Think and Tell: Preview Network for Image Captioning

Zhihao Zhu
zzh123@stu.xjtu.edu.cn
Zhan Xue
xx674967@stu.xjtu.edu.cn
Zejian Yuan*
yuan.ze.jian@mail.xjtu.edu.cn

Institute of Artificial Intelligence and Robotics
School of Electronic and Information Engineering, Xi'an Jiaotong University
Xi'an, China

Abstract

Image captioning is a particularly challenging task, which has attracted considerable interest in the community of computer vision. Most of the existing methods follow the single-pass forward encoding-decoding process to generate image description sentence word by word. When generating a specific word, these methods are only able to utilize the previously generated words, but not the un-generated future words. However, for humans to describe a scene, it's a common behaviour to first preview and organize all the observed visual contents in a semantically-meaningful order, and then form a complete description sentence. In such process, humans can obtain a global information of the visual contents related to both previous words and possible future words. In this paper, we propose a preview network that incorporates such preview mechanism in the encoder-decoder framework. The proposed model consists of two visual encoders and two language decoders: one encoder is used to extract image's high-level attributes and feed them into the first-stage decoder to preview image's contents and to generate a first coarse image caption. Then, together with the convolutional features extracted by another encoder, this coarse caption is then fed into the second-stage decoder to generate a second refined image caption. The experimental results on the benchmark Microsoft COCO dataset show that our method yields state-of-the-art performance on various quantitative metrics.

1 Introduction

Automatic image captioning presents as a particular challenge in the field of computer vision. This task needs to interpret from the pixel information to natural languages, which are two completely different information forms. It requires a high level of image understanding that goes beyond image classification and object recognition.

Inspired by the successful application of neural network in machine translation, recently, many works [5, 8, 17, 20, 28] have been proposed to use neural network-based method for image captioning. A popular pipeline of these method is to first use a Convolutional Neural Network (CNN) to encode pixel information into high-level abstract features. Then a Recurrent Neural Network (RNN) is used to decode this feature into natural language descriptions.

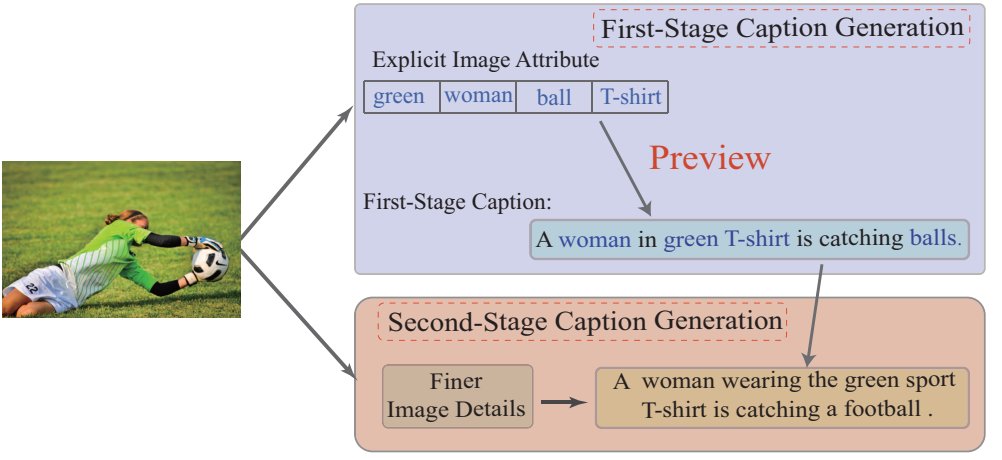

Figure 1: An example illustrating the preview mechanism for generating image descriptions. Image's explicit semantic attributes are first extracted and used for generating a first-stage coarse caption. Then, the first-stage caption and more image visual details are used for generating a second-stage refined caption.

These methods are able to generate coherent and accurate sentences thanks to the good representation ability of CNN and strong capability of RNN for capturing dynamic sequential information. However, most of these methods follow a single-pass forward architecture. When generating a specific word in an image description sentence, these methods are incapable of knowing the information of the future words, which sets a limit on the performance of image captioning.

In addition, on the basis of the encoder-decoder framework, some works [23, 26] also propose that using high-level semantic attributes can boost the image captioning performance. However, these methods either feed image's attributes into the language generator alone, or feed image attributes together with image's convolutional features into a same language generator. For the first case, the problem is that only using the explicit high-level attributes will inevitably leave out some visual details that do not appear in the pre-defined attribute vocabulary. For the second case, as semantic attributes are a form of image's explicit high-level representation while the convolutional features are abstract and do not have semantical meaning, using a single language decoder is hard to well leverage these two levels of image features simultaneously.

Generally, after seeing the visual contents in the image, humans tend to perform a preview of the visual information: organize each visual element and their attributes in a semantically-meaningful order. In such a manner, humans are able to establish a global knowledge of the previously generated words as well as the possible future words. During the preview stage, humans are mainly dealing with image's explicit high-level contents. Then, guided by the preview results, humans are able to describe the image in a finer level by referring to more low-level visual details in the image. Figure 1 offers an example illustrating the above preview process for generating image descriptions.

In this paper, we propose a preview network that is not only able to incorporate the information of both the past and the future words, but can also well leverage the advantages of image's semantic attributes and convolutional features. An overview of our proposed model is shown in Figure 2. Our preview network consists of two visual encoders and two

LSTM-based decoders. The process of generating the caption of an image is as follows: the image's high-level semantic attributes are extracted by an encoder and are fed into the first LSTM-based decoder, which we note as preview LSTM (p-LSTM). The p-LSTM generates a coarse image caption that provides a general information of what the word sequence might be. Then, we apply attention mechanism to select the most important parts of image's convolutional features extracted by another encoder, as well as the word sequence generated at the first stage. The outputs of the attention model are fed into the second-stage decoder: refinement LSTM (r-LSTM) to generate a final refined image caption. By previewing image's high-level visual contents in the first-stage and capturing low-level visual details in the second stage, our model is able to generate more accurate image captions. On the popular benchmark Microsoft COCO dataset, our model improves performance consistently against a very strong baseline and outperforms many published state-of-the-art results.

2 Related Work

The problem of describing images with natural languages at the scene level has long been studied in both the field of computer vision and natural language processing. So far, many pioneering methods have been proposed to tackle this task. Generally, these methods can be divided into three categories according to the way of generating sentence: template-based method, transfer-based method, and neural network-based method.

The template-based methods [9, 10, 25] first detect objects, actions, and attributes by using several classifiers respectively, then fill them in a fixed sentence template, which follows certain predefined syntactic rules, e.g. using a subject-verb-object template. This category of method is simple and intuitive, but lacks the flexibility to generate diverse sentences due to the limitation set by the pre-fixed sentence template.

Given a query image, the transfer-based methods first search for visually-similar image in the database, finds and transfers the best language descriptions from the nearest neighbor captions for the description of the query image [11, 14]. This kind of method is able to generate more natural and human-like sentences than template-based methods. However, the generated captions may not correctly describe the visual content of the query image, as it's hard to accurately leverage the visual similarity between images.

Most neural network-based methods follow the Encoder-Decoder framework, which first uses a deep CNN to encode an image into an abstract representation, and then uses a RNN to decode that information into a natural language sentence which can describe image in details. Mao et al. [13] propose a Multimodal Recurrent Neural Network (MRNN) that uses an RNN to learn the text embedding, and a CNN to learn the image representation. Vinyals et al. [20] use LSTM as the decoder to generate sentences, and provide the image features as input to the LSTM directly. Chen et al. [3] learned a bi-directional mapping between images and their sentence-based descriptions using RNN. What's more, inspired by the successful application of attention mechanism in machine language translation [1], spatial attention [24, 29] has also been widely adopted in the task of image captioning. It's a feedback process that selectively maps a representation of partial regions or objects in the scene. Although there exist many more approaches to improve the image captioning system, most of them follow a single-stage encoding-decoding approach, which directly decodes the extracted image features into word sequence. Thus, it's impossible for the system to know what the future words might be when generating a specific word at the current time step.

In our model, we use a two-stage caption generation strategy: in the first stage, our model

generates a coarse version of the word sequence which carries the information of possible word selection and words' sequential orders. Then in the second stage, this first-generated word sequence is used as guidance information to generate a second refined version of image caption.

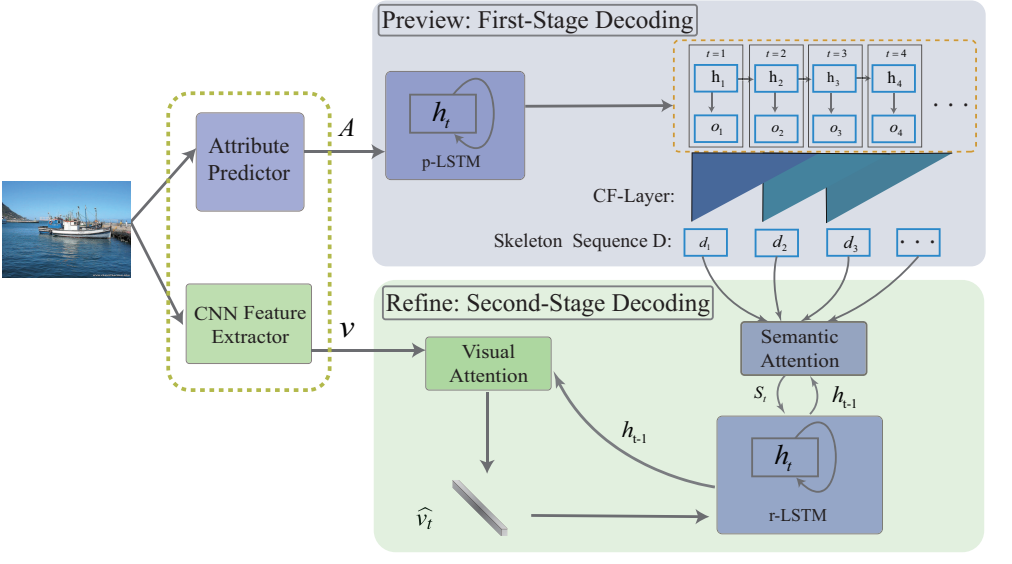

Figure 2: The framework of the proposed image captioning method. CF-Layer represents the convolutional filtering layer, and v is the convolutional feature vectors, A is the semantic attributes of the image.

The idea of preview has not been well explored in the image captioning task. Several related works employing two-stages of sequence generation can be found in machine translation. [2, 16] employ the strategy of post-editing: one model translates the source language into target language, and a separate model refines this translated sentence. As a comparison, we use an integral model where two decoding stages are coupled together. The work that is closest to ours is the Skeleton-Attribute Decomposition by Yufei Wang et al. [21]. They also employ a coarse-to-fine strategy by generating the image caption sentence in two parts: (1) a skeleton sentence describing objects and relationships, and (2) objects' attribute. However, they set a hard-division between the objects and objects' attributes, which pre-defines a sentence template and limits sentence's variety and flexibility. Different from theirs, our model does not manually decompose sentences into different components, but instead treats each sentence component equally during each decoding stage. In such a manner, our model is able to generate more accurate and natural image descriptions.

3 Proposed Method

In this section, we describe our preview network for image captioning in details. Our method consists of the following two processes:

(a) Preview: First-stage decoding Feeding the image's attributes as input and generating a coarse image caption. Then, applying a convolutional filtering to generate a Skeleton Sequence.

(b) Refine: Second-stage decoding Feeding the Skeleton Sequence and image's convolutional features into two separate attention networks, whose outputs are then used as context information to generate a refined image caption.

3.1 Preview

The first-stage decoder consists of a single LSTM with 512 hidden state. Its initial state is image's high-level semantic attributes *A (attribute extraction will be specified in section 3.4)*. In the following time steps, the input is a non-linear projection of the p-LSTM's previous output. A START token and an END token are assigned to each sentence like in many other works [7, 18, 26]. The output word sequence generated by the p-LSTM is noted as $O = \{o_1, o_2, ..., o_T\}$, $o_i \in \mathbb{R}^E$, where T is the length of the sequence and E is the dimension of each output word o_i in the embedding form.

Convolutional filtering: Prior to feeding the first-stage caption into next stage of decoding, it's necessary to perform a filtering of the words. Note that the first-stage decoder's role is to organize the most important visual contents in a semantically-meaningful order and to generate a word sequence that provides a general information of what the final caption might be. However, the words in the whole sentence are not equally important. For example, we notice that the words like "the", "a", "and" appeared very frequently in the training captions, but they are relatively weaker in its semantical importance. Their changes of position in the sentence have little impact on the general meaning of the sentence. Therefore, we add a convolutional filtering on the sequence O, followed by attention mapping (attention mapping is detailed in section 3.2) to select the most important parts of O.

The convolutional layer consists of several same-structured blocks: each contains a convolution kernel parameterized as $W \in \mathbb{R}^{k \times E \times n}, b_w \in \mathbb{R}^n$, where k is the convolutional kernel width, n is convolutional kernel's output channel number. We apply the convolutional filtering on the sequence O to obtain a new sequence $D = \{d_1, d_2, ..., d_T\}$, which we refer as Skeleton Sequence. We apply padding to keep sequence $D's$ length the same with the sequence O. Therefore, D has a length of T and a dimension of n. Mathematically, d_i is obtained by:

$$d_i = tanh(W^i[o_i, o_{i+1}, ..., o_{i+k-1}] + b_w^i), \tag{1}$$

where W^i and b_w^i correspond to the parameters of convolutional kernel in i-th block, and $tanh(\cdot)$ is used as the nonlinear activation function.

3.2 Refinement

Once the Skeleton Sequence D is generated by the first-stage decoder, we feed it into an attention network to obtain the semantic contextual information for the second-stage decoder. At the time step t, the input of the second-stage decoder consists of following information: its previous hidden state h_{t-1}, its previous output word y_{t-1}, the visual contextual information $\widehat{v}_t$ and the semantic contextual information s_t.

Visual Attention: Visual contextual information $\widehat{v}_t$ are obtained via a visual attention mechanism. First we use a deep convolutional neural network in order to extract a set of feature vectors which we refer to as convolutional feature vectors. The extractor produces M vectors $\mathbf{v} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_M\}$, each $\mathbf{v}_i$ is a L-dimensional representation corresponding to a part of the image.

Then in the visual attention process, a score α_t^i is assigned to each convolutional feature vector based on their relevance with r-LSTM's previous hidden state h_{t-1}. we use the commonly-used bilinear function to model the relevance in vector space:

$$\alpha_t^i \propto exp(h_{t-1}\tilde{U}v_i), \tag{2}$$

where the length P and width L of the parameter matrix $\tilde{U} \in \mathbb{R}^{P\times L}$ represents the dimension of r-LSTM's hidden state and the dimension of convolutional feature vector, respectively. The exponent is taken to normalize over all the $\mathbf{v}_i$ in a softmax fashion. Then, we gather all the visual features to obtain the visual contextual information $\widehat{\mathbf{v}}$ by using the weighted sum:

$$\widehat{\mathbf{v}_\mathbf{t}} = \sum_{i=1}^{M} \alpha_t^i \mathbf{v}_i. \tag{3}$$

Semantic Attention: Semantic contextual information s_t is calculated in a similar approach with the visual contextual information. The score β_t^i assigned to each d_i is computed as:

$$\beta_t^i \propto exp(h_{t-1}\tilde{Z}d_i), \tag{4}$$

where $\tilde{Z} \in \mathbb{R}^{P\times n}$ is the parameter matix. Then, the semantic contextual information s_t is obtained as following:

$$s_t = \sum_{i=1}^{T} \beta_t^i d_i. \tag{5}$$

As can be seen from the above computation, the semantic contextual information makes use of the whole Preview Sequence D generated by the first-stage decoder. In other words, it considers the global information including both the words proceeding and after it.

Once the visual contextual information $\widehat{\mathbf{v}_t}$ and semantic contextual information s_t are extracted, r-LSTM's hidden state h_t can be calculated as $h_t = LSTM([y_{t-1};\widehat{v_t};s_t],h_{t-1})$, and its output y_t can be calculated by transforming the matrix $[y_{t-1};\widehat{v_t};s_t;h_t]$.

3.3 High-level Semantic attributes extraction

Similar to [23], we first establish the attributes vocabulary by selecting c most common words in the captions. To reduce the information redundancy, we perform a manual filtering of plurality (e.g. "woman" and "women") and semantic overlapping (e.g. "child" and "kid"), by classifying those words as the same attribute. Finally, we obtain a vocabulary of 196 attributes, which is more compact than [23]. Given this attribute vocabulary, we can associate each image with a set of attributes according to its captions.

We then wish to predict the attributes given a test image. This can be viewed as a multi-label classification problem. We follow [22] to use a Hypotheses-CNN-Pooling (HCP) network to learn attributes from local image patches. It produces the probability score for each attribute that an image may contain, and the top-ranked ones are selected to form the attribute vector A as the input of the preview network.

3.4 Training

We train our model in two consecutive steps: (1) Given the extracted image attributes A as input and image's ground-truth caption as supervision information, the p-LSTM is first

	B-1	B-2	B-3	B-4	M	R	C
Google NIC [20]	66.6	46.1	32.9	24.6	-	-	-
Soft attention [24]	70.7	49.2	34.4	24.3	23.90	-	-
Semantic attention [27]	73.1	56.5	42.4	31.6	25.00	53.5	94.3
Skeleton-Key [21]	74.2	57.7	44.0	33.6	26.8	55.2	107.3
PG-SPIDEr-TAG [12]	**75.1**	**59.1**	45.7	34.2	25.5	55.1	104.2
Baseline	74.8	55.8	41.1	30.2	**27.0**	57.8	109.8
PrevN (Ours)	74.6	57.9	**46.7**	**34.8**	25.9	**58.6**	**110.9**

Table 1: Comparison of different methods on standard evaluation metrics: BLEU-1 (B-1), BLEU-2 (B-2), BLEU-3 (B-3), BLEU-4 (B-4), METEOR (M), ROUGE (R), CIDEr (C). PrevN stands for our Preview Network. Missing numbers are marked by -.

trained alone. (2) After obtaining the first-stage caption and image's convolutional features, the convolutional filtering layer, visual attention network, semantic attention network and r-LSTM are trained jointly.

A same loss function is used in two training steps:

$$L^{(s)} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{L^{(i)}} \log p^{(s)}(w_t^{(i)}) + \lambda \cdot \left\|\theta^{(s)}\right\|_2^2, \quad s \in \{P,R\} \tag{6}$$

where the superscript (s) in loss function $L^{(s)}$ represents the training stage, namely, preview stage (P) and refine stage (R). N is the number of training examples and $L^{(i)}$ is the length of the sentence for the i-th training example. $p(w_t^{(i)})$ corresponds to the Softmax activation of the t-th output of the LSTM. θ represents all the model parameters that need to be trained, and $\lambda \cdot \|\theta\|_2^2$ is a regularization term.

4 Experiment

In this section, we will specify our experimental methodology and verify the effectiveness of our preview mechanism for image caption generation. *(We plan to publish our code in* [1]*.)*

4.1 Setup

Data and Metrics: We conduct the experiment on the popular benchmark: Microsoft COCO dataset. For fair comparison, we follow the commonly used split in many other works: 82,783 images are used for training, 5,000 images for validation, and 5,000 images for testing. Some images have more than 5 corresponding captions, the excess of which will be discarded for consistency. We directly use the publicly available code [2] provided by Microsoft for result evaluation, which includes BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, CIDEr, and ROUGH-L.

Implementation details: For the encoding part in our Preview Network: **1)** The image's convolutional visual features $\mathbf{v}$ are extracted from the last 512-dimensional convolutional layer of the VGGNet [19]. **2)** For the attribute extractor, after obtaining the 196-dimensional output from the last fully-connected layer, we keep the top 10 attributes with the highest

[1] https://github.com/ZhihaoZhu/Image-Captioning-with-Preview-Network/
[2] https://github.com/tylin/coco-caption

	B-1	B-2	B-3	B-4	M	R	C
PrevN	74.6	57.9	46.7	34.8	25.9	58.6	110.9
PrevN+Res	75.1	59.2	47.3	35.0	26.2	59.3	111.6
PrevN+Beam	75.3	59.0	47.3	35.3	26.1	58.6	111.2
PrevN+Res+Beam	**76.0**	**59.6**	**48.0**	**35.7**	**27.2**	**59.9**	**112.5**

Table 2: Performance comparison of several different systematic variants of our method on the MSCOCO data set.

scores to form the attribute vector A. For the decoding part, the dimension of the first-stage decoder's input and hidden state are both set to 258, and the *tanh* is used as the nonlinear activation function. In addition, we use Glove feature representation [15] with 300 dimensions as our word embedding E for both LSTM's input and output word vectors. For the convolutional filtering layer, we use a convolutional kernel size of $3 \times 300 \times 512$. The LSTM used as our second-stage decoder has an input dimension of 1024 and a hidden state dimensions of 1024 as well.

In the training procedure, we use Adam [9] algorithm for model updating with a mini-batch size of 128. We set two language models' learning rate to 0.001 and the dropout rate to 0.5. The whole training process takes about 16 hours on a single NVIDIA TITAN X GPU.

Baseline: In order to demonstrate the effectiveness of our method, we also present the results generated by baseline method. The baseline method is trained and tested on the same dataset without using the preview mechanism. For each dataset, we use the same network architecture for language generator. The CNN features and image attributes are both fed into the baseline model as inputs. All the hyper-parameters and CNN encoder remain the same for our baseline model.

4.2 Quantitative evaluation results

Table 1 compares our PrevN method to several other state-of-the-art methods on the task of image captioning on MSCOCO dataset. We note that we obtain comparable BLEU-2 and METEOR score with PG-SPIDEr-TAG [12] and better BLEU-3, BLEU-4, ROUGE-L and CIDEr scores than [12, 20, 21, 24, 27] on the test set. Our BLEU-1 score is lower than [12] method, and METEOR is lower than [21] , but the margins are very small. The better performance on metrics like BLEU-3, BLEU-4 and CIDEr indicates that our preview network is good at predicting captions that have longer n-gram matching with the ground-truth captions. This well reflects the improvement on the model's language modeling ability. It can be explained by the fact that our model uses two stages of language decoding, where the preview sequence generated by the first decoding stage offers an useful guidance for the second stage decoding.

In the experiments, several systematic variants of our method are considered: (1) PrevN+Res replaces the VGGNet-based encoder with the more powerful ResNet [6]. (2) PrevN+Beam performs beam search instead of using greedy search for sampling the maximum-probability words. (3) PrevN+Res+Beam, as its name suggests, combines ResNet and beam search. We show the results of comparison among different systematic variants in the Table 2. The results show that the benefit of using ResNet as encoder and applying beam search strategy are addictive, which can be demonstrated by the further performance improvement by PrevN+Res+Beam. What's more, we note that the performance improvement brought by using more powerful feature extractors is considerably smaller than that brought by adding

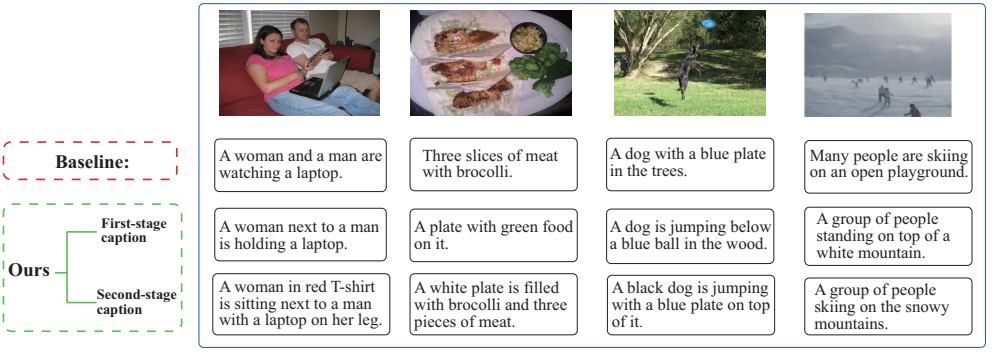

Figure 3: The sample images and their descriptions. The original caption is generated without preview by our baseline method. The new caption is generated with our Preview Network.

preview mechanism, which further demonstrate the advantages of our proposed method.

4.3 Qualitative evaluations

See Figure 3 for a qualitative comparison of captions generated by our method and the baseline model. We observe that our model can better capture details in the target image. For example, in the first image, our model is able to describe woman's wearing in details using "in red T-shirt", which has not been explored by the baseline model. In the fourth image, we predict precisely the background setting as "snowy mountain", while the baseline model made a wrong prediction as "open playground". Moreover, our model explores the spatial relationship between objects more accurately. For example, in the third image, "on top of it" well describes the relationship between the dog and the plate, which however, is not discovered by the baseline model. A similar example can be found in the first image too: "with a laptop on her leg" in our model compared to the inaccurate "watching a laptop" in the baseline model. These examples demonstrate that the preview mechanism has a beneficial influence on the image caption generation.

4.4 Analysis of Two Stages of Captions

The role of preview LSTM and refinement LSTM might seem to be very similar, as they are both used for generating word sequences. But we clarify that they decode the image's visual content in two different levels. From the examples shown in Figure 3, we note that the captions generated by the preview LSTM are less natural and coherent than the captions generated by refinement LSTM. However, most first-stage captions do retain almost the same semantic order of image's contents that also appear in the second-stage captions, such as the objects, objects' attributes and relationships. The comparison results in the Figure 3 show that using such "previewed" word sequences containing the global information of the image's visual contents are beneficial for increasing the accuracy of the final generated image descriptions.

5 Conclusion

In this paper, we propose a novel method for image captioning, which has achieved state-of-the-art performance. Different from other methods who follow a single-forward encoding-decoding approach, our image captioning model applies two stages of decoding: the first-stage decoding is used for generating a coarse word sequence, and the second-stage decoding is used to refine that coarse sentence. The experimental results show that by adding the preview mechanism, the image captioning model is able to obtain significant improvements on multiple evaluation metrics. What's more, captions generated by our method are also more natural and contain more accurate details in the image. For next steps, we plan to test whether stacking more decoders in the same approach can improve the image captioning performance. We also would like to test the application of preview mechanism in other sequence generation tasks.

Acknowledgement: This work was supported by the National Key R&D Program of China (No.2016YFB1001001) and the National Natural Science Foundation of China (No.91648121, No.61573280).

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[2] Rajen Chatterjee, José G. C. de Souza, Matteo Negri, and Marco Turchi. The FBK participation in the WMT 2016 automatic post-editing shared task. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 745–750, 2016.

[3] Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *CoRR*, abs/1411.5654, 2014.

[4] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, pages 15–29, 2010.

[5] Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, and Changshui Zhang. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2321–2334, 2017.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[7] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2407–2415, 2015.

[8] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, 2017.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[10] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12): 2891–2903, 2013.

[11] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 359–368, 2012.

[12] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Optimization of image description metrics using policy gradient methods. *CoRR*, abs/1612.00370, 2016.

[13] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632, 2014.

[14] Rebecca Mason and Eugene Charniak. Nonparametric method for data-driven image captioning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 592–598, 2014.

[15] Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014. ACL.

[16] Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. Pre-translation for neural machine translation. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1828–1836, 2016.

[17] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 817–834, 2016.

[18] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 817–834, 2016.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[20] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):652–663, 2017.

[21] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W. Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7378–7387, 2017.

[22] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. CNN: single-label to multi-label. *CoRR*, abs/1406.5726, 2014.

[23] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. What value do explicit high level concepts have in vision to language problems? In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 203–212, 2016.

[24] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057, 2015.

[25] Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2011.

[26] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4904–4912, 2017.

[27] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *CoRR*, abs/1603.03925, 2016.

[28] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3107–3115, 2017.

[29] Zhihao Zhu, Zhan Xue, and Zejian Yuan. Topic-guided attention for image captioning. *arXiv preprint arXiv:1807.03514*.