RESEARCH ARTICLE OPEN ACCESS

Image Captioning Generated by Zero Shot Learning

N. Kalyani [1], G. Pradeep Reddy [2], K. Sandhya [3]

Computer Science and Engineering, Jawaharlal Nehru Technological University – Ananthapuramu

ABSTRACT

Image captioning includes machines that describe images through sentences. Training pairs demand picture-sentence annotations. Pre-trained models fail in new domains with novel objects, requiring human annotations. Accurate capture of novel objects requires human input. For mining multimedia annotations, descriptive captions are a great source for training automatic image annotation systems. Past research indicates about 20% image-caption relevance, leading to consumer frustration with irrelevant captions. Efficiency needs improvement, especially to avoid repeated captchas due to wrong titles. Introduce zero-shot novel object captioning to improve performance, using adaptive LSTM to incorporate key-value object memory and object knowledge into sentences. The novel algorithm further improves image captioning by focusing on describing unique objects that are not present in the training data.

Keywords: Convolutional Neural Network, Recurrent Neural Networks, Long Short-Term Memory, Novel Object Captioner.

I. INTRODUCTION

As a classical task in vision and language research, image captioning aims to automatically describe an image using natural language sentences or phrases. Encoder decoder architectures have proven to be a general framework for image captioning task [5], [4], [7], [2], [3], [8], [6], [11], [9], [10],in which 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. Because CAPT models are trained on parallel data of image-sentence pairs, they fail to recognize title words if these words are not present in the training sentences. In recent years, generalizing captioning models to describe novel objects that only occur during testing has been a key research direction in picture captioning research. For example, as illustrated in Fig. 1, although the captioning model (LRCN [2]) can correctly generate captions for the object "giraffe", it fails for the similar object "zebra" due to the lack of training sentences. zebra is any word.

Some works have been proposed to solve this problem [14], [15], [12]. In general, these methods attempt to improve model generalization by incorporating extralinguistic knowledge about the new object. This is achieved by using pre-trained language models [14], [15] or additional unpaired training sentences of novel objects. For example, Hendrick’s et al.[14] trained a caption model by using a pre-trained image tagger and a pre-trained language sequence model from an external text corpora. Existing works mitigate this problem by removing the dependency on parallel training data of paired images and sentences, which becomes more difficult to collect. A strict definition of a novel object in existing works is that the object does not appear in parallel training sentences, but must still exist in the training data in the form of dissimilar sentences. In other words, they all assume that during training there will always be training sentences of novel objects. Still, this supposition doesn't hold in numerous real- world scripts. Descriptions for the newest products, such as self-balancing scooters, robot vacuums, and drones, are usually rare, time- honoured. Moreover, and perhaps more importantly, language generation is learned in conjunction with the objects seen and therefore inevitably introduces linguistic biases to the caption model for illustration, if the training rulings are each about bass (an ocean fish), the captioning model won't learn to term the instrument bass and may produce awkward rulings like “a man-eating bass with a guitar amplifier”.

This article resolves captions for new objects that do not require a training set of new objects. We refer to this as "zero-shot novel object captioning" to distinguish it from the traditional problem posed in [14], [15], [12], [13]. In a traditional environment, an additional training set of new objects is injected alongside the pre-trained object recognition model. There is no training set for new objects in the zero-shot captions setting. H. There is no information about the semantic meaning, meaning, or context of the object. The only external knowledge in the proposed environment is a pre-trained object recognition model capable of recognizing new objects. This was also required in the traditional problem setting.

The novel algorithm in image captioning introduces a unique approach to generating textual descriptions for images. It leverages advanced neural network architectures, often incorporating attention mechanisms, to better understand image content. This algorithm surpasses traditional methods by learning contextual relationships between objects, actions, and scenes within images, leading to more accurate and contextually rich captions. Additionally, it might incorporate techniques from reinforcement learning or transformer models to improve caption quality and diversity. Overall, this novel algorithm significantly enhances the ability to generate relevant and coherent captions for a wide range of images.

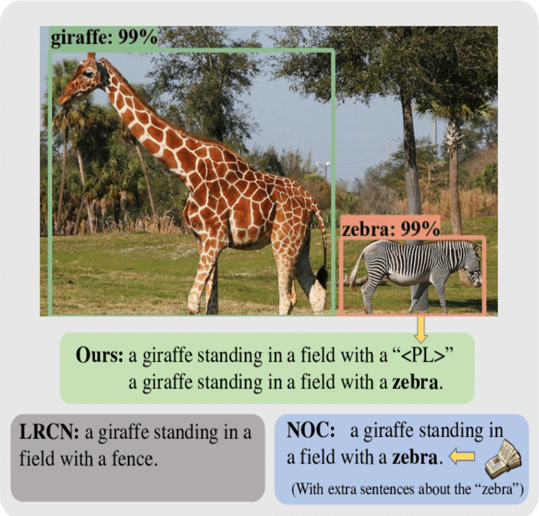

giraffe: 99%
zebra: 99%
Ours: a giraffe standing in a field with a "<PL>"
a giraffe standing in a field with a zebra.
LRCN: a giraffe standing in a field with a fence.
NOC: a giraffe standing in a field with a zebra.
(With extra sentences about the "zebra")

Fig 1. An example of the novel object captioning. The colored bounding boxes show the object discovery results. The new object" zebra" isn't present in the training data. LRCN (8) fails to describe the image with the new object" zebra". NOC (34) can induce the correct caption but requires redundant textbook training data containing the word" zebra" to learn this conception. Our algorithm can induce correct captions, and more importantly, we don't need any redundant judgment data. Specifically, we first induce the caption template with a placeholder"" that represents the new object. We also fill in the placeholder with the word" zebra" from the object discovery model.

Because novel objects are completely invisible during training, zero-shot captioning presents a new challenge for dissociating language production from visual detection. To address this challenge, we propose a solution that mimics the way children speak. When describing an unseen object, the infant tries its best to use objects it has seen before. For example, a child might say, "A horse is standing in a field" to describe a zebra. If the ambiguous word "horse" is replaced by the correct word "zebra" the sentence will be accurate.

Using the above as inspiration, we propose a framework called Switchable Novel Object Captioner (SNOC).This framework aims to generate natural language sentences extracted from training object classes to describe new objects during testing. Unlike existing work, our model learns to completely decouple speech generation from training objects, so new objects do not require a training set. SNOC follows the standard encoder/decoder architecture but has a new decoder. The decoding stage first creates a key-value object store via a recognition model containing visual information and corresponding words for each object displayed in the image. For these new words, use words that describe very similar objects instead. We call it "surrogate visual words". Next, we propose a switchable LSTM that incorporates object memory for sentence generation. A switchable LSTM alternates between two modes of operation. H. 1) generate sentences as a standard LSTM [17] and 2) retrieve the correct nouns from a key-value store driven by a newly developed index on the LSTM cells. Finally, through a switchable LSTM, we first generate fake sentences using only visible words, and then replace visual substitute words with real object labels.

For example, in Fig. 1, the "zebra" object does not appear during the training phase. Our method describes the invisible object "zebra" in a coarse-to-fine manner. It first remembers its similar object from the training data, i.e., "giraffe" in this case. Next, it generates an incorrect sentence using its known knowledge of the proxy visual word "giraffe" and finally corrects it using the correct word "zebra" provided by the external recognition model.

The proposed model is based on our previous work DNOC [3], in which we directly use the special token "<PL>" to represent all unseen words. However, this strategy ignores the visual appearance of novel objects, as using one token term (i.e., <PL>) to represent all novel objects is ambiguous. We have made two important extensions to DNOC. First, we replace the placeholder commemorative by a deputy visual word, which helps construct a better judgment by adding the visual parallels between the new objects and the seen objects. We extensively extend the experiments and analysis of the proposed method. We additionally evaluate our methods on two large-scale 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 reinforcement learning (RL) to directly train our model for optimization on language metrics (CIDer and METEOR).

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

To encapsulate, the primary innovations of this investigation encompass:

• 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 make efforts from the following three aspects, we first design a switchable LSTM to determine where to place the object words (via a switch indicator).

• We prize the semantic information from the LSTM retired countries to find which visual object to relate to then from all honoured object memory.

• To ensure consistency in sentences and reduce out-of-vocabulary problem, we design proxy visual words and avoid unfamiliarity effect by imported novel object labels in LSTM.

II. RELATED WORK

2.1 Image Captioning

Automatic caption generation is the task of describing the content of an image through a complete and natural sentence. This is a fundamental problem in the multi-modal perception field [2], [9], [18], [16], [20], [19], [22]. Some early works such as template-based approaches [23], [21] and search-based approaches [23], [24] generate headings through a sentence template and a pool of sentences. Recently, inspired by deep learning and sequence modelling in computer vision, language-based models have achieved good performance. Most of them are based on encoder-decoder architecture to learn the probability distribution of both visual embedding and textual embeddings [4], [2], [3], [8], [6], [11], [27], [26], [29], [28], [31], [30]. In this architecture, the encoder is a CNN model that processes and encodes the input image as an embedding representation, while the decoder is an RNN model that takes a CNN representation as initial input and sequentially predicts the next word given the previous word. In recent works, Kiros et al. [20] proposed a multi-model log-bilinear neural language model to jointly learn word representations and image feature embeddings. Vinales et al. [11] proposed an end-to-end neural network consisting of a vision CNN, which then generates an RNN. Zhu et al. [9] improved [11] by incorporating an attention mechanism into captioning. The attention mechanism focuses on important image regions while generating related words. In general, these methods are designed to describe visible objects with many training examples. A decoder's vocabulary remains stable after training and cannot be further expanded by external knowledge More Lately, in the SCST (31), all objects and words were present in training, but their combination was unusual in test (e.g., a blue boat in front of a structure). However, in training we focus on a more challenging task that does not contain these objects and words.

2.2 Novel Object Captioning

Novel object captioning is a challenging task where training lacks paired visual-sentence data for the novel object. Only a few works have been proposed to solve this title problem Heinz Dricks et al. [14] proposed a deep composition captioner (DCC) as a pilot work to address the task of generating descriptions of new objects not present in paired image set datasets. Venugopalan et al. (15) bandied a new object captioner (OC) to further ameliorate DCC into an end- to- end system by concertedly training a visual bracket model, a language sequence model, and a captioning model. Anderson at all. [13] used an approximate search algorithm to forcefully guarantee the inclusion of selected words in the evaluation phase of a caption generation model. Yao et al. [12] used a mechanism to copy the recognition results to the output sentence with a pre-trained linguistic model. Lu and others. [34] also proposed to generate a sentence template with slot positions, which are then populated by visual concepts from object detectors. Wang et al. [32] proposed a new zero-shot video captioning with the aim of describing videos outside the domain by composing different experts based on different topic embeddings and implicitly transferring knowledge learned from seen activities to unseen ones. Feng et al. [36] proposed a cascade revision module to generate better sentences by considering both visual similarity and semantic similarity on ambiguous words. Aggarwal and others. [35] collected a large-scale novel object captioning dataset and extended existing novel object captioning models to establish robust baselines. Cao et al. [38] proposed to adapt the heading model to novel object features discovered by assisted recognition.

Note that all of the above methods require the use of additional data of novel objects to train their word embeddings. In contrast to existing methods, our method focuses on a 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 that are not visible during training [32], [39], [37], [41], [40], [43]. Zero-shot literacy islands the gap between visual and textual semantics by learning a wordbook of conception sensors on external data sources [42]. Recently, some works have focused on the zero-shot novel object captioning task, where no additional training sentences are available in learning to caption novel objects. Wu et al. [7] proposed a decoupled captioning framework DNOC to generate a sentence template, which allows model object labels to be freely introduced into the generated sentence template. DNOC simply uses a unique token to represent all novel objects, leading to ambiguous title results. Differently, we propose to use 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 enhance the LSTM cell with flexible working modes, which enable the use of existing and external knowledge.

III. THE PROPOSED METHOD

3.1 Preliminaries

Let t represent the time step in the caption generation process, and let wt_t denote the word predicted at time step

t. The objective is to predict the next word wt, considering the context of the previously generated words (w0, w1, ..., wt-1) and the encoded image features fe(I). The prediction probability distribution over the vocabulary for the next word can be expressed as follows:

p(wt | w0, w1, ..., wt-1, fe(I)) = Softmax(W * [w0, w1, ..., wt-1, fe(I)]). (1)

In this equation:

• w0, w1, ..., wt-1 represent the embeddings or representations of the words generated up to time step t-1.

• fe(I) denotes the encoded features of the input image.

• [w0, w1, ..., wt-1, fe(I)] is the concatenation of the word embeddings and image features, serving as input to the prediction layer.

• W is a learnable weight matrix that transforms the concatenated input to the vocabulary size.

The Softmax function then converts the raw logits into a probability distribution over the vocabulary, enabling the model to predict the next word based on the accumulated context from the previous words and the image features.

This equation outlines the process of generating the next word in the image captioning task, taking into account the context provided by both the previously generated words and the visual features of the input image. It aligns with the architecture and components you've described in your original text.

The Long Short- Term Memory (LSTM) [17] is a classical decoder in visual captioning and natural language processing tasks. Let t denote the time step in the decoding process, and let xt be the input representation at time step t, which includes both the previous predicted word ot-1 and the hidden state ht-1. The goal is to predict the next word wt at time step t. The LSTM unit updates and predicts the output as follows:

(ct, ht) = LSTM_Cell(xt, ht-1, ct-1),

ot = Softmax(Wo * ht),

Where:

- LSTM Cell (xt, ht-1, ct-1) represents the LSTM cell operation that takes the input xt, the previous hidden state ht-1, and the previous cell state ct-1, and produces the current cell state ct and hidden state ht.
- ct is the current cell state.
- ht is the current hidden state.
- Wo is a learnable weight matrix for the output prediction.
- Softmax function converts the raw logits into a probability distribution over the vocabulary for the next word.

Zero-shot novel object captioning. We study a zero-shot novel object captioning task where a model has to caption novel objects without additional training sentence data about the object. Novel object words were not shown in the paired picture-sentence training data P or in the unpaired sentence training data. A notable challenge to this task is dealing with out-of-vocabulary (OOV) words. The learned word embedding function Øw is unable to encode unseen words because these words cannot be found in the training vocabulary. As a result, these unnoticeable words aren't fed into the decoder for caption generation. former workshop [14], [15], [12] overcome this problem by learning embeddings of unseen words using additional sentences containing the words. However, in our zero-shot novel object captioning task, we do not anticipate the availability of additional training sentences of the novel object.

3.2 Building the Key-Value Object Memory

To describe an image with new objects, we use a pretrained object discovery model as an external knowledge source that provides object name information for objects in the input image. Specifically, for the i-th detected object $obj_{i,}$ we extract its CNN feature f_i Ꞓ$R^{1\times N_f}$ from the ROI pooling layer of the detection model. Then the CNN features fi and the predicted semantic class labels l_i Ꞓ$R^{1\times N_d}$ are used to generate a key-value pair, with the CNN feature as the key and the label as the value. N_D is the number of detection candidates.

We use these identified key-value pairs to construct a key-value object memory that associates semantic class labels (descriptions of novel objects) with their visual representation. The maximum memory size is set to N_M. For the images with more than N_M of detected objects, according to the object detection confidence, we select the top N_M detected objects in the image into memory. Memory μ is initialized again for each input image. It contains all objects seen and detected in the input image, including zero-shot objects. During the generation of a title sentence, the memory μ is held constant during the production process of repeated words.

There are two types of objects in key-value object memory during evaluation, i.e., objects seen during training and novel objects not previously seen in memory. For observed objects, we write the feature-name pairs in memory as,

$$\mu \leftarrow \mathrm{WRITE}(\mu,(f_i,l_i)), \qquad (3)$$

A WRITE operation is to insert a key-value pair into a new slot of the existing memory μ.

Proxy Visual Words. For novel objects, we propose to use a proxy visual word instead to reduce the out-of-vocabulary problem. The main idea is to represent an invisible object by some known object having a similar visual appearance. Specifically, for each object shown in the training data, we extract a visual representation of the image patch. These properties are then clustered according to their object labels. By averaging the visual features of objects belonging to the same class, we obtain a prototypical visual representation v_o for each object class viewed, where represents the o-th object class. When a novel word is encountered, we take the visual feature fi of the new image patch to find the most similar one in the prototypical representation set {Vo}. So, we've the similarity between the new object(obji) and the o- th class,

$$s^o_i = \text{cosine}(f_i,v_o), \qquad (4)$$

where cosine (·) denotes the cosine distance function. Similarity soi is defined by the cosine distance between a feature of a novel object and a prototype feature of the seen object class. By searching the database of seen objects, we can find the object category Îi most similar to the novel object. We named it proxy visual word to distinguish it from exact words for visible objects. So, for the novel object obj$_i$ we insert the pair of visual feature f_i and proxy visual word $\hat{I}_i$ into memory,

$$\mu \leftarrow \text{write}(\mu,(fi,\hat{I}_i)), \qquad (5)$$

3.3 Switchable LSTM

In the zero-shot novel object caption task, language modelling influences both being 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 handles switching between two modes, i.e.,1) generating mode, in which the model generates a simple term like the standard LSTM; and 2) retrieving mode, in which the key-value object is to retrieve the noun from memory μ. In generating mode, we use a memory cell from a standard LSTM to induce a judgment grounded on being knowledge. When in reacquiring mode, rather of generating words, we propose to apply a content- grounded address to object memory to find the correct noun word with external knowledge. An index inside an LSTM cell switches two modes.

3.3.1 Standard LSTM Revisit

The prediction p^l_t of the LSTM cell at step t given the hidden state h_t,

$$p^l_t=W_ph_t+b_p \qquad (6)$$

For the image entitling task, the retired state h0 is initialized as a decoded image point Øc(I) and the original input x0 is a unique token. also, the LSTM iteratively labours a word and takes this word as a new input to the coming step. If the model labours the special commemorative the process of generating repeated words is terminated. We name the process of generating terms by Eqn. (6) As a generating mode.

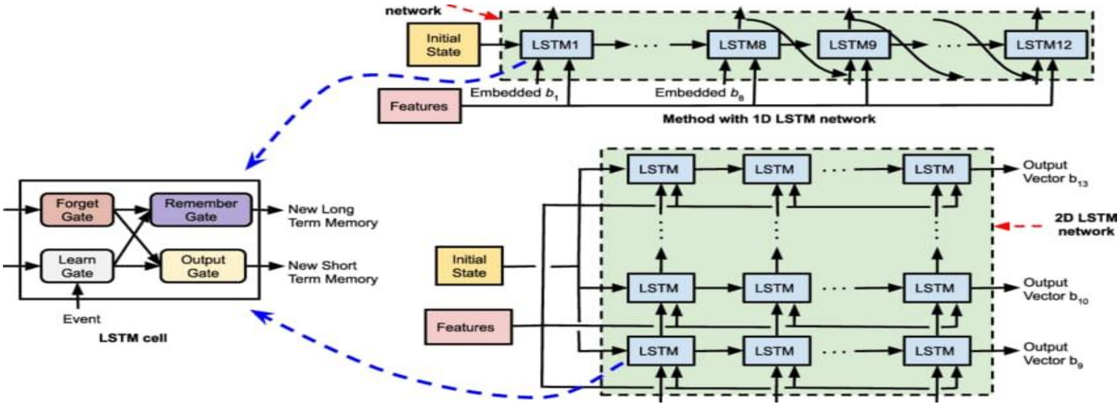


Fig 2. Structure of LSTM

3.3.2 Retrieving Nouns from External Knowledge

Standard LSTM does not consider external knowledge information when generating the caption. To solve this problem, we propose an attention-based operation to incorporate knowledge from object memory μ into sentence production. We name it retrieval mode to distinguish it from the standard LSTM workflow. In retrieving mode, we use the hidden state h_{t-1}as a semantic query to search the object memory μ. The query retrieves the matched noun as the prognosticated word at this time step. The whole retrieving mode can be considered as a grounding operation connecting the semantic language representation and the visual CNN feature. Specifically, at the t-th time step, we define the query qt as the linear transformation of the previous hidden state $h_{t-1,}$

$q_t = W_q h_{t-1};$ (7)

where ht-1 is the previous hidden state at the (t-1)-th step from the sequence model, and Wq is a linear transformation that transforms the hidden state from linguistic semantic space to CNN visual feature space. With this query, we perform content-based addressing operations on object memory µ with the aim of finding relevant object information according to a similarity metric. Formally, the content- grounded addressing operation is defined as,

$p^r{}_t = (q_t K^T) V;$ (8)

where KT and V are perpendicular combinations of all keys and values in memory, independently. The affair pr t ЄRND is a smooth address over all seeker semantic markers. In evaluation, we take the term with maximum probability as the result of the query.

3.3.3 Modes Switching

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

The switch indicator in the t-th stage depends on the hidden state h_{t-1}.

$a_t = W_a h_t + b_a,$ (9)

where W_a and b_a are the trainable weight and bias, respectively. The switch indicator is designed to estimate the probability of selecting the recovery mode at the current time step. We compare the switch indicator with the prediction from the generating mode. Denote the maximum likelihood at $p^l{}_t$ as $p^l{}_t$.If $p^l{}_t$ is greater than the switch index, we choose to estimate the term based on Eqn. (6) (mode of production); Otherwise, we turn on the switch and leverage the object memory to find the correct noun word using Eqn. (8) (retrieving mode). Thus the output of our switchable LSTM at the t-th step is,

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

3.4 Framework Overview

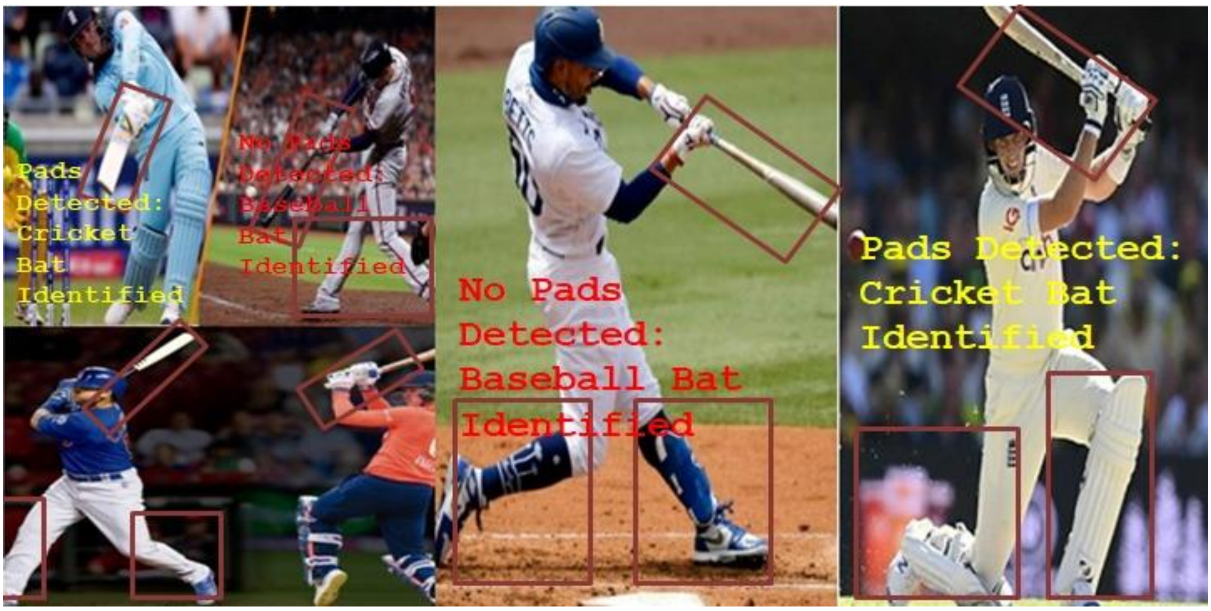
With the design of proxy visual words, the word embedding function Øw can encode all input tokens, regardless of the presence of novel words. So we can generate a naive sentence for the novel objects using the words seen by our switchable LSTM. Finally, we replace the proxy word with the exact name of the novel object, which is provided by the external object detection model. Thus, training our model does not require seeing these novel words and addresses a critical limitation in prior works.

The following steps are used to create a proper title sentence:

1) We use an external object detection model to build a key-value object memory µ for the input image. To overcome the out-of-vocabulary problem, for an invisible object, we use the label of its most similar object as a proxy visual word.

2) We use adaptive LSTM to generate the captioning sentence. The model operates in two working modes jointly to leverage both internal knowledge and external knowledge. A switch indicator is used inside the memory cell to control both modes. In retrieving mode, the guessed word is generated by soft content-oriented addressing on memory µ.

3)Eventually, we replace the sentence's judgment's by corresponding object descriptions.


Pads Detected: Cricket Bat Identified
No Pads Detected: Baseball Bat Identified
No Pads Detected: Baseball Bat Identified
Pads Detected: Cricket Bat Identified

Correct word: a person is holding a cricket bat ….

Fig. 3. Outline of proposed method. In the example, the object " cricket bat" does not appear during training. We first work the object discovery model to make a crucial-value object memory. For the unseen object " cricket bat", we find its most similar candidate from the seen objects by calculating the visual feature distance. A very similar object in this context is a "baseball bat," which is used in memory building. Adaptive LSTM uses both global image feature and object memory as input. When evaluating the second term ("person"), the index inside the cell turns on retrieval mode. Thus the model takes the hidden state as a query to determine object memory. When sentence generation is complete, we replace the proxy visual words with the exact label name provided by the external recognition model.

Taking the input image in Fig. 3 as an example, suppose that the object " cricket bat " is a novel object. We first work the object discovery model and make a crucial-value object memory μ grounded on the discovery results, which contains both visual information and a matching word (discovery class marker). For an unseen object " cricket bat", we find its most similar candidate from seen objects by calculating the distance between visual features, in this case "baseball bat". We use the name "baseball bat" in memory building. Next, our switchable LSTM uses both global image feature and object memory as input. When predicting the second word ("person"), a switch inside the cell turns on the indicator retrieving mode. Therefore, our model takes the retired state h1 as a query to detect the object memory M. Our model finds the perfect noun "person" for the object denoted by Eqn. (8) LSTM iteratively takes the previous output as input to the next step. When sentence generation is complete, we replace the proxy visual word ("baseball bat") with its exact label name ("cricket bat").

3.5 Training

At the core of the zero-shot novel object captioning problem is how to properly integrate object information into sentence production. Toward this goal, we propose to simulate mode switching by placing all object words in retrieval mode during training. In other words, we treat all detected objects as "novel objects" when optimizing the retrieving module.

Specifically, we take into the vocabulary of the retrieval mode all words of recognized objects, including visible objects such as "apple" and "cat". When an object word is encountered in training, we train our model to spark reclamation mode via a switch indicator. Otherwise, we optimize the model to activate the generating mode for regular words other than object words. In this way, the model learns to access external recognition knowledge for help if it wants to specify an object in the image. This training strategy allows our method to activate the retrieval mode even if we do not know the novel objects before.

IV. EXPERIMENTS

4.1 Datasets

Holdouts are MSCOCO records. MSCOCO is a large captioning benchmark containing 123,287 images. Each MSCOCO image has a five-sentence description annotated by humans. According to [14], [15], [12], [13], we use a subset of the MSCOCO dataset to evaluate the model's ability to describe new objects, i.e., the holdout MSCOCO dataset [14]. The retained MSCOCO data set excludes all images that contain at least one of the eight MSCOCO objects. Eight objects are selected by clustering the word2vec embeddings on all 80 objects of the MSCOCO recognition challenge. This leads to the final eight novel objects for evaluation, namely "bottle", "bus", "bed", "microwave", "pizza", "rocket", "suitcase" and "zebra". These eight objects are kept in the training partition and appear only in the test partition. For a fair comparison, we use the same training, validation and test partition as in [14]. Note that visual information of a novel object may be present in the training set, even if there are no training sentences about the novel object. We manually inspected the training data and found that some images contained novel objects without annotated sentences in the training set. These novel objects were not significant in these images as the five human annotators did not mention them during annotation. Specifically, we found only 15 images containing a zebra.

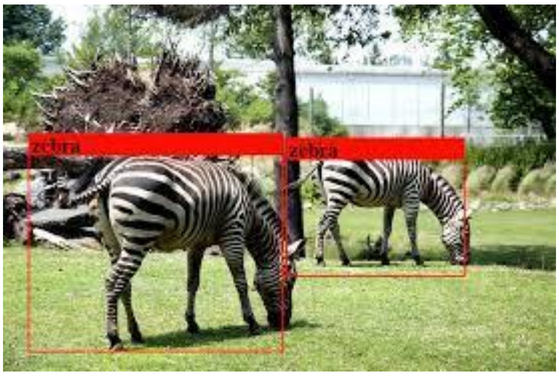


Figure 4. An example of training with the new object "Zebra". Red boxes indicate zebra locations. Note that the training sentences contain the novel object, but the subtitle sentences do not contain the object. Among COCO's 70,194 training images. Also, as shown in Figure 4, the novel object "zebra" is blurred in the 15 training images.. The red box in the figure represents the object zebra. Because the objects were so small, the five mortal evaluators ignored them and didn't mention" zebra" when giving their verbal description. These examples suggest that our model does not rely on the visual presence of these novel objects in training.

Scaling to ImageNet dataset. Following [15], [12], we use the same subset from ImageNet, which contains 646 objects not present in the MSCOCO dataset. This resulted in 164,909 images from the ImageNet dataset for testing. Similar to the previous methods, we take paired picture-sentence patterns in the MSCOCO training set as training data. We apply the trained model to generate caption sentences for images in the test subset of ImageNet. Since the ImageNet dataset does not contain paired Image Sentence data, we empirically assess the ability of our method to describe novel objects.

Nocaps dataset. The NoCaps dataset consists of images collected from validation and test sets of open images. It is a large-scale novel object captioning dataset containing 4,500 Confirmation images and 10,600 test images. Each image is annotated by 11 annotations. In total, the nocaps dataset contains 600 object classes. [35] Next, the model is trained using the COCO training data and directly tested on nocaps without finetuning. Covers many concepts outside the domain of NoCaps that are visually and linguistically similar to COCO but rarely described in COCO (e.g., seahorse, sewing machine).

4.2 Experimental Settings

The Object detection model. We use intimately available-trained object discovery models to make the crucial-value object memory. For experiments on the MSCOCO dataset, we use the Faster R-CNN [47] model with Inception-ResNet-V2 [46] to generate detection bounding boxes and scores. The object discovery model was pre-trained on all MSCOCO training images of 80 objects, including eight novel objects. We use pre-trained models released by [49] which are publicly available. As for the experiments on the ImageNet dataset, [12], we use the same object classifiers (16-layer VGG model) trained on the ImageNet ILSVRC12 dataset.

Evaluation metrics. To evaluate the quality of the generated title sentences, in our experiments, we use an efficient machine translation metric, the Metric for Evaluation of Translation with Explicit Ordering (METEOR) [48]. We also use F1-score as an evaluation metric [14], [15], [12]. The F1-score considers false positives, false negatives and true positives, indicating whether the generated sentence contains a new object or not. For the results in the ImageNet dataset, since there are no annotated ground verity rulings, we follow [15], [12] and use two more metrics for the novel object captioning task, namely describing novel objects (novel) and accuracy scores. As introduced in [15], the novelty score is the percentage of all novel objects mentioned in the predicted title sentences. The accuracy score was the percentage of pictures in which the novel object shown was correctly described by addressing the object in our generated title sentence.

Implementation details. For fair comparisons, we use VGG16 pre-trained as a visual encoder on the ImageNet dataset [23]. 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 fed to a decoder (switchable LSTM) as an initial state. For word embedding, unlike [14], [12], we do not require the counter-trained word embedding with additional knowledge data. Instead, we embed the term Øw with 1,024 dimensions for all terms. We use Tensorflow [49] to implement our framework. We optimize the model using the ADAM [50] optimizer with a learning rate of 1 × 10^{-3}. The weight decay was set to 5 × 10^{-5} to avoid overfitting. 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 in the hold-out MSCOCO dataset. All baseline methods, except LRCN, use additional semantic data consisting of eight novel object words. However, without external sentence data, our method achieves state-of-the-art competitive performance. Our model provides a higher average F1-score than the previous state-of-the-art result (60.08% versus 54.4%). The improvement is significant, as our model uses less training data. Our METEOR score is slightly worse than CBS [13]. The reason is twofold. On the one hand, CBS uses a beam search strategy, which is known to improve sentence performance. On the other hand, it uses many training sentences containing novel words in training. Consequently, it works in a more advantageous setting than our zero-shot setting, in which there were zero training sentences of novel objects. Compared to methods with additional sentence data, our method generates better titles for novel objects without this data. In addition, compared to our previous work DNOC [7], a zero-shot novel object captioning method, the improved version (switchable LSTM) significantly and consistently outperforms the previous version (DNOC) in all evaluation metrics. These results demonstrate the effectiveness of our SNOC framework and its ability to utilize both external and internal knowledge.

Describing Domain Objects. In addition to unseen (outside the domain) objects, we also verify the ability to describes

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	57.2	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]	38.1	78.40	55.93	53.77	18.43	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	75.58	49.25	51.28	76.85	30.68	58.32	82.60	57.13	21.19

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]	4.66	28.76	33.56	26.09	60.43	50.65	12.24	75.98	35.23	20
	NOC [12]										
	-(One hot)	14.21	60.76	49.87	31.98	65.43	56.44	28.23	85.31	54.77	19.7
	-(One hot +Glove)	12.76	44.65	54.98	32.31	66.44	59.54	26.22	87.66	50.22	20
	LSTM-C [13]										
	-(One hot)	23.87	65.98	22.55	12.87	70.32	44.87	50.43	89.76	53.65	21
	CBS [14]	15.4	57.67	44.98	23.54	66.56	53.2	45.32	76.87	55.98	22

		5									
	NBT+G[32]	6.1	67.98	32.66	56.43	50.66	22.9	42.65	77.54	38.2	21.8
	CRN [34]	32.1	75.54	50.43	49.77	78.65	56.44	54.98	86.91	53.49	20.9
	FDM-net [36]	-	-	-	-	-	-	-	-	60.44	20.5
Zero-shot	LRCN [4]	0	0	0	0	0	0	0	0	0	17.93
	DNOC [3]	30.13	65.54	45.43	53.76	66.98	33.24	54.09	76.93	44.93	20.76
	Novel	43.87	74.65	53.46	60.54	73.98	44.87	65.98	88.65	50.77	23.19

seen (in-domain) objects. In-domain testing focuses on describing objects seen during training. As the proxy visual words are computed by the cosine distance with the training objects (Eqn.(4)), it finds the category for the domain (training) object itself. So in this experiment, the proxy visual words were actually objects. Table 2 shows a comparison of our SNOC with the baseline method LRCN [4] and our previous version DNOC [7] on the hold-out MSCOCO dataset. Our switchable LSTM achieves higher F1-scores on known objects than these methods. Our method significantly outperforms the baseline LRCN [2] by 14.8 points on average F1 scores. The comparison results strongly support that our method can better describe objects in images even on seen (in domain) objects.

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

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]	67.87	49.23	54.98	50.54	66.44	59.09
DNOC[3]	76.45	67.43	70.53	64.23	68.76	74.21
Ours	80.65	76.33	74.79	72.67	81.34	80.44

TABLE 3

Results on the ILSVRC ImageNet dataset

Model	Novel	Accuracy
DCC [11]	54.34	10.76
NOC (One hot+ Glove) [12]	60.32	9.32
LSTM-C (One hot+ Glove) [13]	65.43	15.43
Ours	**93.21**	**35.66**

	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

Results on the No caps dataset. To enable our model to caption novel objects in no caps, we replaced our COCO pre-trained detection model with a detection model pretrained on an open image dataset. Others were kept identical to the COCO experiments, i.e., using the same pre-trained VGG-16 model and vanilla LSTM and image features from word embeddings with random initialization. The out-of-domain test results are shown in Table 4. Compared to the robust baseline Updown [10], our method significantly outperforms it on both the validation set and the test set under out-domain evaluation. Up down [10] uses enhanced image features (bottom-up features using a fast-RCNN detector pre-trained on the visual genome) and glove word embeddings, resulting in significant performance improvement. Note that NBT and CBS used additional glove word embeddings and an ELMo model that was pre-trained using an external large-scale corpus dataset. So their language models already saw sentences with unseen objects. For a fair comparison, we retrain NBT using the same pretrained models and input features as ours (indicated by * in Table 4). Our model outperformed the NBT by 7.2 points on a nocaps wall set at the zero-shot setting. Oscar [52] trained first fine-tuned in external 6.5 million text-image pairs and nocaps. Therefore, they are not suitable comparisons for our method. In contrast, we follow our zero-shot setting and use 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 descriptive and accurate sentences about novel objects.

TABLE 4

CIDE$_r$ 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 settings

Settings	Model	Val	Test
With External Semantic Data	NBT [32]	50.32	46.32
	NBT [32] + CBS [14]	61.55	50.21
	Oscar [51]	43.67	-
	Oscar [51] + CBS	76.21	-
Zero-shot	UpDown	30.54	29.08
	NBT [32] *	29.50	-
	Ours	44.25	40.54

Fig .5 Objects and sentences generation on

ImageNet.

4.4 Ablation Studies

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

Effect of proxy visual word. We propose a proxy visual term to represent an invisible object by some known object that has a similar visual appearance. Fig. 5 shows some examples of proxy words generated for the novel objects in the test set. The figure shows that these surrogate words are very close to the ground truth of these novel objects, suggesting that our surrogate ablation study is based on the deferred MSCOCO dataset.

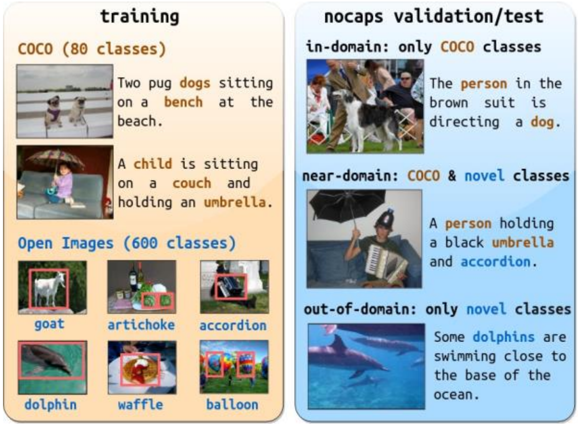

training
COCO (80 classes)
Two pug dogs sitting on a bench at the beach.
A child is sitting on a couch and holding an umbrella.
Open Images (600 classes)
goat
artichoke
accordion
dolphin
waffle
balloon
nocaps validation/test
in-domain: only COCO classes
The person in the brown suit is directing a dog.
near-domain: COCO & novel classes
A person holding a black umbrella and accordion.
out-of-domain: only novel classes
Some dolphins are swimming close to the base of the ocean.

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

TABLE 5

"We without Acquisition" means the SNOC Framework without Acquisition Mode enabled.. “Ours w/o addressing” indicates that we remove the addressing operation (Eqn. (8))

Model	F1$_{average}$	METEOR
Ours w/o Retrieving	0	17.31
Ours w/o addressing	35.32	16.98
Ours	61.43	20.88

visual words can partially describe an invisible object given limited knowledge. For example, we could construct a proxy visual word “sandwich” to represent it when completing a naive sentence, even though “pizza” did not appear during training. The sentences produced are reasonable and meaningful because the meaning and phrasal templates for these two objects are similar.

Effect of retrieving mode. In ablation experiments, we verify the effectiveness of this module by removing the entire retrieving module. In this setting, LSTM has only generating mode. As a result, LSTM detection does not leverage external knowledge from the model and thus performs poorly on these novel words. It can be seen from Table 5 that the model does not specify the invisible object either (F1$_{average}$ = 0). The results are far from our full model.

Effectiveness of content-based addressing operation. Our variable LSTM performs content-based addressing (Eqn (8)) on the object memory to select the correct noun word to describe the invisible object. The addressing operation combines a semantic language representation and a visual CNN feature such as a grounding operation. In Table 5, we also verify the effectiveness of the content-based addressing operation. "Ours w/o addressing" implies that we replace the content-based addressing operation by randomly selecting an identified object as the retrieved noun pr t. With the addressing operation, our full model outperforms "hours w/o addressing" by 19.91% in F1-score and 1.91% in METEOR score. The comparison suggests that the addressing operation in the retrieving mode improves the semantic understanding of the visual content and makes it easier to find the object of interest.

Analysis on different language patterns. We experimented with different language models such as BERT and GRU. Since our switching model design is based on a recurrent neural network, we only take the pre-trained BERT model for initialization of input embeddings. Table 6 shows that BERT improves our method on METEOR from 21.88 to 22.41. This suggests that better language models can further improve the performance of our method. We also replaced our base LSTM model with a gated recurrent unit (GRU). We found in experiments that GRU leads to similar F1-scores in captioning novel objects, but worse METEOR scores. The reason may be the lesser parameters of GRU compared to LSTM.

Analysis on different object detection models. We tried different identification models in building our framework. We compared three types of detection models in experiments, i.e., Faster RCNN with Inception ResNet v2, Mask-RCNN with Inception ResNet v2, SSD with ResNet50 FPN. The results are shown in Table 7. We find that our detection model (Faster RCNN with Inception ResNet v2) achieves the highest performance among all competitors.

TABLE 6

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

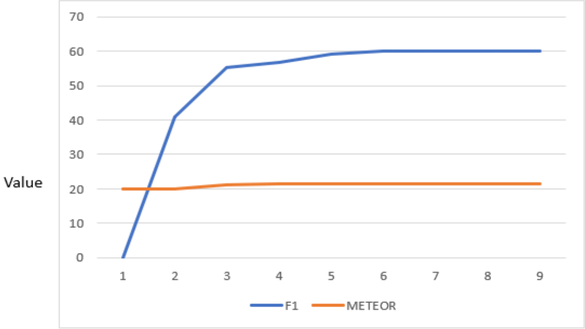
Model	$F1_{average}$	METEOR
Ours (LSTM)	56.09	20.43
Ours + BERT	58.32	21.66
Ours (GRU)	59.87	22.43

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)	56.09	21.55
Ours + SSD (ResNet-50 FPN)	57.33	20.44
Ours + Mask-RCNN (Inception-ResNet)	59.76	21.98

Analysis of maximum object memory size N_M. N_M represents the maximum number of slots in our object memory, i.e., the number of object-label pairs per image. If the memory size is too small, the external knowledge considered in our SNOC will be less. If we set the memory size too large, it introduces too many noisy candidates and thus limits performance. We show the average F1-scores and METEOR scores over different memory sizes in Fig. 7. N_M=0 means no detection output is used in the framework. We can see from the figure that F1-scores are relatively low when N_M is less than three. The reason is that introduced external knowledge is insufficient because only very few recognized objects are in memory as candidates. As N_M increases above three the performance curve flattens. We also notice a slight performance drop (F1-scores drop from 60.1 to 59.8) when the memory size is too large. The reason is that too many noisy candidate objects may be written to memory, making content-based addressing less reliable. The results for different object memory sizes also confirm the effectiveness of our retrieving mode. It can be seen that by introducing enough external recognition knowledge, our SNOC is capable of describing novel objects even without any related training sentences.



Memory Size N_m

TABLE 8

Impact of Reinforcement Learning on the held-out MSCOCO dataset

Model	$F1_{average}$	METEOR
Ours	58.55	20.33
Ours + SCST [31] (CLD$_{er}$)	57.21	21.47
Ours + SCST [31] (METEOR)	61.77	22.99

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

4.5 Qualitative Results

We qualitatively show an example from the test set of the held-out MSCOCO dataset in Fig. 8.in an image detecting the tennis racket and badminton racket based on their differences. The lines in the red colour are generated in the retrieving mode, while text is from the generating mode. We use Switchable LSTM predicts words by words to form a naive sentence. 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 finish a naive sentence with proxy visual words, and then replace them with the accurate unseen words.

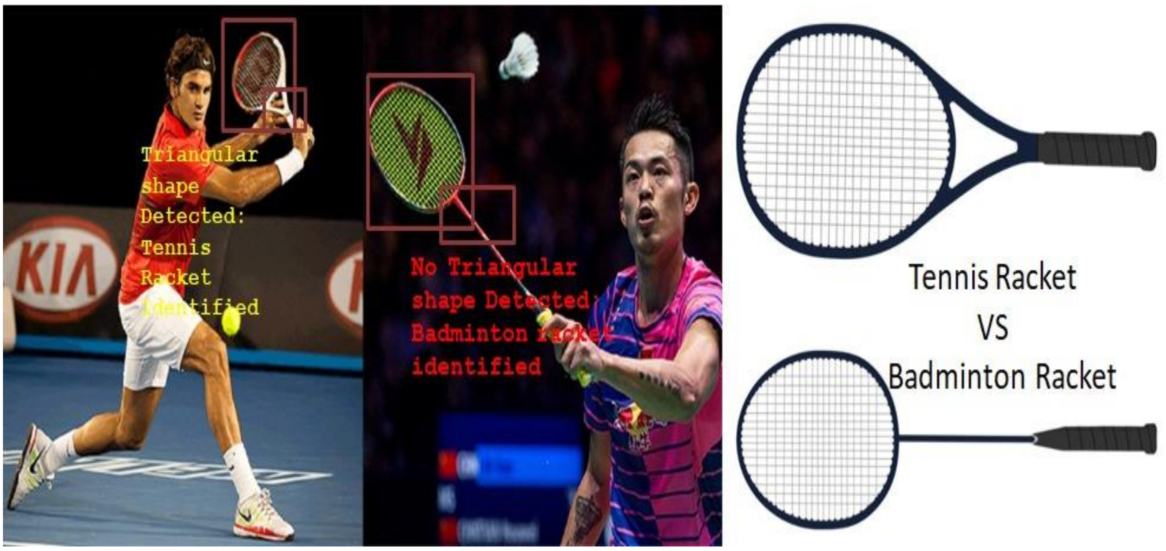

KIA
Triangular shape Detected: Tennis Racket identified
No Triangular shape Detected: Badminton racket identified
Tennis Racket
VS
Badminton Racket

Fig 8. Correct word: a person is holding a tennis racket.

In the fig.8, we observe that the Tennis racket is in oval shape along with a triangular shape whereas, in badminton racket it contains only oval shape and it has no triangular edge. The previous system detects the object only based on the shape but it doesn’t find out any differences. In present system it detects the objects based upon the differences using Switchable LSTM and novel along with python it generates the caption of the above fig as Tennis racket.

The test set of the held-out MSCOCO dataset. Generalization ability and detects the picture successfully.

V. CONCLUSION

We have presented Long Short-Term Memory with Pointing (LSTM-P) architecture which produces novel objects in image captioning. 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 medium is cooked to balance the word generation from RNN- grounded decoder and the word taken directly from the learnt objects. also, the judgment - position content of objects is further exploited to cover further objects in the judgment and therefore ameliorate the captions. To introduce external knowledge into sentence generation, we propose a switchable LSTM that has two switchable working modes, i.e., 1) generating sentences like a standard LSTM and 2) retrieving the correct noun from key-value memory. We generate a new index in the LSTM cell to transform the two modes. Our switchable LSTM can utilize both internal knowledge and external knowledge. Our experiments confirm its effectiveness on both the hold-out MSCOCO dataset and the ImageNet dataset. Without additional sentence data, our method also outperforms state-of-the-art methods that use additional linguistic data.

REFERENCES

[1] Y. Wu, L. Jiang and Y. Yang, "Switchable Novel Object Captioner," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1,pp.1162-1173, 1 Jan. 2023, doi: 10.1109/TPAMI.2022.3144984.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

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

[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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] Y. Wu and Y. Yang, "Exploring heterogeneous clues for weaklysupervised audio-visual video parsing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 1326–1335.

[20] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in Proc. Int. Conf. Mach. Learn., 2014, pp. 595–603.

[21] M. Mitchell et al., "Midge: Generating image descriptions from computer vision detections," in Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics, 2012, pp. 747–756.

[22] 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.

[23] 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.

[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] A. Farhadi et al., “Every picture tells a story: Generating sentences from images,” in Proc. Eur. Conf. Comput. Vis., 2010, pp. 15–29.

[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] 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.

[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] 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.

[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] 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.

[32] 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.

[33] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Selfcritical sequence training for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 7008–7024.

[34] J. Lu, J. Yang, D. Batra, and D. Parikh, “Neural baby talk,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7219–7228.

[35] H. Agrawal et al., “nocaps: Novel object captioning at scale,” in Proc. Int. Conf. Comput. Vis., 2019, pp. 8948–8957

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[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] 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.

[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] 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.

[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] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Representations, 2015.

[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] 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.

[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] 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.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Representations, 2015.

[51] 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.

[52] X. Li et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in Proc. Eur. Conf. Comput. Vis., 2020, pp. 121–137.