This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Pointing Novel Objects in Image Captioning*

Yehao Li [†§], Ting Yao [‡], Yingwei Pan [‡], Hongyang Chao [†§], and Tao Mei [‡]

[†] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

[‡] JD AI Research, Beijing, China

[§] Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education

{yehaoli.sysu, tingyao.ustc, panyw.ustc}@gmail.com, isschhy@mail.sysu.edu.cn, tmei@live.com

Abstract

Image captioning has received significant attention with remarkable improvements in recent advances. Nevertheless, images in the wild encapsulate rich knowledge and cannot be sufficiently described with models built on image-caption pairs containing only in-domain objects. In this paper, we propose to address the problem by augmenting standard deep captioning architectures with object learners. Specifically, we present Long Short-Term Memory with Pointing (LSTM-P) — a new architecture that facilitates vocabulary expansion and produces novel objects via pointing mechanism. Technically, object learners are initially pre-trained on available object recognition data. Pointing in LSTM-P then balances the probability between generating a word through LSTM and copying a word from the recognized objects at each time step in decoder stage. Furthermore, our captioning encourages global coverage of objects in the sentence. Extensive experiments are conducted on both held-out COCO image captioning and ImageNet datasets for describing novel objects, and superior results are reported when comparing to state-of-the-art approaches. More remarkably, we obtain an average of 60.9% in F1 score on held-out COCO dataset.

1. Introduction

Automatic caption generation is the task of producing a natural-language utterance (usually a sentence) that describes the visual content of an image. Practical applications of automatic caption generation include leveraging descriptions for image indexing or retrieval, and helping those with visual impairments by transforming visual signals into information that can be communicated via text-to-speech technology. Recently, state-of-the-art image captioning methods tend to be monolithic deep models essentially of "encoder-decoder" style [13, 28, 36]. In general, a Convolutional Neural Network (CNN) is employed to encode an image into a feature vector, and a caption is then decoded from this vector using a Long Short-Term Memory (LSTM) Network, which is one typical Recurrent Neural Network (RNN). Such models have indeed demonstrated promising results on image captioning task. However, one of the most critical limitations is that the existing models are often built on a number of image-caption pairs, which contain only a shallow view of in-domain objects. That hinders the generalization of these models in real-world scenarios to describe novel scenes or objects in out-of-domain images.

The difficulty of novel objects prediction in captioning mainly originates from two aspects: 1) how to facilitate word vocabulary expansion? 2) how to learn a hybrid network that can nicely integrate the recognized objects (words) into the output captions? We propose to mitigate the first problem through leveraging the knowledge from visual recognition datasets, which are freely available and easier to be scalable for developing object learners. Next, *pointing* mechanism is devised to balance the word generation from decoder and the word taken directly from the learnt objects. In other words, such mechanism controls when to directly put the learnt objects at proper places in the output sentence, i.e., *when to point*. Moreover, despite having high quantitative scores, qualitative analysis shows that automatically generated captions by deep captioning models are often limited to describing very generic information of objects, or rely on prior information and correlations from training examples, and resulting frequently in undesired effects such as object hallucination [14]. As a result, we further take the coverage of objects into account to cover more objects in the sentence and thus improve the captions.

By consolidating the idea of pointing mechanism and the coverage of objects into image captioning, we present a new Long Short-Term Memory with Pointing (LSTM-P) architecture for novel object captioning. Given an image, a CNN is utilized to extract visual features, which are fed into LSTM at the initial time step as a trigger of sentence generation. The output of LSTM is probability distribution over all the

*This work was performed at JD AI Research.

words in the vocabulary. The pre-trained object recognizers are employed in parallel to detect objects in the input image. A Copying layer then takes the prediction scores of objects and the current hidden state of LSTM as its inputs. It outputs the probability distribution of being copied over all the recognized objects. To dynamically accommodate word generation through LSTM and word copying from the learnt objects, pointing mechanism as a multi-layer perceptron is exploited to balance the output probability distribution from LSTM and copying layer at each time step. Moreover, the coverage of objects is encouraged to talk about more objects found in the image, which is independent of the position in the sentence. As such, the measure of coverage is performed on the bag-of-objects on sentence level. The whole LSTM-P is trained by jointly minimizing the widely-adopted sequential loss on the produced sentence plus sentence-level coverage loss.

The main contribution of this work is the proposal of LSTM-P architecture for addressing the issue of novel objects prediction in image captioning. This issue also leads to the elegant view of how to expand vocabulary, and how to nicely point towards the placements and moments of copying novel objects in the sentence, which are problems not yet fully understood in the literature.

2. Related Work

Image Captioning. Inspired from deep learning [10] in computer vision and sequence modeling [24] in Natural Language Processing, modern image captioning methods [6, 21, 28, 31, 34, 35, 36] mainly exploit sequence learning models to produce sentences with flexible syntactical structures. For example, [28] presents an end-to-end CNN plus RNN architecture which capitalizes on LSTM to generate sentences word-by-word. [31] further extends [28] by integrating soft/hard attention mechanism to automatically focus on salient regions within images when producing corresponding words. Moreover, instead of calculating visual attention over image regions at each time step of decoding stage, [13] devises an adaptive attention mechanism in encoder-decoder architecture to additionally decide when to rely on visual signals or language model. Recently, [29, 35] verify the effectiveness of injecting semantic attributes into CNN plus RNN model for image captioning. Moreover, [36] utilize the semantic attention measured over attributes to boost image captioning. Most recently, [3] proposes a novel attention based captioning model which exploits object-level attention to enhance sentence generation via bottom-up and top-down attention mechanism.

Novel Object Captioning. The task of novel object captioning has received increasing attention most recently, which leverages additional image-sentence paired data [15] or unpaired image/text data [8, 26] to describe novel objects. Existing works mainly remould the RNN-based image captioning frameworks towards the scenario of novel object captioning by additionally leveraging image taggers/object detectors to inject novel objects for describing. Specifically, [15] is one of early attempts that describes novel objects by enlarging the original limited vocabulary based on only a few paired image-sentence data. A transposed weight sharing strategy is especially devised to avoid extensive re-training. In contrast, [8] presents Deep Compositional Captioner (DCC) which utilizes the largely available unpaired image and text data (e.g., ImageNet and Wikipedia) to facilitate novel object captioning. The knowledge of semantically related objects is explicitly exploited in DCC to compose the sentences containing novel objects. Venugopalan *et al.* [26] further extend DCC by simultaneously optimizing the visual recognition network, language model, and image captioning network in an end-to-end manner. Recently, [33] integrates the regular RNN-based decoder with copying mechanism which can simultaneously copy the detected novel objects to the output sentence. Another two-stage system is proposed in [17] by firstly building a multi-entity-label image recognition model for predicting abstract concepts and then leveraging such concepts as an external semantic attention & constrained inference for sentence generation. Furthermore, Anderson *et al.* [2] devise constrained beam search to force the inclusion of selected tag words in the output of RNN-based decoder, facilitating vocabulary expansion to novel objects without re-training. Most recently, [14] first generates a hybrid template that contains a mix of words and slots explicitly associated with image region, and then fills in the slots with visual concepts identified in the regions by object detectors.

Summary. In short, our approach focuses on the latter scenario, that leverages object recognition data for novel object captioning. Similar to previous approaches [17, 33], LSTM-P augments the standard RNN-based language model with the object learners pre-trained on object recognition data. The novelty is on the exploitation of pointing mechanism for dynamically accommodating word generation via a RNN-based language model and word copying from the learnt objects. In particular, we utilize the pointing mechanism to elegantly point when to copy the novel objects to target sentence, targeting for balancing the influence between copying mechanism and standard word-by-word sentence generation conditioned on the contexts. Moreover, the measure of sentence-level coverage is adopted as an additional training target to encourage the global coverage of objects in the sentence.

3. Method

We devise our Long Short-Term Memory with Pointing (LSTM-P) architecture to facilitate novel object captioning by dynamically integrating the recognized novel objects into the output sentence via pointing mechanism. In particu-

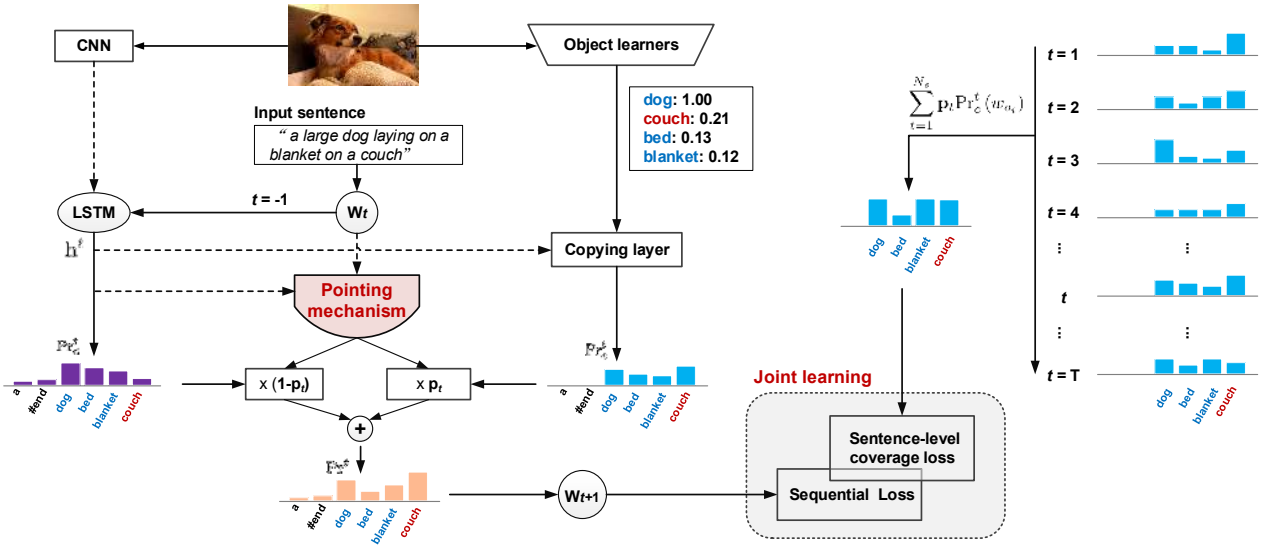


Figure 1. An overview of our Long Short-Term Memory with Pointing (LSTM-P) for novel object captioning (better viewed in color). The image representation extracted by CNN is firstly injected into LSTM at the initial time for triggering the standard word-by-word sentence generation. The output of LSTM is the probability distribution over all the words in the vocabulary at each decoding time. Meanwhile, the object learners pre-trained on object recognition data are utilized to detect the objects within the input image. Such predicted score distribution over objects are further injected into a copying layer along with the current hidden state of LSTM, producing the probability distribution of being copied over the recognized objects. To dynamically accommodate word generation via LSTM and word copying from learnt objects, a pointing mechanism is specially devised to elegantly point when to copy the object depending on contextual information (i.e., current input word and LSTM hidden state). The whole LSTM-P is trained by minimizing two objectives in an end-to-end manner: (1) the widely-adopted sequential loss that enforces the syntactic coherence of output sentence, and (2) the sentence-level coverage loss that encourages the maximum coverage of all objects found in the image, which is independent of the position in the sentence.

lar, LSTM-P firstly utilizes a regular CNN plus RNN language model to exploit the contextual relationships among the generated words. Meanwhile, the object learners trained on object recognition data are leveraged to detect objects for the input image and a copying layer is further adopted to directly copy a word from the recognized objects. Next, the two pathways for generating target word, i.e., the standard word-by-word sentence generation and the direct copying from recognized objects, are dynamically accommodated through the pointing mechanism, which can point when to copy the novel objects to target sentence conditioned on the context. The overall training of LSTM-P is performed by simultaneously minimizing the sequential loss that enforces the syntactic coherence of output sentence, and the sentence-level coverage loss that encourages the maximum coverage of all objects found in the image. An overview of our framework is illustrated in Figure 1.

3.1. Notation

For novel object captioning task, we aim to describe an input image I with a textual sentence $\mathcal{S} = \{w_1, w_2, ..., w_{N_s}\}$ which consists of N_s words. Note that we represent each image I as the D_v-dimensional visual feature $\mathbf{I} \in \mathbb{R}^{D_v}$. Moreover, $\mathbf{w}_t \in \mathbb{R}^{D_w}$ denotes the D_w-dimensional textual feature of the t-th word in sentence $\mathcal{S}$. Let $\mathcal{W}_d$ denote the vocabulary on the paired image-sentence data. Furthermore, we leverage the freely available visual recognition datasets to develop the object learners which will be integrated into standard deep captioning architecture for novel object captioning. We denote the object vocabulary for the object recognition dataset as $\mathcal{W}_c$, and $\mathbf{I}_c \in \mathbb{R}^{D_c}$ represents the probability distribution over all the D_c objects in $\mathcal{W}_c$ for image I via object learners. Hence the whole vocabulary for our system is denoted as $\mathcal{W} = \mathcal{W}_d \cup \mathcal{W}_c$. In addition, to facilitate the additional measure of object coverage in the sentence, we distill all the objects in textual sentence $\mathcal{S}$ as another training target, denoted as the bag-of-objects $\mathcal{O} = \{w_{o_1}, w_{o_2}, ..., w_{o_K}\}$ with K object words.

3.2. Problem Formulation

In the novel object captioning problem, on one hand, the words in the sentence should be organized coherently in language, and on the hand, the generated descriptive sentence must be able to address all the objects within image. As such, we can formulate the novel object captioning problem by minimizing the following energy loss function:

$$E(I, \mathcal{S}) = E_d(I, \mathcal{S}) + \lambda \times E_c(I, \mathcal{O}), \quad (1)$$

where λ is the tradeoff parameter, $E_d(I, \mathcal{S})$ and $E_c(I, \mathcal{O})$ are the sequential loss and sentence-level coverage loss, respectively. The former measures the contextual dependency among the generated sequential words in the sentence

through a CNN plus RNN language model which is introduced below. The latter estimates the coverage degree of all objects within image for output sentence, which is presented in Section 3.4.

Specifically, inspired from the sequence learning models in image/video captioning [6, 11, 18, 19, 28, 31, 32] and copying mechanism [33], we equip the regular CNN plus RNN language model with the copying layer, which predicts each target word through not only the word-by-word generation by LSTM-based decoder, but also the direct copying from the recognized objects via copying layer. Hence, the sequential loss $E_d(I, \mathcal{S})$ can be measured as the negative log probability of the correct textual sentence given the image and recognized objects:

$$E_d(I, \mathcal{S}) = -\log \Pr\left(\mathcal{S}|\mathbf{I}, \mathbf{I}_c\right). \tag{2}$$

As the whole captioning model generates sentence word-by-word, we directly apply chain rule to model the joint probability over the sequential words. Therefore, the log probability of the sentence is calculated as the sum of the log probabilities over target words:

$$\log \Pr\left(\mathcal{S}|\mathbf{I}, \mathbf{I}_c\right) = \sum_{t=1}^{N_s} \log \Pr^t\left(\mathbf{w}_t|\, \mathbf{I}, \mathbf{I}_c, \mathbf{w}_0, \ldots, \mathbf{w}_{t-1}\right). \tag{3}$$

Here the probability of each target word $\Pr^t(\mathbf{w}_t)$ is measured depending on both the probability distribution over the whole vocabulary from LSTM-based decoder and the probability distribution of being copied over the recognized objects from copying layer. To dynamically integrate the influence of such two different probability distributions, we devise a pointing mechanism to adaptively make the decision of which score distribution to focus at each time step, which will be elaborated in Section 3.3.

3.3. Pointing Mechanism

When humans have a limited information on how to call an object of interest, it seems natural for humans (and also some primates) to have an efficient behavioral mechanism by drawing attention to objects of interest, i.e., *Pointing* [16]. Such pointing behavior plays the major role in the information delivery and can naturally associate context to a particular object without knowing how to call it, i.e., the novel object that never seen before. Inspired from the pointing behavior and the pointer networks [27], we design a pointing mechanism to deal with the novel objects in image captioning scenario. More precisely, the pointing mechanism is a hybrid between the regular LSTM-based language model plus a copying layer and a pointing behavior. It facilitates directly copying recognized objects, which concentrates on the handling of novel objects, while retraining the ability to generate coherence words in grammar via a language model. The interactions between LSTM plus copying layer and the pointing mechanism is depicted in the left part of Figure 1.

Specifically, in the decoding stage, given the current LSTM cell output $\mathbf{h}^t$ at the t-th time step, two probability distributions over the whole vocabulary $\mathcal{W}$ and the object vocabulary $\mathcal{W}_c$ are firstly calculated with regard to the regular sequence modeling in LSTM and the direct copying of objects in copying layer, respectively. For the probability distribution over the whole vocabulary of LSTM, the corresponding probability of generating any target word $w_{t+1} \in \mathcal{W}$ is measured as

$$\Pr_d^t\left(w_{t+1}\right) = \mathbf{w}_{t+1}^\top \mathbf{M}_d \mathbf{h}^t, \tag{4}$$

where D_h is the dimensionality of LSTM output and $\mathbf{M}_d \in \mathbb{R}^{D_w \times D_h}$ is the transformation matrix for textual features of word. For the probability distribution of being copied over the object vocabulary, we directly achieve the probability of copying any object $w_{t+1} \in \mathcal{W}_c$ conditioned on the current LSTM cell output $\mathbf{h}^t$ and the output of object learners $\mathbf{I}_c$:

$$\Pr_c^t\left(w_{t+1}\right) = \mathbf{w}_{t+1}^\top \mathbf{M}_c^1 \left(\mathbf{I}_c \odot \sigma\left(\mathbf{M}_c^2 \mathbf{h}^t\right)\right), \tag{5}$$

where $\mathbf{M}_c^1 \in \mathbb{R}^{D_w \times D_c}$ and $\mathbf{M}_c^2 \in \mathbb{R}^{D_c \times D_h}$ are the transformation matrices, σ is the sigmoid function and $\odot$ is the element-wise dot product function.

Next, the pointing mechanism encapsulates dynamic contextual information (current input word and LSTM cell output) to learn when to point novel objects for copying, which is applied with feature transformation, to produce a weight value and followed by a sigmoid function to squash the weight value to a range of $[0, 1]$. Such output weight value $\mathbf{p}_t$ in pointing mechanism is computed as

$$\mathbf{p}_t = \sigma(\mathbf{G}_s \mathbf{w}_t + \mathbf{G}_h \mathbf{h}^t + \mathbf{b}_p), \tag{6}$$

where $\mathbf{G}_s \in \mathbb{R}^{D_w}$, $\mathbf{G}_h \in \mathbb{R}^{D_h}$ are the transformation matrices for textual features of word and cell output of LSTM respectively, and $\mathbf{b}_p$ is the bias. Here the weight value $\mathbf{p}_t$ is adopted as a soft switch to choose between generating a word through LSTM, or directly copying a word from the recognized objects. As such, the final probability of each target word w_{t+1} over the whole vocabulary $\mathcal{W}$ is obtained by dynamically fusing the two probability distributions in Eq.(4) and Eq.(5) with the weight value $\mathbf{p}_t$:

$$\begin{gathered} \Pr^t\left(w_{t+1}\right) = \mathbf{p}_d^t \cdot \phi(\Pr_d^t\left(w_{t+1}\right)) + \mathbf{p}_c^t \cdot \phi(\Pr_c^t\left(w_{t+1}\right)), \\ \mathbf{p}_d^t = 1 - \mathbf{p}_t, \;\; \mathbf{p}_c^t = \mathbf{p}_t, \end{gathered} \tag{7}$$

where $\mathbf{p}_d^t$ and $\mathbf{p}_c^t$ denote the weight for generating a word via LSTM or copying a word from recognized objects. ϕ represents a softmax function.

3.4. Coverage of Objects

While high quantitative scores have been achieved through RNN-based image captioning systems in encoder-decoder paradigm, there is increasing evidence [5, 14] revealing that such paradigm still lacks visual grounding (i.e.,

do not associate mentioned concepts to pixels of image). As such, the generated captions are more prone to describe generic information of objects or even copy the most frequently phrases/captions in training data, resulting in undesired effects such as object hallucination. Accordingly, we further measure the coverage of objects as additional training target to holistically cover more objects in the sentence, aiming to emphasis the correctness of mentioned objects regardless of syntax structure and thus improve the captions.

In particular, measuring the coverage of objects is formulated as the multi-label classification problem. We firstly accumulate all the probability distributions of being copied on the object vocabulary generated at decoding stage. The normalized sentence-level probability distribution for copying is thus obtained via aggregating all the probability distributions for copying weighted by the weight value $\mathbf{p}_t$ in pointing mechanism, followed by a sigmoid normalization:

$$\Pr_s(w_{o_i}) = \sigma\left(\sum_{t=1}^{N_s} \mathbf{p}_t \Pr_c^t(w_{o_i})\right). \tag{8}$$

Here the sentence-level probability for each object $w_{o_i} \in \mathcal{W}_c$ represents how possible the object been directly copied in the generated sentence regardless of the position in the sentence. Thus, the sentence-level coverage loss is calculated as the cross entropy loss in multi-label classification:

$$E_c(I, \mathcal{O}) = -\sum_{i=1}^{K} \log \Pr_s(w_{o_i}). \tag{9}$$

By minimizing this sentence-level coverage loss, the captioning system is encouraged to talk about more objects found in the image.

3.5. Optimization

Training. The overall training objective of our LSTM-P integrates the widely-adopted sequential loss in Eq.(2) and sentence-level coverage loss in Eq.(9). Hence we obtain the following optimization problem:

$$\mathcal{L} = -\sum_{t=1}^{N_s} \log \Pr^t(\mathbf{w}_t) - \lambda \sum_{i=1}^{K} \log \Pr_s(w_{o_i}), \tag{10}$$

where λ is tradeoff parameter. With this overall loss objective, the crucial goal of this optimization is to encourage the generated sentence to be coherent in language and meanwhile address all the objects within image.

Inference. In the inference stage, we choose output word among the whole vocabulary $\mathcal{W}$ with maximum probability at each time step with the guidance from pointing mechanism. The embedded textual feature of output word is set as LSTM input for the next time step. This process continues until the end sign word is emitted or the pre-defined maximum sentence length is reached.

4. Experiments

We conduct extensive evaluations of our proposed LSTM-P for novel object captioning task on two image datasets, including the held-out COCO image captioning dataset (**held-out COCO**) [8], a subset of image captioning benchmark—COCO [12], and **ImageNet** [22] which is a large-scale object recognition dataset.

4.1. Dataset and Experimental Settings

Dataset. The **held-out COCO** consists of a subset of COCO by excluding all the image-sentence pairs which contain at least one of eight specific objects in COCO: "bottle," "bus," "couch," "microwave," "pizza," "racket," "suitcase," and "zebra". In this dataset, each image is annotated with five descriptions by humans. Since the annotations of the official testing set are not publicly available, we follow the split in [8] and take half of COCO validation set as validation set and the other half for testing. In the experiments, we firstly train the object learners with all the images in COCO training set including the eight novel objects, and the LSTM is pre-trained with all the sentences from COCO training set. Next, all the paired image-sentence data from held-out COCO training set are leveraged to optimize our novel object captioning system. Our LSTM-P model is finally evaluated over the testing set of held-out COCO to verify the ability of describing the eight novel objects.

ImageNet is the large-scale object recognition dataset and we adopt a subset of ImageNet containing 634 objects that are not present in COCO for evaluation, as in [26]. Specifically, we take about 75% of images in each class for training and the rest for testing. Hence the training and testing sets include 493,519 and 164,820 images in total. In the experiments, we firstly train the object learners with the entire ImageNet training set, and the LSTM is pre-trained with all the sentences from COCO training set. After that, our novel object captioning system is optimized with all the paired image-sentence data from COCO training set. During inference, we directly produce sentences for testing images in ImageNet and evaluate the ability of describing the 634 novel objects for our LSTM-P.

Implementation Details. For fair comparison with other state-of-the-art methods, we take the output of 4,096-dimensional fc7 from 16-layer VGG [23] pre-trained on ImageNet [22] as image representation. Each word in the sentence is represented as Glove embeddings [20]. For the object learners on COCO, we select only the 1,000 most common words from COCO and utilize MIL model [7] to train the object learners over the whole training data of COCO. For the object learners on ImageNet, we directly finetune 16-layer VGG pre-trained on ImageNet to obtain the 634 object learners. The hidden layer size in LSTM is set as 1,024. The tradeoff parameter λ to balance the sequential loss and the sentence-level coverage loss is empirically set

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

Model	$F1_{bottle}$	$F1_{bus}$	$F1_{couch}$	$F1_{microwave}$	$F1_{pizza}$	$F1_{racket}$	$F1_{suitcase}$	$F1_{zebra}$	$F1_{average}$	SPICE	METEOR	CIDEr
LRCN [6]	0	0	0	0	0	0	0	0	0	-	19.3	-
DCC [8]	4.6	29.8	45.9	28.1	64.6	52.2	13.2	79.9	39.8	13.4	21.0	59.1
NOC [26]	14.9	69.0	43.8	37.9	66.5	65.9	28.1	88.7	51.8	-	20.7	-
NBT [14]	7.1	73.7	34.4	61.9	59.9	20.2	42.3	88.5	48.5	15.7	22.8	77.0
Base+T4 [2]	16.3	67.8	48.2	29.7	77.2	57.1	49.9	85.7	54	15.9	23.3	77.9
KGA-CGM [17]	26.4	54.2	42.1	50.9	70.8	75.3	25.6	90.7	54.5	14.6	22.2	-
LSTM-C [33]	29.7	74.4	38.8	27.8	68.2	70.3	44.8	91.4	55.7	-	23.0	-
DNOC [30]	33.0	76.9	54.0	46.6	75.8	33.0	59.5	84.6	57.9	-	21.6	-
LSTM-P$^-$	26.7	74.5	46.2	50.5	81.7	47.2	61.1	91.9	60.0	16.5	23.2	88.0
LSTM-P	28.7	75.5	47.1	51.5	81.9	47.1	62.6	93.0	**60.9**	**16.6**	**23.4**	**88.3**

to 0.3. Following [26], we implicitly integrate the overall energy loss with a text-specific loss on external sentence data for maintaining the model's ability to address novel objects among sentences and a binary classification loss to guide the learning of pointing mechanism. Our novel object captioning model is mainly implemented on Caffe [9], one of widely adopted deep learning frameworks. Specifically, we set the initial learning rate as 0.0005 and the mini-batch size is set as 512. For all experiments, the maximum training iteration is set as 50 epoches.

Evaluation Metrics. To quantitatively evaluate our LSTM-P on held-out COCO dataset, we utilize the most common metrics of image captioning task, i.e., **METEOR** [4], **CIDEr** [25], and **SPICE** [1], to evaluate the quality of generated description. In addition, **F1-score** [8] is adopted to further evaluate the ability of describing novel objects. Note that the metric of F1-score indicates whether the novel object is addressed in the generated sentences of the given image which contains that novel object. In our experiments, for fair comparison, both of the METEOR and F1-score metrics are calculated by utilizing the codes[1] released by [8]. For the evaluation on ImageNet which contains no ground-truth sentences, we follow [26] and adopt another two metrics: describing novel objects (**Novel**) and **Accuracy** scores. Here the Novel score calculates the percentage of all the 634 novel objects addressed in the generated sentences. In other words, for each novel object, the model should mention it within at least one description for the image containing this object. The Accuracy score of each novel object denotes the percentage of images containing this novel object which can be correctly described by mentioning that novel object in generated descriptions. We obtain the final Accuracy score by averaging all the accuracy scores of 634 novel objects.

4.2. Compared Approaches

We compare our LSTM-P model with the following state-of-the-art methods, which include both the regular image captioning methods and novel object captioning models: (1) **LRCN** [6] is a basic LSTM-based captioning model which triggers sentence generation by injecting input image and previous word into LSTM at each time step. We directly train LRCN on the paired image-sentence data without any novel objects. (2) **DCC** [8] leverages external unpaired data to pre-train lexical classifier and language model. Next, the whole captioning framework is trained with paired image-sentence data. (3) **NOC** [26] presents a novel object captioning system consisting of visual recognition network, LSTM-based language model, and image captioning network. The three components are simultaneously optimized in an end-to-end fashion. (4) **NBT** [14] first generates a hybrid template that contains a mix of words and slots associated with image region, and then fills in the slots with detected visual concepts. (5) **Base+T4** [2] designs constrained beam search to force the inclusion of predicted tag words in the output of RNN-based decoder without re-training. (6) **KGA-CGM** [17] takes the predicted concepts as an external semantic attention and constrained inference for sentence generation. (7) **LSTM-C** [33] integrates the standard RNN-based decoder with copying mechanism which can directly copy the predicted objects into the output sentence. (8) **D-NOC** [30] generates the caption template with placeholder and then fill in the placeholder with the detected objects via a key-value object memory. (9) **LSTM-P** is the proposal in this paper. Moreover, a slightly different version of this run is named as **LSTM-P$^-$**, which is trained without the sentence-level coverage loss.

4.3. Performance Comparison

Evaluation on held-out COCO. Table 1 shows the performances of compared ten models on held-out COCO dataset. Overall, the results across all the four general evaluation metrics consistently indicate that our proposed LSTM-P exhibits better performance than all the state-of-the-art techniques including regular image captioning model (LRCN) and seven novel object captioning systems. In particular, the $F1_{average}$ score of our LSTM-P can achieve 60.9%, making the relative improvement over the best com-

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

tennis: 1.00
court: 0.92
racket: 0.78
woman: 0.71
player: 0.69

GT: a woman walking on a tennis court holding a tennis racket
LRCN: a young boy holding a baseball bat on a court
LSTM-P: a tennis player holding a racket on a court

bus: 0.93
people: 0.77
city: 0.49
building: 0.38
street: 0.35

GT: a small group of people that are in front of a bus
LRCN: a woman is standing in front of a truck
LSTM-P: a group of people standing around a bus

dog: 1.00
couch: 0.21
bed: 0.13
blanket: 0.12
head: 0.11

GT: a large dog laying on a blanket on a couch
LRCN: a dog is laying down in a bed
LSTM-P: a dog laying on a couch with a blanket

Figure 2. Objects and sentence generation results on held-out COCO. The detected objects are predicted by MIL model in [7], and the output sentences are generated by 1) Ground Truth (GT): one ground truth sentence, 2) LRCN and 3) our LSTM-P.

lawnmower: 0.97
man: 0.81
grass: 0.78
trees: 0.49
person: 0.27

GT: lawnmower
LRCN: a man walking down a road next to a truck
LSTM-P: a man sitting on a lawnmower in the grass

orangutan: 1.00
grass: 0.95
ground: 0.21
animal: 0.20
face: 0.19

GT: orangutan
LRCN: a brown bear that is in the grass
LSTM-P: a brown orangutan is laying on a grass field

abacus: 1.00
child: 0.53
boy: 0.39
kid: 0.15
baby: 0.14

GT: abacus
LRCN: a little boy sitting in front of a table
LSTM-P: a young child is holding a abacus in his hand

Figure 3. Objects and sentence generation results on ImageNet. GT denotes the ground truth object. The detected objects are predicted by the standard CNN architecture [23], and the output sentences are generated by 1) LRCN and 2) our LSTM-P.

petitor by 5.2%, which is generally considered as a significant progress on this dataset. As expected, by additionally utilizing external object recognition data for training, all the latter nine novel object captioning models outperform the regular image captioning model LRCN on both description quality and novelty. By augmenting the standard RNN-based language model with the object/concept learners, LSTM-C leads to a performance boost against NOC that produces novel objects purely depending on generative mechanism in LSTM. The results basically indicate that the advantage of directly "copying" the predicted objects/concepts into output sentence via copying mechanism. However, the performances of LSTM-C are still lower than our LSTM-P$^-$, which leverages the pointing mechanism to balance the influence between copying mechanism and standard word-by-word sentence generation conditioned on the contexts. This confirms the effectiveness of elegantly pointing when to copy the novel objects to target sentence for novel object captioning. In addition, by further integrating sentence-level coverage loss into overall training objective, LSTM-P exhibits better performance than LSTM-P$^-$, which demonstrates the merit of encouraging the generated sentence to be coherent in language and meanwhile address all the objects within image.

Evaluation on ImageNet. To further verify the scalability of our proposed LSTM-P, we additionally perform experiment on ImageNet to describe hundreds of novel objects that outside of the paired image-sentence data. Table 2 shows the performance comparison on ImageNet. Similar to the observations on held-out COCO, our LSTM-P exhibits better performance than other runs. In particular, the Novel, F1, and Accuracy scores for LSTM-P can

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 [26]			
-COCO	69.08	15.63	10.04
-BNC&Wiki	87.69	31.23	21.96
LSTM-C [33]			
-COCO	72.08	16.39	11.83
-BNC&Wiki	89.11	33.64	31.11
LSTM-P			
-COCO	90.06	17.67	11.91
-BNC&Wiki	91.17	52.07	44.63

reach 90.06%, 17.67%, and 11.91%, making the relative improvement over LSTM-C by 24.9%, 7.8%, and 0.7%, respectively. The results basically indicate the advantage of exploiting pointing mechanism to balance the word generation from decoder and the word copied from learnt objects, and the global coverage of objects in output sentence, for novel object captioning, even when scaling into ImageNet images with hundreds of novel objects. Moreover, we follow [26, 33] and include the external unpaired text data (i.e., British National Corpus and Wikipedia) for training our LSTM-P. The performance gains are further attained.

4.4. Experimental Analysis

In this section, we further analyze the qualitative results, the weights visualization in pointing mechanism, and the effect of the tradeoff parameter λ for novel object captioning task on held-out COCO dataset.

Qualitative Analysis. Figure 2 and Figure 3 showcase a few sentence examples generated by different methods, the detected objects and human-annotated ground truth on

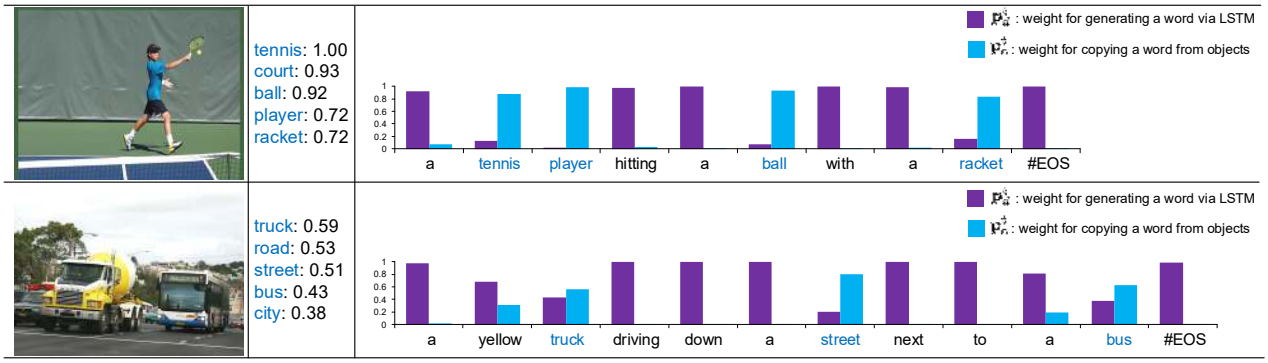
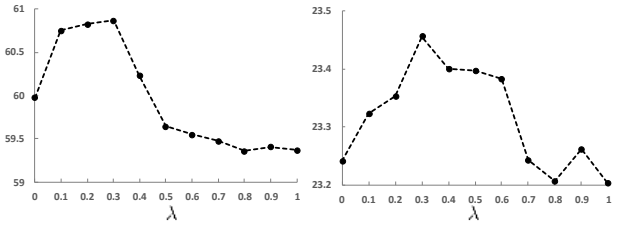


Figure 4. Sentence generation results with visualized weights learnt in pointer mechanism of our LSTM-P at each decoding step on held-out COCO dataset. The bar plot at each decoding step corresponds to the weights for generating a word via LSTM or copying a word from recognized objects when the corresponding word was generated.



(a) $F1_{average}$ (b) METEOR

Figure 5. The effect of the tradeoff parameter λ in our LSTM-P over (a) $F1_{average}$ (%) and (b) METEOR (%) on held-out COCO.

held-out COCO and ImageNet, 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-P can correctly describe the novel objects by learning to point towards the placements and moments of copying novel objects via pointing mechanism. For example, compared to object term "bed" in the sentence generated by LRCN, "couch" in our LSTM-P is more precise to describe the image content in the last image on held-out COCO dataset, since the novel object "couch" is among the top object candidates and directly copied to the output sentence at the corresponding decoding step. Moreover, by additionally measuring the coverage over the bag-of-objects on sentence level, our LSTM-P is encouraged to produce sentences which cover more objects found in images, leading to more descriptive sentence with object "blanket."

Visualization of weights in pointing mechanism. To better qualitatively evaluate the generated results with pointing mechanism of our LSTM-P, we further visualize the generated weights of generating a word via LSTM or copying a word from recognized objects for a few examples in Figure 4. We can easily observe that our LSTM-P correctly chooses to copy word from recognized objects when the object word to be generated. For instance, in the first image, when LSTM-P is about to generate object word (i.e., "tennis," "player," "ball," and "racket"), it mostly prefer to copy the object word from recognized objects with higher weight value $\mathbf{p}_c^t$. Also, for the second video, the pointer mechanism attends to direct copying from objects when the object terms (i.e., "truck," "street," and "bus") are about to be generated at decoding stage.

Effect of the Tradeoff Parameter λ. To clarify the effect of the tradeoff parameter λ in Eq.(10), we illustrate the performance curves over two evaluation metrics with a different tradeoff parameter in Figure 5. As shown in the figure, we can see that all performance curves are generally like the "$\wedge$" shapes when λ varies in a range from 0 to 1. Hence we set the tradeoff parameter λ as 0.3 in our experiments, which can achieve the best performance. This again proves that it is reasonable to encourage both the syntactic coherence and the global coverage of objects in the output sentence for boosting novel object captioning.

5. Conclusions

We have presented Long Short-Term Memory with Pointing (LSTM-P) architecture which produces novel objects in image captioning via pointing mechanism. Particularly, we study the problems of how to facilitate vocabulary expansion and how to learn a hybrid network that can nicely integrate the recognized objects into the output caption. To verify our claim, we have initially pre-trained object learners on free available object recognition data. Next the pointing mechanism is devised to balance the word generation from RNN-based decoder and the word taken directly from the learnt objects. Moreover, the sentence-level coverage of objects is further exploited to cover more objects in the sentence and thus improve the captions. Experiments conducted on both held-out COCO image captioning and ImageNet datasets validate our model and analysis. More remarkably, we achieve new state-of-the-art performance of single model: 60.9% in $F1_{average}$ score on held-out COCO dataset.

Acknowledgments. This work is partially supported by NSF of China under Grant 61672548, U1611461, 61173081, and the Guangzhou Science and Technology Program, China, under Grant 201510010165.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

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

[5] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *EMNLP*, 2016.

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[7] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015.

[8] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016.

[9] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[11] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, 2018.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[13] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.

[14] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018.

[15] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *ICCV*, 2015.

[16] Danielle Matthews, Tanya Behne, Elena Lieven, and Michael Tomasello. Origins of the human pointing gesture: a training study. *Developmental science*, 2012.

[17] Aditya Mogadala, Umanga Bista, Lexing Xie, and Achim Rettinger. Describing natural images containing novel objects with knowledge guided assitance. *arXiv preprint arXiv:1710.06303*, 2017.

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

[19] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, 2017.

[20] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[21] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

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

[25] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[26] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *CVPR*, 2017.

[27] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *NIPS*, 2015.

[28] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[29] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR*, 2016.

[30] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. *arXiv preprint arXiv:1804.03803*, 2018.

[31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[32] Ting Yao, Yehao Li, Zhaofan Qiu, Fuchen Long, Yingwei Pan, Dong Li, and Tao Mei. Msr asia msm at activitynet challenge 2017: Trimmed action recognition, temporal action proposals and dense-captioning events in videos. In *CVPR ActivityNet Challenge Workshop*, 2017.

[33] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*, 2017.

[34] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[35] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, 2017.

[36] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.