Densely Supervised Hierarchical Policy-Value Network for Image Paragraph Generation

Siying Wu , **Zheng-Jun Zha*** , **Zilei Wang** , **Houqiang Li** and **Feng Wu**

National Engineering Laboratory for Brain-inspired Intelligence Technology and Application, University of Science and Technology of China

wsy315@mail.ustc.edu.cn, {zhazj,zlwang,lihq,fengwu}@ustc.edu.cn

Abstract

Image paragraph generation aims to describe an image with a paragraph in natural language. Compared to image captioning with a single sentence, paragraph generation provides more expressive and fine-grained description for storytelling. Existing approaches mainly optimize paragraph generator towards minimizing word-wise cross entropy loss, which neglects linguistic hierarchy of paragraph and results in "sparse" supervision for generator learning. In this paper, we propose a novel Densely Supervised Hierarchical Policy-Value (DHPV) network for effective paragraph generation. We design new hierarchical supervisions consisting of hierarchical rewards and values at both sentence and word levels. The joint exploration of hierarchical rewards and values provides dense supervision cues for learning effective paragraph generator. We propose a new hierarchical policy-value architecture which exploits compositionality at token-to-token and sentence-to-sentence levels simultaneously and can preserve the semantic and syntactic constituent integrity. Extensive experiments on the Stanford image-paragraph benchmark have demonstrated the effectiveness of the proposed DHPV approach with performance improvements over multiple state-of-the-art methods.

1 Introduction

Describing visual content with natural language is an emerging and crucial interdisciplinary task at the intersection of computer vision, natural language processing and artificial intelligence. Recent works have steadily improved the performance of describing images with a single sentence[Vinyals *et al.*, 2015; Xu *et al.*, 2015; Mao *et al.*, 2014; You *et al.*, 2016]. However, compressing an image into a single sentence could only present a coarse description of the rich visual content. One recent alternative is image paragraph generation which aims to describe images with a paragraph of detailed and fine-grained stories [Krause *et al.*, 2017; Liang *et al.*, 2017; Chatterjee and Schwing, 2018; Che *et al.*, 2018; Wang *et al.*, 2018; Mao *et al.*, 2018; Zha *et al.*, 2019].

A few of image paragraph generation approaches have been proposed very recently. For example, [Krause *et al.*, 2017] proposed a hierarchical Recurrent Neural Network for image paragraph generation. [Chatterjee and Schwing, 2018] designed "coherence vectors," "global topic vectors," and incorporated the *Variational Autoencoder* [Kingma and Welling, 2013] into paragraph generator. [Che *et al.*, 2018] explored visual relationships among objects to improve paragraph generation. However, these approaches optimize paragraph generator by minimizing cross-entropy loss between the generated paragraph and ground truth. Such word-wise matching loss neglects the hierarchical linguistic structure within paragraph. This results in "sparse" supervision cues, hindering the learning of effective paragraph generator, especially when it is complicated. Moreover, these methods enforce the generator to maximize the likelihood of generating target sequence given ground-truth paragraph, resulting in "exposure bias" in inference [Bengio S, 2015]. In particular, [Liang *et al.*, 2017] proposed an adversarial model between paragraph generator and multi-level discriminators, towards diverse and coherent paragraph descriptions. However, the guiding signal from discriminators is sparse and not necessarily preserve effective information for generator to sufficiently learn.

In this paper, we propose a novel Densely Supervised Hierarchical Policy-Value (DHPV) network for effective image paragraph generation. We design novel hierarchical dense supervisions, consisting of hierarchical rewards and values at both word and sentence levels, for effective generator learning by exploiting the "token-sentence-paragraph" hierarchy. Moreover, the proposed hierarchical policy-value network models compositionality at token-to-token and sentence-to-sentence levels, leading to preservation of semantic and syntactic integrity. In particular, the hierarchical rewards with respect to a certain evaluation metric consist of sentence-level and word-level rewards. While the former measures the metric gain after generating a new sentence conditioned on preceding sentences, the latter evaluates the metric gain after generating a new word in current sentence. The hierarchical values consist of values at sentence and word levels. The sentence-level value evaluates the possibility of producing correct paragraph started from preceding sentences under current policy. The word-level value evaluates the possibility

*Corresponding Author

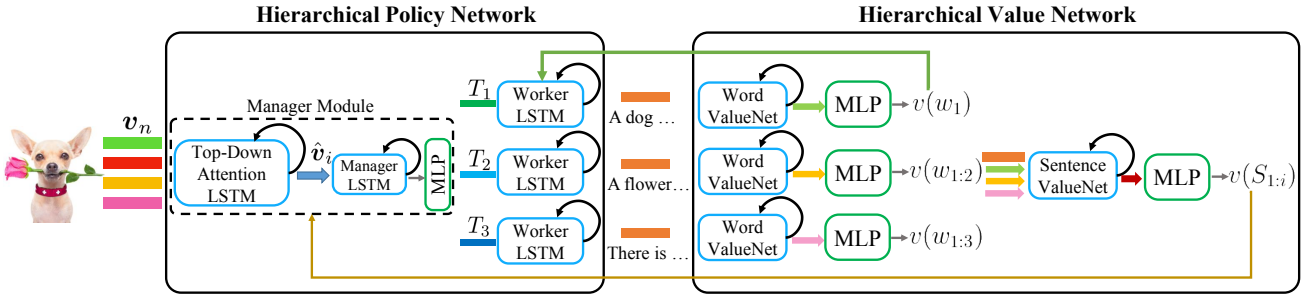

Figure 1: Overview of the proposed Densely supervised Hierarchical Policy-Value network . The hierarchical policy network, consisting of "Manager" and "Worker", produces paragraph based on visual features $\boldsymbol{v}_n$. The hierarchical value network, consisting of sentence-level and word-level value modules, evaluates the values of preceding sentences and words respectively towards generating descriptive paragraph. The hierarchical policy network is learned by exploiting both hierarchical rewards and hierarchal values.

of producing correct sentence started from preceding words. We first train the hierarchical policy network towards optimizing hierarchical rewards and then reinforce it using hierarchical values. The joint exploration of hierarchical rewards and values provides dense supervision cues for learning effective generator. The metric-specific rewards empower the generator to produce paragraphs satisfying the target metric, the metric-nonspecific values give the generator robustness across various metrics. Moreover, we avoid "exposure bias" by using model predicts during training. We conduct extensive experiments to evaluate the proposed DHPV approach on the Stanford image-paragraph benchmark [Krause *et al.*, 2017] and report superior performance over state-of-the-art methods.

The main contribution of this paper is three-fold: (1)we propose a new hierarchical policy-value network for image paragraph generation; (2)we propose new hierarchical dense supervisions, consisting of hierarchical rewards and values, towards learning effective paragraph generator; (3)the proposed approach achieves superior performance in terms of various performance metrics.

2 Related Work

Image paragraph generation aims to describe a given image with a descriptive paragraph in natural language. It has attracted increasingly attentions in recent years due to its importance for various applications. [Krause *et al.*, 2017] proposed a hierarchical recurrent neural network (RNN) to produce a generic paragraph for an image. [Liang *et al.*, 2017] proposed an adversarial model between paragraph generator and multi-level paragraph discriminations based on Generative Adversarial Network (GAN) [Goodfellow *et al.*, 2014], to encourage coherence among successive sentences within a paragraph. [Chatterjee and Schwing, 2018] proposed to augment paragraph generator with "coherence vectors," "global topic vectors," and modeling of the inherent ambiguity of associating paragraphs with images via a variational autoencoder. [Che *et al.*, 2018] explicitly explored visual relationships among objects to improve paragraph generation. [Wang *et al.*, 2018] utilized auxiliary depth maps of images to guide the linguistic decoder to reveal spatial relationships among objects, towards generating logical and coherent paragraphs. [Mao *et al.*, 2018] combined the topic embedding extracted by Latent Dirichlet Allocation (LDA) , image and context to conduct image paragraph generation. [Zha *et al.*, 2019] proposed a hierarchical context-aware visual policy network which consider previous visual attentions for the current sentence generation to product coherent and fine-grained paragraphs.

Image captioning is to describe images with a single sentence [Farhadi *et al.*, 2010; Li *et al.*, 2011; Vinyals *et al.*, 2015; Xu *et al.*, 2015; Mao *et al.*, 2014]. It can be viewed as a subtask of paragraph generation. Previous approaches for image captioning could be divided into two categories. The first category [Farhadi *et al.*, 2010; Li *et al.*, 2011; Kulkarni *et al.*, 2011] uses bottom-up paradigm to combine generated words with language models. While the second category [Vinyals *et al.*, 2015; Karpathy and Fei-Fei, 2015; Xu *et al.*, 2015; Mao *et al.*, 2014] utilizes an encoder-decoder framework, which encodes an image as visual features and then decodes features to a sentence. Recently, some works utilized sequence-based evaluation metrics as optimization objective incorporating techniques from reinforcement learning to avoid "exposure bias". For example, [Rennie *et al.*, 2017] proposed a Self-Critic Sequence Training (SCST) method to conduct image captioning. [Zhang *et al.*, 2017] proposed to train image captioning models under actor-critic framework, towards directly optimizing non-differentiable evaluation metrics of interest. Moreover, [Ren *et al.*, 2017] utilized a value network to evaluate the value of all possible extensions of the current state and provided global and lookahead guidance for inference.

3 Approach

The goal of image paragraph captioning is to describe a given image I by a paragraph $P = \{S_1, S_2, \cdots, S_L\}$. The i-th sentence, $i \in \{1, 2, \cdots, L\}$, consists of a series of words $S_i = \{w_{i,1}, w_{i,2}, \cdots, w_{i,T}\}$. In order to apply REINFORCE algorithm during training, we first cast our problem in the reinforcement learning framework. The agent, i.e. paragraph generator, chooses an action to execute according to the state. The state x_t is the given image I and the generated sentences and words $x_t = \{I, S_1, \cdots, S_{i-1}, \{w_{i,1}, w_{i,2}, \cdots, w_{i,t}\}\}$, $t \in \{1, 2, \cdots, T\}$. The action is defined as choosing next

word $w_{i,t+1}$ from the dictionary.

3.1 Hierarchical Policy-Value Network

We propose a hierarchical policy-value network to generate paragraphs for images, which consists of hierarchical policy network and hierarchical value network. The policy network is used to predict paragraph for a given image, while the value network evaluates the values of preceding sentences and words within a generated paragraph. As shown in Fig. 1, the hierarchical policy network consists of a "Manager" and a "Worker", while the hierarchical value network contains value modules at sentence and word levels.

Hierarchical Policy Network

The Manager module is used to infer latent topics for generating a paragraph, where each topic corresponds to a sentence within the paragraph. The Worker successively generates sentences following the topics provided by Manager to form a complete paragraph. The Manager consists of a top-down attention LSTM [Anderson *et al.*, 2018] which selectively focuses on relevant regions with image, a Manager LSTM combined with a Multilayer Perceptron(MLP) produces topics of sentences. The image is represented by a set of feature vectors $\boldsymbol{v} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$. The top-down attention LSTM takes the concatenation of $\mathbf{E}_{S_{1:i-1}}$ and $\bar{\boldsymbol{v}}$ as input and produces an attended ensemble $\hat{\boldsymbol{v}}_i$ by visual attention modeling as follows,

$$\begin{aligned} \boldsymbol{h}_i^{\mathrm{A}} &= \mathrm{LSTM}_{\mathrm{A}}(\boldsymbol{h}_{i-1}^{\mathrm{A}}, [\mathbf{E}_{S_{1:i-1}}, \bar{\boldsymbol{v}}]), \\ a_{n,i} &= \boldsymbol{w}_a^T tanh(W_{va}\boldsymbol{v}_n + W_{ha}\boldsymbol{h}_i^{\mathrm{A}}), \\ \hat{\boldsymbol{v}}_i &= \sum_{n=1}^{N} a_{n,i}\boldsymbol{v}_n, \end{aligned} \tag{1}$$

where $\mathbf{E}_{S_{1:i-1}}$ is the encoding of preceding sentences, $\bar{\boldsymbol{v}} = \frac{1}{N}\sum \boldsymbol{v_n}$ is the mean pooled image features. $\boldsymbol{w}_a$, W_{va},W_{ha} are trainable parameters, and $\boldsymbol{h}_i^{\mathrm{A}}$ is the hidden state.

The Manager LSTM takes the concatenation of $\hat{\boldsymbol{v}}_i$ and $\boldsymbol{h}_i^{\mathrm{A}}$ as input. The latent topics are then produced by the MLP.

$$\boldsymbol{h}_i^{\mathrm{M}} = \mathrm{LSTM}_{\mathrm{M}}(\boldsymbol{h}_{i-1}^{\mathrm{M}}, [\hat{\boldsymbol{v}}_i, \boldsymbol{h}_i^{\mathrm{A}}]), \tag{2}$$

$$\mathbf{T}_i = \mathrm{MLP}(\boldsymbol{h}_i^{\mathrm{M}}). \tag{3}$$

A fully connection layer is used to infer the probability μ_i of S_i being the last sentence of the paragraph.

$$\mu_i = \mathrm{FC}(\boldsymbol{h}_i^{\mathrm{M}}). \tag{4}$$

Worker module is a single-layer LSTM, which produces sentences S_i by sampling next word $w_{i,t+1}$ from the dictionary under the guidance of topic vector $\mathbf{T}_i$.

$$\boldsymbol{h}_{i,t}^{\mathrm{W}} = \begin{cases} \mathrm{LSTM}_{\mathrm{W}}(\mathbf{T}_i) & t = 0 \\ \mathrm{LSTM}_{\mathrm{W}}(\boldsymbol{h}_{i,t-1}^{\mathrm{W}}, w_{i,t}) & t > 0, \end{cases} \tag{5}$$

The hidden state of Worker LSTM $\boldsymbol{h}_{i,t}^{\mathrm{W}}$ is then successively fed into a fully connection layer and a softmax layer to predict next word.

$$w_{i,t+1} = \mathrm{Softmax}(\mathrm{FC}(\boldsymbol{h}_{i,t}^{\mathrm{W}})). \tag{6}$$

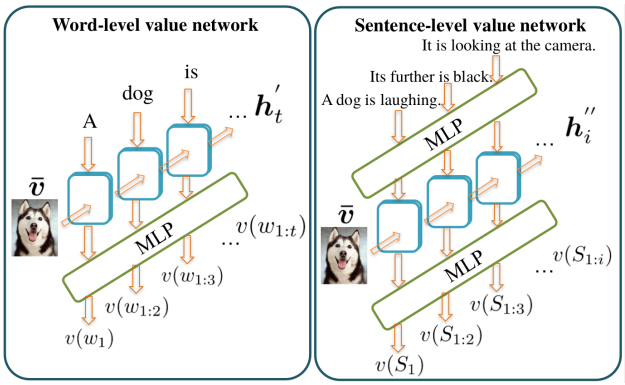

Figure 2: Illustration of hierarchical value network consisting of value evaluators at word and sentence levels. The word and sentence values are used to reinforce the "Worker" and "Manager" in hierarchical policy network, respectively.

Hierarchical Value Network

The sentence-level value network estimates the value of preceding complete sentence $S_{1:i}$ in context of the generated paragraph. The word-level value network evaluates the value of preceding sequence $w_{i,1:t}$ in context of the current sentence S_i.

As shown in Fig. 2, the word-level value network contains a LSTM and a MLP. The mean pooled image features $\bar{\boldsymbol{v}}$ and each word in the i-th sentence are fed into the LSTM in turn. The MLP computes the value of $w_{i,1:t}$ based on the output of LSTM.

$$\boldsymbol{h}_{i,t}^{'} = \begin{cases} \mathrm{LSTM}(\bar{\boldsymbol{v}}) & t = 0 \\ \mathrm{LSTM}(\boldsymbol{h}_{i,t-1}^{'}, w_{i,t}) & t > 0, \end{cases} \tag{7}$$

$$v(w_{i,1:t}) = \mathrm{MLP}(\boldsymbol{h}_{i,t}^{'}) \tag{8}$$

The sentence-level value network is also comprised of a LSTM and a MLP. For a paragraph generated by the policy network, image features $\bar{\boldsymbol{v}}$ and each sentence in paragraph are fed into the sentence-level value LSTM successively. The sentence is represented by the last hidden state of word-level value LSTM. The output of sentence-level value LSTM is processed by MLP to evaluate the value of $S_{1:i}$ in the context of paragraph.

$$\boldsymbol{h}_i^{''} = \begin{cases} \mathrm{LSTM}(\bar{\boldsymbol{v}}) & i = 0 \\ \mathrm{LSTM}(\boldsymbol{h}_{i-1}^{''}, \mathrm{MLP}(\boldsymbol{h}_{i,T}^{'})) & i > 0, \end{cases} \tag{9}$$

$$v(S_{1:i}) = \mathrm{MLP}(\boldsymbol{h}_i^{''}) \tag{10}$$

3.2 Hierarchical Dense Supervisions

We formulate new hierarchical rewards and hierarchical values, together providing dense supervisions for learning effective paragraph generator. The reward is with respect to a certain evaluation metric, while the value is independent on any metric. The metric-specific reward empowers the generator to produce paragraphs satisfying the target metric, the metric-nonspecific value gives the generator robustness across various metrics.

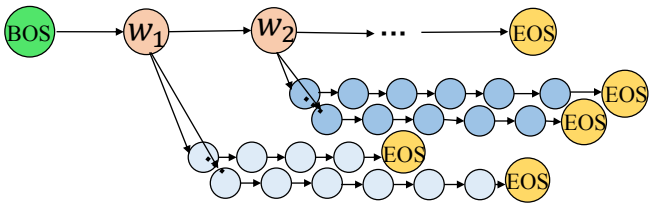

Figure 3: Illustration of the rollouts policy. The tokens in green and yellow are BOS (Beginning Of Sentence) and EOS (End Of Sentence), respectively. The pink tokens represent the words produced by current model via greedy inference, while blue tokens denote the corresponding rollout words sampling from the model.

Hierarchical Rewards

We design a hierarchical reward function, which computes rewards for each sentence in a paragraph as well as that for each word in a sentence. These rewards are used as optimization objective to train the policy network. The sentence-level reward is formulated as the metric gain after generating a new sentence started from the preceding sentences as follows,

$$r(S_i) = f(S_{1:i}, G) - f(S_{1:i-1}, G), \tag{11}$$

where $S_{1:i-1}$ is the preceding sentence generated by greedy inference, G refers to the ground truth paragraph of the given image, and $f(\cdot)$ denotes a certain evaluation metric. Positive $r(S_i)$ indicates it is benefit to generate i-th sentence, while negative means the opposite.

For each word in a sentence, we design a word-level reward to measure the metric gain brought by each word. For the sake of simplicity, we take i-th sentence in the paragraph for example. As shown in Fig. 3, at time step t, a partial sequence, denoted as $w_{1:t} = \{w_1, w_2, \cdots, w_t\}$ is generated by the current model using the greedy inference algorithm.

To allocate a reward to w_t, we apply rollouts policy [Liu *et al.*, 2017] based on $w_{1:t}$ for K times and get a series of rollout sequences:

$$\hat{S}^k_{1:t} = \{w_{1:t}; \hat{w}^k_{t+1:T}\}, k \in \{1, 2, ..., K\}. \tag{12}$$

We apply the specific metric $f(\cdot)$ to calculate scores for rollout sentences. The average of the scores is used to estimate the value of $w_{1:t}$ as follows,

$$v(w_{1:t}|g) = \frac{1}{K}\sum_{k=1}^{K} f(\hat{S}^k_{1:t}, g), \tag{13}$$

where g represents the corresponding sentence in ground truth paragraph.

Using the policy gradient to update policy network, the expected gradient exhibits high variance with $v(w_{1:t}|g)$. Further, we design word-level reward function as in Eq.(14), where the value of partial sequence $v(w_{1:t-1})$ plays the role of baseline to reduce variance.

$$r(w_t) = \begin{cases} v(w_{1:t}|g) & t = 1 \\ v(w_{1:t}|g) - v(w_{1:t-1}|g) & t > 1 \end{cases} \tag{14}$$

Intuitively, by using the policy gradient with word-level reward to reinforce the policy network, the words assigned with positive reward are more likely to be sampled. By punishing words with a negative reward can reduce its sampling probability. The word-level reward function encourages the policy network to generate more accurate descriptions.

Hierarchical Values

The hierarchical values at sentence and word levels are estimated by the value network as in Eq.(10) and Eq.(8). In order to learn the sentence-level value network, we construct training pairs, (I, P^+) and (I, P^-). P^+ is the ground truth paragraph of image I and P^- is the paragraph of other images. The sentence-level value network is trained by minimizing the MSE loss $\|v(S_{1:i}) - \alpha_s\|^2$. The α_s equals to $+0.1$ for the sentence sequence $S_{1:i}$ randomly selected from P^+. And α_s equals to -0.1 for the sentence sequence $S_{1:i}$ randomly selected from P^-.

To train the word-level value network, we construct training pairs, (I, S^+) and (I, S^-). S^+ and S^- are arbitrary sentences in P^+ and P^-, respectively. The optimization objective of word-level value network is $\|v(w_{1:t}) - \alpha_w\|^2$. Each sentence in P^+ is matching with the image I with $\alpha_w = +0.1$. However, S^- which selected randomly from P^- may match with the given image I and it is hard to define α_w for S^-. To track this issue, we adopt the retrieval model [Frome *et al.*, 2013] to embed sentence S^- and image I into a shared semantic space and calculate visual-semantic similarity $Sim(I, S^-)$, which is viewed as α_w for S^-. Our retrieval model is comprised of a RNN and two linear mapping layer. The embedding feature of sentence S is represented by the last hidden state of RNN, denoted as $h_T(S)$. Embedding sentence feature $h_T(S)$ and extracted mean pooled image feature $\bar{\boldsymbol{v}}$ are projected into a joint space by W_I and W_S. The similarity between I and S is computed as follows,

$$Sim(I, S) = \frac{W_I\bar{\boldsymbol{v}} \cdot W_S h_T(S)}{\|W_I\bar{\boldsymbol{v}}\| \cdot \|W_S h_T(S)\|}. \tag{15}$$

The parameters of RNN, W_I and W_S are learned by optimizing following loss function:

$$L = max(0, \beta - Sim(I, S^+) + Sim(I, S^-)), \tag{16}$$

where β is the margin. The optimization objective in Eq.(16) encourages model assign higher score to matching pair (I, S^+). The matching sentence S^+ used to train retrieval model can be randomly selected from the P^+. To avoid choosing a sentence similar to the given image from P^-, based on the assumption that at least one mismatching sentence exists in the P^-, we utilize current retrieval model to calculate the similarity between each sentence in P^- and I and choose the most mismatched sentence as S^- to train the retrieval model.

Training Hierarchical Policy Network

We train the policy network in two steps. We first pre-train it using standard supervised learning with the cross entropy loss in [Krause *et al.*, 2017]. The loss consists of a sentence-level loss l_s on the distribution μ_i and a word-level loss l_w on the distribution $p_{i,t}$ for the t-th word in the i-th sentence.

$$L(I, P) = \lambda_s \sum_{i=1}^{L} l_s(\mu_i, \mathbb{1}_{i=L}) + \lambda_w \sum_{i=1}^{L}\sum_{t=1}^{T} l_w(p_{i,t}, w_{i,t}), \tag{17}$$

where λ_s and λ_w are hype-parameters to balance the sentence-level and word-level cross-entropy losses.

Then, we optimize the policy network using policy gradient algorithm with hierarchical rewards. The Worker policy π_{θ_w} are optimized with Manager policy p_{θ_m} fixed by following loss function:

$$L(\theta_w) = -\mathbb{E}_{w_{i,t}\sim\pi_{\theta_w}}[r(w_{i,t})]. \tag{18}$$

Inspired by DDPG [Lillicrap *et al.*, 2015], the goal of training Manager policy p_{θ_m} is to minimize the negative expected reward with Worker policy π_{θ_w} fixed:

$$L(\theta_m) = -\mathbb{E}_{\mathbf{T}_i\sim p_{\theta_m}}[r(S_i)\pi(S_i|\mathbf{T}_i = p_{\theta_m})]. \tag{19}$$

Afterward, we further reinforce the policy network by optimizing objective function with hierarchical values as follows,

$$L(\theta_w) = -\mathbb{E}_{w_{i,t}\sim\pi_{\theta_w}}[v(w_{i,1:t})], \tag{20}$$

$$L(\theta_m) = -\mathbb{E}_{\mathbf{T}_i\sim p_{\theta_m}}[v(S_{1:i})\pi(S_i|\mathbf{T}_i = p_{\theta_m})]. \tag{21}$$

4 Experiments

4.1 Dataset and Experimental Setting

We conduct experiments on the Stanford image-paragraph dataset released in [Krause *et al.*, 2017], which consists of 19,551 images in total. The dataset was split into training, testing and validation subsets with 14,575 images, 2,489 images and 2,487 images, respectively [Krause *et al.*, 2017]. Each image was annotated with a single paragraph, which contains multiple sentences. Six widely used evaluation metrics BLEU-{1,2,3,4}, METEOR and CIDEr are adopt in our experiments for quantitative evaluation. We utilize the feature extraction technique in [Johnson *et al.*, 2016], which combines VGG-16 and a Region Proposal Network(RPN), to extract 50 regions from an image. Each region is represented by a 4,096 dimensional feature vector. We set the dimension of hidden layers as 512 for all the LSTM cells in our network. We set the hyper-parameters λ_s and λ_w as 5.0 and 1.0, respectively. Following the common setting [Krause *et al.*, 2017], we generate 6 sentences at most with a maximum of 30 words for each to describe a given image. We use SGD solver with an initial learning rate of $1e-4$ to train the network for the first 3 epoch. Then the learning rate stepped decay every 3 training epochs. In addition, we sample $K = 20$ rollout sequences in word-level reward computation. We first pre-train the hierarchical policy network for 50 epochs with the cross-entropy loss in Eq. (17). Then, we optimize the "Manager" and "Worker" alternatively. In particular, we use the sentence-level reward to train "Manager" with "Worker" fixed and use the word-level reward to update "Worker" with "Manager" fixed. Afterwards, we reinforce the policy network using the hierarchical values. The sentence-level and word-level values are used to finetune "Manager" and "Worker," respectively.

4.2 Comparison to the State-of-the-arts

We compare the proposed DHPV approach to the state-of-the-art methods, including 'Image-Flat [Karpathy and Fei-Fei, 2015]', 'Regions-Hierarchical [Krause *et al.*, 2017]', 'RTT-GAN* [Liang *et al.*, 2017]', 'CAPG-VAE [Chatterjee and Schwing, 2018]', 'DAM [Wang *et al.*, 2018]', 'VRD [Che *et al.*, 2018]' and 'TMOS [Mao *et al.*, 2018]'. Table 1 reports the quantitative performance comparison. We can observe that the proposed DHPV achieves the best performance in terms of CIDEr and BLEU-{1,2,3,4}. For example, DHPV improves CIDEr, a specifically designed metric for evaluating captions, by 7.4% over the second best method 'CAPG-VAE [Chatterjee and Schwing, 2018]'. In terms of METEOR, DHPV has a performance degradation as compared to 'RTT-GAN* [Liang *et al.*, 2017]', 'CAPG-VAE [Chatterjee and Schwing, 2018]' and 'TMOS [Mao *et al.*, 2018]'. However, 'RTT-GAN*' uses auxiliary data during training. 'CAPG-VAE' generates several paragraphs for an image and 'TMOS [Mao *et al.*, 2018]' utilizes LDA model to extract topic embedding information. This strategy is beneficial for boosting METEOR. As compared to the other methods [Karpathy and Fei-Fei, 2015] [Krause *et al.*, 2017] [Wang *et al.*, 2018] [Che *et al.*, 2018], which use the benchmark training set without auxiliary data , DHPV obtains the best performance in terms of METEOR.

4.3 Ablation Study

We conduct experiments to evaluate the efficacy of the proposed hierarchical rewards and values for learning paragraph generator as well as the effectiveness of hierarchical policy and value networks. Multiple variants of DHPV are evaluated. Table 2 reports the evaluation results. In particular, the baseline DHPV_C is the hierarchial policy network pre-trained by cross-entropy loss. DHPV_SR denotes the policy network of "Manager" trained by sentence-level reward and "Worker" fixed as pre-trained, while DHPV_WR is that of "Worker" trained by word-level reward and "Manager" fixed as pre-trained. DHPV_HR refers to the policy network trained with hierarchical rewards at both sentence and word levels. Both DHPV_SR and DHPV_WR outperforms DHPV_C. DHPV_HR performs the best compared with them. This indicates that the proposed hierarchical rewards at sentence and word levels are effective in learning paragraph generator. Moreover, DHPV_SV refers to the policy network of "Manager" reinforced by sentence-level value, while DHPV_WV is that of "Worker" reinforced by word-level value. Both DHPV_SV and DHPV_WV obtain better performance compared with DHPV_HR in terms of all the six metrics. This demonstrates that the proposed hierarchical sentence-level and word-level values are able to boost the learning of model. DHPV jointly exploits hierarchical rewards and hierarchial values, augmenting dense supervision cues for model training. Correspondingly, it achieves the best performance.

To better understand the DHPV, we illustrate some qualitative visualization of paragraphs generated by DHPV and its variants in Fig. 4. We can see that the baseline DHPV_C repeats similar sentences in the paragraph and DHPV provides descriptions from different aspects, leading to more informative paragraph.

5 Conclusions

This paper proposed a new Densely Supervised Hierarchical Policy-Value (DHPV) network for generating a descriptive

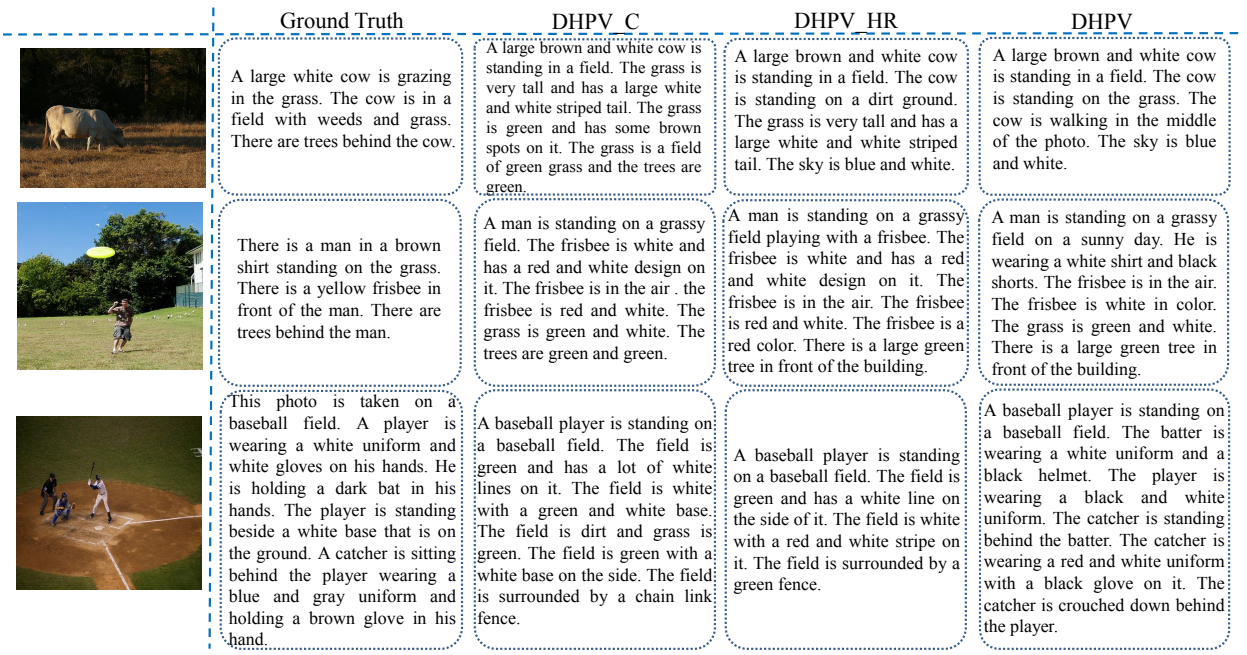

Figure 4: The visual comparisons between Ground Truth, the proposed DHPV and its two baselines, i.e., DHPV_C and DHPV_HR.

Method	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Image-Flat	12.82	11.06	34.04	19.95	12.2	7.71
Regions-Hierarchical	15.95	13.52	41.9	24.11	14.23	8.69
RTT-GAN*	18.39	20.36	42.06	25.35	14.92	9.21
CAPG-VAE	**18.62**	20.93	42.38	25.52	15.15	9.43
DAM	13.90	17.30	35.00	20.20	11.70	6.60
VRD	17.32	14.55	41.74	24.94	14.94	9.34
TMOS	18.6	20.8	43.1	25.8	14.3	8.4
DHPV	17.02	**22.47**	**43.35**	**26.73**	**16.92**	**10.99**

Table 1: Performance comparison to state-of-the-art methods.

Method	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
DHPV_C	16.10	20.70	40.35	24.45	15.41	10.03
DHPV_SR	17.02	21.23	43.23	26.68	16.90	10.94
DHPV_WR	17.05	21.31	43.46	26.66	16.84	10.94
DHPV_HR	17.07	21.92	43.26	26.60	16.86	10.98
DHPV_SV	17.09	22.11	43.33	26.61	16.89	11.00
DHPV_WV	17.03	22.40	43.18	26.65	16.91	11.00
DHPV	17.02	22.47	43.35	26.73	16.92	10.99

Table 2: Comparisons among our methods under different configurations.

paragraph for a given image. The DHPV approach formulates hierarchical rewards and values at both word and sentence levels, which provides dense supervision cues for learning effective paragraph generator. The hierarchical policy network infers descriptive and coherent paragraphs following "token-sentence-paragraph" structure, while hierarchical value network evaluates the values of preceding sentences and words. Moreover, the network is learned by the policy gradient algorithm and thus avoids "exposure bias". Extensive experimental results have shown that the proposed DHPV approach outperforms multiple state-of-the-art methods in terms of various performance metrics.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2017YFB1300201, the National Natural Science Foundation of China (NSFC) under Grants 61622211 and 61620106009 as well as the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

References

[Anderson *et al.*, 2018] 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*, volume 3, page 6, 2018.

[Bengio S, 2015] Jaitly N Shazeer N. Bengio S, Vinyals O. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, pages 1171–1179, 2015.

[Chatterjee and Schwing, 2018] Moitreya Chatterjee and Alexander G Schwing. Diverse and coherent paragraph generation from images. In *ECCV*, pages 729–744, 2018.

[Che *et al.*, 2018] Wenbin Che, Xiaopeng Fan, Ruiqin Xiong, and Debin Zhao. Paragraph generation network with visual relationship detection. In *2018 ACM MM*, pages 1435–1443, 2018.

[Farhadi *et al.*, 2010] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29, 2010.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Johnson *et al.*, 2016] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pages 4565–4574, 2016.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Krause *et al.*, 2017] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, pages 3337–3345. IEEE, 2017.

[Kulkarni *et al.*, 2011] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*, 2011.

[Li *et al.*, 2011] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *ACL*, pages 220–228, 2011.

[Liang *et al.*, 2017] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *ICCV*, pages 3362–3371, 2017.

[Lillicrap *et al.*, 2015] 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. *arXiv preprint arXiv:1509.02971*, 2015.

[Liu *et al.*, 2017] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, pages 873–881, 2017.

[Mao *et al.*, 2014] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[Mao *et al.*, 2018] Yuzhao Mao, Chang Zhou, Xiaojie Wang, and Ruifan Li. Show and tell more: Topic-oriented multi-sentence image captioning. In *IJCAI-18*, pages 4258–4264, July 2018.

[Ren *et al.*, 2017] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *CVPR*, pages 290–298, 2017.

[Rennie *et al.*, 2017] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, volume 1, page 3, 2017.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[Wang *et al.*, 2018] Ziwei Wang, Yadan Luo, Yang Li, Zi Huang, and Hongzhi Yin. Look deeper see richer: Depth-aware image paragraph captioning. In *2018 ACM MM*, pages 672–680, 2018.

[Xu *et al.*, 2015] 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*, pages 2048–2057, 2015.

[You *et al.*, 2016] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016.

[Zha *et al.*, 2019] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[Zhang *et al.*, 2017] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017.