approaches [3, 6, 15], and language-based models [4, 10, 23, 24, 26, 28, 29].

Template-based methods predefine the template for sentence generation and split sentence into several parts (e.g., subject, verb, and object). With such sentence fragments, many works align each part with visual content (e.g., CRF in [11] and HMM in [27]) and then generate the sentence for the image. Obviously, most of them highly depend on the templates of sentence and always generate sentence with syntactical structure. Search-based approaches [3, 6, 15] "generate" sentence for an image by selecting the most semantically similar sentences from sentence pool. This direction indeed can achieve human-level descriptions as all the output sentences are from existing human-generated ones. The need to collect human-generated sentences, however, makes the sentence pool hard to be scaled up.

Different from template-based and search-based models, language-based models aim to learn the probability distribution in the common space of visual content and textual sentence to generate novel sentences with more flexible syntactical structures. In this direction, recent works explore such probability distribution mainly using neural networks and have achieved promising results for image captioning task. Kiros *et al.* [10] employ the neural networks to generate sentence for an image by proposing a multimodal log-bilinear neural language model. In [23], Vinyals *et al.* propose an end-to-end neural networks architecture by utilizing LSTM to generate sentence for an image, which is further incorporated with attention mechanism in [26] to automatically focus on salient objects when generating corresponding words. More recently, in [24], high-level concepts/attributes are shown to obtain clear improvements on image captioning task when injected into existing state-of-the-art RNN-based model. Such high-level attributes are further utilized as semantic attention in [29] and complementary representations to visual features in [17, 28] to enhance image/video captioning.

2.2. Novel Object Captioning

The novel object captioning is a new problem that has received increasing attention most recently, which leverages additional image-sentence paired data [13] or unpaired image/text data [8, 22] to describe novel objects in existing RNN-based image captioning frameworks. [13] is one of the early works that enlarges the original limited word dictionary to describe novel objects by using only a few paired image-sentence data. In particular, a transposed weight sharing scheme is proposed to avoid extensive retraining. In contrast, with the largely available unpaired image/text data (e.g., ImageNet and Wikipedia), Hendricks *et al.* [8] explicitly transfer the knowledge of semantically related objects to compose the descriptions about novel objects in the proposed Deep Compositional Captioner (DCC). The DCC model is further extended to an end-to-end system by simultaneously optimizing the visual recognition network, LSTM-based language model, and image captioning network with different sources in [22].

Our model mainly focuses on the latter scenario, that incorporates the knowledge learnt from freely available unpaired object recognition data for novel object captioning. Different from previous methods which solely rely on the standard word-by-word sentence generation through a decoder RNN, we integrate the regular decoder RNN with copying mechanism which can simultaneously "copy" the novel objects to the output sentence and the framework is trainable in an end-to-end fashion.

3. Image Captioning with Copying Mechanism

The main goal of our Long Short-Term Memory with Copying Mechanism (LSTM-C) framework is to describe novel objects in the output sentences by incorporating the copying mechanism into the decoding stage of image cap-

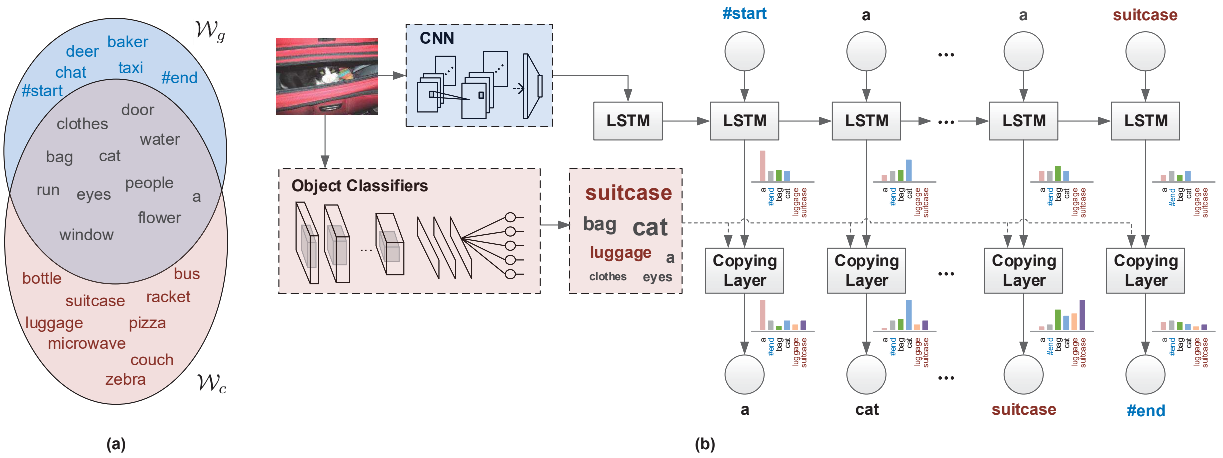


Figure 2. The overview of Long Short-Term Memory with Copying Mechanism (LSTM-C) for describing novel objects (better viewed in color). (a) $\mathcal{W}_g$ and $\mathcal{W}_c$ are the vocabularies on paired image-sentence dataset and unpaired object recognition dataset, respectively. (b) The image representation extracted by CNN is injected into LSTM at the initial time for standard word-by-word sentence generation. Meanwhile, the object classifiers learnt on unpaired object recognition dataset are utilized to detect the object candidates which are additionally incorporated into LSTM for directly "copying" them into the output sentence, enabling the captioning for novel objects. To better leverage both generative mechanism for standard word-by-word sentence generation and our adopted copying mechanism, a copying layer is specially devised to integrate them in an end-to-end trainable architecture.

tioning. The overall training of LSTM-C is similar to regular CNN plus RNN systems by minimizing the energy loss which estimates the contextual relationships among the generated words in the decoding stage. Particularly, we measure the log probability of target word through not only the natural generation by generic RNN decoder, but also the direct "copying" from the detected objects learnt on largely object recognition datasets, enabling the captioning for novel objects. The framework overview is shown in Figure 2.

In the following, we will first define the representations of images, the sequential words in sentence and the detected objects from images, followed by sequence modeling in image captioning. Next, to select words from novel objects and put them at proper places in the output sentence, we present the copying mechanism for image captioning from the viewpoint of rote memorization like a human being. Finally, the overall objective and optimization strategy of LSTM-C are presented in a CNN plus RNN framework. Technically, we devise a copying layer at the top of CNN plus RNN architecture, which incorporates both generative and copying mechanisms to optimize the whole network.

3.1. Notation

Suppose we have an image I to be described by a textual sentence $\mathcal{S}$, where $\mathcal{S} = \{w_1, w_2, ..., w_{N_s}\}$ consisting of N_s words. Let $\mathbf{I} \in \mathbb{R}^{D_v}$ and $\mathbf{w}_t \in \mathbb{R}^{D_w}$ denote the D_v-dimensional visual representations of the image I and the D_w-dimensional textual features of the t-th word in sentence $\mathcal{S}$, respectively. As a sentence consists of a sequence of words, a sentence can be represented by a $D_w \times N_s$ matrix $\mathbf{W} \equiv [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{N_s}]$, with each word in the sentence as its column vector. The vocabulary for the paired image-sentence data is denoted as $\mathcal{W}_g$. Furthermore, we utilize the freely available object recognition datasets to train the object classifiers which will be injected into our CNN plus RNN system for novel object captioning. Let $\mathcal{W}_c$ denote the vocabulary for the unpaired object recognition dataset and the probability of image I containing each object $w_i \in \mathcal{W}_c$ is represented as $\delta(w_i)$. More specifically, for the external images with single label (e.g., ImageNet [19]), the standard CNN architecture [20] is adopted to train the object detectors, while for the image data with multiple objects (e.g., MSCOCO [12]), we follow [5] and learn the detectors by using the weakly-supervised approach of Multiple Instance Learning (MIL).

3.2. Sequence Modeling in Image Captioning

Inspired by the recent successes of probabilistic sequence methods leveraged in statistical machine translation [1, 21], we aim to formulate our image captioning model in an end-to-end fashion based on RNN model which first encodes the given image into a fixed dimensional vector and then decodes it to the target output sentence consisting of sequential words. As such, given the image, the problem of sequence modeling for target sentence we exploit here can be generally formulated by minimizing the following energy loss function:

$$E(\mathbf{I}, \mathbf{W}) = -\log \Pr(\mathbf{W}|\mathbf{I}), \quad (1)$$

which is the negative log probability of the correct textual sentence given the visual image.

Since the model produces one word in the sentence at each time step, it is natural to apply chain rule to model the joint probability over the sequential words. Thus, the $\log$ probability of the sentence is given by the sum of the $\log$ probabilities over the word and can be expressed as

$$\log \Pr\left(\mathbf{W}|\mathbf{I}\right) = \sum_{t=1}^{N_s} \log \Pr\left(\left.\mathbf{w}_t\right|\mathbf{I}, \mathbf{w}_0, \ldots, \mathbf{w}_{t-1}\right). \quad (2)$$

By minimizing this loss, the contextual relationship among the words in the sentence can be guaranteed given the visual content of image.

We formulate this task as a variable-length sequence to sequence problem and model the parametric distribution $\Pr\left(\left.\mathbf{w}_t\right|\mathbf{I}, \mathbf{w}_0, \ldots, \mathbf{w}_{t-1}\right)$ in Eq.(2) with LSTM, which is a widely used type of RNN in image/video captioning [23, 28, 16, 25]. The vector formulas for a LSTM layer forward pass are given as below. For time step t, $\mathbf{x}^t$ and $\mathbf{h}^t$ are the input and output vector respectively, $\mathbf{T}$ are input weights matrices, $\mathbf{R}$ are recurrent weight matrices and $\mathbf{b}$ are bias vectors. Sigmoid σ and hyperbolic tangent ϕ are element-wise non-linear activation functions. The dot product of two vectors is denoted with $\odot$. Given inputs $\mathbf{x}^t$, $\mathbf{h}^{t-1}$ and $\mathbf{c}^{t-1}$, the LSTM unit updates for time step t are:

$$\mathbf{g}^t = \phi(\mathbf{T}_g\mathbf{x}^t + \mathbf{R}_g\mathbf{h}^{t-1} + \mathbf{b}_g),\ \mathbf{i}^t = \sigma(\mathbf{T}_i\mathbf{x}^t + \mathbf{R}_i\mathbf{h}^{t-1} + \mathbf{b}_i),$$

$$\mathbf{f}^t = \sigma(\mathbf{T}_f\mathbf{x}^t + \mathbf{R}_f\mathbf{h}^{t-1} + \mathbf{b}_f), \quad \mathbf{c}^t = \mathbf{g}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t,$$

$$\mathbf{o}^t = \sigma(\mathbf{T}_o\mathbf{x}^t + \mathbf{R}_o\mathbf{h}^{t-1} + \mathbf{b}_o), \quad \mathbf{h}^t = \phi(\mathbf{c}^t) \odot \mathbf{o}^t,$$

where $\mathbf{g}^t$, $\mathbf{i}^t$, $\mathbf{f}^t$, $\mathbf{c}^t$, $\mathbf{o}^t$, and $\mathbf{h}^t$ are cell input, input gate, forget gate, cell state, output gate, and cell output of the LSTM, respectively.

As mentioned above, the LSTM model is utilized to predict each word in the sentence given the image content and previous words. We inject the embedded image representation at the initial time to inform the whole memory cells in LSTM about the visual content. Given the image $\mathbf{I}$ and the corresponding sentence $\mathbf{W} \equiv [\mathbf{w}_0, \mathbf{w}_1, ..., \mathbf{w}_{N_s}]$, the LSTM updating procedure is as following:

$$\mathbf{x}^{-1} = \mathbf{T}_I\mathbf{I}, \quad (3)$$

$$\mathbf{x}^t = \mathbf{T}_s\mathbf{w}_t, \quad t \in \{0, \ldots, N_s - 1\}\ , \quad (4)$$

$$\mathbf{h}^t = f\left(\mathbf{x}^t\right),\ \ t \in \{0, \ldots, N_s - 1\}, \quad (5)$$

where D_e is the dimensionality of LSTM input, and $\mathbf{T}_I \in \mathbb{R}^{D_e \times D_v}$ and $\mathbf{T}_s \in \mathbb{R}^{D_e \times D_w}$ are the transformation matrices for image representation and textual feature of word, respectively, and f is the updating function within LSTM unit. Please note that for the input sentence $\mathbf{W} \equiv [\mathbf{w}_0, \ldots, \mathbf{w}_{N_s}]$, we take $\mathbf{w}_0$ as the start sign word to inform the beginning of sentence and $\mathbf{w}_{N_s}$ as the end sign word which indicates the end of sentence, both of the special sign words are included in the existing vocabulary $\mathcal{W}_g$ for the paired image-sentence data. More specifically, at the initial encoding step, the image representation is transformed as the input for LSTM, and then in the next decoding steps, word embedding $\mathbf{x}^t$ will be input into the LSTM along with the previous step's hidden state $\mathbf{h}^{t-1}$.

In the decoding stage, given the LSTM cell output $\mathbf{h}^t$ at the t-th time step, the widely adopted method for next word prediction is the generative mechanism [1] which calculates the corresponding probability of generating any target word w_{t+1} as

$$\Pr_t^g\left(w_{t+1}\right) = \mathbf{w}_{t+1}^{\top}\mathbf{M}_g\mathbf{h}^t, \quad (6)$$

where D_h is the dimensionality of LSTM output and $\mathbf{M}_g \in \mathbb{R}^{D_w \times D_h}$ is the transformation matrix for textual features of word in the generative mechanism. For the standard word-by-word sentence generation model, a softmax function is applied after the probabilities measured by generative mechanism to produce a normalized probability distribution over all the words in the vocabulary $\mathcal{W}_g$.

3.3. Copying Mechanism

The copying mechanism has been shown effective for sequence learning [7] to address the out-of-vocabulary (OOV) problem in text summarization. The mechanism is regarded as the rote memorization in language processing of human being that directly "copying" existing segments in the input sequence to target sequence. Similar in spirit, we extend the copying mechanism in image captioning to directly "copying" the appropriate objects from the detected candidates in image to compose the output sentence, especially for novel objects which never appear in paired image-sentence data, enabling the novel object captioning. Specifically, at the t-th decoding step, we directly take the similarity between any word w_{t+1} in $\mathcal{W}_c$ and the corresponding LSTM cell output $\mathbf{h}^t$ as the probability for "copying" the target word w_{t+1} to the target sentence, which is calculated as

$$\Pr_t^c\left(w_{t+1}\right) = \varphi\left(\mathbf{w}_{t+1}^{\top}\mathbf{M}_c\right)\mathbf{h}^t\delta\left(w_{t+1}\right), \quad (7)$$

where $\mathbf{M}_c \in \mathbb{R}^{D_w \times D_h}$ is the transformation matrix for mapping textual features of word in the copying mechanism and φ is element-wise non-linear activation function. It is also worth noticing that we additionally incorporate the object classification score $\delta\left(w_{t+1}\right)$ into the formulation of "copying" probability since this classification score reflects the chance of the object appeared in the image. The underlying assumption is that in addition to the effect of LSTM cell output, the larger the classification score of this word in the image, the higher the probability for "copying" this word in the target sentence.

3.4. LSTM with Copying Mechanism

Unlike the existing image captioning approaches which always model the sequence learning with generative mechanism for sentence generation, our proposed LSTM-C architecture further incorporates the copying mechanism into

LSTM at the decoding stage to describe novel objects in sentence. In particular, given the output of LSTM cell at each decoding step, we utilize both generative and copying mechanisms simultaneously to measure the probability of generating any target word. As the vocabulary $\mathcal{W}_c$ of copying mechanism is derived from external image data, it may include the words which are not present in the vocabulary $\mathcal{W}_g$ of image-sentence data, making the copying mechanism able to copy such novel objects to the output sentence. In this case, we directly consider the probability of copying mechanism in Eq.(7) as the final probability of generating these novel objects. Similarly, for the words that only belong to $\mathcal{W}_g$, the final probabilities of them fully depend on their corresponding probabilities of generative mechanism in Eq.(6). In terms of the overlapping words between $\mathcal{W}_g$ and $\mathcal{W}_c$, we linearly fuse the probabilities from both generative and copying mechanisms as the final output probabilities. Hence, at the t-th decoding step, the final output probability $\Pr_t(w_{t+1})$ of any target word w_{t+1} is defined as follows:

$$\Pr_t(w_{t+1}) = \begin{cases} \frac{1}{K} e^{\Pr_t^g(w_{t+1})}, & w_{t+1} \in \mathcal{W}_g \cap \overline{\mathcal{W}_c} \\ \frac{\lambda}{K} e^{\Pr_t^g(w_{t+1})} + \frac{1-\lambda}{K} e^{\Pr_t^c(w_{t+1})}, & w_{t+1} \in \mathcal{W}_g \cap \mathcal{W}_c \\ \frac{1}{K} e^{\Pr_t^c(w_{t+1})}, & w_{t+1} \in \overline{\mathcal{W}_g} \cap \mathcal{W}_c \\ 0, & otherwise \end{cases}, \tag{8}$$

where λ is the tradeoff parameter between the two mechanisms and K is the softmax normalization term.

Accordingly, we define our energy loss function in training stage for each image-sentence pair as follows:

$$E(I, \mathcal{S}) = -\sum_{t=0}^{N_s-1} \log \Pr_t(\mathbf{w}_{t+1}). \tag{9}$$

Let N denote the number of image-sentence pairs in the training set, we have the following optimization problem:

$$\begin{array}{c} \min\limits_{\mathbf{T}_I, \mathbf{T}_s, \mathbf{M}_g, \mathbf{M}_c, \theta} \frac{1}{N} \sum\limits_{i=1}^{N} E(I^{(i)}, \mathcal{S}^{(i)}) \\ + \|\mathbf{T}_I\|_2^2 + \|\mathbf{T}_s\|_2^2 + \|\mathbf{M}_g\|_2^2 + \|\mathbf{M}_c\|_2^2 + \|\theta\|_2^2 \end{array}, \tag{10}$$

where the first term is the overall energy loss, and the rest are regularization terms for image embedding, textual embedding for LSTM input, textual embedding in generative mechanism, textual embedding in copying mechanism, and LSTM, respectively. Moreover, following [22], we also implicitly integrate the overall energy loss with text-specific loss on external sentence data for maintaining the model's ability to address novel objects among sentences.

To solve the optimization according to overall loss objective in Eq.(10), we design a copying layer at the top of LSTM with two textual embedding parameters for generative and copying mechanisms. During training, this copying layer measures the output probability for each word considering both generative and copying mechanisms as defined in Eq.(8), followed by a softmax normalization operation for overall optimization.

In the testing stage for sentence generation, we choose the word among the combination vocabulary of $\mathcal{W}_g$ and $\mathcal{W}_c$ with maximum probability at each time step and set its embedded textual feature as LSTM input for the next time step until the end sign word is outputted.

4. Experiments

We evaluate and compare our proposed LSTM-C with state-of-the-art approaches by conducting novel object captioning task on two image datasets, i.e., the held-out Microsoft COCO Caption dataset (held-out MSCOCO) [8] which is a subset of MSCOCO dataset [12] and ImageNet [19], a large-scale object recognition dataset.

4.1. Datasets

Held-out MSCOCO. The held-out MSCOCO consists of a subset of MSCOCO which excludes all the image-sentence pairs that contain at least one of eight specific objects in MSCOCO. It is worth noting that following [8], the eight specific objects are chosen through the clustering over all the 80 objects in MSCOCO segmentation challenge and each cluster excludes one object, resulting in the final eight novel objects for evaluation: "bottle," "bus," "couch," "microwave," "pizza," "racket," "suitcase," and "zebra." For this subset, there are five human-annotated descriptions per image. As the annotations of the official testing set are not publicly available and thus following [8], we split the MSCOCO validation set into two: 50% for validation and the other 50% for testing. For the experiments on held-out MSCOCO, the object classifiers for copying mechanism are trained with all the MSCOCO training images including the eight novel objects and the LSTM for sequence modeling is pre-trained with all the sentences in MSCOCO training set, while the entire CNN plus RNN system are optimized with the paired image-sentence data only from held-out MSCOCO training set. The testing set of held-out MSCOCO is then utilized to evaluate the ability of our LSTM-C model to describe the eight novel objects.

ImageNet. We also conduct our experiments on the large-scale object recognition dataset, i.e., ImageNet, for evaluation. Similar to [22], a subset from ImageNet with 634 different objects which are not present in the MSCOCO dataset is adopted in our experiments. In particular, about 75% of images in each class are exploited for training and the rest are utilized for testing, resulting in the training and testing set with 493,519 and 164,820 images, respectively. For the experiments on ImageNet, we train the object classifiers for copying mechanism purely on the ImageNet training set and pre-train the LSTM part with all the sentences in MSCOCO training set. In terms of the entire CNN plus RNN system, it is optimized with the paired image-sentence data in MSCOCO training set. Since none of the objects in this subset of ImageNet is addressed in the paired image-sentence data, we generate sentences for images in the testing set of ImageNet and empirically evaluate the ability of our LSTM-C model to describe the 634 novel objects.

4.2. Experimental Settings

Features and Parameter Settings. For image representations, we take the output of 4,096-way fc7 layer from 16-layer VGG [20] pre-trained on Imagenet ILSVRC12 dataset [19]. Each word in the sentence is represented as the combined vector of embedded one-hot representation and Glove [18] representation. For the paired image-sentence data (e.g., MSCOCO), we select the 1,000 most common words on MSCOCO as the objects and train the corresponding object classifiers with MIL model [5] purely on the training data of MSCOCO. The MIL model is mainly designed based on a Fully Convolutional Network (FCN) extended from 16-layer VGG. For the unpaired object recognition data (e.g., ImageNet), 634 object classifiers are trained by directly fine-tuning the 16-layer VGG pre-trained on Imagenet ILSVRC12 dataset. The dimensionality of the input and hidden layers in LSTM are both set to 1,024. The tradeoff parameter λ leveraging both generative and copying mechanisms is empirically set to 0.2. The sensitivity of λ will be discussed later.

Implementation Details. We mainly implement our image captioning models based on Caffe [9], which is one of widely adopted deep learning frameworks. In particular, the initial learning rate and mini-batch size is set as 0.01 and 1,024, respectively. The entire CNN plus RNN system in our LSTM-C is trained for 50 epoches on both datasets or we stop the training until the performance has no longer improvement on the corresponding validation set.

Evaluation Metrics. For quantitative evaluation of our proposed model on held-out MSCOCO, we adopt the most common caption metric, i.e., METEOR [2], to evaluate description quality which computes unigram precision and recall against all ground truth sentences with some pre-processing on WordNet synonyms and stemmed tokens. However, as pointed in [8], it is still possible to achieve high METEOR scores without mentioning the novel objects. Hence, to fully validate the model's ability of describing novel objects, F1-score is exploited as another evaluation metric, which determines whether the specific novel object is mentioned in the generated descriptions for the images containing that novel object. All the metrics above are computed by using the codes[1] released by [8] for fair comparison. To evaluate our model on ImageNet without any ground truth sentences, we utilize another two metrics for novel object captioning task: describing novel objects (Novel) [22] and Accuracy [22] scores. The Novel score measures the percentage of all the 634 novel objects mentioned in generated descriptions, i.e., for each novel object, the model should incorporate it into at least one sentence for the ImageNet image with this object. For the Accuracy score of each novel object, it represents the percentage of images belonging to this novel object which can be described correctly by addressing that novel object in the sentences. The Accuracy score is finally averaged over all the 634 novel objects.

4.3. Compared Approaches

To empirically verify the merit of our LSTM-C model, we compared the following state-of-the-art methods, including both regular image captioning and novel object captioning approaches.

- Long-term Recurrent Convolutional Networks (LRCN) [4]: LRCN is one of the basic RNN-based image captioning models which inputs both visual image and previous word into LSTM at each time step for sentence generation. As a regular image captioning model without any mechanism for considering novel objects, LRCN is trained only on the paired image-sentence data without any novel objects.

- Deep Compositional Captioner (DCC) [8]: DCC firstly pre-trains lexical classifier and language model with external unpaired data, and then integrates both two parts to learn an improved caption model trained with paired image-sentence data. Finally, DCC explicitly transfers the knowledge of semantically related objects to compose the descriptions with novel objects.

- Novel Object Captioner (NOC) [22]: Proposed most recently, NOC extends DCC by jointly optimizing the three parts: visual recognition network, LSTM-based language model, and image captioning network in an end-to-end manner. Please note that for fair comparison with LRCN and DCC which utilize one hot vector as their word representations, we include two runs, i.e., NOC (One hot) and NOC (One hot+Glove) which are our implementations of NOC. The word representations in the latter one are the combination of the embedded one hot vector and Glove vector.

- Long Short-Term Memory with Copying Mechanism (LSTM-C): We design two runs, i.e., LSTM-C (One-hot) and LSTM-C (One hot+Glove), for our proposed end-to-end architecture for novel object captioning.

[1] `https://github.com/LisaAnne/DCC`

Table 1. Per-object F1, averaged F1 and METEOR scores of our proposed model and other state-of-the-art methods on held-out MSCOCO dataset for novel object captioning. All values are reported as percentage (%).

Model	$\text{F1}_{\text{bottle}}$	F1_{bus}	F1_{couch}	$\text{F1}_{\text{microwave}}$	F1_{pizza}	$\text{F1}_{\text{racket}}$	$\text{F1}_{\text{suitcase}}$	F1_{zebra}	$\text{F1}_{\text{average}}$	METEOR
LRCN [4]	0	0	0	0	0	0	0	0	0	19.33
DCC [8]	4.63	29.79	**45.87**	28.09	64.59	52.24	13.16	79.88	39.78	21
NOC [22]										
-(One hot)	16.52	68.63	42.57	32.16	67.07	61.22	31.18	88.39	50.97	20.7
-(One hot+Glove)	14.93	68.96	43.82	**37.89**	66.53	65.87	28.13	88.66	51.85	20.7
LSTM-C										
-(One hot)	29.07	64.38	26.01	26.04	**75.57**	66.54	**55.54**	**92.03**	54.40	22
-(One hot+Glove)	**29.68**	**74.42**	38.77	27.81	68.17	**70.27**	44.76	91.4	**55.66**	**23**

4.4. Performance Comparison

We first conduct the experiment on held-out MSCOCO to examine how our LSTM-C model work on describing the eight novel objects. Then, to further verify the scalability of our proposed model, the second experiment is performed on ImageNet to describe hundreds of novel objects that outside of the paired image-sentence data.

Evaluation on held-out MSCOCO. Table 1 shows the performances of compared six models on held-out MSCOCO dataset. Overall, the results across two general evaluation metrics (averaged F1 and METEOR scores) consistently indicate that our proposed LSTM-C exhibits better performance than all the state-of-the-art techniques including regular image captioning model (LRCN) and two novel object captioning systems (DCC and NOC). In particular, by additionally utilizing external unpaired data for training, all the latter five novel object captioning models outperform the regular image captioning model LRCN on both description quality and novelty. There is a significant performance gap between DCC and LSTM-C (One hot). Although both runs involve the utilization of external image data, they are fundamentally different in the way that DCC leverages explicit transfer mechanism for recognizing novel objects and cannot be trained end-to-end, and LSTM-C (One hot) implicitly addresses the novel objects for sentence generation with copying mechanism in an end-to-end manner. Moreover, by incorporating copying mechanism to standard word-by-word sentence generation model, LSTM-C (One hot) leads to a performance boost against NOC (One hot), indicating that the generative mechanism and copying mechanism are complementary and thus have mutual reinforcement for novel object captioning. Another observation is that when combining the word representations from embedded one hot vector and Glove vector, LSTM-C (One hot+Glove) further increases the performance.

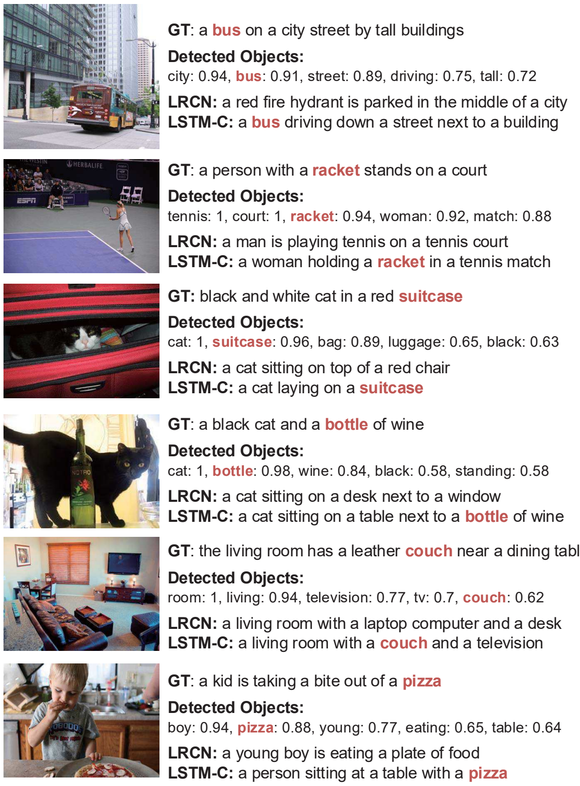
Table 1 also details the F1 scores for all the eight novel objects. Among all the novel objects, our proposed LSTM-C achieves the best performance for describing six novel objects, followed by DCC and NOC for one object, respectively. The improvements can be generally expected by additionally incorporating copying mechanism in sequence learning except "couch" and "microwave" objects. This is not surprise because such novel objects always have high visual similarity with other objects (e.g., "bed" for "couch" and "oven" for "microwave") and thus are not easy to be detected precisely, making LSTM-C fail to copy them to the output sentences.

Table 2. Novel, F1 and Accuracy scores of our proposed model and other state-of-the-art methods on ImageNet dataset. All values are reported as percentage (%).

Model	Novel	F1	Accuracy
NOC (One hot+Glove) [22]			
-MSCOCO	69.08	15.63	10.04
-BNC&Wiki	87.69	31.23	21.96
LSTM-C (One hot+Glove)			
-MSCOCO	72.08	16.39	11.83
-BNC&Wiki	89.11	33.64	31.11

Evaluation on ImageNet. Table 2 summarizes the experimental results on ImageNet dataset. By only adopting the MSCOCO as the training data for the CNN plus RNN system, our LSTM-C (One hot+Glove) makes the relative improvement over NOC (One hot+Glove) by 4.3%, 4.9% and 17.8% in Novel, F1 and Accuracy, respectively. The results basically indicate the advantage of exploiting both generative and copying mechanisms in the CNN plus RNN system for novel object captioning, even when scaling into ImageNet images with hundreds of novel objects. Moreover, following [22], we also include the external unpaired text data (i.e., British National Corpus and Wikipedia) in our LSTM-C (One hot+Glove) and performance improvements are further observed.

Qualitative analysis. Figure 3 and Figure 5 shows a few sentence examples generated by different methods, the detected objects and human-annotated ground truth on held-out MSCOCO and ImageNet dataset, respectively. From these exemplar results, it is easy to see that all of these captioning models can generate somewhat relevant sentences on both datasets, while our proposed LSTM-C can predict the novel objects by incorporating copying mechanism for image captioning. For example, compared to object term "hydrant" in the sentence generated by LRCN, "bus" in our LSTM-C is more precise to describe the image content in



GT: a bus on a city street by tall buildings
Detected Objects:
city: 0.94, bus: 0.91, street: 0.89, driving: 0.75, tall: 0.72
LRCN: a red fire hydrant is parked in the middle of a city
LSTM-C: a bus driving down a street next to a building

GT: a person with a racket stands on a court
Detected Objects:
tennis: 1, court: 1, racket: 0.94, woman: 0.92, match: 0.88
LRCN: a man is playing tennis on a tennis court
LSTM-C: a woman holding a racket in a tennis match

GT: black and white cat in a red suitcase
Detected Objects:
cat: 1, suitcase: 0.96, bag: 0.89, luggage: 0.65, black: 0.63
LRCN: a cat sitting on top of a red chair
LSTM-C: a cat laying on a suitcase

GT: a black cat and a bottle of wine
Detected Objects:
cat: 1, bottle: 0.98, wine: 0.84, black: 0.58, standing: 0.58
LRCN: a cat sitting on a desk next to a window
LSTM-C: a cat sitting on a table next to a bottle of wine

GT: the living room has a leather couch near a dining table
Detected Objects:
room: 1, living: 0.94, television: 0.77, tv: 0.7, couch: 0.62
LRCN: a living room with a laptop computer and a desk
LSTM-C: a living room with a couch and a television

GT: a kid is taking a bite out of a pizza
Detected Objects:
boy: 0.94, pizza: 0.88, young: 0.77, eating: 0.65, table: 0.64
LRCN: a young boy is eating a plate of food
LSTM-C: a person sitting at a table with a pizza

Figure 3. Objects and sentence generation results on held-out MSCOCO. The detected objects are predicted by MIL model in [5], and the output sentences are generated by 1) Ground Truth (GT): one ground truth sentence, 2) LRCN and 3) our LSTM-C.

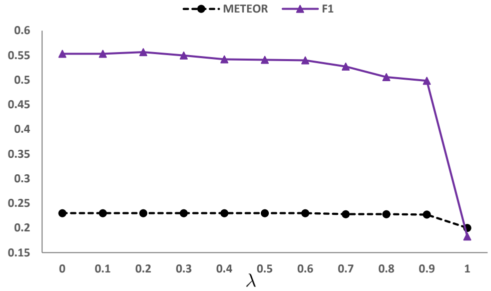
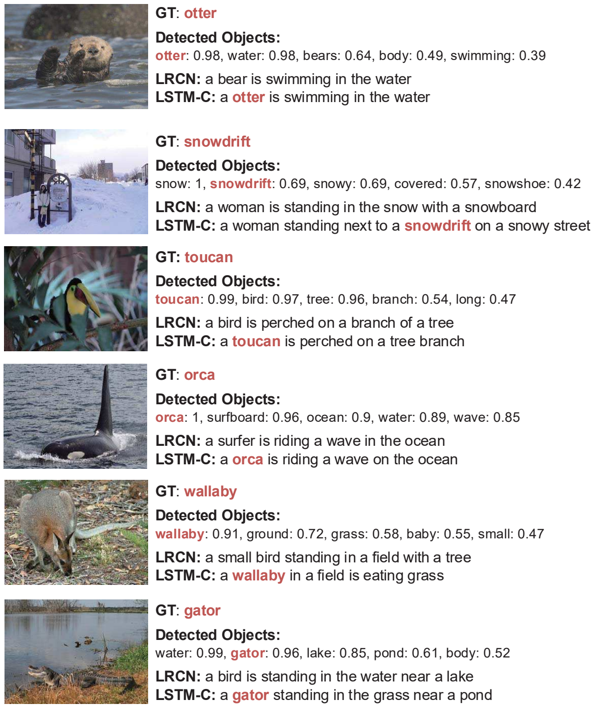


Figure 4. The effect of the tradeoff parameter λ in our LSTM-C (One hot+Glove) framework on held-out MSCOCO.

the first image on held-out MSCOCO dataset, since the novel object "bus" is among the top object candidates and directly copied to the output sentence at the decoding stage.

4.5. Analysis of the Tradeoff Parameter λ

To clarify the effect of the tradeoff parameter λ in Eq.(8), we illustrate the performance curves with different tradeoff parameters in Figure 4. As shown in the figure, we can see that the performance curves of F1 and METEOR scores are both relatively smooth when λ varies in a range from 0 to 0.6. Specifically, the best performance is achieved when λ is about 0.2. Furthermore, when the λ increases more than 0.6, the F1 score begins to drop significantly, again demonstrating the importance of copying mechanism in our LSTM-C for describing novel objects.



GT: otter
Detected Objects:
otter: 0.98, water: 0.98, bears: 0.64, body: 0.49, swimming: 0.39
LRCN: a bear is swimming in the water
LSTM-C: a otter is swimming in the water

GT: snowdrift
Detected Objects:
snow: 1, snowdrift: 0.69, snowy: 0.69, covered: 0.57, snowshoe: 0.42
LRCN: a woman is standing in the snow with a snowboard
LSTM-C: a woman standing next to a snowdrift on a snowy street

GT: toucan
Detected Objects:
toucan: 0.99, bird: 0.97, tree: 0.96, branch: 0.54, long: 0.47
LRCN: a bird is perched on a branch of a tree
LSTM-C: a toucan is perched on a tree branch

GT: orca
Detected Objects:
orca: 1, surfboard: 0.96, ocean: 0.9, water: 0.89, wave: 0.85
LRCN: a surfer is riding a wave in the ocean
LSTM-C: a orca is riding a wave on the ocean

GT: wallaby
Detected Objects:
wallaby: 0.91, ground: 0.72, grass: 0.58, baby: 0.55, small: 0.47
LRCN: a small bird standing in a field with a tree
LSTM-C: a wallaby in a field is eating grass

GT: gator
Detected Objects:
water: 0.99, gator: 0.96, lake: 0.85, pond: 0.61, body: 0.52
LRCN: a bird is standing in the water near a lake
LSTM-C: a gator standing in the grass near a pond

Figure 5. Objects and sentence generation results on ImageNet. GT denotes the ground truth object. The detected objects are predicted by the standard CNN architecture [20], and the output sentences are generated by 1) LRCN and 2) our LSTM-C.

5. Discussions and Conclusions

We have presented Long Short-Term Memory with Copying Mechanism (LSTM-C) framework which leverages external visual recognition for image captioning. Particulary, we study the problem of predicting novel objects in image caption by integrating the detected objects with copying mechanism. To verify our claim, we have devised an end-to-end architecture to accommodate the standard word-by-word sentence generation by LSTM and the mechanism of copying from detected objects. Experiments conducted on MSCOCO image captioning and ImageNet datasets validate our proposal and analysis. Performance improvements are clearly observed when comparing to other novel object captioning techniques.

Our future works are as follows. First, more objects will be learnt on large-scale image benchmarks, e.g., YFCC-100M dataset, and integrated into our LSTM-C architecture. We will further analyze the impact of different sources involved. Second, how to apply our proposal to video domain is worth trying.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005.

[3] 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 *ACL*, 2015.

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[5] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015.

[6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010.

[7] J. Gu, Z. Lu, H. Li, and V. O. Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

[8] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *MM*, 2014.

[10] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *ICML*, 2014.

[11] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. on PAMI*, 2013.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[13] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *ICCV*, 2015.

[14] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. Midge: Generating image descriptions from computer vision detections. In *EACL*, 2012.

[15] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[16] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[17] Y. Pan, T. Yao, H. Li, and T. Mei. Video captioning with transferred semantic attributes. *arXiv preprint arXiv:1611.07675*, 2016.

[18] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[21] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[22] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. *arXiv preprint arXiv:1606.07770*, 2016.

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[24] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v. d. Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR*, 2016.

[25] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[27] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011.

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

[29] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016.