Enhanced Model for Real World Knowledge Image Captioning

Reem AlJunaid
Student ID: g202102170
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Supervised by: Dr. Muzammil Behzad
muzammil.behzad@kfupm.edu.sa
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

***Abstract*—Many image captioning models produce generic captions that lack specificity and fail to capture knowledge-rich concepts. This research enhances the KnowCap framework to address this issue by introducing three key modifications: beam search for more diverse decoding, attention-based modules in the image encoder for better feature representation, and schedulers to stabilize training. These enhancements significantly improved both caption quality and recognition accuracy. On the KnowCap dataset, recognition accuracy increased from 50.40% to 63.30%, with a slight improvement in captioning metrics such as CIDEr (+2.3). The model also demonstrated stronger generalization, achieving 57.69% recognition accuracy on unseen knowledge categories, compared to the baseline of 45.80%.**

***Index Terms*—Image Captioning, Vision-Language Models, Knowledge Recognition**

I. Introduction

A. Background and Significance

Image captioning is the task of generating descriptions of images using computer vision and natural language processing techniques [1]–[3]. It has a variety of applications, such as enhancing the content understanding of multimedia and helping visually impaired people. Current existing models, however, often produce generic captions which lack real-world concepts, such as contextual details and named entities [4], [5]. This missing knowledge often represents key information in understanding the content of the image. Further, this detailed knowledge can also enhance the performance of other models that rely on the output of image captioning systems, such as question answering systems.

Several efforts have 1een placed in order to enhance the descriptions with real-world knowledge [4], [6], [7]. However, they were limited by using external resources, such as image metadata or object recognition models, to detect existing entities before generating the descriptions. The Vision-Language Pretrained (VLP) models present a powerful solution for this era. They were trained on massive data and can capture diverse real-world knowledge. Yet, VLP models suffer from two main problems: (1) zero-shot inference leading to safe but low-quality descriptions and (2) knowledge hallucination due to the noise in image-text pairs in pre-training. Additionally, fine-tuning VLP models on downstream tasks introduces a ”generic bias” that restricts their expression of detailed knowledge.

Addressing these limitations, the Knowledge guided Replay (K-Replay) framework is proposed by [8]. K-Replay preserves the original model structure while helping VLP models retain knowledge during fine-tuning on downstream tasks. It is done by selecting knowledge-rich samples from the pretraining data and computing a coverage loss based on the existence of required keywords in the sentence to reinforce memory of this knowledge and hence avoid generic descriptions. Moreover, to reduce hallucination and ensure faithful descriptions, a constraint based on knowledge distillation is applied from a model that has already been fine-tuned. K-Replay showed strong performance, especially in unseen scenarios, proving it effectively helps the model to recall and express pre-learned knowledge.
In this work, we enhance the K-Replay framework [8] by replacing greedy decoding with beam search to generate higher-quality pseudo-captions during replay. Second, we introduce learning rate schedulers to stabilize training and promote smoother convergence. Finally, we integrate attention layers into the model architecture to improve its ability to focus on relevant image regions and contextual cues. These enhancements help improve the model's capacity to produce detailed and accurate knowledge-based captions.

B. Problem Statement

Despite significant progress in image captioning, existing models often produce generic captions which overlook real-world concepts, such as contextual information and named entities. While VLP models have shown promise in capturing this knowledge through training on large-scale datasets [9]–[14], they suffer from zero-shot

inference and knowledge hallucination [8]. Furthermore, finetuning VLP models in downstream tasks suffers from generic bias, which inhibits their knowledge expression [8]. The K-Replay framework was introduced to address these challenges by helping the model preserve previously learned information during fine-tuning for image captioning [8]. Nonetheless, the effectiveness and stability of this method could still be further improved.

C. Objectives

The main objective of this research is to enhance image captioning models by improving their ability to produce knowledge-rich and accurate descriptions. Specifically, we plan to:

1) Mitigate the issue of knowledge hallucination in fine-tuned VLP models through an improved fine-tuning approach.
2) Enhance the K-Replay framework by incorporating beam search for higher-quality pseudo-captions.
3) Improve training stability and convergence by introducing learning rate schedulers.
4) Integrate attention layers to enhance the model's focus on relevant image areas and contextual cues.
5) Evaluate the performance of these enhancements with experiments on the KnowCap and COCO datasets.

D. Scope of Study

This study focuses on enhancing VLP models to produce knowledge-rich and accurate image captions. The evaluation is conducted using the KnowCap dataset, and as such, the scope of knowledge assessed is limited to the types of real-world knowledge represented by the predefined keyword categories in KnowCap. These include four categories, foods, brands, landmarks, and movie characters. The study does not attempt to capture or evaluate all forms of general knowledge but rather focuses on the ability of the model to incorporate in the caption the specific knowledge keywords defined within this benchmark.

II. Literature Review

A. Overview of Existing Techniques

Image captioning has evolved through various techniques, starting with traditional encoder-decoder architecture to transformer-based architecture [15]. Each architecture has its advantages and disadvantages. In traditional techniques, images are encoded with convolutional neural networks (CNNs) and texts decoded with recurrent neural networks (RNNs), converting visual features into language linearly. Even though these models are effective, they often fail to capture extended dependencies within captions and the complexity of visual scenes. Transformer-based frameworks, on the other hand, offer attention mechanisms for identifying semantic relations within visual scenes rather than considering them separately. Using self-attention, transformers create accurate, knowledge-rich captions that capture detailed scenes despite lengthy or complex descriptions. Compared to traditional approaches, these transformer-based models provide superior accuracy and knowledge-rich captions.

Traditional encoder-decoder architecture plays a key role in the development of image captioning models. Vinyals et al. [3] pioneered the use of a deep learning encoder-decoder architecture for image captioning. Their approach involved using pre-trained CNNs for image encoding, with the last hidden layer output passed into an RNN to generate descriptive captions. Their work provided a baseline for subsequent studies, which have expanded and refined the model by incorporating various enhancements, including the visual attention mechanism introduced by [16]. Another popular framework that serves as a baseline for many image captioning models is the Bottom-Up and Top-Down framework [1]. By leveraging Faster R-CNN, it combines two types of attention: Bottom-Up, which identifies important parts of the image using feature vectors, while Top-Down decides which parts to focus on while generating the next word of the caption. Huang et al. [17] developed the attention-on-attention (AoA) framework, which improves traditional attention by adding another layer to help the model choose more relevant words based on context. Another encoder-decoder-based framework is the Deep Hierarchical Encoder-Decoder Network (DHEDN). DHEDN comprises a three-layer LSTM architecture: S-LSTM handles text encoding, VSE-LSTM combines visual and textual features into a common semantic space, and SF-LSTM is responsible for generating captions.

Recent research in image captioning has shifted towards transformer-based architecture. The self-attention in deep learning models enables simultaneous processing of images and captions, thus helping generate more accurate and knowledge-rich descriptions. Unlike RNNs, transformers process sequences in parallel, making training much faster and more scalable. Vaswani et al. [18] introduced self-attention for efficient parallel processing, replacing recurrence entirely. It treats image features as input tokens and uses standard encoder-decoder layers for generation. However, the model lacks visual enhancements. Later, Cornia et al. presented the M2 meshed-memory transformer, which extends this by introducing memory vectors to encode multi-level relationships among image regions. The encoder analyzes these image regions and their connections, while the decoder generates captions dynamically.

Building on advances in NLP, the BERT architecture employs a masked language modeling objective to enable attention in both directions. This design choice has

substantially boosted its ability to understand context. Although BERT was originally developed to perform text-based tasks, its architecture has inspired its use in multimodal applications such as image captioning. Researchers have developed hybrid models combining BERT-like components with visual encoders to link visual and textual features. These models use pre-trained language representations to enhance the connection between images and captions [19], [20].

Recent advancements in image captioning have been driven by Vision-Language Pre-training (VLP) models. The model is initially trained on a large-scale dataset using self-supervised learning and later adapted to perform a downstream task. One of the most utilized pre-trained models is Contrastive Language-Image Pre-Training (CLIP) by Radford et al. [21]. CLIP has been trained using a contrastive loss on a vast dataset of image-caption pairs. Building upon CLIP, Mokady et al. [22] proposed ClipCap, a method that maps CLIP image embeddings into a prefix used to condition a pre-trained language model (GPT-2) for caption generation. The method demonstrated the effectiveness of combining visual and linguistic pre-trained models. Another influential VLP model is OSCAR (Object-Semantics Aligned Pretraining for Vision-and-Language Tasks) introduced by Li et al. [23]. OSCAR improves cross-modal representation learning by injecting object tags as anchor points during pretraining. This approach enhances the alignment between image regions and descriptions, thus enhancing the image captioning performance. Recently, several VLP models, such as BLIP [10], GIT [24], and OFA [12], have been trained for image captioning tasks and have outperformed existing methods across multiple benchmarks.

Despite advances in image captioning, existing models still struggle to generate knowledge-rich and context-aware descriptions. Transformer and VLP-based models often produce generic captions due to fine-tuning on limited datasets, which can override pre-trained knowledge. Additionally, many methods rely on external tools to inject knowledge, making them less scalable. Hallucination and forgetting of real-world knowledge remain key challenges, especially in zero-shot scenarios, highlighting the need for better knowledge retention and expression during fine-tuning.

B. Related Work

Although image captioning has improved fluency and grammaticality, many studies have shown that the generated captions remain generic and lack real-world knowledge. To overcome this limitation, several studies have attempted to enrich captions with knowledge through a retrieve-and-generate framework using external resources. For example, entity-aware models integrate named entities into the captioning pipeline, either through predefined templates [25] or by modifying decoder architectures [26], [27]. Other methods leverage visual recognition outputs [6], [5] or contextual metadata such as geolocation [4] to guide caption generation. However, these methods require a lot of extensive supervision and annotated datasets.

Other studies fine-tuned VLP models for image captioning; however, this approach, on the contrary, suffers from generic bias, which limits its informativeness. In contrast, this paper leverages a VLP model by fine-tuning it in a controlled manner to mitigate generic bias and hallucinations, while enhancing informativeness.

C. Limitations in Existing Approaches

Highlight shortcomings and justify the need for enhancement.

III. Proposed Methodology

This section describes the proposed improvements to enhance the existing model. It outlines the challenges and limitations of the current approach, the key enhancements made to address these challenges, and provides an overview of the algorithm, data processing techniques, as well as the loss functions and optimization strategy used to improve performance.

A. Existing Model and Challenges

The baseline model used in this work is the K-Replay framework [8], which was originally proposed to address the issue of knowledge retention during the fine-tuning of VLP models for image captioning. While this framework showed promise in preserving real-world knowledge during training, it has notable limitations. One major issue lies in its method of generating pseudo-captions for the replay samples. The original implementation uses greedy decoding, which often leads to low-diversity and generic outputs. These captions may lack detail and fail to fully represent the knowledge concepts intended for replay, thereby limiting the effectiveness of the knowledge-guided learning process. Another challenge observed with the baseline model is the instability of the training process. The K-Replay framework combines multiple loss terms, a cross-entropy loss, a replay loss, and a distillation loss, which can lead to fluctuating training curves and difficulty in convergence. This lack of smooth optimization raises the need for mechanisms that can balance learning dynamics. Additionally, while the original OFA architecture includes attention mechanisms across both visual and textual modalities, it does not explicitly model self-attention among image patches prior to fusion with text. This may limit the model's ability to capture detailed visual patterns that are important for understanding real-world knowledge. These limitations highlight the need for further refinement.

B. Proposed Enhancements

To address the limitations in the original K-Replay framework [8], we introduce three key improvements aimed at boosting the quality of pseudo-captions, improving training stability, and helping the model better understand visual content. First, we replace the greedy decoding strategy used for generating pseudo-captions with beam search (beam size = 5). This allows the model to explore multiple caption possibilities and choose the most informative one, resulting in higher-quality replay samples during training. Second, we apply learning rate schedulers, specifically using cosine annealing, to control the pace of learning. Since the training involves multiple loss components, the loss curves were previously unstable. By dynamically adjusting the learning dynamically over time, the training process becomes more stable and converges more smoothly. This prevents drastic updates and helps the model avoid local minima. This contributes further to better final performance and more efficient training. Lastly, we enhance the encoder by adding a self-attention layer over the image patch embeddings before they are fused with text. In the original framework, the encoder processed image patches without explicitly modeling the relationships between different regions of the image. By incorporating self-attention, the model can attend to relevant areas of the image based on their semantic importance, helping it focus on detailed visual patterns. These enhancements collectively can help generate accurate, knowledge-rich captions.

C. Algorithm and Implementation

The **Knowledge-guided Replay (K-Replay)** method is a continual learning strategy designed to prevent catastrophic forgetting. It works by selectively replaying knowledge from previous tasks while learning new ones. K-Replay maintains a buffer of knowledge-rich examples from earlier tasks and reuses them during training on new tasks to ensure the model retains essential prior knowledge.

In our setup, we selected a pretrained VLP model, **OFA-Large**, which was pretrained on multi-modal tasks and datasets, including CC12M. The model we aim to train is denoted as M_θ, while a fine-tuned version, ofa-large-caption, is used as a teacher model, denoted M_{ref}. Training of the model M_θ was accomplished using both **CC12M** and **COCO** datasets. The COCO samples contain image-caption pairs $\mathcal{S}_c$, and the CC12M samples contain image-keyword pairs $\mathcal{S}_k$, which serve as replay samples.

During each training iteration, the algorithm operates on a mini-batch containing samples from both the current task ($\mathcal{S}_c$) and the replay buffer containing past knowledge-related samples ($\mathcal{S}_k$). If the current batch contains an image from $\mathcal{S}_c$, it is passed through M_θ to generate a caption. The generated caption is then compared with the ground truth caption, and the cross-entropy loss $\mathcal{L}_{txt}$ is computed by comparing the logits of the predicted caption with the true caption.

However, if the current batch contains an image from $\mathcal{S}_k$, the teacher model M_{ref} is first used to generate a pseudo-caption using beam search. The image and the pseudo-caption are then fed into both M_{ref} and M_θ to compute their logits. These logits are used to calculate the knowledge distillation loss $\mathcal{L}_{distill}$. Additionally, the presence of the original keyword in the generated caption is checked to calculate the knowledge prediction loss $\mathcal{L}_{kpred}$. The overall loss is the total of the cross-entropy loss $\mathcal{L}_{txt}$, the knowledge distillation loss $\mathcal{L}_{distill}$, and the knowledge prediction loss $\mathcal{L}_{kpred}$. The hyperparameters λ_k and λ_d are employed to control the balance between the different losses. The step-by-step procedure of this training loop is outlined in Algorithm 1, which provides a pseudocode representation of the K-Replay training mechanism used in our setup.

Algorithm 1 Enhanced Knowledge-guided Replay Training with Attention and Beam Search

Input: Training mini-batch $b \in \mathcal{S}_c \cup \mathcal{S}_k$,
$b = \{(i_u, t_u)\}_u \cup \{(i_v, k_v)\}_v$; enhanced captioning model M_θ;
reference model M_{ref} (frozen); loss weights λ_k, λ_d
Output: Updated model parameters θ'

for each sample $[i, t, (k)]$ in b **do**
  if $i \in \mathcal{S}_c$ **then**
    apply self-attention on image patches in M_θ
    $z \leftarrow M_\theta(i, t)$
    $\mathcal{L}_{txt} \leftarrow \text{CrossEntropy}(z, t)$
  end if
  if $i \in \mathcal{S}_k$ **then**
    $\hat{t} \leftarrow \text{BeamDecode}(M_\theta(i), b{=}5)$
    $\tilde{z} \leftarrow M_{ref}(i, \hat{t})$
    $z \leftarrow M_\theta(i, \hat{t})$
    $\mathcal{L}_{kpred} \leftarrow \text{MSE}(z, k)$
    $\mathcal{L}_{distill} \leftarrow \text{KL}(z, \tilde{z})$
  end if
end for
$\mathcal{L}_{total} \leftarrow \mathcal{L}_{txt} + \lambda_k \cdot \mathcal{L}_{kpred} + \lambda_d \cdot \mathcal{L}_{distill}$
update θ using AdamW optimizer and learning rate scheduler

D. Loss Function and Optimization

The proposed framework combines multiple objectives into a unified multitask learning setup. The total loss function comprises three key components:

1. **Cross-Entropy Loss:**
Cross-entropy loss is used during the fine-tuning of the

M_θ model for the image captioning task for the images from the COCO dataset. It measures the difference between the generated caption and the ground truth captions, with the goal of maximizing the likelihood of generating the correct caption t given an input image i under the current model parameters θ.

$$\mathcal{L}_{txt} = -\frac{1}{T}\sum_{t=1}^{T}\log p(w_t \mid w_{<t}, i; M_\theta). \quad (1)$$

2. **Knowledge Prediction Loss:**
Knowledge Prediction Loss is introduced to encourage the model to incorporate knowledge-related tokens during caption generation. It is computed over the generated pseudo-captions for images from the CC12M dataset by prompting the M_θ model to identify named entities present in the image. Given an image-keyword pair (i, k), where the keyword k is tokenized into BPE subwords $\{w_1^k, w_2^k, ..., w_N^k\}$, the objective is to ensure that these tokens are reflected in the generated caption $\hat{t} = M_\theta(i)$.

$$\mathcal{L}_{cov} = -\sum_{i=1}^{N}\log\sigma(p(w_i^k)), \quad (2)$$

To further prevent the model from excessively repeating the keywords, a repetition penalty is added to the loss.

$$\mathcal{L}_{rep} = \sum_{i=1}^{N}\left[1 - p(w_i^k)\right]^2. \quad (3)$$

The final knowledge prediction loss combines both components, knowledge coverage and repetition penalty.

$$\mathcal{L}_{know} = \mathcal{L}_{cov} + \mathcal{L}_{rep}. \quad (4)$$

3. **Knowledge Distillation Loss:**
The Knowledge Distillation Loss helps the updated model keep the pre-trained network's knowledge and generalization abilities as it adapts to captioning. This loss applies the Kullback–Leibler (KL) divergence between the output logits of the teacher model (pre-trained M_{ref}) and the student model (the M_θ model undergoing fine-tuning). It quantifies how much the student's output distribution deviates from that of the teacher. It acts as a regularization term to ensure the fine-tuning does not lead to catastrophic forgetting of the original model's learned representations.

$$\mathcal{L}_{distill} = D_{kl}[\varphi(z_t), \varphi(z_s)], \varphi(z_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (5)$$

IV. Experimental Design and Evaluation

This section presents datasets, evaluation metrics, training setup, and results. It includes comparisons with baselines and an ablation study to assess model components.

A. Datasets and Preprocessing

This study employs three datasets, as follows:

- **MS-COCO** [28]: This image captioning dataset follows the Karpathy split. It contains 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. Each image has five human-written captions.
- **Replay CC12M Subset** [29]: A curated subset of more than 20,000 samples which is extracted from CC12M by filtering image-text pairs that mention any of 122 predefined keywords. These samples are used as replay exemplars during training.
- **KnowCap** [8]: This dataset enhances captioning with real-world knowledge. It contains 1,424 image-caption pairs with 240 knowledge categories. A total of 424 samples are used for validation, and 1,000 samples are used for testing. Among the test set, 520 samples with 120 knowledge categories include concepts that do not appear in the predefined keyword list. This subset, which we refer to as the unseen set, is used to evaluate the model's generalization to new knowledge.

As part of preprocessing, all datasets are first formatted into the standard pycocoevalcap format. For training, we randomly selected 27,000 image-caption pairs from the COCO training set and 5,000 image-keyword pairs from the Replay CC12M subset. These were then combined and shuffled to form the 32,000-sample training set. The models were evaluated during training using both the COCO and KnowCap validation sets. For testing, we use the COCO test set, the full KnowCap test set, and the unseen subset from KnowCap.

B. Performance Metrics

The evaluation of the generated captions was carried out using these metrics:

- **BLEU** [30]: measures the precision of n-grams in the generated caption that appear in the ground truth captions.
- **METEOR** [31]: is similar to BLEU, but it incorporates semantic similarity by considering stemming, synonyms, and exact matches rather than just exact word matches.
- **ROUGE** [32]: primarily evaluates recall by comparing the overlap of n-grams, longest common subsequences (LCS), or word sequences between the generated and ground truth captions.
- **CIDEr** [33]: scores the generated captions by comparing the TF-IDF weighted n-gram similarity with a set of human-generated reference captions.
- **Knowledge Recognition Accuracy**: used to evaluate the inclusion of real-world concepts in generated captions. It measures the percentage of captions that

correctly include valid knowledge-related keywords from the KnowCap dataset.

C. Experiment Setup

All experiments were conducted using Google Colab with an NVIDIA A100 GPU. We used the official checkpoint of the PyTorch OFA-Large model with a batch size of 8, trained for 10 epochs using a learning rate of 7e-6 and label smoothing of 0.1. Knowledge distillation was performed using a temperature $T = 16$, and the loss weights for the distillation loss and the knowledge prediction loss were both set to 1.0. We report results for the best-performing model checkpoint based on validation performance on the KnowCap dataset.

D. Comparative Analysis

We evaluate our model on the KnowCap and MSCOCO datasets, placing primary emphasis on the CIDEr score (C) and Recognition Accuracy (Rec), which best capture the quality and real-world grounding of generated captions. The results are presented in Table I. In the zero-shot setting, the base OFA model performs poorly on both datasets, achieving a CIDEr of 39.2 on KnowCap and only 22.1 on MSCOCO. Its recognition accuracy is also limited, with just 39.8% on KnowCap. These results highlight the model's inability to generalize to real-world concept captioning without fine-tuning.

Once fine-tuned (+OFA-Finetuned), the model shows substantial improvements in general captioning metrics. On KnowCap, the CIDEr score increases from 39.2 to 41.7, and MSCOCO sees a large jump from 22.1 to 134.6. However, this gain comes at a cost: Recognition Accuracy on KnowCap slightly drops from 39.8% to 38.5%, despite the overall improvement in fluency and structure of captions. This indicates signs of catastrophic forgetting, where the model becomes less effective at grounding captions in real-world concepts due to overfitting to MSCOCO-style language patterns during fine-tuning. It highlights the necessity of mechanisms like K-Replay to retain domain-specific knowledge without compromising language generation quality.

Incorporating K-Replay yields significant gains on both datasets. CIDEr on KnowCap more than doubles to 90.3, and the recognition accuracy increases to 50.4%, demonstrating that the model now retains better knowledge of real-world concepts. On MSCOCO, the CIDEr score remains strong at 130.6, showing that knowledge retention does not come at the cost of general captioning performance.

The use of a learning rate scheduler (+Scheduler) brings moderate improvements in CIDEr (92.6) and a noticeable increase in the recognition accuracy (54.2%) on KnowCap. This suggests that a more stable optimization process can positively influence both caption fluency and concept recognition.

Introducing beam search decoding (+Beam) significantly boosts the recognition accuracy to 63.3%, the highest recorded on KnowCap, although the CIDEr score slightly drops to 92.6. This indicates a trade-off between linguistic diversity and precision. Nevertheless, beam search enhances the model's ability to recover conceptually meaningful captions.

Lastly, augmenting the model with additional attention layers (+Attention) offers a balanced improvement, achieving 92.0 CIDEr and 58.9% recognition accuracy on KnowCap.

Figure 1 showcases the improvements made by the Enhanced K-Replay model over the original K-Replay captions. Each image is presented with the captions generated by the model, both before and after the enhancement. The Enhanced K-Replay model significantly improves upon the original K-Replay captions by offering more detailed descriptions, incorporating real-world concepts, which are highlighted in green for emphasis.

E. Generalization Results

In addition to the improvements on the KnowCap dataset, we assess the model on unseen KnowCap categories, a portion of the test set containing 120 categories and 520 images that were not present in the replay data. This enables us to evaluate the model's capacity to adapt to new, unseen scenarios. The results are illustrated in Table II. Starting with the OFA zero-shot baseline, we observe CIDEr of 36.9 and Recognition Accuracy (Rec) of 39.2%, which serves as the starting point for our evaluation. After fine-tuning with OFA-Finetuned, CIDEr increases to 42.0, and the recognition accuracy remains stable at 39.0%, confirming the benefits of fine-tuning on this dataset.

When K-Replay is added to the fine-tuned model, we see a remarkable jump in CIDEr to 83.9, and a noticeable increase in Recognition Accuracy to 45.8%. This result demonstrates that K-Replay is not only helping the model retain learned knowledge but also significantly improving its performance on the unseen categories. Introducing the learning rate scheduler (+Scheduler) boosts CIDEr further to 91.0 and the recognition accuracy to 55.6%, showing that adjusting the learning rate aids in better fine-tuning and model convergence.

However, after adding beam search (+Beam), we see a slight drop in CIDEr, which decreases to 82.6, but the recognition accuracy increases to 57.7%. This suggests that although beam search improves generation quality, it helps the the recognition accuracy even further, leading to higher recognition accuracy, even though the CIDEr metric slightly decreases. Finally, adding attention layers (+Attention) leads to a slight improvement in CIDEr, bringing it to 84.4, but Recognition Accuracy (Rec) decreases slightly to 55.0%. The attention layers allow

TABLE I: Comparison of Different Techniques on KnowCap and COCO Datasets

Dataset	KnowCap								COCO						
Technique	**B1**	**B2**	**B3**	**B4**	**M**	**R**	**C**	**Rec**	**B1**	**B2**	**B3**	**B4**	**M**	**R**	**C**
OFA zero-shot	32.8	20.0	13.5	9.4	11.5	24.6	39.2	39.80%	24.5	14.6	9.3	6.1	10.3	21.0	22.1
OFA-Finetuned	35.6	22.0	15.2	10.8	12.1	26.6	41.7	38.50%	79.6	64.7	50.9	39.8	30.4	59.7	134.6
+K-Replay	58.8	41.5	30.5	22.7	20.2	43.0	90.3	50.40%	77.3	62.0	48.6	37.8	30.6	59.2	130.6
+Scheduler	58.6	42.3	31.7	23.9	20.9	44.1	92.6	54.20%	77.8	62.4	48.7	37.7	30.4	59.1	130.4
+Beam	57.7	40.7	29.6	21.9	20.4	42.5	92.6	63.30%	77.2	62.1	48.7	37.9	30.7	59.2	130.8
+Attention	58.1	41.1	30.1	22.1	20.5	43.2	92.0	58.90%	77.0	61.7	48.3	37.6	30.6	59.0	129.4

the model to focus on more relevant visual features, which enhances the overall concept grounding, though the recognition accuracy slightly drops compared to the beam search configuration.

TABLE II: Performance on unseen categories of KnowCap.

Technique	**B1**	**B2**	**B3**	**B4**	**M**	**R**	**C**	**Rec**
OFA zero-shot	31.7	19.5	13.5	9.8	11.3	24.1	36.9	39.20%
OFA-Finetuned	35.0	22.2	15.7	11.5	12.3	26.5	42.0	39.00%
+K-Replay	57.9	40.8	30.5	23.2	19.6	42.6	83.9	45.80%
+Scheduler	58.7	42.4	31.9	24.3	20.8	44.2	91.0	55.60%
+Beam	56.2	39.5	28.9	21.7	19.5	41.8	82.6	57.69%
+Attention	57.5	40.8	30.1	22.5	20.1	43.0	84.4	55.00%

F. Comparison with Prior Techniques

Catastrophic forgetting remains a major challenge in the fine-tuning of Vision-Language Pretraining (VLP) models, as the process of adapting to downstream tasks like image captioning can result in the loss of pre-trained knowledge. In this section, we compare our proposed methods with several previous approaches aimed at mitigating catastrophic forgetting, particularly in the context of concept-aware image captioning.

- **EWC** [34] introduces a regularization term that penalizes significant changes to parameters important for the original task, aiming to preserve prior knowledge during adaptation.
- **Recall and Learn** [35] employs regularization weights to help the model retain previously learned knowledge.
- **Child-Tuning** [36] focuses on fine-tuning only a specific set of parameters that are considered essential for the downstream task. This approach helps enhance the model's generalization and minimizes the risk of forgetting.
- **Adapter** [37] uses lightweight bottleneck modules to inject task-specific parameters while leaving the original pre-trained model largely intact, thus preventing overfitting and catastrophic forgetting.

Table III provides a quantitative comparison of these methods on the KnowCap dataset. While all methods introduce strategies to alleviate forgetting, their effectiveness varies considerably. Notably, even strong baselines like EWC and Child-Tuning struggle to maintain both high recognition accuracy (Rec) and CIDEr (C) scores simultaneously. By contrast, K-Replay shows substantial improvements, especially in recognition-focused metrics. Our enhanced variant further pushes these gains, achieving the highest recognition accuracy and maintaining competitive scores across all metrics. This highlights the robustness of our approach in preserving concept-awareness during fine-tuning, outperforming previous techniques without compromising language quality.

TABLE III: Comparison of Different Techniques on the KnowCap Dataset

Technique	**B1**	**B4**	**M**	**R**	**C**	**Rec**
EWC [34]	56.9	21.8	19.1	42.0	73.6	30.40%
Recall and Learn [35]	53.7	20.1	18.1	39.3	70.6	37.20%
Child-Tuning [36]	55.8	21.7	18.8	41.5	74.7	33.80%
Adapter [37]	54.4	20.5	17.6	40.1	63.7	30.40%
Vanilla Fine-tuning [8]	35.6	10.8	12.1	26.6	41.7	38.50%
K-Replay [8]	58.8	22.7	20.2	43.0	90.3	50.40%
+Scheduler	58.6	23.9	20.9	44.1	92.6	54.20%
+Beam	57.7	21.9	20.4	42.5	92.6	63.30%
+Attention	58.1	22.1	20.5	43.2	92.0	58.90%

V. Extended Contributions

This research enhances the KnowCap framework by incorporating beam search decoding, attention-based mechanisms, and training schedulers, resulting in improved caption quality and greater accuracy in knowledge recognition within images. The findings demonstrate significant improvements in model performance, especially in its ability to generalize to unseen knowledge categories. These advancements have the potential to make a significant impact in a wide range of areas including robotics, autonomous systems, medical imaging and diagnostics, and assistive technologies, where precise and context-aware image descriptions are essential.

VI. Conclusion and Future Work

This research proposed a set of enhancements to the KnowCap framework by replacing greedy decoding with beam search, integrating attention-based modules within the image encoder, and employing schedulers to stabilize training and promote smoother convergence.

K-Replay: A close up of a fast food meal in a box

Enhanced K-Replay: A mcdonalds breakfast sandwich with scrambled eggs and hash browns

K-Replay: A close up of a person cutting up a blackberry

Enhanced K-Replay: A close up of a person scooping seeds from a caviar dish

K-Replay: A concept car is displayed at show

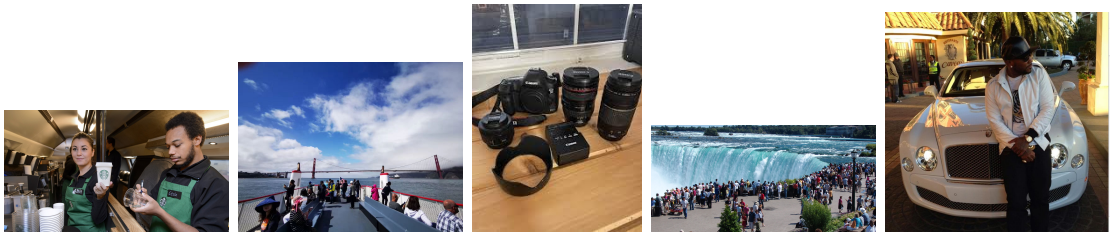
Enhanced K-Replay: A audi concept car is displayed during the shanghai auto show in shanghai

K-Replay: A man driving a car with a rear view mirror

Enhanced K-Replay: A audi a4 is being driven by a man

K-Replay: A bear in the big blue house

Enhanced K-Replay: A scooby doo cartoon with a castle in the background



K-Replay: A man and a woman wearing a green apron are holding a cup and a glass

Enhanced K-Replay: A man and a woman wearing a starbucks apron are holding a cup and a glass

K-Replay: A bridge over the water with a bunch of people on a boat

Enhanced K-Replay: caption: A view of the golden gate bridge from a boat

K-Replay: A group of cameras sitting on top of a wooden table

Enhanced K-Replay: A canon dslr and some other cameras on a table

K-Replay: a group of people standing on top of a waterfall

Enhanced K-Replay: A large group of people standing at the edge of niagara falls

K-Replay: A man standing in front of a white car

Enhanced K-Replay: A man standing in front of a bentley car

K-Replay: A deer and a rabbit looking at each other

Enhanced K-Replay: A bambi deer and a rabbit

K-Replay: A group of people standing at the top of a hill

Enhanced K-Replay: A group of people standing at the hollywood sign

K-Replay: a man standing next to a car on the side of a road

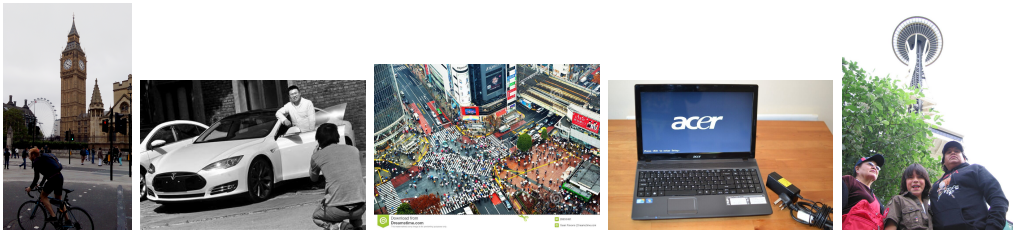
Enhanced K-Replay: A man standing next to a audi car on a road

K-Replay: A woman in a pink shirt is working on a truck

Enhanced K-Replay: A woman in a pink shirt is working on a ford truck

K-Replay: A group of people standing in front of a store

Enhanced K-Replay: two women standing in front of a walmart display



K-Replay: A clock tower with a ferris wheel in the background

Enhanced K-Replay: the big ben clock tower towering over the city of london

K-Replay: A couple of people standing next to a white car

Enhanced K-Replay: A couple of people standing in front of a white tesla

K-Replay: An aerial view of a shinjuku intersection with a lot of people

Enhanced K-Replay: A shibuya scramble crossing in tokyo city

K-Replay: A laptop computer sitting on top of a wooden table

Enhanced K-Replay: A acer laptop computer sitting on top of a wooden table

K-Replay: A group of people standing in front of a tower

Enhanced K-Replay: A group of people standing in front of the space needle

Fig. 1: Comparison of captions generated by K-Replay and Enhanced K-Replay across diverse image examples.

These changes led to substantial improvements in both the caption quality and the model's ability to recognize and describe knowledge-related concepts. The enhanced models were evaluated on both COCO and KnowCap datasets. The results showed that the enhanced model improved the recognition accuracy from 50.40% to 63.30%, while maintaining or slightly improving overall captioning scores such as CIDEr (+2.3). Notably, the generalizability of the models was evaluated on unseen knowledge categories, and it showed significant results with recognition accuracy increasing from the baseline of 45.80% to 57.69%, demonstrating the model's enhanced generalization to unseen concepts.

For future work, focusing on prompt-augmented keyword generation can improve the model's ability to generate more relevant and context-specific keywords for image captioning tasks. By leveraging prompts, the model can focus on specific aspects of the input image or its associated text. This approach may enhance the specificity and relevance of the generated keywords.

REFERENCES

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[2] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 578–10 587.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[4] S. Nikiforova, T. Deoskar, D. Paperno, V. Seggev *et al.*, "Geo-aware image caption generation," in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020, pp. 3143–3154.

[5] S. Zhao, P. Sharma, T. Levinboim, and R. Soricut, "Informative image captioning with external sources of information," *arXiv preprint arXiv:1906.08876*, 2019.

[6] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and C. Sienkiewicz, "Rich image captioning in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 49–56.

[7] S. Whitehead, H. Ji, M. Bansal, S.-F. Chang, and C. Voss, "Incorporating background knowledge into video description generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3992–4001.

[8] K. Cheng, W. Song, Z. Ma, W. Zhu, Z. Zhu, and J. Zhang, "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)*. ACM, 2023, pp. 5038–5047. [Online]. Available: https://arxiv.org/abs/2308.01126

[9] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 11 336–11 344.

[10] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 12 888–12 900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.

[12] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 23 318–23 340.

[13] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.

[14] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.

[15] Y. Zhang, Z. Wang, and J. Zhang, "A survey on enhancing image captioning with advanced strategies and techniques," *Computational and Mathematical Engineering Sciences*, vol. 142, no. 3, p. 59756, 2023. [Online]. Available: https://www.techscience.com/CMES/v142n3/59756

[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[17] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4634–4643.

[18] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 2407–2415.

[19] F. Chen, R. Ji, J. Su, Y. Wu, and Y. Wu, "Structcap: Structured semantic embedding for image captioning," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 46–54.

[20] F. Chen, R. Ji, X. Sun, Y. Wu, and J. Su, "Groupcap: Group-based image captioning with structured relevance and diversity constraints," 2017.

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, vol. 139, 2021, pp. 8748–8763.

[22] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[23] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

[24] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[25] D. Lu, S. Whitehead, L. Huang, H. Ji, and S.-F. Chang, "Entity-aware image caption generation," *arXiv preprint arXiv:1804.07889*, 2018.

[26] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! context driven entity-aware captioning for news images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 466–12 475.

[27] A. Tran, A. Mathews, and L. Xie, "Transform and tell: Entity-aware news image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 035–13 045.

[28] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015. [Online]. Available: https://arxiv.org/abs/1504.00325

[29] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3558–3568.

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[31] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005, pp. 65–72.

[32] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, 2004, pp. 74–81.

[33] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[34] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[35] S. Chen, Y. Hou, Y. Cui, W. Che, T. Liu, and X. Yu, "Recall and learn: Fine-tuning deep pretrained language models with less forgetting," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 7870–7881. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.634

[36] R. Xu, F. Luo, Z. Zhang, C. Tan, B. Chang, S. Huang, and F. Huang, "Raise a child in large language model: Towards effective and generalizable fine-tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 9514–9528.

[37] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.