Beyond Generic: Enhancing Image Captioning with Real-World Knowledge using Vision-Language Pre-Training Model

Kanzhi Cheng
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
chengkz@smail.nju.edu.cn

Wenpo Song
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
songwp@smail.nju.edu.cn

Zheng Ma
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
maz@smail.nju.edu.cn

Wenhao Zhu
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
zhuwh@smail.nju.edu.cn

Zixuan Zhu
University of Glasgow
Glasgow, Scotland
zzx349313@gmail.com

Jianbing Zhang*
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
zjb@nju.edu.cn

ABSTRACT

Current captioning approaches tend to generate correct but "generic" descriptions that lack real-world knowledge, e.g., named entities and contextual information. Considering that Vision-Language Pre-Training (VLP) models master massive such knowledge from large-scale web-harvested data, it is promising to utilize the generalizability of VLP models to incorporate knowledge into image descriptions. However, using VLP models faces challenges: zero-shot inference suffers from knowledge hallucination that leads to low-quality descriptions, but the generic bias in downstream task fine-tuning hinders the VLP model from expressing knowledge. To address these concerns, we propose a simple yet effective method called Knowledge-guided Replay (K-Replay), which enables the retention of pre-training knowledge during fine-tuning. Our approach consists of two parts: (1) a knowledge prediction task on automatically collected replay exemplars to continuously awaken the VLP model's memory about knowledge, thus preventing the model from collapsing into the generic pattern; (2) a knowledge distillation constraint to improve the faithfulness of generated descriptions hence alleviating the knowledge hallucination. To evaluate knowledge-enhanced descriptions, we construct a novel captioning benchmark KnowCap, containing knowledge of landmarks, famous brands, special foods and movie characters. Experimental results show that our approach effectively incorporates knowledge into descriptions, outperforming strong VLP baseline by 20.9 points (78.7→99.6) in CIDEr score and 20.5 percentage points (34.0%→54.5%) in knowledge recognition accuracy. Our code and data is available at https://github.com/njucckevin/KnowCap.

CCS CONCEPTS

• **Information systems** → *Multimedia content creation.*

KEYWORDS

Image Captioning; Vision-Language Pre-Training; Knowledge

*Corresponding author.

ACM Reference Format:
Kanzhi Cheng, Wenpo Song, Zheng Ma, Wenhao Zhu, Zixuan Zhu, and Jianbing Zhang. 2023. Beyond Generic: Enhancing Image Captioning with Real-World Knowledge using Vision-Language Pre-Training Model. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3611987

1 INTRODUCTION

Image captioning aims to automatically describe the content of an image using natural language [3, 15, 57]. It has a variety of applications such as assisting visually impaired people and multi-modal content understanding for social media. Most existing approaches generate image descriptions in a "generic" manner [43, 69], i.e., describing only the common objects in an image, while lacking real-world knowledge such as named entities and contextual information. However, in many situations, such specific knowledge is the key to understanding the image. Taking Figure 1 as an example, the knowledge-enhanced description containing visual entity *Daisy Duck* and contextual knowledge *Disneyland* might evoke the memory of a wonderful journey. In contrast, the generic description generated by the advanced VLP+fine-tuning model seems tedious.

There have been several efforts attempt to incorporate knowledge into image descriptions [43, 55, 62]. Most of them follow a retrieve-and-generate methodology, first retrieving visual entities in images using external resources (e.g. entity recognition model or image metadata), and then adding the retrieved entities into descriptions. Moreover, they require the collection of additional caption data for supervised training. Such approaches are limited by the capacity of external resources and the availability of annotated caption data. By contrast, powerful VLP models learn from large-scale web-harvested data in an unsupervised manner, with the potential to master infinite real-world knowledge. In this paper, we pave a new way to leverage the VLP model's generalizability to recognize real-world knowledge and incorporate them into image description generation. Using VLP models has significant advantages over the aforementioned methods: (1) not limited by the capacity of external

Figure 1: Comparison of the result of VLP zero-shot, VLP+ fine-tuning and knowledge-enhanced description. We expect to properly integrate knowledge into descriptions while avoiding knowledge hallucination caused by pre-training noise. Knowledge words are marked by font colors and hallucination words are marked by blue font background.

resources; (2) no extra data collection for training; (3) no specific model architecture design.

Despite its potential, using the VLP model still faces some challenges. As depicted in Figure 1, the zero-shot inference of the VLP model yields low-quality description, and the noisy correspondence in pre-training image-text pairs causes the generation of harmful knowledge hallucinations. Therefore, fine-tuning the VLP model on a general captioning task is indispensable. However, we identified that fine-tuning on a captioning dataset with limited semantic diversity leads to the generic bias, which further markedly inhibits the VLP model's ability to express knowledge.

To tackle the above challenges, we propose the Knowledge-guided Replay (K-Replay), which guides the VLP model to retain pre-training knowledge during downstream task fine-tuning. We first filter a handful of images containing knowledge from the pre-training data as knowledge-related replay exemplars, then a sentence-level knowledge coverage loss is applied to evoke the VLP model's memory about knowledge thus preventing the model from collapsing into the generic pattern. In addition, we implement a knowledge distillation constraint using a fine-tuned VLP model to encourage the generation of faithful descriptions, thus alleviating the knowledge hallucination problem. Notably, K-Replay has a significant performance improvement in the replay unseen scenario, demonstrating that it is not only learning the knowledge from replay exemplars, but activating the VLP model to express the knowledge it has mastered during pre-training.

To evaluate the quality of the generated knowledge-enhanced descriptions, we constructed a new captioning benchmark, the KnowCap dataset. The dataset contains 1400+ images and 4100+ descriptions, covering 240 knowledge categories of landmarks, famous brands, special foods and movie characters. We extensively tested a series of representative captioning models on KnowCap. Our approach outperforms strong VLP baselines by a large margin in CIDEr score and knowledge recognition accuracy, establishing a new state-of-the-art on KnowCap.

The main contributions of this work are summarized as follows:

- We propose to exploit VLP model's generalizability for knowledge-enhanced image captioning, which is more efficient compared to previous retrieve-and-generate methods.
- We find that the generic bias in downstream task fine-tuning inhibits VLP models from expressing knowledge, thus designing the K-Replay to continuously evoke the model's memory about knowledge, while reducing knowledge hallucination through knowledge distillation constraint.
- We constructed the novel KnowCap dataset for evaluation, consisting of more than 1400 images containing various types of knowledge.
- Experimental results show that our approach can effectively retain the knowledge mastered by the pre-trained model during downstream task fine-tuning, finally outperforming a series of strong baselines on the KnowCap dataset.

2 RELATED WORK

2.1 Vision-Language Pre-Training

Pre-training technique has revolutionized the NLP and CV research community in recent years [7, 18, 22, 48]. Meanwhile, Vision-Language Pre-Training has been shown to significantly improve performance on a wide range of uni-modal and multi-modal tasks [30, 31, 47, 60, 61, 68, 70]. Recently, several studies show that the Large Language Models (LLMs) can store and predict knowledge about the world [20, 26, 41, 46]. However, the real-world knowledge contained in VLP is largely overlooked by existing research, with only [10] exploring using a knowledge base to assist pre-training. In this paper, we focus on generating knowledge-enhanced image descriptions with the knowledge of VLP.

2.2 knowledge-enhanced Image Captioning

Previous image captioning systems adopted the encoder-decoder approach and achieved success through carefully designed model architectures [3, 15, 28, 57, 66] and training methods [5, 13, 38, 50]. Benefiting from pre-training techniques, the pre-training and fine-tuning paradigm [30, 32, 61, 68, 70] further takes the model capability to another level. Among them, autoregressive generative models [14, 59, 60] achieved remarkable performance through the simple sequence-to-sequence learning framework.

However, [38, 58, 62, 69] revealed that existing captioning approaches suffer from the over-generic problem, which limits the informativeness of descriptions. A line of effort attempted to incorporate knowledge into descriptions to alleviate this shortcoming. These works adopt a retrieve-and-generate methodology, first retrieving entities contained in images using additional visual entity recognition model [55, 69] or image metadata [43, 62], followed by supervised training to integrate these entities into the decoding process [43, 54, 69] or template caption [6, 37]. These methods are limited by the ability of external resources and the difficulty of labeling data. This paper solves these difficulties through VLP model's generalizability. Recently, Universal Captioner [14] and GIT [59] have shown that VLP models have certain generalizability to in-the-wild entities but lack quantitative analysis. In this paper, we construct a benchmark for knowledge-enhanced image captioning task and demonstrate the superiority of our approach.

2.3 Catastrophic Forgetting

We argue that the difficulty of expressing knowledge in VLP models is associated with catastrophic forgetting [21, 40] when fine-tuning

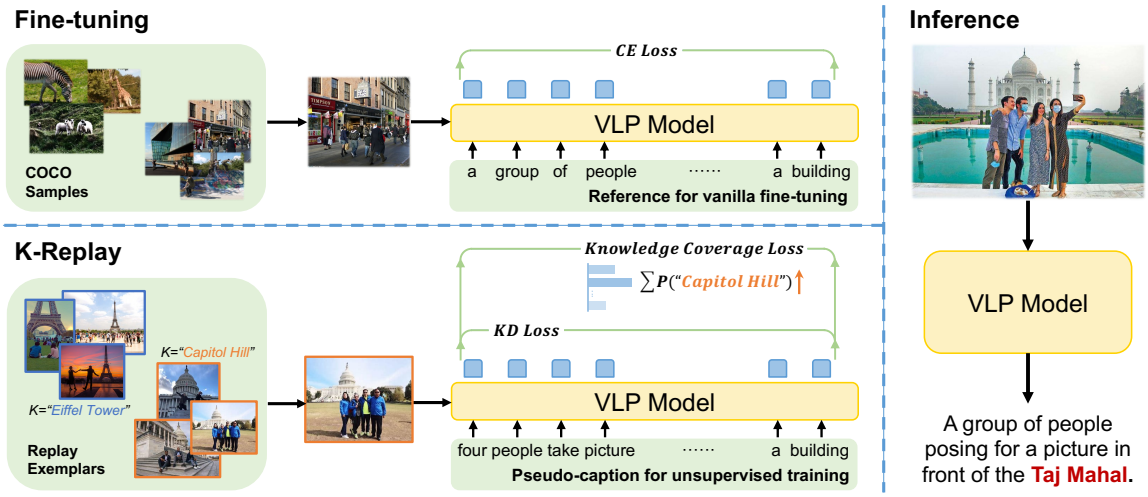

Figure 2: Illustration of our K-Replay method. K-Replay is performed simultaneously with downstream task fine-tuning to activate the model's memory about knowledge. Notice that the knowledge generated in the inference stage (Taj Mahal) does not need to appear in the replay exemplars.

downstream tasks. Methods designed to alleviate catastrophic forgetting fall into three categories [42]: *regularization*, *replay*, and *dynamic architecture*. Regularization methods mitigate the forgetting of prior knowledge by limiting the change in model parameters, using regular terms [2, 11, 27] or knowledge distillation [8, 63]. Replay methods store exemplars from the past task and continuously review them while learning new tasks [8, 42, 49]. Dynamic architecture methods use different modules to learn different tasks and thus avoid forgetting [17, 39, 51]. Other efforts seek an alternative strategy to alleviate the forgetting of pre-trained models by fine-tuning part of the parameters to enhance generalizability , such as adapter [19, 24], prompt tuning [29, 33] and child tuning [65]. In Section 6.2 we systematically compare the effects of these methods.

3 APPROACH

3.1 Overview

Visual-Language Pre-Training allows the VLP model to learn a large amount of knowledge from web-scale data. However, knowledge hallucination and the generic bias in downstream task fine-tuning invisibly inhibit the expression of such knowledge. To address these concerns, we propose K-Replay, which continuously stimulates the model to express knowledge while downstream task fine-tuning, and reduces the hallucination through knowledge distillation constrain. Note that K-Replay does not introduce additional model design, but rather seeks to guide the behavior of the VLP model through learning tasks, thus enabling the retention of pre-trained knowledge during downstream task fine-tuning.

Given a VLP model P and a downstream captioning dataset $D_C = \{(x_i, y_i)|_i\}$, with the i-th image x_i along with its corresponding caption y_i, vanilla fine-tuning trains P on dataset D_C to obtain an image captioning model P_C. In addition, we automatically filter a small portion of knowledge-related samples using knowledge keywords from the pre-training data to constitute the replay exemplars set $D_K = \{(x_i, k_i)|_i\}$, with the i-th image x_i along with its knowledge keyword k_i (details in Section 3.3). Next, our approach uses D_C and D_K to train the VLP model P for generating knowledge-enhanced image descriptions.

The overview of proposed K-Replay method is presented in Figure 2. It is implemented in three main specially designed parts: (1) a knowledge prediction task on the replay exemplar to awaken the knowledge of the VLP model (in Section 3.3); (2) a knowledge distillation constraint to alleviate the knowledge hallucination caused by pre-training noise (in Section 3.4); and (3) simultaneous training on D_C and D_K achieved by constructing pseudo-caption data (in Section 3.5).

3.2 Vanilla Fine-tuning

We start by introducing the process of vanilla fine-tuning the VLP model P to the downstream captioning task. Formally, given the captioning dataset $D_C = \{(x_i, y_i)|_i\}$, where $y = \{w_1, w_2, ...w_T\}$ represents the word sequence of the reference description of length T, by omitting the sample subscript i, as default in the later. Then, we model the image captioning task as an end-to-end sequence generation paradigm, by maximizing the probability of image-caption pairs (x, y) in D_C:

$$\arg\max_{\theta} E_{(x,y)\sim D_C} \prod_{t=1}^{T} p(w_t|w_{\tau<t}, x; \theta), \tag{1}$$

where θ is the parameters of VLP model P. Finally, the loss of vanilla fine-tuning is the cross-entropy between generated caption and groundtruth:

$$\mathcal{L}_{ce} = -\frac{1}{T}\sum_{t=1}^{T} log\, p(w_t|w_{\tau<t}, x; P). \tag{2}$$

3.3 Knowledge Prediction Task

We aim to preserve the knowledge learned by the VLP model during pre-training while downstream task fine-tuning. A straightforward