Visual Informatics 5 (2021) 41–48

Contents lists available at ScienceDirect

Visual Informatics

journal homepage: www.elsevier.com/locate/visinf

Image Captioning with multi-level similarity-guided semantic matching

Jiesi Li [a], Ning Xu [a,*], Weizhi Nie [a], Shenyuan Zhang [b]

[a] The School of Electrical and Information Engineering, Tianjin University, China
[b] People's Daily, China

ARTICLE INFO

Article history:
Received 15 September 2021
Received in revised form 25 November 2021
Accepted 30 November 2021
Available online 11 December 2021

Keywords:
Image Captioning
Cross-modal semantic matching
Reinforcement learning

ABSTRACT

Image Captioning is a cross-modal task that needs to automatically generate coherent natural sentences to describe the image contents. Due to the large gap between vision and language modalities, most of the existing methods have the problem of inaccurate semantic matching between images and generated captions. To solve the problem, this paper proposes a novel multi-level similarity-guided semantic matching method for image captioning, which can fuse local and global semantic similarities to learn the latent semantic correlation between images and generated captions. Specifically, we extract the semantic units containing fine-grained semantic information of images and generated captions, respectively. Based on the comparison of the semantic units, we design a local semantic similarity evaluation mechanism. Meanwhile, we employ the CIDEr score to characterize the global semantic similarity. The local and global two-level similarities are finally fused using the reinforcement learning theory, to guide the model optimization to obtain better semantic matching. The quantitative and qualitative experiments on large-scale MSCOCO dataset illustrate the superiority of the proposed method, which can achieve fine-grained semantic matching of images and generated captions.

1. Introduction

Image Captioning task has received widespread attention from the computer vision and natural language processing communities (Vinyals et al., 2015; Xu et al., 2015; Zhang et al., 2017; Jiang et al., 2018a; Gu et al., 2019). The goal of it is to generate semantically faithful descriptions for the given images. It is a challenging cross-modal problem that not only needs to model the visual contents in depth, but also to translate the captured visual information into relevant language descriptions. Image Captioning algorithms can support various potential applications, like semantic image retrieval (Karaoglu et al., 2017), visually impaired assistance (Wu et al., 2017).

To generate coherent descriptions, most of the existing methods first extract visual semantic units from images and then design the RNN-based generator to transfer visual units into textual words (Karpathy and Li, 2015; Chen et al., 2017; Anderson et al., 2018; Yao et al., 2018; Yang et al., 2019). However, they neglect the correspondence of visual semantic units and corresponding textual words, which results in inaccurate semantic matching between images and generated captions. For instance, in Fig. 1, the model (Yang et al., 2019) describes the given image as caption "A vase with an umbrella sitting on a table". It can be seen that the generated caption has obvious defects, where the essential visual semantic unit "flower" is wrongly described as "umbrella" and the corresponding word description of visual semantic unit "glass" is absent.

To solve the above problem and obtain more accurate captions, we propose a novel multi-level similarity-guided semantic matching method for image captioning. Particularly, we extract the fruitful visual semantic units composed of visual objects. Meanwhile, a generated sentence is also decomposed to get the textual words describing objects, called textual semantic units. We design a local semantic similarity evaluation mechanism to explicitly measure the correlation between visual and textual semantic units, which can help to optimize the model to accurately learn the visual–textual correspondence and achieve the fine-grained semantic matching of images and generated captions.

Specifically, our method consists of a caption generation network and a multi-level similarity fusion network, as shown in Fig. 2. The caption generation network involves the extraction and encoding of visual semantic units based on Faster R-CNN, as well as the language decoding based on LSTM and attention mechanism. The multi-level similarity fusion network obtains the visual semantic units of the image and the textual semantic units of a generated caption, respectively. It calculates the

* Corresponding author.
E-mail addresses: lijs980211@tju.edu.cn (J. Li), ningxu@tju.edu.cn (N. Xu), weizhinie@tju.edu.cn (W. Nie), zhangshenyuan0@gmail.com (S. Zhang).

https://doi.org/10.1016/j.visinf.2021.11.003

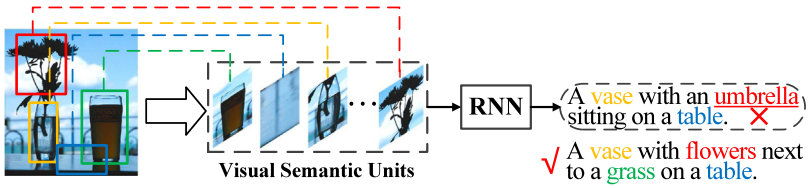

Fig. 1. Illustration of the problem with the existing methods.

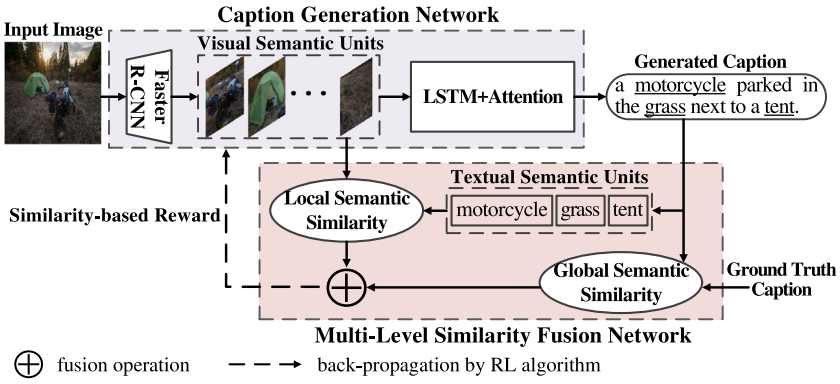

Fig. 2. Framework of the proposed method, including a Caption Generation Network and a Multi-Level Similarity Fusion Network. The former aims to extract and encode the visual semantic units of image and decode them into the sentence. The latter calculates the local and global semantic similarities respectively and fuses them to build the multi-level similarity-based reward, which is back-propagated to guide the parameter update of caption generation network by Reinforcement Learning (RL) algorithm, to achieve better semantic matching.

correlation between visual and textual semantic units as local semantic similarity, and CIDEr score as global semantic similarity. The two-level similarities are effectively fused as a reward by applying the reinforcement learning (RL) mechanism, to guide the optimization of the caption generation network to achieve better semantic matching of images and generated captions.

Our contributions are summarized as follows:

- We propose a multi-level similarity-guided semantic matching method for image captioning, which can fuse the local and global two-level similarities to learn the latent semantic correlation of images and generated captions.
- We design a local semantic similarity evaluation mechanism, which can explicitly measure the correlation between the visual semantic units of image and the textual semantic units of the generated caption to achieve fine-grained semantic matching.
- We verify the proposed method by conducting comparative and ablative experiments on the large-scale MSCOCO dataset. The great performances prove its effectiveness. Besides, sample visualization examples further validate its ability.

2. Related work

2.1. Image Captioning

In the literature, numerous image captioning methods have been proposed, where an encoder–decoder fashion is the mainstream framework (Vinyals et al., 2015; Karpathy and Li, 2015; Jiang et al., 2018a; Chen et al., 2018). Convolutional Neural Network(CNN) has been broadly used as the encoder to extract image features, while LSTM, a variant of RNN is often employed as a decoder to generate natural language sentences. Vinyals et al. (2015) first introduced a deep learning CNN-LSTM image captioning method. Later, a lot of works were proposed. Karpathy and Li (2015) generated coherent descriptions by integrating object detection with Region-CNN and corresponding sentence snippets to the visual regions. However, the model was not end-to-end and too complicated to operate. Inspired by Karpathy and Li (2015), Wang et al. (2016) presented a deep end-to-end bidirectional LSTM model that considered both history and future context to summarize visual-language long range interactions. Some works innovated inside the network architecture to enhance captioning, such as Gu et al. (2017) designed a competitive language CNN. Different from the previous models that predicted the next word according to only one previous word, it can model the long range dependencies by feeding with all the previous words. Chen et al. (2018) focused on RNN that regularized the transition dynamics of it and mitigated its discrepancy for sequence prediction. In contrast, Jiang et al. (2018a) added a guiding network component into the CNN-RNN framework to improve caption quality without changing internal structures of the networks.

Recently, to further improve the model ability, attention-based approaches were employed, where the captioning model can selectively emphasize features depending on the current context (Xu et al., 2015; Chen et al., 2017; Lu et al., 2017; Anderson et al., 2018; Herdade et al., 2019; Zhou et al., 2020). For instance, Xu et al. (2015) built the attention mechanism to locate salient image parts during sentence generation. Lu et al. (2017) introduced an adaptive attention that can automatically determine whether to rely on the image (visual information) or only a language model to generate each word. Anderson et al. (2018) combined bottom-up and top-down visual attention, with the former used to extract region features and the latter determined feature weights, resulting in a more natural calculation of attention at both object-level and other salient region-level. Zhou et al. (2020) regularized the visual attention in captioner by extracting knowledge from a more complex image-text matching model. In general, attention-based methods can better integrate visual information into language processing and significantly improve captioning accuracy.

Additionally, some researchers incorporated the scene graph into captioning model for fine-grained understanding. Scene graph

is a graphical structured representation of scene semantics composed of object nodes and directed relationship edges. The typical work such as Yao et al. (2018) leveraged the scene graph to model spatial and semantic relationships of visual objects for captioning. Yang et al. (2019) embedded language inductive bias through scene graphs, which is prior knowledge for generating more fruitful descriptions. However, they all decoded the graph features into textual words disorderly, ignoring the visual–textual correspondence. Based on the work of Yang et al. (2019), Gu et al. (2019) presented a cross-modal graph feature alignment method for unpaired image captioning.

2.2. Reinforcement Learning

Reinforcement Learning (RL) involves the interaction between an agent and an environment, learning a policy that maximizes the cumulative future rewards, to solve the sequential decision problems. At present, RL algorithms have successfully solved many challenging problems (Kong et al., 2017; Krull et al., 2017; Yun et al., 2017). The RL mechanism is also used in captioning task, which greatly improves the performances of image captioning (Ranzato et al., 2016; Rennie et al., 2017; Liu et al., 2017; Ren et al., 2017; Liu et al., 2018; Gao et al., 2019; Xu et al., 2020). For instance, Ranzato et al. (2016) first introduced REINFORCE algorithm (Williams, 1992) to train an RNN-based text generation model by directly optimizing an evaluation metric. Rennie et al. (2017) modified the REINFORCE algorithm to build a more stable self-critical sequence training RL system. Liu et al. (2017) directly optimized the combination of various metrics through a policy gradient approach to enhance the caption quality. Liu et al. (2018) improved the visual policy by incorporating the visual context into sequential reasoning.

At present, although the fluent and semantically rich caption can be generated, it still suffers from the visual–textual inaccurate semantic matching. In this paper, we explore the latent semantic correlation of images and generated captions. Particularly, we extract the visual semantic units of the image and the textual semantic units of generated caption individually. We use the designed local semantic similarity evaluation mechanism to explicitly measure their correlation, which is helpful for model optimization and generating accurate descriptions.

3. Method

In this section, we elaborate the proposed method that is illustrated in Fig. 2. We first introduce the caption generation network, followed by the multi-level similarity fusion network, including the textual semantic units extraction, the local and global semantic similarity calculation, and the two-level similarities fusion. Our training procedure is then briefly described.

3.1. Caption generation network

Visual Semantic Units Encoding. An image can be seen as a structured combination of regions containing visual objects that we call visual semantic units. Input an image I, we detect the visual objects in I, and extract their region features. Specifically, we employ the pre-trained Faster R-CNN (Ren et al., 2015) to locate and classify object instances, whose operation can be summarized into three steps. Firstly, the image feature map is extracted through convolutions, and sent to RPN to generate multiple box proposals. Then using the location information of the box to extract the feature representation of each candidate region from the image feature map, and convert it into fixed-length output by ROI pooling. The fixed-length features are finally batched together as input to a fully connected layer for position regression and object classification.

We take the final output of the Faster R-CNN that includes refined bounding boxes of regions and a probability distribution over the class labels of objects contained in the regions. We keep each region which detected object class probability exceeds a confidence threshold, and get its mean-pooled convolutional feature vector $\boldsymbol{v}_i$ and object class label o_i. Thus, the region feature representation $V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ and semantic units representation $L_I = \{o_1, \ldots, o_k\}$ of the image I are obtained.

Language Decoding. Given the image region features $V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$, inspired by the region-level attention mechanism (Anderson et al., 2018), we built two stacked LSTM (Hochreiter and Schmidhuber, 1997) layers with an attention module to decode the features into a corresponding sentence $S = \{w_1, \ldots, w_T\}$, where w_i represents a generated word and T is sentence length.

Take the operation on single time step t as an example, the first LSTM layer concatenates the mean-pooled feature $\bar{\boldsymbol{v}} = \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{v}_i$, the embedding vector $\boldsymbol{e}_{t-1} = \boldsymbol{W}_e w_{t-1}$ of previously generated word w_{t-1}, and the previous output $\boldsymbol{h}_{t-1}^2$ of second LSTM layer as its input and output a hidden state $\boldsymbol{h}_t^1$:

$$\boldsymbol{h}_t^1 = \mathrm{LSTM}_1\left(\left[\bar{\boldsymbol{v}}; \boldsymbol{e}_{t-1}; \boldsymbol{h}_{t-1}^2\right], \boldsymbol{h}_{t-1}^1\right) \tag{1}$$

where $\boldsymbol{W}_e$ is a word2vec (Church, 2017) initialized embedding matrix; [;] represents the concatenation operation of vectors. For each feature $\boldsymbol{v}_i$ in V at time step t, we use the $\boldsymbol{h}_t^1$ as an index vector to generate its attention weight $\boldsymbol{a}_{i,t}$ through the attention module:

$$\begin{aligned} \boldsymbol{\alpha}_{i,t} &= \boldsymbol{W}_\alpha \tanh\left(\boldsymbol{W}_v \boldsymbol{v}_i + \boldsymbol{W}_h \boldsymbol{h}_t^1\right) \\ \boldsymbol{a}_{i,t} &= \mathrm{softmax}\left(\boldsymbol{\alpha}_{i,t}\right) \end{aligned} \tag{2}$$

where $\boldsymbol{W}_\alpha$, $\boldsymbol{W}_v$ and $\boldsymbol{W}_h$ are trainable parameter matrices; $\boldsymbol{a}_{i,t}$ is the final normalized attention weight among [0, 1]. Take the weighted sum $\hat{\boldsymbol{v}}_t = \sum_{i=1}^{k}\boldsymbol{a}_{i,t}\boldsymbol{v}_i$ of all visual features, which is concatenated with $\boldsymbol{h}_t^1$ as the input of the second LSTM layer:

$$\boldsymbol{h}_t^2 = \mathrm{LSTM}_2\left(\left[\hat{\boldsymbol{v}}_t; \boldsymbol{h}_t^1\right], \boldsymbol{h}_{t-1}^2\right) \tag{3}$$

where $\boldsymbol{h}_t^2$ denotes the output (hidden state) of the second LSTM layer. The $\boldsymbol{h}_t^2$ is then gone through softmax to predict the current word w_t:

$$p\left(w_t \mid w_{1:t-1}\right) = \mathrm{softmax}\left(\boldsymbol{W}_p \boldsymbol{h}_t^2 + \boldsymbol{b}_p\right) \tag{4}$$

where $w_{1:t-1} = \{w_1, \ldots, w_{t-1}\}$ is the sequence of previously generated words; $\boldsymbol{W}_p$ and $\boldsymbol{b}_p$ are learnable parameters (weight and bias). when $t = T$, a complete description sentence $w_{1:T} = \{w_1, \ldots, w_T\}$ can be obtained.

3.2. Multi-level similarity fusion network

3.2.1. Textual semantic units extraction

We have obtained the visual semantic units of the image above. Similar to images, a natural language sentence is a combination of various linguistic words with different part of speech, where nouns represent objects, verbs or adjectives describe the action states or attributes of the objects. Given the generated sentence $S = \{w_1, \ldots, w_T\}$, we identify the part of speech of each word in it and select the nouns, that is, the textual objects, to form the textual semantic units. Particularly, we use the Stanford CoreNLP (Manning et al., 2014) parser, which can extract the noun phrases (NP) in a sentence through syntactic analysis. Take the sentence "a motorcycle parked in the grass next to a tent" as an example, it can be parsed by CoreNLP to get the noun phrases "a motorcycle", "the grass" and "a tent", where the last word in the phrase is the object, and the previous word is the attribute modifier of the object. For each generated sentence S, we keep the last words in noun phrases extracted by CoreNLP to form the textual semantic units $L_S = \{w_1, \ldots, w_n\}$.

3.2.2. Local and global semantic similarity calculation

Local Semantic Similarity. In this paper, we design a local semantic similarity evaluation mechanism based on the visual and textual semantic units comparison, to evaluate the semantic similarity of images and generated captions in a fine-grained manner. In particular, given the extracted visual semantic units $L_I = \{o_1, \ldots, o_k\}$ and textual semantic units $L_S = \{w_1, \ldots, w_n\}$, we employ the natural language processing toolkit spaCy (Honnibal and Montani, 2017) to measure the correlation between any two elements from L_I and L_S. spaCy can calculate the cosine similarity of two words by embedding them into corresponding vector representations through the built-in word2vec module. Specifically, for each w_i in L_S, we calculate its similarity score with all o_j in L_I by the function spaCy $(\cdot, \cdot)$ and obtain a size of $n \times k$ score matrix $M_s = [s_{i,j}]$:

$$s_{i,j} = \text{spaCy}\left(w_i, o_j\right) \quad i \epsilon 1, \ldots, n; j \epsilon 1, \ldots, k \tag{5}$$

where w_i, o_j are the word labels of the ith textual semantic unit (textual object) in L_S and jth visual semantic unit (visual object) in L_I, respectively; spaCy $(\cdot, \cdot)$ returns the semantic similarity score of two input words, the higher the scalar score, the more similar two words are to each other.

Further, we take the maximum value of each row in M_s, which means for each textual semantic unit w_i, its best match with multiple visual semantic units $\{o_1, \ldots, o_k\}$ is selected. The n maximum values are finally summed to get the local semantic similarity Sim_{loc}:

$$Sim_{loc}(I, S) = \sigma \sum_{i=1}^{n} \max_{j \in [1,k]} \left[s_{i,j}\right] \tag{6}$$

where max means the operation of selecting maximum; σ is the sigmoid function.

Global Semantic Similarity. Different from the word-level fine-grained comparison based on the semantic units, here we use the sentence-level CIDEr (Vedantam et al., 2015) score to characterize the global semantic similarity between images and generated captions. CIDEr, as an automatic evaluation metric, measures the similarity between the generated sentences and human-labeled ground truths. Therefore, it can be applied to reflect the global semantic consistency of the generated sentences and real images. Given the ground truth S^* of image I, the calculation of global semantic similarity Sim_{glo} can be formulated as:

$$Sim_{glo}(I, S) = \text{CIDEr}\left(S, S^*\right) \tag{7}$$

3.2.3. Two-level similarities fusion

After obtaining the local semantic similarity Sim_{loc} and global semantic similarity Sim_{glo}, we fuse them for model optimization using the Reinforcement Learning (RL) mechanism. Based on the RL theory, we regard our caption generation network as an "agent", which can interact with the external "environment" and do "actions" to change its state to optimize the goal. The "environment" consists of the input image I and the sequence $\{w_1, \ldots, w_{t-1}\}$ of previously generated words, while the "action" is the prediction of the next word w_t. When a sentence $S = \{w_1, \ldots, w_T\}$ prediction is completed, the "agent" will receive an expected "reward" score, which can indicate the quality of the generated sentence. Particularly, we build the reward function r for agent network optimization by fusing the two-level similarities, which is the weighted sum of the Sim_{loc} and Sim_{glo}:

$$r(S) = \lambda Sim_{loc}(I, S) + (1 - \lambda) Sim_{glo}(I, S) \tag{8}$$

where λ is the adjustable fusion weight.

Table 1
The implementation details of the Caption Generation Network.

Definition	Dimension
visual feature $\boldsymbol{v}_i$	2048
mean-pooled feature $\bar{\boldsymbol{v}}$	2048
word embedding matrix $\boldsymbol{W}_e$	1000 × 10369
word embedding vector $\boldsymbol{e}_{t-1}$	1000
LSTM hidden state $\boldsymbol{h}_t^1/\boldsymbol{h}_t^2$	1000/1000
parameter metrics $\boldsymbol{W}_\alpha/\boldsymbol{W}_v/\boldsymbol{W}_h$	512/512 × 2048/512 × 1000
attention weight $\boldsymbol{a}_{i,t}$	512

3.3. Training procedure

Define the parameter of the caption generation network (agent) by θ, we first minimize the cross-entropy (XE) loss by following the traditional captioning training:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log p_\theta\left(w_t \mid w_{1:t-1}\right) \tag{9}$$

where $w_{1:t-1} = \{w_1, \ldots, w_{t-1}\}$ is the sequence of previously generated words.

Then based on RL mechanism, we use the built reward function $r(S)$ that fuses the local and global semantic similarities to further optimize the parameter θ, which is the process of minimizing the negative expected reward:

$$L_{RL}(\theta) = -E_{S \sim p_\theta}[r(S)] \tag{10}$$

where $E_{S \sim p_\theta}$ is the expected reward objective function. According to the policy gradient algorithm in SCST (Rennie et al., 2017), we calculate the gradient of the above loss as:

$$\nabla_\theta L_{RL}(\theta) \approx -\left(r\left(S^s\right) - r\left(S^b\right)\right) \nabla_\theta \log p_\theta\left(S^s\right) \tag{11}$$

where $S^s = \left\{w_1^s, \ldots, w_T^s\right\}$ is a generated sample caption; $S^b = \left\{w_1^b, \ldots, w_T^b\right\}$ is a baseline sentence acquired by greedy decoding the current model.

4. Experiments

4.1. Dataset and evaluation metrics

Dataset. The proposed method is verified on a large-scale MSCOCO (Lin et al., 2014) dataset. MSCOCO contains various types of images from daily complex scenes, with over 2 million instances, 80 object categories and each image contains 5 human description (ground-truth) captions. We adopt the widely followed Karpathy split (Karpathy and Li, 2015) that utilizes different 113,287 and 5000 images for model training and validation, another 5000 for testing.

Metrics. We evaluate the generated captions by standard evaluation metrics: BLEU-N (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin and Hovy, 2003), CIDEr-D (Vedantam et al., 2015), SPICE (Anderson et al., 2016), shorted as B-N, M, R, C and S in the following experiments. The higher the metric value, the better the captioning model.

4.2. Experimental details

We trim the ground-truth captions of more than 16 words in MSCOCO, and delete the words occurring less than 5 times to obtain a dictionary of 10369 words. For the caption generation network, We follow Yang et al. (2019) to train Faster R-CNN, and set the detection confidence threshold of object class as 0.2 to select salient image regions. The number of regions k selected per image is at least 10 and varies with the image complexity. The

Table 2
Performances of the proposed model and existing advanced models on MSCOCO. All values are percentages and achieved by single model.

	Model	B-1	B-4	M	R	C
CNN + LSTM	Up-Down (Anderson et al., 2018) (CVPR2018)	79.8	36.3	27.7	56.9	120.1
	RFNet (Jiang et al., 2018b) (ECCV2018)	79.1	36.5	27.7	57.3	121.9
	POS-SCAN (Zhou et al., 2020) (CVPR2020)	80.1	37.8	28.3	–	125.9
	CapNet (Yang et al., 2021) (TMM2021)	80.4	38.5	**28.8**	58.3	127.6
	GCN-LSTM (Yao et al., 2018) (ECCV2018)	80.5	38.2	28.5	58.3	127.6
	SGAE (Yang et al., 2019) (CVPR2019)	80.8	38.4	28.4	58.6	127.8
	DGLC (Dong et al., 2021) (ACM MM2021)	81.4	38.8	28.0	58.4	127.6
	SCST (Rennie et al., 2017) (CVPR2017)	–	34.2	26.7	55.7	114.0
	N-step (Gao et al., 2019) (CVPR2019)	77.9	35.0	26.9	56.3	115.2
	CAVP (Liu et al., 2018) (ACM MM2018)	–	38.6	28.3	58.5	126.3
CNN + Transformer	ORT (Herdade et al., 2019) (NIPS2019)	80.5	38.6	28.7	58.4	128.3
	DGT (Dong et al., 2021) (ACM MM2021)	**81.8**	**39.2**	28.7	**58.9**	128.0
Ours	Full-model	81.2	39.0	28.5	**58.9**	**128.5**

size of the visual feature $\boldsymbol{v}_i$ is set to 2048-dimension, the word embedding $\boldsymbol{e}_{t-1}$ and LSTM hidden state $\boldsymbol{h}_t^1/\boldsymbol{h}_t^2$ are all set to 1000. We embed the attention weight $\boldsymbol{a}_{i,t}$ through the attention module as 512-dimension. The corresponding implementation settings of the caption generation network are summarized in Table 1 for facilitate observation. In the multi-level similarity fusion network, we set the fusion weight λ in Eq. (8) as 0.2 to build the reward function $r(S)$ for RL optimization. As mentioned above, our model are successively trained by the XE loss (Eq. (9)) and the built multi-level similarity-based reward (Eq. (10)). Specifically, in the phase of training, Adam (Kingma and Ba, 2015) optimizer is employed to update the model parameter with a batch size of 25. The initial learning rate is set to $5e^{-4}$ and decays by 0.8 every 5 epochs. In the inference process of language decoding, we use beam search with a beam size of 5.

4.3. Quantitative analysis

Since the proposed method is based on the LSTM network, for fair comparison, we first compare it with the existing advanced CNN-LSTM methods on MSCOCO Karpathy test split. Specifically, as shown in Table 2, the compared methods include 1) four Attention-based models: Up-Down (Anderson et al., 2018), RFNet (Jiang et al., 2018b), POS-SCAN (Zhou et al., 2020), CapNet (Yang et al., 2021); 2) three Graph-based models: GCN-LSTM (Yao et al., 2018), SGAE (Yang et al., 2019), DGLC (Dong et al., 2021); 3) three RL-based Models: SCST (Rennie et al., 2017), CAVP (Liu et al., 2018), N-step (Gao et al., 2019). Besides, we also make a comparison with two transformer-based model, i.e., ORT (Herdade et al., 2019), DGT (Dong et al., 2021).

It can be seen that our model realizes competitive performances. Especially, compared with the methods that are also based on LSTM, our model can obtain the best results in terms of BLEU-1 (81.2), BLEU-4 (39.0), ROUGE-L (58.9), and CIDEr-D (128.5). To be specific, we have four key observations:

- RFNet uses multiple encoders to extract diverse representations for boosting captioning; POS-SCAN enhances the image captioning model by distilling knowledge from a more complex image-text matching model; CapNet improves LSTM specifically for image captioning and designs a memory initialization method to obtain richer visual semantics. However, their performances are worse than our model. It proves the effectiveness of the proposed method, which can achieve fine-grained semantic matching of images and generated captions by effectively mining the visual–textual latent semantic correlation under the guidance of local and global multi-level similarities.
- Our model can outperform the Graph-based methods GCN-LSTM, SGAE and DGLC, which exploit the semantic information of structured graphs to bridge the visual–textual semantic gap. By comparison, our method performs the visual–textual semantic matching by explicitly measuring the correlation between visual semantic units of the image and textual semantic units of the generated sentence. The results show that even without the advanced graph representation, our method can considerably improve the caption quality through fine-grained semantic matching.
- Our model improves the performances on all metrics against the RL-based methods. In particular, our model wins the gains of 14.0%/12.7%, 11.4%/11.5% in terms of BLEU-4/CIDEr-D compared with SCST and N-step, respectively. Significantly, CAVP applies a more complex multi-agent framework and incorporates visual context for better visual reasoning. Even so, our model is still superior in all metrics. This indicates the advantage of our method, which can fuse the local and global semantic similarities to explicitly optimize the agent to better learn the latent semantic correlation between images and generated captions and achieve their fine-grained semantic matching.
- ORT, DGT are based on transformer, which is more powerful than the LSTM we used. Moreover, both of them improve the current Transformer from different aspects to promote the performances, where the former incorporates geometric attention and the latter integrates the structured graphs. Hence, it is unfair to compare with them. Nevertheless, our model is still superior in ROUGE-L and CIDEr-D, which fully demonstrates its capacity.

4.4. Ablation studies

In the above experiment, our full model uses the reward that fuses the local and global semantic similarities (the fusion weight λ is set to 0.2) for RL training. To more intuitively verify the effectiveness of the designed local semantic similarity evaluation mechanism, we decompose the fused multi-level similarities in this section. Specifically, we remove the local semantic similarity Sim_{loc} by setting the λ in Eq. (8) to be 0, that is, only the CIDEr-based global semantic similarity Sim_{glo} is remained to optimize the agent (caption generation network). The setup details of the caption generation network are the same to Table 1. As shown in Table 3, we give the corresponding variant results.

From Table 3, we can observe that when the local semantic similarity Sim_{loc} is removed, the performances on all metrics are dropped considerably. This confirms the effectiveness

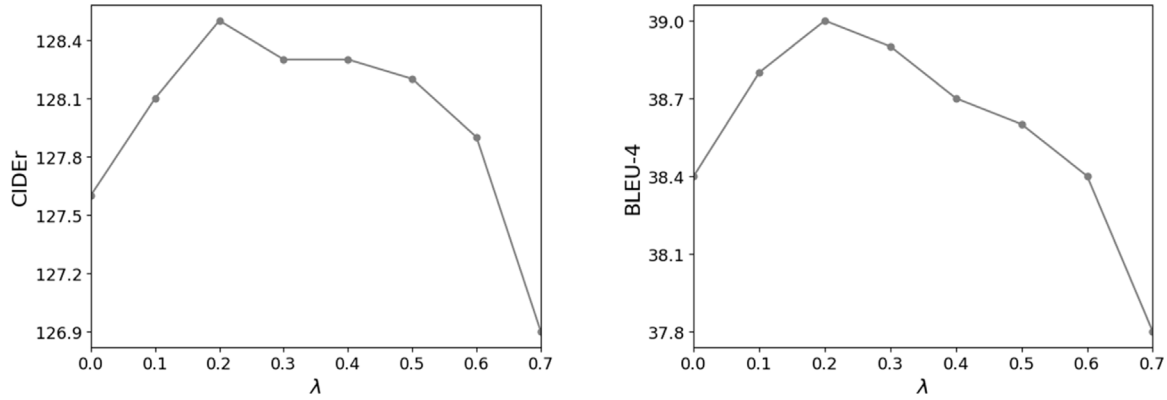

Fig. 3. Analysis of the fusion weight λ on MSCOCO. From the Figure, we can observe that $\lambda = 0.2$ is the best choice.

Fig. 4. Visualization examples of the proposed method on MSCOCO. GT stands for one human annotated caption, SGAE (Yang et al., 2019) is a considerable comparison model, and Our is the generated caption of our model. The visual and textual semantic units are highlighted in red, respectively.

Table 3
Ablation analysis of the proposed method on MSCOCO. ✓means "used", while None means "removed".

Sim_{loc}	Sim_{glo}	B-1	B-4	M	R	C	S
	✓	80.9	38.4	28.2	58.6	127.6	21.5
✓	✓	**81.2**	**39.0**	**28.5**	**58.9**	**128.5**	**21.9**

of matching visual and textual semantic units, which is helpful for explicitly optimizing the agent to learn the latent visual–textual semantic correlation. The joint learning with the local and global two-level similarities can achieve better semantic matching between images and generated captions.

Besides, we analyze the fusion weight λ and evaluate its effect on caption quality. The corresponding results are presented in Fig. 3. We can see that the evaluation results are the best when $\lambda = 0.2$. Moreover, when λ is too high ($\lambda > 0.6$), the performances drop sharply. This indicates that the weight of local semantic similarity cannot be too large and the fusion must be performed on the basis of efficient global semantic. Thus, we set λ to 0.2 during the experiments.

4.5. Qualitative analysis

To better reveal the proposed method, Fig. 4 shows some visualization examples on MSCOCO. In the figure, the ground truth and generated sentence of each image, as well as the corresponding visual and textual semantic units (red highlight) are presented. We can observe that the generated sentence of our model can be consistent with the ground truth and reflect the image global content well. For example in Fig. 4(f), compared with the false description (“on a field”, “talking on a cell phone”) of SGAE, our model can accurately capture the “suitcase” and describe the scene as “in the woods”. Furthermore, benefit from the guiding of visual–textual local semantic similarity, the textual semantic units (textual objects) of generated sentence can be highly matched with the visual semantic units (visual objects) of image, such as “cat”, “bed”, “book” in Fig. 4(b) and “truck”, “hay” (SGAE wrongly describe as “hill”) in Fig. 4(c). It indicates the proposed method can accurately find the correspondence between visual and textual objects, to achieve the fine-grained semantic matching of images and generated captions.

5. Conclusion

This paper presents a Multi-Level Similarity-Guided Semantic Matching image captioning method, including a caption generation network and a multi-level similarity fusion network. In particular, our method can fuse the local and global two-level similarities to achieve better semantic matching between images and generated captions. To this end, we simultaneously extract the visual semantic units of the image and the textual semantic units of the generated sentence. We design a local semantic similarity evaluation mechanism to explicitly measure the correlation between the visual and textual semantic units, which is helpful to reward the agent(caption generation network) to implement the fine-grained semantic matching. We fully prove the advantage of the proposed method through both quantitative and qualitative experiments on the MSCOCO dataset.

Ethical approval

This study does not contain any studies with Human or animal subjects performed by any of the authors.

CRediT authorship contribution statement

Jiesi Li: Conceptualization, Methodology. **Ning Xu:** Data curation, Investigation. **Weizhi Nie:** Methodology, Writing – review & editing. **Shenyuan Zhang:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62002257), and the China Postdoctoral Science Foundation (2021M692395).

References

Anderson, P., Fernando, B., Johnson, M., Gould, S., 2016. Spice: Semantic propositional image caption evaluation. In: ECCV. 9909, pp. 382–398.
Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR. pp. 6077–6086.
Banerjee, S., Lavie, A., 2005. METEOR: AN automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop on MT. pp. 65–72.
Chen, X., Ma, L., Jiang, W., Yao, J., Liu, W., 2018. Regularizing RNNs for caption generation by reconstructing the past with the present. In: CVPR. pp. 7995–8003.
Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T., 2017. SCA-CNN: SPatial and channel-wise attention in convolutional networks for image captioning. In: CVPR. pp. 6298–6306.
Church, K.W., 2017. Word2Vec. Natural Lang. Eng. 23 (1), 155–162.
Dong, X., Long, C., Xu, W., Xiao, C., 2021. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In: ACM MM. pp. 2615–2624.
Gao, J., Wang, S., Wang, S., Ma, S., Gao, W., 2019. Self-critical n-step training for image captioning. In: CVPR. pp. 6300–6308.
Gu, J., Joty, S.R., Cai, J., Zhao, H., Yang, X., Wang, G., 2019. Unpaired image captioning via scene graph alignments. In: ICCV. pp. 10322–10331.
Gu, J., Wang, G., Cai, J., Chen, T., 2017. An empirical study of language CNN for image captioning. In: ICCV. pp. 1231–1240.
Herdade, S., Kappeler, A., Boakye, K., Soares, J., 2019. Image captioning: Transforming objects into words. In: NIPS. pp. 11135–11145.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Honnibal, M., Montani, I., 2017. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. 7, (1), pp. 411–420, (in press).
Jiang, W., Ma, L., Chen, X., Zhang, H., Liu, W., 2018a. Learning to guide decoding for image captioning. In: AAAI. pp. 6959–6966.
Jiang, W., Ma, L., Jiang, Y., Liu, W., Zhang, T., 2018b. Recurrent fusion network for image captioning. In: ECCV. 11206, pp. 510–526.
Karaoglu, S., Tao, R., Gevers, T., Smeulders, A.W.M., 2017. Words matter: Scene text for image classification and retrieval. IEEE Trans. Multimedia 19 (5), 1063–1076.
Karpathy, A., Li, F., 2015. Deep visual-semantic alignments for generating image descriptions. In: CVPR. pp. 3128–3137.
Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: ICLR.
Kong, X., Xin, B., Wang, Y., Hua, G., 2017. Collaborative deep reinforcement learning for joint object search. In: CVPR. pp. 7072–7081.
Krull, A., Brachmann, E., Nowozin, S., Michel, F., Shotton, J., Rother, C., 2017. Poseagent: Budget-constrained 6d object pose estimation via reinforcement learning. In: CVPR. pp. 2566–2574.
Lin, C., Hovy, E.H., 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: NAACL.
Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: ECCV. pp. 740–755.
Liu, D., Zha, Z., Zhang, H., Zhang, Y., Wu, F., 2018. Context-aware visual policy network for sequence-level image captioning. In: ACM MM. pp. 1416–1424.
Liu, S., Zhu, Z., Ye, N., Guadarrama, S., Murphy, K., 2017. Improved image captioning via policy gradient optimization of spider. In: ICCV. pp. 873–881.
Lu, J., Xiong, C., Parikh, D., Socher, R., 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: CVPR. pp. 3242–3250.
Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D., 2014. The stanford corenlp natural language processing toolkit. In: ACL. pp. 55–60.
Papineni, K., Roukos, S., Ward, T., Zhu, W., 2002. Bleu: A method for automatic evaluation of machine translation. In: ACL. pp. 311–318.
Ranzato, M., Chopra, S., Auli, M., Zaremba, W., 2016. Sequence level training with recurrent neural networks. In: ICLR.
Ren, S., He, K., Girshick, R.B., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99.
Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L., 2017. Deep reinforcement learning-based image captioning with embedding reward. In: CVPR. pp. 1151–1159.
Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., 2017. Self-critical sequence training for image captioning. In: CVPR. pp. 1179–1195.
Vedantam, R., Zitnick, C.L., Parikh, D., 2015. Cider: Consensus-based image description evaluation. In: CVPR. pp. 4566–4575.
Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator. In: CVPR. pp. 3156–3164.
Wang, C., Yang, H., Bartz, C., Meinel, C., 2016. Image captioning with deep bidirectional LSTMs. In: ACM MM. pp. 988–997.

Williams, R.J., 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. 8 (3–4), 229–256.
Wu, S., Wieland, J., Farivar, O., Schiller, J., 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In: ACM CSCW. pp. 1180–1192.
Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: ICML. 37, pp. 2048–2057.
Xu, N., Zhang, H., Liu, A., Nie, W., Su, Y., Nie, J., Zhang, Y., 2020. Multi-level policy and reward-based deep reinforcement learning framework for image captioning. IEEE Trans. Multimedia 22 (5), 1372–1383.
Yang, X., Tang, K., Zhang, H., Cai, J., 2019. Auto-encoding scene graphs for image captioning. In: CVPR. pp. 10685–10694.
Yang, L., Wang, H., Tang, P., Li, Q., 2021. CaptionNet: A Tailor-made recurrent neural network for generating image descriptions. IEEE Trans. Multimedia 23, 835–845.
Yao, T., Pan, Y., Li, Y., Mei, T., 2018. Exploring visual relationship for image captioning. In: ECCV. 11218, pp. 711–727.
Yun, S., Choi, J., Yoo, Y., Yun, K., Choi, J.Y., 2017. Action-decision networks for visual tracking with deep reinforcement learning. In: CVPR. pp. 1349–1358.
Zhang, L., Sung, F., Liu, F., Xiang, T., Gong, S., Yang, Y., Hospedales, T.M., 2017. Actor-critic sequence training for image captioning, in: NIPS Workshop on Visually-Grounded Interaction and Language.
Zhou, Y., Wang, M., Liu, D., Hu, Z., Zhang, H., 2020. More grounded image captioning by distilling image-text matching model. In: CVPR. pp. 4776–4785.