applied sciences

Article

Structure Preserving Convolutional Attention for Image Captioning

Shichen Lu ^{1,2,†,‡}, Ruimin Hu ^{1,2,*,‡}, Jing Liu ^{3}, Longteng Guo ^{3} and Fei Zheng ^{4}

[1] National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan 430072, China

[2] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China

[3] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[4] China General Technology Research Institute, Beijing 100190, China

* Correspondence: hrm@whu.edu.cn

† Current address: Information Department, Wuhan University, Dormitory 8, Room 617, Hongshan District, Wuhan 430072, China.

‡ These authors contributed equally to this work.

Received: 11 June 2019; Accepted: 16 July 2019; Published: 19 July 2019

Abstract: In the task of image captioning, learning the attentive image regions is necessary to adaptively and precisely focus on the object semantics relevant to each decoded word. In this paper, we propose a convolutional attention module that can preserve the spatial structure of the image by performing the convolution operation directly on the 2D feature maps. The proposed attention mechanism contains two components: convolutional spatial attention and cross-channel attention, aiming to determine the intended regions to describe the image along the spatial and channel dimensions, respectively. Both of the two attentions are calculated at each decoding step. In order to preserve the spatial structure, instead of operating on the vector representation of each image grid, the two attention components are both computed directly on the entire feature maps with convolution operations. Experiments on two large-scale datasets (MSCOCO and Flickr30K) demonstrate the outstanding performance of our proposed method.

Keywords: image captioning; attention; spatial structure; deep learning; computer vision

1. Introduction

Image captioning is to automatically generate a natural language sentence given an image [1–6], for which an encoder-decoder framework with attention mechanisms has achieved great progress in recent years. Usually, Convolutional Neural Network (CNN) is used to encode visual features and a recurrent neural network (RNN) is used to generate a caption [7,8]. The attention mechanism [5,6,9] can generate a dynamical and spatial localized image representation focusing on certain parts of an input image. As a typical solution of most existing work, the image parts are encoded as a set of vectorial representations corresponding to different grids on the feature maps, which are considered as independent from each other [5,10], we called this grid-based attention. The grid-based attention realized by fully connected layer treats the image features as a set of independent vectors, each of which corresponds to a region in the image grids and then calculates attention weights for each vector and aggregates them with weighted sum. However, this operation totally breaks the spatial structure between each grid, which could be harmful to the model to fully understand the scene.

This motivates us to explore an alternative operation for grid-based attention in image captioning. Instead of operating on the vector representation of each image grid, our attention is computed directly on the entire 2D feature map with convolution operations. As opposed to the standard formulation, this alternation is capable of preserving spatial locality, and therefore it may strengthen the role of visual structures in the process of caption generation. In this paper, we propose a convolutional attention module called Structure Preserving Convolutional Attention (SPCA) that can preserve the spatial structure of the image by convolution operations directly on the 2D feature maps. Our SPCA has two submodules: convolution spatial attention and cross channel attention, which can adaptively determine the intended regions to describe the image according to the current decoding state. As shown in Figure 1, the top row is the visualized results with grid-based attention. We can observe that the attentive regions are inaccurate, because the spatial structure of the image features is not preserved when calculating the attention, resulting in partial deviation. However, in our SPCA (the bottom row), the resulting attention area is precise.

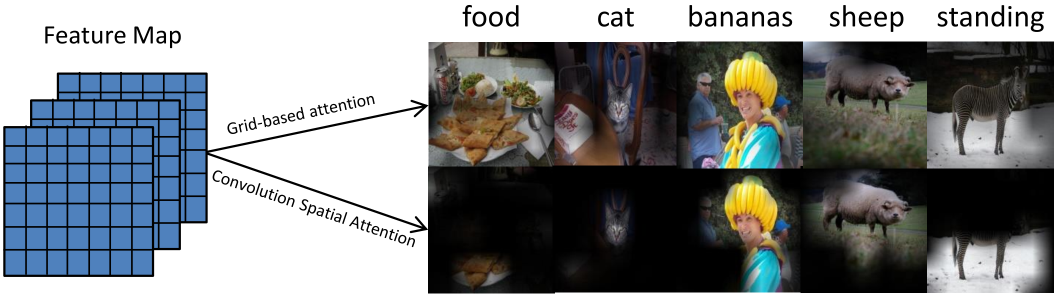


Figure 1. The illustration of the generated attention maps when predicting certain words for our convolutional spatial attention mechanism, comparing with the traditional grid-based attention mechanism.

To verify the effectiveness of the proposed attention module for image captioning, we apply it to two distinctive models, including a standard 1D-Long Short Term Memory (LSTM) model [5] and a recently proposed novel model which represents the latent states of LSTM with 2D dimensional maps. Experiments on two large-scale datasets (MSCOCO and Flickr30K) show that our attention model performs great in the two models.

The contributions of this paper are presented as follows:

- We propose a convolutional spatial attention for preserving spatial structures in the attention map.
- Two attention components, namely cross-channel attention and convolutional spatial attention, are designed to adaptively determine 'what' and 'where' to focus on when predicting each word.
- Extensive experiments on Flickr30K [11] and MSCOCO [12] show the effectiveness of spatial and channel attention mechanism. In addition, our approach demonstrates great performance and generalization ability when applied to two distinctive models with both 1D and 2D LSTM latent states.

2. Related Work

2.1. Image Captioning

Image captioning is a task of generating short descriptions for given images, and it has been an active research topic in computer vision. To generate captions, early techniques mainly rely on detection results, first extracting a set of attributes related to elements within an image and then generating language description. In recent years, in view of Deep Neural Networks'(DNNs) great successes in computer vision, a number of works [2,5,6,13–16] have developed neural network based methods to generate image captions. Specifically, these methods all use encoder-decoder paradigm [17], which uses Convolution Neural Networks (CNNS) to encode the images as features, and then generates captions with Recurrent Neural Networks (RNNs) or one of its variants, e.g., Gated Recurrent Unit (GRU) and Long Short Term Memory (LSTM).

2.2. Attention Mechanism in Captioning

Visual attention has been widely used in various image captioning models in order to allow models to selectively concentrate on objects of interest. Xu et al. [5] combine the memory vector of LSTM with visual features from CNN and feed the fused features to an attention network to compute the weights for features at different spatial locations. Yang et al. [18] propose a reviewer module that applies the visual attention mechanism multiple times while generating the next word. In [10], an adaptive attention mechanism is proposed to determine when to look and where to look at, and no visual information words such as 'a' and 'the' should attend to the visual features. Chen et al. [13] introduce channel-wise attention which is operated on different filters within a convolutional layer. Most of these models generate visual attention in a vector and pay slight attention to temporal information. Without spatial attention structure and temporal information, the compute attention will fail to catch objects accurately and pay attention to what we are not interested in in the next step.

2.3. 2D-Latent-State LSTM

As in [19], a 2D-latent-state LSTM is proposed. For image caption task, it is important to capture and preserve properties of the visual content in the latent states, representing the latent states with 2D maps and connecting them via convolutions. As opposed to the standard formulation, this variant is capable of preserving spatial locality, and therefore it may strengthen the role of visual structures in the process of caption generation. This motivates us to rethink the attention mechanism, and a convolution spatial attention is proposed as follows.

3. Proposed Method

3.1. Overview

We start by briefly describing the encoder-decoder image captioning framework [5,6], and then we describe our SPCA modules.

Encoder-decoder framework: Given an image and the corresponding caption, the encoder-decoder model is directly optimized by the following objective:

$$\theta^* \quad = \quad \underset{\grave{}}{\arg\max} \sum_{(I,y)} \log p(y|I;\theta), \tag{1}$$

where θ is the parameters of the model, I is the given image, and $y = y_1, \ldots, y_n$ is the corresponding caption words.

We adopt LSTM for decoding image features into a sequence of words. The update for its hidden units and cells of an LSTM are defined as:

$$h_t \quad = \quad LSTM\left([x_t; c_t], h_{t-1}\right) \tag{2}$$

where x_t is the embedded word representation, c_t is the context representation, $[\ ;\ ]$ represents concatenate. Furthermore, h_t is the hidden state and memory cell at time t.

Commonly, the context representation c_t, is an important factor in the neural encoder-decoder framework, which provides visual evidence for generating caption. Attention mechanism [5] has been proven to be crucial in producing c_t:

$$c_t = ATT\left(V, h_t\right) \tag{3}$$

where ATT is the attention function, $V \in \mathbb{R}^{C \times W \times H}$ $(C, W, H$ represents the channel, width and height, respectively) is the image feature map, which is the output of a CNN image encoder.

As for conventional spatial attention models, ATT is a grid-based attention. In these models, the context representation c_t is a vector, which is formulated as:

$$
\begin{aligned}
z_{ti} &= \mu^T \tanh(W_v V_i + W_h h_{t-1}) &(4)\\
\alpha_t &= softmax\,(z_t) &(5)\\
c_t &= \sum_{i=1}^{H\times W} \alpha_{ti} V_i &(6)
\end{aligned}
$$

where $V_i \in \mathbb{R}^C$ is a vector representation corresponding to the i-th grid of the image features, W_v, W_h and μ are parameters to be learnt. α_{ti} is the attention weight for V_i.

3.2. Structure Preserving Convolutional Attention

As defined in Equation (6), attention weight α_{ti} is calculated on the vector representation of each grid independently by a fully connected layer. However, this operation totally discards the spatial structure of the image regions, which are known to be significant in image captioning. To overcome such a problem, instead of calculating on vector level, our SPCA attention is computed with convolution operations on the 2D feature maps, and thus can preserve the spatial structure of the image regions. Our SPCA module is composed of two attention components, named convolutional spatial attention and cross channel attention. Figure 2 depicts the framework of our attention module.

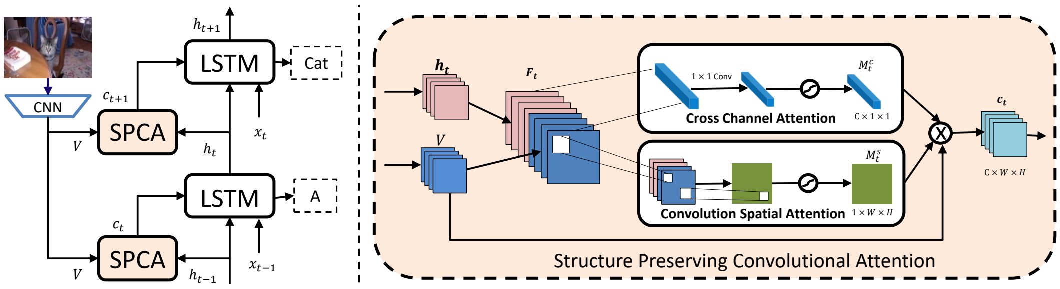


Figure 2. The left side is an illustration of the proposed attention module. The right side is a diagram of our Structure Preserving Convolutional Attention sub-module.

Note that our SPCA's input latent state h, input image feature V and output feature C are all represented as 3D tensors of size $C \times W \times H$. Such a tensor can be considered as a multi-channel map, which comprises C channels, each with a size of $H \times W$. In general 1D-LSTM which has latent state $h' \in \mathbb{R}^{1\times C}$, we utilize tile operation to copy h' into a new tensor $h \in \mathbb{R}^{C\times W\times H}$. As for output C in our SPCA, we finally use a pooling layer to reduce dimension as $C' \in \mathbb{R}^{1\times C}$. However, we can omit these operations at 2D-LSTM.

At the t-th step, given the obtained image feature whose size is $V \in \mathbb{R}^{C\times W\times H}$. After the above mentioned dimensional change operation, we concatenate V in channel dimensions with 2D latent state $h_{t-1} \in \mathbb{R}^{C\times W\times H}$ to form a new feature maps $F_t \in \mathbb{R}^{2C\times W\times H}$. Then, our attention process can be summarized as:

$$
\begin{aligned}
F_t &= [V; h_{t-1}] &(7)\\
M_t^C &= SPCA^C(F_t) &(8)\\
M_t^S &= SPCA^S(F_t) &(9)\\
C_t &= V \otimes \left(M_t^C \otimes M_t^S\right) &(10)
\end{aligned}
$$

where $\otimes$ denotes element-wise multiplication. $SPCA^C, SPCA^S$ represents the attention function for channel and spatial attention modules. $M_t^C \in \mathbb{R}^{C\times 1\times 1}, M_t^S \in \mathbb{R}^{1\times W\times H}$ are the attention weights of resulting channel and spatial attentions, respectively. $C_t \in \mathbb{R}^{C \times W \times H}$ is the resulting context representation.

3.2.1. Convolutional Spatial Attention

This submodule generates M_t^S to tell 'where' in the image should be focused on. While grid-based attention has been proven to be effective on image captioning, it does not take into consideration the spatial structure of image regions and treat them as independent vectors. In fact, lack of spatial structure will lead to inaccurate positioning, which affects the quality of the generated captions.

To enable preservation of the spatial structure of image regions, a convolutional spatial attention is proposed.

We use a convolution operation instead of the original fully connected layer and use 2D latent state and 2D image features instead of those used before. On the one hand, our convolutional spatial attention preserves the structure of the image. On the other hand, we use convolution operation with 3×3 kernel size to provide larger receptive field to accurately determine 'where' we should focus at t step.

In short, our convolutional spatial attention is computed as:

$$M_t^S = \sigma(Conv2(Relu(Conv1(F_t)))) \tag{11}$$

where $M_t^S \in R^{1 \times W \times H}$ is the convolution attention weights, *Conv1* and *Conv2* are both convolution operation with 3×3 kernel sizes, *Relu* denotes RELU activation function and σ denotes the sigmoid function.

3.2.2. Cross Channel Attention

This submodule generates M_t^C to tell 'what' content in the image to describe when generating the current word. While spatial attention has been widely used in previous work, cross channel attention has not been paid much attention. As is described in [19], different channels have different activated regions, which means only several channel will be activated when predicting a word. Thus, we add information h_{t-1} to decide 'what' should be followed at the next step. The cross channel attention is also computed based on $F_t \in \mathbb{R}^{2C \times W \times H}$. First we apply average pooling for each channel to obtain the channel feature $F_c \in \mathbb{R}^{2C \times 1 \times 1}$, then we obtain the cross channel attention map $M_t^C \in \mathbb{R}^{C \times 1 \times 1}$ by convolution layers. In short, the cross channel attention is computed as:

$$M_t^C = \sigma(Conv2(Relu(Conv1(AvgPool(F_t))))) \tag{12}$$

where r denotes the reduction ratio, *Conv1*'s input dimension is 2C and output dimension is 2C/r, *Conv2*'s input dimension is 2C/r and output dimension is C and Kernel sizes of them are both 1×1, *Relu* denotes RELU activation function and σ denotes the sigmoid function.

4. Experiments

4.1. Dataset and Evalution

We evaluate our method on two well-known datasets: (1) **MS-COCO** [12], where most images contain multiple objects in complex natural scenes with abundant context information. Each image has five corresponding captions. We use the split as follows [2]. It uses all 113,287 training set images for training and selects 5000 images for validation and 5000 images for a test from an official validation set. (2) **Flickr-30K** [11], totally contains 31,000 images from Flickr. Because of the lack of official split, we use the split given in [2]. It uses 29,000 images for training, 1000 images for validation and 1000 images for testing. To evaluate our method, four automatic metrics are used in our experiments. Including BLEU4(B4) [20], ROUGEL(RG) [21], METEOR(MT) [22] and CIDER(CD) [23].

4.2. Implementation Details

In our captioning model, for the encoding part, given the powerful capabilities of Resnet-101 and its convenience in controlling variables to compare other methods, we finally adopted widely-used CNN architecture: Resnet-101 [24] to extract image features as input for our SPCA module. When extracting the features, no cropping or re-scaling is applied to the original images, instead, an adaptive spatial average pooling layer is utilized to produce features with a fixed size of $2048 \times 14 \times 14$. For the decoding part, we used the LSTM [25] to generate caption words. For both 1D-LSTM [5] and 2D-LSTM [19], word embedding and attention dimensions are set as 512. We use the Adam [26] to optimize our network with learning rate set as 4×10^{-4}. During the supervised learning, the learning rate is decayed by a factor of 0.5 every five epochs. Each mini-batch contains 20 images. In the test period we adopted BeamSearch [6] strategy, which selects the best caption from some candidates—the beam size is 2. We show the parameters and speed results in Table 1.

Table 1. The parameters and speed results on Structure Preserving Convolutional Attention (SPCA). Speed results based on beam size 3, batch size 20 on an NVIDIA Titan X GPU.

Method	Param	Millisecond/Image
Soft-Attention [5]	248.8 M	57.6 ± 0.7
1D-LSTM + SPCA	264.8 M	70.95 ± 0.9
2D-latent state [19]	205.7 M	55.7 ± 0.4
2D-latent-LSTM + SPCA	223.9 M	64.7 ± 0.3

4.3. Performance and Analysis

4.3.1. Attention Structure Selection

We explore the respective roles of cross channel attention and convolutional spatial attention, and the performance of their various combinations. We choose 2D-LSTM model as our baseline because it has 2D-latent-state that is suitable for our purposes. At the same time, it does not add an attention module so that we can observe the role of different attractions in the captioning system. As we show in Figure 3. (1) **S**: only spatial attention model. (2) **C**: only channel attention model. (3) **C-S**: we first compute the channel attention weights M_t^S. Then based on M_t^S, the image feature V and spatial attention model, we can calculate the modulated feature C_t. (4) **S-C**: we first compute the spatial attention then calculate channel attention weight. (5) **C**||**S**: this model as illustrated in Figure 2. We calculate channel attention and spatial attention in parallel, as Equation (10). We use convolutional spatial attention to preserve spatial structure of attention and concatenate hidden state at each step to get more temporally contextual information.

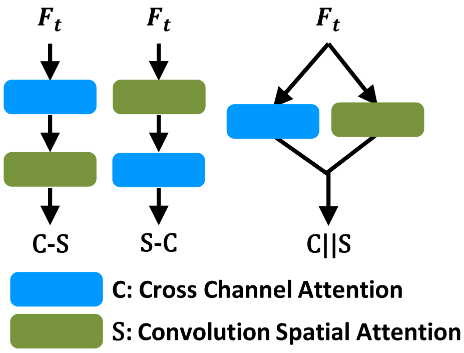


Figure 3. Different structures of our SPCA.

Based on the results listed in Table 2, we have the following observations: (1) Comparing to the performance of S, the performance of C shows that more channel information can help improve in image captioning. Furthermore, all of S-C, C-S and S||C can achieve better performance than only S or C, which proves spatial attention and channel attention are complementary to each other, one for 'where', the other for 'what'. (2) For both datasets, S||C gets better scores than the other sequentially combining pattern.

Table 2. The result of different way use attention module in Flickr-30K and MS-coco datasets.

Dataset	Method	B4	MT	RG	CD
Flickr_30K	Baseline	22.6	19.6	44.5	42.2
	S	22.2	19.3	44.5	44.3
	C	22.9	19.8	45	46.9
	C-S	**23.8**	20.1	45.4	47.5
	S-C	23.4	19.9	45.1	46.9
	S||C	23.7	**20.3**	**45.6**	**49.9**
MS-COCO	Baseline	30.1	24.9	52.8	94.8
	S	30.5	24.6	52.3	91.2
	C	31.9	24.7	52.6	96
	C-S	**33.1**	25.2	53.3	96.9
	S-C	32.5	25	53	96.5
	S||C	32.4	**25.8**	**53.9**	**99.3**

4.3.2. Convolution Kernel Size

In our SPCA the kernel size in convolutional operation has an effect on the receptive field of the attention to get more completely semantic, thus in this experiment we explored the effect of different kernel sizes on model performance. We set three different kernel sizes including 1 × 1, 3 × 3 and 5 × 5. From the result in Table 3, we have the following observations: performance of 1 × 1 is the worst, 5 × 5 is the second and 3 × 3 is the best. For 1 × 1 kernel size, it equals to grid-based attention, which ignores the spatial structures during the attention computed. For 5 × 5 kernel size, it indicates the larger receptive field cannot get better performance. Large receptive field will provide more global information, but including much background information to confuse our attention mechanisms. So the 3 × 3 kernel size gets the best performance—not only does it preserve the spatial structure of the feature, but it also removes the interference information.

Table 3. The result of spatial structure effective in spatial attention.

Dataset	Kernel Size	B4	MT	RG	CD
Flickr-30K	1 × 1	**23.9**	20.2	45.6	48.7
	3 × 3	23.7	**20.3**	**45.6**	**49.9**
	5 × 5	23.6	**20.3**	45.5	49.3
MS-COCO	1 × 1	31.9	25.4	53.2	97.3
	3 × 3	**32.4**	**25.8**	**53.9**	**99.3**
	5 × 5	31.8	25.6	53.6	98.4

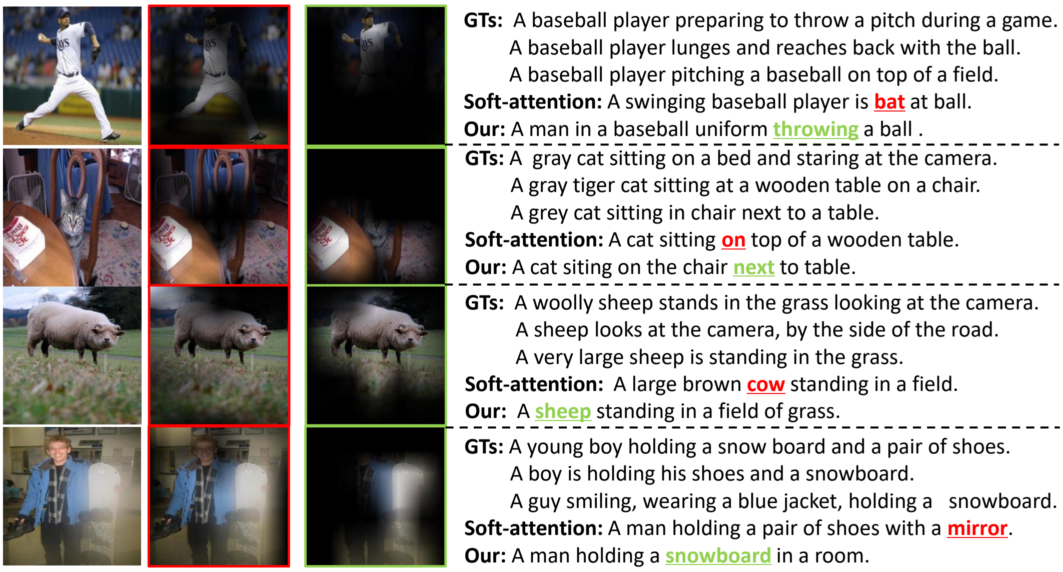
4.3.3. Performance Comparisons

We compare our proposed SPCA with the state-of-art methods on image captioning, including Google NIC [6], Soft-Attention [5], SCA-CNN [13], ATT2in [27] and 2D-latent state [19]. The first three models are based on 1D-LSTM and the last one is for 2D-LSTM. As listed in Table 4, our SPCA outperforms the other models. The '*' represents the model we reproduce ourselves whose setting as above. We apply our SPCA module based on 1D-LSTM: Soft Attention and enhance performance by replacing the VGG [28] based visual encoder with a more powerful ResNet-101 [24] based one. As for 2D-LSTM: 2D-latent state, we only train the model solely on cross-entropy loss without reinforcement learning or fine-tune CNN. (1) Comparing with Soft-attention: we get the higher score, indicating the fact that SPCA preserves the spatial structures during calculating spatial attention with convolutional operation. (2) Comparing with 2D-latent state: a significant increase in scores, due to the fact that our SPCA exploits spatial and cross-channel attentions. (3) Comparing with the other model: our SPCA method greatly exceeds the previous method in performance.

Table 4. Performances compared with state-of-art in Flickr_30K and MS-COCO dataset.

Model	Flikr-30K				MS-COCO			
	B4	MT	RG	CD	B4	MT	RG	CD
Google NIC [6]	-	-	-	-	24.6	-	-	-
Soft-Attention [5]	19.1	18.5	-	-	24.3	23.9	-	-
Soft-Attention*	22.2	19.7	45.0	47.0	33.4	26.1	54.2	101.1
Hard-Attention [5]	19.9	18.5	-	-	25	23	-	-
SCA-CNN [13]	22.3	19.5	-	-	31.1	25	52.4	91.2
ATT2in [27]	-	-	-	-	31.3	26.0	54.3	101.3
2D-latent state [19]	22	-	44.4	42.7	31.9	-	53.8	99.4
2D-latent state*	22.6	19.6	44.5	42.2	30.1	24.9	52.8	94.8
1D-LSTM + SPCA	**23.4**	**20.2**	**46.2**	**50.0**	**33.9**	**26.4**	**55.1**	**106.3**
2D-latent-LSTM + SPCA	**23.7**	**20.3**	**45.6**	**49.9**	**32.1**	**25.8**	**53.9**	**99.3**

We provide some qualitative examples in Figure 4 for a better understanding of our model. We visualized results at one word prediction step. For example, in the first sample, when SPCA module tries to predict the word 'throwing', our attention focuses on the player's hand and the upper part of his body accurately. However, in soft-attention when predicting the word 'bat', it regards him as a 'batter' rather than a 'bowler'. This indicates that our SPCA preserves the image structure with 2D latent state and learns a more accurate attention region.



GTs: A baseball player preparing to throw a pitch during a game.
A baseball player lunges and reaches back with the ball.
A baseball player pitching a baseball on top of a field.
Soft-attention: A swinging baseball player is **bat** at ball.
Our: A man in a baseball uniform **throwing** a ball .

GTs: A gray cat sitting on a bed and staring at the camera.
A gray tiger cat sitting at a wooden table on a chair.
A grey cat sitting in chair next to a table.
Soft-attention: A cat sitting **on** top of a wooden table.
Our: A cat siting on the chair **next** to table.

GTs: A woolly sheep stands in the grass looking at the camera.
A sheep looks at the camera, by the side of the road.
A very large sheep is standing in the grass.
Soft-attention: A large brown **cow** standing in a field.
Our: A **sheep** standing in a field of grass.

GTs: A young boy holding a snow board and a pair of shoes.
A boy is holding his shoes and a snowboard.
A guy smiling, wearing a blue jacket, holding a snowboard.
Soft-attention: A man holding a pair of shoes with a **mirror**.
Our: A man holding a **snowboard** in a room.

Figure 4. Qualitative results for our method performance. From left to right: original image, soft-attention map (red), our attention map (green) corresponding the underline word.

5. Conclusions

In this paper, we propose a structure preserving convolutional attention (SPCA) module for image captioning. There are two submodules: convolutional spatial attention and cross channel attention, which can adaptively determine 'where' and 'what' should be attended to, respectively. Different from the grid-based attention, our SPCA preserves spatial structure and fuses channel information at each step when calculating the attention. Thus, we can get more accurate attention and achieve better performance than popular benchmarks.

Author Contributions: Conceptualization, S.L., R.H. and J.L.; Data curation, S.L., L.G. and F.Z.; Formal analysis, S.L., R.H., J.L., L.G. and F.Z.; Investigation, S.L., L.G. and F.Z.; Methodology, S.L., R.H., J.L., L.G. and F.Z.; Resources, R.H. and J.L.; Software, S.L. and L.G.; Supervision, R.H. and J.L.; Writing—original draft, S.L.; Writing—review and editing, R.H., J.L., L.G. and F.Z.

Funding: This work is supported by the National Nature Science Foundation of China (No. U1736206).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1473–1482.
2. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
3. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *arXiv* **2014**, arXiv:1412.6632.
4. Socher, R.; Karpathy, A.; Le, Q.V.; Manning, C.D.; Ng, A.Y. Grounded compositional semantics for finding and describing images with sentences. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 207–218. [CrossRef]
5. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
6. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
7. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
8. Graves, A. Generating sequences with recurrent neural networks. *arXiv* **2013**, arXiv:1308.0850.
9. Wu, Z.; Cohen, R. Encode, review, and decode: Reviewer module for caption generation. *arXiv* **2016**, arXiv:1605.07912.
10. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 6, p. 2.
11. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
12. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
13. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6298–6306.
14. Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; Deng, L. Semantic Compositional Networks for Visual Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
15. Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; Murphy, K. Improved image captioning via policy gradient optimization of spider. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; Volume 3, p. 3.
16. Ren, Z.; Wang, X.; Zhang, N.; Lv, X.; Li, L.J. Deep reinforcement learning-based image captioning with embedding reward. *arXiv* **2017**, arXiv:1704.03899.

17. Bengio, S.; Vinyals, O.; Jaitly, N.; Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 1171–1179.
18. Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W.W.; Salakhutdinov, R.R. Review networks for caption generation. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2016; pp. 2361–2369.
19. Dai, B.; Ye, D.; Lin, D. Rethinking the form of latent states in image captioning. *Work* **2018**, *8*, 10–13.
20. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
21. Lin, C.Y. *Rouge: A Package for Automatic Evaluation of Summaries*; Text Summarization Branches Out; Association for Computational Linguistics: Barcelona, Spain, 2004.
22. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*; Association for Computational Linguistics: Ann Arbor, MI, USA, 2005; pp. 65–72.
23. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Graves, A. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.
26. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
27. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical Sequence Training for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.