Exploiting Cross-Modal Prediction and Relation Consistency for Semisupervised Image Captioning

Yang Yang, Hongchen Wei, Hengshu Zhu, Senior Member, IEEE, Dianhai Yu, Hui Xiong, Fellow, IEEE, and Jian Yang, Member, IEEE

Abstract—The task of image captioning aims to generate captions directly from images via the automatically learned cross-modal generator. To build a well-performing generator, existing approaches usually need a large number of described images (i.e., supervised image-sentence pairs), requiring a huge effort on manual labeling. However, in real-world applications, a more general scenario is that we only have limited amount of described images and a large number of undescribed images. Therefore, a resulting challenge is how to effectively combine the undescribed images into the learning of cross-modal generator (i.e., semisupervised image captioning). To solve this problem, we propose a novel image captioning method by exploiting the cross-modal prediction and relation consistency (CPRC), which aims to utilize the raw image input to constrain the generated sentence in the semantic space. In detail, considering that the heterogeneous gap between modalities always leads to the supervision difficulty while using the global embedding directly, CPRC turns to transform both the raw image and corresponding generated sentence into the shared semantic space, and measure the generated sentence from two aspects: 1) prediction consistency: CPRC utilizes the prediction of raw image as soft label to distill useful supervision for the generated sentence, rather than employing the traditional pseudo labeling and 2) relation consistency: CPRC develops a novel relation consistency between augmented images and corresponding generated sentences to retain the important relational knowledge. In result, CPRC supervises the generated sentence from both the informativeness and representativeness perspectives, and can reasonably use the undescribed images to learn a more effective generator under the semisupervised scenario. The experiments show that our method outperforms state-of-the-art comparison methods on the MS-COCO “Karpathy” offline test split under complex nonparallel scenarios, for example, CPRC achieves at least 6% improvements on the CIDEr-D score.

Index Terms—Cross-modal learning, image captioning, relation consistency, semisupervised learning.

I. INTRODUCTION

In real-world applications, object can always be represented by multiple source information, that is, multiple modalities [1], [2]. For example, the news always contains image and text information, the video can be divided into image, audio, and text modalities. Along this line, the study of cross-modal learning has emerged for bridging the connections among different modalities, so as to better perform downstream tasks, in which the image captioning is one of the important research directions. Specifically, image captioning aims to automatically generate natural language descriptions for images, and has emerged as a prominent research problem in both academia and industry [3]–[6]. For example, we can automatically broadcast road conditions by learning visual images to assist driving, and can also help visually impaired users to read more conveniently. In fact, the challenge of image captioning is to learn the generator between two heterogeneous modalities (i.e., the image and text modalities), which needs to recognize salient objects in an image using computer vision techniques and generate coherent descriptions using natural language processing.

To solve this problem, researchers first explore the neural encoder–decoder models [3], [7], which are composed of a CNN encoder and an LSTM (or transformer) decoder. In detail, these methods first encode the image into a set of feature vectors using a CNN based model, with each segmentation captures semantic information about an image region, then sequentially decode these feature vectors to words via an LSTM-based or transformer-based network. Furthermore, Xu et al. [4], Lu et al. [8], and Huang et al. [9] adopted the single or hierarchical attention mechanism that enables the model to focus on particular image regions during decoding process. To mitigate the incorrect or repetitive content,
The key difficulty of semisupervised image captioning is to design the pseudo supervision for the generated sentences. Actually, there have been some preliminary attempts recently. For example, [11], [12] propose unsupervised captioning methods, which combine the adversarial learning [13] with traditional encoder–decoder models to evaluate the quality of generated sentences. In detail, based on the traditional encoder–decoder models, these approaches employ adversarial training to generate sentences indistinguishable from the sentences within auxiliary corpus. In order to ensure that the generated captions contain the visual concepts, they particularly distill the knowledge provided by a visual concept detector into the image captioning model. However, the domain discriminator and visual concept distiller do not fundamentally evaluate the matching degree and structural rationality of the generated sentence, so the captioning performance is poor. As for semisupervised image captioning, a straightforward way is utilizing the undescribed images together with their machine-generated sentences directly [14], [15] as the pseudo image–sentence pair to train the model. However, limited amount of parallel data can hardly establish a proper initial generator to generate precisely pseudo descriptions, which may have negative affection to the training of mapping function.

To circumvent these issues, we attempt to utilize the raw image as pseudo supervision. However, heterogeneous gap between modalities always leads the supervision difficulty if we directly constrain the consistency between global embedding of image and sentence. Thereby, we switch to use the broader and more effective semantic prediction information, rather than directly utilize the embedding, and introduce a novel approach, dubbed semisupervised image captioning by exploiting the cross-modal prediction and relation consistency (CPRC). In detail, there are two common approaches in traditional semisupervised learning: 1) pseudo labeling: it minimizes the entropy of unlabeled data using predictions and 2) consistency regularization: it transforms the unlabeled raw images using data augmentation techniques, then constrains the consistency of transformed instances’ outputs. Different form these two techniques, we design CPRC by comprehensively considering the informativeness and representativeness: 1) prediction consistency: we utilize the soft label of image to distill effective supervision for generated sentence and 2) relation consistency: we work on constraining the generated sentences and the augmented image inputs to have similar relational distributions. Consequently, CPRC can effectively qualify the generated sentences from perspectives of both the prediction confidence and distribution alignment, thereby to learn a more effective mapping function. Note that CPRC can be implemented with any current captioning model, and we adopt several typical approaches for verification [16], [17]. The source code is available at https://github.com/njustkmg/CPRC.

In summary, the contributions in this article can be summarized as follows.

1) We propose a novel semisupervised image captioning framework for processing undescribed images, which is universal for any captioning model.
2) We design the CPRC to measure the undescribed images, which maps the raw image and corresponding generated sentence into the shared semantic space, and supervises the generated sentence by distilling the soft label from image prediction and constraining the cross-modal relation consistency.
3) In experiments, our approach improves the performance under the semisupervised scenario, which validates that knowledge hidden in the content and relation is effective for enhancing the generator.

II. RELATED WORK

A. Image Captioning

Image captioning approaches can be roughly divided into three categories as follows.

1) Template-based methods, which generate slotted captioning templates manually, and then utilize the detected keywords to fill the templates [18], but their expressive power is limited because of the need for designing templates manually.
2) Encoder–decoder-based methods, which are inspired by the neural machine translation [19]. For example, Vinyals et al. [20] proposed an end-to-end framework with a CNN encoding the image to feature vector and an LSTM decoding to caption; Huang et al. [9] added an attention-on-attention module after both the LSTM...
and the attention mechanism, which can measure the relevance between attention result and query.

3) Editing-based methods, which consider editing inputs independent from generating inputs. For example, Hashimoto et al. [10] learned a retrieval model that embeds the input in a task-dependent way for code generation; Sammani and Elsayed [5] introduced a framework that learns to modify existing captions from a given framework by modeling the residual information.

However, all these methods need huge amount of supervised image–sentence pairs for training, whereas the scenario with a large amount of undescribed images is more general in real applications. To handle the undescribed images, several attempts propose unsupervised image captioning approaches. Feng et al. [11] distilled the knowledge in visual concept detector into the captioning model to recognize the visual concepts, and adopted sentence corpus to teach the captioning model; Gu et al. [12] developed an unsupervised feature alignment method with adversarial learning that maps the scene graph features from the image to sentence modality. Nevertheless, these methods mainly depend on employing the domain discriminator for learning plausible sentences, an thus are difficult for generating matched sentences. On the other hand, considering the semisupervised image captioning, Mithun et al. [14] and Huang et al. [15] proposed to extract regional semantics from un-annotated images as additional weak supervision to learn visual-semantic embeddings. However, the generated pseudo sentences are always unqualified to train the generator in real experiments.

B. Semisupervised Learning

Recently, deep networks achieve strong performance by supervised learning, which requires a large number of labeled data. However, it comes at a significant cost when labeling by human labor, especially by domain experts. To this end, semisupervised learning, which permits harnessing the large amounts of unlabeled data in combination with typically smaller sets of labeled data, attracts more and more attention. Existing semisupervised learning mainly considers two aspects as follows.

1) Self-Training [21]: The generality of self-training is to use a model’s predictions to obtain artificial labels for unlabeled data. A specific variant is the pseudo labeling, which converts the model predictions of unlabeled data to hard labels for calculating the cross entropy. Besides, pseudo labeling is often used along with a confidence thresholding that retains sufficiently confident unlabeled instances. In result, pseudo labeling results in entropy minimization, has been used as a component for many semisupervised algorithms, and been validated to produce better results [22].

2) Consistency Regularization [23]: Early extensions include exponential moving average of model parameters [24] or previous model checkpoints [25]. Recently, data augmentation, which integrates these techniques into the self-training framework, has shown better results [26], [27]. A mainstream technology is to produce random perturbations with data augmentation [28], and then enforce consistency between the augmentations. For example, Xie et al. [26] proposed unsupervised data augmentation with distribution alignment and augmentation anchoring, which encourages each output close to the weakly augmented version of the same input; Berthelot et al. [27] used a weakly augmented example to generate an artificial label and enforce consistency against strongly augmented example. Furthermore, Sohn et al. [29] combined the pseudo labeling and consistency regularization into a unified framework, which generates pseudo labels using the model’s predictions on weakly augmented unlabeled images, and constrains the prediction consistency between weakly augmented and strongly augmented versions. Note that the targets in previous semisupervised methods are uniform and simple, that is, the label ground truths. However, cross-modal semisupervised learning is more complicated, that is, each image has the corresponding sentence and label ground truth. It is more difficult to build cross-modal generator than single modal classifier with limited supervised data, thereby it may causes noise accumulation if we directly employ the traditional semisupervised technique for the generated sentences.

The remainder of this article is organized as follows. Section III presents the proposed method, including the model, solution, and extension. Section IV shows the experimental results under different semisupervised settings. Section V concludes this article.

III. PROPOSED METHOD

A. Notations

Without any loss of generality, we define the semisupervised image–sentence set as: $D = \{\{v_i, w_i, y_i\}_{i=1}^{N_l}, \{v_j\}_{j=1}^{N_u}\}$, where $v_i \in \mathcal{R}^{d_v}$ denotes the $i$th image instance, $w_i \in \mathcal{R}^{d_w}$ represents the aligned sentence instance, $y_i \in \mathcal{R}^{C}$ denotes the instance label, $y_{i,k} = 1$ if the $i$th instance belongs to the $k$th label, otherwise is 0. $v_j$ is the $j$th undescribed image. $N_l$ and $N_u$ ($N_l \ll N_u$) are the numbers of described and undescribed instances, respectively.

Definition 1 (Semisupervised Image Captioning): Given limited parallel image–sentence pairs $\{v_i, w_i, y_i\}_{i=1}^{N_l}$ and a huge number of undescribed images $\{v_j\}_{j=1}^{N_u}$, we aim to construct a generator $G$ for image captioning by reliably utilizing the undescribed images.

B. Framework

It is notable that CPRC focuses on employing the undescribed images, and is a general semisupervised framework. Thereby the image–sentence generator, that is, $G : v \rightarrow w$, can be represented as any state-of-the-art captioning model. In this article, considering the effectiveness and reproducibility, we adopt the attention model, that is, AoANet [9], for $G$ as the base model. As shown in Fig. 2, AoANet applies an attention on attention (AoA) module to the image encoder and the caption decoder. AoA first generates an “information
Fig. 2. Diagram of the proposed CPRC. For example, three weakly augmented images and the raw image are fed into the generator, which can generate four corresponding sentences. Without any loss of generality, we adopt the AoANet [9] as the generator. Then, the embeddings of image inputs and generated sentences are fed into the shared classifier to obtain the predictions. The model is trained by considering two objectives: 1) \textit{supervised loss} includes the generation and prediction losses for described images. In detail, generation loss measures the quality of generated sentence sequence, and prediction loss considers the multilabel prediction of generated sentence and 2) \textit{unsupervised loss} includes the prediction consistency and relation consistency for undescribed images. In detail, prediction consistency utilizes the image’s prediction as pseudo labels for corresponding generated sentence, and relation consistency constrains the generated sentences’ distribution with image inputs’ distribution.

vector” (i.e., \(I\)) and an “attention gate” (i.e., \(P\)) with two linear transformations. The information vector is derived from the current context (i.e., the query \(Q\)) and the attention result (i.e., \(\hat{V}\)) via a linear transformation, and stores the newly obtained information from the attention result together with the information from the current context. The attention gate is also derived from the query and the attention result via another linear transformation with sigmoid activation followed. Subsequently, AoA adds another attention by applying the attention gate to the information vector using element-wise multiplication and finally obtains the “attended information” (i.e., \(\hat{I}\)). Details can refer to [9]. Note that the CPRC is a general framework, which can apply to any existing captioning model. AoANet only trains with the generation loss, which has not made effective use of category labels. Thereby, AoANet has no label predictor, and can not employ the undescribed images.

The learning process of CPRC is shown in Fig. 2. Specifically, CPRC first samples a minibatch of images from the dataset \(D\) (including described and undescribed images), and adopts the data augmentation techniques for each undescribed image (i.e., each image has \(K\) variants). Then we can acquire the generated sentences for both augmented images and the raw image using the \(G\), and compute the predictions for image inputs and generated sentences using the shared prediction classifier \(f\). The model is trained through two main objects:

1) \textit{supervised loss}, which is designed for described images, that is, supervised image–sentence pairs. In detail, supervised loss considers both the label and sentence predictions, including: a) \textit{generation loss}, which employs the cross-entropy loss or reinforcement learning-based reward [16] of generated sentence and ground truth sentence and b) \textit{prediction loss}, which calculates the multilabel loss between image/sentence’s prediction and label ground truth.

2) \textit{unsupervised loss}, which is designed for undescribed images. In detail, unsupervised loss considers both the informativeness and representativeness: a) \textit{prediction consistency}, which uses the image’s prediction as pseudo label to distill effective information for generated sentence, so as to measure the instance’s informativeness and b) \textit{relation consistency}, which adopts the relational structure of the augmented images as the supervision distribution for generated sentences, so as to measure the instance’s representativeness. Therefore, in addition
to the traditional loss for described images, we constrain the sentences generated from undescribed images by comprehensively using the raw image inputs. The details are described as follows.

C. Supervised Loss

1) Generation Loss: Given an image \( \mathbf{v} \), the decoder generates a sequence of sentence \( \mathbf{w} = \{w_1, w_2, \ldots, w_T\} \) describing the image, \( T \) is the length of sentence. Then, we can minimize the cross-entropy loss (i.e., \( \ell_{XE} \)) or maximize a reinforcement learning-based reward \cite{13} (i.e., \( \ell_{RL} \)), according to ground truth caption \( \mathbf{w} \)

\[
\ell_{XE} = - \sum_{t=1}^{T} \log p(w_t | w_{1:t-1})
\]

\[
\ell_{RL} = - \mathbb{E}_{w_{1:T}}[p(r(w_{1:T}))]
\]

where \( w_{1:T} \) denotes the target ground truth sequence, \( p(\cdot) \) is the prediction probability. The reward \( r(\cdot) \) is a sentence-level metric for the sampled sentence and the ground truth, which always uses the score of some metric (e.g., CIDEr-D \cite{30}). In detail, as introduced in \cite{13}, captioning approaches traditionally train the models using the cross-entropy loss. On the other hand, to directly optimize NLP metrics and address the exposure bias issue, \cite{13} casts the generative models in the Reinforcement Learning terminology as \cite{31}. In detail, traditional decoder (i.e., LSTM) can be viewed as an “agent” that interacts with the “environment” (i.e., words and image features). The parameters of the network define a policy, that results in an “action” (i.e., the prediction of the next word). After each action, the agent updates its internal “state” (i.e., parameters of the LSTM, attention weights etc). Upon generating the end-of-sequence (EOS) token, the agent observes a “reward” that is, for example, the CIDEr-D score of the generated sentence.

2) Prediction Loss: On the other hand, we can measure the generation with classification task using label ground truth \( \mathbf{y} \). We extract the embeddings of image input and generated sentence from the representation output layer. Considering that the image and corresponding sentence share the same semantic representations, the embeddings of image input and generated sentence can be further put into the shared classifier \( f \) for predicting. Thereby, the forward prediction process can be represented as

\[
\mathbf{p}^y = f(\mathcal{E}_d(\mathbf{v})), \quad \mathbf{p}^w = f(\mathcal{D}_e(\mathcal{E}_e(\mathbf{v})))
\]

where \( \mathbf{p}^y \) and \( \mathbf{p}^w \) are normalized prediction distribution of image input and generated sentence. \( f(\cdot) \) denotes the shared classification model for text and image modalities. Without any loss of generality, we utilize a network with three fully connected layers here. \( \mathcal{E}_d \) denotes the encoder, \( \mathcal{D}_e \) represents the decoder. \( \mathcal{E}_e(\mathbf{v}), \mathcal{D}_e(\mathcal{E}_e(\mathbf{v})) \in \mathbb{R}^d \) represents the embeddings of image input and generated sentence. Note that \( \mathcal{E}_e(\mathbf{v}) \) and \( \mathcal{D}_e(\mathcal{E}_e(\mathbf{v})) \) are the final embeddings of image/text region embedding with mean(\( \cdot \)) operator. The commonly used image captioning dataset (i.e., the COCO dataset) is a multilabel dataset, that is, different from the multiclass dataset that each instance only has one ground truth, each instance has multiple labels. Therefore, we utilize the binary cross-entropy loss (BCELoss) here

\[
\ell_p = \sum_{m \in \{v, w\}} H(p^m, y^m)
\]

\[
H(p^m, y^m) = - \sum_{j} (y_j^m \log p_j^m + (1 - y_j^m) \log (1 - p_j^m))
\]

where \( H(\cdot) \) denotes the BCELoss for multilabel prediction, and the model’s predictions are encouraged to be low-entropy on supervised data.

D. Unsupervised Loss

1) Prediction Consistency: First, we introduce the augmentation technique for transforming the images. Existing methods usually leverage two kinds of augmentations: 1) weak augmentation is a standard flip-and-shift strategy, which does not significantly change the content of the input and 2) strong augmentation always refers to the AutoAugment \cite{32} and its variant, which uses reinforcement learning to find an augmentation strategy comprising transformations from the Python Imaging Library.\footnote{https://www.pythonware.com/products/pil/} Considering that “strong” augmented (i.e., heavily augmented) instances are almost certainly outside the data distribution, which leads to the low quality of generated sentence, we leverage the “weak” augmentation instead. In result, each image can be expanded to \( K + 1 \) variants, that is, \( \Psi(\mathbf{v}) = \{\mathbf{v}, \mathbf{v}_1, \ldots, \mathbf{v}_K\} \), \( \mathbf{v}_0 \) denotes the raw input.

Then, we input the augmented image set to the image–sentence generator \( G \), and extract the embeddings of generated sentences from the representation output layer. The embeddings are further put into the shared classifier for prediction. Thereby, the prediction process can be represented as

\[
\mathbf{p}^y_k = f(\mathcal{D}_e(\mathcal{E}_e(\mathbf{v}_k))), \quad k \in \{0, 1, \ldots, K\}.
\]

Similarly, we can acquire the prediction of image inputs: \( \mathbf{p}^w_k = f(\mathcal{E}_e(\mathbf{v}_k)), \quad k \in \{0, 1, \ldots, K\} \). Considering that the commonly used image captioning datasets are multilabel datasets, traditional pseudo labeling that leverages “hard” labels (i.e., the arg max of model’s output) is inappropriate, because it is difficult to determine the number of hard labels for each instance. As a consequence, we directly utilize the prediction of image for knowledge distillation \cite{33}

\[
\ell_{pc} = \sum_{k \in \{0, 1, \ldots, K\}} H(p^y_k, p^w_k)
\]

\[
H(p^y_k, p^w_k) = - \sum_{j} (p^y_{kj} \log p^w_{kj} + (1 - p^y_{kj}) \log (1 - p^w_{kj}))
\]

where \( H(\cdot) \) denotes the BCELoss.

2) Relation Consistency: Inspired by the linguistic structuralism \cite{34} that relations can better present the knowledge than individual example, the primary information actually lies in the structure of the data space. Therefore, we define a new relation consistency loss \( \ell_{rc} \) using a metric learning-based constraint, which calculates the KL divergence of the similarity vectors between the image inputs and generated sentences. The relation consistency aims to ensure the structural knowledge
using mutual relations of data examples in the raw inputs. Specifically, each image input can be denoted as a bag of \( K + 1 \) instances, that is, \( \Psi(v) \), while the corresponding generated sentences can also be represented as a bag of instances, that is, \( G(\Psi(v)) \). With the shared classifier, the image and sentence prediction can be formulated as

\[
\begin{align*}
    p_k^v &= f(E_v(v_k)), \quad k \in \{0, 1, \ldots, K\} \\
    p_k^w &= f(D_v(v_k)), \quad k \in \{0, 1, \ldots, K\}.
\end{align*}
\]

With the predictions of image inputs and generated sentences, the objective of relation consistency can be formulated as

\[
\ell_{rc} = KL(\Phi(p_0^v, p_1^v, \ldots, p_K^v), \Phi(p_0^w, p_1^w, \ldots, p_K^w)).
\] (5)

KL(a, b) = \( a \log(a/b) \) is the KL divergence that penalizes difference between the similarity distributions of image inputs and generated sentences. \( \Phi \) is a relation function, which measures a relation energy of the given tuple. In detail, \( \Phi \) aims to measure the similarities formed by the examples in semantic prediction space

\[
\begin{align*}
    \Phi(p_0^v, p_1^v, \ldots, p_K^v) &= [q_{mn}^v] \quad m, n \in \{0, 1, \ldots, K\} \\
    \Phi(p_0^w, p_1^w, \ldots, p_K^w) &= [q_{mn}^w] \quad m, n \in \{0, 1, \ldots, K\}
\end{align*}
\]

\[
\begin{align*}
    q_{mn}^v &= \frac{\exp(d_{mn}^v)}{\sum \exp(d_{ij}^v)} \\
    q_{mn}^w &= \frac{\exp(d_{mn}^w)}{\sum \exp(d_{ij}^w)}
\end{align*}
\] (6)

where \( d_{mn}^v = -\text{Dist}(p_m^v, p_n^v), d_{mn}^w = -\text{Dist}(p_m^w, p_n^w) \) measures the distance between \( (p_m^v, p_n^v) \), \( (p_m^w, p_n^w) \), respectively. \( \text{Dist}(p_m^v, p_n^v) = \|p_m^v - p_n^v\|_2 \) and \( \text{Dist}(p_m^w, p_n^w) = \|p_m^w - p_n^w\|_2 \). \( q_{mn}^v \) and \( q_{mn}^w \) denote the relative instance-wise similarity. Finally, we pull the \([q_{mn}^v]\) and \([q_{mn}^w]\) into vector form. In result, the relation consistency loss can deliver the relationships of examples by penalizing structure differences. Since the structure has higher order properties than single output, it can transfer knowledge more effectively, and is more suitable for consistency measure.

E. Overall Function

In summary, with the limited amount of parallel image-sentence pairs and large amount of undescribed images, we define the total loss by combining (1), (2), (4), and (5)

\[
L = \sum_{i=1}^{N_l} \ell_s(v_i, w_i, y_i) + \sum_{j=1}^{N_u} \left( \lambda_1 \ell_{pc}(v_j) + \lambda_2 \ell_{rc}(v_j) \right)
\] (7)

where \( \ell_s \) denotes the captioning loss, which can be adopted as \( \ell_{XE} \) or \( \ell_{RL} \) in (1). Note that \( \ell_c \) and \( \ell_p \) are with same order of magnitude, so we do not add hyperparameter here. \( \lambda_1 \) and \( \lambda_2 \) are scale values that control the weights of different losses. In \( \ell_s \), we use labeled images and sentences to jointly train the shared classifier \( f \), which increases the amount of training data, as well as adjust the classifier to better suit subsequent prediction of augmented images and generated sentences. Furthermore, considering that the pseudo labels \( p^s \) and \( p^w \) may exist noises, we can also adopt a confidence threshold that retains confident instances. Equation (7) can be reformulated as

\[
L = \sum_{i=1}^{N_l} \ell_s(v_i, w_i, y_i) + \sum_{j=1}^{N_u} \left( \lambda_1 \ell_{pc}(v_j) + \lambda_2 \ell_{rc}(v_j) \right)
\]

\[
\{ \lambda_1 \ell_{pc}(v_j) + \lambda_2 \ell_{rc}(v_j) \} \geq \tau
\]

\[
\ell_s(v_i, w_i, y_i) = \ell_s(v_i, w_i) + \ell_p(v_i, w_i, y_i)
\] (8)

where \( p_{ij}^s \) denotes the prediction probability of the \( j \)-th raw image input, and \( \tau \) is a scalar hyperparameter denoting the threshold above which we retain the generated sentences. \( \mathbf{1}(\cdot) \) denotes the indicator function. Details are shown in Algorithm 1. In each epoch, we randomly sample minibatch data containing described and undescribed images, then utilize the data in the minibatch to calculate both the supervised and unsupervised losses. Finally, we calculate the gradient using overall loss for back propagation. Loop the entire process until reaching the defined epochs.

---

**Algorithm 1 Code of CPRC**

**Input:**
- Data: \( D = \{\{v_i, w_i, y_i\}_{i=1}^{N_l}, \{v_j\}_{j=1}^{N_u}\} \)
- Parameters: \( \lambda_1, \lambda_2, \tau \), epoch number \( T \)

**Output:**
- Image captioning mapping function: \( G \)

1. Initialize the \( G \) and \( f \) randomly;
2. for \( t = 1:T \) do
3. for sample mini-batch \( B_{i,k} \) from \( D \) do
4. Calculate supervised loss \( \ell_s \) with described images according to Equation 1 or Equation 2;
5. Calculate prediction consistency \( \ell_{pc} \) with undescribed images according to Equation 4;
6. Calculate relation consistency \( \ell_{rc} \) with undescribed images according to Equation 5;
7. Calculate overall loss \( L \) with \( \ell_s, \ell_{pc} \) and \( \ell_{rc} \) according to Equation 8;
8. Update model parameters of \( G, f \) using SGD;
9. end for
10. end for
IV. EXPERIMENTS

A. Datasets

We adopt the popular MS-COCO dataset [35] as mostly former related methods [9], [16], [17], [36], [37]. The MS-COCO dataset contains 123,287 images (82,783 training images and 40,504 validation images), each labeled with five captions. The popular test sets are divided into two categories: 1) online evaluation and 2) offline evaluation. Considering that all methods are evaluated under semisupervised scenario, online evaluation cannot be used, so we only use offline evaluation. The offline Karpathy data split [38] contains 5000 images for validation, 5000 images for testing, and the rest for training. To construct the semisupervised scenario, we randomly selected examples with artificially set proportions as supervised data from the training set, and the rest are unsupervised data.

B. Implementation Details

The target of CPRC is to train the generator $G$. In detail, we employ the AoANet [9] structure for $G$ as base model. Meanwhile, we adopt fully connected networks for $f$ with three fully connected layers (with 1024 dimension for the hidden layers). The dimension of original image vectors is 2048 and we project them to a new space with the dimension of 1024 following [9]. The $K = 3$, that is, each image has three augmentations using random occlusion technique. As for the training process, we train AoANet for 40 epochs with a minibatch size of 16, and the ADAM [39] optimizer is used with a learning rate initialized by $10^{-4}$ and annealed by 0.8 every three epochs. The parameters $\lambda_1$ and $\lambda_2$ are tuned in $\{0.01, 0.1, 1, 10\}$, and $\tau = 0.1$. The entire network is trained on an Nvidia TITAN X GPU.

C. Baselines and Evaluation Protocol

The comparison models fall into three categories: 1) state-of-the-art supervised captioning methods: SCST [16], AoANet [9], AAT [36], ORT [37], GIC [17], Anchor [40] and RSTNet [41]. Note that these methods can only utilize the supervised image–sentence pairs; 2) unsupervised captioning methods: Graph-align [12] and UIC [11]. These approaches utilize the independent image set and corpus set for training; and 3) state-of-the-art semisupervised method: A3VSE [15].

Moreover, we conduct extra ablation studies to evaluate each term in our proposed CPRC: 1) AoANet+P, we combine the label prediction consistency with the original AoANet generation loss as multitask loss (only using the supervised data); 2) AoANet+c, we combine the relation consistency loss with the original AoANet generation loss as multitask loss (only using the supervised data); 3) PL, we replace the prediction consistency with pseudo labeling as traditional semisupervised methods; 4) AC, we replace the relation consistency with augmentation consistency as traditional semisupervised methods; 5) Embedding+, we replace the relation consistency loss with embedding consistency loss, which minimizes the difference between the embedding of image inputs and generated sentences; 6) Semantic+, we replace the relation consistency loss with prediction consistency loss, which minimizes the difference between the predictions of image inputs and generated sentences; 7) Strong+, we replace the weak augmentation with strong augmentation for CPRC; 8) w/o Prediction, CPRC only retains the relation consistency loss in (8); 9) w/o Relation, CPRC only retains the prediction consistency in (8); and 10) w/o $\tau$, CPRC removes the confidence threshold in (7). For evaluation, we use different metrics, including BLEU [42], METEOR [43], ROUGE-L [9], CIDEr-D [30], and SPICE [44], to evaluate the proposed method and comparison methods. All the metrics are computed with the publicly released code.2 In fact, the CIDEr-D and SPICE metric is more suitable for the image captioning task [30], [44]. One of the problems with using metrics, such as BLEU, ROUGE-L, CIDEr-D, and METEOR is that these metrics are primarily sensitive to n-gram overlap. However, n-gram overlap is neither necessary nor sufficient for two sentences to convey the same meaning [45].

D. Qualitative Analysis

Table I presents the quantitative comparison results with state-of-the-art methods (i.e., 1% supervised data and 99% unsupervised in the training set), it is notable that supervised captioning methods can only develop the mapping functions with supervised data, and leave out the unsupervised data. For fairness, all the models are first trained under cross-entropy loss and then optimized for CIDEr-D score as [9]. “—” represents the results have not given in the raw paper. The results reveal the following.

1) Unsupervised approach, that is, UIC, achieve the worst performance on all metrics under different losses. This phenomenon verifies that the generated sentence may mismatch the image with a high probability when only considering the domain discriminator. Graph-align performs better than supervised approaches, but worse than A3VSE on most metrics, because it ignores to measure specific matching.

2) Semisupervised method, that is, A3VSE, has little effect on improving the captioning performance, for example, cross-entropy loss/CIDEr-D score optimization only improves 0.4/2.0 and 0.2/0.1 on CIDEr-D and SPICE scores compared to AoANet, because it is more difficult to ensure the quality of generated sentences.

3) CPRC achieves the highest scores among all compared methods in terms of all metrics, on both the cross-entropy loss and CIDEr-D score optimization stage, except ROUGE-L on cross-entropy loss. For example, CPRC achieves a state-of-the-art performance of 77.9/78.8 (CIDEr-D score) and 16.2/16.8 (SPICE score) under two losses (cross entropy and CIDEr-D score), that acquires 8.4/8.1 and 1.9/1.5 improvements compared to RSTNet (i.e., SOTA supervised model), 8.4 and 1.8 improvements compared to Graph-align (i.e., SOTA unsupervised model), and 8.3/6.4 and 1.7/1.5 improvements compared to A3VSE (i.e., SOTA semisupervised model). The phenomena indicates that, with limited amount of supervised data, existing methods cannot construct a well mapping function, whereas CPRC can reliably utilize the undescribed image to enhance the model.

2https://github.com/tylin/coco-caption
4) CPRC performs better than w/o $\tau$ on all metrics, which indicates the effectiveness of threshold confidence.

### E. Ablation Study

To quantify the impact of the proposed CPRC modules, we compare CPRC against other variant models under various settings. The bottom half of Table I presents the results.

1) AoANet$+P$ and AoANet$+C$ achieve better performance than AoANet, which indicates that the prediction loss and relation consistency loss can improve the generator learning, because the labels can provide extra semantic information; meanwhile, AoANet$+P$ performs better than AoANet$+C$ on most metric, which indicates that prediction loss is more significant than relation consistency.

2) PL and AC perform worse than the w/o Prediction and w/o Relation, which verifies that traditional semisupervised techniques considering pseudo labeling are not as good as cross-modal semisupervised techniques considering raw image as pseudo supervision.

3) Embedding$+$ performs worse than the Semantic$+$, which reveals that embeddings are more difficult to compare than predictions since image and text have heterogeneous representations.

4) Strong$+$ performs worse than CPRC, which validates that the strong augmentation may impacts the generated sentence, and further affects the prediction.

5) Both the w/o Prediction and w/o Relation can improve the captioning performance on most criteria, especially on the important criteria, that is, CIDEr-D and SPICE. The results indicate that both the prediction and relation consistencies can provide effective supervision to ensure the quality of generated sentences.

### F. CPRC With Different Captioning Model

To explore the generality of CPRC, we conduct more experiments by incorporating CPRC with different supervised captioning approaches, that is, SCST (encoder-decoder-based model) and GIC (attention-based model). Note that we have not adopted the editing-based method considering the reproducibility, the results are recorded in Table II. We find that all
methods, that is, SCST, GIC, and AoANet (results can refer to the Table I), have improved the performance after combining the CPRC framework. This phenomenon validates that CPRC can well combine the undescribed images for existing supervised captioning models.

G. Influence of the Supervised and Unsupervised Images

To explore the influence of supervised data, we tune the ratio of supervised data, and the results are recorded in Figs. 4 and 5 with different metrics. Here, we find that, with the percentage of supervised data increases, the performance of CPRC improves faster than other state-of-the-art methods. This phenomenon indicates that CPRC can reasonably utilize the undescribed images to improve the learning of generator. Furthermore, we validate the influence of unsupervised data, that is, we fix the supervised ratio to 1%, and tune the ratio of unsupervised data in \{10\%, 40\%, 70\%, 100\%\}, the results are recorded in Fig. 6. Note that one of the problems by using metrics, such as BLEU, ROUGE-L, CIDEr-D, and METEOR to evaluate captions, is that these metrics are primarily sensitive to n-gram overlap [9], [44]. Therefore, we only give the results of CIDEr-D and SPICE here (refer to the supplementary material for more details). We find that with the percentage of unsupervised data increases, the performance of CPRC also improves. This phenomenon indicates that CPRC can make full use of undescribed images for positive training.

H. Influence of the Augmentation Number

To explore the influence of augmentation number, that is, \(K\), we conduct more experiments. In detail, we tune \(K\) in \{1, 2, 3, 4, 5\} and recorded the results in Table III. The results reveal that the CPRC achieves the best performance with \(K = 3\), for the reason that additional inconsistent noises between images and sentences may be introduced with the number of augmentations increases.

I. Influence of the Confidence Threshold

To explore the influence of confidence threshold, that is, \(\tau\), we conduct more experiments. In detail, we tune the \(\tau\) in \{0, 0.1, 0.4, 0.7\} and recorded the results in Table IV. The
Results reveal that the performance of CPRC increases first, and then decreases with the increasing of \( \tau \). The reason is that fewer undescribed images are used with the increasing of \( \tau \), thereby the generator training has not fully explored the information in the unsupervised data.

**J. Visualization and Analysis**

Fig. 7 shows a few examples with captions generated by our CPRC and two baselines, A3VSE and AoANet, as well as the human-annotated ground truths. From these examples, we find that the generated captions of baseline models lack language logic and lose accuracy for the image objects, while CPRC can generate accurate captions in high quality.

Fig. 8 shows an example of augmented images and corresponding generated captions. We find that the generated captions basically have similar semantic information, which can help the prediction and relation consistencies for the undescribed images.

**K. Influence of Label Prediction**

To explore the effect of prediction loss, we conduct more experiments and exhibit several cases. Fig. 9 shows a few examples with captions generated by our CPRC and two baselines, A3VSE and AoANet, as well as the human-annotated ground truths. From these examples, we find that the label prediction helps the generator to understand the image from these words marked in red within the sentence generated by CPRC, for example, in Fig. 9(a), the content of the image is complicated and the bird is not obvious, which causes the inconsistency of sentences generated by AoANet and A3VSE with the ground truths. But CPRC can generate a good description of “bird” and “umbrella” by combining label prediction information.
L. Discussion

In this article, we concentrate on the semisupervised image captioning task, in which the key challenge is the reasonable usage of undescribed images to improve the performance of generator. Compared to current semisupervised approaches that only consider the pseudo descriptions, we design a novel captioning model by constraining the undescribed images from both the prediction and relation consistencies. In experiments, we comprehensively compare with state-of-the-art captioning methods to validate the effectiveness of proposed CPRC on popular datasets. Furthermore, we conduct adequate ablation studies.
which verify that: 1) the designed prediction and relation consistencies can positively promote the use of undescribed images; 2) CPRC can be effectively applied to any supervised models; 3) CPRC is highly interpretable for parameter selection; and 4) CPRC has a good visualization effect, and the significance of each module is reflected.

V. Conclusion

Since traditional image captioning methods usually work on supervised multimodal data, in this article, we investigated how to use undescribed images for semisupervised image captioning. Specifically, our method can take CPRC into consideration. CPRC employs prediction distillation for the predictions of sentences generated from undescribed images, and develops a novel relation consistency between augmented images and generated sentences to retain the important relational knowledge. As demonstrated by the experiments, CPRC outperforms state-of-the-art methods in various complex semisupervised scenarios.

References

[1] T. Baltrusaitis, C. Abuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, Feb. 2019.
[2] E. S. Debie et al., “Multimodal fusion for objective assessment of cognitive workload: A review,” IEEE Trans. Cybern., vol. 51, no. 3, pp. 1542–1555, Mar. 2021.
[3] A. Karpathy and F. Li, “Deep visual-semantic alignments for generating image descriptions,” in Proc. CVPR, 2015, pp. 3128–3137.
[4] K. Xu et al., “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. ICML, 2015, pp. 2048–2057.
[5] F. Samman and M. Elsayed, “Look and modify: Modification networks for image captioning,” in Proc. BMVC, 2019, p. 75.
[6] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, “Describing video with attention-based bidirectional LSTM,” IEEE Trans. Cybern., vol. 49, no. 7, pp. 2631–2641, Jul. 2019.
[7] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. Salakhutdinov, “Review networks for caption generation,” in Proc. NeurIPS, 2016, pp. 2361–2369.
[8] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” in Proc. CVPR, 2017, pp. 3242–3250.
[9] J. Huang, W. Wang, J. Chen, and X. Wei, “Attention on attention for image captioning,” in Proc. ICCV, 2019, pp. 4633–4642.
[10] T. B. Hashimoto, K. Guu, Y. Oren, and P. Liang, “A retrieve-and-edit framework for predicting structured outputs,” in Proc. NeurIPS, 2018, pp. 10073–10083.
[11] Y. Feng, L. Ma, W. Liu, and J. Luo, “Unsupervised image captioning,” in Proc. CVPR, Long Beach, CA, USA, 2019, pp. 4125–4134.
[12] J. Gu, S. R. Joty, J. Cui, H. Zhao, X. Yang, and G. Wang, “Unpaired image captioning via scene graph alignments,” in Proc. ICCV, 2019, pp. 10322–10331.
[13] I. J. Goodfellow et al., “Generative adversarial nets,” in Proc. NeurIPS, Montreal, QC, Canada, 2014, pp. 2672–2680.
[14] N. C. Mithun, R. Panda, E. E. Papalexakis, and A. K. Roy-Chowdhury, “Websly supervised joint embedding for cross-modal image-text retrieval,” in Proc. ACMMM, 2018, pp. 1856–1864.
[15] P. Huang, G. Kang, W. Liu, X. Chang, and A. G. Hauptmann, “Annotation efficient cross-modal retrieval with adversarial attention alignment,” in Proc. ACM MM, 2019, pp. 1758–1767.
[16] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in Proc. CVPR, 2017, pp. 1179–1195.
[17] Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, “More grounded image captioning by distilling image-text matching model,” in Proc. CVPR, 2020, pp. 4776–4785.
Yang Yang received the Ph.D. degree in computer science from Nanjing University, Nanjing, China, in 2019.

He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing. He has published over 20 papers in leading international journals and conference proceedings. His research interests lie primarily in machine learning and data mining, including heterogeneous learning, model reuse, and incremental mining.

Hengshu Zhu (Senior Member, IEEE) received the B.E. and Ph.D. degrees in computer science from the University of Science and Technology of China, Hefei, China, in 2009 and 2014, respectively.

He is currently a Principal Data Scientist and an Architect with Baidu Inc., Beijing, China. He has published prolifically in refereed journals and conference proceedings, including the IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Mobile Computing, ACM Transactions on Information Systems, ACM Transactions on Knowledge Discovery from Data, ACM SIGKDD, ACM SIGIR, WWW, IJCAI, and AAAI. His general area of research is data mining and machine learning, with a focus on developing advanced data analysis techniques for innovative business applications.

Dr. Zhu was the recipient of the Distinguished Dissertation Award of CAS in 2016, the Distinguished Dissertation Award of CAAI in 2016, the Special Prize of President Scholarship for Postgraduate Students of CAS in 2014, the Best Student Paper Award of KSEM-2011, WAIM-2013, and CCDM-2014, and the Best Paper Nomination of ICDM-2014. He has served regularly on the organization and program committees of numerous conferences, including as a Program Co-Chair for the KDD Cup-2019 Regular ML Track, and a Founding Co-Chair for the first International Workshop on Organizational Behavior and Talent Analytics and the International Workshop on Talent and Management Computing, in conjunction with ACM SIGKDD. He is the Senior Member of ACM and CCF.

Hong Chen Wei is currently pursuing the M.Sc. degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests lie primarily in machine learning and data mining, including cross-modal learning.

Hui Xiong (Fellow, IEEE) received the Ph.D. degree from the University of Minnesota, Minneapolis, MN, USA, in 2005.

He is currently a Full Professor with the Rutgers University, Newark, NJ, USA. He is currently a Chang-Jiang Professor with the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2002.

In 2003, he was a Postdoctoral Researcher with the University of Zaragoza, Zaragoza, Spain. From 2004 to 2006, he was a Postdoctoral Fellow with the Biometrics Centre, The Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Postdoctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Chang-Jiang Professor with the School of Computer Science and Engineering, NUST. He has authored more than 200 scientific papers in pattern recognition and computer vision. His papers have been cited more than 6000 times in the Web of Science and 15,000 times in the Scholar Google. His current research interests include pattern recognition, computer vision, and machine learning.

Dr. Yang is currently an Associate Editor of Pattern Recognition, Pattern Recognition Letters, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and Neurocomputing. He is a Fellow of IAPR.