

What and Where the Themes Dominate in Image

Xinyu Xiao,^{1,2} Lingfeng Wang,¹ Shiming Xiang,^{1,2} Chunhong Pan¹

¹National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

²School of Artificial Intelligence, University of Chinese Academy of Sciences

{xinyu.xiao, lfwang, smxiang, chpan}@nlpr.ia.ac.cn

Abstract

The image captioning is to describe an image with natural language as human, which has benefited from the advances in deep neural network and achieved substantial progress in performance. However, the perspective of human description to scene has not been fully considered in this task recently. Actually, the human description to scene is tightly related to the endogenous knowledge and the exogenous salient objects simultaneously, which implies that the content in the description is confined to the known salient objects. Inspired by this observation, this paper proposes a novel framework, which explicitly applies the known salient objects in image captioning. Under this framework, the known salient objects are served as the themes to guide the description generation. According to the property of the known salient object, a theme is composed of two components: its endogenous concept (what) and the exogenous spatial attention feature (where). Specifically, the prediction of each word is dominated by the concept and spatial attention feature of the corresponding theme in the process of caption prediction. Moreover, we introduce a novel learning method of Distinctive Learning (DL) to get more specificity of generated captions like human descriptions. It formulates two constraints in the theme learning process to encourage distinctiveness between different images. Particularly, reinforcement learning is introduced into the framework to address the exposure bias problem between the training and the testing modes. Extensive experiments on the COCO and Flickr30K datasets achieve superior results when compared with the state-of-the-art methods.

Introduction

The technique of image description is to simulate human description of a scene content. Benefited from the development of the deep neural network, the performance of image captioning has achieved considerable progress in recent years (Vinyals et al. 2015; Xu et al. 2015; Rennie et al. 2017). Current structure of image captioning methods are typically developed by taking the Convolutional Neural Network (CNN) plus Recurrent Neural Network (RNN) as the encoding-decoding framework. Based on the framework, various variants have been proposed to achieve the goal of description with natural language (Lu et al. 2016; Gan et al. 2016).

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The attention mechanism has been widely used in many image captioning models (Xu et al. 2015; Lu et al. 2016; Gu et al. 2018). There are two broad categories of visual attention mechanisms: exogenous objects mainly based on the visual stimulus and endogenous knowledge determined by cognitive phenomena. The normal attention operation is depended on the exogenous objects which focuses on specific parts of the visual input to compute the adequate responses. Differently, according to the research in (Borji, Sihite, and Itti 2013), human visual attention is what you see as well as what you known which means that it is decided by the endogenous knowledge and the exogenous salient objects simultaneously. It indicates that, when human describes a scene with a sentence, the known salient objects (including instances, actions, scenes, etc.) are captured and served as the themes to guide the description generation. Since the unknown objects might be difficult to interpret in human brain, the attended objects would be always known about in advance to human. Rather than human attention in scene description, the recently implemented attention mechanism in image captioning only focus on the exogenous salient objects that come to the forefront dynamically. But what semantics the attended objects refer to can not be confirmed. Some models (Pedersoli et al. 2016; Li et al. 2017) utilize the development of object detection (Ren et al. 2015), and apply the features of the detected regions to image captioning. These regions are known about in advance, which is approaching to human behavior. But the annotated of the detected regions are limited and cannot meet the requirements of all the image captioning.

To human being, the distinctive description to a scene is specially mentioned when distinguished from others. The distinctiveness between every two different images is containing. If the image captioning model could capture it, the diversity of the overall generated captions can improve. Depending on the distinctiveness contained in image, the themes in image have distinctiveness as well. Therefore, improving the distinctiveness of the theme learning can affect the quality of the generated description.

Reinforcement learning (RL) has become a standard method for training agents to interact with an environment, even in fields like sequence learning (Rennie et al. 2017; Liu et al. 2017). In convention, because of the application of Maximum Likelihood Estimation (MLE), the log-likelihood

score of the prediction in training does not correlate well with the standard evaluation metrics when testing. But RL agents can optimize their action policies of predicting word to maximize the expected reward, which is received from the environment. Moreover, the sentence level evaluation metrics can be used as the reward in training with the RL method of image captioning. In this way, the exposure bias between the training and the testing sets can be eliminated.

In the perspective of human description, we propose a novel framework to explicitly apply the known salient objects as the themes to guide the caption generation. A theme is composed of two components:

- **What:** the concept which defines the endogenous knowledge of theme.
- **Where:** the spatial attention feature which indicates the exogenous salient object of theme.

Specifically, our proposed method extracts the concept and spatial feature of the corresponding theme to dominate the prediction of each word in the process of caption generation. A Distinctive Learning (DL) is introduced to image captioning to achieve the specificity like human description. Moreover, we design a improved RL method from (Rennie et al. 2017) to train our framework, where the back propagated rewards are added with balance weight. The weight is depended on the difficulty level of the training samples.

The major contributions in this paper are as follows:

- We design a new framework which contains two modules. One is attending to the concept of theme and the other is attending to the correlative dominating region in image. Moreover, the channel of the visual feature map is defined by the concepts. Each channel is responding to the corresponding concept of the image themes.
- A novel Distinctive Learning (DL) is introduced to utilize the distinctiveness of themes between different images. Technically, denoted the true image-theme pairs as (\mathbf{I}, \mathbf{t}) , where \mathbf{t} is the attended concept of theme. Meanwhile, the negative image-theme pairs are denoted as $(\neg\mathbf{I}, \mathbf{t})$ or $(\mathbf{I}, \neg\mathbf{t})$, where $\neg\mathbf{I}$ and $\neg\mathbf{t}$ are the mismatched objects of the concept of theme \mathbf{t} and image \mathbf{I} , respectively. The goals of the distinctive learning is to increase the correlation probability of true pairs and decrease the correlation probability of negative pairs.
- The introduction of the improved RL by the balance weight in training can effectively address the exposure bias problem between the training and the testing sets. The extensive experiments on the COCO and Flickr30K datasets demonstrate the effectiveness of our model.

Related Work

Recent public methods based on neural networks in image captioning achieve great success. Here, we review the most relevant works of image captioning to our work.

Neural network based methods are inspired by the success of RNN to sequence-to-sequence learning, and have already played a huge role in image captioning. Attention mechanism is used to determine the spatial map highlighting

regions in image over time according to the textual context. Xu et al. (Xu et al. 2015) proposed the “soft” and “hard” variants of attention mechanism and combined visual attentions with the hidden states of LSTM when generating corresponding words. Lu et al. (Lu et al. 2016) proposed an adaptive attention model which can decide the salient features to dynamically input to the language model to generate the next word.

Semantic-based approaches use the high-level image semantic information in the caption generation model to fill the gaps ahead between vision and language. In Wu et al. (Wu et al. 2016), a new high-level concepts extracted method to clear improve the performance of image captioning model with the injection of the high-level concepts to the state-of-the-art LSTM-based model. Gan et al. (Gan et al. 2016) attempted to effectively compose the semantic concepts in the process of image caption generation, and proposed a Semantic Compositional Network (SCN) to combine the semantic concepts to the decoder LSTM in the decoding process.

Learning methods for image captioning have attracted more and more attentions recently. Different from a structural model, The method of learning to model is to define a set of rules to training the constructed model. To many classical models, the Maximum Likelihood Estimation (MLE) is always applied as the learning method to maximize the conditional log-likelihood of the training samples. But the defects of MLE including high resemblance and overfitting are limiting the effect of image captioning models. Yang et al. (Yang et al. 2016) noticed the distinctive supervision is beneficial for captioning, except introducing a review network to global modeling the attended inputs, and output a global properties captured thought vector, imported a distinctive supervision to improve the performance of description generation. Dai et al. (Dai and Lin 2017) argued the distinctiveness is significant in natural language descriptions and proposed a contrastive learning method, which explicitly encourages distinctiveness to maintain the quality of the overall generated captions.

Reinforcement learning (RL) has been taken seriously since the work (Silver et al. 2017) was proposed, and due to (Ranzato et al. 2015), the methods of RL have been introduced into image captioning. For RL, the prediction of word can be seen as an “action” when interacting with the “environment” by the “agent” LSTM. Ranzato et al. (Ranzato et al. 2015) first proposed a RL based approach to calculate the sentence-level reward by the Monte-Carlo technique and reverse propagation the policy gradient in training. A self-critical sequence learning method was proposed by Rennie et al. (Rennie et al. 2017) to use a policy network with a normalized reward of metric obtained by the model against the baseline under the test-time inference algorithm. Chen et al. (Chen et al. 2018), instead of using the Monte-Carlo technique, applied a temporal difference method to model the RL value function to estimate the value of actions at each time step.

Unlike the existing researches, we base on the human description to scene and combine the advanced learning methods, to propose a framework which predetermines the concept and location of theme in image to guide the word gen-

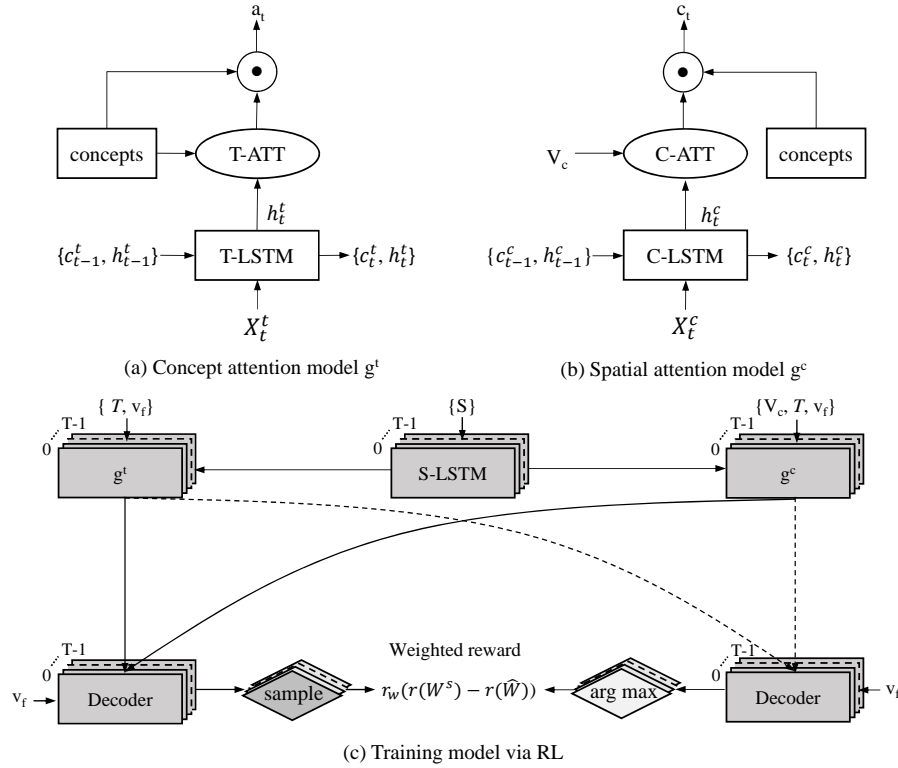


Figure 1: Model architecture of our method. (a) is an illustration of the concept attention process g^t of theme, (b) is presenting the model of our proposed spatial attention model g^c of theme. (c) illustrates the whole proposed image captioning model, S-LSTM and Decoder are the Sentence-LSTM and decoder LSTM, respectively. The rewards are balanced by the weight r_w .

eration at each time step.

Our Approach

The normal operator of the image captioning is named an encoder-decoder, which encodes an image and decodes it to a sentence. Through the constructing of network and designing of learning method, we expect to reset the connotation of conventional structure to simulate the perspective of human description. In the following sections, we will first present the outputs of the CNN before introducing the main framework in detail.

CNN feature extractor. We take a widely-used CNN architecture as the CNN feature extractor of images. The output of the last fully connected layer is treated as the context vector v_f , and the last convolutional output is indicated to the context CNN feature map, which is denoted by V_c .

What Dominates the Captioning

In general, a caption word is mainly associated with a theme, and the theme may lead to the formation of several related words in a sentence. Therefore, considering what dominates the process of captioning at each step of word generation may lead to sub-optimal results.

Before theme is confirmed, we first build a theme vocabulary. A group of words which contain rich semantic cues are chosen to construct the vocabulary of theme. The vocabulary contains most of the visual concepts of an image set, includ-

ing various parts of speech like nouns, verbs, adjectives and so on. To reduce the sensitivity of vocabulary, we merge the words with similar semantic. For instance, “pictures”, “photos” and “photographs” which have the same semantic are classified to the same category. This operation decreases the size of the theme vocabulary and enriches its connotation.

Given this N_t length theme vocabulary, we introduce a concept attention mechanism to attend what concept of the theme dominates the final caption generation at each time step. Before deciding on the theme, we wish to predict the whole visual concepts of a given image I . Motivated by (Fang et al. 2015), we follow the operation of Fang et al. to adopt a weakly supervised approach of Multiple Instance Learning (MIL) (Viola, Platt, and Zhang 2005) to train a detector of image concepts. Finally, the obtained vector T can represent the probability distribution over the set of concepts for the image.

The concept attention model $g^t(T, h_t^t)$ which can be seen in Fig. 1 (a), is proposed to compute the concept of the theme a_t . $T = [t_1, t_2, \dots, t_{N_a}]$ is the global concept vector, h_t^t is the hidden state of LSTM at time t . The updating procedure of h_t^t is defined as follows:

$$\begin{aligned} h_t^s &= \text{LSTM}^s(W_e s_t, h_{t-1}^s), \\ h_t^t &= \text{LSTM}^t(T, v_f, h_t^s, h_{t-1}^t), \end{aligned} \quad (1)$$

where s_t denotes the one-hot vector of the t -th word in the caption, $W_e \in \mathbb{R}^{N_s \times V}$ is the word embedding matrix and

V is the vocabulary size; $LSTM^s$ means Sentence-LSTM unit which is used to encode the sentence inputs; $LSTM^t$ is to extract the discriminative information of the theme concept for current word. Formally, for the t time step, the concept attention model g^t can be defined as follows:

$$\begin{aligned} \mathbf{z}_t^t &= \tanh((\mathbf{W}^t \mathbf{T} + \mathbf{b}^t) \oplus \mathbf{W}^{ht} \mathbf{h}_t^t), \\ \boldsymbol{\alpha}_t^t &= \text{softmax}(\mathbf{W}^{\alpha t} \mathbf{z}_t^t + \mathbf{b}^{\alpha t}), \\ \mathbf{a}_t &= f(\mathbf{T}, \boldsymbol{\alpha}_t^t), \end{aligned} \quad (2)$$

where $\mathbf{W}^t \in \mathbb{R}^{k \times N_t}$, $\mathbf{W}^{ht} \in \mathbb{R}^{k \times N_s}$, $\mathbf{W}^{\alpha t} \in \mathbb{R}^{k \times k}$ are the transformation matrices that map the concept vector and hidden state to a same dimension; $\mathbf{b}^t \in \mathbb{R}^k$ and $\mathbf{b}^{\alpha t} \in \mathbb{R}^k$ are the model biases; we denote \oplus as a sum module; $f(\cdot)$ is an element-wise multiplication to the global concept vector and its corresponding position attention weights and the dimensions of $\boldsymbol{\alpha}^t$ is the same with \mathbf{T} , i.e., $k = N_t$.

Where Dominates the Captioning

According to the work in (Xu et al. 2015), the region attention mechanism attending to the irrelevant regions of the global image may lead to the sub-optimal results in the caption generation. But most region attention models are not taking the pattern of different channels into consideration. Inspired by the research in (Chen et al. 2017), the filters of different channels can be seen as the semantic filters, and each channel is a response activation of the corresponding semantic. Moreover, we associate the channel of the global image feature map with the concept vector. Therefore, semantic information of the concept vector can be attached to the feature map in channel.

The context CNN feature \mathbf{V}_c is reshaped to $\mathbf{V}_c = [v_{c1}, v_{c2}, \dots, v_{cm}]$ by flattening the width and height of the original \mathbf{V}_c , where $v_{ci} \in \mathbb{R}^{N_t}$ equals to the \mathbf{T} in dimension. To confirm the theme dominates the captioning at t time step, as in Fig. 1 (b), the corresponding spatial attention process $g^c(\mathbf{V}_c, \mathbf{T}, \mathbf{h}_t^c)$ is introduced. \mathbf{h}_t^c is generated by $LSTM^c$ to encode the distinctive information of the irrelevant regions. Below is the updating procedure of \mathbf{h}_t^c :

$$\mathbf{h}_t^c = LSTM^c(\mathbf{T}, \mathbf{v}_f, \mathbf{h}_t^s, \mathbf{h}_{t-1}^c). \quad (3)$$

Formally, for the t time step, we define the attention part of the model g^c as follows:

$$\begin{aligned} \mathbf{z}_t^c &= \tanh((\mathbf{W}^c \mathbf{V}_c + \mathbf{b}^c) \oplus \mathbf{W}^{hc} \mathbf{h}_t^c), \\ \boldsymbol{\alpha}_t^c &= \text{softmax}(\mathbf{W}^{\alpha c} \mathbf{z}_t^c + \mathbf{b}^{\alpha c}), \end{aligned} \quad (4)$$

where $\mathbf{W}^c \in \mathbb{R}^{l \times N_t}$, $\mathbf{W}^{hc} \in \mathbb{R}^{l \times N_s}$, $\mathbf{W}^{\alpha c} \in \mathbb{R}^l$ are the transformation matrices that map the CNN feature and hidden state to a same dimension; $\mathbf{b}^c \in \mathbb{R}^l$ and $\mathbf{b}^{\alpha c} \in \mathbb{R}^1$ are the model biases. Based on the attention distribution, the explicit representation \mathbf{a}_t^c can be extracted through:

$$\mathbf{a}_t^c = \sum_{i=1}^m \boldsymbol{\alpha}_{ti}^c \mathbf{v}_{ci}, \quad (5)$$

where the context vector \mathbf{a}_t^c can be filtered by the global concept vector to select the semantic information. The element-wise multiplying function is $\mathbf{c}_t = f^c(\mathbf{T}, \mathbf{a}_t^c)$.

Distinctive Learning

The sentences produced by many recent methods lack in variability in general. The problem is perhaps due to the rules which object is to maximize only the probabilities of the given captions. The rules may lead to high resemblance of the generations. In essence, the models neglect some subtle but significant aspects in training. And these aspects are always distinctive. From (Fang et al. 2015), the distinctive supervision has shown to be useful and common. According to Jas et al. (Jas and Parikh 2015), the specificity is universal in human descriptions, which implies that the distinctive aspects should be reflected in the scene descriptions.

In this paper, the distinctive learning is introduced into our model in a novel way. The themes between two samples are always different. It means that the learned themes contain distinctiveness of the image. To learn and reinforce this distinctiveness, two sets of data are required: (1) the positive set R , which is the right theme pairs to ground-truth image $((\mathbf{a}_1^t, \mathbf{T}), (\mathbf{a}_2^t, \mathbf{T}), \dots, (\mathbf{a}_{T_r}^t, \mathbf{T}))$; (2) the negative set F , which is the noisy theme pairs $((\mathbf{a}_1^{t-}, \mathbf{T}^-), (\mathbf{a}_2^{t-}, \mathbf{T}^-), \dots, (\mathbf{a}_{T_f}^{t-}, \mathbf{T}^-))$. And the pairs are generated by randomly sampling from $\mathbf{T}^- \in \mathbf{T}^c$, where \mathbf{T}^c is sampled from the complementary set of the concept vector \mathbf{T} . The length of positive or negative pairs is up to the length of the sample caption.

As we known, each \mathbf{a}_t^t is extracted from the \mathbf{T} . For the pair $(\mathbf{a}_t^t, \mathbf{T})$, we wish that the similarity between \mathbf{a}_t^t and \mathbf{T} is greater than any other pairs like $(\mathbf{a}_t^{t-}, \mathbf{T})$, $(\mathbf{a}_t^t, \mathbf{T}^-)$ and so on. Following this intuition, the distinctive learning is to differentiate the positive set with the negative set, which is defined as follows:

$$D(R, F) = \beta - (p(\text{match}) - p(\text{mismatch})), \quad (6)$$

where β is the margin cross-validated, and the Distinctive Learning Loss function can be set as:

$$\begin{aligned} L_d &= \frac{1}{N} \sum_{n=1}^N \sum_{t=1}^{\min(T_r, T_f)} \max(0, \beta - (\sum_{l=1}^{N_t} \mathbf{a}_t^t \mathbf{T} - \sum_{l=1}^{N_t} \mathbf{a}_t^{t-} \mathbf{T})) \\ &+ \frac{1}{N} \sum_{n=1}^N \sum_{t=1}^{\min(T_r, T_f)} \max(0, \beta - (\sum_{l=1}^{N_t} \mathbf{a}_t^t \mathbf{T} - \sum_{l=1}^{N_t} \mathbf{a}_t^t \mathbf{T}^-)), \end{aligned} \quad (7)$$

where N is the number of the training samples, which is used as a normalizer. Except for the positive pair, the negative set is matched with the positive set in two types $(\mathbf{a}_t^{t-}, \mathbf{T})$ and $(\mathbf{a}_t^t, \mathbf{T}^-)$.

Model Training via RL

The Decoder is applied the decoder LSTM to decode the encoding information mentioned above to generate the target caption. In convention, the learning method of the language model is applied the MLE. The learning objective is to learn parameters through minimizing the negative log-likelihood of the target sentence of the ground-truth. But because of the limitations of the MLE in the gradient vanishing and overfitting, the discrepancy exists between training and testing.

Table 1: Performance comparisons on COCO with variable variants of our model under the optimization of MLE and RL.

Model	MLE				RL			
	B-3	B-4	Meteor	CIDEr	B-3	B-4	Meteor	CIDEr
LSTM-ATT	44.4	33.9	26.5	106.5	46.8	34.8	26.9	113.3
LSTM+TATT	44.4	34.1	26.5	106.7	47.5	35.4	27.0	114.5
LSTM+RATT	45.2	34.7	26.9	108.6	47.6	35.5	27.2	115.5
WWT	45.5	34.9	27.0	109.0	48.3	36.1	27.3	116.4
WWT+D	45.9	35.4	27.2	110.9	48.2	36.1	27.4	117.0
WWT+DW	46.2	35.7	27.3	111.4	48.4	36.3	27.4	117.5

Table 2: Performance comparisons with state-of-the-art methods on Flickr30K and COCO datasets.

Model	Flickr30K				COCO				
	B-3	B-4	Meteor	CIDEr	B-3	B-4	Meteor	Rouge-L	CIDEr
SCN (Gan et al. 2016)	40.3	28.8	22.3	—	44.4	34.1	26.1	—	104.1
Adaptive (Lu et al. 2016)	35.4	25.1	20.4	53.1	43.9	33.2	26.6	—	108.5
SCST (Rennie et al. 2017)	—	—	—	—	—	33.3	26.3	55.3	111.4
CL (Dai and Lin 2017)	—	—	—	—	43.7	33.4	26.2	55.9	105.9
TD (Chen et al. 2018)	—	—	—	—	45.6	34.0	26.3	55.5	111.6
Stack-Cap (Gu et al. 2018)	—	—	—	—	47.9	36.1	27.4	56.9	120.4
Up-Down (Anderson et al. 2018)	—	—	—	—	—	36.3	27.7	56.9	120.1
WWT(MLE)	40.8	29.7	22.3	73.5	46.5	35.9	27.3	56.2	112.1
WWT(RL)	42.5	32.2	22.8	88.2	49.4	37.4	27.7	57.5	118.8

To solve this problem, the reinforcement learning which applies the evaluation metrics as the optimizing objects is taken into consideration. In (Sutton and Barto 1998), the reinforcement learning is interacting with the “environment” (here contains the image features, theme concepts, hidden states, and previous words) by an “agent” (e.g. LSTM). The “agent” takes an action according to the policy p_θ of the parameters θ of the network to select an “action”, which is to predict the next word in our case. After each action, the agent updates its internal state (cells and hidden states of the LSTM). After generating a complete sentence, the agent observes a sentence-level reward. The reward can be any of the evaluation metrics to the “agent”. As the Fig. 1 (c) illustrating, following the implementation in (Rennie et al. 2017), the objective in learning is to minimize the negative expected rewards of the complete sampled caption $\mathcal{W}^s = \{w_1^s, \dots, w_T^s\}$:

$$L_\theta = -\frac{\lambda_\theta}{N} \sum_{n=1}^N E_{\mathcal{W}^s \sim p_\theta} [r(\mathcal{W}^s)], \quad (8)$$

where λ_θ is used to balance the weight of the loss functions; the \mathcal{W}^s is calculated by comparing sampled caption with the reference caption in the specified evaluation metric. We calculate the expected gradient by applying a single Monte-Carlo sample:

$$\nabla_\theta L_\theta \approx -\frac{\lambda_\theta}{N} \sum_{n=1}^N \Delta r(\mathcal{W}^s) \nabla_\theta \log[p_\theta(\mathcal{W}^s)], \quad (9)$$

where $\Delta r(\mathcal{W}^s)$ is the relative reward, which is computed by relating to a baseline reward $r(\hat{\mathcal{W}})$, $r(\hat{\mathcal{W}})$ is obtained by

performing greedy decoding:

$$r(\hat{\mathcal{W}}) = \arg \max p(w_t | h_t^d),$$

$$\nabla_\theta L_\theta \approx -\frac{\lambda_\theta}{N} \sum_{n=1}^N (r(\mathcal{W}^s) - r(\hat{\mathcal{W}})) \nabla_\theta \log[p_\theta(\mathcal{W}^s)]. \quad (10)$$

If the dominating theme is the appropriate choice, the sum of products between \mathbf{a}_t^t and \mathbf{T} could be plus one. And the inappropriate choice indicates the corresponding sample is a hard learning sample. Inspired by the study in (Lin et al. 2018), to improve the learning efficiency, we add a dynamic reward weight r_w into the learning process:

$$r_w = \frac{\|\sum_{t=1}^T \mathbf{a}_t^t\| \|\mathbf{T}\|}{(\sum_{t=1}^T \mathbf{a}_t^t) \cdot \mathbf{T}},$$

$$\nabla_\theta L_\theta \approx -\frac{\lambda_\theta}{N} \sum_{n=1}^N r_w (r(\mathcal{W}^s) - r(\hat{\mathcal{W}})) \nabla_\theta \log[p_\theta(\mathcal{W}^s)]. \quad (11)$$

The above formulas indicates that the sample needs a large weight to training when the r_w is high.

Experiments

Implementation Details

We evaluate our method on two widely used datasets. The first one is Flickr30K (Young et al. 2014), which contains 31,783 images. In the dataset, each image is paired with five sentences. Following the publicly splits¹, we divide 29,014, 1,000 and 1,000 images for training, validation and testing, respectively. The other more challenging dataset COCO (Lin et al. 2014), which is consisting of 123,287 images, where

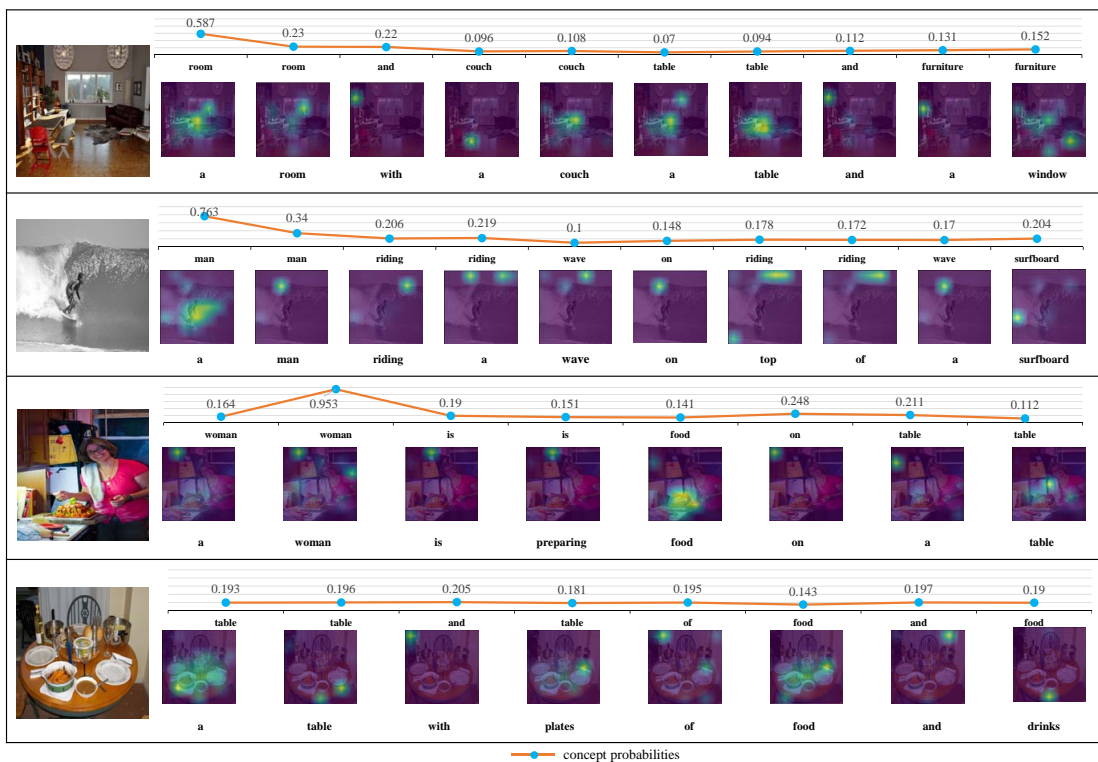


Figure 2: Visualization of the generated captions, the probabilities of the concept and corresponding spatial attention map of each dominating theme.

Table 3: Leaderboard on the online COCO testing server of the published state-of-the-art image captioning models.

Model	B-1		B-2		B-3		B-4		Meteor		Rouge-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCN (Gan et al. 2016)	74.0	91.7	57.5	83.9	43.6	73.9	33.1	63.1	25.7	34.8	54.3	69.6	100.3	101.3
Adaptive (Lu et al. 2016)	74.6	91.8	58.2	84.2	44.3	74.0	33.5	63.3	26.4	35.9	55.0	70.6	103.7	105.1
CL (Dai and Lin 2017)	74.2	91.0	57.7	83.1	43.6	72.8	32.6	61.7	26.0	35.0	54.4	69.5	101.0	102.9
SCST (Rennie et al. 2017)	78.1	93.1	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.8	116.7
TD (Chen et al. 2018)	75.7	91.3	59.1	83.6	44.1	72.6	32.4	60.9	25.9	34.2	54.7	68.9	105.9	109.0
Stack-Cap (Gu et al. 2018)	77.8	93.2	61.6	86.1	46.8	76.0	34.9	64.6	27.0	35.6	56.2	70.6	114.8	118.3
WWT(RL)	80.2	94.3	63.2	87.5	48.1	77.7	35.9	66.5	27.2	36.1	56.9	71.9	113.2	115.9

each image has at least five reference captions. According to the setting of¹: 5,000 images are for testing and 5,000 images for offline testing, which are all splitted from the validation, the rest images are used for training. In addition, we further test 40,775 images of the official COCO test set for online testing to against the state-of-the-art methods.

The yielding sizes of vocabulary are 9,487 and 7,000 for COCO and Flickr30k, respectively. In evaluation, we report the following metrics: B-N (N=1,2,3,4) (Papineni et al. 2002), Meteor (Banerjee and Lavie 2005), Rouge-L (Lin 2004), CIDEr (Vedantam, Zitnick, and Parikh 2015).

Parameter Settings and Training We adopt the ResNet-152 model (He et al. 2016), which is pretrained on the ImageNet dataset (Deng et al. 2009), as the CNN model to extract the visual feature of image. The dimensions of the

visual feature channel and all the LSTM hidden states are set to the same length as the concept vector. And the length of the theme vector is determined by the complexity of the dataset, which is set to 1,000, 200 for COCO and Flickr30k, respectively. The proposed RL-based method is applied to optimize the just MLE trained model with the CIDEr metric, and λ_θ is set to 2. At each epoch, the validation set is used to evaluate the model, and the model with the best CIDEr score is selected for testing. All the experiments are implemented with Pytorch (Paszke et al. 2017).

Variant Models for Comparison To gain insight into the effectiveness of our proposed approach, the variable variants of our model are described as follows: **LSTM-ATT**: We implement the 3 layers LSTM-based model which is not adding any attention operations. **LSTM+TATT**: We implement the 3 layers LSTM-based model which is adding the concept attention, “TATT” means the concept attention

¹<http://cs.stanford.edu/people/karpathy/deeimagesent/>









			
<p>LSTM-ATT: A kitchen with a refrigerator and a refrigerator.</p> <p>WWT: A kitchen with a refrigerator and a refrigerator.</p> <p>WWT+DW: A white refrigerator freezer sitting inside of a kitchen.</p>	<p>LSTM-ATT: A herd of sheep grazing on a beach.</p> <p>WWT: A herd of sheep grazing on a lush green field.</p> <p>WWT+DW: A herd of cattle standing on top of a grass covered field.</p>	<p>LSTM-ATT: A display of various types of different colors.</p> <p>WWT: A display case with a variety of pastries.</p> <p>WWT+DW: A display case filled with different types of pastries.</p>	<p>LSTM-ATT: A man wearing a tie and tie holding a camera.</p> <p>WWT: A man with glasses is holding a blue and white tie.</p> <p>WWT+DW: A man wearing glasses and a tie in front of a brick wall.</p>
			
<p>LSTM-ATT: A large jet sitting on top of an airport runway.</p> <p>WWT: A large airplane sitting on top of a runway.</p> <p>WWT+DW: An airplane is parked on the runway with its door open.</p>	<p>LSTM-ATT: Two children are playing tennis in a park.</p> <p>WWT: Two young boys are playing with a tennis racket.</p> <p>WWT+DW: Two young boys holding tennis rackets in their hands.</p>	<p>LSTM-ATT: A plate of food with a cup of coffee.</p> <p>WWT: A white plate with a sandwich on it.</p> <p>WWT+DW: A white plate topped with a pastry next to a cup of coffee.</p>	<p>LSTM-ATT: A traffic light hanging from a street light.</p> <p>WWT: A city street with traffic lights and a building.</p> <p>WWT+DW: A traffic light hanging from the side of a building.</p>

Figure 3: This figure illustrates several images with captions generated by different variants of our model. Compared with LSTM-ATT and WWT, WWT+DW generates more distinctive captions in these samples.

mechanism. **LSTM+RATT:** The region attention is applied into the LSTM-ATT. “RATT” denotes the region attention mechanism. **WWT:** The full structure of our model does not involve the Distinctive Learning. **WWT+D:** We add the distinctive Learning learning in the training process for our full model, “D” represents the learning module. **WWT+DW:** To balance the training of the hard samples with the easy samples, we add the “W”, which is a dynamic reward weight, to the loss function of caption generation.

Evaluation and Analysis

In Table 1, it can be seen that the WWT+DW achieves the best performances in all metrics, which indicates the introduced dynamic reward and the distinctive Learning learning of our model can significantly improve the performance of image captioning. Compared to LSTM-ATT, LSTM+TATT and LSTM+RATT can get the better performance. Moreover, when combining the TATT and the RATT mechanisms to WWT, the model can significantly improve the performance of the results. It demonstrates that the simulation of human description to scene in our method is effective. Comparing with the results of MLE-based and RL-based methods, the RL method can improve the performance of MLE-based model by significant margins across all metrics. Specifically, because the RL-based models is training with CIDEr metric, the improvements on CIDEr are over 3% to all variants compare to the MLE based models.

From Table 2, on Flickr30K, our method achieves su-

perior performance. Furthermore, on the CIDEr metric, our method achieves 88.2%, which is the highest known value and improving over 66% against the performance of (Lu et al. 2016). On COCO, whether using the MLE or RL, compared with the state-of-the-art, our method outperforms them by obvious margins. Comparing to the MLE-based 3 ensembled methods, our MLE-based results present high competitiveness. Moreover, our RL-based 3 ensembled model obtains significant gains across all metrics. To further evaluate our model, we upload the results of our RL ensembled model to the online COCO test server. The results in Table 3 show that the present results of our method achieve the highest performance with other state-of-the-art methods.

We further visualize how the themes dominate the caption generation. In Fig. 2, the concept probability and corresponding spatial attention map of the dominating theme for each word are visualized. From these examples, we can see the generated words are closely linked to their dominating themes. For example, in the first image, there are four main themes “room”, “couch”, “table” and “furniture” dominate the caption generation. The result “a living room with a couch a table and a window” and corresponding spatial attentions perfectly validate the dominating process of themes. It shows that the captured theme has the ability to predict the relevant words in caption as human beings.

Some qualitative results are shown in Fig. 3. These captions are generated by three variants LSTM-ATT, WWT and WWT+DW, respectively. From these examples, we know

that the proposed structure of the framework and the designed learning method are significant to obtain better results. Depending on the way of human description to scene, the model can capture more significant information. For example, in the fourth image, the WWT+DW captures rich content in the image contains “man”, “glasses”, “tie” and “wall”, the WWT misses the theme of “wall”. But the performer of LSTM-ATT is the worst, which has a grammar mistake and captures wrong information. Depending on the distinctive Learning method, more distinctive descriptions can be generated. From the third and sixth examples, compared to the other two variants, the sentences of the WWT+DW are more distinctive and appropriate.

Conclusion

In this paper, we proposed a framework to explicitly apply the known salient objects as the themes in image captioning. The concept (what) and spatial attention feature (where) of the corresponding theme are extracted to dominate the word prediction at each time step. Moreover, while maintaining the quality of the generated sentences, a novel learning method is introduced, Distinctive Learning, to encourage distinctiveness between captions. Under the improved reinforcement learning method, our model achieves comparable performance with the state-of-the-art approach on two challenging datasets, namely COCO and Flickr30K.

The vocabulary of the theme is predefined in this framework, which means that the number of the visual concepts is fixed and the semantic cues are limited. The limitation of the endogenous knowledge could influence the model’s description of the exogenous objects in image. Therefore, we tend to learn a knowledge base of the theme in future, which contains more visual concepts and richer semantic cues. The learned endogenous knowledge base can meet the needs of the description to the more diverse exogenous objects.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 91646207, 61773377, and 61573352, and the Beijing Natural Science Foundation under Grants 4162064 and L172053.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, 6.

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, volume 29, 65–72.

Borji, A.; Sihite, D. N.; and Itti, L. 2013. What stands out in a scene? a study of human explicit saliency judgment. *Vision research* 91:62–77.

Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T. 2017. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 6298–6306.

Chen, H.; Ding, G.; Zhao, S.; and Han, J. 2018. Temporal-difference learning with sampling baseline for image captioning. In *AAAI*.

Dai, B., and Lin, D. 2017. Contrastive learning for image captioning. In *NIPS*, 898–907.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.; Zitnick, C.; and Zweig, G. 2015. From captions to visual concepts and back. In *CVPR*, 1473–1482.

Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; and Deng, L. 2016. Semantic compositional networks for visual captioning. *CoRR* abs/1611.08002.

Gu, J.; Cai, J.; Wang, G.; and Chen, T. 2018. Stack-captioning: Coarse-to-fine learning for image captioning. In *AAAI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Jas, M., and Parikh, D. 2015. Image specificity. In *CVPR*, 2727–2736.

Li, L.; Tang, S.; Deng, L.; Zhang, Y.; and Tian, Q. 2017. Image caption with global-local attention. In *AAAI*, 4133–4139.

Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. 2014. Microsoft COCO: common objects in context. In *ECCV*, 740–755.

Lin, L.; Wang, K.; Meng, D.; Zuo, W.; and Zhang, L. 2018. Active self-paced learning for cost-effective and progressive face identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(1):7–19.

Lin, C. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*, volume 8.

Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved image captioning via policy gradient optimization of spider. In *ICCV*, 873–881.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2016. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *CoRR* abs/1612.01887.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *NIPS Workshop*.

Pedersoli, M.; Lucas, T.; Schmid, C.; and Verbeek, J. 2016. Areas of attention for image captioning. *CoRR* abs/1612.01033.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2015. Sequence level training with recurrent neural networks. *CoRR* abs/1511.06732.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, 91–99.

Rennie, S.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*, 1179–1195.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; and Hassabis, D. 2017. Mastering the game of go without human knowledge. *Nature* 550.

Sutton, R., and Barto, A. 1998. Reinforcement learning: An introduction. *IEEE Transactions Neural Networks* 9(5):1054–1054.

Vedantam, R.; Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*, 3156–3164.

Viola, P.; Platt, J.; and Zhang, C. 2005. Multiple instance boosting for object detection. In *NIPS*, 1417–1424.

Wu, Q.; Shen, C.; Liu, L.; Dick, A.; and Hengel, A. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*, 203–212.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W.; and Salakhutdinov, R. 2016. Review networks for caption generation. In *NIPS*, 2361–2369.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* 2:67–78.