Discrete-continuous Action Space Policy Gradient-based Attention for Image-Text Matching

Shiyang Yan[1]*, Li Yu[1], Yuan Xie[2]*
[1]Nanjing University of Information Science and Technology, Nanjing, China
[2]East China Normal University, Shanghai, China
elyotyan@gmail.com, li.yu@nuist.edu.cn, yxie@cs.ecnu.edu.cn

Abstract

Image-text matching is an important multi-modal task with massive applications. It tries to match the image and the text with similar semantic information. Existing approaches do not explicitly transform the different modalities into a common space. Meanwhile, the attention mechanism which is widely used in image-text matching models does not have supervision. We propose a novel attention scheme which projects the image and text embedding into a common space and optimises the attention weights directly towards the evaluation metrics. The proposed attention scheme can be considered as a kind of supervised attention and requiring no additional annotations. It is trained via a novel Discrete-continuous action space policy gradient algorithm, which is more effective in modelling complex action space than previous continuous action space policy gradient. We evaluate the proposed methods on two widely-used benchmark datasets: Flickr30k and MS-COCO, outperforming the previous approaches by a large margin.

1. Introduction

Computer Vision and Natural Language Processing are two important areas of modern artificial intelligence, which can be processed jointly in cross-modal tasks. A large amount of research has been conducted to bridge the vision and language modalities [32, 35, 5, 19, 18]. Image-text matching or retrieval is one of the critical topics in this area, which has a huge application scope in many real-world scenarios. The image-text matching requires a machine learning model to extract the high-level semantic representations and measure the similarities across modalities accordingly.

Existing methods use deep learning models to extract the image and language features, and apply various metric learning techniques to automatically find the semantic similarities between the samples from the two modalities [5, 19, 18]. Metric learning is powerful in visual semantic embedding as it tries to measure and manipulate the similarities between samples regardless of the domain differences. However, it is not designed to perform an explicit transformation from one modality to another, often leading to sub-optimal performance. Though there are approaches apply Instance Loss [37], i.e., a classification over image and text categories, to form a multi-task learning approach with metric learning loss for image-text matching, the performance gain is limited as the Instance Loss optimises the embedding in the category domain, which does not perform explicit transformation either. An image often contains many fine-grained objects. A flat vector representation from a vanilla deep CNN model such as ResNet [9] is not powerful enough to discover these objects and their relationships. Hence, advanced methods employ image features from a pre-trained object detector [28] and apply the visual attention mechanism [35] on these features to discriminate the important features over irrelevant ones [1]. Attention mechanism plays a significant role in varying computer vi-

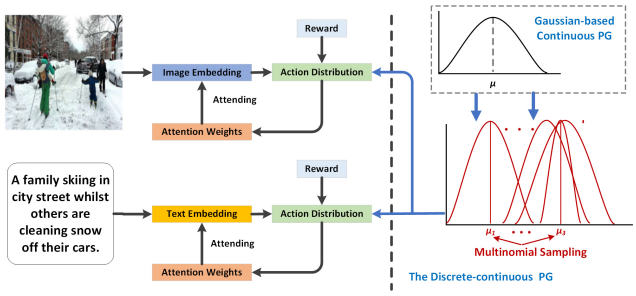

Figure 1: Our Motivations: The attention weights are utilised as a projection from each modality to a common space. Existing continuous PG assumes a simple Normal distribution. Instead, we considered the mean values as discrete actions first and then use multiple Normal distributions to form a compound distribution, which is more realistic. *Best Viewed in Colour.*

*Corresponding Author

sion tasks. It is considered as hidden neurons in these models, but often leads to incorrect selection of image features for lacking a direct supervision [23].

In this paper, to make an explicit transformation and provide supervision to the attention mechanism in image-text matching, we propose a policy gradient (PG) [30] optimised attention adjustment over the visual and text features in image-text matching. The attention weights in our approach can be considered as a transformation from a specific modality to a common space, as the attention weights perform a vector transformation in the last image and text vectors used for matching, instead of selecting important features in the previous layers of the deep learning models [35]. The attention weights are trained by the PG method with the batch-wise ranking metrics and the instance-wise Average Precision (AP) as the reward function. These attention weights are directly optimised via PG algorithm to achieve optimal ranking results and higher AP metrics. It can be considered as a kind of supervised attention mechanism, and this supervision does not need any additional annotations. This PG-based attention mechanism is straightforward and is optimised towards the evaluation metrics. It is also more accurate than the conventional soft attention which is only a regular neuron.

To be more specific, as shown in Figure 1, we consider the attention weights generation as an action selecting process in PG, whose space can be flexibly pre-defined. The action space in conventional PG is discrete, which is not suitable for the feature adjustment like in the attention mechanism. One solution is applying a continuous action space PG algorithm [20], which consider the action space as a Gaussian distribution and sample action values from this distribution. Restricting the action distribution to Normal is not optimal, and such a hypothesis lacks theoretical and practical support. In reality, the distribution of the action space might be very complex, which cannot be described via a simple Normal distribution. Hence, we consider the action is continuous and sampled from multiple Normal distributions with a different mean (μ) and standard deviation (σ) values. We first treat the μ as discrete actions, sampled from a pre-defined action space while σ is obtained from a neural model as it is continuous. We want to use this μ and σ to form a Normal distribution and sample continuous action from this distribution, which is applied as the attention weights to adjust the feature representations for both the visual embedding and the text embedding. Usually, in conventional PG, we do not need the μ to be trainable as we only back-propagate the gradient to the log-probabilities. In contrast, in this case, the subsequently obtained Normal distributions need the μ being able to be back-propagated to make the Normal distribution learn-able. As there involves sampling in obtaining the μ, it is not trainable in this current form. To make this μ differentiable, instead of directly using greedy sampling or ϵ-greedy sampling. We use a Gumbel-softmax to relax the discreteness [12] and make the sampled μ trainable together with the Normal distribution. We call this method "Discrete-continuous PG" as it involves both discrete and continuous action space, making them benefit from each other. In fact, by using the discrete and continuous action space, the action space used to sample the attention weights is a compound distribution, which can model a high complex distribution. We evaluate our algorithm and model on image-text matching task, achieving state-of-the-art performance on two widely-used benchmark datasets. To summarise, our contributions are threefold: (1) We propose a novel attention supervision scheme for image-text matching task based on policy gradient. (2) We propose a new Discrete-continuous policy gradient algorithm by leveraging both the discrete and continuous action space. (3) The achieved state-of-the-art results validate the effectiveness of the attention supervisions scheme and the novel policy gradient algorithm.

2. Related Works

2.1. Image-text Matching

Frome et al. [6] propose a feature embedding method via CNNs and Skip-Gram for cross-modal matching. They also utilise a ranking loss to measure the distance between similar pairs. Faghri et al. [5] focus on the hard negative mining in the Triplet loss, with improved results. Zheng et al. [37] utilise an Instance Loss over a large number of categories. They find that the Instance Loss is helpful in image-text matching. Gu et al. [8] improve the cross-modal problem by looking into the generative models. Li et al. [19] propose a visual semantic reasoning framework by leveraging graph neural networks and image captioning loss. The visual semantic reasoning model can reason on the semantic relationship of the image features, with good performance.

2.2. Attention Mechanism

The visual attention mechanism [35] has been widely applied in many types of computer vision applications. Notably, bottom-up attention model [1] is the current mainstream for image captioning, visual question answering and image-text matching. However, there is not much research on supervised attention. Gan et al. [7] propose a supervised attention scheme on visual question answering using attention annotations. Kamigaito et al. [13] also use attention annotations for supervised attention in natural language processing task. Instead, we propose a supervised attention mechanism based on reinforcement learning, which can make the attention module directly optimise towards a specific goal such as AP. Also, the proposed attention module does not need any additional annotations.

2.3. Continuous Action Space Policy Gradient

The continuous control problem has long been investigated. For example, Lillicrap et al. [20] propose the deep deterministic policy gradient by considering a continuous action space. Previous research has exploited the relationship between discrete and continuous action space. For instance, Dulacc-Arnold et al. [3] leverage the continuity in the underlying continuous action space for generalisation on discrete actions. Pazis et al. [27] convert the continuous control problem into discrete ones, by using a binary discrete action space. Tang et al. [31] show that discretizing action space for continuous control is a simple yet powerful technique for on-policy optimisation. We also consider the combination of discrete and continuous action space for on-policy optimisation. We prove that a compound distribution is superior to a strict assumption of one Normal distribution.

3. The Proposed Method

Our goal is to adjust the generated visual and text features to facilitate the image-text matching. We first apply Graph Convolutional Neural Networks [34] on the bottom-up attention [1] features of the images, which is similar to the Visual Semantic Reasoning Networks (VSRN) [19]. Once the visual features are obtained, we then use our Discrete-continuous action space PG to generate the attention weights, which are used to adjust the visual features. Similarly, the text features are also adjusted via the proposed Discrete-continuous PG-based attention mechanism. The obtained image and text embedding are trained via multi-task loss, including the Triplet Loss, Instance Loss and Text Decoding Loss. A schematic diagram of the proposed method is shown in Figure 2.

3.1. Image and Text Features Extraction

GCN for image region features reasoning. We apply a GCN model similar to VSRN approach [19]. Specifically, the semantic relationship between image region features is measured via pairwise affinity.

$$Relation(F_i, F_j) = E_i(F_i)^T E_i(F_j), \tag{1}$$

where F_i and F_j are two bottom-up image region features obtained from Faster R-CNN detectors. E_i and E_j are embedding functions, which are usually matrices multiplications, which are can be learnt via backpropagation.

Then a fully-connected relationship graph $G_r = (V, E)$ is constructed. V is the set of detected image region features and E is the set of edges where each of the edges is described by the affinity matrices $Relation(F_i, F_j)$, which is presented in Equation 1. We apply the GCN to perform reasoning on this fully-connected graph. The output of the GCN reasoning is denoted as $Image = \{I^1, ..., I^t, ..., I^T\}$.

Text Embedding. Given one-hot text representations, represented as w, a linear word embedding layer is constructed to obtain the word representations, represented as $W_e = \{w_e^1, ..., w_e^i, ..., w_e^N\}$, where $w_e^i = word\text{-}embedding(w^i)$.

3.2. The Proposed Discrete-continuous Action Space PG

PG is usually with discrete action space for two reasons: many control problems are modeled in discrete action space which leads to high performance as it can model complex action distribution. However, when meeting with continuous action space control problem, we have to develop corresponding PG algorithms. However, as discussed previously, continuous action space PG normally assumes the actions follow a Normal distribution, which is too strict. We propose an approach to essentially sample the continuous action from a compound distribution, which can better model the real distribution.

Discrete Action Sampling. As shown in Figure 2, we first model the attention weights generation process as a finite Markov Decision Process (MDP) and sample a discrete action by using Multinomial Sampling. We define n action categories, i.e., $A = \{a_1, a_2, ..., a_n\}$, The state space contains the input region features and the attention weights generated so far, which are $s_t = \{I^0, Att^0 ..., I^{t-1}, Att^{t-1}\}$. The policy is parametrised via a GRU model to explore the environment and sample the action. More formally:

$$\begin{aligned}
&h^t = GRU_{mdp}(I^t, h^{t-1}), \;\; t = 1, ..., T, \\
&F_I^t = h^t, \\
&a^t = F_I^t * W_\mu^t, \\
&a_g^t = Gumbel\text{-}softmax(a^t), \\
&a_s^t \sim Multinomial(a_g^t), \\
&logprob_a^t = \log a_g^t(a_s^t),
\end{aligned} \tag{2}$$

where I^t is the t_{th} image feature after the GCN reasoning. GRU_{mdp} is the Gated Recurrent Unit (GRU) used to model the attention weights generation problem as MDP. $W_\mu^t \in \mathcal{R}_{s \times n}$ are the weights need to be learnt. s is the size of the feature vector. a_g^t is the probability of each actions after the $Gumbel\text{-}softmax$ activation.

$$\begin{aligned}
&\mu^t = Logistic\left(\frac{a_s^t}{n}\right), \\
&std^t = F_I^t * W_{std}^t,
\end{aligned} \tag{3}$$

where $W_{std} \in \mathcal{R}_{s \times 1}$ are the weights need to be learnt.

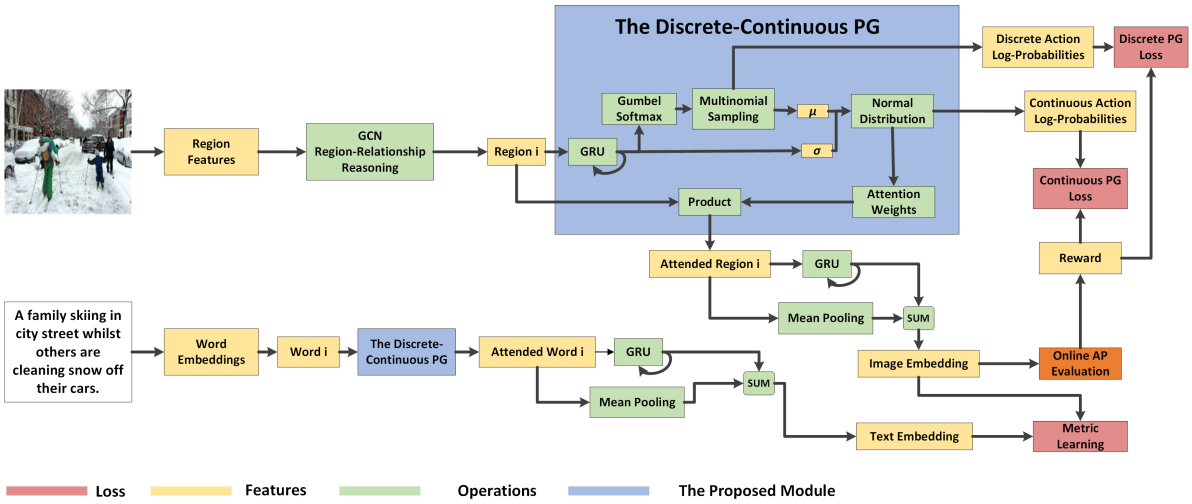

Figure 2: A schematic diagram of the proposed method: The image and text are forwarded into the model. The extracted image features are first processed via a GCN model to reason on the semantic relationships. The region features are then inputted to the proposed Discrete-continuous PG algorithm to generate attention maps, which are applied to adjust and fuse the region features subsequently. Similarly, the text embedding is also adjusted via the attention maps generated by the Discrete-continuous PG algorithm. The final image and text embedding are then connected with Metric Learning losses, the Discrete PG Loss and the Continuous PG loss for training. *Best Viewed in Colour.*

Continuous Action Sampling. The sampled μ and σ form a Normal Distribution, described as follows:

$$\begin{aligned} &Sample \sim \mathcal{N}(\mu^t, \sigma^t), \\ &Att^t = Sigmoid(Sample), \end{aligned} \tag{4}$$

where Att^t are the attention weights sampled from this particular Normal Distribution. The log probabilities of this Normal Distribution is expressed as:

$$\begin{aligned} &\log\left(f\left(Att^t; \mu^t, {\sigma^t}^2\right)\right) = \\ &-\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\left(\sigma^t\right)^2\right) \\ &-\frac{1}{2(std^t)^2}\sum\left(Att - \mu^t\right)^2. \end{aligned} \tag{5}$$

Discrete PG Optimisation. To be simple and efficient, we formulate the PG as an on-line learning method, specifically, the REINFORCE algorithm [33]. The PG for discrete action space is then to maximise the long-term reward with the following expression:

$$\begin{aligned} &\nabla_\theta J(\theta) = \\ &\mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\left(\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right)\left(\sum_{t=0}^{T} r(s_t, a_t)\right)\right], \end{aligned} \tag{6}$$

we use the one sample Monte-Carlo to approximate the accumulative reward, i.e., $\sum_{t=0}^{T} r(s_t, a_t) = \sum_{t=0}^{T}\mathcal{R}$, where $\mathcal{R}$ is the reward and will be defined later. Also, $\log \pi_\theta(a_t \mid s_t) = logprob_a^t$, which is obtained from Equation 2. Hence, Equation 6 can lead to a PG loss function as follows,

$$Loss_{PG_D} = -\sum_{b=1}^{B}\left[\left(\sum_{t=0}^{T}\nabla_\theta logprob_a^t\right)\left(\sum_{t=0}^{T}\mathcal{R}\right)\right], \tag{7}$$

where B is the size of each mini-batch. Note that the negative notation on the right-hand side means that we want to minimise the loss so as to maximise the $\mathcal{R}$.

Continuous PG Optimisation. Equation 5 provides a straightforward definition of the log probabilities of a Normal Distribution. Similarly, the PG loss for the continuous action space is presented as follows:

$$\begin{aligned} &Loss_{PG_C} = \\ &-\sum_{b=1}^{B}\left[\left(\sum_{t=0}^{T}\log\left(f\left(Att^t; \mu^t, {\sigma^t}^2\right)\right)\right)\left(\sum_{t=0}^{T}\mathcal{R}\right)\right]. \end{aligned} \tag{8}$$

Reward Function Formulation. The reward signal is of vital significance as it guides the attention generation process, which is the initial goal of PG method. The reward signal is obtained from an on-line evaluation of the image and text embedding using R@K and Average Precision (AP).

We consider a batch of samples as the gallery, and each sample as a query to compute the instance-wise AP. Specifically, we treat each of the samples as one category, and calculate the R@1 and AP of it on-line in a batch of samples. The reward signal can thus be expressed as a linear combination of the R@1 and the AP results:

$$\mathcal{R} = R@1 + AP, \tag{9}$$

we then use this reward to guide the proposed PG algorithm to generate attention weights to automatically adjust the image and text features to formulate a more effective embedding for the image-text matching task. To further reduce the variance and make the PG training more stable, we additionally apply a PG baseline, which is an average of the rewards from all the other instances within a batch, expressed as:

$$b_k = \frac{1}{K-1}\sum_{j \neq k} \mathcal{R}_j, \tag{10}$$

where K is the batch size, b_k is the baseline for k_{th} instance and R_j is the reward of j_{th} instance. We apply a coefficient $\beta = 0.5$ over the baseline, which is empirically better.

3.3. Feature Fusion

The image embedding can be adjusted by using the generated attention weights. Recall the image region features as $Image = \{I^1, ..., I^t, ..., I^T\}$, and the generated attention weights are $ATT = \{Att^1, ..., Att^t, ..., Att^T\}$, we use element-wise multiplication to adjust the image region features with the attention weights.

$$\begin{aligned} I_A &= Image * (\lambda * ATT^I), \\ h_g^t &= GRU_{gr}^I(I_A^t, h_g^{t-1}),\ t = 1, ..., T, \\ I_E &= h_g^T + [\sum_{t=1}^{T} I_A]/T, \end{aligned} \tag{11}$$

where I_A stands for adjusted image region features. GRU_{gr}^I is used to perform global reasoning of the adjusted image features. The fused features involve a summation of the outputs of the GRU_{gr} and the adjusted image region features. I_E is the image embedding.

Similarly, we apply the same approach to the text embedding generation. Note that we directly apply the proposed Discrete-continuous PG on the word embedding W_e.

Then the feature adjustment and fusion of text embedding generation can be presented as follows:

$$\begin{aligned} T_A &= W_e * (\lambda * ATT^T), \\ h_g^i &= GRU_{gr}^T(T_A^i, h_g^{i-1}),\ i = 1, ..., N, \\ T_E &= h_g^N + [\sum_{i=1}^{N} T_A]/N, \end{aligned} \tag{12}$$

Networks	Methods	Caption Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
AlexNet	DSVA [14]	22.2	48.2	61.4	15.2	37.7	50.5
	HMlstm [25]	38.1	-	76.5	27.7	-	68.8
VGG	FV [16]	35.0	62.0	73.8	25.0	52.7	66.0
	VQA [22]	33.9	62.5	73.8	25.0	52.7	66.0
	SMlstm [10]	42.5	71.9	81.5	30.2	60.4	72.3
	2wayN [4]	49.8	67.5	-	36.0	55.6	-
ResNet	RRF [24]	47.6	77.4	87.1	35.4	68.3	79.9
	VSE [5]	52.9	79.1	87.2	39.6	69.6	79.5
	SCO [11]	55.5	82.0	89.3	41.1	70.5	80.1
Faster R-CNN	SCAN [18]	67.4	90.3	95.8	48.6	77.7	85.2
	VSRN [19]	71.3	90.6	96.0	54.7	81.8	88.2
	Ours	**82.8**	**95.9**	**97.9**	**62.2**	**89.3**	**93.8**

Table 1: Comparison of the image-text matching on Flickr30k Dataset.

where T_A is the adjusted text features and ATT^T are the attention weights generated for text embedding. T_E is the text embedding.

3.4. Loss Functions

To fulfill the image-text matching task, we apply cross-modal Triplet Loss, Instance Loss, Text Decoding Loss, and together with the proposed PG loss, to train the model. The final loss objective function of the model is described as follows:

$$\begin{aligned} \mathbb{L} =& Loss_{triplet} + Loss_{xe} + loss_{td}^I + loss_{td}^T \\ &+ Loss_{PG_c}^I + Loss_{PG_d}^I + Loss_{PG_c}^T + Loss_{PG_d}^T, \end{aligned} \tag{13}$$

where $Loss_{triplet}$ is the hinge-based Triplet ranking loss [5, 14, 18]. The $Loss_{xe}$ is the cross-entropy classification loss which treats each instance as one class categories [37]. The $Loss_{td}^I$ and $Loss_{td}^T$ are the Image-to-Text Decoding Loss and Text-to-Text Decoding Loss, respectively. They decode the image or text embedding into sentences. Note the weights of the Text Decoding Module are shared between image and text branches.

The Triplet loss is expressed as follows:

$$\begin{aligned} Loss_{metric} =& [\alpha - S(I,T) + S(I,\hat{T})]_+ + \\ & [\alpha - S(I,T) + S(\hat{I},T)]_+, \end{aligned} \tag{14}$$

where α is the margin hyper-parameter.$[x]_+ = max(x, 0)$. $S(\cdot)$ is the similarity function. $\hat{I}$ and $\hat{T}$ are the hardest negatives for a positive pair (I, T).

For the Text Decoding Loss, We apply the convolutional image captioning model [2] as the decoder of the image and text decoding module. We use the same loss functions as in [2], which has a parallel training capability for text decoding and is much efficient than the RNN-based one.

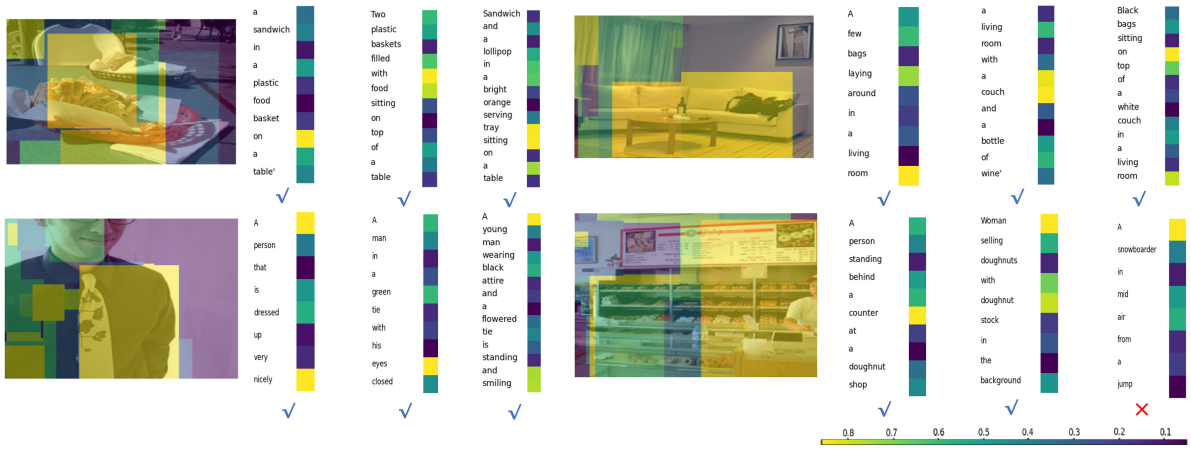

Figure 3: Visualisation of the caption retrieval results and attention mechanism. We select top 3 retrieval results where a ✓means the retrieval is correct whilst the × indicates a wrong result. *Best Viewed in Colour.*

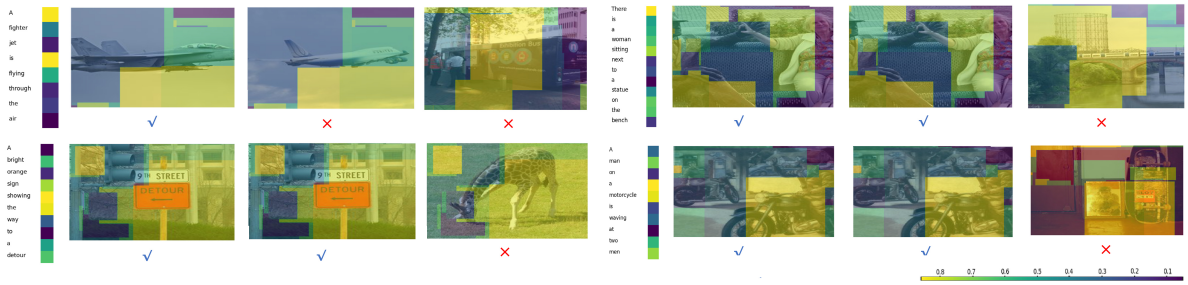

Figure 4: Visualisation of the image retrieval results and attention mechanism. *Best Viewed in Colour.*

Networks	Methods	Caption Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
AlexNet	DSVA [14]	38.4	69.9	80.5	27.4	60.2	74.8
	HMlstm [25]	43.9	-	87.8	36.1	-	86.7
VGG	FV [16]	39.4	67.9	80.9	25.1	59.8	76.6
	VQA [22]	50.5	80.1	89.7	37.0	70.9	82.9
	SMlstm [10]	53.2	83.1	91.5	40.7	75.8	87.4
	2wayN [4]	55.8	75.2	-	39.7	63.3	-
ResNet	RRF [24]	56.4	85.3	91.5	43.9	78.1	88.6
	VSE [5]	64.6	89.1	95.7	52.0	83.1	92.0
	GXN [8]	68.5	-	97.9	56.6	-	94.5
	SCO [11]	69.9	92.9	97.5	56.7	87.5	94.3
Faster R-CNN	SCAN [18]	72.7	94.8	98.4	58.8	88.4	94.8
	VSRN [19]	76.2	94.8	**98.2**	62.8	**89.7**	95.1
	Ours	**84.0**	**95.8**	97.8	**63.9**	88.9	**95.6**

Table 2: Comparison of the image-text matching on MS-COCO Dataset of 1K test set.

Networks	Methods	Caption Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
AlexNet	DSVA [14]	11.8	32.5	45.4	8.9	24.9	36.3
	FV [16]	17.3	39.0	50.2	10.8	28.3	40.1
VGG	VQA [22]	23.5	50.7	63.6	16.7	40.5	53.8
	OEM [10]	23.3	-	84.7	31.7	-	74.6
ResNet	VSE [5]	41.3	69.2	81.2	30.3	59.1	72.4
	GXN [8]	42.0	-	84.7	31.7	-	74.6
	SCO [11]	42.8	72.3	83.0	33.1	62.9	75.5
Faster R-CNN	SCAN [18]	50.4	82.2	90.0	38.6	69.3	80.4
	VSRN [19]	53.0	81.1	89.4	40.5	70.6	81.1
	Ours	**68.7**	**88.7**	**93.0**	**46.2**	**77.8**	**85.5**

Table 3: Comparison of the image-text matching on MS-COCO Dataset of 5K test set.

4. Experiments

To evaluate the effectiveness of the proposed Discrete-continuous PG algorithm, we follow previous research and perform two kinds of experiments which include sentence retrieval using image and image retrieval using a sentence.

4.1. Datasets and Protocols

We evaluate the performance of our method on the Flickr30K [36] and Microsoft-COCO datasets [21]. Flickr30K contains 31,783 images. Each image corresponds to 5 human-annotated text descriptions. We use the standard training, validation and testing split [14], which consist of 28,000 images, 1, 000 images and 1, 000 images, respectively. We follow the splits of [5, 8, 14, 18]

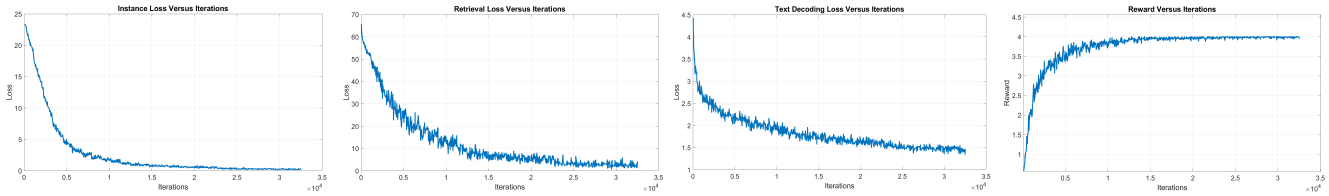

Figure 5: The Instance Loss, Triplet (Retrieval) Loss, Text Decoding Loss and Reward curves are shown in the figures.

Methods	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Triplet Loss	68.2	88.8	93.6	52.6	75.5	86.0
Triplet + Instance	69.3	89.5	93.5	52.1	77.8	87.8
Triplet + Instance + Text (**Baseline**)	70.9	89.0	93.5	52.2	78.1	87.2
Baseline + Discrete PG	78.0	93.4	94.6	56.0	80.6	89.1
Continuous PG	76.8	90.2	91.6	54.3	78.4	87.4
Multi-head Continuous PG	78.1	91.6	92.0	56.2	79.9	89.2
Baseline + Our PG (**Our scheme**)	81.0	94.5	97.1	60.6	86.5	92.4
Our Scheme + Reward R@1	79.3	94.5	96.0	57.8	82.3	90.2
Our Scheme + Reward AP	80.4	95.2	96.8	60.8	84.9	92.7
Our Scheme + Reward R@1+AP	81.0	94.5	97.1	60.6	86.5	92.4
Our scheme + ($\lambda = 10$)	80.2	95.3	96.9	58.4	82.2	90.6
Our scheme + ($\lambda = 20$)	81.0	94.5	97.1	60.6	86.5	92.4
Our scheme + ($\lambda = 30$)	79.2	94.7	96.9	57.1	83.0	90.3
Our scheme (+PG baseline)	80.3	94.6	97.5	60.8	86.9	92.4
Our scheme (+multi-head)	81.4	94.9	97.7	61.2	87.5	92.6
Our scheme (+GloVe Embedding)	**82.8**	**95.9**	**97.9**	**62.2**	**89.3**	**93.8**

Table 4: Ablation Studies on Fickr30k Dataset.

for MS-COCO dataset, which includes 113, 287 images for training, 5, 000 images for validation and 5, 000 images for testing. Each image has five captions. We use the same evaluation protocol as the previous research [5, 8, 18, 19], which is the recall performance at K (R@K) defined as the proportion of queries for which the correct item is retrieved in the nearest K samples to each query.

4.2. Implementation Details

We build our model based on PyTorch [26]. We use the pre-trained bottom-up attention image features provided by [19]. The word embedding size is 300 and the dimension of the image and text embedding is 2048. The hidden size of the GRU modules used in our model is 2048. We pre-define 100 discrete action categories which are $\{0, 1, 2, ..., a_i, ...100\}$, where a_i corresponds an action of enlarging the features with a value of a_i/λ, where λ is a hyper-parameter. Note the choice of the number of action categories is mostly empirical. We choose 100 as it is close to the maximum number of regions of an image, and also close to the maximum number of words of a sentence, which is powerful enough to describe the difference between each item of the image regions and the sentence. The detailed explanation is presented in Equation 11 and 12. For training, we apply Adam optimiser [15] to train the model with 30 epochs with a mini-batch size of 128. We start the training with a learning rate of 4e-4 for 15 epochs and lower the learning rate to 4e-5 for another 15 epochs. We apply the early stopping tricks to select the model which performs the best in the validation set. For the cross-modal Triplet ranking loss, the margin is set 0.2 for all the experiments. For the classification loss, there are 29, 783 categories for Flickr30K dataset and 113, 287 categories for MS-COCO dataset. We perform all the experiments on a server equipped with an Nvidia Geforce 2080-TI GPU card, and a Windows 10 operating system.

4.3. Comparison with the State-of-the-art

Results on Flickr30k. We show the results on the Flickr30k dataset and comparison with the current state-of-the-art methods in Table 1. We also indicate the backbone networks that used for each of the state-of-the-art methods, such as AlexNet [17], VGG [29], ResNet [9], Faster R-CNN [28]. The proposed method outperforms other approaches by a large margin. SCAN [18] and VSRN [19] are two approaches that close to ours. Our method is different from them mainly on the proposed PG-based supervised feature attention mechanism as both VSRN and our method use the same cross-modal Triplet loss and the Text Decoding Loss. Hence, the main performance gain is from the proposed Discrete-continuous PG algorithm, which is effective in improving the existing baseline model that is similar to the VSRN model [19]. Specifically, we achieve 82.8% R@1 in captioning retrieval using the image, and 62.2% R@1 image retrieval using the caption.

Results on MS-COCO. We present the experimental results on the 1K and 5K MS-COCO dataset and comparison with the state-of-the-art models in Table 2 and Table 3, respectively. For the 1K testing protocol, the results are obtained by averaging over 5 folds of 1K test images. When comparing with the current best method SCAN [18] and VSRN [19], we follow the same strategy to combine results from two trained proposed models by averaging their predicted similarity scores. As shown in Table 2, our proposed model achieves 84.0% R@1 on caption retrieval using an image, and 63.9 % R@1 on image retrieval using the caption, respectively. The results outperform the VSRN and SCAN by a large margin. For the 5K testing protocol, we evaluate the proposed model by using the whole 5K testing samples. From Table 3, it is obvious that our method

achieves the new state-of-the-art, with 68.7% R@1 and on 46.2% R@1 on caption retrieval using image and image retrieval using the caption, respectively.

4.4. Ablation Studies

Baseline. We perform ablation studies on each component of the proposed model. which are shown in Table 4. We first evaluate the model with only Triplet Loss, with relatively poor results. Adding an Instance Loss to the model brings an limited increase in the ranking results. Similarly, the Text Decoding Loss also improves the performance of the model, which proves that it is helpful to narrow the domain gap between different modalities. Our baseline model include all of the three Loss functions.

The Impact of the Discrete-continuous PG Method. Based on the baseline model, to validate the superiority of the proposed Discrete-continuous action space policy gradient algorithm, we first compare it with the conventional discrete action space policy gradient scheme. To realise the Discrete PG scheme, we remove the continuous action space sampling and utilise the discrete action directly as the attention weights. The proposed method yields better results than the Discrete PG scheme. Second, we solely apply a single Gaussian-based continuous action space PG scheme. The results of our scheme is also better than the single Gaussian PG as we form a complex distribution which better describe the real distribution of the action space, the results are shown in Table 4.

The Impact of Different Reward Function. We then perform ablation studies on the reward function, the results show that using the batch-wise R@1 combined the instance-wise AP as the reward has the best performance. Note that AP alone is better than R@1 reward, as the AP evaluation is more comprehensive and instance-wise reward is more accurate than the batch-wise one. To further reduce the variance and make the PG training more stable, we additionally apply a PG baseline. The impact of the PG baseline is evaluated subsequently, which yields a slightly better performance as the PG baseline can stabilise the training and reduce the variance of this on-line PG method.

The Impact of the Different Values of λ. We evaluate the proposed method which largely improves the performance in our ablation studies, with more than 5% increase on the R@1 metric of both the image and caption retrieval. The value of λ controls the scale of the attention weights, which is with significant importance. The ablation studies show that a suitable value of λ (20) is critical in maintaining good performance, though our method with different λ is all with superior results.

The Impact of Applying a Multi-head Mechanism. Multi-head Mechanism is widely applied in well-known models like Transformer, often with extra improvement. We validate the positive effect of this multi-head mechanism on the proposed PG algorithm. Specifically, we apply a multi-head mechanism on the latent discrete μ and σ values with a head number of 2. Increasing the head number would increase the computing burden, which is less practical. The empirical results reveal that the multi-head mechanism can improve the performance, by essentially reflecting different aspects of the sampled latent distribution.

The Impact of Utilising a Pre-trained GloVe Word Embedding. In the vanilla VSRN baseline, the word embedding module is trainable. We investigate the impact of a pre-trained GloVe Word Embedding module as shown in the table. Applying a pre-trained GloVe embedding can improve the matching performance slightly as it embeds some prior information.

4.5. Visualisations

We visualise the retrieval results and attention maps of both the image and text in Figure 3 and Figure 4. It is clear from the figures that the attention maps can capture the expected image regions, and the language attention maps can reflect the important semantics. Some incorrect examples are also provided in the figures, which have similar semantic contents or have similar visual layouts. Visualisation on the training loss curves and the reward function curve are presented in Figure 5. The Triplet loss, Instance Loss and Text Decoding Loss all decrease as the training is performed. The reward value increases which validates the proposed Discrete-continuous PG method.

5. Conclusions

In this paper, we propose a novel policy gradient-based attention mechanism to transform the image and text embedding to a common space and optimise them towards higher AP. To model complex action space in the attention weights sampling, we propose a Discrete-continuous action space policy gradient algorithm, with a compound action space distribution. Comprehensive experiments on two widely-used benchmark datasets validate the effectiveness of the proposed method, leading to state-of-the-art performance.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (61772524, 62002172), the Natural Science Foundation of Shanghai (20ZR1417700) and CAAI-Huawei MindSpore Open Fund.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Jyoti Aneja, Aditya Deshpande, and Alexander Schwing. Convolutional image captioning. In *CVPR*, 2018.

[3] Gabriel Dulac-Arnold, Richard Evans, Hado van Hasselt, Peter Sunehag, Timothy Lillicrap, Jonathan Hunt, Timothy Mann, Theophane Weber, Thomas Degris, and Ben Coppin. Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679*, 2015.

[4] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017.

[5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[7] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *ICCV*, 2017.

[8] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, 2018.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016.

[10] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, 2017.

[11] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, 2018.

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

[13] Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. Supervised attention for sequence-to-sequence constituency parsing. In *IJCNLP*, 2017.

[14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[16] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

[19] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019.

[20] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *ICLR*, 2016.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[22] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *ECCV*, 2016.

[23] Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Neural machine translation with supervised attention. In *COLING*, 2016.

[24] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *ICCV*, 2017.

[25] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *ICCV*, 2017.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[27] Jason Pazis and Michail G Lagoudakis. Binary action search for learning continuous-action control policies. In *ICML*, 2009.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 2000.

[31] Yunhao Tang and Shipra Agrawal. Discretizing continuous action space for on-policy optimization. In *AAAI*, 2020.

[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[33] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[34] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[36] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New

similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[37] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.