Fine-Grained Cross-Modal Semantic Consistency in Natural Conservation Image Data from a Multi-Task Perspective

Rui Tao , Meng Zhu , Haiyan Cao , Hong-e Ren *

Posted Date: 12 April 2024

doi: 10.20944/preprints202404.0847.v1

Keywords: cross-modal; multi-task; image captioning; cross-modal retrieval; cross-modal alignment

Article

Fine-Grained Cross-Modal Semantic Consistency in Natural Conservation Image Data from a Multi-Task Perspective

Rui Tao [1,2], Meng Zhu [3], Haiyan Cao[2] and Hong-e Ren [1,2,4,*]

[1] College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China; trlx20@nefu.edu.cn
[2] College of Artificial Intelligence and Big Data, Hulunbuir University, Hulunbuir 021008, China; ske159@163.com
[3] College of Information Engineering, Harbin University, Harbin 150076, China; zhum913@163.com
[4] Heilongjiang Forestry Intelligent Equipment Engineering Research Center, Harbin 150040, China; nefu_rhe@163.com
* Correspondence: nefu_rhe@163.com

Abstract: The essence of cross-modal generation and retrieval modeling for image-text interactions lies in semantic consistency. Among these, cross-modal representation alignment serves as the foundational requirement for achieving cross-modal semantic consistency. The crux of our proposed method entails the collaborative training of an image captioning model(referred to as the 'captioner') and a pair of contrastive encoders for image-text matching(referred to as the 'concoder'). This synergistic approach yields concurrent enhancements in concoder's performance, captioner's proficiency, and cross-modal semantic coherence. We coin the proposed method as 'ReCap'. To begin, we initialize the concoder through knowledge distillation from a pre-trained model. During the training process, we use the output of the initialized concoder as input to a residual attention network, jointly training it with the captioner to achieve semantic consistency. Subsequently, we iteratively update the concoder during the image-text momentum encoding phase using the residual attention network, creating a closed-loop for semantic consistency resolution. Upon completion of training, the concoder and captioner modules can be used independently to perform tasks related to cross-modal retrieval and cross-modal generation of text and images. In order to substantiate the efficacy of our proposed method, we initially conducted empirical experiments encompassing image-text retrieval and image captioning tasks, employing a widely recognized benchmark dataset, MSCOCO. Subsequently, we assessed the fine-grained cross-modal alignment performance of the concoder through an image classification task on the iNaturalist 2018 dataset. The achieved performance metrics notably surpassed those achieved by several incumbent state-of-the-art models, thus validating the proficiency of our method. Finally, motivated by the practical requirements of handling the nature conservation image data, we performed caption annotation for the iNaturalist 2018 dataset. Subsequently, we trained the ReCap on this annotated data. Experimental results demonstrate that our proposed method can maintain semantic consistency between cross-modal retrieval and image caption generation for species with similar visual features but distinct semantics. This achievement signifies a meaningful exploration in the field of cross-modal semantic consistency representation, holding significant practical value in the domain of biological research.

Keywords: cross-modal; multi-task; image captioning; cross-modal retrieval; cross-modal alignment

1. Introduction

Neural networks function as parameterized databases, typically driven by specific tasks, with each network dedicated to fulfilling a corresponding task. However, there are instances where our requirements transcend single-task boundaries. Consider the context of rapidly accumulating natural conservation area image data. We seek not only to retrieve a single image but also to attach essential descriptions when summoning an image. Furthermore, we aspire to employ textual descriptions as queries to sift through our image repository, locating images that align with our specific needs. This scenario necessitates simultaneous engagement with two tasks: cross-modal image-text retrieval and image captioning.

As these data accumulate over time, the volume becomes formidable. For example, the Snapshot Serengeti Project at Serengeti National Park, Tanzania deployed hundreds of camera traps to

 doi:10.20944/preprints202404.0847.v1

understand the dynamics of African animal species. From 2010 to 2013, the project collected 3.2 million images from 225 camera traps[1]. And it was found very costly to manually process the images and add annotation labels given such a large amount of data. The project carried out by Literature [2]requires thousands of technical volunteers to work for 2-3 months to annotate image data. With the improvement of camera manufacturing technology, each camera deployed in the field can record more than 40,000 photos per day due to a single trigger event[3], and there were a lot of camera traps deployed in related projects. Literature[4] and Literature[5] deployed hundreds of camera traps in their project. Literature[6] and Literature[7] deployed about 50 cameras at water sources in the natural conservation and recorded more than 800,000 wildlife images within a few weeks.

When we resort to two separate models to independently address these tasks, we encounter suboptimal outcomes. Specifically, the images retrieved through descriptive text queries may not align with the descriptive text generated by the model for the same image. In other words, these two models exhibit inconsistent encoding and decoding for the same data. Can we train a model that maintains consistency during both encoding and decoding, all while meeting task requirements, thus mitigating semantic ambiguity within our cross-modal parameterized database?

To address this, we propose a multi-task model for joint training in cross-modal image-text retrieval and image captioning. Through the collaborative optimization of parameters, we achieve cross-module information sharing, thereby facilitating semantic-consistency encoding and decoding modeling. Post-training, the encoder and decoder can be independently employed to perform cross-modal image-text retrieval and image captioning tasks while maintaining semantic consistency between the two tasks. This is made possible because our model is constructed upon a foundation of shared semantic-consistency representation space. Of course, the prerequisite is the construction of a dataset aligning with our specific needs and the judicious design of the model's structure.

As illustrated in Figure 1, we are able to retrieve corresponding images from the dataset using a customized textual input and subsequently generate descriptive text for the retrieved images. In this paper, our objective is to preserve semantic consistency in the context of fine-grained visual features and rich textual descriptions by jointly training a retrievaler and a captioner.

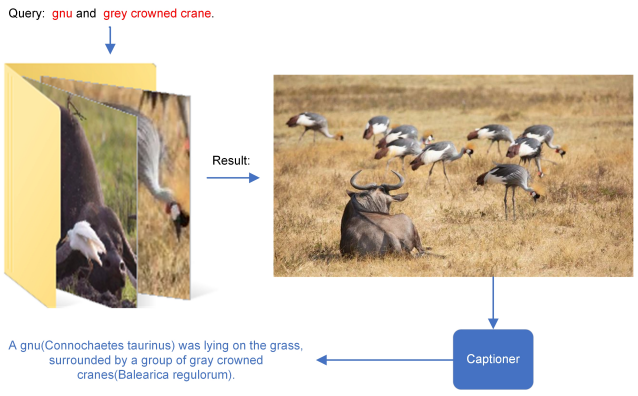


Figure 1. An Application Instance of the ReCap Model

The contributions of this work include 1) creation of a dataset of image-text pairs for natural conservation; 2) propose a combined offline and online training approach; 3) introduce a method for information transfer through collaborative parameter solving within a multi-task module; and 4) present a technique for cross-modal alignment and semantic consistency preservation based on a shared representation space for cross-modal tasks.

2. Related Work

The cross-modal semantic consistency between images and text in our research is primarily achieved through the model design and joint training of two tasks: cross-modal retrieval and image captioning. The essence of this approach lies in the optimization of the cross-modal shared space

 doi:10.20944/preprints202404.0847.v1

embedding of images and text. On one hand, optimization is performed from the perspective of cross-modal alignment between image and text entities. On the other hand, the model needs to reorganize tokens related to the input image representation in the shared space in an autoregressive manner and output them in natural language, thereby achieving semantic consistency between image and text descriptions at a broader and deeper semantic level. The encoder and decoder constitute the core modules of our designed model, involving popular techniques in cross-modal alignment and cross-modal representation fusion. Subsequently, the literature review will delve into both cross-modal representation alignment and cross-modal representation fusion.

2.1. Cross-modal Alignment

Currently, research on cross-modal alignment of image and text representations is predominantly centered around contrastive learning methods. These studies achieve the embedding and alignment of image and text representations in a shared cross-modal space by training encoders separately for each modality using a contrastive learning loss. ConVIRT[8] demonstrated the potential of contrastive objectives to learn image representations from text. Inspired by ConVIRT, CLIP[9] performs pre-training on a dataset containing 4 billion image-text pairs and becomes a milestone of vision-language models with excellent cross-modal representation. CLIP4Clip[10] demonstrated the CLIP model with high performance in cross-modal retrieval. ALIGN[11] performs pre-training on massive noisy web data. The above methods all use contrastive loss which is the most effective loss for cross-modal alignment[12–15].

Intuitively, performing cross-modal contrastive learning by treating corresponding visual and textual entities as inputs to image and text encoders, respectively, can achieve better cross-modal alignment. Therefore, some research in this domain utilizes object detection models as visual unit extractors. The extracted target pixel regions are then fed to the image encoder for contrastive learning with the text encoder, enhancing the performance of cross-modal representations. Often, these studies require the integration of a pre-trained object detection model at the front end of the visual data input[16–18]. An intuitive approach is to align the visual features of the region where the object is located with the label. For example, Oscar[19] uses Faster R-CNN[20] to detect the object in the image and then aligns it with the word embeddings of the object tags. However, they are not suitable for fine-grained cross-modal alignment as the object tags are too limited to align the vision features suitably. With a properly designed prompt, CLIP can be used for open-vocabulary classification, which solves the problem of limited object tags. ViLD[21] designed an open vocabulary object detection model by knowledge distillation from the CLIP. [22] achieved a language-driven zero-shot semantic segmentation by directly using the representation of CLIP. Groupvit[23] implements unsupervised image segmentation by using the text representation of CLIP as a pseudo label.

Contrastive learning with dual encoders, while excelling in cross-modal retrieval tasks involving images and text, encounters challenges in adapting to fine-grained cross-modal retrieval tasks with natural conservation images due to the following reasons. First, certain species' visual features in natural conservation images exhibit high intra-class and inter-class similarities, resulting in dense distributions of these highly similar representations in the shared space. This necessitates encoders with finer discriminative capabilities. Second, these encoders, trained on image-text pair datasets using contrastive learning, are often constrained by the representation of text descriptions alone and struggle to adapt well to cross-modal retrieval tasks where the semantics are similar but the expression methods differ.

2.2. Cross-Modal Fusion

With the successful application of the transformer[24] architecture in the fields of natural language processing, computer vision, and multimodal, ViLT[25]proposed a transformer-based multimodal encoder which focused on cross-modal feature fusion, and take the masked language modeling loss [26] for visual Embedding as future work. This work has been achieved by VL-BEiT[27] after VIT[28]

 doi:10.20944/preprints202404.0847.v1

and MAE[29]. From then on, a big convergence of language, vision, and multimodal pretraining is emerging. BLIP[30] proposed a new vision-language pre-training framework that transfers flexibly to both vision-language understanding and generation tasks. The multiway transformer proposed by BEiT-V3[31] has achieved state-of-the-art transfer performance in both vision and vision-language tasks. FLIP[32], which is called Fast Language-Image Pre-training, presents a simple and more efficient method for training CLIP by dropping a part of masked tokens. VLMo[33] jointly learns a dual encoder and a fusion encoder with a modular Transformer network. Coca [34] is a minimalist design to pre-train an image-text encoder-decoder foundation model jointly with contrastive loss and captioning loss like CLIP and SimVLM[35] respectively.

Cross-modal feature fusion is not suitable for cross-modal retrieval tasks due to the lack of effective optimization for unimodal encoders. However, when applied to image captioning tasks for the same input image, this method generates descriptions that share the same semantics but have different expressions. This indicates that such methods contribute to solving cross-modal semantic consistency. Our research goal is to explore the joint application of cross-modal feature fusion and cross-modal feature alignment, aiming to leverage their respective strengths and compensate for weaknesses, fostering mutual enhancement. This objective is emphasized in the Method section for in-depth discussion.

3. Materials and Methods

The entire model design and objective function construction are aimed at harnessing redundant textual information to eliminate cross-modal ambiguity in image-text pairs, thereby achieving fine-grained semantic consistency. This results in a cross-modal shared embedding space from the perspectives of both image-text retrieval and image captioning tasks. The core concept involves initially extracting representations from various modalities and aligning them preliminarily through ITC (Image Text Contrastive Learning) loss. Subsequently, image features are further semantically disambiguated within a cross-modal residual attention network through MLM(Mask Language Modeling) loss, using richer textual descriptions. This results in a further refinement of the shared embedding space. Finally, the semantically aligned image embeddings are fed to the captioner to generate textual descriptions by the LM(Language modeling) loss, enhancing the model's deeper understanding of cross-modal semantics in image-text contexts. As illustrated in Figure 2, we employed neural network architecture to seamlessly integrate multiple tasks for joint training. The joint solving of captioning and retrieval tasks allows the model to delve deeper into the understanding of cross-modal interactions between images and text. The semantic consistency sought in this context is an understanding based on image-text matching, as well as a matching grounded in joint understanding of image and text. Its essence lies in the cross-modal shared space embedding from a multi-task perspective. The model's design and training revolve precisely around this objective.

 doi:10.20944/preprints202404.0847.v1

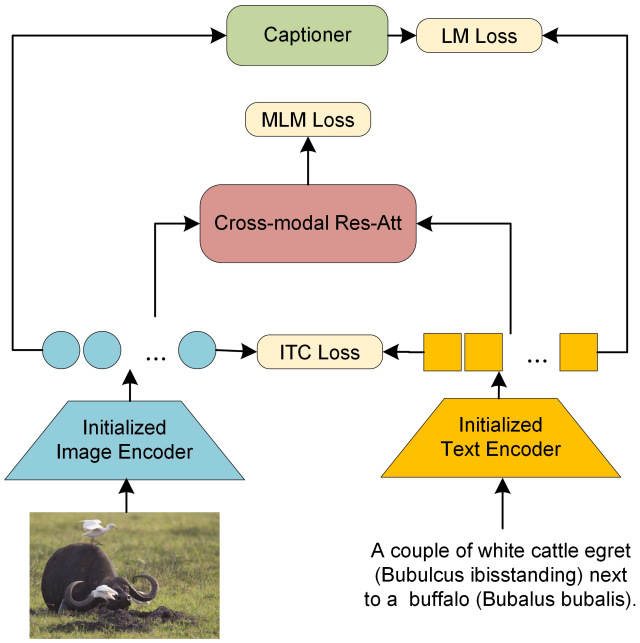


Figure 2. An Overview of training ReCap for Cross-modal Semantics Consistancy

In the initial phase, we employ knowledge distillation to initialize the encoder (Section 3.1). Subsequently, we engage in image-text pair mask learning in a self-supervised environment to achieve richer semantic consistency(Section 3.2). Finally, during the text generation phase(Section 3.3), we aim to further enhance the model's comprehension of deep-level information from the image-text context.

3.1. Contrastive Encoder Initialization

First, we designed and initialized a pair of encoders, one for images and one for text, to extract representations of image and text data (the initialized image encoder and initialized text encoder as shown in Figure 2). These unimodal encoders serve as projectors that embed each modality into a shared sematic space.

As illustrated in Figure 3, we leverage the knowledge from the pre-trained CLIP[9] model to initialize our lightweight transformer encoder. The encoder initialization is performed offline to reduce the computational requirements throughout the entire model training process. Distillation from the CLIP pre-trained encoder to the target encoder is achieved through the calculation of the L1 losss. Let the image encoder of the pre-trained CLIP model be denoted as $V()$, and the text encoder as $T()$. The distilled image encoder is denoted as $D_V()$, and the distilled text encoder as $D_T()$. The input image-text pairs are respectively denoted as i_n and t_n. The loss functions for distilling the image and text encoders are denoted as $\mathcal{L}_V$ and $\mathcal{L}_T$, respectively. The $\mathcal{L}_1$ loss for the model distillation is expressed as:

$$\begin{aligned} \mathcal{L}_V &= |D_V(i_n) - V(i_n)| \\ \mathcal{L}_T &= |D_T(t_n) - T(t_n)|. \end{aligned} \tag{1}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 April 2024 doi:10.20944/preprints202404.0847.v1

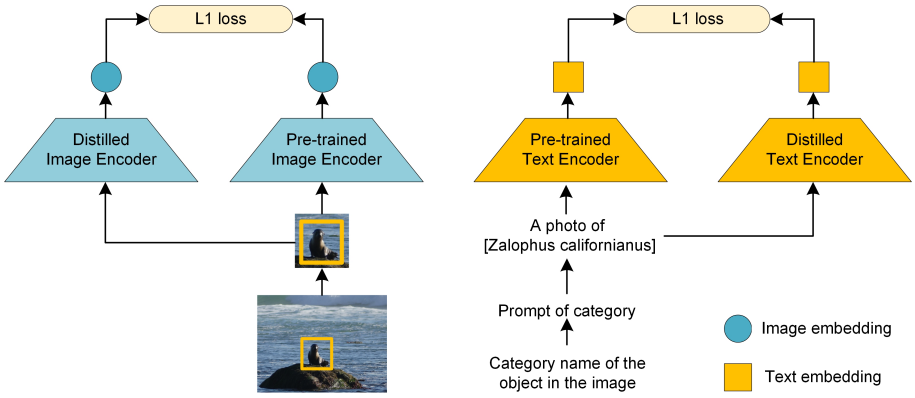


Figure 3. An Example of Knowledge Distillation from a Pre-trained Model

As illustrated in Figure 4, the unimodal encoder consists of L layers of stacked self-attention and feed-forward modules. The projector is employed to adjust the output dimensions of each module to ensure compatibility, while the normalization layer serves to balance the scale differences among various modal data, enhancing model robustness and facilitating subsequent momentum calculations.

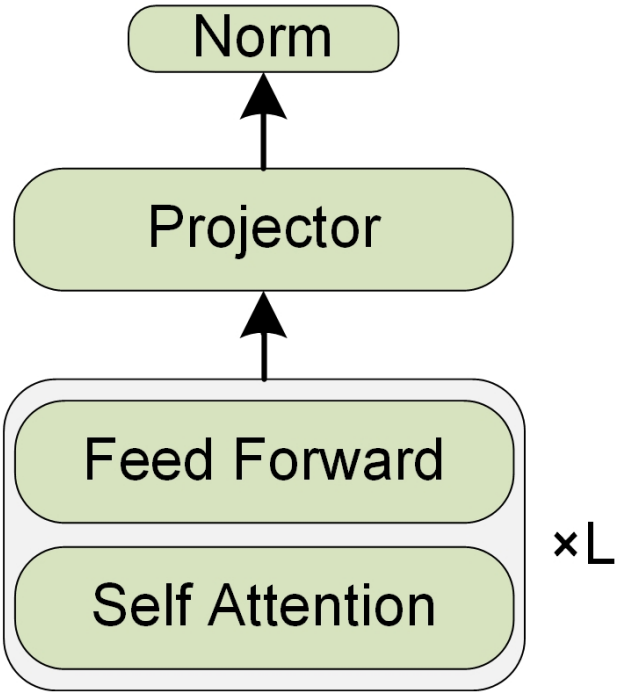


Figure 4. The Structural Details of the Distilled Encoder Module

3.2. Residual Attention Neural Network

In accordance with [36] we adopt the practice of concatenating every S Asymmetric Co-Attention (AC) block with a Connected Attention (CA) block, thereby creating a Cross-Modal Skip-Connection (CK) module. Furthermore, the Cross-modal Res-Att is concatenated with N CK modules. As visually represented in Figure 5, we represent the Self-Attention layer, Cross Attention layer, Feed Forward Network, Layer Normalization, and Concatenation layer as SA, CA, FFN, LN, and Cat, respectively. The image embedding is denoted as $v = \{\mathcal{V}_{cls}, \mathcal{V}(I), \mathcal{V}(R_1), \ldots \mathcal{V}(R_i)\}$, while the text embedding is represented as $l = \{\boldsymbol{w}_{\text{cls}}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$, consisting of word vectors corresponding to the input caption paired with I. Here, 'I' signifies the input image, 'R_i' refers to the i-th patch within it, and an additional

 doi:10.20944/preprints202404.0847.v1

[CLS] token is utilized to summarize the input sequence. Let l^{S-1}, v^{S-1} and l^S represent the input word vectors, visual features, and output of the S-th AC layer respectively. Then

$$l_{SA}^S = LN\left(SA\left(l^{s-1}\right) + l^{s-1}\right), \tag{2}$$

$$l_{CA}^S = LN\left(CA\left(l_{SA}^S, v^{N-1}\right) + l_{SA}^S\right), \tag{3}$$

$$l^s = LN(FFN(l_{CA}^s) + l_{CA}^s). \tag{4}$$

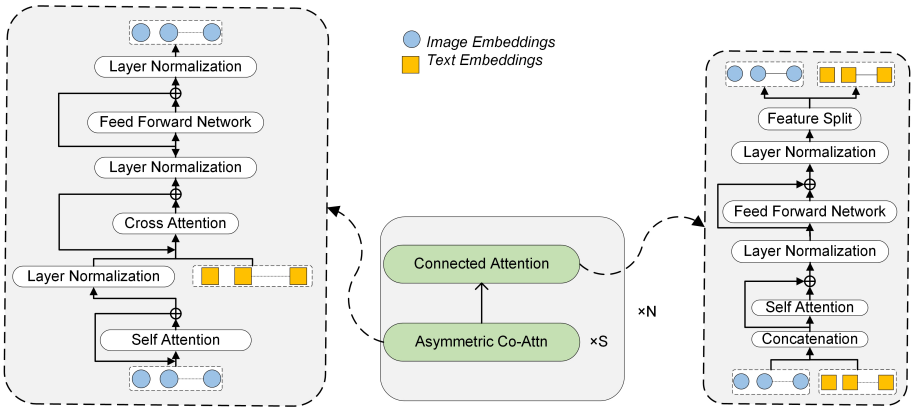


Figure 5. Residual Attention Network Architecture

Subsequently, we feed both l^s and v^{N-1} into a CA block to facilitate cross-modal information interaction. The computation $[v^N; l^S]$ of the CK module's output, denoted as follows:

$$[v^{N-1}; l^{N-1}] = Cat([v^{N-1}, l^S) \tag{5}$$

$$[v_{SA}^N; l_{SA}^N] = LN(SA([v^{N-1}; l^{N-1}]) + [v^{N-1}; l^{N-1}]), \tag{6}$$

$$[v^N; l^N] = LN(FFN([v_{SA}^N; l_{SA}^N]) + [v_{SA}^N; l_{SA}^N]). \tag{7}$$

The primary training objective for Cross-modal Res-Att is Masked Language Modeling (MLM). In this context, let us denote a caption as CnP and the set of randomly masked positions as M_CnP. The MLM loss can be formally defined as follows:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in M_{CnP}} \log p\left(CnP_i \mid CnP_{\setminus M_{CnP}}\right), \tag{8}$$

where $CnP_{\setminus M_{CnP}}$ is the masked version of the input caption, i.e., *Two* [MASK] *are*[MASK] *on* [MASK] *of a pond*. The Cross-modal Res-Att module predicts the masked tokens based on image and text context.

3.3. Captioner Training Objectives

The cross-modal image features output by the Cross-modal Res-Att module are denoted as$F^k = F_1^k, ...F_u^k$.

The objective of the Captioner module is to produce output C^k based on the input F^k.Here, C^k represents the vector sequence form of the image description generated by the Captioner module,

 doi:10.20944/preprints202404.0847.v1

recorded as $C^k = C_1^k ... C_m^k$, while the trainable parameters are denoted as θ. Subsequently, the training objective of the captioner can be defined as follows:

$$\max_{\theta} \sum_{k=1}^{N} \log p_\theta \left(\mathrm{C}_1^{\mathrm{k}}, \ldots\ldots, \mathrm{C}_m^{\mathrm{k}} \mid \mathrm{F}^{\mathrm{k}} \right). \tag{9}$$

The generated word, in conjunction with F^k, serves as the condition for predicting the subsequent word. Subsequently, the Language Modeling (LM) loss function can be expressed as:

$$\mathcal{L}_{\mathrm{LM}} = \max_{\theta} \sum_{k=1}^{N} \sum_{j=1}^{m} \log p_\theta \left(\mathrm{C}_{\mathrm{j}}^{\mathrm{k}} \mid \mathrm{F}^{\mathrm{k}}, \mathrm{C}_{\mathrm{j}}^{\mathrm{k}}, \ldots \mathrm{C}_{\mathrm{j-1}}^{\mathrm{k}} \right). \tag{10}$$

The loss function for the joint training of ResAtt and Captioner is the sum of Equation 8 and Equation 10, which is:

$$\mathcal{L}_{\mathbf{Res\&Cap}} = \mathcal{L}_{\mathrm{MLM}} + \mathcal{L}_{\mathrm{LM}}. \tag{11}$$

3.4. Offline Cross-Module Information Propagation Based on Momentum Encoder

Incorporating the Momentum Encoder Method Proposed by MOCO[12] for Offline Parameter Updates of Unimodal Encoders: Due to the high compression ratio of text during the contrastive learning phase, many details are lost in cross-modal learning between text and images. Upon entering the ResAtt module, increasing the redundancy of text helps eliminate ambiguity in cross-modal learning, thereby enhancing the semantic consistency of cross-modal representations. The information-theoretic basis for this lies in the fact that redundant information aids in ambiguity elimination. As shown in Figure 5, the ResAtt network module preserves word vectors for text, unlike the unimodal contrastive learning phase, which compresses the entire text expression into a single vector. Furthermore, the ResAtt network exhibits higher redundancy in cross-modal learning of text-image representations, including longer forward channels and the use of text as residuals to repeatedly solve for semantic consistency with images, reducing ambiguities in the process of cross-modal semantic representation.

3.4.1. Momentum Encoder

In brief, the principle of the Momentum Encoder is that the training of the encoder in unsupervised learning can be simplified as a look-up table problem. In other words, an encoded query should have high similarity to its corresponding key and low similarity to other keys. This simplifies the entire process to minimizing the contrastive loss. During the solving process, contrastive learning requires a queue containing keys for both positive and negative samples to look up for queries. To maintain the consistency of encoding for positive and negative samples in the queue, the Momentum Encoder is employed.

The encoding update rule for the Momentum Encoder is shown in Equation 12, where the momentum parameter $m \in [0, 1)$ is used. The query encoding θ_q is updated based on gradient back propagation, while the key encoding θ_k is updated using momentum. Typically, m takes a value greater than 0.9, which is equivalent to taking a moving average of the encoding updates. The slow-changing Momentum Encoder reduces the difference between the encoding of positive and negative samples in the queue, thereby improving the cross-task transfer performance of the encoder optimization process based on momentum in contrastive learning.

$$\theta \leftarrow m\theta_k + (1 - m)\theta_q \tag{12}$$

3.4.2. Image-Text Contrastive Loss Funtion

Following [8], The image-text contrastive learning (ITC) formulate the loss function according to InfoNCE[37]. Let T denote a certain species class embedding and V denote its visual embedding,

 doi:10.20944/preprints202404.0847.v1

then we have the embedding pair(V, T). We use (V_i, T_i) to denote the i-th pair of positive samples and $(V_i, T_j)j \neq i$ a pair of negative samples. The ITC training objective of ReCap consists of two loss functions to make the distance of the positive pair closer than the negative one in the embedding space. Since ITC is asymmetric for each modality, it needs to be computed separately from both directions for images and text. The contrastive loss for the i-th pair in the image $\rightarrow$ text direction:

$$\ell_i^{(V \rightarrow T)} = -\log \frac{sim(\mathbf{V}_i, \mathbf{T}_i)/\tau}{\sum_{k=1}^{n} sim(\mathbf{V}_i, \mathbf{T}_k)/\tau}, \tag{13}$$

Where $sim(\cdot)$ is the cosine similarity, i.e., $sim(a, b) = a^\top b/(\|a\|\|b\|)$,and τ is a temperature parameter. Similarly we formulate the text $\rightarrow$ image loss as:

$$\ell_i^{(T \rightarrow V)} = -\log \frac{sim(\mathbf{T}_i, \mathbf{V}_i)/\tau}{\sum_{k=1}^{M} sim(\mathbf{T}_i, \mathbf{V}_k/)\tau}, \tag{14}$$

Finally, the training objective is a weighted sum:

$$\mathcal{L}_{\mathbf{ITC}} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \ell_i^{V \rightarrow T)} + (1 - \lambda) \ell_i^{(T \rightarrow V)} \right), \tag{15}$$

where $\lambda \in [0, 1]$ is a hyperparameter weight, and N is batch size.

3.4.3. Offline Cross-Module Information Propagation

The cross-module joint solving of parameters constitutes the inter-module propagation of information. Deep learning models are essentially parameterized databases, with relationships among data implicitly encoded within the model's parameters. Therefore, cross-module operations on parameters represent the propagation of information across modules.

Firstly, as illustrated in Figure 6, we feed the image-text paired dataset to the unimodal encoders, obtaining image encodings $(Ve0, Ve1, Ve2...)$ through the ITC loss. Subsequently, as depicted in Figure 7, we feed $(Ve0, Ve1, Ve2...)$ to the ResAtt module, applying Equation 12 to compute the visual momentum encoding queue $(Vm0, Vm1, Vm2...)$ and the text momentum encoding queue $(Tm0, Tm1, Tm2...)$. Then, as shown in Figure 8, we feed back the momentum encodings to update the unimodal encoders. Repeating these steps forms a closed-loop for information propagation, all of which can be conducted offline. Of course, if computational resources permit, online processing is also feasible.

 doi:10.20944/preprints202404.0847.v1

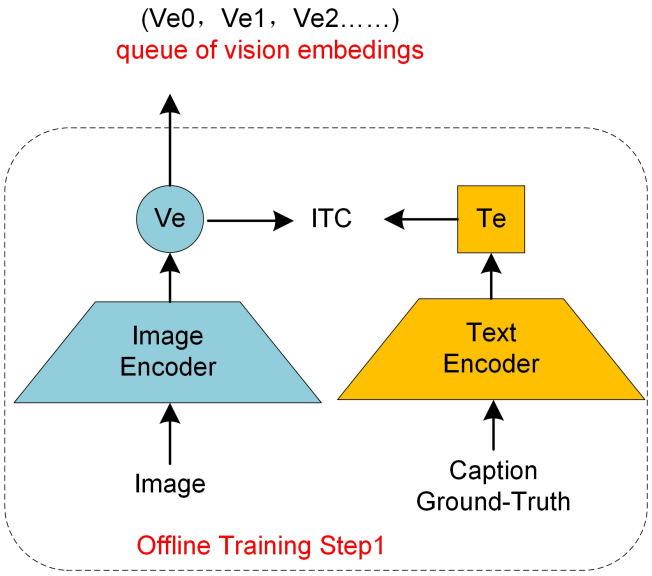


Figure 6. Initial Visual Encoding

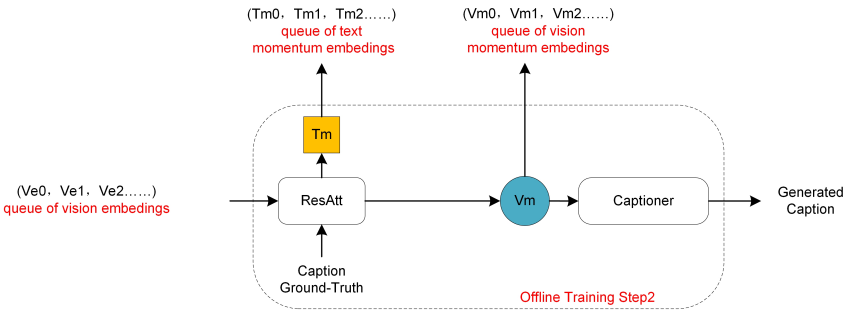


Figure 7. Redundant Disambiguation Momentum Encoding

 doi:10.20944/preprints202404.0847.v1

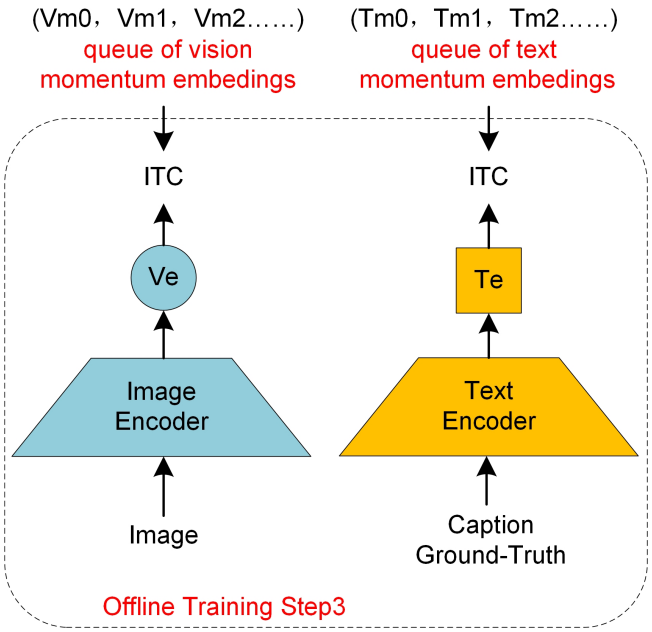


Figure 8. Unimodal Encoding Momentum Update

3.4.4. Why Contrastive learning and Momentum Encoding

Contrastive learning is a multi-classification approach that treats each sample as a class. This classification paradigm, under the context of cross-modal embeddings, enhances the flexibility of representations in a shared space. However, it inevitably introduces ambiguity, particularly as granularity increases and feature vectors become more abundant, thereby complicating the challenge of solving semantic consistency. Since representation embeddings are determined by encoders, our unimodal encoders highly compress image and text features, deliberately for the sake of embedding and retrieval efficiency. Nonetheless, higher compression ratios inevitably lead to the loss of some information, posing a fundamental question of how to balance encoder performance and efficiency.

Image captioning is a typical task in cross-modal understanding, where input image features are reorganized into textual descriptions. Essentially, this task involves computing similarities between image and text representations to embed them into a shared space. Both the unimodal encoder and captioner used for retrieval tasks are based on cross-modal shared space embeddings, but their embedding results differ. Can we share information between the captioner's embedding space and the unimodal encoder's embedding space to optimize each other? Typically, in the absence of a multitask mapping network, the captioner directly takes the image encoding output by the unimodal encoder as input and generates a textual description based on it. However, the problem lies in the fact that the same image encoding is encoded and decoded in different cross-modal shared spaces, which is the fundamental reason for the semantic inconsistency in the encoding and decoding processes described earlier. How can information be exchanged between these two spaces, complementing each other, in order to assimilate the representation spaces of the two tasks into a single shared representation space with semantic consistency? This is the focal point of our deep learning network design. To achieve this, an information transfer feedback channel is required. Deep learning models are parameterized databases, and joint parameter solving enables information transfer, meaning that the model updates its understanding of the data and the embeddings of various modalities in the shared space through parameter updates. This brings us to the topic of ResAtt network. Since the captioner only generates text and does not modify the image embedding, the image embedding output by the unimodal encoder is not affected by the captioner. Alternatively, one could say that the captioner simply uses the image embedding as a prefix and does not perform any operations on it. However, cross-modal representation requires shared space embedding based on repeated cross-validation of information

 doi:10.20944/preprints202404.0847.v1

from both modalities. Therefore, it is necessary for the image representation encoded by the unimodal encoder to interact with more richly correlated textual descriptions across modalities to optimize its representation in the shared embedding space. Such optimization also facilitates the captioner in smoothly generating corresponding descriptions for images. The ResAtt network serves as a bridge and information conduit between the captioner and the unimodal encoder.

Why adopt momentum encoding? Contrastive learning is based on the differentiation between positive and negative samples. If embeddings of positive and negative samples come from different encoders or different training stages of the same encoder, the model learns the differences between encoders rather than the differences between positive and negative samples. To ensure a "fair" comparison between positive and negative samples and enable the model to finely learn the data's features, it's necessary to maintain the consistency of encodings in a longer queue of positive and negative sample encodings (in our case, a queue length of 4096). This is why we employ momentum encoders, a key technique in our multi-task deep learning network design.

In general, the concoder module serves as a mechanism for projecting image and text sequences into vectors to achieve cross-modal alignment at a macroscopic level. Within the ResAtt module, images are represented as sequences of patches, while text is represented as sequences of words, facilitating micro-level alignment. Co-training the ResAtt module with the captioner module enables the model to achieve alignment in the context of cross-modal understanding.

4. Experiments

The content of this section primarily focuses on validating the effectiveness of our approach (ReCap) through comparisons on mainstream tasks using mainstream datasets. Specifically, these tasks encompass image captioning and image-text retrieval on the COCO dataset, classification tasks on the iNaturalist2018 dataset, as well as image captioning and image-text retrieval tasks on the iNaturalist2018 dataset.

4.1. Dataset Settings

we utilized the Karpathy split[38] of the MSCOCO dataset[39], comprising 123,000 images, with each image accompanied by five sentences as annotations.The iNaturalist2018 dataset comprises 8,142 distinct species, each serving as an individual image classification category. It encompasses a total of 437,513 training images and 24,426 validation images. As this dataset initially lacked caption annotations, we conducted a comprehensive annotation effort, providing five sentences of description for each image. Furthermore, we annotated both the common name and the Latin name for each species.In Table 1, we present some examples of our annotated data.

Table 1. Samples of nature conservation image-text pair dataset

Images				
Captions	Two geese are walking on the shore of a pond.	A bunch of yellow flowers are sitting in a field.	A Catasticta nimbice is sitting on an Ageratum houstonianum in the sun.	An Aepyceros melampus grazing in a field.

4.2. Implementation Details

We utilized eight NVIDIA 3090 24G GPUs for the image-text encoder contrastive learning training process, with a queue length set to 4096 and a momentum parameter of 0.995. We employed the

 doi:10.20944/preprints202404.0847.v1

AdamW optimizer with a decay weight set to 0.02. The learning rate was warm-up set to 1e-4 for the first 1000 iterations and decayed in a cosine function manner to 1e-5 for the subsequent iterations. The total training duration for the model was approximately 127 hours.

4.3. Evaluation Metrics

The image caption model employs four widely recognized evaluation metrics, namely BLEU (Bilingual Evaluation Understudy) [40], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [41], CIDEr (Consensus-based Image Description Evaluation) [42], and SPICE (Semantic Propositional Image Captioning Evaluation) [43]. Among these, BLEU4 segments sentences into four-word chunks to gauge the descriptive accuracy of the model-generated captions. METEOR, building on the foundations of BLEU, addresses the issue of excessive word matching while emphasizing word recall and precision.

CIDEr, primarily applied in the domain of image description, employs TF-IDF (Term Frequency-Inverse Document Frequency) to weigh each sentence fragment. It encodes the frequency (E_r) of a fragment in the reference description and the frequency (E_c) in the generated description. Subsequently, it computes the similarity between E_r and E_c to generate an evaluation score for the model.

SPICE, on the other hand, is an evaluation metric based on scene graphs and semantic concepts. It assesses the extent to which the model-generated description aligns with the entities, attributes, and relationships present in the image.

The image classification task on iNaturalist has only one label for each picture, denote as g_i, the result predicted by the model is denoted as p_i, and the error rate is

$$e_i = \min_i d(g_i, p_i), \tag{16}$$

where $d(\cdot)$ is

$$d(x, y) = \begin{cases} 0 & \text{if} \quad x = y \\ 1 & \text{otherwise} \end{cases}, \tag{17}$$

the total score is

$$\text{score} = \frac{1}{N} \sum_i e_i. \tag{18}$$

4.4. Evaluation on the MSCOCO Dataset

We trained models on the MSCOCO dataset to perform image captioning and image-text retrieval tasks, in order to validate the effectiveness of the proposed method. Table 2 presents the performance comparison of ReCap with state-of-the-art models in the context of image captioning. Here, B4 denotes BLEU-4, C represents CIDEr, M stands for METEOR, and S corresponds to SPICE. Further details are provided in Section 4.3. Table 3 illustrates the performance comparison of ReCap in image-text retrieval tasks against high-level models. Here, I2T denotes image-to-text retrieval, while T2I represents text-to-image retrieval. R@1, R@5, and R@10 respectively indicate recall rates for the top 1, top 5, and top 10 retrieval recommendations. The experimental results demonstrate that ReCap outperforms several state-of-the-art models, thereby validating the efficacy of the proposed method.

Table 2. Quantitative Analysis of Image Captioning on MSCOCO Dataset(%)

Method	B4	C	M	S
Oscar[19]	36.6	124.1	30.4	23.2
BUTD[44]	36.2	113.5	27.0	20.3
UnifiedVLP[45]	33.53	113.1	27.5	21.1
ClipCap[46]	33.5	113.1	27.5	21.1
ReCap	39.8	126.7	31.6	24.4

 doi:10.20944/preprints202404.0847.v1

Table 3. Quantitative Analysis of Cross-modal Retrieval on COCO Dataset(%)

Method	Retrieval I2T			Retrieval T2I		
	R@1	R@5	R@10	R@1	R@5	R@10
Oscar[19]	57.5	82.8	89.8	73.5	92.2	96.0
METER[47]	57.1	82.7	90.1	76.2	93.2	96.8
ViSTA[48]	52.6	79.6	87.6	68.9	90.1	95.4
ALADIN[49]	51.3	79.2	87.5	64.9	88.6	94.5
ReCap	65.5	89.2	92.9	77.1	92.6	96.3

Based on the comparative data in Table 2, it is evident that ReCap demonstrates improved performance compared to others. Taking the scores in the B4 column as an example, ReCap's score has increased by nearly 7 points. This improvement can be attributed to two main enhancements. Firstly, the incorporation of an open vocabulary, meaning there is no restriction on the number of categories. Secondly, the ResAtt network excels in the fusion of cross-modal features, effectively emulating the representation style of the dataset. This results in a higher overlap between the generated captions and the ground truth.

As shown in Table 3, in the retrieval task of image to text, the R@1 score exhibits an improvement of approximately 8 to 14 percentage points compared to others. In the text to image retrieval task, there is an improvement of approximately 1 to 12 percentage points compared to others. This indicates a significant effect of the proposed method in the cross-modal alignment of image and text features. The improvement in text to image retrieval performance is relatively challenging due to the high information compression in textual data and the sparse nature of image data. When calculating mutual information, the same textual representation often exhibits similarity to a larger number of image representations. For instance, different models of cars appearing in images with similar backgrounds would have high similarity. To effectively differentiate between the brand and model of cars in the image, a finer-grained cross-modal alignment is required for text to image retrieval. Therefore, adopting an open vocabulary approach during the training of the image encoder is essential, as it avoids limitations to a finite set of categories and proves crucial in the cross-modal modeling tasks involving image and text.

4.5. Evaluation on the iNaturalist Dataset

In accordance with the introduction, the motivation behind this study is to address the need for cross-modal processing of vast quantities of imagery data from natural conservations. In order to assess the cross-modal alignment of the model's representations between images and text, we opted to employ the image classification task on the iNaturalist2018 dataset. This section's experiments were conducted independently using the concoder module (as detailed in Section 3.1). Notably, concoder was originally designed without a classification head. To achieve classification, we employed a method that involves comparing the representations output by the image encoder with the prompt encodings generated by the text encoder.

The format of the prompts used is: 'a photo of <category >' where 'category' corresponds to the category names in the dataset. In other words, for as many categories as there are in the dataset, there are corresponding prompts. In essence, our image classification approach assigns an image to the category with the highest similarity to its image representation. Specific experimental results are presented in Table 4. The experimental outcomes demonstrate that ReCap outperforms several state-of-the-art models, thereby confirming the the proposed method's cross-modal alignment capability between image features and textual representations for species.

 doi:10.20944/preprints202404.0847.v1

Table 4. Comparison on Image Classification on iNaturalist 2018(%)

Method	Top1 Accuracy
MetaFormer[50]	84.3
OMNIVORE[51]	84.1
RegNet-8GF[52]	81.2
VL-LTR[53]	81.0
μ2Net+[54]	81.0
MixMIM-L[55]	80.3
DeiT-B[56]	79.5
CeiT-s[57]	79.4
GPaCo[58]	78.1
ReCap	**85.1**

As shown in Table 4, ReCap demonstrates a performance improvement of approximately 1 to 7 percentage points compared to others. This indicates that our proposed method, employing an open vocabulary approach, is capable of handling image classification tasks on the iNaturalist Dataset. The experimental results not only affirm the effectiveness of our method in cross-modal representation alignment but also validate the feasibility of applying this approach to open vocabulary image classification tasks.

4.6. Qualitative Evaluation

The top-5 results for Text to Image Retrieval are illustrated in Figure 9. Both non-target images and target images contain relevant content related to grassland and the target species. From the perspective of our application, we seek relatively open-ended retrieval results. This approach allows the model to continuously improve through small-sample learning in real-world applications. If the model were confined to strict one-to-one retrieval, it would lack practical utility.

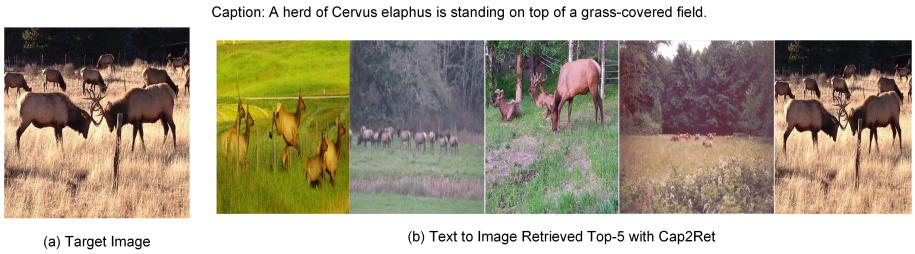
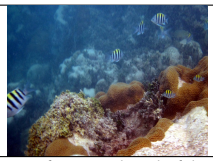


Figure 9. Examples of Text to Image Retrieval on Validation Dataset

As shown in Table 5, the captions generated by the model align well with the content of the test images, and the species names are consistent with the Latin names used in the training set. This intuitively demonstrates the model's learning capability in the domain of image-text cross-modal alignment. In the fourth prediction, the bear species (Ursus arctos horribilis) occurred 24 times in the training set, but there were no caption annotations for "cubs" in the training data prior to GPT-2 fine-tuning. This underscores the importance of pre-existing knowledge within NLP models for image captioning tasks, as it can provide additional information that is subsequently expressed in the form of generated descriptions. In the context of our approach, aligning image representations cross-modally into the pre-trained NLP decoder representation space leverages the rich knowledge of the NLP decoder for a deeper understanding of the images.

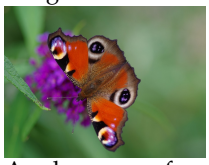
 doi:10.20944/preprints202404.0847.v1

Table 5. Examples Sentences Generated by ReCap for Test Images

Images				
Captions	A few Abudefduf saxatilis swim in the stony water.	There are some red Castilleja indivisa in the grass.	A Libellula quadrimaculata is flying over the water.	A Ursus arctos horribilis and her cubs on a green field.

We conducted zero-shot experiments using three datasets related to natural conservations. These three datasets are as follows: birds 525 species - image classification (downloadable from https://www.kaggle.com/datasets/gpiosenka/100-bird-species), Animals Detection Images Dataset (available at https://www.kaggle.com/datasets/antoreepjana/animals-detection-images-dataset), and Wildlife Conservation Society (accessible from https://library.wcs.org/Library/Science-Data/Datasets.aspx). The experimental procedure is as follows: Firstly, we designed sentences resembling "A photo of <species >" based on the dataset content. Subsequently, we performed text to image retrieval with these sentences and provided the retrieved images to the captioner for generating descriptive text. The experimental results are presented in Table 6. The experimental results indicate that the species names on the retrieval side, the species within the images, and the species names on the generation side are all consistent. This observation underscores that the features extracted by the image encoder and text encoder are aligned, and the semantics of the encoder and decoder are in harmony, visually demonstrating the model's capabilities in cross-modal alignment and semantic consistency between text and images. Examining the generated captions reveals the decoder's capacity for systematic descriptions of foreground and background elements. This is a result of the combined influence of the model's prior knowledge and fine-tuning.

Table 6. Examples of ReCap Zero-shot Retrieval & captioning

Query	A photo of Leopardus pardalis.	A photo of Phoenicopterus rubber.	A photo of Aglais io.
Dataset	Wildlife Conservation Society	Birds 510 Species-Image Classification	Animals Detection Images Dataset
Result			
Caption	A small Leopardus pardalis walking through a forest at night.	A pink Phoenicopterus ruber standing in the water.	A close-up of an Aglais io is sitting on top of a flower.

4.7. Ablation Study

The results of the ablation experiments are presented in Table 7. In the table, the term "C+C" indicates a direct connection between the conocoder and captioner, where the visual representations generated by the encoder are used as input for the captioner. "C+R+C" signifies the bridging of conocoder and captioner through the ResAtt module. Finally, "C+R+M+C" represents an extension of "C+R+C" with the addition of a momentum feedback loop.

From the experimental results in the "C+C" row, it can be observed that the I2T and T2I performance on both datasets is relatively consistent, maintaining an average level. In comparison to the performance of ReCap, there is a slight decrease in T2I, while I2T and image captioning exhibit more substantial performance degradation. This suggests that when the encoder and decoder operate independently, the model's performance heavily relies on the knowledge inherited from pre-trained

 doi:10.20944/preprints202404.0847.v1

models and the training process. However, without a channel for information transfer between them, they cannot leverage distinct task perspectives from each other to enhance each other's performance.

Looking at the experimental results in the "C+R+C" row, there is a noticeable improvement in the performance of image captioning compared to the "C+C" row. This indicates that after a finer-grained cross-modal alignment of image and text representations at the micro-level, it becomes more favorable for the captioner to generate descriptions for images. It is evident that the ResAtt module significantly contributes to the optimization of cross-modal representation alignment and the refinement of shared semantic space embedding for text and images.

ReCap and the "C+R+C" configuration only differ in the presence of a momentum feedback loop in their model structures. From the experimental results, it is evident that there are overall performance improvements in the model, particularly in the I2T and image captioning tasks. This suggests that the feedback information on the decoding side significantly aids in enhancing the performance of the encoder, resulting in substantial gains in cross-modal alignment of image and text representations.

The improvement in image captioning performance further illustrates that, after optimizing the encoder's performance, it is possible to further enhance the decoder's performance. From a structural perspective, this involves optimizing the pre-modules while also elevating the performance of various components along the forward path, thus establishing a beneficial feedback optimization cycle.

Table 7. Ablation Study of ReCap on the MSCOCO and iNaturalist 2018 Datasets

	MSCOCO			**iNaturalist2018**		
Module Composition	I2T-R@1	T2I-R@1	Cap-B4	I2T-R@1	T2I-R@1	Cap-B4
C+C	51.5	75.2	31.9	54.1	68.9	32.3
C+R+C	51.3	75.7	35.3	53.7	69.5	36.1
ReCap	65.5	77.1	39.8	63.6	72.2	41.0

5. Conclusion

The image-text representation initially undergoes coarse alignment through the encoder, followed by fine-grained alignment by the decoding side consisting of ResAtt and Captioner. Subsequently, the encoder is momentum-updated based on decoding side information, forming feedback from the decoding side to the encoding side, enhancing both the quality of the encoder and caption generation. The essence of this process lies in the sharing of a semantic space, where the decoder imparts its understanding of embedding similarities and categorization to the encoder. These insights are propagated to the encoder's network parameters through momentum-based backpropagation. Furthermore, contrastive learning on the encoding side plays a crucial role. As mentioned earlier, the classification in contrastive learning is open-ended, with as many categories as there are samples. Such a classification method has no upper limit on granularity, compelling the encoder to learn subtle distinctions among samples as much as possible. Achieving this solely from the encoding side would be information bottlenecked, and this is where feedback from the decoding side effectively bridges the information gap. Experimental results also confirm the contribution of prior knowledge in the decoder during this process. In summary, the feedback from the decoding side, the prior knowledge in the decoder, and momentum updates collectively enhance the quality of feature extraction in the encoder. All of this coalesces into a shared semantic space embedding for the encoder-decoder, where both entities possess a shared and aligned embedding space, embodying the essence of semantic consistency.

The performance of both cross-modal retrieval in image-text pairs and generative models fundamentally depends on the quality of shared space embeddings. The main contribution of our proposed method lies in the effective fusion of the advantages of both tasks in the cross-modal shared space embedding of images and text through thoughtful model design. This approach is particularly suitable for scenarios where there are strict alignment requirements between the objects in the image and the vocabulary in the text. Moreover, it demands that the model can further associate the input image

 doi:10.20944/preprints202404.0847.v1

representation with a more extensive and semantically rich textual description along a longer logical chain. Our proposed method is well-suited for such scenarios.

Author Contributions: R. Tao was responsible for model design, model training, dataset construction, and code writing and debugging. M. Zhu was responsible for model design and code debugging. H. Cao contributed to dataset annotation checking, article editing, and revision. H. Ren contributed to conceptualization, methodology, draft writing, reviewing, and provided experimental conditions, including the artificial intelligence laboratory and experimental equipment.

Funding: This work was supported by the Natural Science Foundation of Heilongjiang Province (LH2020F040), the Young Doctoral Research Initiation Fund Project of Harbin University "Research on Wood Recognition Methods Based on Deep Learning Fusion Model" Project (HUDF2022110), the Self-funded project of Harbin Science and Technology Plan Research on Computer Vision Recognition Technology of Wood Species Based on transfer learning Fusion Model Project (2022ZCZJCG022), and the Fundamental Research Funds for the Central Universities (2572017PZ10).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Matin, M.; Shrestha, T.; Chitale, V.; Thomas, S. Exploring the potential of deep learning for classifying camera trap data of wildlife: a case study from Nepal. In Proceedings of the AGU Fall Meeting Abstracts, 2021, pp. GC45I–0923.
2. Norouzzadeh, M.S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M.S.; Packer, C.; Clune, J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* **2018**, *115*, E5716–E5725. https://doi.org/10.1073/pnas.1719367115.
3. Zett, T.; Stratford, K.J.; Weise, F. Inter-observer variance and agreement of wildlife information extracted from camera trap images. *Biodiversity and Conservation* **2022**, *31*, 3019–3037. https://doi.org/10.1007/s10531-022-02472-z.
4. Swanson, A.; Kosmala, M.; Lintott, C.; Simpson, R.; Smith, A.; Packer, C. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific data* **2015**, *2*, 1–14. https://doi.org/10.1038/sdata.2015.26(2015).
5. McShea, W.J.; Forrester, T.; Costello, R.; He, Z.; Kays, R. Volunteer-run cameras as distributed sensors for macrosystem mammal research. *Landscape Ecology* **2016**, *31*, 55–66. https://doi.org/10.1007/s10980-015-0262-9.
6. Edwards, S.; Portas, R.; Hanssen, L.; Beytell, P.; Melzheimer, J.; Stratford, K. The spotted ghost: density and distribution of serval Leptailurus serval in Namibia. *African Journal of Ecology* **2018**, *56*, 831–840. https://doi.org/10.1111/aje.12540.
7. Stratford, K.; Stratford, S.; Périquet, S. Dyadic associations reveal clan size and social network structure in the fission–fusion society of spotted hyaenas. *African Journal of Ecology* **2020**, *58*, 182–192. https://doi.org/10.1111/aje.12641.
8. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive learning of medical visual representations from paired images and text (2020). *arXiv preprint arXiv:2010.00747* **2020**. https://doi.org/10.48550/arXiv.2010.00747.
9. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. Proceedings of Machine Learning Research, 2021, pp. 8748–8763.
10. Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; Li, T. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* **2022**, *508*, 293–304. https://doi.org/10.1016/j.neucom.2022.07.028.
11. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International conference on machine learning. Proceedings of Machine Learning Research, 2021, pp. 4904–4916.

 doi:10.20944/preprints202404.0847.v1

12. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR 2020, 2020, pp. 9729–9738.
13. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International conference on machine learning. PMLR, 2020, pp. 1597–1607.
14. Li, J.; Zhou, P.; Xiong, C.; Hoi, S.C. Prototypical Contrastive Learning of Unsupervised Representation. In Proceedings of the International Conference on Learning Representations. ICLR2021, 2021.
15. Li, J.; Xiong, C.; Hoi, S. MoPro: Webly Supervised Learning with Momentum Prototypes. In Proceedings of the International Conference on Learning Representations. ICLR2021, 2021.
16. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX. Springer, 2020, pp. 104–120.
17. Xu, X.; Wang, T.; Yang, Y.; Zuo, L.; Shen, F.; Shen, H.T. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems* **2020**, *31*, 5412–5425. https://doi.org/10.1109/TNNLS.2020.2967597.
18. Diao, H.; Zhang, Y.; Ma, L.; Lu, H. Similarity reasoning and filtration for image-text matching. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence. AAAI, 2021, pp. 1218–1226.
19. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. Springer, 2020, pp. 121–137.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.
21. Gu, X.; Lin, T.Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv:2104.13921* **2021**. https://doi.org/10.48550/arXiv.2104.13921.
22. Li, B.; Weinberger, K.Q.; Belongie, S.; Koltun, V.; Ranftl, R. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546* **2022**. https://doi.org/10.48550/arXiv.2201.03546.
23. Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; Wang, X. Groupvit: Semantic segmentation emerges from text supervision. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2022, 2022, pp. 18134–18144.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017.
25. Kim, W.; Son, B.; Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Proceedings of the International Conference on Machine Learning. PMLR, 2021, pp. 5583–5594.
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**. https://doi.org/arXiv:1810.04805.
27. Bao, H.; Wang, W.; Dong, L.; Wei, F. Vl-beit: Generative vision-language pretraining. *arXiv preprint arXiv:2206.01127* **2022**. https://doi.org/arXiv:2206.01127.
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**. https://doi.org/arXiv:2010.11929.
29. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2022, 2022, pp. 16000–16009.
30. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning. PMLR, 2022, pp. 12888–12900.

 doi:10.20944/preprints202404.0847.v1

31. Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S.; et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442* **2022**. https://doi.org/arXiv:2208.10442.
32. Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; He, K. Scaling Language-Image Pre-training via Masking. *arXiv preprint arXiv:2212.00794* **2022**. https://doi.org/arXiv:2212.00794.
33. Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O.K.; Aggarwal, K.; Som, S.; Piao, S.; Wei, F. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. In Proceedings of the Advances in Neural Information Processing Systems; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds. Curran Associates, Inc., 2022, pp. 32897–32912.
34. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* **2022**. https://doi.org/arXiv:2205.01917.
35. Wang, Z.; Yu, J.; Yu, A.W.; Dai, Z.; Tsvetkov, Y.; Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904* **2021**. https://doi.org/arXiv:2108.10904.
36. Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. *arXiv preprint arXiv:2205.12005* **2022**. https://doi.org/arXiv:2205.12005.
37. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* **2018**. https://doi.org/arXiv:1807.03748.
38. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.
39. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision – ECCV 2014; Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T., Eds. Springer International Publishing, 2014, pp. 740–755.
40. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
41. Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Proceedings of the ninth workshop on statistical machine translation, 2014, pp. 376–380.
42. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR 2015, 2015, pp. 4566–4575.
43. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer, 2016, pp. 382–398.
44. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
45. Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; Gao, J. Unified vision-language pre-training for image captioning and vqa. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2020, pp. 13041–13049.
46. Mokady, R.; Hertz, A.; Bermano, A.H. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734* **2021**. https://doi.org/arXiv:2111.09734.
47. Dou, Z.Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18166–18176.
48. Cheng, M.; Sun, Y.; Wang, L.; Zhu, X.; Yao, K.; Chen, J.; Song, G.; Han, J.; Liu, J.; Ding, E.; et al. ViSTA: vision and scene text aggregation for cross-modal retrieval. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5184–5193.
49. Messina, N.; Stefanini, M.; Cornia, M.; Baraldi, L.; Falchi, F.; Amato, G.; Cucchiara, R. ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval. In Proceedings of the Proceedings of the 19th International Conference on Content-based Multimedia Indexing, 2022, pp. 64–70.
50. Diao, Q.; Jiang, Y.; Wen, B.; Sun, J.; Yuan, Z. Metaformer: A unified meta framework for fine-grained recognition. *arXiv preprint arXiv:2203.02751* **2022**. https://doi.org/arXiv:2203.02751.

 doi:10.20944/preprints202404.0847.v1

51. Girdhar, R.; Singh, M.; Ravi, N.; van der Maaten, L.; Joulin, A.; Misra, I. Omnivore: A single model for many visual modalities. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16102–16112.
52. Touvron, H.; Sablayrolles, A.; Douze, M.; Cord, M.; Jégou, H. Grafit: Learning fine-grained image representations with coarse labels. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 874–884.
53. Tian, C.; Wang, W.; Zhu, X.; Dai, J.; Qiao, Y. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV. Springer, 2022, pp. 73–91.
54. Gesmundo, A. A Continual Development Methodology for Large-scale Multitask Dynamic ML Systems. *arXiv preprint arXiv:2209.07326* **2022**. https://doi.org/arXiv:2209.07326.
55. Liu, J.; Huang, X.; Liu, Y.; Li, H. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137* **2022**. https://doi.org/arXiv:2205.13137.
56. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 10347–10357.
57. Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; Wu, W. Incorporating convolution designs into visual transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, 2021, pp. 579–588.
58. Cui, J.; Zhong, Z.; Tian, Z.; Liu, S.; Yu, B.; Jia, J. Generalized Parametric Contrastive Learning. *arXiv preprint arXiv:2209.12400* **2022**. https://doi.org/arXiv:2209.12400.