Image Difference Captioning with Pre-training and Contrastive Learning

Linli Yao, Weiying Wang, Qin Jin*

School of Information, Renmin University of China
{linliyao, wy.wang, qjin}@ruc.edu.cn

Abstract

The Image Difference Captioning (IDC) task aims to describe the visual differences between two similar images with natural language. The major challenges of this task lie in two aspects: 1) **fine-grained** visual differences that require learning stronger vision and language association and 2) **high-cost** of manual annotations that leads to limited supervised data. To address these challenges, we propose a new modeling framework following the pre-training-finetuning paradigm. Specifically, we design three self-supervised tasks and contrastive learning strategies to align visual differences and text descriptions at a fine-grained level. Moreover, we propose a data expansion strategy to utilize extra cross-task supervision information, such as data for fine-grained image classification, to alleviate the limitation of available supervised IDC data. Extensive experiments on two IDC benchmark datasets, CLEVR-Change and Birds-to-Words, demonstrate the effectiveness of the proposed modeling framework. The codes and models will be released at https://github.com/yaolinli/IDC.

Introduction

Endowing machines with the ability to automatically perceive and understand visual information and express in natural language is a goal that researchers have long aspired to achieve. Image captioning (Vinyals et al. 2015; Xu et al. 2015; Rennie et al. 2017), which aims at generating natural language description of a given image, has been one of the classic research tasks. Image Difference Captioning (IDC), which generates natural descriptions of the differences between two similar images, is a further extension of the general image captioning task, and it is more challenging (Jhamtani and Berg-Kirkpatrick 2018; Park, Darrell, and Rohrbach 2019; Tan et al. 2019). IDC has rich potential in real world applications, such as assisting ornithologists to distinguish species with similar appearances, detecting and describing lesions automatically, and reporting salient changes in media assets and surveillance etc.

Intuitively, image difference captioning involves the steps of first perception, then comparison, and finally description, which is more complex and challenging than the general image captioning task. The key challenges of image difference captioning task relate to two aspects. First, the IDC task requires **fine-grained** semantic comprehension. Different from general image captioning that describes a single image, IDC needs to perceive the fine-grained contents of two similar images to identify the usually subtle differences. As shown in Figure 1 (b), the focused differences lie in the tiny body parts of bird species (i.e."feather" and "abdomen"). Moreover, the fine-grained visual differences can be very diverse in different scenarios. In case (a), we focus on the change of geometry objects in the scene, while in case (b), we only attend to the appearance of bird species regardless of the complex natural environment. Second, the annotation for the IDC task is particularly **high-cost**. Compared with general image caption annotation, it requires higher cognition load from annotators, including first observing two images, then comparing the differences, and then using natural language to annotate these differences in order to produce the annotation in a triplet format *(img1, img2, description)*. Therefore, existing manually annotated benchmark datasets are limited in data size (Jhamtani and Berg-Kirkpatrick 2018; Forbes et al. 2019; Tan et al. 2019).

Figure 1: Image difference captioning examples. (a) an example from CLEVR-Change that involves an object change; (b) an example from Birds-to-Words that involves detailed appearance differences of two birds.

*Qin Jin is the corresponding author.

There have been previous endeavors to address the above-mentioned challenges, which mainly focus on designing various attention mechanisms or improving the image features to better capture the subtle image difference based on the typical encoder-decoder structure (Park, Darrell, and Rohrbach 2019; Tan et al. 2019; Shi et al. 2020). However, these works have not paid sufficient attention on fully interacting cross-modal fine-grained representations. Inspired by the recent vision-language pre-training works (Li et al. 2020a; Chen et al. 2020; Li et al. 2020d), in this paper, we propose a new training schema for image difference captioning, which uses self-supervised learning to learn stronger associations between visual differences and language.

Our proposed training schema follows the pre-training and fine-tuning paradigm to align the visual differences with textual semantics at a fine-grained level. In the pre-training stage, we design three self-supervised tasks including Masked Language Modeling (MLM), Masked Visual Contrastive Learning (MVCL) and Fine-grained Difference Aligning (FDA). In MLM and MVCL tasks, we mask some parts of one modality and recover it by the other, thereby promoting the semantic interaction between visual differences and language. We replace the common feature regression objective (Qi et al. 2020) on the visual side with a Noise Contrastive Estimation (NCE) loss. In the FDA task, we introduce contrastive learning strategies and construct negative contrasts to further enhance the fine-grained cross-modal association. Specifically, we carefully design three hard negatives construction strategies, namely Retrieve, Replace, and Confuse. Considering the high annotation cost, we leverage extra cross-task data in our framework to learn additional background knowledge. To be specific, we utilize datasets from general image captioning (GIC) and fine-grained visual classification (FGVC) (Ge, Lin, and Yu 2019; Liu et al. 2020; Dubey et al. 2018). The GIC task provides image-text pairs and can benefit learning the alignment between images and textual descriptions. In the FGVC task, the fine-grained image labels can drive the model to learn more discriminative visual representations. Our model architecture is flexible to handle the different-form cross-task data with an image difference encoder and a multi-layer cross-modal transformer.

Extensive experiments are carried out on two benchmark datasets from different scenarios: CLEVR-Change and Birds-to-Words. Our model significantly outperforms the state-of-the-art methods on the main metrics on both benchmark datasets.The major contributions of this work are as follows:

- We propose a new training schema with the pre-training-finetuning paradigm for the IDC task to better align the visual difference and language by three self-supervised tasks with contrastive learning.
- The proposed model has a flexible structure that can utilize extra cross-task data to alleviate the limitation of supervised data due to high annotation cost.
- The proposed model achieves the state-of-the-art performances on the CLEVR-Change and Birds-to-Words benchmark datasets.

Method

In this section, we introduce our proposed pre-training and fine-tuning paradigm for the image difference captioning task. The overall modeling framework is illustrated in Figure 2. It consists of an image difference encoder to capture subtle visual differences and a multi-layer cross-modal transformer to align cross-modal representations. Three pre-training tasks including Masked Language Modeling (MLM), Masked Visual Contrastive Learning (MVCL) and Fine-grained Difference Aligning (FDA) are designed to take most advantage of the given data. To handle the problem of limited supervised IDC data, we expand cross-task data in our flexible framework.

Model Architecture

Input Representation We define the input of the IDC task, which contains a pair of images and a text description, as $\{V^{(1)}, V^{(2)}, T\}$. For the text description, we tokenize each word in the sentence and convert it to word embedding trained from scratch, denoted as:

$$T = \{[\text{CLS}], [\text{BOS}], w_0, \ldots, w_M, [\text{EOS}]\} \tag{1}$$

where the special token [CLS] is added to capture the global semantics of the sentence. Following previous works (Park, Darrell, and Rohrbach 2019; Tan et al. 2019; Forbes et al. 2019), we use pre-trained ResNet101 (He et al. 2016) to extract grid features for the two images, denoted as:

$$V^{(1)} = \{[\text{IMG1}], v_0^{(1)}, \ldots, v_i^{(1)}, \ldots, v_N^{(1)}\} \tag{2}$$

$$V^{(2)} = \{[\text{IMG2}], v_0^{(2)}, \ldots, v_i^{(2)}, \ldots, v_N^{(2)}\} \tag{3}$$

Similar to text representations, we also add two special tokens [IMG1] and [IMG2] to extract global image semantics respectively. A linear layer is applied to keep the visual feature dimension the same as the word embedding.

To explicitly indicate the positions of tokens inside each modality, we add fixed positional embeddings (Vaswani et al. 2017) to each token as well. In addition, we employ type embeddings to distinguish whether a token belongs to $V^{(1)}$, $V^{(2)}$, or T.

Image Difference Encoder Based on the intuitive cognition process of image difference captioning, observing two images and then comparing them, we design an image difference encoder, which consists of a single-image encoder and a pair-image encoder. The single-image encoder $\mathcal{F}_{\text{sing}}$ takes visual tokens from an image as input to embed the semantics of different regions in the image. Then embeddings of the two images are fed into the pair-image encoder $\mathcal{F}_{\text{pair}}$, which can interact the visual semantics between two images and implicitly learn to locate image differences. We use the transformer architecture for both $\mathcal{F}_{\text{sing}}$ and $\mathcal{F}_{\text{pair}}$.

$$\widetilde{V}^{(1)}, \widetilde{V}^{(2)} = \mathcal{F}_{\text{pair}}\left(\mathcal{F}_{\text{sing}}(V^{(1)}), \mathcal{F}_{\text{sing}}(V^{(2)})\right) \tag{4}$$

Cross-modal Transformer We utilize self-attention based multi-layer transformer for the cross-modal encoder to align context between visual and textual modalities.

$$\hat{V}^{(1)}, \hat{V}^{(2)}, \hat{T} = \mathcal{F}_{\text{cross}}\left(\widetilde{V}^{(1)}, \widetilde{V}^{(2)}, T\right) \tag{5}$$

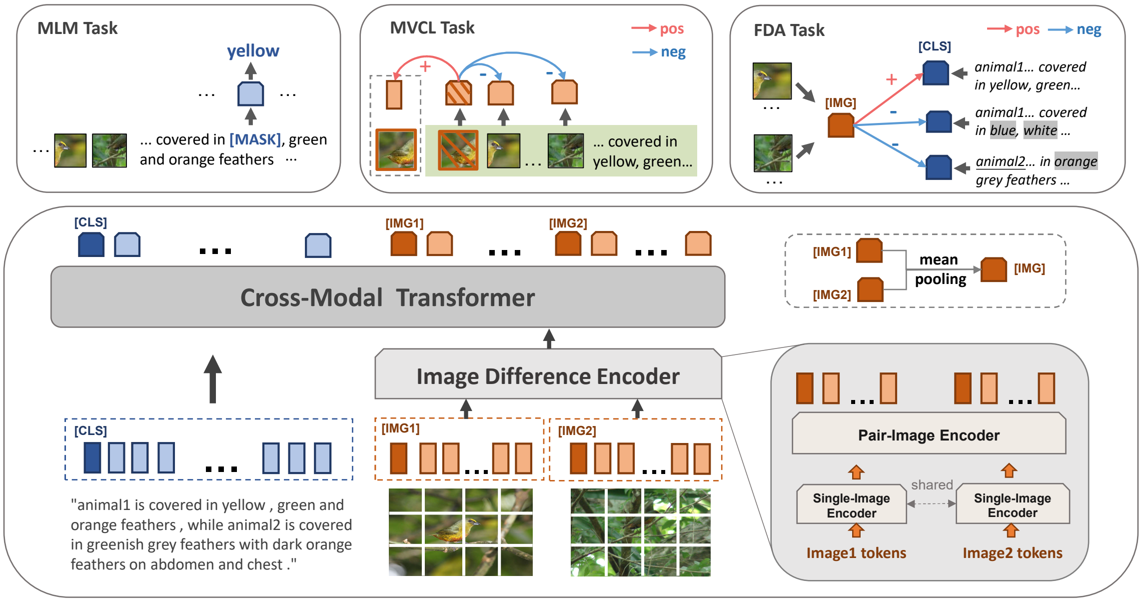

Figure 2: Overview of our proposed framework with an image difference encoder and a multi-layer cross-modal transformer (best viewed in color). Given the input in triplet format *(img1, img2, description)*, the image pair is first fed into the image difference encoder to capture the fine-grained image differences. It consists of a single-image encoder and a pair-image encoder, which enhances the intra-image and inter-image visual representations respectively. Then the enhanced visual representations are aligned with the textual representations in the cross-modal transformer through three pre-training tasks: Masked Language Modeling (MLM), Masked Visual Contrastive Learning (MVCL) and Fine-grained Difference Aligning (FDA).

Pre-training Tasks

We design three pre-training tasks to enhance the fine-grained alignment between image differences and captions so that to learn better feature representations.

Masked Language Modeling (MLM) For textual side, we apply Masked Language Modeling to promote the context mapping from vision to language, following existing VLP works (Chen et al. 2020; Huang et al. 2021b). The objective of the MLM task is to predict the masked word w_m based on the surrounding unmasked words $w_{\backslash m}$ and visual difference context $\{\widetilde{V}^{(1)}, \widetilde{V}^{(2)}\}$. Similar to BERT, we mask 15% of the input word tokens in which 80% are replaced with a special token [MASK], 10% with random words and 10% unchanged. The masked hidden output of cross-modal transformer are fed into a classifier to predict the original words. We formulate the training objective of MLM task as:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}_{V,T \in D} \left[-\log P_\theta \left(w_m \mid w_{\backslash m}, \widetilde{V}^{(1)}, \widetilde{V}^{(2)} \right) \right] \tag{6}$$

where D denotes the entire training set and θ is the model parameters to be learned.

Masked Visual Contrastive Learning (MVCL) Similar to the MLM task, we also apply masking and recovering strategy on the visual side. Since visual representation is continuous and high-dimensional, the goal of the MVCL task is to reconstruct the masked image features according to the difference caption and the remaining visual semantics. Specifically, we mask 15% input image features and replace the masked features with zero vectors. Note that we only mask features from one image each time to ensure the masked content can be recovered by the other two in triplet $\{V^{(1)}, V^{(2)}, T\}$. The general objective of MVCL task is:

$$\mathcal{L}_{\text{MVCL}} = \mathbb{E}_{V,T \in D} f_\theta \left(v_m \mid v_{\backslash m}, T \right) \tag{7}$$

where $V = \{V^{(1)}, V^{(2)}\}$. Inspired by the video language pre-training work (Luo et al. 2020; Li et al. 2020b), we introduce contrastive learning and use a NCE loss (Sun et al. 2019) to define $f_\theta \left(v_m \mid v_{\backslash m}, T \right)$ as:

$$-\log \frac{\exp\left(d(v_m, v_m^+)/\tau_1\right)}{\exp\left(d(v_m, v_m^+)/\tau_1\right) + \sum_{v' \in \mathcal{N}(v_m)} \exp\left(d(v_m, v')/\tau_1\right)} \tag{8}$$

where $d(.)$ denotes the cosine similarity, τ_1 is the temperature hyper-parameter and v_m^+ denotes the original image feature of v_m before masking. We define the unmasked image features in the batch as negative samples $\mathcal{N}(v_m)$. The contrastive loss pushes the model to identify the positive sample v_m^+ of v_m from negative samples $\mathcal{N}(v_m)$ in the batch, which enforces the reconstructed image representations of v_m to be more discriminative.

Fine-grained Difference Aligning (FDA) To explicitly bridge visual and textual modalities in a more fine-grained way, we introduce contrastive learning and construct hard negative samples. To be specific, we rewrite the original difference caption in three ways, as illustrated in Figure 3:

- **Retrieve** For each triplet sample $\{V^{(1)}, V^{(2)}, T\}$, we use TF-IDF similarity to retrieve the most similar difference descriptions $\{T^-\}$ from other samples for T and consider them as hard negative samples.

Figure 3: An example of constructed negative sentences in FDA task on Birds-to-Words Dataset.

- **Replace** We replace the most important difference-related words in the caption to facilitate fine-grained alignment. Empirically, we observe that the attribute words in the caption are more related to the difference (e.g. "grey", "beak"). Therefore, we first annotate adjectives and nouns in each sentence using Stanford CoreNLP tool. Then we sort the annotated words by TF-IDF score to measure their importance. Top K (50%) annotated words are selected and replaced by other words with the same POS-Tags that are randomly sampled from pre-defined vocabularies.
- **Confuse** If we change the subject in the difference description, the semantics of the sentence will totally change while the structure remains similar. For example, changing "*animal1 has a longer tail than animal2*" to "*animal2 has a longer tail than animal1*". We achieve this by changing the subjects between sentences or switching the subject and object within a sentence.

Based on the positive sample (V, T^+) and constructed negative samples (V, T^-), we employ a contrastive loss to define the training objective $\mathcal{L}_{\text{FDA}} = \mathbb{E}_{V,T\in D}\left[-\log \text{NCE}(V,T)\right]$, in which $\text{NCE}(V,T)$ is:

$$\frac{\exp\left(d\left(V,T^+\right)/\tau_2\right)}{\exp\left(d\left(V,T^+\right)/\tau_2\right)+\sum_{T^-\in\mathcal{N}_T}\exp\left(d\left(V,T^-\right)/\tau_2\right)} \tag{9}$$

where $d(.)$ denotes the cosine similarity, τ_2 is a hyperparameter and $\mathcal{N}_T$ is the negative text set. We take the average of special token [IMG1] and [IMG2] from the output as the global visual representation, and the special token [CLS] from the output as the textual representation.

In the pre-training stage, we use bi-directional attention mask, thus the special tokens [IMG1], [IMG2], [CLS] can learn global joint representations from images and text. We train different pre-training tasks by alternating batch with different ratio.

Finetuning and Inference

After pre-training, we fine-tune our model to generate difference captions based on effective visual semantics across two similar images. Similar to (Li et al. 2020d), we adapt the MLM task in the fine-tuning stage. Different from the pre-training stage, we employ a uni-directional attention mask to restrict the attention on the text side, which means that each word can only attend to its previous words. While the attention on the visual side does not change and each word can attend to all visual tokens. In this way, we can keep fine-tuning and pre-training as consistent as possible while adapting the model to sentence generation as much as possible.

During the inference stage, the model generates the difference caption word by word based on visual difference semantics. The embeddings of all visual features $\{V^{(1)}, V^{(1)}\}$ and the special token [CLS] are used as input. The start token [BOS] with a [MASK] token are then fed to the model to trigger the sentence generation. The model samples a word w_0 from the vocabulary based on the likelihood output of the [MASK] token. At step t, the [BOS] token, previous generated words $w_{<t}$ and a new [MASK] token are fed to the model to generate next word w_t. The model can thus generate the whole sentence until [EOS] is sampled.

Expansion of Cross-task Data

To alleviate the limitation of available IDC triplet data, we propose a data expansion strategy to utilize extra cross-task data. Specifically, we expand data from general image captioning and fine-grained visual classification tasks.

General image captioning(GIC) task aims to describe a single image with sentences. The GIC data is in the *(image, text)* format, which can facilitate the model to learn preliminary cross-modal alignment. In order to adapt to the triplet data format of the IDC task, we pad an empty image with zero vectors to form the pseudo triplet. In cross-modal transformer, the padded vectors will not be involved in self-attention calculation by a special attention mask. The pseudo triplet input can naturally adapt to three pre-training tasks. Note that in the MVCL task, we only mask tokens from the real image.

Fine-grained visual classification(FGVC) is a challenging task to classify images with subtle inter-class differences. The slightly different images with different class labels can enhance our image difference encoder to learn more discriminative visual representations. Specifically, we construct image pairs *(img1, img2)* by random sampling, half of which are from the same class label, and the other half are not. Each single image (i.e. *img1* or *img2*) in the obtained pairs is used to refine the single-image encoder $\mathcal{F}_{\text{sing}}$ with a fine-grained classification objective and a contrastive loss (He et al. 2021). Then the dual visual representations are fed to the pair-image encoder $\mathcal{F}_{\text{pair}}$ to optimize it with a matching loss which verifies whether the two images *(img1, img2)* are from the same class. The matching loss enhances the ability of the difference encoder to compare two similar images. Note that we select cross-task datasets with similar semantic domain as the IDC benchmark dataset, so they can provide more similar background knowledge for multi-modal representation learning and cross-modal alignment.

Experiments

We evaluate our model on two IDC benchmark datasets from different domains including CLEVR-Change (Park, Darrell, and Rohrbach 2019) and Birds-to-words (Tan et al. 2019).

Model	B4	M	C(D)	**R**
Neural Naturalist (2019)	22.0	-	25.0	43.0
Relational Speaker (2019)	21.5	22.4	5.8	43.4
DUDA (2019)	23.9	21.9	4.6	44.3
L2C (2021)	31.3	-	15.1	45.3
L2C(+CUB) (2021)	**31.8**	-	16.3	45.6
Ours	28.0	23.1	18.6	48.4
Ours(+Extra Data)	31.0	**23.4**	**25.3**	**49.1**

Table 1: Comparison with state-of-the-art models on Birds-to-Words dataset. B4, M, R, and C(D) are short for BLEU-4, METEOR, ROUGE-L and CIDEr(D). The main metric ROUGE-L on this dataset is highlighted.

Model	B4	M	R	**C**
Capt-Dual-Att (2019)	43.5	32.7	-	108.5
DUDA (2019)	47.3	33.9	-	112.0
VAM (2020)	50.3	37.0	69.7	114.9
VAM+ (2020)	**51.3**	**37.8**	70.4	115.8
IFDC (2021a)	49.2	32.5	69.1	118.7
DUDA+Aux (2021)	51.2	37.7	70.5	115.4
Ours	51.2	36.2	**71.7**	**128.9**

Table 2: Comparison with state-of-the-art models on CLEVR-Change dataset. B4, M, R, and C are short for BLEU-4, METEOR, ROUGE-L and CIDEr. The main metric CIDEr on this dataset is highlighted.

Experimental Settings

Benchmark Datasets CLEVR-Change dataset (Park, Darrell, and Rohrbach 2019) is automatically built via the CLEVR engine and describes scene changes of geometry objects with clean background. It has 67,660, 3,976 and 7,970 image pairs for training, validation and test split respectively. Each image pair is annotated with 6.2 captions on average. Birds-to-Words dataset (Tan et al. 2019) describes the fine-grained difference of various bird species collected in the wild. It has 4,860 image pairs and each pair corresponds to 3.31 annotated captions on average.

Cross-task Datasets CUB (Wah et al. 2011) serves as a single image captioning dataset. It contains 11,788 images of 200 bird species and each image is annotated with 10 captions. We use the split of 8855 training images and 2933 validation images following (Yan et al. 2021). NABirds (Van Horn et al. 2015) is a fine-grained visual classification dataset which contains 48,562 images of North American birds with 555 categories.

Metrics We report results on the standard image captioning metrics including BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), ROUGE-L (Lin 2004) and CIDEr(CIDEr-D) (Vedantam, Lawrence Zitnick, and Parikh 2015) following previous works. In addition, we particularly emphasize the main metric that has been commonly recognized in previous works, which refers to: CIDEr for CLEVR-Change and ROUGE-L for Birds-to-Words.

Implementation Details We extract grid image features in the shape of (7,7,2048) using the pre-trained ResNet101 (He et al. 2016) and flatten it to a feature sequence in the shape of (49, 2048). The word embedding is learned from scratch and its dimension is 512. For the cross-modal transformer, the hidden size is 512, the attention head is 8, and the layer number is 2 for Birds-to-Words and 3 for CLEVR-Change. We set τ_1, τ_2 in contrastive learning to 1. In FDA task, we rewrite 6 negative sentences for each image pair, among which retrieve:replace:confuse=2:2:2. For CLEVR-Change, we sample the batch from the three pre-training tasks with ratio of MLM:MVCL:FDA=8:1:2. We pre-train the model with 8K warm-up steps and 250K iterations in total. For Birds-to-words, the ratio of pre-training tasks is MLM:MVCL:FDA=9:1:2. The warm-up steps are 4K and total training steps are 50K. In the pre-training stage, we apply Adam (Kingma and Ba 2014) optimizer with learning rate 1e-4. In the finetuning stage, the learning rate is set as 3e-5. Early-stop is applied on the main metric to avoid overfitting. The sentence is generated with greedy search in inference. More details can be found in the supplementary material.

For CLEVR-Change, we notice that half of the data samples belong to a *distractor* type, which means that there are only non-semantic differences between the images, e.g. angle, zoom, or illumination changes. Accurately distinguishing semantic changes from distractors has great impact on model performance. We therefore jointly train a distractor judging task in the finetuing stage. Specifically, we concatenate the representation of special token [IMG1] and [IMG2] and feed it to a binary classifier to judge whether the visual change is a distractor or not.

Comparison with the State-of-the-Arts

Results on Birds-to-Words We evaluate our method on Birds-to-Words compared with other state-of-the-art models, including Neural Naturalist (Forbes et al. 2019), Relational Speaker (Tan et al. 2019), DUDA (Park, Darrell, and Rohrbach 2019) and L2C/L2C(+CUB) (Yan et al. 2021). As shown in Table 1, our model without extra cross-task data has achieved significant improvement on the main metric ROUGE-L (from 45.6 to 48.4), even outperforming L2C with extra CUB data. The extra cross-task data further improves the model performance on all metrics, which demonstrates that more data from similar domain can provide beneficial background knowledge. Our model with extra data achieves the new state-of-the-art performance on METEOR, CIDEr-D and ROUGE-L.

Results on CLEVR-Change We compare the proposed model with other state-of-the-art models with typical encoder-decoder structure, including Capt-Dual-Att (Park, Darrell, and Rohrbach 2019), DUDA (Park, Darrell, and Rohrbach 2019), VAM/VAM+ (Shi et al. 2020), IFDC (Huang et al. 2021a) and DUDA+Aux (Hosseinzadeh and Wang 2021). Table 2 shows that our proposed model significantly outperforms all previous models on the main

Model	C	T	M	A	D	DI
DUDA	120.4	86.7	56.4	108.2	103.4	110.8
VAM+	122.1	98.7	82.0	126.3	115.8	122.6
IFDC	**133.2**	99.1	**82.1**	128.2	**118.5**	114.2
Ours	131.2	**101.1**	81.7	**133.3**	116.5	**145.0**

Table 3: Breakdown CIDEr performance on different type of changes of CLEVR-Change Dataset: C(Color), T(Texture), M(Move), A(Add), D(Drop) and DI(Distractor).

Pre-training Tasks	DE	B4	M	R	C
1 None	✓	32.7	27.7	57.2	89.8
2 MLM	✓	36.7	28.2	60.9	94.9
3 MLM + MVCL	✓	50.3	37.6	70.6	119.7
4 MLM + MVCL + FDA	✓	51.2	36.2	71.7	128.9
5 MLM + MVCL + FDA	✗	49.2	35.8	68.8	107.9
6 w/o Distractor Judging	✓	49.8	36.9	69.2	123.5

Table 4: Ablation study results on CLEVR-Change dataset. DE is short for Image Difference Encoder module in our model. B4, M, R, and C are short for BLEU-4, METEOR, ROUGE-L and CIDEr. The main metric CIDEr on this dataset is highlighted.

metric CIDEr, boosting previous best CIDEr score of 118.7 to 128.9. Compared with models employing well-designed attention mechanism or visual encoder, our architecture is more straightforward and flexible which achieves improved performance via effective self-supervised tasks. We also evaluate results on breakdown change types as shown in Table 3. Our model achieves better CIDEr scores on three types including Texture, Add and Distractor due to the fine-grained alignment across modalities in pre-training.

Since CLEVR-Change and Birds-to-Words are from entirely different domains, the experiment results indicate that our model can robustly capture diverse visual differences.

Results and Analysis

Pre-training for IDC We first validate the effectiveness of three self-supervised tasks on CLEVR-Change dataset in Table 4. Line 1 shows the result from the baseline without pre-training, which can be considered as performing the finetuning stage only. The CIDEr score is obviously low in this case without any further interaction across modalities. When more self-supervised tasks are added in the pre-training stage (Line 2∼4), the model performance increases accordingly, which proves the effectiveness of our pretraining-finetuning paradigm. It is worth noting that adding MVCL task brings large performance improvement (CIDEr from 94.9 to 119.7), which benefits from the mask-recover schema on the visual side. The FDA task further improves the CIDEr score from 119.7 to 128.9, which indicates that contrastive learning with well-designed hard negative samples helps fine-grained cross-modal alignment. Furthermore, we analyze the impact of the image difference encoder module in Line 5, where the results show that the performance drops significantly without the image difference encoder. It indicates that the image difference encoder can effectively learn to capture subtle visual differences. Compared to Line 4, Line 6 is the result without distractor judging task in the finetuning stage, which proves that this task can help model distinguish semantic changes from distractors and improve the model performance.

Model	B2W	CUB	NAB	B4	M	C(D)	R
L2C	✓			31.3	-	15.1	45.3
	✓	✓		31.8	-	16.3	45.6
Ours	✓			28.0	23.1	18.6	48.4
	✓	✓		29.3	23.1	23.8	48.5
	✓		✓	27.5	23.3	21.9	48.5
	✓	✓	✓	31.0	23.4	25.3	49.1

Table 5: Model performance on Birds-to-Words(B2W) dataset using two cross-task dataset including CUB and NABirds(NAB). B4, M, R, and C(D) are short for BLEU-4, METEOR, ROUGE-L and CIDEr-D. The main metric ROUGE-L on this dataset is highlighted.

Cross-task Data Usage We evaluate the effectiveness of cross-task data usage on Birds-to-Words in Table 5. CUB is a general image captioning (GIC) dataset that describes bird species, and NABirds is a fine-grained bird image classification dataset. The experiment results show that leveraging CUB can bring more performance improvement because it promotes learning the alignment between images and text descriptions, while NABirds only enhances the visual side. A combination of both cross-task datasets provides more background knowledge and thus achieves the best result.

Qualitative Results We visualize some cases in Figure 4. The top two cases are from CLEVR-Change. (a) The attention maps show that our proposed model can correctly capture the fine-grained semantic changes. (b) Our model can accurately identify the distractor scenario. The bottom cases (c) are from Birds-to-Words, which demonstrate that our model can capture the fine-grained bird appearance difference regardless of the complex background. We also observe that the proposed model tends to repeat the same differences or ignore the conflicts in the generated sentences, which could be further addressed in our future work. More visualizations and analysis are available in the supplementary material.

Related Works

Image Difference Captioning More challenging than general image captioning (Vinyals et al. 2015; Xu et al. 2015; Jiang et al. 2018; Gu et al. 2018; Zhao, Wu, and Zhang 2020; Fei 2021), the image difference captioning task promotes the research of vision and language to achieve a more fine-grained understanding. Recently, several benchmark datasets of IDC have emerged (Jhamtani and Berg-Kirkpatrick 2018; Park, Darrell, and Rohrbach 2019; Tan

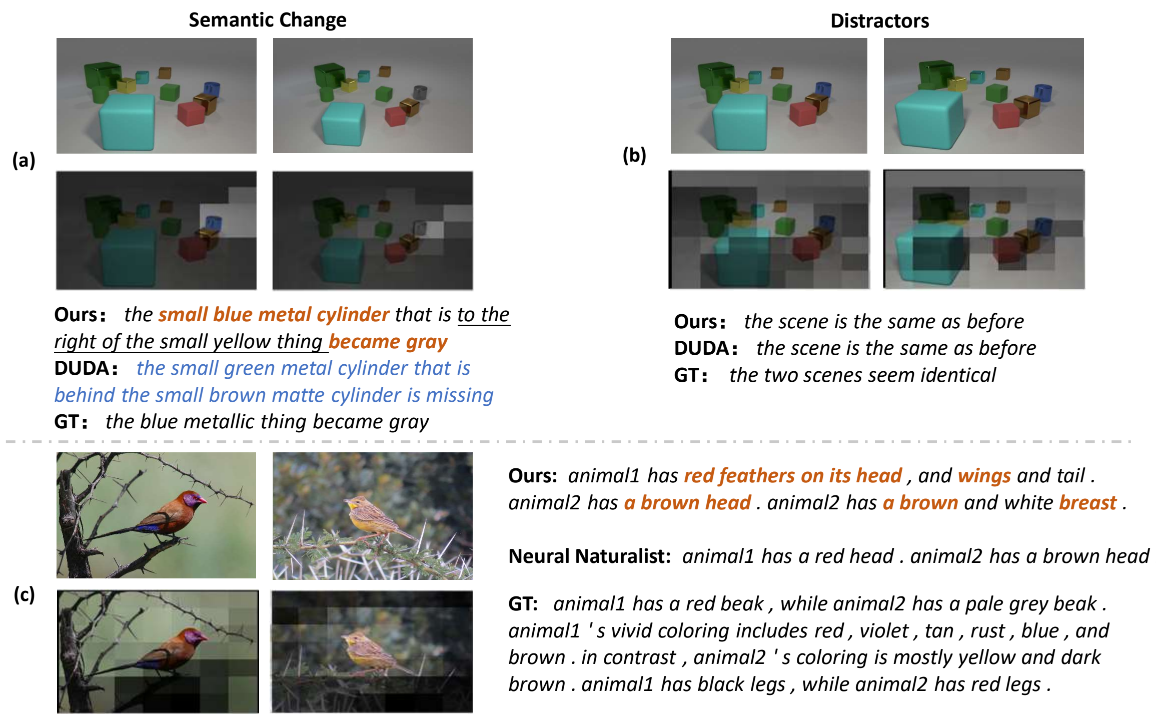

Figure 4: Visualization of generated cases (best viewed in color). The first block illustrates two cases from CLEVR-Change. Case (a) involves semantic changes while case (b) is a distractor with only viewpoint shift. For each case, the first row is the original image pair and the second row is the corresponding attention maps. When the attention score is higher, the region is brighter. Case (c) is from Birds-to-Words and it involves fine-grained bird appearance difference with complex background. Orange bold words are correct generated difference-related words while blue words are wrong.

et al. 2019; Forbes et al. 2019), which are collected from different domains with different focus. However, most of the datasets are limited in scale due to the high annotation cost. Prior works mainly focus on employing well-designed attention mechanisms (Park, Darrell, and Rohrbach 2019; Tan et al. 2019; Shi et al. 2020) or improving visual features (Yan et al. 2021; Huang et al. 2021a) to better capture the subtle visual difference. Besides, (Hosseinzadeh and Wang 2021) use the composed query image retrieval as an auxiliary task to enhance the training process, which can be seen as a self-supervised method. (Yan et al. 2021) employ extra general image captioning data to improve semantic understanding. However, none of these works attend to fully explore the interaction across modalities and take most advantage of the given data. Instead, we propose a new pre-training and fine-tuning schema to align visual and textual representations at a fine-grained level.

Vision-Language Pre-training Recently, Vision and Language Pre-training (VLP) methods (Lu et al. 2019; Chen et al. 2020; Li et al. 2020a,d; Zhou et al. 2020; Kim, Son, and Kim 2021; Huang et al. 2021b; Hu et al. 2021) have shown their success on multi-modal tasks by learning cross-modal representations. However, these pre-trained models are not applicable to the IDC task as they lack the ability of comparing, and the learned representations are too coarse. We follow these works to tailor self-supervised tasks for IDC to enhance the cross-modal alignment and make best use of the given data. Moreover, several VLP related works (Radford et al. 2021; Li et al. 2020c; Huo et al. 2021; Lee et al. 2021) introduce contrastive learning to unify different modal representations. The major idea is to pull closer the distances between positive image-text pairs and push away negative pairs. Motivated by these works, we introduce Contrastive Learning to our pre-training tasks and construct negative samples in different granularity, which enhances more fine-grained cross-modal alignment.

Conclusions

In this paper, we propose a new pretraining-finetuning paradigm for image difference captioning (IDC), which can align visual differences and textual semantics at a fine-grained level. We specifically design three self-supervised tasks with contrastive learning strategies to learn stronger association across modalities. Experiments on CLEVR-Change and Birds-to-Words demonstrate the effectiveness of our proposed pre-training tasks. To address the limitation of supervised IDC data, we explore to utilize two cross-task datasets on Birds-to-Words and prove that they can provide more background knowledge. Our proposed model architecture is flexible to accommodate more data of different forms. In the future work, we will further explore automatic construction of large-scale feasible data to enhance the pre-training stage and improve model generalization ability.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 61772535 and No. 62072462) and Beijing Natural Science Foundation (No. 4192028).

References

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Dubey, A.; Gupta, O.; Guo, P.; Raskar, R.; Farrell, R.; and Naik, N. 2018. Pairwise confusion for fine-grained visual classification. In *Proceedings of the European conference on computer vision (ECCV)*, 70–86.

Fei, Z. 2021. Partially non-autoregressive image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1309–1316.

Forbes, M.; Kaeser-Chen, C.; Sharma, P.; and Belongie, S. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 708–717.

Ge, W.; Lin, X.; and Yu, Y. 2019. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3034–3043.

Gu, J.; Cai, J.; Wang, G.; and Chen, T. 2018. Stack-captioning: Coarse-to-fine learning for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

He, J.; Chen, J.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C.; and Yuille, A. 2021. TransFG: A Transformer Architecture for Fine-grained Recognition. *arXiv preprint arXiv:2103.07976*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hosseinzadeh, M.; and Wang, Y. 2021. Image Change Captioning by Learning From an Auxiliary Task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2725–2734.

Hu, X.; Yin, X.; Lin, K.; Zhang, L.; Gao, J.; Wang, L.; and Liu, Z. 2021. VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1575–1583.

Huang, Q.; Liang, Y.; Wei, J.; Yi, C.; Liang, H.; Leung, H.-f.; and Li, Q. 2021a. Image Difference Captioning with Instance-Level Fine-Grained Feature Representation. *IEEE Transactions on Multimedia*, 1–1.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021b. Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12976–12985.

Huo, Y.; Zhang, M.; Liu, G.; Lu, H.; Gao, Y.; Yang, G.; Wen, J.; Zhang, H.; Xu, B.; Zheng, W.; et al. 2021. WenLan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*.

Jhamtani, H.; and Berg-Kirkpatrick, T. 2018. Learning to Describe Differences Between Pairs of Similar Images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4024–4034.

Jiang, W.; Ma, L.; Chen, X.; Zhang, H.; and Liu, W. 2018. Learning to guide decoding for image captioning. In *Thirty-second AAAI conference on artificial intelligence*.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, H.; Yoon, S.; Dernoncourt, F.; Bui, T.; and Jung, K. 2021. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. *arXiv preprint arXiv:2106.14019*.

Li, G.; Duan, N.; Fang, Y.; Gong, M.; and Jiang, D. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11336–11344.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020b. Hero: Hierarchical Encoder for Video+ Language Omni-representation Pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2046–2065.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2020c. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. *arXiv preprint arXiv:2012.15409*.

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

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, C.; Xie, H.; Zha, Z.-J.; Ma, L.; Yu, L.; and Zhang, Y. 2020. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11555–11562.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 13–23.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. *arXiv preprint arXiv:2002.06353*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *Proceedings of the IEEE international conference on computer vision*, 4624–4633.

Qi, D.; Su, L.; Song, J.; Cui, E.; Bharti, T.; and Sacheti, A. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7008–7024.

Shi, X.; Yang, X.; Gu, J.; Joty, S.; and Cai, J. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *European Conference on Computer Vision*, 574–590. Springer.

Sun, C.; Baradel, F.; Murphy, K.; and Schmid, C. 2019. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*.

Tan, H.; Dernoncourt, F.; Lin, Z.; Bui, T.; and Bansal, M. 2019. Expressing Visual Relationships via Language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1873–1883.

Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. J. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

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

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

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

Yan, A.; Wang, X.; Fu, T.-J.; and Wang, W. Y. 2021. L2C: Describing Visual Differences Needs Semantic Understanding of Individuals. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2315–2320.

Zhao, W.; Wu, X.; and Zhang, X. 2020. Memcap: Memorizing style knowledge for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12984–12992.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13041–13049.