

simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions

Fenglin Liu^{1*}, Xuancheng Ren^{2*}, Yuanxin Liu¹, Houfeng Wang² and Xu Sun²

¹School of ICE, Beijing University of Posts and Telecommunications

²MOE Key Laboratory of Computational Linguistics, School of EECS, Peking University

lfl@bupt.edu.cn, renxc@pku.edu.cn, yuanxinLIU@bupt.edu.cn

{wanghf, xusun}@pku.edu.cn

Abstract

The encode-decoder framework has shown recent success in image captioning. Visual attention, which is good at detailedness, and semantic attention, which is good at comprehensiveness, have been separately proposed to ground the caption on the image. In this paper, we propose the Stepwise Image-Topic Merging Network (*simNet*) that makes use of the two kinds of attention at the same time. At each time step when generating the caption, the decoder adaptively merges the attentive information in the extracted topics and the image according to the generated context, so that the visual information and the semantic information can be effectively combined. The proposed approach is evaluated on two benchmark datasets and reaches the state-of-the-art performances.¹

1 Introduction

Image captioning attracts considerable attention in both natural language processing and computer vision. The task aims to generate a description in natural language grounded on the input image. It is a very challenging yet interesting task. On the one hand, it has to identify the objects in the image, associate the objects, and express them in a fluent sentence, each of which is a difficult sub-task. On the other hand, it combines two important fields in artificial intelligence, namely, natural language processing and computer vision. More importantly, it has a wide range of applications, including text-based image retrieval, helping visually impaired people see (Wu et al., 2017), human-robot interaction (Das et al., 2017), etc.

Models based on the encoder-decoder framework have shown success in image captioning. According to the pivot representation, they can be

*Equal Contributions

¹ The code is available at <https://github.com/lancopku/simNet>



Soft-Attention: a open laptop computer sitting on top of a table

ATT-FCN: a dog sitting on a desk with a laptop computer and mouse

simNet: a open laptop computer and mouse sitting on a table with a dog nearby

Figure 1: Examples of using different attention mechanisms. Soft-Attention (Xu et al., 2015) is based on visual attention. The generated caption is **detailed** in that it knows the visual attributes well (e.g. *open*). However, it omits many objects (e.g. *mouse* and *dog*). ATT-FCN (You et al., 2016) is based on semantic attention. The generated caption is more **comprehensive** in that it includes more objects. However, it is bad at associating details with the objects (e.g. missing *open* and mislocating *dog*). *simNet* is our proposal that effectively merges the two kinds of attention and generates a detailed and comprehensive caption.

roughly categorized into models based on visual information (Vinyals et al., 2015; Chen and Zitnick, 2015; Mao et al., 2014; Karpathy and Li, 2015, 2017), and models based on conceptual information (Fang et al., 2015; You et al., 2016; Wu et al., 2016). The later explicitly provides the visual words (e.g. *dog*, *sit*, *red*) to the decoder instead of the image features, and is more effective in image captioning according to the evaluation on benchmark datasets. However, the models based on conceptual information have a major drawback that it is hard for the model to associate the details with the specific objects in the image, because the visual words are inherently unordered in semantics. Figure 1 shows an example. For semantic attention, although *open* is provided as a visual word, due to the insufficient use of visual information, the model gets confused about what objects *open* should be associated with and thus discards *open* in the caption. The model may even associate the details incorrectly, which is the case

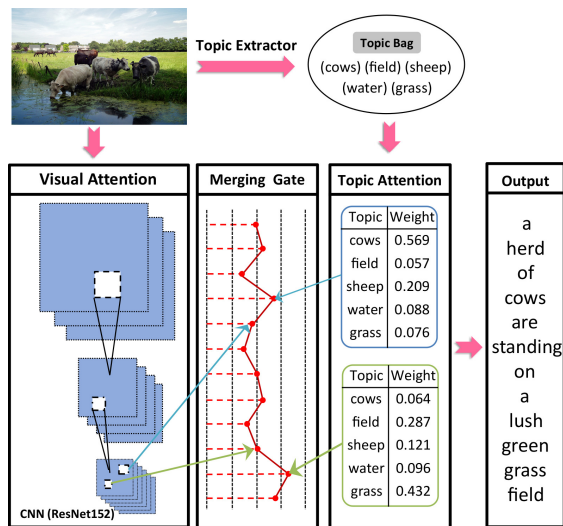


Figure 2: Illustration of the main idea. The visual information captured by CNN and the conceptual information in the extracted topics are first condensed by attention mechanisms respectively. The merging gate then adaptively adjusts the weight between the visual information and the conceptual information for generating the caption.

for the position of the dog. In contrast, models based on the visual information often are accurate in details but have difficulty in describing the image comprehensively and tend to only describe a subregion.

In this work, we get the best of both worlds and integrate visual attention and semantic attention for generating captions that are both detailed and comprehensive. We propose a **Stepwise Image-Topic Merging Network** as the decoder to guide the information flow between the image and the extracted topics. At each time step, the decoder first extracts focal information from the image. Then, it decides which topics are most probable for the time step. Finally, it attends differently to the visual information and the conceptual information to generate the output word. Hence, the model can efficiently merge the two kinds of information, leading to outstanding results in image captioning.

Overall, the main contributions of this work are:

- We propose a novel approach that can effectively merge the information in the image and the topics to generate cohesive captions that are both detailed and comprehensive. We refine and combine two previous competing attention mechanisms, namely visual attention and semantic attention, with an importance-based merging gate that effectively combines

and balances the two kinds of information.

- The proposed approach outperforms the state-of-the-art methods substantially on two benchmark datasets, Flickr30k and COCO, in terms of SPICE, which correlates the best with human judgments. Systematic analysis shows that the merging gate contributes the most to the overall improvement.

2 Related Work

A large number of systems have been proposed for image captioning. Neural models based on the encoder-decoder framework have been attracting increased attention in the last few years in several multi-discipline tasks, such as neural image/video captioning (NIC) and visual question answering (VQA) (Vinyals et al., 2015; Karpathy and Li, 2015; Venugopalan et al., 2015; Zhao et al., 2016; Zhang et al., 2017). State-of-the-art neural approaches (Anderson et al., 2018; Liu et al., 2018; Lu et al., 2018) incorporate the attention mechanism in machine translation (Bahdanau et al., 2014) to generate grounded image captions. Based on what they attend to, the models can be categorized into visual attention models and semantic attention models.

Visual attention models pay attention to the image features generated by CNNs. CNNs are typically pre-trained on the image recognition task to extract general visual signals (Xu et al., 2015; Chen et al., 2017; Lu et al., 2017). The visual attention is expected to find the most relevant image regions in generating the caption. Most recently, image features based on predicted bounding boxes are used (Anderson et al., 2018; Lu et al., 2018). The advantages are that the attention no longer needs to find the relevant generic regions by itself but instead find relevant bounding boxes that are object orientated and can serve as semantic guides. However, the drawback is that predicting bounding boxes is difficult, which requires large datasets (Krishna et al., 2017) and complex models (Ren et al., 2015, 2017a).

Semantic attention models pay attention to a predicted set of semantic concepts (Fang et al., 2015; You et al., 2016; Wu et al., 2016). The semantic concepts are the most frequent words in the captions, and the extractor can be trained using various methods but typically is only trained on the given image captioning dataset. This kind

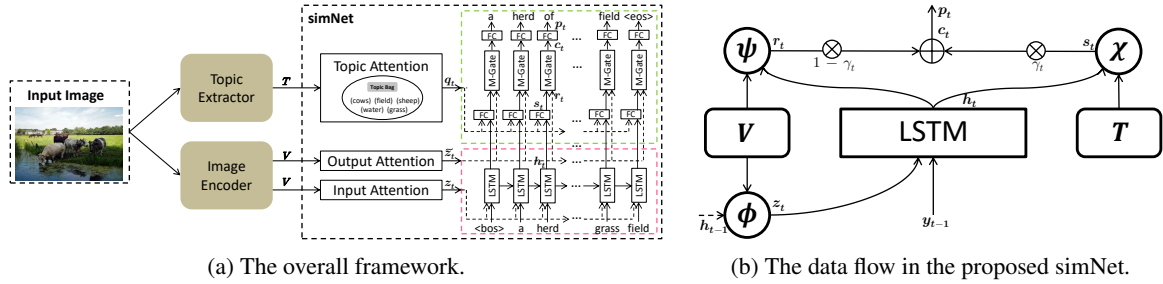


Figure 3: Illustration of the proposed approach. In the right plot, we use ϕ , ψ , χ to denote input attention, output attention, and topic attention, respectively.

of approach can be seen as the extension of the earlier template-based slotting-filling approaches (Farhadi et al., 2010; Kulkarni et al., 2013).

However, few work studies how to combine the two kinds of attention models to take advantage of both of them. On the one hand, due to the limited number of visual features, it is hard to provide comprehensive information to the decoder. On the other hand, the extracted semantic concepts are unordered, making it hard for the decoder to portray the details of the objects correctly.

This work focuses on combining the visual attention and the semantic attention efficiently to address their drawbacks and make use of their merits. The visual attention is designed to focus on the attributes and the relationships of the objects, while the semantic attention only includes words that are objects so that the extracted topics could be more accurate. The combination is controlled by the importance-based merging mechanism that decides at each time step which kind of information should be relied on. The goal is to generate image captions that are both detailed and comprehensive.

3 Approach

Our proposed model consists of an image encoder, a topic extractor, and a stepwise merging decoder. Figure 3 shows a sketch. We first briefly introduce the image encoder and the topic extractor. Then, we introduce the proposed stepwise image-topic merging decoder in detail.

3.1 Image Encoder

For an input image, the image encoder expresses the image as a series of visual feature vectors $V = \{v_1, v_2, \dots, v_k\}$, $v_i \in \mathbb{R}^g$. Each feature corresponds to a different perspective of the image. The visual features serve as descriptive guides of the objects in the image for the decoder. We use a

ResNet152 (He et al., 2016), which is commonly used in image captioning, to generate the visual features. The output of the last convolutional layer is used as the visual information:

$$V = W^{V,I} \text{CNN}(I) \quad (1)$$

where I is the input image, and $W^{V,I}$ shrinks the last dimension of the output.²

3.2 Topic Extractor

Typically, identifying an object requires a combination of visual features, and considering the limited capacity of the visual features, it is hard for the conventional decoder to describe the objects in the image comprehensively. An advance in image captioning is to provide the decoder with the semantic concepts in the image directly so that the decoder is equipped with an overall perspective of the image. The semantic concepts can be objects (e.g. *person*, *car*), attributes (e.g. *off*, *electric*), and relationships (e.g. *using*, *sitting*). We only use the words that are objects in this work, the reason of which is explained later. We call such words **topics**. The topic extractor concludes a list of candidate topic embeddings $T = \{w_1, w_2, \dots, w_m\}$, $w_i \in \mathbb{R}^e$ from the image, where e is the dimension of the topic word embeddings. Following common practice (Fang et al., 2015; You et al., 2016), we adopt the weakly-supervised approach of Multiple Instance Learning (Zhang et al., 2006) to build a topic extractor. Due to limited space, please refer to Fang et al. (2015) for detailed explanation.

Different from existing work that uses all the most frequent words in the captions as valid semantic concepts or visual words, we only include the object words (nouns) in the topic word list. Existing work relies on attribute words and rela-

²For conciseness, all the bias terms of linear transformations in this paper are omitted.

relationship words to provide visual information to the decoder. However, it not only complicates the extracting procedure but also contributes little to the generation. For an image containing many objects, the decoder is likely to combine the attributes with the objects arbitrarily, as such words are specific to certain objects but are provided to the decoder unordered. In contrast, our model has visual information as additional input and we expect that the decoder should refer to the image for such kind of information instead of the extracted concepts.

3.3 Stepwise Image-Topic Merging Decoder

The essential component of the decoder is the proposed stepwise image-topic merging network. The decoder is based on an LSTM (Hochreiter and Schmidhuber, 1997). At each time step, it combines the textual caption, the attentive visual information, and the attentive conceptual information as the context for generating an output word. The goal is achieved by three modules, the visual attention, the topic attention, and the merging gate.

Visual Attention as Output The visual attention attends to attracting parts of the image based on the state of the LSTM decoder. In existing work (Xu et al., 2015), only the previous hidden state $\mathbf{h}_{t-1} \in \mathbb{R}^d$ of the LSTM is used in computation of the visual attention:

$$\mathbf{Z}_t = \tanh(\mathbf{W}^{Z,V} \mathbf{V} \oplus \mathbf{W}^{Z,h} \mathbf{h}_{t-1}) \quad (2)$$

$$\boldsymbol{\alpha}_t = \text{softmax}(\mathbf{Z}_t \mathbf{w}^{\alpha,Z}) \quad (3)$$

where $\mathbf{W}^{Z,V} \in \mathbb{R}^{k \times g}$, $\mathbf{W}^{Z,h} \in \mathbb{R}^{k \times d}$, $\mathbf{w}^{\alpha,Z} \in \mathbb{R}^k$ are the learnable parameters. We denote the matrix-vector addition as \oplus , which is calculated by adding the vector to each column of the matrix. $\boldsymbol{\alpha}_t \in \mathbb{R}^k$ is the attentive weights of \mathbf{V} and the attentive visual input $\mathbf{z}_t \in \mathbb{R}^g$ is calculated as

$$\mathbf{z}_t = \mathbf{V} \boldsymbol{\alpha}_t \quad (4)$$

The visual input \mathbf{z}_t and the embedding of the previous output word \mathbf{y}_{t-1} are the input of the LSTM.

$$\mathbf{h}_t = \text{LSTM}\left(\begin{bmatrix} \mathbf{z}_t \\ \mathbf{y}_{t-1} \end{bmatrix}, \mathbf{h}_{t-1}\right) \quad (5)$$

However, there is a noticeable drawback that the previous output word \mathbf{y}_{t-1} , which is a much stronger indicator than the previous hidden state \mathbf{h}_{t-1} , is not used in the attention. As \mathbf{z}_t is used as the input, we call it **input attention**. To overcome that drawback, we add another attention that incorporates the current hidden state \mathbf{h}_t , which is

based on the last generated word \mathbf{y}_{t-1} :

$$\tilde{\mathbf{Z}}_t = \tanh(\tilde{\mathbf{W}}^{Z,V} \mathbf{V} \oplus \tilde{\mathbf{W}}^{Z,h} \mathbf{h}_t) \quad (6)$$

$$\tilde{\boldsymbol{\alpha}}_t = \text{softmax}(\tilde{\mathbf{Z}}_t \tilde{\mathbf{w}}^{\alpha,Z}) \quad (7)$$

$$\tilde{\mathbf{z}}_t = \mathbf{V} \tilde{\boldsymbol{\alpha}}_t \quad (8)$$

The procedure resembles the input attention, and we call it **output attention**. It is worth mentioning that the output attention is essentially the same with the spatial visual attention proposed by Lu et al. (2017). However, they did not see it from the input-output point of view nor combine it with the input attention.

The attentive visual output is further transformed to $\mathbf{r}_t = \tanh(\mathbf{W}^{s,z} \tilde{\mathbf{z}}_t)$, $\mathbf{W}^{s,z} \in \mathbb{R}^{e \times g}$, which is of the same dimension as the topic word embedding to simplify the following procedure.

Topic Attention In an image caption, different parts concern different topics. In the existing work (You et al., 2016), the conceptual information is attended based on the previous output word:

$$\boldsymbol{\beta}_t = \text{softmax}(\mathbf{T}^\top \mathbf{U} \mathbf{y}_{t-1}) \quad (9)$$

where $\mathbf{U} \in \mathbb{R}^{e \times e}$, $\boldsymbol{\beta}_t \in \mathbb{R}^m$. The profound issue is that this approach neglects the visual information. It should be beneficial to provide the attentive visual information when selecting topics. The hidden state of the LSTM contains both the information of previous words and the attentive input visual information. Therefore, the model attends to the topics based on the hidden state of the LSTM:

$$\mathbf{Q}_t = \tanh(\mathbf{W}^{Q,T} \mathbf{T} \oplus \mathbf{W}^{Q,h} \mathbf{h}_t) \quad (10)$$

$$\boldsymbol{\beta}_t = \text{softmax}(\mathbf{Q}_t \mathbf{w}^{\beta,Q}) \quad (11)$$

where $\mathbf{W}^{Q,T} \in \mathbb{R}^{m \times e}$, $\mathbf{W}^{Q,h} \in \mathbb{R}^{m \times d}$, $\mathbf{w}^{\beta,Q} \in \mathbb{R}^m$ are the parameters to be learned. $\boldsymbol{\beta}_t \in \mathbb{R}^m$ is the weight of the topics, from which the attentive conceptual output $\mathbf{q}_t \in \mathbb{R}^e$ is calculated:

$$\mathbf{q}_t = \mathbf{T} \boldsymbol{\beta}_t \quad (12)$$

The topic attention \mathbf{q}_t and the hidden state \mathbf{h}_t are combined as the contextual information \mathbf{s}_t :

$$\mathbf{s}_t = \tanh(\mathbf{W}^{s,q} \mathbf{q}_t + \mathbf{W}^{s,h} \mathbf{h}_t) \quad (13)$$

where $\mathbf{W}^{s,q} \in \mathbb{R}^{e \times e}$, $\mathbf{W}^{s,h} \in \mathbb{R}^{e \times d}$ are learnable parameters.

Merging Gate We have prepared both the visual information \mathbf{r}_t and the contextual information \mathbf{s}_t . It is not reasonable to treat the two kinds of information equally when the decoder generates different types of words. For example, when generating descriptive words (e.g., *behind*, *red*), \mathbf{r}_t should matter more than \mathbf{s}_t . However, when generating

object words (e.g., *people*, *table*), \mathbf{s}_t is more important. We introduce a novel score-based merging mechanism to make the model adaptively learn to adjust the balance:

$$\gamma_t = \sigma(S(\mathbf{s}_t) - S(\mathbf{r}_t)) \quad (14)$$

$$\mathbf{c}_t = \gamma_t \mathbf{s}_t + (1 - \gamma_t) \mathbf{r}_t \quad (15)$$

where σ is the sigmoid function, $\gamma_t \in [0, 1]$ indicates how important the topic attention is compared to the visual attention, and S is the scoring function. The scoring function needs to evaluate the importance of the topic attention. Noticing that Eq. (10) and Eq. (11) have a similar purpose, we define S similarly:

$$S(\mathbf{s}_t) = \tanh(\mathbf{W}^{S,h} \mathbf{h}_t + \mathbf{W}^{S,s} \mathbf{s}_t) \cdot \mathbf{w}^S \quad (16)$$

$$S(\mathbf{r}_t) = \tanh(\mathbf{W}^{S,h} \mathbf{h}_t + \mathbf{W}^{S,r} \mathbf{r}_t) \cdot \mathbf{w}^S \quad (17)$$

where \cdot denotes dot product of vectors, $\mathbf{W}^{S,s} \in \mathbb{R}^{m \times e}$, $\mathbf{W}^{S,r} \in \mathbb{R}^{m \times e}$ are the parameters to be learned, and $\mathbf{W}^{S,h}$, \mathbf{w}^s share the weights of $\mathbf{W}^{Q,h}$, $\mathbf{w}^{\beta,Q}$ from Eq. (10) and Eq. (11), respectively.

Finally, the output word is generated by:

$$y_t \sim \mathbf{p}_t = \text{softmax}(\mathbf{W}^{p,c} \mathbf{c}_t) \quad (18)$$

where each value of $\mathbf{p}_t \in \mathbb{R}^{|D|}$ is a probability indicating how likely the corresponding word in vocabulary D is the current output word. The whole model is trained using maximum log likelihood and the loss function is the cross entropy loss.

In all, our proposed approach encourages the model to take advantage of all the available information. The adaptive merging mechanism makes the model weigh the information elaborately.

4 Experiment

We describe the datasets and the metrics used for evaluation, followed by the training details and the evaluation of the proposed approach.

4.1 Datasets and Metrics

There are several datasets containing images and their captions. We report results on the popular Microsoft COCO (Chen et al., 2015) dataset and the Flickr30k (Young et al., 2014) dataset. They contain 123,287 images and 31,000 images, respectively, and each image is annotated with 5 sentences. We report results using the widely-used publicly-available splits in the work of Karpathy and Li (2015). There are 5,000 images each in the validation set and the test set for COCO, 1,000 images for Flickr30k.

We report results using the COCO captioning evaluation toolkit (Chen et al., 2015) that reports the widely-used automatic evaluation metrics SPICE, CIDEr, BLEU, METEOR, and ROUGE. SPICE (Anderson et al., 2016), which is based on scene graph matching, and CIDEr (Vedantam et al., 2015), which is based on n-gram matching, are specifically proposed for evaluating image captioning systems. They both incorporate the consensus of a set of references for an example. BLEU (Papineni et al., 2002) and METOR (Banerjee and Lavie, 2005) are originally proposed for machine translation evaluation. ROUGE (Lin and Hovy, 2003; Lin, 2004) is designed for automatic evaluation of extractive text summarization. In the related studies, it is concluded that SPICE correlates the best with human judgments with a remarkable margin over the other metrics, and is expert in judging detailedness, where the other metrics show negative correlations, surprisingly; CIDEr and METEOR follows with no particular precedence, followed by ROUGE-L, and BLEU-4, in that order (Anderson et al., 2016; Vedantam et al., 2015).

4.2 Settings

Following common practice, the CNN used is the ResNet152 model (He et al., 2016) pre-trained on ImageNet.³ There are 2048 7×7 feature maps, and we project them into 512 feature maps, i.e. g is 512. The word embedding size e is 256 and the hidden size d of the LSTM is 512. We only keep caption words that occur at least 5 times in the training set, resulting in 10,132 words for COCO and 7,544 for Flickr30k. We use the topic extractor pre-trained by Fang et al. (2015) for 1,000 concepts on COCO. We only use 568 manually-annotated object words as topics. For an image, only the top 5 topics are selected, which means m is 5. The same topic extractor is used for Flickr30k, as COCO provides adequate generality. The caption words and the topic words share the same embeddings. In training, we first train the model without visual attention (freezing the CNN parameters) for 20 epochs with the batch size of 80. The learning rate for the LSTM is 0.0004. Then, we switch to jointly train the full model with a learning rate of 0.00001, which exponentially decays with the number of epochs so that it is halved every 50 epochs. We also use momen-

³We use the pre-trained model from `torchvision`.

| Flickr30k | SPICE | CIDEr | METEOR | ROUGE-L | BLEU-4 |
|--|--------------|--------------|--------------|--------------|--------------|
| HardAtt (Xu et al., 2015) | - | - | 0.185 | - | 0.199 |
| SCA-CNN (Chen et al., 2017) | - | - | 0.195 | - | 0.223 |
| ATT-FCN (You et al., 2016) | - | - | 0.189 | - | 0.230 |
| SCN-LSTM (Gan et al., 2017) | - | - | 0.210 | - | 0.257 |
| AdaAtt (Lu et al., 2017) | 0.145 | 0.531 | 0.204 | 0.467 | 0.251 |
| NBT (Lu et al., 2018) | 0.156 | 0.575 | 0.217 | - | 0.271 |
| SR-PL (Liu et al., 2018)* [†] | 0.158 | 0.650 | 0.218 | 0.499 | 0.293 |
| simNet | 0.160 | 0.585 | 0.221 | 0.489 | 0.251 |

Table 1: Performance on the Flickr30k Karpathy test split. The symbol * denotes directly optimizing CIDEr. The symbol [†] denotes using extra data for training, thus not directly comparable. Nonetheless, our model supersedes all existing models in SPICE, which correlates the best with human judgments.

| COCO | SPICE | CIDEr | METEOR | ROUGE-L | BLEU-4 |
|---|--------------|--------------|--------------|--------------|--------------|
| HardAtt (Xu et al., 2015) | - | - | 0.230 | - | 0.250 |
| ATT-FCN (You et al., 2016) | - | - | 0.243 | - | 0.304 |
| SCA-CNN (Chen et al., 2017) | - | 0.952 | 0.250 | 0.531 | 0.311 |
| LSTM-A (Yao et al., 2017) | 0.186 | 1.002 | 0.254 | 0.540 | 0.326 |
| SCN-LSTM (Gan et al., 2017) | - | 1.012 | 0.257 | - | 0.330 |
| Skeleton (Wang et al., 2017) | - | 1.069 | 0.268 | 0.552 | 0.336 |
| AdaAtt (Lu et al., 2017) | 0.195 | 1.085 | 0.266 | 0.549 | 0.332 |
| NBT (Lu et al., 2018) | 0.201 | 1.072 | 0.271 | - | 0.347 |
| DRL (Ren et al., 2017b)* | - | 0.937 | 0.251 | 0.525 | 0.304 |
| TD-M-ATT (Chen et al., 2018)* | - | 1.116 | 0.268 | 0.555 | 0.336 |
| SCST (Rennie et al., 2017)* | - | 1.140 | 0.267 | 0.557 | 0.342 |
| SR-PL (Liu et al., 2018)* [†] | 0.210 | 1.171 | 0.274 | 0.570 | 0.358 |
| Up-Down (Anderson et al., 2018)* [†] | 0.214 | 1.201 | 0.277 | 0.569 | 0.363 |
| simNet | 0.220 | 1.135 | 0.283 | 0.564 | 0.332 |

Table 2: Performance on the COCO Karpathy test split. Symbols, * and [†], are defined similarly. Our model outperforms the current state-of-the-art Up-Down substantially in terms of SPICE.

tum of 0.8 and weight decay of 0.999. We use Adam (Kingma and Ba, 2014) for parameter optimization. For fair comparison, we adopt early stop based on CIDEr within maximum 50 epochs.

4.3 Results

We compare our approach with various representative systems on Flickr30k and COCO, including the recently proposed NBT that is the state-of-the-art on the two datasets in comparable settings. Table 1 shows the result on Flickr30k. As we can see, our model outperforms the comparable systems in terms of all of the metrics except BLEU-4. Moreover, our model overpasses the state-of-the-art with a comfortable margin in terms of SPICE, which is shown to correlate the best with human judgments (Anderson et al., 2016).

Table 2 shows the results on COCO. Among the directly comparable models, our model is arguably the best and outperforms the existing models except in terms of BLEU-4. Most encouragingly, our model is also competitive with Up-Down (Anderson et al., 2018), which uses much larger dataset,

Visual Genome (Krishna et al., 2017), with dense annotations to train the object detector, and directly optimizes CIDEr. Especially, our model outperforms the state-of-the-art substantially in SPICE and METEOR. Breakdown of SPICE F-scores over various subcategories (see Table 3) shows that our model is in dominant lead in almost all subcategories. It proves the effectiveness of our approach and indicates that our model is quite data efficient.

For the methods that directly optimize CIDEr, it is intuitive that CIDEr can improve significantly. The similar improvement of BLEU-4 is evidence that optimizing CIDEr leads to more n-gram matching. However, it comes to our notice that the improvements of SPICE, METEOR, and ROUGE-L are far less significant, which suggests there may be a gaming situation where the n-gram matching is wrongfully exploited by the model in reinforcement learning. As shown by Liu et al. (2017), it is most reasonable to jointly optimize

| Methods | SPICE | | | | | | | CIDEr | METEOR | ROUGE-L | BLEU-4 |
|--|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | All | Objects | Attributes | Relations | Color | Count | Size | | | | |
| Baseline (Plain Encoder-Decoder Network) | 0.150 | 0.295 | 0.048 | 0.039 | 0.022 | 0.004 | 0.023 | 0.762 | 0.220 | 0.495 | 0.251 |
| Up-Down (Anderson et al., 2018) ^{*†} | 0.214 | 0.391 | 0.100 | 0.065 | 0.114 | 0.184 | 0.032 | 1.201 | 0.277 | 0.569 | 0.363 |
| Baseline + Input Att. | 0.164 | 0.316 | 0.060 | 0.044 | 0.030 | 0.038 | 0.024 | 0.840 | 0.233 | 0.512 | 0.273 |
| Baseline + Output Att. | 0.181 | 0.329 | 0.094 | 0.053 | 0.089 | 0.184 | 0.044 | 0.968 | 0.253 | 0.534 | 0.301 |
| Baseline + Input Att. + Output Att. | 0.187 | 0.338 | 0.101 | 0.055 | 0.115 | 0.161 | 0.048 | 1.038 | 0.259 | 0.542 | 0.311 |
| Baseline + Topic Att. | 0.184 | 0.348 | 0.074 | 0.051 | 0.047 | 0.064 | 0.037 | 0.915 | 0.250 | 0.517 | 0.260 |
| Baseline + Topic Att. + MGate | 0.189 | 0.355 | 0.080 | 0.051 | 0.055 | 0.090 | 0.033 | 0.959 | 0.256 | 0.527 | 0.281 |
| Baseline + Input Att. + Output Att. + Topic Att. | 0.206 | 0.381 | 0.091 | 0.060 | 0.075 | 0.094 | 0.045 | 1.068 | 0.273 | 0.556 | 0.320 |
| simNet (Full Model) | 0.220 | 0.394 | 0.109 | 0.070 | 0.088 | 0.202 | 0.045 | 1.135 | 0.283 | 0.564 | 0.332 |

Table 3: Results of incremental analysis. For a better understanding of the differences, we further list the breakdown of SPICE F-scores. *Objects* indicates comprehensiveness, and the others indicate detailedness. Additionally, we report the performance of the current state-of-the-art Up-Down for further comparison, which uses extra dense-annotated data for pre-training and directly optimizes CIDEr.

| Method | Precision | Recall | F1 |
|----------------------|--------------|--------------|--------------|
| Topics ($m=5$) | 49.95 | 38.91 | 42.48 |
| All words ($m=5$) | 84.01 | 17.99 | 29.49 |
| All words ($m=10$) | 70.90 | 30.18 | 42.05 |
| All words ($m=20$) | 52.51 | 44.53 | 47.80 |

Table 4: Performance of visual word extraction.

| Method | S | C | M | R | B |
|----------------------|--------------|--------------|--------------|--------------|--------------|
| Topics ($m=5$) | 0.220 | 1.135 | 0.283 | 0.564 | 0.332 |
| All words ($m=5$) | 0.197 | 1.047 | 0.264 | 0.550 | 0.314 |
| All words ($m=10$) | 0.201 | 1.076 | 0.256 | 0.528 | 0.293 |
| All words ($m=20$) | 0.209 | 1.117 | 0.276 | 0.561 | 0.329 |

Table 5: Effect of using different visual words.

all the metrics at the same time.

We also evaluate the proposed model on the COCO evaluation server, the results of which are shown in Appendix A.1, due to limited space.

5 Analysis

In this section, we analyze the contribution of each component in the proposed approach, and give examples to show the strength and the potential improvements of the model. The analysis is conducted on the test set of COCO.

Topic Extraction The motivation of using objects as topics is that they are easier to identify so that the generation suffers less from erroneous predictions. This can be proved by the F-score of the identified topics in the test set, which is shown in Table 4. Using top-5 object words is at least as good as using top-10 all words. However, using top-10 all words introduces more erroneous visual words to the generation. As shown in Ta-

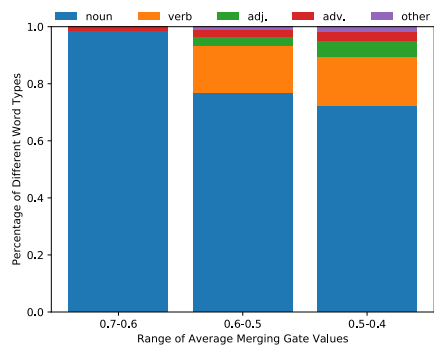







Figure 4: Average merging gate values according to word types. As we can see, object words (noun) dominate the high value range, while attribute and relation words are assigned lower values, indicating the merging gate learns to efficiently combine the information.

ble 5, when extracting all words, providing more words to the model indeed increases the captioning performance. However, even when top-20 all words are used, the performance is still far behind using only top-5 object words and seems to reach the performance ceiling. It proves that for semantic attention, it is also important to limit the absolute number of incorrect visual words instead of merely the precision or the recall. It is also interesting to check whether using other kind of words can reach the same effect. Unfortunately, in our experiments, only using verbs or adjectives as semantic concepts works poorly.

To examine the contributions of the submodules in our model, we conduct a series of experiments. The results are summarized in Table 3. To help with the understanding of the differences, we also report the breakdown of SPICE F-scores.

Visual Attention Our input attention achieves similar results to previous work (Xu et al., 2015),

| Comparison of Models |  |  |  |  |  |
|-------------------------|---|---|---|---|---|
| Topics | woman girl baby bear kitchen | computer keyboard laptop mouse desk | buildings bus clock tower street | pizza cheese table plate toppings | motorcycle street car bike motorcycles |
| Visual Attention | a girl and a baby are holding a stuffed animal | a computer keyboard sitting on top of a wooden desk | two green buses is parked on the side of the road | two pizzas with toppings on a table | a row of motorcycles parked next to each other |
| Topic Attention | a woman holding a teddy bear in a kitchen | a computer keyboard and a mouse sitting on a desk | a large double decker bus is parked in front of a building | a pizza with a lot of toppings on it | a motorcycle parked in a parking lot next to a car |
| simNet | a woman and a baby are holding a stuffed animal | a computer keyboard and mouse on a wooden desk | two green double decker buses parked in front of a large building | two pizzas sitting on a table with two different kinds of toppings | a row of motorcycles parked in a street |




| Error Analysis |  |  |  |
|-------------------|---|---|---|
| Topics | clock tower building street city | people bus truck street train | garden bench park forest plants |
| Reference | a tall building that has a clock on it (near a large building) | tour buses driving down a street lined with cheering people | an old wooden bench in nature surrounded by plants |
| simNet | a large building with a clock tower in the background | a group of people standing around a parked bus at a bus stop | a wooden bench sitting in the middle of a lush green garden |
| Error Type | distance | movement | object |

Figure 5: Examples of the generated captions. The left plot compares simNet with visual attention and topic attention. Visual attention is good at portraying the relations but is less specific in objects. Topic attention includes more objects but lacks details, such as material, color, and number. The proposed model achieves a very good balance. The right plot shows the error analysis of the proposed simNet.

if not better. Using only the output attention is much more effective than using only the input attention, with substantial improvements in all metrics, showing the impact of information gap caused by delayed input in attention. Combining the input attention and the output attention can further improve the results, especially in color and size descriptions.

Topic Attention As expected, compared with visual attention, the topic attention is better at identifying objects but worse at identifying attributes. We also apply the merging gate to the topic attention, but it now merges q_t and h_t instead of s_t and r_t . With the merging gate, the model can balance the information in caption text and extracted topics, resulting in better overall scores. While it overpasses the conventional visual attention, it lags behind the output attention.

Merging Gate Combing the visual attention and the topic attention directly indeed results in a huge boost in performance, which confirms our motivation. However, directly combining them also causes lower scores in attributes, color, count, and size, showing that the advantages are not fully made use of. The most dramatic improvements come from applying the merging gate to the combined attention, showing that the proposed balance mechanism can adaptively combine the two kinds of information and is essential to the overall performance. The average merging gate value summarized in Figure 4 suggests the same.

We give some examples in the left plot of Figure 5 to illustrate the differences between the models more intuitively. From the examples, it is clear that the proposed simNet generates the best captions in that more objects are described and many informative and detailed attributes are included, such as the quantity and the color.

Visualization Figure 6 shows the visualization of the topic attention and the visual attention with running examples. As we can see, the topic attention is active when generating a phrase containing the related topic. For example, *bathroom* is always most attended when generating *a bathroom*. The merging gate learns to direct the information flow efficiently. When generating words such as *on* and *a*, it gives lower weight to the topic attention and prefers the visual attention. As to the visual attention, the output attention is much more focused than the input attention. As we hypothesized, the conventional input attention lacks the information of the last generated word and does not know what to look for exactly. For example, when generating *bathroom*, the input attention does not know the previous generated word is *a*, and it loses its focus, while the output attention is relatively more concentrated. Moreover, the merging gate learns to overcome the erroneous topics, as shown in the second example. When generating *chair*, the topic attention is focused on a wrong object *bed*, while the visual attention attends correctly to the chair, and especially the output attention attends to the armrest. The merging gate effectively remedies

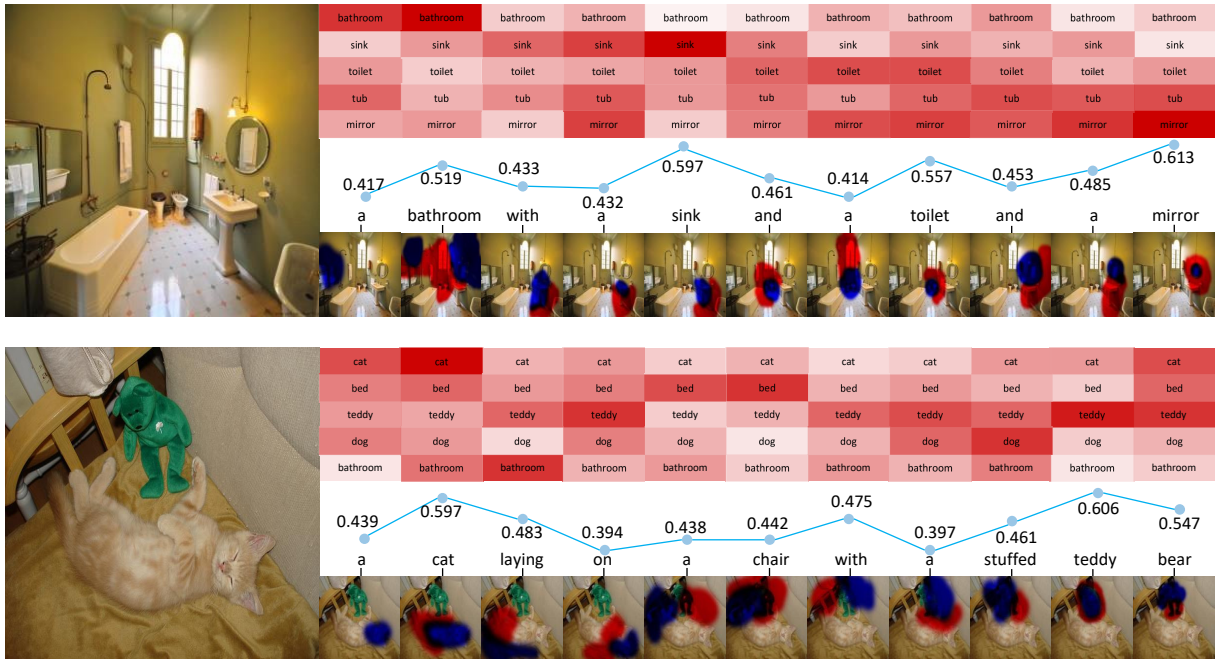


Figure 6: Visualization. Please view in color. Here, we give two running examples. The upper part of each example shows the attention weights of each of 5 extracted topics. Deeper color means larger in value. The middle part shows the value of the merging gate that determines the importance of the topic attention. The lower part shows the visualization of visual attention. The attended region is covered with color. The blue shade indicates the output attention. The red shade indicates the input attention.

the misleading information from the topic attention and outputs a lower weight, resulting in the model correctly generating the word *chair*.

Error Analysis We conduct error analysis using the proposed (full) model on the test set to provide insights on how the model may be improved. We find 123 out of 1000 generated captions that are not satisfactory. There are mainly three types of errors, i.e. distance (32, 26%), movement (22, 18%), and object (60, 49%), with 9 (7%) other errors. Distance error takes place when there is a lot of objects and the model cannot grasp the foreground and the background relationship. Movement error means that the model fails to describe whether the objects are moving. Those two kinds of errors are hard to eliminate, as they are fundamental problems of computer vision waiting to be resolved. Object error happens when there are incorrect extracted topics, and the merging gate regards the topic as grounded in the image. In the given example, the incorrect topic is *garden*. The tricky part is that the topic is seemingly correct according to the image features or otherwise the proposed model will choose other topics. A more powerful topic extractor may help with the problem but it is unlikely to be completely avoided.

6 Conclusions

We propose the stepwise image-topic merging network to sequentially and adaptively merge the visual and the conceptual information for improved image captioning. To our knowledge, we are the first to combine the visual and the semantic attention to achieve substantial improvements. We introduce the stepwise merging mechanism to efficiently guide the two kinds of information when generating the caption. The experimental results demonstrate the effectiveness of the proposed approach, which substantially outperforms the state-of-the-art image captioning methods in terms of SPICE on COCO and Flickr30k datasets. Quantitative and qualitative analysis show that the generated captions are both detailed and comprehensive in comparison with the existing methods.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (No. 61673028). We thank all the anonymous reviewers for their constructive comments and suggestions. Xu Sun is the corresponding author of this paper.

References

- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and VQA. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, June 29, 2005*, pages 65–72. Association for Computational Linguistics.
- Hui Chen, Guiguang Ding, Sicheng Zhao, and Jungong Han. 2018. Temporal-difference learning with sampling baseline for image captioning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press.
- Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6298–6306. IEEE Computer Society.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.
- Xinlei Chen and C. Lawrence Zitnick. 2015. Mind’s eye: A recurrent visual representation for image caption generation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2422–2431. IEEE Computer Society.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.
- Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1473–1482. IEEE Computer Society.
- Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. 2010. 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*, volume 6314 of *Lecture Notes in Computer Science*, pages 15–29. Springer.
- Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1141–1150. IEEE Computer Society.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. 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. IEEE Computer Society.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society.
- Andrej Karpathy and Fei-Fei Li. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

- Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. BabyTalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 35(12):2891–2903.
- Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain, July, 2004*, pages 74–81. Association for Computational Linguistics.
- Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics.
- Junyang Lin, Xu Sun, Xuancheng Ren, Shuming Ma, Jinsong Su, and Qi Su. 2018. Deconvolution-based global decoding for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3260–3271. Association for Computational Linguistics.
- Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 873–881. IEEE Computer Society.
- Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. *CoRR*, abs/1803.08314.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3242–3250. IEEE Computer Society.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*.
- Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4251–4257. ijcai.org.
- Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-RNN). *CoRR*, abs/1412.6632.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318. ACL.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017a. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.
- Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017b. Deep reinforcement learning-based image captioning with embedding reward. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1151–1159. IEEE Computer Society.
- Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDER: consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.
- Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence - video to text. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4534–4542. IEEE Computer Society.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society.

- Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W. Cottrell. 2017. 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. IEEE Computer Society.
- Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. 2016. 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. IEEE Computer Society.
- Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, pages 1180–1192. ACM.
- Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018a. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 979–988. Association for Computational Linguistics.
- Jingjing Xu, Yi Zhang, Qi Zeng, Xuancheng Ren, Xiaoyan Cai, and Xu Sun. 2018b. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, Brussels, Belgium, October 31-November 4, 2018*. Association for Computational Linguistics.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.
- Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4904–4912. IEEE Computer Society.
- Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4651–4659. IEEE Computer Society.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Cha Zhang, John C. Platt, and Paul A. Viola. 2006. Multiple instance boosting for object detection. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada*, pages 1417–1424. MIT Press.
- Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. 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. IEEE Computer Society.
- Zhou Zhao, Hanqing Lu, Deng Cai, Xiaofei He, and Yueting Zhuang. 2016. Partial multi-modal sparse coding via adaptive similarity structure regularization. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 152–156. ACM.

| COCO | BLEU-1 | | BLEU-2 | | BLEU-3 | | BLEU-4 | | METEOR | | ROUGE-L | | CIDEr | |
|--|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | c5 | c40 | c5 | c40 | c5 | c40 | c5 | c40 | c5 | c40 | c5 | c40 | c5 | c40 |
| HardAtt (Xu et al., 2015) | 0.705 | 0.881 | 0.528 | 0.779 | 0.383 | 0.658 | 0.277 | 0.537 | 0.241 | 0.322 | 0.516 | 0.654 | 0.865 | 0.893 |
| ATT-FCN (You et al., 2016) | 0.731 | 0.900 | 0.565 | 0.815 | 0.424 | 0.709 | 0.316 | 0.599 | 0.250 | 0.335 | 0.535 | 0.682 | 0.943 | 0.958 |
| SCA-CNN (Chen et al., 2017) | 0.712 | 0.894 | 0.542 | 0.802 | 0.404 | 0.691 | 0.302 | 0.579 | 0.244 | 0.331 | 0.524 | 0.674 | 0.912 | 0.921 |
| LSTM-A (Yao et al., 2017) | 0.739 | 0.919 | 0.575 | 0.842 | 0.436 | 0.740 | 0.330 | 0.632 | 0.256 | 0.350 | 0.542 | 0.700 | 0.984 | 1.003 |
| SCN-LSTM (Gan et al., 2017) | 0.740 | 0.917 | 0.575 | 0.839 | 0.436 | 0.739 | 0.331 | 0.631 | 0.257 | 0.348 | 0.543 | 0.696 | 1.003 | 1.013 |
| AdaAtt (Lu et al., 2017) [†] | 0.748 | 0.920 | 0.584 | 0.845 | 0.444 | 0.744 | 0.336 | 0.637 | 0.264 | 0.359 | 0.550 | 0.705 | 1.042 | 1.059 |
| TD-M-ATT (Chen et al., 2018) ^{*†} | 0.757 | 0.913 | 0.591 | 0.836 | 0.441 | 0.726 | 0.324 | 0.609 | 0.259 | 0.342 | 0.547 | 0.689 | 1.059 | 1.090 |
| SCST (Rennie et al., 2017) ^{*†} | 0.781 | 0.937 | 0.619 | 0.860 | 0.470 | 0.759 | 0.352 | 0.645 | 0.270 | 0.355 | 0.563 | 0.707 | 1.147 | 1.167 |
| Up-Down (Anderson et al., 2018) ^{*†‡} | 0.802 | 0.952 | 0.641 | 0.888 | 0.491 | 0.794 | 0.369 | 0.685 | 0.276 | 0.367 | 0.571 | 0.724 | 1.179 | 1.205 |
| simNet | 0.766 | 0.941 | 0.605 | 0.874 | 0.462 | 0.778 | 0.350 | 0.671 | 0.267 | 0.362 | 0.558 | 0.716 | 1.087 | 1.111 |

Table 6: Performance on the online COCO evaluation server. The SPICE metric is unavailable for our model, thus not reported. c5 means evaluating against 5 references, and c40 means evaluating against 40 references. The symbol * denotes directly optimizing CIDEr. The symbol [†] denotes model ensemble. The symbol [‡] denotes using extra data for training, thus not directly comparable. Our submission does not use the three aforementioned techniques. Nonetheless, our model is second only to Up-Down and surpasses almost all the other models in published work, especially when 40 references are considered.

A Supplementary Material

A.1 Results on COCO Evaluation Server

Table 6 shows the performance on the online COCO evaluation server⁴. We put it in the appendix because the results are incomplete and the SPICE metric is not available for our submission, which correlates the best with human evaluation. The SPICE metrics are only available at the leaderboard on the COCO dataset website⁵, which, unfortunately, has not been updated for more than a year. Our submission does not directly optimize CIDEr, use model ensemble, or use extra training data. The three techniques typically result in orthogonal improvements (Lu et al., 2017; Rennie et al., 2017; Anderson et al., 2018). Moreover, the SPICE results are missing, in which the proposed model has the most advantage. Nonetheless, our model is second only to Up-Down (Anderson et al., 2018) and surpasses almost all the other models in published work, especially when 40 references are considered.

⁴<https://competitions.codalab.org/competitions/3221>

⁵<http://cocodataset.org/#captions-leaderboard>