ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230381

Reasoning Guided by a Manual: Context-Aware Image Captioning with Novel Objects

Peiyao Hua[a], Haifeng Sun [a;*], Jiachang Hao[a], Cong Liu[b], Jingyu Wang[a], Qi Qi[a] and Jianxin Liao[a]

[a]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[b]China Mobile

Abstract. Novel object captioning task aims at describing objects that are absent from training data. Due to the scarcity of novel objects, it's challenging to find a way to utilize external data to improve model's reasoning ability. While previously designed methods all follow deep learning approach, we boost novel object captioning by incorporating reasoning with traditional deep learning framework. We design a manual from dictionaries that provides our model with sufficient and accurate external information on novel objects. We propose Manual-guided Context-aware Novel Object Captioning model (MC-NOC) that utilizes image and caption context to generate novel object captions. It contains a Manual-Guided Novel Object Reasoning module to reason about novel objects based on other objects of the given image and a Caption Reconstruction module to incorporate novel objects into generated captions according to caption context. We validate MC-NOC with state-of-the-art performance on the challenging Held-out COCO and Nocaps dataset, leading their leaderboard. In particular, we improved the CIDER metric by 6.4 points on the held-out coco dataset. Comprehensive experiments demonstrate our model's reasoning capability and the quality of generated captions.

1 Introduction

Image captioning is an essential task that describes image content, holding the potential to advance human-computer interaction and image comprehension. Deep learning approaches have demonstrated the ability to learn from large volumes of data and generate precise captions. However, they encounter challenges when attempting to caption novel objects absent from training sets. Examples of these objects include charging piles, robot vacuums, and drones, which are new products or expressions in fast-growing industrial fields lacking annotated labels, making them difficult for models to caption. To address this challenge, the novel object captioning task has been introduced, which aims to efficiently obtain visual information from an image and align it with corresponding linguistic descriptions of the novel object.

Currently, there are three main approaches to novel object captioning task. The first approach involves using pre-trained models [16, 36, 21, 25, 30, 31] that align each word representation with corresponding object representation in a common space, facilitating the mapping of novel objects to their descriptions. While these models can caption objects in general scenarios, they still require adaptation for specific domains or datasets to achieve better performance. The second approach is partially supervised training [6, 8, 4], which uses object features from object detection datasets to enhance original images in image caption datasets. The third approach involves utilizing pre-trained object detectors to identify novel objects, which are then incorporated into generated captions [33, 7, 9, 32, 13, 11, 22, 24, 34, 32, 3, 14]. Object detection datasets directly influences the accuracy of captions produced by the latter two methods. These approaches skip the role of other objects in the image, which could serve as essential cues when predicting novel objects. Therefore, we aim to incorporate caption context to improve our model's robustness and generalizability.

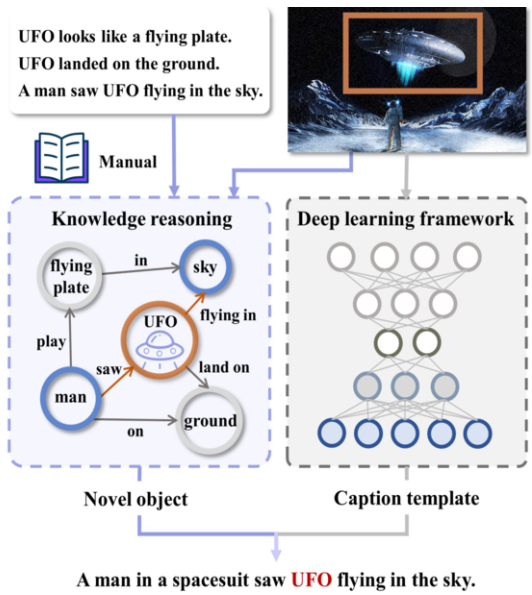

Figure 1: An example illustrating how we integrate knowledge-based reasoning into novel object caption process.

We gain inspiration from human reasoning process. Humans leverage prior knowledge to reason about unfamiliar objects based on other objects in the scene. As illustrated in Figure 1, when encountering a flying plate-shaped craft in the sky with a man in a spacesuit, humans can recognize it as a UFO and generate corresponding captions. Unlike reciting descriptive sentences, humans remember

* Corresponding Author. Email: hfsun@bupt.edu.cn

objects in context with other objects, as if using a manual that lists objects and their relationships to one another. This knowledge can bridge the gap between familiar and novel objects, enabling effective reasoning. Recent studies [10] have shown that incorporating reasoning into deep learning models can improve their efficiency, generalizability, interpretability, robustness, and adaptability to real-world scenarios. The difficulties of incorporating reasoning into novel object captioning tasks are as follows: first, how to combine explicit knowledge reasoning process with implicit deep learning framework, then how to integrate reasoned results in caption context.

Our approach utilizes visual and linguistic contextual information extracted from a deep learning framework and a manual construct of objects and their relations. We perform unsupervised knowledge reasoning via Conditional Random Fields (CRF) to obtain the probability distribution for each novel object across all object categories. The location of novel object within each caption is determined based on linguistic contextual information. We then modify the following context to fit novel object. As a result, our approach generates captions that not only include novel objects but are also fluent and contextually relevant.

We propose a Manual-guided Context-aware Novel Object Captioning model (MC-NOC) for knowledge reasoning and caption generation. MC-NOC consists of a Manual-Guided Novel Object Reasoning module (M-NOR) and a Caption Reconstruction (CR) module. M-NOR module extracts objects visual feature, and refer to the manual to reason possible novel objects based on other objects in the image context. CR module calculates position attention according to caption context to determine the most appropriate position for each novel object and generate coherent captions. In this case, our model utilizes context information from both vision and language modality to reason richer and more accurate novel object captions. In summary, our contributions are as follows:

- Provide a novel solution for novel object captioning tasks that integrates knowledge-based reasoning into traditional deep learning framework.
- Design a MC-NOC framework to reason about novel object captions with a structured manual and contextual information.
- Extensive experiment results demonstrate that our MC-NOC model outperforms state-of-the-art methods in terms of accuracy and coherence.

Caveat: Our manual plays a critical role in our approach by providing the model with necessary knowledge to reason about novel objects. To ensure that our manual contains accurate and sufficient knowledge, we utilize dictionaries as our knowledge sources. We employ a triplet parser that identifies triplets from example sentences of dictionaries. Our manual can be updated online during training. With these triplets, we are able to create a straightforward and easily accessible manual for our model to consult.

2 Related work

In the past few years, the problem of novel object captioning has been proposed and rapidly developed [25]. Our work is related to novel object captioning methods that leverage external data to describe objects unseen during training phase. In general, there are three approaches to handling novel object caption task.

Pre-training approaches: Pre-training approaches use external visual and language data that contain potential information about novel objects, allowing models to acquire novel object captioning ability. NOC model [30] proposes minimizing a joint objective that can learn from these diverse data sources and leverage distributional semantic embeddings. NOC-REK [31] utilizes a large amount of natural language data for pre-training and views object detection process as retrieval from a dictionary The following models pre-train large-scale vision language transformers models and fine-tune the pre-trained model to adopt downstream tasks. Among them, OSCAR [21] uses object labels detected from images as anchor points in the learning process of semantic alignment between text and images. VinVL [36] designed a new target detection model for better visual features. VIVO [16] model was pre-trained using image-tag to align semantic tags with regional features of the image. The pre-trained approach performs well on generic datasets, but for specific scenarios, additional data is needed for fine-tuning.

Partially-supervised approaches: Partially-supervised approaches use object features from object detection datasets to reform the original image, thereby providing a much wider variety of object classes. In order to improve the generalizability of image caption models, PS3 [4] employs labels and objects of each image solely to generate its description. FDM-net [8] leverages external object features to deform and amplify the training data. PS-NOC [6] utilizes the context within existing image-caption pairs to generate pseudo-label descriptions for novel objects. These methods still face difficulties in generating descriptions of novel objects that are not included in object detection datasets.

Object detector approaches: Object detector approaches use an object detector pre-trained on object detection datasets that can identify a greater number of objects and incorporate them into captions. CBS [3] and Region Selector [7] use the output of object detector to restrict generated captions. LSTM-C [35] and LSTM-P [22] incorporate novel objects into captions using a copy mechanism. NBT [24] and ZSC [11] generate a sentence template that is filled with visual concepts to form captions. DNOC [34] and SNOC [33] generate sentences that contain placeholders and use a key-value object memory to retrieve words through querys. CRN [13] is a method that replaces an incorrectly identified object in the generated caption with a novel object. ECOL-R[32] uses copy networks and reinforcement learning methods to integrate novel objects into sentences. It should be noted that these approaches rely solely on the object detector as external knowledge. Therefore, the accuracy of the novel object description depends on the object detector's accuracy.

To improve the model's accuracy, some models use external visual or linguistic knowledge in addition to object detectors to help describe novel objects. ANOC [9] incorporats human attention features that capture essential information. DCC [14] leverages large object detection datasets and external text corpora to transfer knowledge between semantically similar concepts. [28] expands the vocabulary for captioning by using word embeddings of novel objects estimated from a small number of image features.

Although these approaches can describe a certain range of novel objects, we aim to enhance the novel object captioning method further by incorporating knowledge-based reasoning and leveraging manual knowledge from dictionaries to improve the quality and coherence of the generated captions.

3 Proposed Method

3.1 Framework

The MC-NOC model is composed of three modules: Manual-Guided Novel Object Reasoning (M-NOR), Template Generation, and Caption Reconstruction (CR), illustrated in Figure 2. Given an input

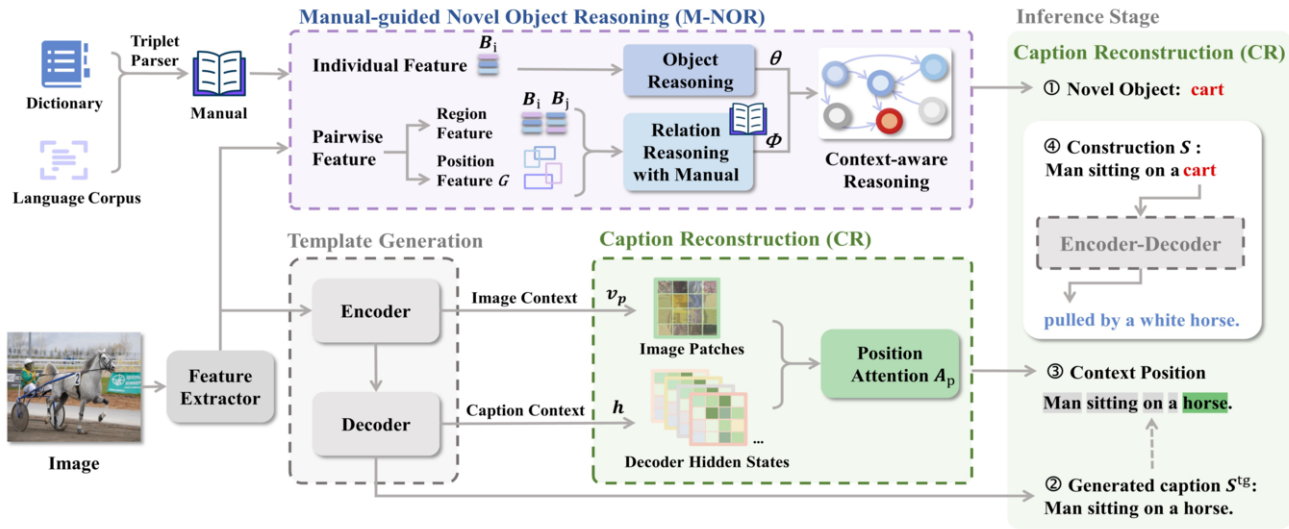

Figure 2: This is the framework for MC-NOC. Manual-Guided Novel Object Reasoning module (M-NOR) utilizes image features and manual knowledge to predict objects and their relationships. Object labels are updated using CRF during context-aware reasoning. Visual features are then inputted into an encoder-decoder Template Generation module. The Caption Reconstruction (CR) module uses encoder image patches and decoder hidden states to predict the position of each novel object. During inference, we replace novel objects according to predicted positions in caption templates and reconstruct output caption S.

image I, M-NOR module takes in region features B and position features G to make object reasoning and relation reasoning to predict class labels c. Template Generation module employs UpDown framework [5] to generate template caption, denoted as S^{tg}. Note that other image caption frameworks can also be compatible. Using CR module, we modify template captions to include novel objects. This module takes image patches v_p and decoder hidden states h as inputs and uses a position attention mechanism to predict positions of novel objects in corresponding caption. During inference stage, we refer to the novel object and its position to reconstruct template caption S^{tg}, generating the final output caption S.

To ensure the accuracy and completeness of the external knowledge, we extract knowledge from example sentences in dictionaries for general settings and collect corresponding corpora for specific scenarios. We use a triplet parser [37] to extract subjects, relations, and objects from example sentences, then combine them into triplets $\langle subject, relation, object \rangle$. In the process of reasoning, we increase the probability of object pairs that have the predicted relationship in the manual.

3.2 Manual-Guided Novel Object Reasoning

We investigate the use of knowledge-based reasoning to identify novel objects in conjunction with other objects in an image. By leveraging image context information and manual knowledge, we aim to improve the performance of novel object captioning.

To achieve our objective, we utilized the widely used object detector, Faster R-CNN, to extract object region feature B and position feature G. It should be noted that other detectors or image encoders could also be compatible with our approach. Specifically, Object Reasoning block takes object region feature B as input to predict the class of each object individually. The probability of class label c given by region B is represented by $\theta(c_i|B_i)$. On the other hand, Relation Reasoning block takes paired regions (B_i, B_j) as input to jointly predict class labels according to relation information provided by manual. The joint probability of paired class labels c_i and c_j given regions B_i and B_j, is represented by $\phi(c_i, c_j|B_i, B_j)$. Finally, we leveraged two characteristic equations of object reasoning and relation reasoning to perform context-aware reasoning with Conditional Random Fields (CRF) [12]:

$$P(c_N|B_N) = \sum_i \theta(c_i|B_i) + \sum_{i \neq j} \phi(c_i, c_j|B_i, B_j). \quad (1)$$

Specifically, we use a fully connected layer in Object Reasoning block to reason class probability:

$$P_c(c_i) = softmax(WB_i), \quad (2)$$

where W is a weight matrix. Thus, the potential of individual class label is given by

$$\theta_i(c_i) = \log P_c(c_i|B_i). \quad (3)$$

The manual records objects and their relationships in a triplet format. Relation Reasoning block increases the probability of paired objects recorded in manual with an indicator function δ. Let r_k be the likelihood of the relation between two objects, $\delta(r_k; c_i, c_j)$ represents the triplet indicator to judge whether $\langle c_i, r_k, c_j \rangle$ is reasonable in the current scene. Moreover, based on [15], we use t_η to embed the position feature of two objects g_{ij} into high-dimensional relation space to generate relation potential:

$$\mathcal{L}(r_k; B_i, B_j) = MLP(t_\eta(g_{ij})). \quad (4)$$

Thus, the binary potential is given by:

$$\phi(c_i, c_j|B_i, B_j) = \sum_k \delta(r_k; c_i, c_j)\mathcal{L}(r_k; B_i, B_j). \quad (5)$$

Since novel objects could be incorrectly identified or undetected by Faster R-CNN, we add a position on the CRF chain for novel object and take image feature as its region feature B_i.

To train the CRF in Context-aware Reasoning block, we utilize a loss function based on pseudo-likelihood, which calculates the likelihood of each object class given ground truth of other class labels.

During training, we extract ground truth labels of each novel object from caption annotations. The objective of the training process is to minimize following loss function L_{nor}. Here, c_i^* represents ground truth label of B_i, while $c_{\setminus i}^*$ denotes ground truth labels of other regions in the image:

$$L_{nor} = -\sum_i \log P(c_i^* | c_{\setminus i}^*), \tag{6}$$

where

$$P(c_i^* | c_{\setminus i}^*) = \frac{\exp \sum_{j \neq i} [\theta_i(c_i^*) + \phi_{ij}(c_i^*, c_j^*) + \phi_{ji}(c_j^*, c_i^*)]}{\sum_c \exp \sum_{j \neq i} [\theta_i(c) + \phi_{ij}(c, c_j^*) + \phi_{ji}(c_j^*, c)]}. \tag{7}$$

3.3 Caption Reconstruction

Caption Reconstruction module is designed to integrate identified objects into template captions based on contextual information present in template captions. Directly inserting a novel object into the template caption, or replacing a novel object with an existing object in the caption, can result in poorly constructed sentences that are difficult to comprehend. To overcome this challenge, we leveraged contextual information in the encoder-decoder architecture to identify the most appropriate position in the sentence to introduce novel object. Then, we generated a smooth context based on novel object and content present above, ensuring that output caption S is both accurate and comprehensible.

Intuitively, description information of a novel object is often implied in the context of its caption, and novel objects are characterized with low confidence detected by Faster R-CNN. To leverage these features, we compute position attention scores between the caption words and image patches. Empirically, we figure that novel objects are more likely to appear when the corresponding word's position attention score is high, but object confidence in the corresponding image patch is low. We identify such positions in the caption as the likely locations where novel objects should be located.

Specifically, the encoder in Template Generation provides image patches v_p of each image and Faster R-CNN detector provides confidence s_c for each image patch. We use decoder hidden state h to query image patch feature v_p. Thus we can get the attention score for each image patch:

$$A_p = atten(h, v_p). \tag{8}$$

Since novel object is more likely to be in a place where object confidence is low and patch attention is high. The possible position s_p of a novel object in the caption is given by:

$$s_p = softmax((1 - s_c) \times A_p). \tag{9}$$

Given ground truth novel object position s_p^*, we minimize the following binary cross-entropy loss for all positions P_{total}:

$$L_{cr} = -\sum_{p=1}^{P_{total}} s_p^* \log s_p + (1 - s_p^*) \log(1 - s_p). \tag{10}$$

Similarly, we calculate cross-entropy loss of Template Generation model. Given a ground truth sequence of length T denoted as $y_{1:T}^*$, the loss can be expressed as:

$$L_{tg} = -\sum_{t=1}^{T} log P(y_t | y_{1:t-1}^*). \tag{11}$$

By summing the loss of Template Generation module L_{tg}, M-NOR module L_{nor}, and CR module L_{cr} with corresponding bias b_{tg}, b_{nor}, b_{cr}, we obtain the loss L of MC-NOC:

$$L = b_{tg} L_{tg} + b_{nor} L_{nor} + b_{cr} L_{cr}. \tag{12}$$

After training, we feed novel objects according to their positions into the beam search decoding process to reconstruct captions. Position attention guarantee that novel object in this position is consistent with the former context's semantics. We replace novel objects with their corresponding positions. The following context is generated according to novel object in beam search process. This guarantees that the reconstructed caption can appropriately describe novel objects and solves the problem of caption incoherency.

4 Experiments

4.1 Dataset and Evaluation Metrics

Datasets. We compared our MC-NOC model with state-of-the-art approaches on two public datasets specifically designed for the novel object captioning task: the Held-out COCO dataset [14] and the Nocaps dataset [1]. The Held-out COCO dataset is a subset of the MS COCO dataset [23]. It comprises 70,194 fully paired data instances that exclude image-caption pairs describing any of eight novel object types: bottle, bus, couch, microwave, pizza, racket, suitcase, and zebra. The Nocaps training set consists of image-caption pairs from the MS COCO dataset as well. Moreover, Nocaps provides 166,100 human-generated captions describing 15,100 images from the Open Images dataset [19].It contains 600 image classes, including 119 in-domain classes that frequently appear in the MS COCO dataset, and the rest are out-domain classes. Captions containing both in-domain and out-domain classes are referred to as near-domain captions.

Evaluation metrics. To evaluate the quality of generated captions, we utilized three metrics: CIDEr [29], METEOR [20], and SPICE [2]. CIDEr measures the similarity between reference captions and generated outputs using TF-IDF weighted n-gram overlap. METEOR focuses on aligning words in the reference captions with generated outputs. SPICE is based on scene graphs matching between words in reference sentences and generated outputs. A higher score in any of these metrics indicates better performance of the model in generating accurate, diverse, and informative captions that describe novel objects in the image. We also evaluated F1-score of generated novel objects in both datasets to measure the accuracy and recall of generated novel objects. This evaluation helps to ensure that our model correctly identifies novel objects.

4.2 Implementation Details

We utilized UpDown baseline [5] as Template Generation module. To ensure a fair comparison, we employed the commonly used Faster R-CNN as Feature Extractor, which was pre-trained on Visual Genome dataset [27].

Our manual contains both in-domain and out-domain objects from the Held-out COCO and Nocaps datasets. To construct the manual, we collected raw illustrative sentences from Wiktionary and VG training set [17]. We then used a triplet parser to extract triplets containing novel objects. To ensure the relevance of extracted triplets, we filtered them by selecting the 70 most frequently occurring relations. This process resulted in a manual containing 331 objects. Due to the scarcity of novel objects in training set, we randomly masked

Table 1: Evaluation results on the Nocaps dataset.

Method	Validation set								Test set							
	in-domian		near-domian		out-domian		Overall		in-domian		near-domian		out-domian		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
NBT [24]	62.3	10.3	61.2	9.9	63.7	9.1	61.9	9.8	63.0	10.1	62.0	9.8	58.5	8.8	61.5	9.7
UpDown [5]	78.1	11.6	57.7	10.3	31.3	8.3	55.3	10.1	74.3	11.5	56.9	10.3	30.9	8.1	54.3	10.1
UpDown+ELMo [26]	79.3	**12.4**	73.8	11.4	71.7	9.9	74.3	11.2	76.0	11.8	74.2	11.5	66.7	9.7	73.1	11.2
VIT-GPT2 [18]	78.2	12.0	64.1	11.1	69.4	10.0	75.0	11.4	73.7	12.0	64.0	11.1	43.2	9.4	61.5	10.9
ANOC [9]	85.8	11.8	78.9	11.6	73.0	9.7	78.7	11.2	82.0	11.9	77.4	11.6	68.3	**10.0**	76.4	11.4
MC-NOC	**87.1**	12.0	**79.5**	**11.8**	**73.3**	**10.1**	**79.3**	**11.4**	**85.1**	**12.3**	**78.5**	**11.7**	**69.5**	**10.0**	**77.7**	**11.5**

Table 2: Evaluation results on the Held-out COCO dataset.

Method	Avg. F1-score	SPICE	METEOR	CIDER
DCC [14]	39.8	3.4	21.0	59.1
NOC [30]	48.8	-	21.4	-
NBT [24]	48.5	15.7	22.8	77.0
CBS [3]	54.0	15.9	23.3	79.9
DNOC [34]	57.9	-	21.6	-
ZSC [11]	29.8	14.2	21.9	-
LSTM-P [22]	60.9	16.6	23.4	88.3
CRN [13]	64.1	-	21.3	-
ANOC [9]	64.3	18.2	25.2	94.7
SNOC [33]	60.1	-	21.9	-
MC-NOC	**66.5**	**20.9**	**28.1**	**101.1**

off 10 in-domain object classes as novel objects during training to improve the generalization ability of our MC-NOC model.

During training, we trained Template Generation module with SGD optimizer, while Manual-Guided Novel Object Reasoning and Caption Reconstruction modules were trained with Adam optimizer and a Lambda learning rate scheduler. Let N_{iter} be the total number of iterations, and the learning rate of each iteration is given by:

$$lr_\lambda = \frac{1 - n_{iter}}{N_{iter}}. \tag{13}$$

We initialized the learning rate for each module to 0.0005, with loss bias $b_{tg} = 0.05$, $b_{cr} = 1$, and $b_{nor} = 0.5$. Following Nocaps baseline [1], we trained our model for $N_{iter} = 70,000$ with a batch size of 50, and set beam size to $k = 5$ during test time. It takes approximately 5 hours to train our MC-NOC with a GTX-3090 GPU.

4.3 Quantitative Evaluation

Result on the Held-out COCO dataset. We compared MC-NOC with state-of-the-art models on Held-out COCO dataset, to validate accuracy and quality of the generated caption. These approaches can be divided into four categories: DCC [14] and NOC [30] transfer knowledge from unannotated text corpora to generated captions. ANOC [9] and SNOC [33] adopt external visual information to enhance model's novel object recognition ability. NBT [24], DNOC [34], CRN [13], ZSC [11] and LSTM-P [22] design template based method to incorporate novel objects into captions. NBT, DNOC, CRN, CBS [3] adopt pre-trained object detectors to predict novel objects. For a fair comparison, we apply the same Faster R-CNN detector and CBS [3] method on all compared models. We use F1-score, SPICE, METEOR, and CIDEr metrics to measure how well MC-NOC can recognize and describe the novel object in the image. As shown in Table 2, our model outperforms the state-of-the-art methods in F1-score, SPICE, METEOR, and CIDEr metrics on Held-out COCO dataset. Particularly, our MC-NOC method outperforms ANOC by 2.2 points in F1-score, 2.7 points in SPICE, 2.9 points in METEOR, and 6.4 points in CIDEr. The average F1-score demonstrates that our method can improve the accuracy of predicted novel objects. Having a higher score in SPICE, METEOR, and CIDEr metrics indicates that the quality of our generated sentences has also been improved.

Result on the Nocaps dataset. We compare MC-NOC with the following state-of-the-art models on Nocaps dataset, including NBT [24], UpDown [5], UpDown+ELMo [26], VIT-GPT2 [18] and ANOC [9]. As mentioned earlier, we applied the same CBS method to all the compared models to ensure a fair comparison. Our model outperformed all other state-of-the-art approaches on 14 out of 16 evaluation metrics, as shown in Table 1. Although MC-NOC did not surpass the SPICE score of the pre-trained models UpDown+ELMo and VIT-GPT2 on the in-domain validation set, we believe this may be attributed to data bias arising from the small validation set size. Nonetheless, MC-NOC surpassed all other models in other metrics. This result is noteworthy as the Nocaps dataset comprises a more diverse range of novel objects, demonstrating that our model can improve the quality of novel object descriptions in a wider range of scenarios. Remarkably, our model achieved a 1.2 point improvement over ANOC on the out-domain Nocaps test set.

Table 3: Evaluation of novel object reasoning ability.

Method	out-domain			Overall		
	Accuracy	Recall	F1	Accuracy	Recall	F1
UpDown [5]	8.5	11.6	9.8	39.5	24.1	29.9
VIT-GPT2 [18]	10.9	18.4	13.7	40.6	30.3	34.7
ANOC [9]	25.2	51.0	33.7	44.2	58.7	50.4
MC-NOC	**28.4**	**51.6**	**36.5**	**44.4**	**59.1**	**50.7**

The evaluation results on the Held-out COCO and Nocaps datasets demonstrate the accuracy and quality of captions generated by MC-NOC. To further verify the effectiveness of the reasoning process and the accuracy of the reasoned novel objects, we conducted experiments to compare the accuracy, recall, and F1-score of captions generated from both out-domain and entire-domain validation images. The ground truth novel object labels were extracted from the caption annotations of the Open Images dataset, which provides a comprehensive description of various types of novel objects. As shown in Table 3, We compared these ground truth labels with the object classes in captions generated by three different models: Our method surpass UpDown [5], VIT-GPT2 [18] and ANOC [9]. To prevent novel objects from being overwhelmed by in-domain objects, we make comparisons on both out-domain and entire-domain data. The results demonstrate the effectiveness of knowledge-based reasoning, bringing a 3.2 points improvement on the out-domain accuracy score for our MC-NOC model.

Figure 3: Examples of generated captions on Nocaps dataset.

Table 4: Different sizes of the manual.

number of	100%	70%	50%	20%
object	331	232	166	66
relation	70	70	70	70
triplet	8475	4854	2004	398

Table 5: Impact of the manual size on caption generation.

Size of Manual	out-domain			Overall		
	Accuracy	Recall	F1	Accuracy	Recall	F1
100%	28.4	51.2	36.5	44.4	59.1	50.7
70%	27.6	51.0	35.8	44.2	58.5	50.4
50%	27.3	51.0	35.6	44.0	58.4	50.2
20%	27.2	50.8	35.4	44.0	58.4	50.2

4.4 Ablation Studies

Impact of the size of the manual. We design an ablation experiment to investigate the impact of manual size on MC-NOC's reasoning ability. We counted the number of objects, relations, and triplets in manuals of varying sizes and recorded the statistical information in Table 4. We use the same experimental setup as Table 3. According to the results presented in Table 5, we found that when the manual size is reduced, the reasoning ability of MC-NOC decreases. Even though the number of relations remains constant, the number of triplets matching the image context decreases. When the manual size is reduced from 100% to 50%, the accuracy score drops by 1.1 points in the out-domain and by 0.4 points in the entire domain. However, if the number of triplets is reduced by 50% or more, the reasoning ability of MC-NOC remains unaffected since MC-NOC tends to predict relation triplets in-domain. Due to the scarcity of novel objects, this manual has less impact on the entire domain. These results suggest that the M-NOR module can effectively enhance the performance of novel object captioning by accurately integrating novel objects based on context information.

Ablation study of the position attention and caption context. We have also assessed the importance of the Caption Reconstruction module in generating coherent captions. We compared three different caption reconstruction methods. The first method, Position Prediction + Context Information approach used in MC-NOC, involves inputting predicted positions and novel objects into the Caption Reconstruction module to reconstruct captions according to caption context. The second method Position Prediction + Replacement approach only replaces original word in the template sentence with the predicted novel object according to predicted position without modifying other caption content. The third approach adopts a conventional copy mechanism to incorporate novel object. We ensured that all other modules remained constant to enable a fair comparison. The results presented in Table 6 show that Position Prediction + Replacement underperforms Position Prediction + Context Information method since the predicted position can only accommodate the context before the position, not after it. Hence, the caption reconstruction process that utilizes caption context during inference is critical. Compared with the traditional copy network, the Position Prediction + Context Information approach can significantly improve the CIDEr metric by nearly 10 points. Indicating that MC-NOC's predicted position of novel objects is more accurate than the copy network. This promising result implies that the Caption Reconstruction module can precisely predict the novel object position and the caption reconstruction process can guarantee contextual fluency of generated captions.

4.5 Qualitative Analysis

We provide a quantitative comparison of the generated captions among MC-NOC, ANOC, and UpDown baseline in Figure 3. These examples illustrate that (1) the generated captions include appropriate novel objects such as parking lot, lobster, stretcher, pajamas, and wheelchair. (2) MC-NOC generates more specific captions and makes reasonable extensions based on novel objects and context in-

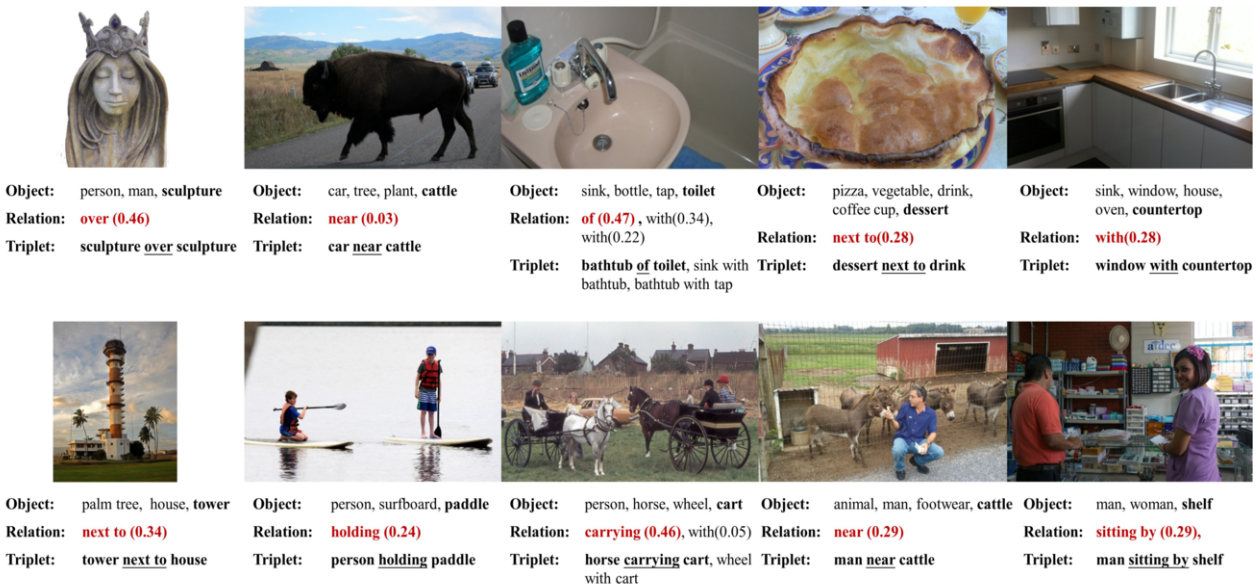

Figure 4: Examples of reasoned triplets.

Table 6: Ablation result of Caption Reconstruction module.

Method	Nocaps Validation set							
	in-domian		near-domian		out-domian		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Position Prediction+Context Information	**87.1**	**12.0**	**79.5**	**11.8**	**73.3**	**10.1**	**79.3**	**11.4**
Position Prediction+Replacement	84.8	11.7	78.4	11.5	71.8	9.7	78.0	11.1
Copy Mechanism [35]	76.7	11.5	70.9	10.9	67.4	9.5	71.0	10.7

formation. For instance, in the first example, MC-NOC adds "jeans" to describe what the men are wearing and "parking lot" as scene information. In the second example of the second row, MC-NOC brings in "basket" from image context. (3) MC-NOC makes reasonable guesses based on image context. For instance, as a result of reasoning, the first example of the second row predicts "carrots" based on "lobster" since they are closely related in the manual. As the manual expands, MC-NOC will make more rational reasoning and pays more attention to image context to describe image scenes at a fine-grained level. With the help of the manual, generated captions incorporate a boarder range of object categories. These quantitative results prove that our model significantly improves the diversity and explicitness of generated captions, leading to more coherent and accurate descriptions of novel objects.

We also visualized the intermediate triplets predicted by the M-NOR module to demonstrate MC-NOC's ability to reason about novel objects. Figure 4 presents detected objects, inferred relations, and predicted triplets. Instead of using supervised information to train the model's triplets recognition ability, we only use manual knowledge to enhance the probability of object categories paired with relationships documented in the manual during training. MC-NOC provides reasonable representations of triplets, and we observe that the model tends to predict more general and straightforward relations such as "next to", "near", and "with". However, the model also predicts relatively complex and precise relationships such as "carrying" and "holding". Visualization results demonstrate the accuracy of reasoned multiple triplets, and MC-NOC selects the most appropriate novel object for each caption.

5 CONCLUSIONS

In this paper, we proposed a novel approach to address the challenges of novel object captioning, often hindered by the scarcity of novel objects. By incorporating knowledge-based reasoning into traditional deep learning framework, our Manual-guided Context-aware Novel Object Captioning model (MC-NOC) achieves promising results in generating captions for novel objects. Our approach utilizes a manual from dictionaries to provide MC-NOC with sufficient and accurate external information. Experimental results demonstrate the effectiveness of our reasoning approach and the significance of considering image and caption context when captioning novel objects. Our work provides a direction for future efforts in novel object captioning and related cross-modality tasks. Overall, we believe that our approach can pave the way for generating more robust and reliable captioning for novel objects, enabling more sophisticated applications in natural language processing and computer vision.

Limitations

We face limitations in accurately inferring objects with limited or no information in the manual, like orcas. We filtered out relationships that appeared fewer than 20 times during manual construction. Orcas are represented by just one triple: "orcas in the sea." As a result, our model often misclassifies orcas. Our future research will focus on improving and exploring alternative methods beyond Object Reasoning. Our aim is to enable the model to generate better descriptions of novel objects and handle instances with limited information for more reliable predictions.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants (62201072, 62171057, 62071067, 62001054), in part by the Ministry of Education and China Mobile Joint Fund (MCM20200202), in part by the Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

References

[1] Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee, 'nocaps: novel object captioning at scale', in *ICCV*, pp. 8947–8956, (2019).

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, 'SPICE: semantic propositional image caption evaluation', in *ECCV*, pp. 382–398, (2016).

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, 'Guided open vocabulary image captioning with constrained beam search', in *EMNLP*, pp. 936–945, (2017).

[4] Peter Anderson, Stephen Gould, and Mark Johnson, 'Partially-supervised image captioning', in *NeurIPS*, eds., Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, pp. 1879–1890, (2018).

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, 'Bottom-up and top-down attention for image captioning and visual question answering', in *CVPR*, pp. 6077–6086, (2018).

[6] Shashank Bujimalla, Mahesh Subedar, and Omesh Tickoo, 'Partially-supervised novel object captioning using context from paired data', in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, p. 649. BMVA Press, (2022).

[7] Marco Cagrandi, Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara, 'Learning to select: A fully attentive approach for novel object captioning', in *ICMR*, eds., Wen-Huang Cheng, Mohan S. Kankanhalli, Meng Wang, Wei-Ta Chu, Jiaying Liu, and Marcel Worring, pp. 437–441. ACM, (2021).

[8] Tingjia Cao, Ke Han, Xiaomei Wang, Lin Ma, Yanwei Fu, Yu-Gang Jiang, and Xiangyang Xue, 'Feature deformation meta-networks in image captioning of novel objects', in *AAAI*, pp. 10494–10501, (2020).

[9] Xianyu Chen, Ming Jiang, and Qi Zhao, 'Leveraging human attention in novel object captioning', in *IJCAI*, pp. 622–628, (2021).

[10] Zijun Cui, Tian Gao, Kartik Talamadupula, and Qiang Ji, 'Knowledge-augmented deep learning and its applications: A survey', *CoRR*, (2022).

[11] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis, 'Image captioning with unseen objects', in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, p. 146. BMVA Press, (2019).

[12] Santosh Kumar Divvala, Derek Hoiem, James Hays, Alexei A. Efros, and Martial Hebert, 'An empirical study of context in object detection', in *CVPR*, pp. 1271–1278, (2009).

[13] Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, Mingliang Xu, and Yi Yang, 'Cascaded revision network for novel object captioning', *IEEE Trans. Circuits Syst. Video Technol.*, **30**(10), 3413–3421, (2020).

[14] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond J. Mooney, Kate Saenko, and Trevor Darrell, 'Deep compositional captioning: Describing novel object categories without paired training data', in *CVPR*, pp. 1–10, (2016).

[15] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei, 'Relation networks for object detection', in *CVPR*, pp. 3588–3597, (2018).

[16] Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu, 'VIVO: visual vocabulary pre-training for novel object captioning', in *AAAI*, pp. 1575–1583, (2021).

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei, 'Visual genome: Connecting language and vision using crowdsourced dense image annotations', *Int. J. Comput. Vis.*, **123**(1), 32–73, (2017).

[18] Ankur Kumar, 'The illustrated image captioning using transformers', *ankur3107.github.io*, (2022).

[19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari, 'The open images dataset V4', *Int. J. Comput. Vis.*, **128**(7), 1956–1981, (2020).

[20] Alon Lavie and Abhaya Agarwal, 'METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments', in *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL*, eds., Chris Callison-Burch, Philipp Koehn, Cameron S. Fordyce, and Christof Monz, pp. 228–231. Association for Computational Linguistics, (2007).

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao, 'Oscar: Object-semantics aligned pre-training for vision-language tasks', in *ECCV*, eds., Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, volume 12375 of *Lecture Notes in Computer Science*, pp. 121–137. Springer, (2020).

[22] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei, 'Pointing novel objects in image captioning', in *CVPR*, pp. 12497–12506, (2019).

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, 'Microsoft COCO: common objects in context', in *ECCV*, eds., David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, (2014).

[24] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, 'Neural baby talk', in *CVPR*, pp. 7219–7228, (2018).

[25] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L. Yuille, 'Learning like a child: Fast novel visual concept learning from sentence descriptions of images', in *ICCV*, pp. 2533–2541, (2015).

[26] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 'Deep contextualized word representations', in *NAACL-HLT*, pp. 2227–2237, (2018).

[27] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, 'Faster R-CNN: towards real-time object detection with region proposal networks', in *NIPS*, pp. 91–99, (2015).

[28] Mikihiro Tanaka and Tatsuya Harada, 'Captioning images with novel objects via online vocabulary expansion', *CoRR*, **abs/2003.03305**, (2020).

[29] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh, 'Cider: Consensus-based image description evaluation', in *CVPR*, pp. 4566–4575, (2015).

[30] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J. Mooney, Trevor Darrell, and Kate Saenko, 'Captioning images with diverse objects', in *CVPR*, pp. 1170–1178, (2017).

[31] Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama, 'NOC-REK: novel object captioning with retrieved vocabulary from external knowledge', in *CVPR*, pp. 17979–17987. IEEE, (2022).

[32] Yufei Wang, Ian D. Wood, Stephen Wan, and Mark Johnson, 'ECOL-R: encouraging copying in novel object captioning with reinforcement learning', in *EACL 2021, Online, April 19 - 23, 2021*, eds., Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, pp. 1222–1234. Association for Computational Linguistics, (2021).

[33] Yu Wu, Lu Jiang, and Yi Yang, 'Switchable novel object captioner', *IEEE Trans. Pattern Anal. Mach. Intell.*, **45**(1), 1162–1173, (2023).

[34] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang, 'Decoupled novel object captioner', in *2018 ACM Multimedia Conference on Multimedia Conference, MM*, eds., Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei, pp. 1029–1037. ACM, (2018).

[35] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, 'Incorporating copying mechanism in image captioning for learning novel objects', in *CVPR*, pp. 5263–5271, (2017).

[36] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao, 'Vinvl: Revisiting visual representations in vision-language models', in *CVPR*, pp. 5579–5588, (2021).

[37] Yu Zhang, Zhenghua Li, and Min Zhang, 'Efficient second-order treecrf for neural dependency parsing', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, eds., Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, pp. 3295–3305. Association for Computational Linguistics, (2020).