Switchable Novel Object Captioner

Yu Wu, Lu Jiang, and Yi Yang

Abstract—Image captioning aims at automatically describing images by sentences. It often requires lots of paired image-sentence data for training. However, trained captioning models can hardly be applied to new domains in which some novel words exist. In this paper, we introduce the zero-shot novel object captioning task, where the machine generates descriptions about novel objects without extra training sentences. To tackle the challenging task, we mimic the way that babies talk about something unknown, i.e., using the word of a similar known object. Following this motivation, we build a key-value object memory by detection models, containing visual information and corresponding words for objects in the image. For those novel objects, we use words of most similar seen objects as proxy visual words to solve the out-of-vocabulary issue. We then propose a Switchable LSTM that incorporates knowledge from the object memory into sentence generation. The model has two switchable working modes, i.e., 1) generating the sentences like standard LSTMs and 2) retrieving proper nouns from the key-value memory. Thus our model is learned to fully disentangle language generation from training objects, and requires zero training sentence in describing novel objects. Experiments on three large-scale datasets demonstrate the ability of our method to describe novel concepts.

Index Terms—Image captioning, novel object captioning, zero-shot learning

1 Introduction

As a classical task in vision and language research, image captioning aims at automatically describing an image using natural language sentences or phrases. Encoder-decoder architectures prove to be a common framework for the image captioning task [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], in which the Convolutional Neural Networks (CNN) are often used as the image encoder, and the decoder is usually a Recurrent Neural Network (RNN) to sequentially predict the next word given the previous word.

As captaining models are trained on parallel data of image-sentence pairs, they fail to caption words if these words do not exist in the training sentences. In recent years, a pivotal research direction in image captioning research is to generalize captioning models to describe novel objects that only occur at test time. For example, as illustrated in Fig. 1, although the captioning model (LRCN [4]) is able to correctly generate captions for the object "truck", it fails for a similar object "bus" merely because the training sentences do not contain any word of bus.

A few works have been proposed to address this problem [11], [12], [13]. Generally, these methods attempt to improve model generalization by incorporating external linguistic knowledge about the new object. This is achieved by either using pre-trained language models [11], [12] or additional unpaired training sentences of the novel objects. For example, Henzdricks *et al.* [11] trained a captioning model by leveraging a pre-trained image tagger and a pre-trained language sequence model from external text corpora.

Existing works mitigate this problem by removing the dependency to parallel training data of paired images and sentences, which turns out to be very difficult to collect. The precise definition of *novel object* in existing works is that the object is *unseen* in the *paralleled* training sentences but still needs to exist in the training data in the form of the *unparalleled* sentences. In other words, they all assume training sentences of novel objects always exist during training. This assumption, however, does not hold in many real-world scenarios. The descriptions are typically rare, in a timely manner, for brand new products such as self-balancing scooters, robot vacuums, drones. Moreover, perhaps more importantly, the language generation are learned closely coupled with seen objects, and hence will inevitably introduce language biases to the captioning model. For example, if training sentences are all about the bass (a sea fish), the captioning model will never learn to caption the instrument bass, and may generate awkward sentences like "A man is eating a bass with a guitar amplifier."

In this paper, we tackle the image captioning for novel objects in which zero training sentences of novel objects are needed. We call it *zero-shot* novel object captioning to distinguish it from the traditional problem setting in [11], [12], [13], [14]. In the traditional setting, extra training sentences of the novel object are provided in addition to the pre-trained object detection model. In the zero-shot captioning setting, there are *zero training sentences* about the novel object, i.e., there is no information about the object's semantic meaning, sense, or context. The only external knowledge in the proposed setting is a pre-trained object detection model that can detect the novel object, which is also required in the traditional problem setting.

- *Yu Wu is with Baidu Research, Beijing 100000, China, and also with the School of Computer Science, Princeton University, Princeton, NJ 08540 USA. E-mail: yw5952@princeton.edu.*
- *Lu Jiang is with Google Research, Mountain View, CA 94043 USA. E-mail: lujiang@google.com.*
- *Yi Yang is with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310058, China. E-mail: yangyics@zju.edu.cn.*

Manuscript received 24 June 2020; revised 7 Jan. 2022; accepted 18 Jan. 2022. Date of publication 25 Jan. 2022; date of current version 5 Dec. 2022. (Corresponding author: Yi Yang.) Recommended for acceptance by L. Wang. Digital Object Identifier no. 10.1109/TPAMI.2022.3144984

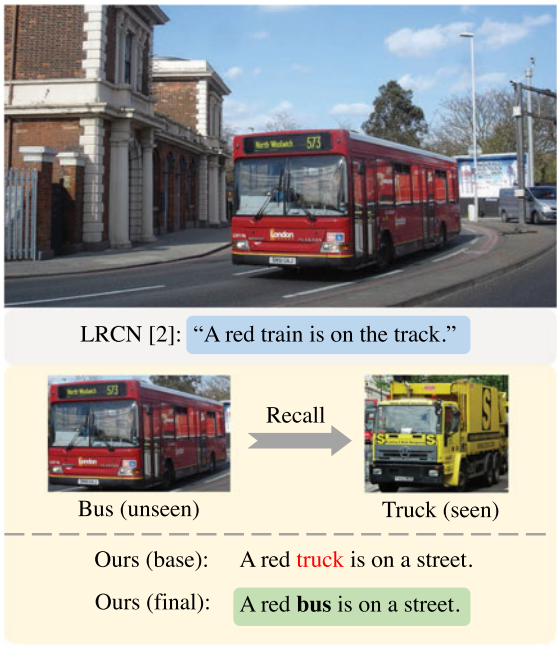


Fig. 1. An example of the novel object captioning. Suppose the novel object "bus" is not present in the training data. Traditional image captioning model LRCN [4] fails to describe the image with the novel object "bus". Our algorithm can generate precise captions, and *more importantly*, we do not need any training data containing bus. Specifically, when seeing the unseen object, we recall the most similar object "truck", which is a seen object in training. Then our designed Switchable LSTM is capable of incorporating the object name into sentence generating. It first generates an inaccurate sentence using known knowledge ("A red truck is on a street"), and then correct it by the exact word provided by the external detection model.

Since novel objects are completely unseen during training, zero-shot captioning presents a new challenge to disentangle language generation from visual detection. To tackle this challenge, we propose a solution mimicking the way of babies talking. When describing an unseen object, a baby tries her best to use similar objects that she has seen before. For example, a baby might say "a horse is standing in a field" to actually describe a *zebra*. The sentence will be accurate if we further replace the ambiguous word "horse" with the exact word "zerba".

Following this motivation, we propose a framework called Switchable Novel Object Captioner (SNOC) that aims at generating natural language sentences disentangled from training object classes in order to describe novel objects at test time. Unlike existing works, our model is learned to fully disentangle language generation from training objects, thus requiring zero training sentences for the novel object. SNOC follows the standard encoder-decoder architecture but with a novel decoder. At the decoding stage, we first build a key-value object memory by a detection model, containing the visual information and the corresponding word for each object shown in the image. For those novel words, we use the word of the most similar seen objects instead. We call it *proxy* visual word. Then we propose a Switchable LSTM that incorporates the object memory into the sentence generation. The Switchable LSTM switches between two working modes, i.e., 1) generating a sentence as a standard LSTM [15] and 2) retrieving a proper noun from the key-value memory, controlled by our newly designed indicator in the LSTM cell. Finally, through the Switchable LSTM, we first generate an inaccurate sentence only using the seen words, and then replace the proxy visual word by the true object label.

For example, in Fig. 1, the object "bus" is unseen in the training stage. Our method describes the unseen object "bus" in a coarse-to-fine manner. It first recalls its most similar object from the training data, i.e., "truck" in this case. Next, it generates an inaccurate sentence using known knowledge of the proxy visual word "truck", and finally correct it by the exact word "bus" provided by the external detection model.

The proposed model is based on our previous work DNOC [3], in which we directly use a special token "<PL>" to represent all unseen words. However, this strategy ignores the visual appearance of novel objects as it is ambiguous to use one token word (i.e., <PL>) to represent all the novel objects. We made two important extensions to DNOC. First, we replace the placeholder token by the proxy visual word, which helps generate a better sentence by leveraging visual similarities between novel objects and seen objects. We extensive expand the experiments and analysis of the proposed method. We additionally evaluate our methods on two large-scaled datasets, ImageNet and nocaps. We also tried different variants of our models, e.g., using different language models and different object detection models. We also tested the Reinforcement Learning (RL) to directly train our model for optimization on the language metrics (CIDer and METEOR).

Experiments on three representative datasets show that our method is effective for zero-shot novel object captioning. Without extra training data, our model even significantly outperforms state-of-the-art methods (with additional training sentences) on the F1-score metric.

In summary, the main contributions of this work are listed as follows:

- We introduce the zero-shot novel object captioning task, an important yet neglected research direction of image captioning.
- To generate sentences with correct word orders, we have made efforts from the following three aspects We first design the switchable LSTM to figure out *where* we should place the object words (by the switch indicator).
- We then take the semantic information from the LSTM hidden states to find *which* visual object should be mentioned here from all the recognized object memory.
- To ensure the consistency in sentences and alleviate out-of-vocabulary issue, we design the proxy visual words and avoid the *unknown impact* brought by the imported novel object labels on LSTM.

2 Related Work

2.1 Image Captioning

Automatic caption generation is the task of describing the content of an image by a complete and natural sentence. This is a fundamental problem in the multi-modal perception field [1], [9], [16], [17], [18], [19], [20]. Some early works such as template-based approaches [21], [22] and search-based

approaches [23], [24] generate captioning by the sentence template and the sentence pool. Recently, inspired by deep learning and sequence modeling in computer vision, language-based models have achieved promising performance. Most of them are based on the encoder-decoder architecture to learn the probability distribution of both visual embedding and textual embeddings [2], [4], [5], [6], [7], [8], [25], [26], [27], [28], [29], [30]. In this architecture, the encoder is a CNN model that processes and encodes the input image into an embedding representation, while the decoder is an RNN model that takes the CNN representation as the initial input and sequentially predicts the next word given the previous word. Among recent contributions, Kiros *et al.* [18] proposed a multi-modal log-bilinear neural language model to jointly learn word representations and image feature embeddings. Vinyals *et al.* [8] proposed an end-to-end neural network consisting of a vision CNN followed by a language generating RNN. Xu *et al.* [9] improved [8] by incorporating the attention mechanism into captioning. The attention mechanism focuses on the salient image regions when generating corresponding words. In general, these methods are designed to describe seen objects with lots of training examples. The vocabulary of the decoder is fixed after training and can not be further extended by external knowledge. Recently, in SCST [31], all the objects and words are existing in training, while the combination of them are unusual in testing (e.g., a blue boat in front of a building). However, we focus on a more challenging task that these objects and words do not exist in training.

2.2 Novel Object Captioning

Novel object captioning is a challenging task where there is no paired visual-sentence data for the novel object in training. Only a few works have been proposed to address this captioning problem. Henzdricks *et al.* [11] proposed the Deep Compositional Captioner (DCC), a pilot work to address the task of generating descriptions of novel objects which are not present in paired image-sentence datasets. Venugopalan *et al.* [12] discussed a Novel Object Captioner (NOC) to further improve the DCC to an end-to-end system by jointly training the visual classification model, language sequence model, and the captioning model. Anderson *et al.* [14] leveraged an approximate search algorithm to forcibly guarantee the inclusion of selected words during the evaluation stage of a caption generation model. Yao *et al.* [13] exploited a mechanism to copy the detection results to the output sentence with a pre-trained language sequence model. Lu *et al.* [32] also proposed to generate a sentence template with slot locations, which are then filled in by visual concepts from object detectors. Wang *et al.* [33] proposed a new zero-shot video captioning that aims at describing out-of-domain videos by composing different experts based on different topic embeddings and implicitly transfer the knowledge learned from seen activities to unseen ones. Feng *et al.* [34] proposed a cascaded revision module to generate better sentences by considering both visual similarity and semantic similarity on uncertainty words. Agrawal *et al.* [35] collected a large-scaled novel object captioning dataset, and extend existing novel object captioning models to establish strong baselines. Cao *et al.* [36] proposed to adapt the captioning model to the novel object features detected via the auxiliary detection.

Note that all of the above methods have to use extra data of the novel object to train their word embedding. Different from existing methods, our method focuses on the zero-shot novel object captioning task in which there are no additional sentences or pre-trained models to learn such embeddings for novel objects.

2.3 Zero-Shot Novel Object Captioning

Zero-shot learning aims to recognize objects whose instances may not have been seen during training [33], [37], [38], [39], [40], [41]. Zero-shot learning bridge the gap between the visual and the textual semantics by learning a dictionary of concept detectors on external data sources [42]. Recently, some works focus on the zero-shot novel object captioning task, where there are no additional training sentences available in learning to caption novel objects. Wu *et al.* [3] proposed a decoupled captioning framework DNOC to generate a sentence template, which enables the model to freely introduce the novel object labels into the generated sentence template. DNOC simply uses a special token to represent all the novel objects, leading to ambiguous captioning results. Differently, we propose to leverage visual similarities between novel objects and seen objects to generate more accurate sentences. In addition, instead of the standard LSTM used in DNOC, we propose to improve the LSTM cell with flexible working modes, enabling it to utilize both existing and external knowledge.

3 The Proposed Method

3.1 Preliminaries

Given an input image $\mathbf{I}$, the goal is to generate an associated natural language sentence $\mathbf{s}$ to describe the image. A sentence $\mathbf{s}$ containing N_l words is denoted as $\mathbf{s} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n_l})$, where $\mathbf{w}$ represents a word. The length N_l usually varies for different sentences. The training data is a set of image-sentences pairs $\mathcal{P} = \{(\mathbf{I}_1, \mathbf{s}_1), \ldots, (\mathbf{I}_{n_p}, \mathbf{s}_{n_p})\}$. The vocabulary of $\mathcal{P}$ is $\mathcal{W}_p = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{N_t}\}$ which contains all N_t unique words in the training sentences.

We represent each word $\mathbf{w}_i \in \{0,1\}^{N_t}$ as a one-hot vector of N_t dimension and then embed it into a D_w-dimensional real-valued vector $\mathbf{x}_i = \phi_w(\mathbf{w}_i) \in \mathbb{R}^{D_w}$. The embedding function $\phi_w(\cdot)$ is a linear transformation $\mathbf{x}_i = \mathbf{T}_w \mathbf{w}_i$, where $\mathbf{T}_w \in \mathbb{R}^{D_w \times N_t}$ is the learnable embedding matrix. Our model follows the classical encoder-decoder architecture for image captioning.

The Encoder. We obtain the representation for an input image $\mathbf{I}$ by $\phi_e(\mathbf{I})$, where $\phi_e(\cdot)$ is the visual encoder function. ϕ_e is implemented as an ImageNet pre-trained CNN model with the classification layer removed. In our experiments, we use a 16-layer VGG [43] pre-trained on the ImageNet ILSVRC12 dataset [22] as the visual encoder ϕ_e.

The Decoder. The decoder is a word-by-word sequence model that recurrently predicts the next word given the previous word and encoder features as input,

$$p(\mathbf{s}|\mathbf{I}) = \prod_{t=1}^{n_l} p(\mathbf{w}_t|\mathbf{w}_0, \ldots, \mathbf{w}_{t-1}, \phi_e(\mathbf{I})). \quad (1)$$

The Long Short-Term Memory (LSTM) [15] is a classical decoder in visual captioning and natural language processing tasks [11], [13], [44]. Given the inputs $\mathbf{x}_t$ and hidden

states $\mathbf{h}_{t-1}$, we get the predicted output word $\mathbf{o}_t$ by updating the LSTM unit at time step t as follows:

$$\mathbf{o}_t, \mathbf{h}_t = \texttt{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}). \tag{2}$$

In training, we feed the ground-truth word as the model input. In the evaluation stage, we take the previous output $\mathbf{o}_{t-1}$ of the model as the input $\mathbf{x}_t$ at the t-th step.

Zero-Shot Novel Object Captioning. We study the zero-shot novel object captioning task, where the model needs to caption novel objects *without* additional training sentence data about the object. The novel object words are shown neither in the paired image-sentence training data $\mathcal{P}$ nor unpaired sentence training data. A notable challenge for this task is to deal with the out-of-vocabulary (OOV) words. The learned word embedding function ϕ_w is unable to encode the unseen words, since these words cannot simply be found in the training vocabulary. As a result, these unseen words cannot be fed into the decoder for caption generation. Previous works [11], [12], [13] circumvented this problem by learning the word embeddings of unseen words using additional sentences that contain the words. However, in our zero-shot novel object captioning task, we do not assume the availability of additional training sentences of the novel object.

3.2 Building the Key-Value Object Memory

To describe the image with novel objects, we use a pre-trained object detection model as the external knowledge source, which provides the object name information for objects in the input image. Specifically, for the i-th detected object $\texttt{obj}_i$, we extract its CNN feature $\mathbf{f}_i \in \mathbb{R}^{1\times N_f}$ from the ROI pooling layer of the detection model. Then the CNN features $\mathbf{f}_i$ and the predicted semantic class labels $\mathbf{l}_i \in \mathbb{R}^{1\times N_D}$ are used to form a key-value pair, with the CNN feature as key and the label as value. N_D is the number of detection candidates.

We build a key-value object memory $\mathcal{M}$ using these detected key-value pairs, which associates the semantic class labels (descriptions of the novel objects) with their visual appearance. The maximum memory size is set to N_M. For those images with more than N_M detect objects, we select the top N_M detected objects in an image into the memory, according to the object detection confidences. The memory $\mathcal{M}$ is re-initialized for each input image. It contains all the detected objects in the input image, including both seen and zero-shot objects. During generating the captioning sentence, the memory $\mathcal{M}$ is kept fixed during the recurrent words generation process.

There are two kinds of objects in the key-value object memory $\mathcal{M}$ during the evaluation, i.e., the *seen objects* that show up during training, and the *novel objects* that have never seen in the memory before. For seen objects, we simply write the feature-name pairs into the memory by,

$$\mathcal{M} \leftarrow \texttt{WRITE}(\mathcal{M}, (\mathbf{f}_i, \mathbf{l}_i)), \tag{3}$$

where `WRITE` operation is to insert the key-value pair into a new slot of the existing memory $\mathcal{M}$.

The Proxy Visual Words. For novel objects, we propose to use the *proxy visual word* as instead to mitigate the out-of-vocabulary issue. The main idea is to represent an unseen object by some known objects that have a similar visual appearance. Specifically, for each object that shows in the training data, we extract the visual representation of the image patch. Then these features are clustered according to their object labels. By simply averaging the visual features of objects that belong to the same class, we obtain the prototypical visual representation $\mathbf{v}_o$ for each seen object category, where o indicates the o-th object category. When meeting a novel word, we take the visual feature $\mathbf{f}_i$ of the new image patch to find a most similar one within the prototypical representation set $\{\mathbf{v}_o\}$. Therefore, we have the similarity between the novel object $\texttt{obj}_i$ and the o-th class,

$$s_i^o = \texttt{cosine}(\mathbf{f}_i, \mathbf{v}_o), \tag{4}$$

where $\texttt{cosine}(\cdot)$ denotes the cosine distance function. The similarity s_i^o is defined by the cosine distance between the feature of a novel object and the prototypical feature of the seen object class. By searching over the seen objects database, we can find the most similar object category $\hat{\mathbf{l}}_i$ for the novel object. We name it the proxy visual word to distinguish from the accurate words for those seen objects. Therefore, for the novel object $\texttt{obj}_i$, we insert the pair of the visual feature $\mathbf{f}_i$ and the proxy visual word $\hat{\mathbf{l}}_i$ to the memory,

$$\mathcal{M} \leftarrow \texttt{WRITE}(\mathcal{M}, (\mathbf{f}_i, \hat{\mathbf{l}}_i)). \tag{5}$$

3.3 Switchable LSTM

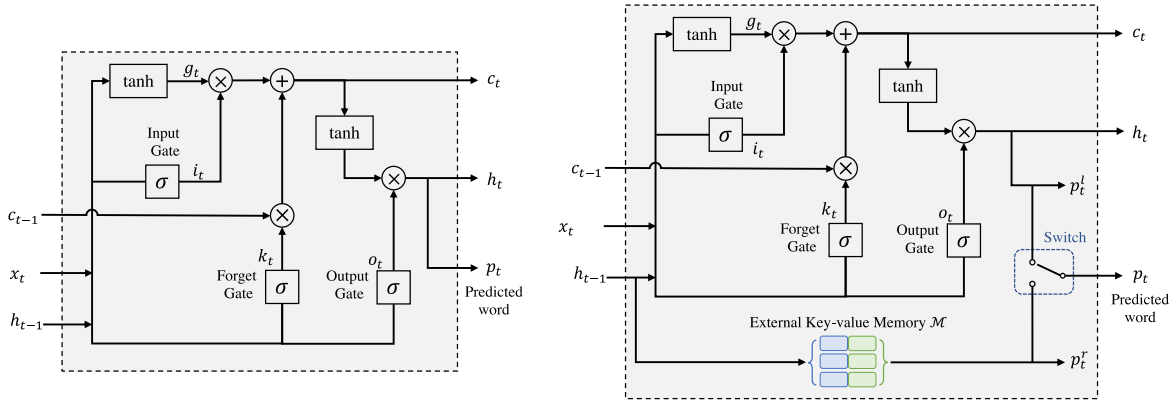
In the zero-shot novel object captioning task, the language model is supposed to leverage both the existing knowledge and external knowledge. Therefore, we propose a Switchable LSTM with two working modes to leverage both knowledge sources. Different from the standard LSTM, our Switchable LSTM operates switching between two modes, i.e., 1) the `Generating` mode, in which the model generates a common word like a standard LSTM; and 2) the `Retrieving` mode, in which retrieving a noun from the key-value object memory $\mathcal{M}$. In the `Generating` mode, we use the memory cell from the standard LSTM to generate the sentence based on the existing knowledge. While in the `Retrieving` mode, instead of generating words, we propose to apply content-based addressing on the object memory to find a proper noun word with the external knowledge. An indicator inside the LSTM cell switches the two modes.

3.3.1 Standard LSTM Revisit

Given the hidden state $\mathbf{h}_t$ the prediction $\mathbf{p}_t^l$ of the LSTM cell at t-th step is,

$$\mathbf{p}_t^l = \mathbf{W}_p \mathbf{h}_t + \mathbf{b}_p. \tag{6}$$

For the image captioning task, the hidden state $\mathbf{h}_0$ is initialized to be the encoded image feature $\phi_e(\mathbf{I})$, and the initial input $\mathbf{x}_0$ is a special token <GO>. Then the LSTM recurrently outputs a word and takes this word as a new input for the next step. The recurrent word generation process is terminated if the model outputs a special token <EOS>. We name the process of generating words based on via Eqn. (6) as the `Generating` mode.



(a) Traditional LSTM (b) Switchable LSTM

Fig. 2. Comparison of the traditional LSTM (the left one) and our Switchable LSTM (the right one).

3.3.2 Retrieving Nouns From External Knowledge

The standard LSTM does not consider the external knowledge information when generating a captioning. To address this issue, we propose an attention-based operation to incorporate the knowledge from the object memory $\mathcal{M}$ into the sentence generation. We name it as the `Retrieving` mode to distinguish from the standard LSTM workflow. In the `Retrieving` mode, we use the hidden state $\mathbf{h}_{t-1}$ as a semantic query to search the object memory $\mathcal{M}$. The query retrieves a matched noun as the predicted word at this time step. The overall `Retrieving` mode may be regarded as a grounding operation that connects the semantic language representation and the visual CNN feature.

Specifically, at the t-th time step, we define the query $\boldsymbol{q}_t$ to be a linear transformation of previous hidden state $\mathbf{h}_{t-1}$,

$$\boldsymbol{q}_t = \mathbf{W}_q \mathbf{h}_{t-1}, \tag{7}$$

where $\mathbf{h}_{t-1}$ is the previous hidden state at $(t-1)$-th step from the sequence model, and $\mathbf{W}_q$ is a linear transformation that converts the hidden state from language semantic space to CNN visual feature space. With this query, we conduct content-based addressing operations on the object memory $\mathcal{M}$, aiming at finding related object information according to the similarity metric. Formally, the content-based addressing operation is defined as,

$$\boldsymbol{p}_t^r = (\boldsymbol{q}_t \boldsymbol{K}^T)\boldsymbol{V}, \tag{8}$$

where $\boldsymbol{K}^T$ and $\boldsymbol{V}$ are the vertical concatenations of all keys and values in the memory, respectively. The output $\boldsymbol{p}_t^r \in \mathbb{R}^{N_D}$ is a softened addressing on all semantic labels candidates. In evaluation, we take the word with the max probability as the query result.

3.3.3 Modes Switching

We design a switch inside the memory cell to control the two working modes of the Switchable LSTM. The comparison of our proposed Switchable LSTM and the traditional LSTM is illustrated in Fig. 2.

The *switch indicator* a_t at the t-th step is based on the hidden state h_{t-1}.

$$a_t = \mathbf{W}_a \mathbf{h}_t + \boldsymbol{b}_a, \tag{9}$$

where $\mathbf{W}_a$ and $\boldsymbol{b}_a$ are the trainable weight and bias, respectively. The switch indicator is designed to estimate the probability of choosing the `Retrieving` mode at the current time step. We then compare the switch indicator with the prediction from the `Generating` mode. Denote the max probability in $\mathbf{p}_t^l$ as p_t^l, if p_t^l is greater than the switch indicator a_t, we choose to predict the word based on Eqn. (6) (the `Generating` mode); otherwise, we turn on the switch and leverage the object memory to find a proper noun word using Eqn. (8) (the `Retrieving` mode). Thus the output of our Switchable LSTM at t-th step is,

$$\mathbf{p}_t = \begin{cases} \mathbf{p}_t^l & \text{if } p_t^l > a_t, \\ \mathbf{p}_t^r & \text{otherwise.} \end{cases} \tag{10}$$

3.4 Framework Overview

With the design of the proxy visual words, regardless of the existence of novel words, the word embedding function ϕ_w is able to encode all the input tokens. We hence can generate a naive sentence for novel objects using seen words via our Switchable LSTM. Finally, we replace the proxy word by the exact name of the novel object, which is provided by the external object detection model. In this way, our model does not require to see these novel words in training, and addresses the critical limitation in prior works.

The following steps are used to generate a correct captioning sentence:

1) We utilize the external object detection model to build a key-value object memory $\mathcal{M}$ for the input image. To circumvent the out-of-vocabulary problem, for an unseen object, we use the label of its most similar seen object as the proxy visual word.
2) We then exploit the Switchable LSTM to generate a captioning sentence. The model jointly operates in two working modes to leverage both internal knowledge and external knowledge. A switch indicator inside the memory cell is used to control the two modes. In the `Retrieving` mode, the predicted word is generated by a soft content-based addressing on the memory $\mathcal{M}$.
3) Finally, we replace the placeholders of the sentence by corresponding object descriptions.

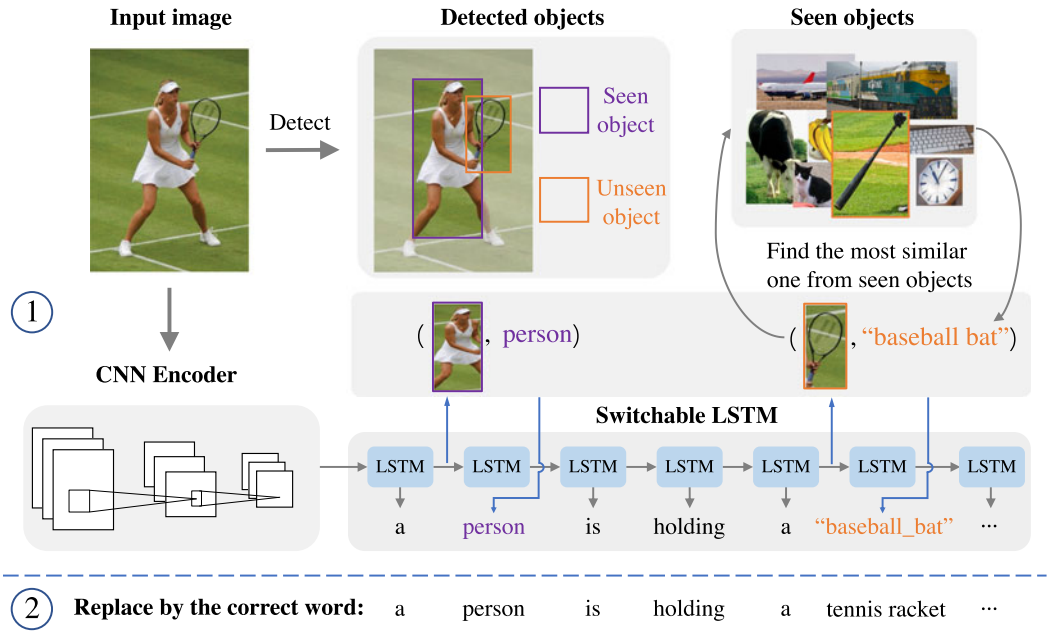


Fig. 3. The overview of the proposed method. In the example, the object "tennis racket" is unseen during training. We first leverage the object detection model to build the key-value object memory. For the unseen object "tennis racket", we find its most similar candidate from seen objects by calculating the visual feature distance. The most similar object is "baseball bat" in this case, which is used in building the memory. The Switchable LSTM takes advantage of both the global image feature and the object memory as input. When predicting the second word ("person"), the indicator insides the cell turns on the *Retrieving* mode. Thus the model takes the hidden state as a query to locate the object memory. When the sentence generation is finished, we replace the proxy visual words by its accurate label name provided by the external detection model.

Taking the input image in Fig. 3 as an example, suppose the object "tennis racket" is the novel object. We first leverage the object detection model and build the key-value object memory $\mathcal{M}$ based on the detection results, which contains both the visual information and the corresponding word (the detection class label). For the unseen object "tennis racket", we find its most similar candidate from seen objects by calculating the distance between visual features, which is "baseball bat" in this case. We then use the name of "baseball bat" in building the memory. Next, our Switchable LSTM takes advantage of both the global image feature and the object memory as input. When predicting the second word ("person"), the switch indicator insides the cell turns on the `Retrieving` mode. Thus our model takes the hidden state $\mathbf{h}_1$ as a query to locate the object memory $\mathcal{M}$. Our model finds an accurate noun "person" for the referred object via Eqn. (8). The LSTM recurrently takes the previous output as the input for the next step. When the sentence generation is finished, we replace the proxy visual word ("baseball bat") by its accurate label name ("tennis racket").

3.5 Training

How to correctly integrate the object information in sentence generating is the core of the zero-shot novel object captioning problem. Towards this target, we propose to simulate the mode switching by putting all object words into the retrieving mode during during training. In other words, we regard all the detected objects as "novel objects" when optimizing the retrieving module.

To be specific, we take all the words of detected objects into the vocabulary of the retrieving mode, including those *seen* objects such as "apple" and "cat". When meeting an object word in training, we train our model to activate the `retrieving` mode via the switch indicator. Otherwise, we optimize the model to activate the `Generating` mode for those common words other than object words. In this way, the model would learn to seek the external detection knowledge for help if it wants to mention an object in the image. This training strategy enables our method to activate the retrieving mode even we don't know the novel objects before.

4 Experiments

4.1 Datasets

The Held-Out MSCOCO Dataset. MSCOCO is a large-scale image captioning benchmark containing 123,287 images. For each image in MSCOCO, there are five human-annotated paired sentence descriptions. Following [11], [12], [13], [14], we employ a subset of the MSCOCO dataset, namely the held-out MSCOCO dataset [11], to evaluate the model's capability in describing the novel objects. The held-out MSCOCO dataset excludes all images that contain at least one of the eight MSCOCO objects. The eight objects are chosen by clustering the word2vec embeddings over all the 80 objects in MSCOCO detection challenge. It results in the final eight novel objects for evaluation, which are *"bottle"*, *"bus"*, *"couch"*, *"microwave"*, *"pizza"*, *"racket"*, *"suitcase"*, *and "zebra"*. These eight objects are held-out in the training split and appear only in the test split. For a fair comparison, we use the same training, validation, and test split as in [11].

Note that although there are no training sentences about the novel object, visual information of the novel object may exist in the training set. We have manually checked the training data and found there are a few images contain novel objects but without annotated sentences in the training set. These novel objects are not salient in these images since the five human annotators did not mention it during annotating. Specifically, we found only **15** images containing the zebra

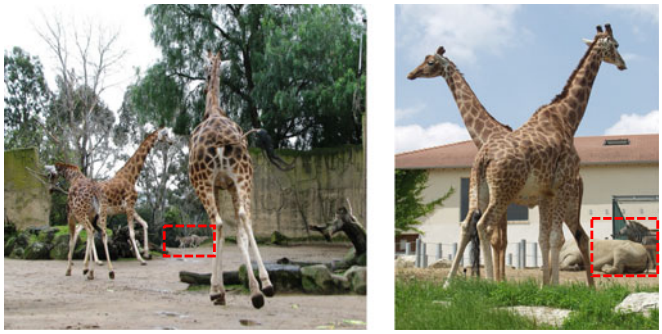


Fig. 4. Training examples containing the novel object "zebra". Red boxes indicate the location of zebras. Note although the novel objects are in the training set, the captioning sentences does not contain the object word.

among the 70,194 training images of COCO. In addition, the novel object "zebra" in the 15 training images is inconspicuous as shown in Fig. 4. The red box in the figure points out the object zebra. Since the objects are too small, all the five human annotators ignored them and did not mention "zebra" when giving their language description. These examples indicates that our model does not rely on the visual existence of these novel objects in training.

Scaling to the ImageNet Dataset. Following [12], [13], we use the same subset from ImageNet, which contains 646 objects that are not present in the MSCOCO dataset. This results in 164,909 images from the ImageNet dataset for testing. Same as previous methods, we take the paired image-sentence samples in MSCOCO training set as the training data. We apply the trained model to generate captioning sentences for images in the test subset of ImageNet. Since there is no paired image-sentence data on the ImageNet dataset, we empirically evaluate the ability of our method to describe the novel objects.

The Nocaps Dataset. The nocaps dataset contains images sourced from the Open Images validation and test sets. It is a large-scale novel object captioning dataset containing 4,500 validation images and 10,600 test images. Each image is annotated by 11 annotators. Totally, the nocaps dataset span 600 object classes. Following [35], the model is trained using COCO training data, and directly tested on nocaps without finetuning. The out-of-domain of nocaps covers many visually and linguistically similar concepts to COCO but rarely described in COCO (e.g., seahorse, sewing machine).

4.2 Experimental Settings

The Object Detection Model. We employ the publicly available pre-trained object detection models to build the key-value object memory. For experiments on the MSCOCO dataset, we use Faster R-CNN [45] model with Inception-ResNet-V2 [46] to generate detection bounding boxes and scores. The object detection model is pre-trained on all the MSCOCO training images of 80 objects, including the eight novel objects. We use the pre-trained models released by [47] which are publicly available. As for experiments on the ImageNet dataset, following [13], we use the same object classifiers (a 16-layer VGG model) trained on the Imagenet ILSVRC12 dataset.

Evaluation Metrics. To assess the quality of the generated captioning sentences, in our experiments, we use an effective machine translation metric, Metric for Evaluation of Translation with Explicit Ordering (METEOR) [48]. We also use the F1-score as an evaluation metric following [11], [12], [13]. F1-score considers false positives, false negatives, and true positives, indicating whether a generated sentence includes a new object. For results on the ImageNet dataset, since there are no annotated ground truth sentences, we follow [12], [13] and use another two metrics for the novel object captioning task, i.e., describing novel objects (Novel) and Accuracy scores. As introduced in [12], the Novel score is the percentage of all novel objects mentioned in predicted captioning sentences. The Accuracy score is the percentage of images where the shown novel object is correctly described by addressing the object in our generated captioning sentence.

Implementation Details. For fair comparisons, we use VGG-16 pretrained on the ImageNet dataset [22] as the visual encoder. The CNN encoder is fixed during model training. The decoder is an LSTM with cell size 1,024 and 15 sequence steps. For each input image, we take the output of the fc7 layer from the pre-trained VGG-16 model with 4,096 dimensions as the image representation. The representations are processed by a fully-connected layer and then fed to the decoder (Switchable LSTM) as the initial state. For the word embedding, unlike [11], [13], we do not need the per-trained word embeddings with additional knowledge data. Instead, we learn the word embedding ϕ_w with 1,024 dimensions for all words. We use TensorFlow [49] to implement our framework. We optimize the model using the ADAM [50] optimizer, with the learning rate of 1×10^{-3}. The weight decay is set to 5×10^{-5} to avoid overfitting. We train the model for 50 epochs. The maximum object memory size N_M is set to four.

4.3 Comparison to the State-of-the-Art Results

Table 1 summarizes the F1 scores and METEOR scores of all methods on the held-out MSCOCO dataset. All the baseline methods, except LRCN, use additional semantic data containing the words of the eight novel objects. Nevertheless, without external sentence data, our method achieves competitive performance to state of the art. Our model yields a higher average F1-score than the previous state-of-the-art result (60.08% versus 54.4%). The improvement is significant, considering our model uses fewer training data. Our METEOR score is slightly worse than the CBS [14]. The reason is two-fold. On the one hand, CBS uses the beam search strategy, which is known for improving sentence performance. On the other hand, it uses many training sentences containing the novel words in the training. Consequently, it operates in a more advantageous setting than our zero-shot setting in which there are zero training sentences of the novel objects. Compared to the methods with additional sentence data, our method generates better captions for the novel objects without these data. In addition, compared to our previous work DNOC [3], which is also a zero-shot novel object captioning method, the improved version (Switchable LSTM) significantly and consistently outperforms the previous version (DNOC) on all evaluation metrics. These results demonstrate the effectiveness of our SNOC framework and its capability of utilizing both the external and internal knowledge.

Describing in-Domain Objects. Besides the unseen (out-of-domain) objects, we also validate the capability of describing

TABLE 1
The comparison with the state-of-the-art methods on the eight novel objects in the held-out MSCOCO dataset. All the results are reported using VGG-16 [43] feature and without beam search except CBS [14]. Note that we adopt the zero-shot novel object captioning setting where no additional language data is used in training. All F1-score values are reported as percentage (%)

Settings	Methods	F_{bottle}	F_{bus}	F_{couch}	$F_{microwave}$	F_{pizza}	F_{racket}	$F_{suitcase}$	F_{zebra}	$F_{average}$	METEOR
With External Semantic Data	DCC [11]	4.63	29.79	45.87	28.09	64.59	52.24	13.16	79.88	39.78	21
	NOC [12]										
	–(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[13]										
	–(One hot)	29.07	64.38	26.01	26.04	75.57	66.54	55.54	92.03	54.40	22
	CBS [14]	16.3	67.8	48.2	29.7	77.2	57.1	49.9	85.7	54.0	23.3
	NBT+G [32]	7.1	73.7	34.4	61.9	59.9	20.2	42.3	88.5	48.5	22.8
	CRN [34]	38.05	78.40	55.93	53.76	81.43	62.02	57.69	85.38	64.08	21.3
	FDM-net [36]	-	-	-	-	-	-	-	-	64.7	25.7
Zero-shot	LRCN [4]	0	0	0	0	0	0	0	0	0	19.33
	DNOC [3]	32.18	75.88	49.25	51.28	76.85	30.68	58.32	82.60	57.13	21.19
	Ours	31.60	76.82	52.87	54.55	78.13	43.74	59.3	83.69	60.08	21.88

seen (in-domain) objects. The in-domain testing focuses on describing the objects that present during training. Since the proxy visual words are computed by the cosine distance with *training* objects (Eqn.(4)), it would find the category itself for an in-domain (training) object. Thus in this experiment, the proxy visual words are indeed the objects themselves. Table 2 shows the comparison of our SNOC with the baseline method LRCN [4] and our previous version DNOC [3] in the held-out MSCOCO dataset. Our Switchable LSTM achieves higher F1-scores on the known objects than these methods. Our method significantly outperforms the baseline LRCN [4] by 14.8 points on the averaged F1 scores. The comparison results strongly support that our method can better describe objects in images, even on the seen (in-domain) objects.

Scaling to the ImageNet Dataset. Table 3 shows the results on the ImageNet dataset. Note that LSTM-C uses huge external unpaired text data (i.e., British National Corpus and Wikipedia). It is surprising to see that without any external data as used in the compared methods, our method achieves higher performances on the ImageNet dataset. The comparison demonstrates that our SNOC can correctly generate captioning for novel objects even when scaling into ImageNet images with hundreds of novel objects.

Results on the Nocaps Dataset. To enable our model in captioning the novel objects in nocaps, we replaced our COCO pre-trained detection model with the detection model pre-trained on the Open Image dataset. Others are kept the same with the COCO experiments, i.e., using image features from the same pre-trained VGG-16 model and the vanilla LSTM and word embeddings with random initialization. The results of the out-of-domain testing are shown in Table 4. Compared to the strong baseline UpDown [10], our method significantly outperforms it on both validation set and test set under out-of-domain evaluation. The performance improvement is considerable, since UpDown [10] exploits much better image features (bottom-up features using a Faster-RCNN detector pre-trained on Visual Genome) and GloVe word embeddings. Note that NBT and CBS utilized additional GloVe word embeddings and ELMo model, which are pretrained using external large-scale corpus dataset. Thus their language models have already seen sentences with the unseen objects. For a fair comparison, we re-train NBT using the same pretrained models and input features as ours (indicated by * in Table 4). Our model outperforms NBT by 7.2 points on the nocaps val set in the zero-shot setting. Oscar [51] is first pre-trained on

TABLE 2
Comparison of some known objects in the held-out MSCOCO dataset

Methods	F_{cat}	F_{dog}	$F_{elephant}$	F_{horse}	$F_{motorcycle}$	$F_{average}$
LRCN [4]	75.73	53.62	65.49	55.20	71.45	64.50
DNOC [3]	86.92	70.66	75.49	72.41	78.13	76.72
Ours	90.01	74.11	76.89	77.38	77.90	79.26

TABLE 3
Results on the ILSVRC ImageNet dataset

Model	Novel	Accuracy
DCC [11]	56.85	11.08
NOC (One hot+Glove) [12]	69.08	10.04
LSTM-C (One hot+Glove) [13]	72.08	16.39
Ours	**95.82**	**36.21**

TABLE 4
CIDEr Scores on the *out-of-domain* validation set and test set of the nocaps dataset. * indicates our re-implementation results in the zero-shot novel object captioning setting

Settings	Model	Val	Test
With External Semantic Data	NBT [32]	54.0	48.7
	NBT [32] + CBS [14]	63.7	58.5
	Oscar [51]	45.1	-
	Oscar [51]+SCST+CBS	80.3	-
Zero-shot	UpDown [10]	31.3	30.1
	NBT [32] *	35.2	-
	Ours	42.6	39.4

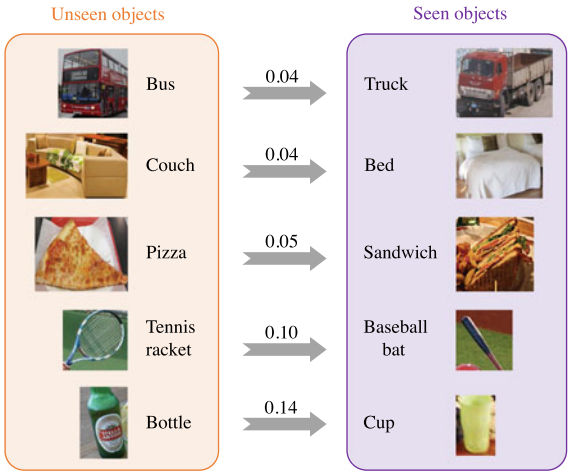


Fig. 5. Examples of the proxy visual words. Left are detection results of the novel objects during testing. Right are the retrieved corresponding proxy visual words from seen objects in training. The numbers above arrows indicate the cosine distance calculated via Eqn. (4).

external 6.5 million text-image pairs and then fine-tuned on nocaps. Therefore, they are not fair comparisons to our method. In contrast, we follow our zero-shot setting and use the image features extracted by VGG-16 and randomly initialized word embeddings. We also show some qualitative results on the nocaps dataset in Fig. 6. Compared to UpDown, our method generates more detailed and precise sentences about the novel objects.

4.4 Ablation Studies

We design ablation studies to evaluate the effectiveness of each component in our framework.

The Effectiveness of the Proxy Visual Word. We propose the proxy visual word to present an unseen object by some known objects that have similar visual appearance. Fig. 5 shows some examples of the generated proxy words for the novel objects in the test set. We can see from this figure that these proxy words are very close to the ground truth of these novel objects, indicating that our proxy visual words are capable of partially describing the unseen object given the limited knowledge. For example, although there is no "pizza" seen during the training, we can generate the proxy visual word "sandwich" to represent it when completing a naive sentence. The generated sentences are reasonable and meaningful since the meaning and phrases templates are similar for these two objects.

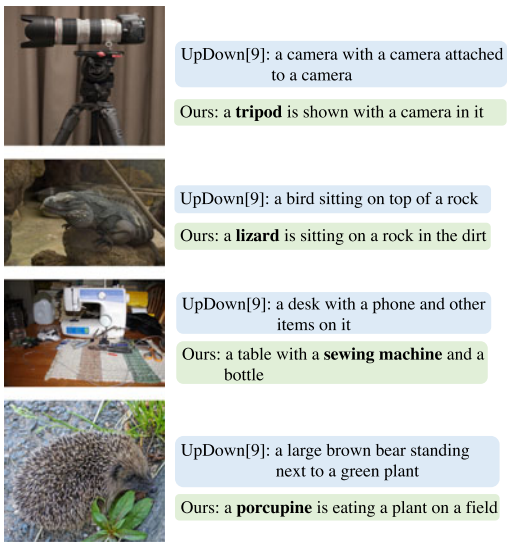


Fig. 6. Quilitative results on the nocaps [35] validation set.

TABLE 5
Ablation studies on the held-out MSCOCO dataset. "Ours w/o `Retrieving`" indicates our SNOC framework but without the activation of the `Retrieving` mode. "Ours w/o addressing" indicates that we remove the addressing operation (Eqn. (8))

Model	$F1_{average}$	METEOR
Ours w/o `Retrieving`	0	19.51
Ours w/o addressing	40.17	19.97
Ours	60.08	21.88

The Effectiveness of the `Retrieving` *Mode.* In the ablation experiments, we validate the effectiveness of this module by removing the whole `Retrieving` module. In this setting, there is only the `Generating` mode in the LSTM. As a result, the LSTM cannot leverage the external knowledge from the detection model and thus performs badly on these novel words. It can be seen from Table 5 that the model even cannot mention the unseen object ($F1_{average} = 0$). The results are far away from that of our full model.

The Effectiveness of the Content-Based Addressing Operation. Our Switchable LSTM conducts content-based addressing (Eqn (8)) on the object memory to select a correct noun word to describe the unseen object. The addressing operation connects the semantic language representation and the visual CNN feature like a grounding operation. In Table 5, we also validate the effectiveness of content-based addressing operation. "Ours w/o addressing" indicates that we replace the content-based addressing operation by randomly selecting a detected object as the retrieved noun word $\boldsymbol{p}_t^r$. With the addressing operation, our full model outperforms "Ours w/o addressing" by 19.91% in F1-score and 1.91% in METEOR score. The comparison indicates that the addressing operation in the `Retrieving` mode can enhance the semantic understanding of the visual content and can easily find the object of interest.

Analysis on Different Language Models. We experimented different language model such as BERT and GRU. Since our switching model design is based on the recurrent neural network, we only take the pretrained BERT model as the initialization of the input embeddings. Table 6 shows BERT improves our method from 21.88 to 22.41 in METEOR. This indicates better language models would bring further performance improvement of our method. We also replaced our base LSTM model by the Gated Recurrent Unit (GRU). We found in experiments GRU leads to similar F1-scores in captioning novel objects, but worse METEOR scores. The reason might be the less parameters of GRU compared to that of LSTM.

Analysis on Different Object Detection Models. We tried different detection models in building our framework. We

TABLE 6
Analysis of different language models in terms of Averaged F1-score and METEOR score on the held-out MSCOCO dataset

Model	$F1_{average}$	METEOR
Ours (LSTM)	60.08	21.87
Ours + BERT	60.81	22.41
Ours (GRU)	60.59	21.52

TABLE 7
Impact of different detection models in terms of Averaged F1-score and METEOR score on the held-out MSCOCO dataset

Model	$F1_{average}$	METEOR
Ours + Faster-RCNN (Inception-ResNet)	60.08	21.87
Ours + SSD (ResNet-50 FPN)	58.98	21.32
Ours + Mask-RCNN (Inception-ResNet)	60.22	21.71

compared three kinds of detection models in experiments, i.e., FasterRCNN with Inception ResNet v2, Mask-RCNN with Inception ResNet v2, SSD with ResNet50 FPN. The results are shown in Table 7. We found the detection model (FasterRCNN with Inception ResNet v2) achieves the highest performance among all the competitors.

Analysis of the Maximum Object Memory Size $\boldsymbol{N_M}$. N_M indicates the maximum slots in our object memory, i.e., the number of object-label pairs for each image. If the memory size is too small, there is little external knowledge considered in our SNOC. If we set the memory size to be very large, it might introduce too many noisy candidates, and thus limits the performance. We show the averaged F1-scores and METEOR scores over different memory sizes in Fig. 7. $N_M = 0$ means there is no detection output used in the framework. We can see from the figure that the F1-scores is relatively low when N_M is less than three. The reason is that the introduced external knowledge is not sufficient since only very few detected objects are in the memory as candidates. The performance curve becomes flatten as N_M increases more than three. We also observe a slight performance drop (F1-scores drops from 60.1 to 59.8) when the memory size is too large. The reason might be that too many noisy candidate objects are written to the memory, making the content-based addressing less reliable. The results of different object memory sizes also validate the effectiveness of our `Retrieving` mode. It can be seen that by introducing sufficient external detection knowledge, our SNOC is capable of describing the novel objects even without any related training sentences.

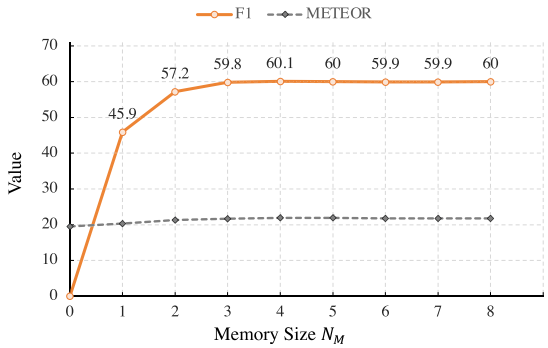


Fig. 7. The performance curves over different memory size N_M.

TABLE 8
Impact of Reinforcement Learning on the held-out MSCOCO dataset

Model	$F1_{average}$	METEOR
Ours	60.08	21.88
Ours + SCST[31] (CIDer)	59.79	22.39
Ours + SCST[31] (METEOR)	59.57	22.76

Analysis of Reinforcement Learning (RL) Training. Recently, RL training has been a standard way to improve the performance for the image captioning task. We follow SCST [31] and apply it to our model. We first train our model using cross-entropy loss and then fine-tune the model using SCST with a small learning rate (1×10^{-5}). We take the greedy decoding as the baseline in RL. We tried two different reward target metrics, i.e., CIDer and METEOR. The results are shown in Table 8. We can see that the RL-based training only improves the captioning metrics but not the F1 scores. The reason might be that the optimization target is for better evaluation scores (e.g., CIDer and METEOR), but not for the novel object captioning (the switching mechanism and memory retrieval).

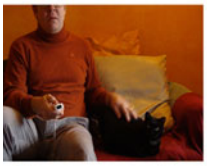
4.5 Qualitative Results

We qualitatively show some examples from the test set of the held-out MSCOCO dataset in Fig. 8. For each image, we first build the object memory based on the detection results. Since the words of novel objects do not exist in training, we use our generated proxy visual words for each novel object in building the memory. Then our Switchable LSTM predicts words by words to form a naive sentence. The words in purple are generated in the `Retrieving` mode, while other words are from the `Generating` mode. It can be seen that our Switchable LSTM successfully switches between the two modes in generating sentences. The words from the `Retrieving` mode are the nouns of detected objects. We first finish a naive sentence with proxy visual words, and then replace them with the accurate unseen words.

Take the image at the lower-left corner as an example. The object microwave does not exist in training data. We first find the proxy visual word "oven" for "microwave" based on the visual similarity. Then We build the object memory with two slots, containing the feature-label pairs for "dog" and "oven", respectively. Our Switchable LSTM generates the words in two modes, with an inside switcher controlling the two modes. The word "dog" and "oven" in the sentence are the retrieved results using Eqn. (8). Finally, we replace the proxy visual word "oven" by its true name "microwave", which circumvents the out-of-vocabulary issue. Without seeing any image and sentence that contain

Detected: bus
Proxy visual word: truck
Ours: a yellow truck is parked in a parking lot
Ours: a yellow bus is parked in a parking lot
GT: a yellow bus is shown parked in a parking lot

Detected: dog, bottle
Proxy visual word: cup
Ours: a dog is standing in the grass with a cup
Ours: a dog is standing in the grass with a bottle
GT: a black dog pants next to a blue water bottle



Detected: person, cat, couch
Proxy visual word: bed
Ours: a person laying on a bed with a cat
Ours: a person laying on a couch with a cat
GT: a man sitting on a red couch with a black cat

Detected: dog, microwave
Proxy visual word: oven
Ours: a dog is sitting in a oven in a kitchen
Ours: a dog is sitting in a microwave in a kitchen
GT: a dog that is siting in a microwave

Detected: cup, pizza
Proxy visual word: sandwich
Ours: a sandwich and a cup of coffee on a table
Ours: a pizza and a cup of coffee on a table
GT: a small pizza next to a glass of water is shown

Detected: zebra
Proxy visual word: sheep
Ours: a herd of sheep grazing on a lush green field
Ours: a herd of zebra grazing on a lush green field
GT: a big field that has a lot of zebras on it

Detected: person, bench, backpack, suitcase
Proxy visual word: handbag
Ours: a person is sitting on a bench with a handbag
Ours: a person is sitting on a bench with a suitcase
GT: a couple of suitcases sitting at the end of some benches

Detected: person, tennis racket
Proxy visual word: baseball bat
Ours: a man in a red shirt holding a baseball bat
Ours: a man in a red shirt holding a tennis racket
GT: a man holding a tennis racket up and smiling

Fig. 8. Qualitative results of our SNOC model on the test set of the held-out MSCOCO dataset. The words in orange are not present during training, i.e., the novel objects. The words in purple are generated in the `Retrieving` mode, while other words in sentences are from the `Generating` mode. "Detected" shows the object detection results from the external detection model. "Proxy visual word" is the proxy visual word for the unseen word used in building the object memory (Eqn. (5)). "Ours" contains two sentences. The first one is the generated sentence of our Switchable LSTM with proxy visual words. The second one is the revised sentence by replacing the proxy visual words with the detected labels. "GT" means the human-annotated ground truth sentences.

microwave during training, our model improves the generalization ability and successfully describes the image.

5 Conclusion

In this paper, we tackle the novel object captioning under a challenging condition where zero sentences of the novel object are available during training. We mimic the way that babies talk about something unknown. When describing an unseen object, a baby usually uses the word of its most similar object. For those novel objects that are never shown in the training stage, we take words of most similar seen objects as proxy visual words. Then we utilize an external detection model to build a key-value object memory, containing the visual information and the corresponding word for each object. To introduce external knowledge into the sentence generation, we propose a Switchable LSTM that has two switchable working modes, i.e., 1) generating the sentences like a standard LSTM and 2) retrieving a proper noun from the key-value memory. We design a new indicator in the LSTM cell to switch the two modes. Our Switchable LSTM is thus capable of leveraging both internal knowledge and external knowledge. Our experiments validate its effectiveness on both the held-out MSCOCO dataset and the ImageNet dataset. Without any additional sentence data, our method even outperforms the state-of-the-art methods that use additional language data.

References

[1] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.

[2] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. Int. Conf. Neural Informat. Process. Syst.*, 2015, pp. 1171–1179.

[3] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, "Decoupled novel object captioner," in *Proc. ACM Multimedia Conf.*, 2018, pp. 1029–1037.

[4] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.

[5] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[6] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," *Proc. Int. Conf. Learn. Representations*, 2015.

[7] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *Proc. Int. Conf. Learn. Representations*, 2016.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[9] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[10] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[11] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1–10.

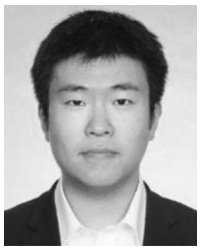
[12] S. Venugopalan, L. Anne Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1170–1178.

[13] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating copying mechanism in image captioning for learning novel objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5263–5271.

[14] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Guided open vocabulary image captioning with constrained beam search," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 936–945.

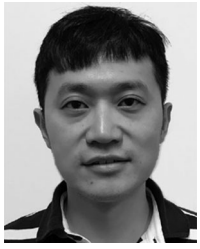
[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.
[17] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6292–6300.
[18] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 595–603.
[19] Y. Wu and Y. Yang, "Exploring heterogeneous clues for weakly-supervised audio-visual video parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1326–1335.
[20] Y. Yang, Y. Zhuang, and Y. Pan, "Multiple knowledge representation for Big Data artificial intelligence: Framework, applications, and case studies," *Front. Informat. Technol. Electron. Eng.*, vol. 22, pp. 1551–1558, 2021.
[21] G. Kulkarni *et al.*, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.
[22] M. Mitchell *et al.*, "Midge: Generating image descriptions from computer vision detections," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2012, pp. 747–756.
[23] A. Farhadi *et al.*, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.
[24] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Int. Conf. Neural Informat. Process. Syst.*, 2011, pp. 1143–1151.
[25] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *Int. J. Comput. Vis.*, vol. 124, no. 3, pp. 409–421, Sep. 2017.
[26] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Understanding*, vol. 163, pp. 21–40, 2017.
[27] Y. Wu, L. Jiang, and Y. Yang, "Revisiting embodiedqa: A simple baseline and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 3984–3992, Jan. 2020.
[28] X. Wang, L. Zhu, and Y. Yang, "T2VLAD: Global-local sequence alignment for text-video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5079–5088.
[29] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4651–4659.
[30] H. R. Tavakoliy, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2506–2515.
[31] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7008–7024.
[32] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7219–7228.
[33] X. Wang, J. Wu, D. Zhang, Y. Su, and W. Y. Wang, "Learning to compose topic-aware mixture of experts for zero-shot video captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8965–8972.
[34] Q. Feng, Y. Wu, H. Fan, C. Yan, M. Xu, and Y. Yang, "Cascaded revision network for novel object captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3413–3421, Oct. 2020.
[35] H. Agrawal *et al.*, "nocaps: Novel object captioning at scale," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8948–8957.
[36] T. Cao, K. Han, X. Wang, L. Ma, Y. Fu, Y.-G. Jiang, and X. Xue, "Feature deformation meta-networks in image captioning of novel objects," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10 494–10 501.
[37] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.
[38] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann, "Bridging the ultimate semantic gap: A semantic search engine for internet videos," in *Proc. Annu. ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 27–34.
[39] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1641–1648.
[40] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.
[41] L. Jiang, S.-I. Yu, D. Meng, Y. Yang, T. Mitamura, and A. G. Hauptmann, "Fast and accurate content-based semantic search in 100m internet videos," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 49–58.
[42] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4622–4630.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
[44] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4534–4542.
[45] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Informat. Process. Syst.*, 2015, pp. 91–99.
[46] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[47] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3296–3297.
[48] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proc. 2nd Workshop Statist. Mach. Transl.*, 2005, pp. 65–72.
[49] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. Operating Syst. Des. Implementation*, 2016, pp. 265–283.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[51] X. Li *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 121–137.



Yu Wu received the PhD degree from the University of Technology Sydney, Australia, in 2021. He was a research intern with Baidu Research. He was the recipient of Google PhD fellowship 2020. His research interests include multi-modal perception and video understanding.

Lu Jiang received the PhD degree from Carnegie Mellon University in 2017. He is currently a staff research scientist in Google Research and an adjunct faculty at Carnegie Mellon University, Language Technologies Institute. His research interests include the interdisciplinary field of multimedia, machine learning, and computer vision, specifically including robust deep learning, content creation, and video understanding.



Yi Yang received the PhD degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a professor with Zhejiang University, Hangzhou, China. His current research interests include machine learning and its applications to computer vision, such as multimedia retrieval and video content understanding.