Watch What You Just Said: Image Captioning with Text-Conditional Attention*

Luowei Zhou
Robotics Institute
University of Michigan
luozhou@umich.edu

Chenliang Xu
Department of Computer Science
University of Rochester
chenliang.xu@rochester.edu

Parker Koch
Electrical Engineering and Computer Science
University of Michigan
pakoch@umich.edu

Jason J. Corso
Electrical Engineering and Computer Science
University of Michigan
jjcorso@eecs.umich.edu

ABSTRACT

Attention mechanisms have attracted considerable interest in image captioning due to their powerful performance. However, existing methods use only visual content as attention and whether textual context can improve attention in image captioning remains unsolved. To explore this problem, we propose a novel attention mechanism, called *text-conditional attention*, which allows the caption generator to focus on certain image features given previously generated text. To obtain text-related image features for our attention model, we adopt the guiding Long Short-Term Memory (gLSTM) captioning architecture with CNN fine-tuning. Our proposed method allows joint learning of the image embedding, text embedding, text-conditional attention and language model with one network architecture in an end-to-end manner. We perform extensive experiments on the MS-COCO dataset. The experimental results show that our method outperforms state-of-the-art captioning methods on various quantitative metrics as well as in human evaluation, which supports the use of our text-conditional attention in image captioning.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; **Neural networks**; *Computer vision representations*;

KEYWORDS

image captioning; multi-modal embedding; LSTM, Neural Network

*Produces the permission block, and copyright information

ThematicWorkshops'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-5416-5/17/10...$15.00
DOI: https://doi.org/10.1145/3126686.3126717

1 INTRODUCTION

Image captioning is drawing increasing interest in computer vision and multimedia [19, 22, 25, 34, 38]. Basically, it requires machines to automatically describe the content of an image using an English sentence. While this task seems obvious for human-beings, it is complicated for machines since it requires the language model to capture various semantic information within an image, such as objects' motions and actions. Another challenge for image captioning, especially for generative models, is that the generated output should be human-like natural sentences.

Recent literature in image captioning is dominated by neural network-based methods [6, 31, 33, 38]. The idea originates from the encoder-decoder architecture in Neural Machine Translation [2], where a Convolutional Neural Network (CNN) is adopted to encode the input image into a feature vector, and a sequence modeling approach (e.g., Long Short-Term Memory (LSTM) [13]) decodes the feature vector into a sequence of words [31]. Most recent work in image captioning relies on this structure, and leverages image guidance [14], attributes [38] or region attention [35] as the extra input to LSTM decoder for better performance. The intuition comes from visual attention, which has been known in Psychology and Neuroscience for a long time [5]. For image captioning, this means the image guidance to the language model should change over time according to the context.

However, these methods using attention lack consideration from the following two aspects. First, attending to the image is only half of the story; watching what you just said comprises the other half. In other words, **visual evidence can be inferred and interpreted by textual context**, especially when the visual evidence is ambiguous. For example, in the sentence "After dinner, the man is comfortably lying on the sofa and watching TV", the objects "sofa" and "TV" are naturally inferred even with weak visual evidences (see Fig. 1). Despite its importance, textual context was not a topic of focus in attention models. Existing attention based methods such as [34, 35, 38] have used implicit text-guiding from an LSTM hidden layer to determine which of the image regions or attributes to attend on. However, as we mentioned in the previous example, the object for attention might be only partially observable, so the attention input could be

"After dinner, the man is comfortably lying on the sofa and watching TV."

Figure 1: We can barely see the *sofa* from the image on the left, and we can only see a corner of the *TV* in the image on the right, but we can infer them from the textual context even with the weak visual evidences. Image credits: left, right.

misleading. This is not the case for our attention model since the textual features are tightly coupled with the image features to compensate for one another. Another work is Jia et al. [14], where they use joint embedding of the text and image as the guidance for the LSTM. However, their approach has pre-specified guidance that is fixed over time and has a linear form. In contrast, our method systematically incorporates the time-dependent text-conditional attention, from 1-gram to n-gram and even to the sentence level.

Second, existing attention based methods separate CNN feature learning (trained for a different task, i.e. image classification) from the LSTM text generation. This leads to a representational disconnection between features learned and text generated. For instance, the attention model proposed by You et al. [38] uses weighted-sum visual attributes to guide the image captioning, while the attributes proposed by the specific predictor are separated from the language model. This makes the attributes guidance lack the ability to adapt to the textual context, which ultimately compromises the end-to-end learning ability the paper claimed.

To overcome the above limitations, we propose a new text-conditional attention model based on the time-dependent guiding LSTM. Our model has the ability to interpret image features based on textual context and is end-to-end trainable. The model learns a text-conditional embedding matrix between CNN image features and previously generated text. Given a target image, the proposed model generates LSTM guidance by directly conditioning the image features on the current textual context. The model hence learns how to interpret image features given the textual content it has recently generated. If it conditions the image features on one previous word, it is a 1-gram word-conditional model. If it is on previous two words, we get a 2-gram word-conditional model. Similarly we can construct an n-gram word-conditional model. The extreme version of our text-conditional model is the sentence-conditional model, which takes advantage of all the previously generated words.

We implement our model [1] based on NeuralTalk2, an open-source implementation of Google NIC [31]. We compare our methods with state-of-the-art methods on the commonly used MS-COCO dataset [23] with publicly available splits [16] of training, validation and testing sets. We evaluate methods on standard metrics as well as human evaluation. Our proposed methods outperform the state-of-the-art approaches across different evaluation metrics and yield reasonable attention outputs.

The main contributions of our paper are as follows. First, we propose text-conditional attention which allows the language model to learn text-specified semantic guidance automatically. The proposed attention model learns how to focus on parts of the image feature given the textual content it has generated. Second, the proposed method demonstrates a less complicated way to achieve end-to-end training of attention-based captioning model, whereas state-of-the-art methods [14, 35, 38] involve LSTM hidden states or image attributes for attention, which compromises the possibility of end-to-end optimization.

2 RELATED WORK

Recent successes of deep neural networks in machine translation [3, 29] catalyze the adoption of neural networks in solving image captioning problems. Early works of neural network-based image captioning include the multimodal RNN [17] and LSTM [31]. In these methods, neural networks are used for both image-text embedding and sentence generating. Various methods have shown to improve performance with region-level information [8, 15], external knowledge [1], and even from question-answering [24]. Our method differs from them by considering attention from textual context in caption generating.

Attention mechanism has recently attracted considerable interest in LSTM-based image captioning [34, 35, 37, 38]. Xu et al. [35] propose a model that integrates visual attention through the hidden state of LSTM model. You et al. [38] and Wu et al. [34] tackle the semantic attention problem by fusing visual attributes extracted from images with the input or the output of LSTM. Even though these approaches achieve state-of-the-art performance, the performances rely heavily upon the quality of the pre-specified visual attributes, i.e., better attributes usually lead to better results. Our method also uses attention mechanism, but we consider the explicit time-dependent text attention and is comprised a clean architecture for the ease of end-to-end learning.

Early works in image captioning focus on either template-based methods or transfer-based methods. Template-based methods [7, 11, 19, 21, 26, 36] specify templates and fill them with detected visual evidences from target images. In Kulkarni et al. [19], visual detections are first put into a graphical model with higher order potentials from text corpora to reduce noise, then converted to language descriptions based on pre-specified templates. In Yang et al. [36], a quadruplet consisting of noun, verb, scene and preposition is used to

[1] Source code: https://github.com/LuoweiZhou/e2e-gLSTM-sc

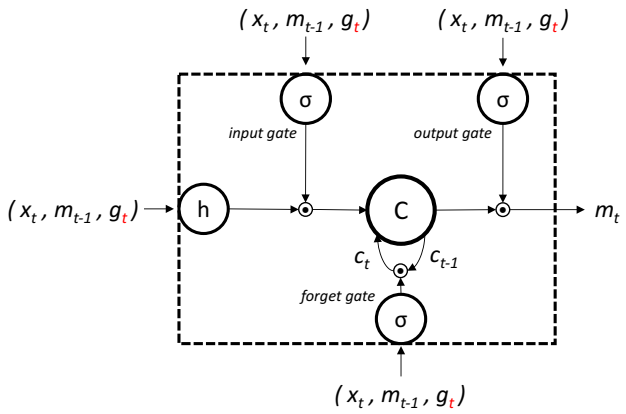

Figure 2: Shown in black is the gLSTM node [14]. The red subscript of g_t represents our td-gLSTM guidance.

describe an image. The drawback of these methods is that the descriptions are not vivid and human-crafted templates do not work for all images. Transfer-based methods [4, 9, 20] rely on image retrieval to assign the target image with descriptions of similar images in the training set. A common issue is that they are less robustness to unseen images.

2.1 Background

The generated sentences by the LSTM model may lose track of the original image content since it only accesses the image content once at the beginning of the learning process, and forgets the image after even a short period of time. Therefore, Jia et al. [14] propose an extension of the LSTM model, named the guiding LSTM (gLSTM), which extracts semantic information from the target image and feeds it into the LSTM model every time step as extra information. The basic gLSTM unit is shown in Fig. 2. Its memory cell and gates are defined as follows:

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{iq}g) \\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fq}g) \\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oq}g) \\
c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1} + W_{cq}g) \\
m_t &= o_t \odot c_t \quad , \qquad (1)
\end{aligned}
$$

where Ws denote weights, $\odot$ represents element-wise multiplication, $\sigma(\cdot)$ is the sigmoid function, $h(\cdot)$ is the hyperbolic tangent function, x_t stands for input, i_t for the input gate, f_t for the forget gate, o_t for the output gate, c_t for state of the memory cell, m_t for the hidden state (also output for one-layer LSTM), and g represents guidance information, which is time-invariant. The subscripts denote time: t is the current time step and $t-1$ is the previous time step. Note that here we omit bias terms in Eq. 1 and Eq. 2 within the embeddings.

3 METHODS

Our text-conditional attention model is based on a time-dependent gLSTM (td-gLSTM). We first describe the td-gLSTM in Sec. 3.1 and show how to obtain semantic guidance through this structure. Then, we introduce our text-conditional attention model and its variants, e.g. n-gram word- and sentence-conditional models, in Sec. 3.2.

3.1 Time-Dependent gLSTM (td-gLSTM)

The gLSTM described in Sec. 2.1 has a time-invariant guidance. In Jia et al. [14], they show three ways of using such guidance, including an embedding of the joint image-text feature by linear CCA. However, the textual context in a sentence is constantly changing while the caption generator is generating the sentence. Obviously, we need the guidance to evolve over time, and hence we propose td-gLSTM. Notice that, despite its simple change in structure, the td-gLSTM is much more flexible in the way it incorporates guidance, e.g. a time-series dynamic guidance such as tracking and actions in a video. Also, notice that the gLSTM is a special case of the td-gLSTM, when the guidance is set as $g_t = g_{t-1}$.

Our proposed td-gLSTM consists of three parts: 1) image embedding; 2) text embedding; and 3) LSTM language model. Figure 3 shows an overview for using td-gLSTM for captioning. First, image feature vector I is extracted using CNN and each word in the caption is represented by a one-hot vector S_t, where t indicates the index of the word in the sentence. We use the text embedding matrix W_e to embed text feature S_t into a latent space, which is the input x_t of the LSTM language model. The text embedding matrix is initialized from a zero-mean Gaussian distribution with standard deviation 0.01. On the other hand, the text feature is jointly embedded with the image feature, denoted as $g_t = h(S_t, I)$, where g_t is the time-dependent guidance. Here, we do not specify the particular form of g_t to make the framework general, and its choices are discussed in Sec. 3.2.

Both the guidance g_t and embedded text features $x_t = W_e S_t$ are used as the inputs to td-gLSTM, which are shown in Fig. 2 (including red) and formulated as follows:

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{iq}g_t) \\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fq}g_t) \\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oq}g_t) \\
c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1} + W_{cq}g_t) \\
m_t &= o_t \odot c_t \quad . \qquad (2)
\end{aligned}
$$

We back-propagate error through guidance g_t for fine-tuning the CNN. One significant benefit of this is that the model allows the guidance information to be more similar to its corresponding text description. Note that the text-conditional guidance g_t keeps changing in each time step, which is a time-dependent variable. The outputs of the language model are the log likelihood of each word from the

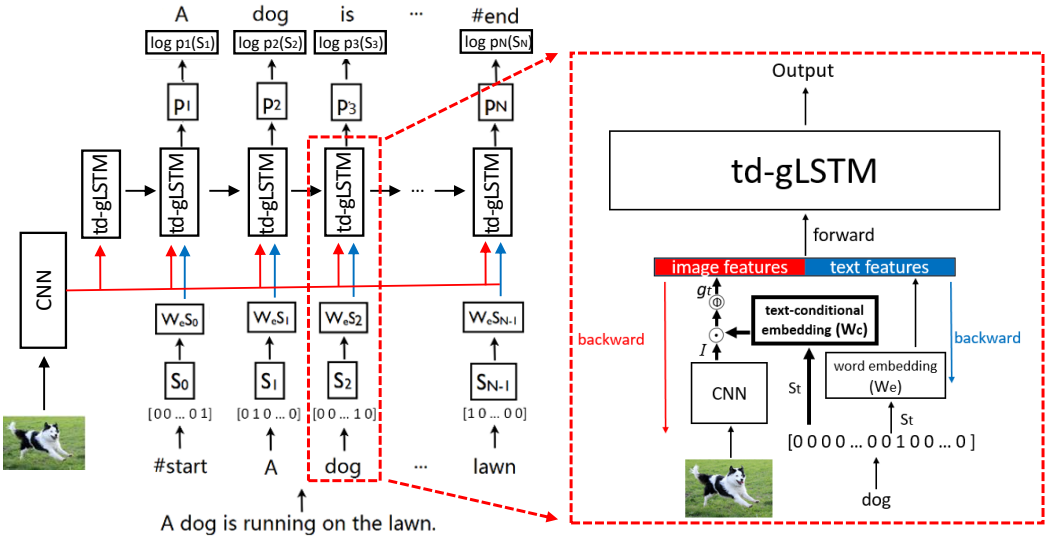

Figure 3: Text-conditional attention. The text-conditional attention part are highlighted in bold. S_t indicates the one-hot vector representation of the t^{th} word in the sentence. W_c is word embedding matrix, W_c is text-conditional embedding matrix, I is image feature and g_t is time-dependent guidance. See text for more details.

target sentence, followed by a Softmax function for normalization. We use the regularized cross-entropy loss function:

$$L(\mathcal{I}, \mathcal{S}) = -\sum_{k=1}^{N} \log(p_k(S_k)) + \frac{\lambda}{2}||W_{conv}||_2^2 \ , \qquad (3)$$

where $\mathcal{I}$ represents the image, $\mathcal{S}$ represents the sentence, S_k denotes the k^{th} word in the sentence, S_N is the stop sign, W_{conv} denotes all the weights in the convolutional net and λ controls the importance of the regularization term. Finally, we back-propagate the loss to LSTM language model, the text embedding matrix and the image embedding CNN. The training detail is described in Sec. 4.1.

3.2 Text-Conditional Semantic Attention

Recently, You et al. [38] use visual attributes as the semantic attention to guide the image captioning. Their semantic guidance consists of top visual attributes of the input image, and the weight of each attribute is determined by the current word, which is the previous output of RNN. However, the attribute predictor adopted in their model has no learning ability and is separated from the encoder-decoder language model. In contrast, following the td-gLSTM model (see Sec. 3.1), we condition the guidance information g_t on the current word S_t (the one-hot vector representation), and use the text-conditional image feature as the semantic guidance. The benefits are twofold: first, the model can learn which part of the semantic image feature should be focused on when seeing a specific word; second, this structure is end-to-end tunable such that the weights of CNNs are tuned for captioning rather than for image classification [27]. For instance, when the caption generator is generating a sequence as "a woman is washing", its attention on the image feature should be automatically switched to objects that can be washed, such as clothes and dishes.

We first consider modeling the text-conditional guidance feature g_t as the weighted-sum of the outer product of image feature I and text feature S_t, therefore each entry in g_t is represented as:

$$g_t^i = \sum_{j,k} W_{ijk} I^j S_t^k + b^i \ , \qquad (4)$$

where I^j denotes the j^{th} entry of the image feature, S_t^k denotes the k^{th} entry of the text feature, and g_t^i is the i^{th} entry of the text-conditional guidance feature. For each g_t^i, the corresponding weights W_i is a 2-D tensor, hence, the total weights W for g_t is a 3-D tensor. In this model, image feature is fully coupled with text feature though the 3-D tensor.

Despite Eq. 4 fully couples the two types of features, it results in a huge amount of parameters, which prohibits its use in practice. To overcome it, we introduce an embedding matrix W_c, which contains various text-to-image masks. Furthermore, in practice, adding one non-linear transfer function layer after the image-text feature embedding boosts the performance. Therefore, we model the text-conditional feature g_t as a text-based mask on image feature I followed by a non-linear function:

$$g_t = \Phi(I \odot W_c S_t) \ , \qquad (5)$$

where W_c is the text-conditional embedding matrix and $\Phi(\cdot)$ is a non-linear transfer function. When W_c is a all-one matrix, the conditioned feature $I \odot W_c S_t$ is identical to I. We transfer

the pre-trained model from gLSTM to initialize the CNN, language model and word embedding of our attention model. For text-conditional matrix, we initialize it with all ones. We show the sensitivity of our model to various transfer functions in Sec. 4.2.

The above model is the 1-gram word-conditional semantic attention owing to the guidance feature is merely conditioned on the previous word. Similarly, we develop the 2-gram word-conditional model, which utilizes previous two words, or even n-gram word-conditional model. The extreme version of the text-conditional model is the sentence-condition model, which takes advantage of all the previously generated words:

$$g_t = \Phi(I \odot W_c \sum_{k=1}^{t} \frac{S_{k-1}}{t}) \ . \tag{6}$$

One benefit of the text-conditional model is that it allows the language model to learn semantic attention automatically though the back-propagation of the training loss while attribute-based method, such as [38], represents semantic guidance by some major components of an image, but other semantic information, such as objects' motions and locations, are discarded.

4 EXPERIMENTS

We describe our experiment settings in Sec. 4.1, analyze the variants of our model and attention in Sec. 4.2, and compare our method with state-of-the-art methods in Sec. 4.3.

4.1 Experiment Setup

We use the MS-COCO dataset [23] with the commonly adopted splits as described in [16]: 113,287 images for training, 5,000 images for validation and 5,000 images for testing. Three standard evaluation metrics, e.g. BLEU, METEOR and CIDER, are used in addition to human evaluation. We implement our model based on the NeuralTalk2 [16], which is an open source implementation of [31]. We use three different CNNs in our experiments, e.g. 34-layer and 200-layer ResNets [12] and 16-layer VGGNet [28]. For a fair comparison, we use 34-layer ResNet when analyzing the variants of our models in Table 1 and 2, 16-layer VGGNet when comparing to state-of-the-art methods in Table 5 and 6, and 200-layer ResNet for leaderboard competition in Table 7. The variation of performance regarding different CNNs are also evaluated in Table 3.

We train our model in three steps: 1) train time-invariant gLSTM (ti-gLSTM) without CNN fine-tuning for 100,000 iterations; 2) train ti-gLSTM with CNN fine-tuning for 150,000 iterations; and 3) train td-gLSTM with initialized text-conditional matrix but without CNN fine-tuning for 150,000 iterations. The reason for this multiple-step training is described in Vinyals et al. [32]: jointly training the system at the initial time causes noise in the initial gradients coming from LSTM that corrupts the CNN unrecoverably. For the hyper-parameters, we set the CNN weight decay rate (λ in Eq. 3) to 10^{-3} to avoid overfitting. The learning rate for CNN fine-tuning is set to 10^{-5} and the learning rate for language model is set to 4×10^{-4}. We use Adam optimizer [18] for updating weights with $\alpha = 0.8$ and $\beta = 0.999$. We adopt 2 and 3 for beam sizes during inference, as recommended by recent studies [6, 32]. The whole training process takes about one day on a single NVIDIA TITAN X GPU.

Table 1: Results of n-gram word- and sentence-conditional models with Tanh transfer function and 34-layer ResNet. Top-2 scores for each metric are highlighted. All values are reported as percentage (%).

	BLEU@4	METEOR	CIDEr
1-gram	29.5	24.6	94.6
2-gram	30.2	24.8	**97.3**
3-gram	29.9	**24.9**	96.1
4-gram	**30.3**	**24.9**	97.0
sentence	**30.6**	**25.0**	**98.1**

Table 2: Results of different transfer functions on sentence-conditional model (denoted as sc) with 34-layer ResNet. Top scores for each metric are highlighted. All values are reported as percentage (%).

	BLEU@4	METEOR	CIDEr
sc-relu	30.5	**25.0**	**98.1**
sc-tanh	**30.6**	**25.0**	**98.1**
sc-softmax	30.2	24.9	97.1
sc-sigmoid	30.1	24.8	96.2

Table 3: The Impact of image encoding CNNs on captioning performance. Tanh is used as transfer function. Top result for each column is highlighted. All values are reported as percentage (%).

Methods	BLEU@4	METEOR	CIDEr
sc-vgg-16	30.1	24.7	97.0
sc-resnet-34	30.6	25.0	98.1
sc-resnet-200	**31.6**	**25.6**	**101.2**

4.2 Model Variants & Attention

N-gram v.s. Sentence. Table 1 shows results with n-gram word- and sentence-conditional models. For conciseness, we only use BLEU@4, METEOR and CIDEr as the evaluation metrics, since they are more correlated with human judgments compared with low-level BLEU scores [30]. It turns out generally, word-conditional models with higher grams yield better results, especially for METEOR. Notice that the $2, 3, 4$-gram models achieve considerably better results than 1-gram model, which is reasonable as the 1-gram has the least context that limits the attention performance. Furthermore, the sentence-conditional model outperforms all word-conditional models in all metrics, which shows the importance of long-term word dependency in attention modeling.

Table 4: Top-6 nearest neighbors for randomly picked words. The results are based on the text-conditional matrix W_c of 1-gram word-conditional model. We highlight similar words in semantics.

dog	**bear** three woman **cat** girl person
banana	it **carrots** **fruits** six onto includes
red	UNK **blue** three several man **yellow**
sitting	**standing** next are **sits** dog woman
man	**woman** **person** **his** three are dog

Transfer Function. We use a non-linear transfer function $\Phi(\cdot)$ in our attention model (see Eq. 5) and we test four different functions: Softmax, ReLU, Tanh and Sigmoid. In all cases, we initialize the text-conditional embedding matrix with noises from one-mean Gaussian distribution with standard deviation 0.001. We base our experiments on the sentence-conditional model and conclude that the model achieves best performance when $\Phi(\cdot)$ is a Tanh or a ReLU function (see Table 2). Notice that it is possible that other transfer functions different than the four we tested may lead to better results.

Image Encoding. We study the impact of image encoding CNNs on captioning performance, as shown in Table 3. In general, the more sophisticated image encoding architecture the higher performance of the captioning.

4.2.1 Attention. It is essential to verify whether our learned text-conditional attention is semantically meaningful. Each column in the text-conditional matrix W_c is an attention mask for image features, and it corresponds to a word in our dictionary. It is expected that similar words should have similar masks (with some variations). To verify, we calculate the similarities among masks using Euclidean distance. We show five *randomly* sampled words w.r.t. different parts of speech (noun, verb and adjective). Table 4 shows their top few nearest words. Most of the neighbors are related to the original word, and some of them are strongly related, such as "cat" for "dog", "blue" for "red", "sits" for "sitting", and "woman" for "man". This shows strong evidence that our model is learning meaningful text-conditional attention.

4.3 Compare to State-of-The-Art Methods

We use LSTM with time-invariant image guidance (img-gLSTM) [14] and NeuralTalk2 [16], an implementation of [31], as baselines. We also compare to a state-of-the-art non-attention-based model—LSTM with semantic embedding guidance (emb-gLSTM) [14]. Furthermore, we compare our method to a set of state-of-the-art attention-based methods including visual attention with soft- and hard-attention [35], and semantic attention with visual attributes (ATT-FCN) [38]. For fair comparison among different attention models, we report our results with 16-layer VGGNet [28] since it is similar to the image encodings used in other methods.

Table 5 shows the comparison results. Our methods, both 1-gram word-conditional and sentence-conditional, outperform our two baselines in all metrics by a large margin, ranging from 1% to 5%. The results are strong evidence that 1) our td-gLSTM is better suited for captioning comparing to time-invariant gLSTM; and 2) modeling textual context is essential for image captioning. Also, our methods yield much higher evaluation scores than emb-gLSTM [14] showing the effectiveness of using textual content in our model.

We further compare our text-conditional methods with state-of-the-art attention-based methods. For 1-gram word-conditional method, the attention on the image feature guidance is merely determined by the previously generated word. Apparently, this results in semantic information loss. Even though, its performances are still on par with or better than state-of-the-art attention-based methods, such as Hard-Attention and ATT-FCN. We then upgrade the word-conditional model to the sentence-conditional model, which leads to improved performance in all metrics, and it outperforms state-of-the-art methods in most metrics. It worth noting that BLEU@1 score is related to single word accuracy, and highly affected by word vocabularies. This might result in our relatively low BLEU@1 score compared with hard-attention [35].

4.3.1 Human Evaluation. We choose three methods for human evaluation, NeuralTalk2, img-gLSTM and our sentence-conditional attention model. A cohort of five well-trained human annotators performed the experiments. Each of the annotators were shown 500 pairs of randomly selected images and three corresponding generated captions. The annotators rate the three captions from 0 to 3 regarding the content quality and grammar (the higher the better). For content quality, a score of 3 is given if the caption describes all the important content, e.g. objects and actions, in the image; a score of 0 is given if the caption is totally wrong or irrelevant. For grammar, a score of 3 denotes human-level natural expression and a score of 0 means the caption is unreadable. The results are shown in Table 6. Our proposed sentence-conditional model lead the baseline img-gLSTM by a large margin of 28.2% in the caption content quality, and 3.1% compared to the baseline Neuraltalk2, showing the effectiveness of our attention mechanism in captioning. As for grammar, all the methods create human-like sentences with a few grammar mistakes, and adding sentence-conditional attention to LSTM yields a slightly higher grammar score, due to the explicitly textual information contained in the LSTM guidance input.

4.3.2 Qualitative Results. Figure 4 shows qualitative captioning results. The fix images in the first three rows are positive examples and the last two are failed cases. Our proposed model can better capture details in the target image, such as "yellow fire hydrant" in the second image, and "soccer" in the fifth image. Also, the text-conditional attention discovers rich context information in the image, such as the "preparing food" followed by "kitchen" in the first image, and the 'in their hand'" followed by "holding" in the sixth image. However, we also show the failed cases, where the objects

Table 5: Comparison to baselines and state-of-the-art methods. For some competing methods, we extract their performance from the corresponding papers. For a fair comparison, we use 16-layer VGGNet for image encoding. Top-two scores for each metric are highlighted. All values are reported as percentage (%).

Methods	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	CIDEr
img-gLSTM [14]	64.7	45.9	31.1	21.4	20.4	67.7
emb-gLSTM [14]	67.0	49.1	35.8	26.4	22.7	81.3
NeuralTalk2 [16]	70.5	53.2	39.2	28.9	24.3	92.3
Hard-Attention [35]	**71.8**	50.4	35.7	25.0	23.0	-
Soft-Attention [35]	70.7	49.2	34.4	24.3	23.9	-
ATT-FCN [38]	70.9	53.7	**40.2**	**30.4**	24.3	-
Our 1-gram-vgg-16	71.5	**54.2**	40.0	29.6	**24.5**	**95.5**
Our sc-vgg-16	**71.6**	**54.5**	**40.5**	**30.1**	**24.7**	**97.0**

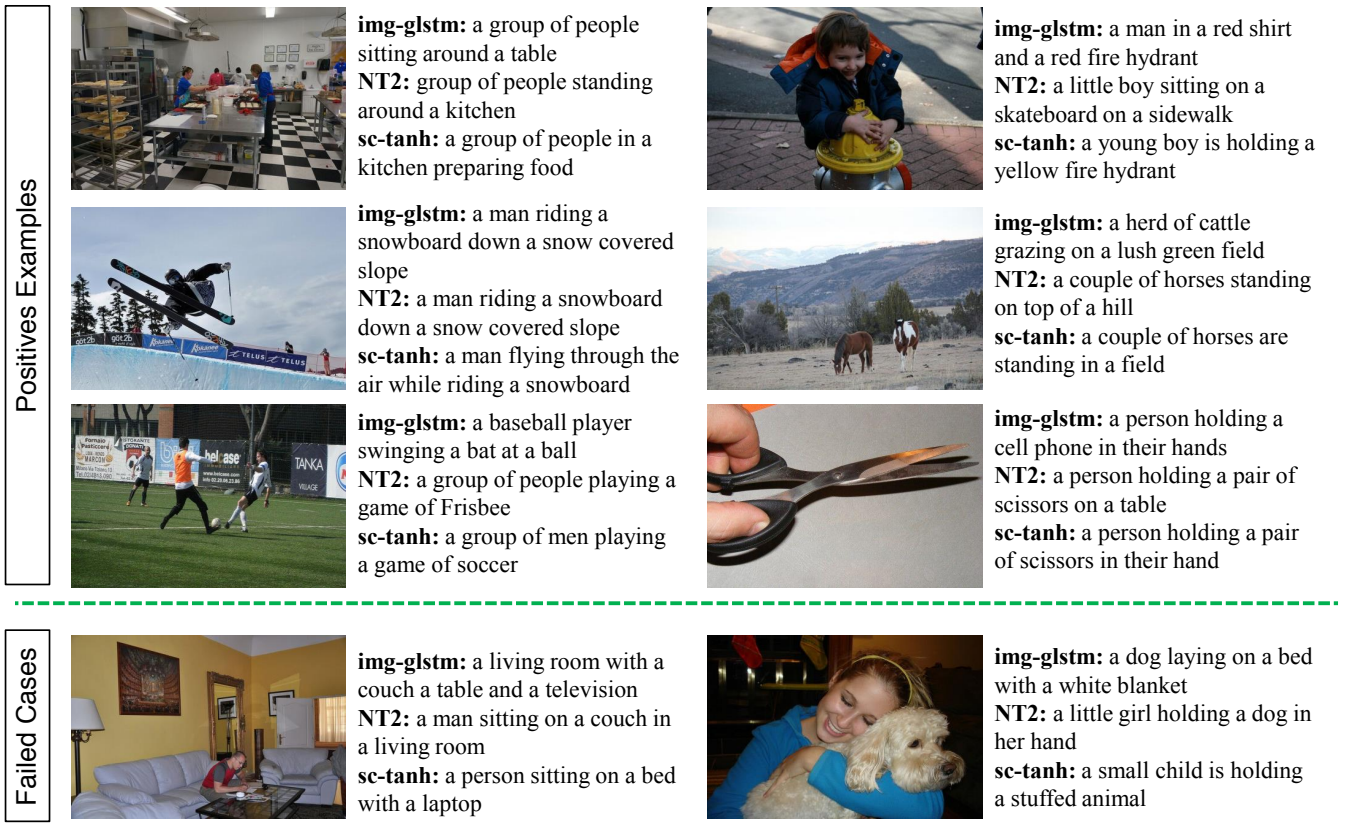

Figure 4: Qualitative results. We show both positive examples and failed cases of our method. NT2 indicates NeuralTalk2 and sc-tanh is our sentence-conditional model. Better viewed in color.

Table 6: Results of human evaluation. The higher the better for both content score and grammar score. The highest score for each column is highlighted.

Methods	Content	Grammar
img-gLSTM [14]	1.56	2.75
NeuralTalk2 [16]	1.94	2.77
Our sc-vgg-16	**2.00**	**2.80**

are mistakenly inferred from the previous words. For the first image, when we feed in the word sequence "a man (is) sitting", our text-conditional attention is triggered by things can be sat by a man; a sofa is a reasonable candidate according to the training data. Similarly, for the second image, the model is trained on some images with stuffed animal held by a person, which in some sense biases the semantic attention model.

Table 7: Evaluation on MS-COCO leaderboard. We list state-of-the-art published results. We highlight our method and our baseline method, NeuralTalk2. Notice that methods highly-ranked in learderboard use better CNNs, inference methods, and more careful engineering than research publishes.

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr-D	
Methods	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
ATT_VC [38]	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
OriolVinyals [32]	0.713	0.895	0.542	0.802	0.407	0.694	0.309	0.587	0.254	0.346	0.530	0.682	0.943	0.946
jeffdonahue [6]	0.718	0.895	0.548	0.804	0.409	0.695	0.306	0.585	0.247	0.335	0.528	0.678	0.921	0.934
SC-Tanh (ours)	**0.717**	**0.887**	**0.546**	**0.794**	**0.405**	**0.682**	**0.300**	**0.569**	**0.248**	**0.330**	**0.515**	**0.667**	**0.923**	**0.929**
Q.Wu [34]	0.725	0.892	0.556	0.803	0.414	0.694	0.306	0.582	0.246	0.329	0.528	0.672	0.911	0.924
Human	0.663	0.880	0.469	0.744	0.321	0.603	0.217	0.471	0.252	0.335	0.484	0.626	0.854	0.910
NeuralTalk2 [16]	**0.706**	**0.877**	**0.530**	**0.776**	**0.388**	**0.657**	**0.284**	**0.541**	**0.238**	**0.317**	**0.515**	**0.654**	**0.858**	**0.865**

4.3.3 Leaderboard Competition. We test our model on the MS-COCO leaderboard competition and summarize the results in Table 7. Our method outperforms the baseline (NeuralTalk2 [16]) across all the metrics and is on par with state-of-the-art methods. It worth noting that our baseline is an open source implementation of [31], shown as OriolVinyals in Tab. 7, but the latter performs much better due to better CNNs, inference methods, and more careful engineering. Also, several methods unreasonably outperform human-annotated captions, which reveals the drawback of the existing evaluation metrics.

5 CONCLUSION

In this paper, we propose a semantic attention mechanism for image caption generation, called text-conditional attention, which provides explicitly text-conditioned image features for attention. We also improve the existing gLSTM framework by introducing time-dependent guidance, opening up a new way for further boosting image captioning performance. We show in our experiments that the proposed methods significantly improve the baseline method and outperform state-of-the-art methods, which supports our argument of explicit consideration of using text-conditional attention modeling.

Future Work. There are several ways in which we can further improve our method. First, combining text-conditional attention with region-based or attribute-based attention, so that the model can learn to attend on regions in feature maps or attributes extracted from the image. Second, one common issue with supervised training is overfitting. As Vinyals et al. [31] pointed out, we cannot access enough training samples, even for the relatively huge dataset such as MS-COCO. One possible solution is to combine weakly annotated images with current dataset, such as [10]. We keep those for our future work.

Acknowledgement. This work was partially support by DARPA W31P4Q-16-C-0091, NSF NRI 1522904, ARO W911NF-15-1-0354 and a gift from Google. This article solely reflects the opinions and conclusions of its authors and not DARPA, NSF, ARO nor Google. We sincerely thank Vikas Dhiman and Suren Kumar for their helpful discussions.

REFERENCES

[1] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep Compositional Captioning: Describing Novel Object Categories Without Paired Training Data. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing*.

[4] Pradipto Das, Rohini K Srihari, and Jason J Corso. 2013. Translating related words to videos and back through latent topics. In *WSDM*.

[5] Robert Desimone and John Duncan. 1995. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18, 1 (1995), 193–222.

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Desmond Elliott and Frank Keller. 2013. Image Description using Visual Dependency Representations. In *Empirical Methods in Natural Language Processing*.

[8] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[9] Ali Farhadi, Mohsen Hejrati, Mohammad Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*.

[10] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*.

[11] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition. In *IEEE International Conference on Computer Vision*.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[14] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the Long-Short Term Memory Model for Image Caption Generation. In *IEEE International Conference on Computer Vision*.

[15] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[16] Andrej Karpathy. 2015. neuraltalk2. https://github.com/karpathy/neuraltalk2. (2015).
[17] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[18] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[19] 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 and Machine Intelligence* 35, 12 (2013), 2891–2903.
[20] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Association for Computational Linguistics*.
[21] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Computational Natural Language Learning*.
[22] Xiangyang Li, Xinhang Song, Luis Herranz, Yaohui Zhu, and Jiang Shuqiang. 2016. Image Captioning with both Object and Scene Information. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1107–1110.
[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.
[24] Xiao Lin and Devi Parikh. 2016. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*.
[25] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *International Conference on Learning Representations*.
[26] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *EACL*. Citeseer.
[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.
[28] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.
[29] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*.
[30] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.
[31] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2016).
[33] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 988–997.
[34] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. 2016. What value do explicit high level concepts have in vision to language problems?. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*.
[36] Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 444–454.
[37] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. 2016. Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1008–1017.
[38] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning With Semantic Attention. In *IEEE Conference on Computer Vision and Pattern Recognition*.