High-Quality Image Captioning With Fine-Grained and Semantic-Guided Visual Attention

Zongjian Zhang, Qiang Wu, *Senior Member, IEEE*, Yang Wang, *Senior Member, IEEE*, and Fang Chen, *Senior Member, IEEE*

***Abstract*—The soft-attention mechanism is regarded as one of the representative methods for image captioning. Based on the end-to-end convolutional neural network (CNN)-long short term memory (LSTM) framework, the soft-attention mechanism attempts to link the semantic representation in text (i.e., captioning) with relevant visual information in the image for the first time. Motivated by this approach, several state-of-the-art attention methods are proposed. However, due to the constraints of CNN architecture, the given image is only segmented to the fixed-resolution grid at a coarse level. The visual feature extracted from each grid indiscriminately fuses all inside objects and/or their portions. There is no semantic link between grid cells. In addition, the large area "stuff" (e.g., the sky or a beach) cannot be represented using the current methods. To address these problems, this paper proposes a new model based on the fully convolutional network (FCN)-LSTM framework, which can generate an attention map at a fine-grained grid-wise resolution. Moreover, the visual feature of each grid cell is contributed only by the principal object. By adopting the grid-wise labels (i.e., semantic segmentation), the visual representations of different grid cells are correlated to each other. With the ability to attend to large area "stuff," our method can further summarize an additional semantic context from semantic labels. This method can provide comprehensive context information to the language LSTM decoder. In this way, a mechanism of fine-grained and semantic-guided visual attention is created, which can accurately link the relevant visual information with each semantic meaning inside the text. Demonstrated by three experiments including both qualitative and quantitative analyses, our model can generate captions of high quality, specifically high levels of accuracy, completeness, and diversity. Moreover, our model significantly outperforms all other methods that use VGG-based CNN encoders without fine-tuning.**

***Index Terms*—Image captioning, attention mechanism, fine-grained resolution, semantic guidance, fully convolutional network-long short term memory framework.**

Manuscript received February 7, 2018; revised July 10, 2018 and October 22, 2018; accepted November 21, 2018. Date of publication December 20, 2018; date of current version June 21, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hantao Liu. *(Corresponding author: Zongjian Zhang.)*

Z. Zhang, Q. Wu, and F. Chen are with the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: Zongjian.Zhang@student.uts.edu.au; Qiang.Wu@uts.edu.au; Fang.Chen@uts.edu.au).

Y. Wang is with the Data61, Commonwealth Scientific and Industrial Research Organisation, Eveleigh, NSW 2015, Australia (e-mail: Yang.Wang@data61.csiro.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2888822

I. Introduction

VISUAL captioning is a challenging multi-modal scene understanding task, requiring a deep understanding of two totally different types of media data, i.e., vision and language. In this sense, this task bridges Computer Vision and Natural Language Processing [1]. "Vision" refers to a raw appearance of open-ended and free-form real-world scenes [2], whereas "language" refers to a high-level extraction with a strict structure. Therefore, the nature of this task makes multi-modal learning [3] on these two types of modal data challenging, specifically shared feature space modeling [4] and semantic alignment learning [5]. Although the task is easily handled by humans, it is difficult for AI. Therefore, visual captioning is regarded as an important AI-complete task, as it aims to achieve the ultimate AI goal. Through the automatic generation of captions based on a comprehensive understanding of real-world scenes, visual captioning can benefit human-machine interaction, autonomous/assisted driving, and intelligent navigation for visually impaired people. Currently, most research focuses on the problems associated with two major tasks: image captioning [6]–[19] and video captioning [20], [21]. Video captioning is the more difficult task of the two, as video involves an extra temporal dimension [4].

An accurate and diverse description requires a comprehensive understanding of objects and/or "stuff" [22], and their mutual relationships/interactions in all the different image regions, which are then selectively dealt with according to their semantic relationships to each generated word. Such a visual attention mechanism has attracted a great deal of research interest, leading to significant performance improvement [6]–[11], [19]–[21]. Generally, the attention mechanism has two roles. The first one is to learn a shared features space, where vision and language can be jointly modeled. The second one is learning semantic alignment, by mapping together related visual elements and words/phrases. This mechanism can be further extended to multiple types of multimedia data [2], [21], [23], Affective Analysis and Retrieval [3], context modeling [5], semi-supervised annotation [24], etc.

Most state-of-the-art spatial visual attention models are based on the Convolutional Neural Network (CNN)-Long Short Term Memory (LSTM) framework in an end-to-end trainable way [7]–[9], [19]–[21]. CNN plays the role of image encoder, responsible for understanding visual regions and encoding them into region-specific features at different locations. There are two

main ways of capturing regions. The most common method is to divide the image into grid cells based on the model structure of CNN [7]–[9], [20], [21], which is a hard way of splitting regions. To encode each grid region in the image, the outputs of the last convolutional layer in CNN are usually extracted as the visual feature representation for each region. Another method involves capturing regions at the object level using bounding box [19], which is adaptive and accurate. Similar to CNN for grid-based region features, Region-CNN (R-CNN) is used for providing object candidates and extracting their visual features. As a caption decoder, LSTM is responsible for understanding all words that have been generated and generating the following word at each time step. The attention mechanism serves as an agent between CNN encoder and LSTM decoder. In generating each word, the mechanism makes joint inferences and adaptively attends to those semantically related image regions by generating a distinct attention weight for each region. Based on this weight map, a visual context feature is firstly summarized through the weighted sum of all region features encoded by CNN and is then fed into the LSTM for language inference. In general, an accurate and comprehensive understanding of all grid regions at the image encoder side is the premise for a stronger attention mechanism, and hence plays a fundamental role in generating high-quality captions in terms of accuracy, completeness, and diversity.

However, to the best of our knowledge, current soft-attention-based approaches only use CNN as the image encoder to create the attention module. Their underlying CNN-LSTM framework has four limitations in providing an accurate attention mechanism and high-quality captions: 1) Due to the constraints of current CNN architecture, the attention mechanism has a fixed low grid resolution in the soft-attention framework. [25] supports attention mechanism in a 14×14 grid resolution using VGG as CNN encoder. [26] and [20] supports a 7×7 grid resolution using ResNet and Inception V3. Moreover, it is impossible to elevate it to a fine-grained level. 2) The representation of each grid cell is indiscriminately a mix of visual information about all objects and/or their portions inside this cell. Therefore, it lacks the semantic correspondence related to the most salient visual cue within the grid cell. 3) Due to the lack of mutual reference information across grid cells, those different grid cells containing partial visual information of the same objects cannot be correlated to each other. The semantic visual guidance just does not exist across grid cells. 4) Due to the object-oriented nature of the CNN encoder, existing soft-attention frameworks are not able to recognize and describe large area stuff, like the sky, beaches, and grass. Hence, the context information cannot be well represented based only on object information. Therefore, overcoming these four limitations would enhance the caption quality.

In addition, there are some special CNN-LSTM variants that use R-CNN as the image encoder [19], [27], [28]. Specifically, [19] proposed an object-level attention mechanism. This improvement can mitigate limitations 1) and 3) to a certain extent by attending to object proposals. The entire object region can be attended in a bounding box, preventing the splitting of one object into several grid-cell regions. However, it still suffers from the other two limitations. Other visual information is still mixed in the bounding box region. Only objects can be attended to by a bounding box, which is not the case for stuff of irregular shape [22]. Moreover, semantic connections between objects are overlooked.

In this paper, we propose a novel image captioning model with fine-grained and semantic-guided visual attention based on a novel Fully Convolutional Network (FCN)-LSTM framework, inspired by the soft-attention framework [9]. It leverages the spatially dense and semantically abundant outputs of FCN to solve the above-mentioned limitations. FCN is particularly designed for semantic segmentation task, specifically the dense pixel-level predictions [29], [30]. Therefore, it naturally excels in generating both visual features and semantic labels in the form of a spatial grid at a fine-grained level, which theoretically can reach the pixel level. Therefore, our model has five strengths:

1) Our model can have a fine-grained visual attention at a higher grid resolution, given the same sized image. It can attend to relevant object regions accurately, and hence can extract a precise context feature with a limited amount of noises. Moreover, the grid resolution of our attention module can be flexibly adjusted.
2) As the FCN encoder is both object-oriented and stuff-oriented, our model can extract a comprehensive representation of context information by attending to large area stuff, such as the sky or a beach. Therefore, the contextual inference is more comprehensive and accurate.
3) Based on pixel-level semantic labels, our model can represent each grid cell based on the dominating area that is associated with an object or its portion inside the cell. This saliency-related semantic correspondence can be maintained when the resolution is adjusted.
4) Guided by the semantic labels of all grid cells, our model can grasp the semantic layout across grid cells, and efficiently associate the grid cells containing different portions of the same object. In this way, incorrect inferences can be mitigated.
5) The semantic context feature can also be summarized from semantic labels to form the joint context feature with a visual context feature. This joint context feature can provide strong context information to the LSTM decoder.

Having the fine-grained and semantic-guided attention mechanism, our FCN-LSTM model demonstrates state-of-the-art performance on the Microsoft Common Objects in COntext (MSCOCO) dataset [31] on metrics such as BLEU@N (BiLingual Evaluation Understudy @ N-gram) [32], METEOR [33], and CIDEr (Consensus-based Image Description Evaluation) [34]. Specifically, this study is based on three experiments, demonstrating that our model can generate high-quality captions with high levels of accuracy, completeness, and diversity. In experiment 1, the high attention resolution can enhance the accuracy of both the attention map and meaningful words, particularly for small area objects. In experiment 2, integrating the semantic guidance into our model can further enhance the attention accuracy, particularly for large area objects and stuff. Iit can also make the single caption more complete by generating new meaningful words. In experiment 3, the diversity of the top three captions is enhanced by semantic guidance.

Regarding semantic guidance in experiment 2 and 3, three forms of integration are combined: Saliency Guidance (Case 2), Explicit Semantic Guidance (Case 3), and Semantic Context Feature (Case 4).

This paper is organized into five sections: This first section is an introduction, which is followed by the second section about related works. In section three, our model will be described in detail. Section four will provide the experiment details. The last section will be a conclusion of this study.

II. Related Works

Image captioning has attracted a great deal of research interests, and many different models have been proposed. Recently, due to substantial advances in Deep Neural Networks (DNNs) [19], [25], [26], [29], [30], [35], most state-of-the-art approaches are mainly based on this framework. In particular, the best one is the encoder-decoder neural framework [6]–[11], [19], [36]–[43] inspired by Machine Translation [36]. In this mainstream framework, CNN is generally used as the visual encoder that is responsible for understanding the visual scene, and RNN serves as the language decoder, understanding and generating language. Specifically, the CNN encoder is responsible for extracting image features at the highest semantic level. These image features are then fed into RNN decoder to generate the natural language caption in a sequential manner, word-by-word.

Attention mechanism is a significant area of research, which can achieve the state-of-the-art performance. Caption generation is a dynamic decoding process in which each different time-step needs a different combination of visual information. To this end, the attention mechanism bridges the CNN encoder and Recurrent Neural Network (RNN) decoder together efficiently by enabling the RNN decoder to adaptively attend to, via a weight map, only those image features that are semantically related to the word to be generated at a certain time-step. Based on this weight map, a context feature is summarized by using the weighted sum, and it is then fed into the RNN decoder for language inference. So far, the attention mechanism has been researched in three respects. They all try to establish an alignment between visual information and word information in an LSTM style. The major difference between these attention methods lies in the outputs of the encoder.

A. Grid-Wise Visual Feature Without a Semantic Label

This type of attention model focuses on which spatial regions need to be attended to. The features of regions at different locations are extracted by the CNN encoder from its last convolutional layer and fed into the attention model for relativity inference. This type of attention mechanism is generally integrated into an end-to-end trainable encoder-decoder framework, and it is trained implicitly without any explicit supervision. As the pioneer in attention mechanism research, [9] proposed a 14×14 grid resolution (VGG) spatial attention model for image captioning using two different pooling methods. The "soft" attention model combines all spatial features based on soft probabilistic attention weights, whereas the "hard" attention model attends to the only one region feature with the highest relevance based on hard binary weights. [20] applied this pipeline to video captioning, extending the attention mechanism from the spatial dimension to the temporal dimension. [8] further proposed a time-wise adaptive attention model, at a 7×7 grid resolution (ResNet), by introducing a visual sentinel. For each word generation, this model can automatically determine when to attend to the image regions and when to simply rely on the decoder knowledge. Based on the nature of CNN structure, [7] proposed a novel channel-wise and multi-layer spatial attention model, which additionally attend to related channels among the multi-layer feature maps. Different channels in a certain layer represent a specific semantic concept, which has a different level of semantic abstraction as a different layer. However, all these spatial attention models have a fixed low grid resolution, which is difficult to convert to high grid resolution. Moreover, being object-oriented due to the nature of the CNN encoder, they are not able to recognize large area stuff as the sky or a beach. Another problem is that it lacks the ability to represent the connections between the grid cells on the image.

B. Attribute-Based Visual Representation

This type of attention model chooses which semantic concepts need to be prioritized. The image feature is represented by a confidence vector for all concepts, which is a mixture of objects, stuff, attributes, interactions, relations, etc. [11] proposed a semantic attention model to attend to related visual attributes for inputs and outputs respectively. These attributes are detail-oriented and are trained by convolutionalized CNN with Multiple Instance Learning (MIL) in a separate stage. Based on the gLSTM model, [44] proposed a text-conditional semantic attention model. Using this attention model, the caption generator can automatically learn on which parts of the image feature it should focus, given previously generated text. Although such models involve rich semantic concepts, they lack the significant spatial layout.

C. Objectness-Based Visual Representation

This model aims to identify the latent correspondence between sentence segments and image regions which corresponds to the objects detected in the image. [28] proposed an alignment model, based on Region-CNN (R-CNN) and Bidirectional RNN (BRNN), to infer the latent alignments between image regions and segments of sentences by treating the sentences as weak labels. Then, an end-to-end multimodal RNN model was proposed to generate descriptions for image regions. To be able to automatically locate and describe object regions, [27] proposed an end-to-end trainable Fully Convolutional Localization Network (FCLN) model to resolve a dense captioning problem, namely localizing and describing the salient regions of images. By further integrating the image-level feature as a global context with object-level features, [19] proposed a global-local attention model. The model can attend to related local objects and global context information simultaneously. However, these methods focus too much on objectness rather than the large area stuff using the bounding box, and they overlook the connections between these detected objects.

To the best of our knowledge, our FCN-LSTM model is the first work to propose a novel attention mechanism that combines

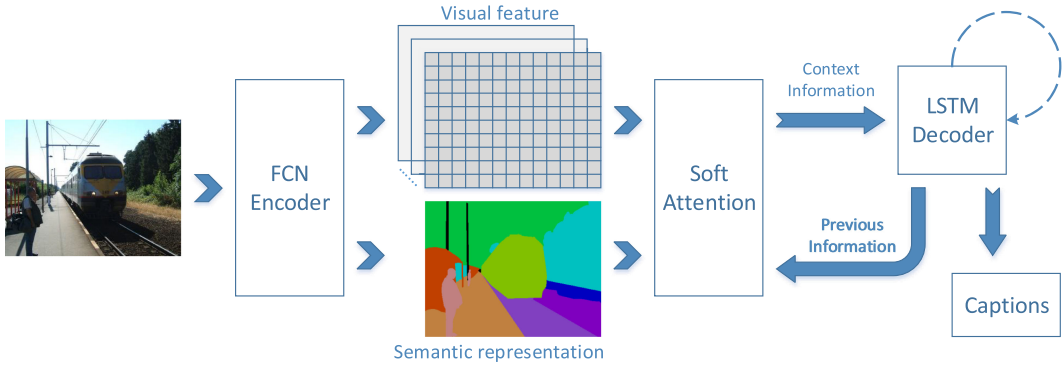


Fig. 1. The overview of our proposed framework.

grid-wise visual representation with the grid-wise semantic label at a fine-grained resolution. Moreover, our model can grasp the semantic connections between all objects and stuff in the image.

III. Method

We firstly describe the overall FCN-LSTM framework for our captioning model in Section A, and then further introduce our fine-grained and semantic-guided attention modules in Section B.

A. FCN-LSTM Framework for Image Captioning

Similar to the mainstream CNN-LSTM framework, our novel FCN-LSTM framework is also a variant of the Encoder-Decoder framework for image captioning. It can be regarded as a translation from vision to language. The FCN encoder firstly extracts both visual representations and semantic labels from the input image at the pixel level, then the LSTM decoder generates caption word-by-word based on joint understanding over the visual and semantic information. Given an image and its corresponding caption, the FCN-LSTM model maximizes the probability of word sequence:

$$\boldsymbol{\theta}^* = \arg\max_{\theta} \sum_{(\boldsymbol{I},\boldsymbol{y})} \log p(\boldsymbol{y}|\boldsymbol{I};\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta}$ represents the model parameters, $\boldsymbol{I}$ is the image, and $\mathbf{y} = \{y_1, y_2, \ldots, y_t\}$ is the word sequence of corresponding caption. Based on chain rule, the log likelihood of the joint probability distribution over $\mathbf{y}$ is comprised of $\boldsymbol{T}$ conditional probabilities:

$$\log p(\boldsymbol{y}) = \sum_{t=1}^{T} \log p(y_t|y_{t-1}, \ldots, y_1, \boldsymbol{I}) \tag{2}$$

where $\boldsymbol{T}$ is the total length of the caption. Here, the dependency on model parameters $\boldsymbol{\theta}$ is removed for convenience. During the training phase, $(\boldsymbol{I}, \boldsymbol{y})$ is a training image-caption pair, and the overall optimization objective is the sum of log probabilities over all training pairs in the training set. During the testing phase, only image $\boldsymbol{I}$ is fed into the model for caption generation.

Specifically, our FCN-LSTM framework consists of three parts: FCN encoder, LSTM decoder, and soft-attention model (Fig. 1). It firstly uses the FCN encoder to extract both spatial visual features and semantic representations from the image at the pixel level. Then, the fine-grained and semantic-guided soft-attention summarizes all outputs of the FCN encoder into a joint context feature for the LSTM decoder to generate captions.

1) FCN Encoder: Particularly designed for the semantic segmentation task, FCN can directly perform the pixel-wise classification. To encode the image, our framework employs the FCN to directly extract both visual feature and semantic label for each different pixel in the image. First of all, the N $\times$ N sized image $\boldsymbol{I}$ can be represented by the spatial visual features:

$$V = FCN_v(\boldsymbol{I}) = \{v_1, v_2, \ldots, v_k\} \tag{3}$$

where $k = \mathrm{N}^2$ is the number of image pixels. Each feature $v_i \in R^d$ is a d dimensional representation corresponding to an image pixel. Specifically, the visual features are taken from the second last layer of FCN. This is similar to what CNN encoder does in the CNN-LSTM framework. Differently, the image $\boldsymbol{I}$ also has corresponding spatial semantic representations:

$$S = FCN_s(\boldsymbol{I}) = \{s_1, s_2, \ldots, s_k\} \tag{4}$$

where s_i is a semantic label for each pixel indicating which object or stuff it may belong to. Note that the concatenation of 2-D image pixels into 1-D form does not break the spatial correspondence.

2) LSTM Decoder: As each conditional probability in Equation 2 can be naturally modeled based on the RNN, our model adopts the LSTM as the caption decoder. At time t, the previous conditional variable-length word sequence $\{y_1, y_2, \ldots, y_{t-1}\}$ and image I are represented by the fixed-length hidden state h_t of LSTM as following:

$$x_t = W_e y_{t-1} \tag{5}$$

$$h_t = LSTM(x_t, h_{t-1}, c_t) \tag{6}$$

Here, y_{t-1} is the output word at time $t-1$. As the current new input, x_t is the word embedding of y_{t-1} based on the embedding matrix W_e. Each word y_i is simply encoded as the one-hot vector. h_{t-1} is the hidden state representing the conditional word sequence $\{y_1, y_2, \ldots, y_{t-2}\}$ and image I. c_t is the context feature extracted from image at time t by the attention mechanism. This context feature represents the dynamic combination of visual and semantic information from image I.

Specifically, the detailed definition of the LSTM decoder is as follows:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_t + b_i) \quad (7)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_t + b_f) \quad (8)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_t + b_o) \quad (9)$$

$$g_t = \tanh(W_{gx}x_t + W_{gh}h_{t-1} + W_{gc}c_t + b_g) \quad (10)$$

$$m_t = f_t \odot m_{t-1} + i_t \odot g_t \quad (11)$$

$$h_t = m_t \odot o_t \quad (12)$$

Here, i_t, f_t, o_t, g_t, m_t, h_t are the input gate, forget gate, output gate, modulated input, memory, and hidden state of the LSTM at time t respectively. Moreover, the operation σ, $\tanh$, $\odot$ are the sigmoid, hyper tangent, and element-wise multiplication respectively.

Finally, the probability of generating word y_t at time t is modeled based on the input (previous word), hidden state, and context feature as follow:

$$p(y_t|y_{t-1}, \ldots, y_1, \boldsymbol{I}) = f(h_t, x_t, c_t)$$
$$= \text{softmax}(W \tanh(W_h h_t + W_c c_t + x_t + b_h) + b) \quad (13)$$

B. Fine-Grained Grid-Wise Soft-Attention

Traditionally, the soft-attention mechanism [36] selectively attends to relevant regions in the image with reference to previously generated words and generates an attention distribution in the form of a weight map over all regions. A higher attention weight indicates that the region has a higher relevance (or importance) to the generation of the next word and vice versa. Then, based on the attention distribution, the information of relevant regions is summarized together and fed into the LSTM decoder as the above-mentioned context feature c_t. Therefore, this attention mechanism serves as an agent between the FCN encoder and the LSTM decoder by sending needed information from the former to the latter. A better attention mechanism provides a more accurate context feature to the LSTM decoder, which can then generate a more correct word for a caption of higher quality.

Our soft-attention mechanism is enhanced by the fine-grained attention resolution based on this novel FCN-LSTM framework. It can attend to relevant regions more accurately based on a higher resolution weight map, which will make the visual context feature c_t more accurate. Specifically, our fine-grained attention inherits the pixel-wise nature of the FCN encoder, which is practically grid-wise so far. Therefore, the fine-grained grid-wise resolution is determined by the resolution of the FCN encoder's grid output, which can theoretically reach up to the pixel level. Actually, most FCN encoders can only reach a certain small-patch level, and each grid cell corresponds to a small patch (n × n pixels) in the image. Due to this, all regions of relevant objects/stuff can be attended to with a high spatial accuracy, as smaller patch can distinguish the object/stuff boundary more precisely. Particularly, at the object boundary, the grid patch contains pixels of both this object and its neighbors (including other objects and stuff). Using a smaller grid patch (i.e., fine-grained grid) can mitigate the noisy information created by neighbor objects and stuff. Hence, the context feature will be more accurate because of less irrelevant information.

Therefore, our fine-grained grid-wise attention is modeled as Equation 14. It requires three inputs: V and S from the FCN encoder, and h_{t-1} from the LSTM decoder. V represents the spatial visual features through which the attention model attends to relevant regions locally. S represents the spatial semantic representations related to pixel-wise semantic labels. M × M is the grid resolution, and $g = \text{M}^2$ is the number of locations. The fine-grained grid-wise nature of V and S contributes to the fine-grained attention. α_{ti} represents the attention weight for the grid cell at the location $i = 1, 2, \ldots, l$ and the time t. h_{t-1} is the hidden state of the LSTM decoder at the time $t-1$, which contains previously generated words and their corresponding relevant image information.

$$c_t = f_{att}(h_{t-1}, V, S) \quad (14)$$

$$V = \{v_1, v_2, \ldots, v_g\} \quad (15)$$

$$S = \{s_1, s_2, \ldots, s_g\} \quad (16)$$

$$\boldsymbol{\alpha_t} = \{\alpha_{t1}, \alpha_{t2}, \ldots, \alpha_{tg}\} \quad (17)$$

Note that we use a different symbol g to indicate the original grid resolution, which equals to the resolution of the FCN encoder's outputs. This would equal to k (in Equation 3,4) when the FCN encoder achieves the pixel level. S plays the role of semantic guidance, which will be illustrated in below parts.

This part aims to enhance the accuracy of the attention map via merely fine-grained visual features, which would further enhance the meaning accuracy of generated keywords. In this way, the caption quality is improved specifically in terms of accuracy. The improvement of this part is demonstrated by Experiment 1 in Subsection C of Experiment.

C. Semantic-Guided Attention

In addition to the fine-grained attention resolution, the grid-wise semantic labels also serve as the semantic guidance for the attention model. Firstly, it provides a global view of semantic relationships among all grid-cell regions and hence can enhance the accuracy of the attention map and keyword. Moreover, it enriches the context feature with semantic context feature, which can benefit the completeness and diversity of captions. Illustrated by Fig. 2, our attention mechanism comprises three layers: the saliency pooling layer, the attention distribution prediction layer, and the joint context computation layer. The saliency pooling layer firstly extracts compact visual features V_c and semantic representations S_c. With both as inputs, the next layer predicts the attention distribution weight map α_t over all regions. Then, the context feature c_t is computed by adding two weighted sums of visual features V_c and semantic features S_e.

1) Saliency Pooling Layer: Ideally, FCN may ultimately provide pixel-level visual features and semantic labels, which then could be fed into rest layers of the attention model for further processing. However, in practice, due to the constraint of GPU memory and computation power, the rest layers can process the limited number of patches, although the patch can be of fine-grained size because of the nature of FCN. That means the visual features and semantic representations of pixels inside a

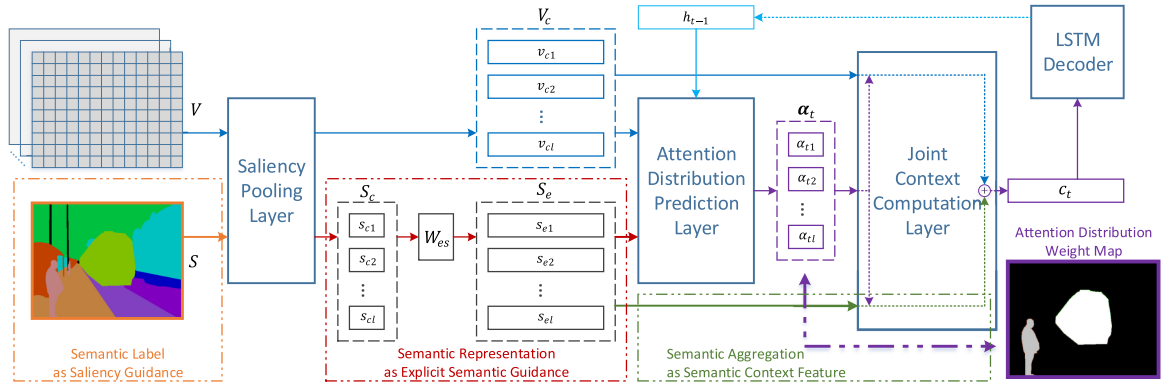


Fig. 2. The detailed structure of our fine-grained and semantic-guided attention model. (Best viewed in color.)

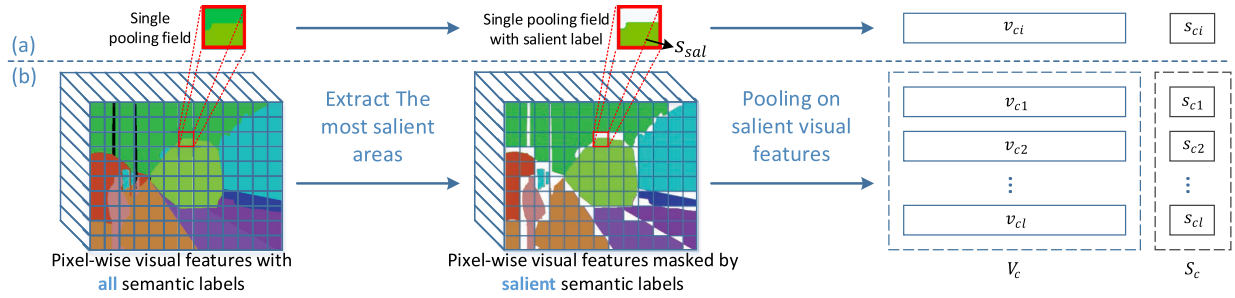


Fig. 3. An illustration of the saliency pooling layer for single field (a) and entire image (b). (Best viewed in color.)

patch of a given size have to be pooled together. Normally, this process on visual features can be carried out through a common average pooling which simply sums up the visual features of all pixels inside the patch with equal weights. In this paper, we propose a novel saliency pooling method which only pools the visual features of those salient pixels. The salient pixels are defined as those pixels whose pixel labels generated by FCN dominate inside the patch. The pooling process can be modeled as:

$$(V_c, S_c) = Pooling_{sp}(V, S) \tag{18}$$

$$V_c = \{v_{c1}, v_{c2}, \ldots, v_{cl}\} \tag{19}$$

$$S_c = \{s_{c1}, s_{c2}, \ldots, s_{cl}\} \tag{20}$$

Displayed in Fig. 3, it pools visual features V of the original grid resolution to compact visual features V_c at an acceptable lower level (i.e., $\mathrm{M}_c \times \mathrm{M}_c$), under the guidance of semantic representations S. S_c is the compact semantic representation. Let s_{sal_i} denote the labels of pixels which dominate the area inside the patch i, where $i = 1, 2, \ldots, l$. Then, s_{ci} in Equation 20 can be defined as:

$$s_{ci} = s_{sal_i} \tag{21}$$

Correspondingly, the number of grid locations is reduced to $l = \mathrm{M}_c^2$. Each v_{ci} is a brief visual feature pooled from those original visual features inside the pooling field i. The saliency pooling layer generates the visual feature v_{ci} in Equation 19 based on the salient pixels only. In the Equation 22 below, v_c is a generic representation of any patch v_{ci}, where $i = 1, 2, , l$.

$$v_c = \frac{1}{\sum_{j=1}^{\mathrm{w}^2} f_{sal}(s_j)} \sum_{j=1}^{\mathrm{w}^2} v_j \cdot f_{sal}(s_j) \tag{22}$$

$$f_{sal}(s_j) = \begin{cases} 1, & s_j = s_{sal} \\ 0, & s_j \neq s_{sal} \end{cases} \tag{23}$$

where j stands for the relevant location of each pixel inside the $\mathrm{w} \times \mathrm{w}$ pooling field. $\mathrm{w} \times \mathrm{w}$ is the size of the patch where the pooling processing is carried out. v_j is the visual feature of each pixel. w^2 represents the number of pixels inside the pooling field. It may be seen that if $f_{sal}(s_j)$ is enforced to be 1, the saliency pooling is equivalent to the common average pooling. Illustrated in Fig. 3, the output of saliency pooling layers are the salient visual features on the patches (i.e., pooling visual feature on salient pixels in the patch) and salient pixel labels of the patches.

This part aims to enhance the accuracy of attention map via compact and accurate visual features at a relatively lower grid resolution, which is implemented by integrating semantic labels as the saliency guidance for the saliency pooling layer. This form plays a role of an implicit semantic guidance based on the common average pooling. Similarly, it would also further enhance the meaning accuracy of generated keywords, and the caption quality is improved specifically in terms of accuracy. The improvement of this part is demonstrated by Experiment 2 in Subsection D of Experiment.

2) Attention Distribution Prediction Layer: In Fig. 2, the inputs of the attention distribution prediction layer include visual feature pooling (i.e., the proposed saliency pooling or the simple common average pooling), explicit semantic guidance, and the hidden state h_{t-1} feedback from LSTM. In the existing CNN-LSTM framework [7]–[9], there is no such explicit semantic guidance.

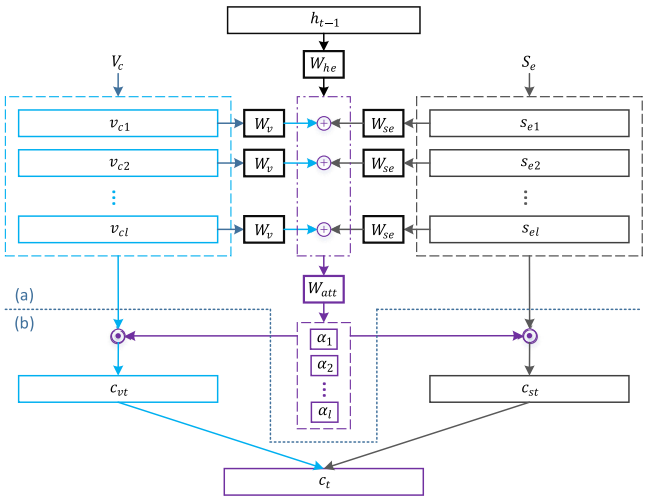


Fig. 4. An illustration of the attention distribution prediction layer (a) and joint context computation layer (b).

Similar to the word embedding for the LSTM decoder, the compact semantic representations S_c are a map of semantic label words, which are encoded as one-hot vectors. Therefore, they need to be embedded into dense semantic features via the embedding matrix W_{es}.

$$S_e = W_{es} S_c = \{s_{e1}, s_{e2}, \ldots, s_{el}\} \tag{24}$$

The attention prediction model is specifically designed as a two-layer perception. The first layer is mainly responsible for feature fusion. From different feature spaces, the hidden state h_{t-1}, compact visual features V_c and the dense semantic features S_e are mapped into a shared feature space by the embedding matrices W_{he}, W_v, and W_{se} respectively. As the hidden state h_{t-1} does not have the spatial dimension, an all-one vector $\hat{\mathbf{1}}$ is used to extend its spatial dimension by simple copying. Then, these three embedded features are merged via the element-wise sum and fed into the hyperbolic tangent activation function. The overall process can be illustrated in Fig. 4. The fused feature z_t is then fed into the second layer with a softmax function to generate the attention weights over l grid regions.

$$z_t = \tanh(W_{he} h_{t-1} \hat{\mathbf{1}} + W_v V_c + W_{se} S_e + b_z) \tag{25}$$

$$\boldsymbol{\alpha_t} = \text{softmax}(W_{att} z_t + b_{att}) = \{\alpha_{t1}, \alpha_{t2}, \ldots, \alpha_{tl}\} \tag{26}$$

where α_{ti} represents the attention distribution for the grid location $i = 1, 2, \ldots, l$ at the time t.

This part aims to enhance the accuracy of the attention map via extra grid-wise semantic representations, particularly for large area objects/stuff. Note that the semantic representations serve as the explicit guidance, as the semantic meanings of labels are fully used for guiding the attention prediction model. Specifically, the semantic labels are firstly embedded into dense semantic features and then mapped into a shared feature space with visual features, so as to guide the layer to predict the attention distribution explicitly. This would further enable the attention model to attend to novel objects/stuff or their relationships. Therefore, the meanings of captions would be more accurate and complete. The improvement of this part is demonstrated by Experiment 2 and Case 3 in Subsection D of Experiment.

3) Joint Context Computation Layer: Based on the attention weights, the visual context feature c_{vt} is computed as the weighted sum of compact visual features V_c, and the semantic context feature c_{st} is calculated as the weighted sum of dense semantic features S_e. See Fig. 4(b). Then, the joint context feature c_t is computed as the element-wise sum of c_{vt} and c_{st} and fed into the LSTM decoder for word generation.

$$c_{vt} = \boldsymbol{\alpha_t} \cdot V_c = \sum_{i=1}^{l} \alpha_{ti} v_{ci} \tag{27}$$

$$c_{st} = \boldsymbol{\alpha_t} \cdot S_e = \sum_{i=1}^{l} \alpha_{ti} s_{ei} \tag{28}$$

$$c_t = c_{vt} + c_{st} \tag{29}$$

This part aims to directly enhance the caption accuracy and completeness, by integrating an extra semantic context feature into the language model. This would further enhance the diversity of top-k captions. Therefore, the meanings of captions would be more accurate and complete. The improvement in accuracy and completeness are demonstrated by Experiment 2 in Subsection D of Experiment. The improvement in diversity is demonstrated by Experiment 3 in Subsection E.

In Equation (29), without considering c_{st}, c_t will become the aggregated visual feature based on attention distribution only. That is, c_t only presents the visual context instead of joint context. For the LSTM decoder, the initial hidden state h_t and memory state m_t are predicted by feeding the global average-pooled visual features into two separate single layer perceptions:

$$m_0 = \tanh(W_{m0} c_0 + b_{m0}) \tag{30}$$

$$h_0 = \tanh(W_{h0} c_0 + b_{h0}) \tag{31}$$

$$c_0 = \frac{1}{l} \sum_{i=1}^{l} v_{ci} \tag{32}$$

IV. Experiment

This section firstly specifies datasets, evaluation metrics, and experiment settings. Then, three experiments are designed to demonstrate three advantages - high levels of accuracy, completeness, and diversity, in terms of fine-grained resolution and semantic guidance. Regarding the semantic guidance, contributions of three forms are further studied. The saliency guidance is used in saliency pooling layer. The explicit semantic guidance is adopted in the attention distribution prediction layer. The joint context computation layer summarizes the context of semantic guidance for the LSTM decoder.

A. Datasets and Metrics

Our experiments use two datasets. **MSCOCO** [31] is the largest dataset for image captioning, with 82,783 training images, 40,504 validation images, and 40,775 testing images. For the offline evaluation, we use the same data split as [9], [11], containing 5000 images for validation and test respectively. The length of the captions is truncated to be no larger than 16. The word vocabulary is built with only those words occurring at

TABLE I
PERFORMANCES COMPARED WITH THE STATE-OF-THE-ART MODELS ON MSCOCO TEST SPLIT VIA ALL METRICS

Method	B@1	B@2	B@3	B@4	METEOR	CIDEr
NIC v1 [45]	0.666	0.461	0.329	0.246	-	-
DeepVS [28]	0.625	0.450	0.321	0.230	0.195	0.660
emb-gLSTM [46]	0.670	0.491	0.358	0.264	0.227	-
m-RNN [47]	0.670	0.490	0.350	0.250	-	-
Soft-Attention [9]	0.707	0.492	0.344	0.243	0.239	0.773
Hard-Attention [9]	**0.718**	0.504	0.357	0.250	0.230	-
SCA-VGG-1layer [7]	-	-	-	**0.281**	0.235	0.847
Our Model (Attention Resolution 27×27 with Joint Context Feature)	0.712	**0.514**	**0.368**	0.265	**0.247**	**0.882**

least 5 times in the training caption set, containing about 8443 words. **COCO-Stuff** [22] is a more semantic-complete dataset for semantic segmentation. In total, it has 10,000 images sampled from MSCOCO training images, and annotations for 80 objects, 91 stuff, and 1 unknown background. Our DeepLab encoder is pre-trained on the MSCOCO 80-object dataset and then finetuned on this COCO-stuff dataset.

We use BLEU@N (B@1, B@2, B@3, B@4) [32], METEOR [33], and CIDEr [34] as the evaluation metrics. Their scores are calculated via the COCO captioning evaluation tool [31]. Among these metrics, CIDEr and METEOR have the highest correlations with human manual evaluation, and CIDEr is used for competition ranking in MSCOCO challenge[45]. Therefore, our performance comparison mainly focuses on CIDEr, METEOR, and BLEU@4.

B. Experiment Settings

This section describes the implementation details of our model and training.

FCN encoder: A elegantly designed DeepLab [29], designed based on VGG-16 [25], is used as the FCN encoder. The spatial visual features are extracted as the mean of four sets of spatial visual features with different Field-Of-View(FOV) from the outputs of the second last layer. Its dimension is 81×81, 1024d. The spatial semantic representations are extracted from the outputs of the final layer, which has dimension of 81×81, 1d.

LSTM encoder: A single-layer LSTM with the hidden size of 1024 is used in our model. The dimension of word embedding is 1024.

Attention model: The output size of the saliency pooling layer is set as 14×14, 1024d and 27×27, 1024d respectively. 14×14 is selected to make comparisons with the Soft-Attention model [9], and 27×27 is selected to demonstrate the improvement of fine-grained attention.

Training details: We use SGD to finetune DeepLab on the dataset COCO-stuff for 20 epochs by learning rate 0.001, momentum 0.9, and weight decay 0.0005. We use the Adam optimizer with a base learning rate of 0.0001 for LSTM language model. We also use weight decay 0.95 and dropout ratio 0.5. There is no finetune for FCN-encoder, as the Soft-Attention model [9] does not finetune CNN. The network is trained for up to 30 epochs with early stopping if the CIDEr [34] score had not improved over the last epochs. We use the beam size of 3 when sampling the caption for MSCOCO.

Compared methods: The proposed method is motivated by Soft-Attention [9] which is based on the idea of spatial visual attention, and the FCN encoder is designed based on the VGG model. Thus, it is basically essential to compare the performance of the proposed method against the Soft-Attention [9]. Moreover, our implementation does not carry out fine-tuning by re-training the visual encoder on the large captioning dataset like Adaptive-Attention [8], MSM [43] and ATT-FCN [11] did. Therefore, this paper also compares with other approaches DeepVS [28], NIC v1 [45], emb-gLSTM [46], m-RNN [47], SCA-VGG-1layer [7], and Hard-Attention [9] that all use the VGG-based encoder and has no fine-tuning training as our methods. This aims to ensure a fair comparison in order to show the performance boosted by fine-grain and semantic-guided attention.

C. Experiment 1 - Evaluation of Fine-Grained Grid-Wise Attention

The qualitative analysis is illustrated in Table III, visualizing the improved quality of attention maps and captions by increasing the attention resolution from 14×14 to 27×27. It is shown that attention at higher resolution can capture related regions more accurately. Taking image 2 and 6 as examples, the 27×27 attention model can attend to the "bat" regions accurately, whereas the 14×14 attention model attends to wrong regions. All blue-colour words, such as the "bear', "hydrant", "player", "boy" and "man", have more accurate attention maps in the 27×27 resolution. Moreover, stuff regions like the "street" and "stone" can also be correctly located. In the meantime, it is noticed that the overall quality of captions in higher attention resolution is improved, which is more meaningful. In Table II, the quantitative analysis is shown by Case 1 in attention resolution 14×14 (Soft-Attention method) and 27×27. The improvements are quite significant, boosting 0.013 in B@2 and B@3, 0.011 in B@4, and 0.064 in CIDEr. Moreover, the large improvement is proved by the comparison that the Case 1 in 27×27 beats most of methods in Table I.

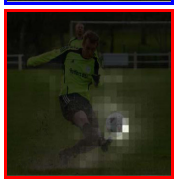
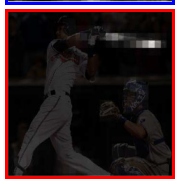
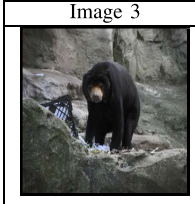
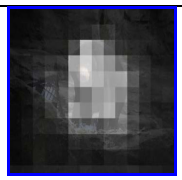
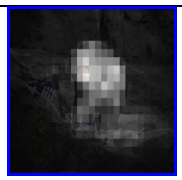
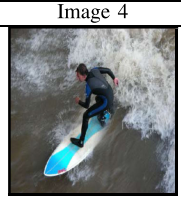
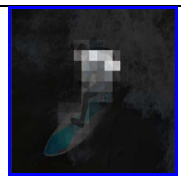
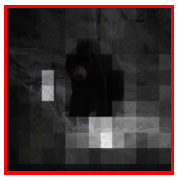
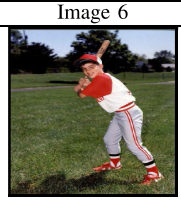
D. Experiment 2 - Evaluation of Semantic Guidance

The quantitative analysis in Table II further demonstrates the performance improvements contributed by three forms of semantic guidance. In this table, two fine-grained attention resolutions are adopted. For semantic guidance, there are four different cases according to the framework design in Fig. 2. **Case 1** - No semantic guidance is used, as the common average pooling scheme is adopted in the pooling layer. The aggregated visual features are fed into the rest layers without considering semantic information at all. This is the base-line scheme of the proposed framework, namely Soft-Attention [9]. **Case 2** - Saliency guidance is used, as the framework adopts the saliency pooling in the

TABLE II
PERFORMANCES OF OUR ABLATED MODELS ON MSCOCO TEST SPLIT ON ALL METRICS

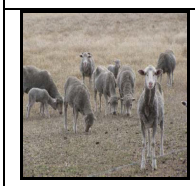
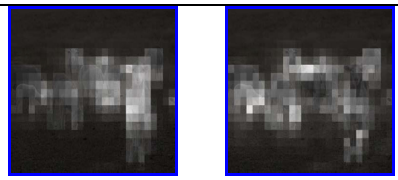
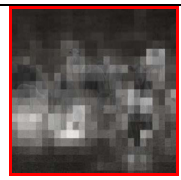
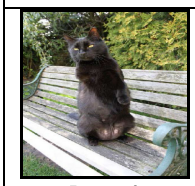
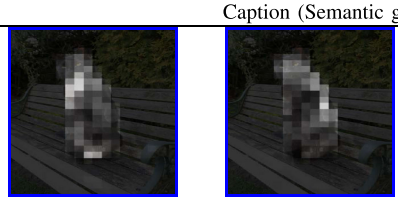
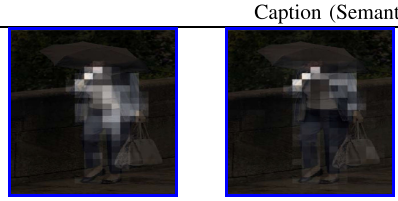
Attention Resolution	Semantic Guidance	B@1	B@2	B@3	B@4	METEOR	CIDEr
14 x 14 (Base-line Soft-Attention)	Case 1 - No Guidance	0.707	0.492	0.344	0.243	0.239	0.773
14 x 14	Case 2 - Saliency Guidance	0.703	0.501	0.354	0.251	0.240	0.827
14 x 14	Case 3 - Explicit Guidance	0.705	0.504	0.358	0.255	0.241	0.846
14 x 14	Case 4 - Joint Context Feature	0.708	0.507	0.360	0.257	0.242	0.846
27 x 27	Case 1 - No Guidance	0.707	0.505	0.357	0.254	0.241	0.837
27 x 27	Case 2 - Saliency Guidance	0.708	0.507	0.359	0.256	0.241	0.839
27 x 27	Case 3 - Explicit Guidance	0.709	0.508	0.361	0.258	0.242	0.844
27 x 27 (Our Best Model)	Case 4 - Joint Context Feature	**0.712**	**0.514**	**0.368**	**0.265**	**0.247**	**0.882**

TABLE III
QUALITATIVE ANALYSIS OF THE ADVANTAGES PROVIDED BY HIGHER ATTENTION RESOLUTION

Image 1	14 × 14 Attention	27 × 27 Attention	Image 2	14 × 14 Attention	27 × 27 Attention
					
					
Caption (14 × 14): A boy is playing baseball on a field. Caption (27 × 27): A young boy is playing soccer on a field.			Caption (14 × 14): A baseball player swinging a bat at a ball. Caption (27 × 27): A baseball player is swinging his bat.		
Image 3	14 × 14 Attention	27 × 27 Attention	Image 4	14 × 14 Attention	27 × 27 Attention
					
					
Caption (14 × 14): A black bear is standing on a rock. Caption (27 × 27): A large brown bear walking across a stone wall.			Caption (14 × 14): A man riding a wave on top of a surfboard. Caption (27 × 27): A man riding a surfboard on a wave in the ocean.		
Image 5	14 × 14 Attention	27 × 27 Attention	Image 6	14 × 14 Attention	27 × 27 Attention
					
					
Caption (14 × 14): A fire hydrant on the side of the street. Caption (27 × 27): A yellow fire hydrant sitting on the side of a street.			Caption (14 × 14): A man in a baseball uniform is holding a bat. Caption (27 × 27): A young boy in a baseball uniform holding a bat.		

Each image provides two attention maps corresponding to the two most meaningful nouns in the captions in two different attention resolutions. Attention maps with blue/red boundaries correspond to the words highlighted by blue/red respectively. (Best viewed in color.)

TABLE IV
QUALITATIVE ANALYSIS OF THE ADVANTAGES PROVIDED BY SEMANTIC GUIDANCE

Original Image	More Precise Attention Provided by Semantic Guidance				New Meaningful Words Discovered by Semantic Guidance
	No Semantic Guidance	Semantic Guidance	No Semantic Guidance	Semantic Guidance	
 Image 1					
	Caption (No semantic guidance): A herd of sheep standing on top of a grass covered field. Caption (Semantic guidance): A herd of sheep gazing on a dry grass field.				
 Image 2					
	Caption (No semantic guidance): A cat sitting on a bench next to a wooden bench. Caption (Semantic guidance): A cat sitting on a bench in the grass.				
 Image 3					
	Caption (No semantic guidance): A man in a suit and tie holding a umbrella. Caption (Semantic guidance): A woman walking down a street holding an umbrella.				

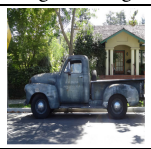
The analysis is carried out on 27×27 attention. The attention maps of different color boundaries (i.e. blue, green and red) correspond to the different words (highlighted by blue, green or red) in the captions. The words by red color are not discovered without semantic guidance, instead, which are captured by the proposed methods by using semantic guidance. (Best viewed in color.)

pooling layer rather than the common average pooling. However, the semantic aggregation for explicit semantic guidance is not fed into rest layers of the framework. **Case 3** - Explicit Guidance is additionally used, as both semantic grid-wise features and visual grid-wise features are fed into the attention distribution prediction layer. However, the aggregated semantic context feature is not fed into the last context computation and the LSTM language model. **Case 4** - The joint context feature is used, as all three forms of semantic guidances are fully integrated into our model, as illustrated in Fig. 2. In both grid resolutions, integrating saliency pooling, explicit guidance, and joint context feature one by one into our model can all lead to better performances steadily and consistently, although they are very modest. In 14×14, the improvements of full semantic guidance are obvious, boosting CIDEr by 0.073, B@4 by 0.014, and METEOR by 0.03. However, the improvements of full semantic guidance are quite modest in 27×27.

To demonstrate the modest improvements of full semantic guidance in 27×27 really make sense, the qualitative analysis is done as illustrated in Table IV. Obviously, the semantic guidance helps the model attend to large area objects/stuff, such as the "grass" in image 1, and the "bench" in image 2. Besides the improvement on the completeness and/or correctness of the attention maps, the semantic guidance can also discover new meanings to make the caption more meaningful. For image 2 and 3, the "grass" and "street" are not captured without using semantic guidance. After introducing semantic guidance in the proposed method, they are exposed in the new captions. In image 1, the word "gazing" is more precise than "standing", and attention has correctly focused on those regions where the sheep are eating grass. Therefore, adding semantic guidance can greatly increase the caption quality.

From Table I, it demonstrates that our model with attention resolution 27×27 and full semantic guidance has the best performance. All results are calculated on the test split of MSCOCO dataset. Our best model significantly outperforms all chosen state-of-the-art models over nearly all metrics. Compared with the base-line model Soft-Attention [9], our best model boosts CIDEr score by 0.109, B@4 score by 0.022, B@3 score by 0.024, and B@2 by 0.022. The METEOR and B@1 scores are slightly boosted by 0.008 and 0.005 respectively. Compared with B@1, larger improvements on B@4, B@3, and B@2 scores indicate that our model can better capture both grammatical properties and richer semantics because of higher resolution and introduced semantic guidance [34]. Moreover, these advantages are further strengthened by the large improvement on CIDEr scores, as it is a metric integrating all four B@N scores based on the human-consensus [34]. Our model has second-best B@1 and B@4 scores. The best B@1 score is obtained by the Hard-Attention [9], which has significantly lower scores on other metrics. The best B@4 score is obtained by the second-best model SCA-VGG-1layer [7], which has significantly lower CIDEr and METEOR scores than the proposed method. However, as CIDEr and METEOR metrics are more authoritative than B@N metric

TABLE V
QUALITATIVE ANALYSIS OF CAPTION DIVERSITY

Original Image	Semantic Guidance	Generated Top-3 Captions
	Yes	Top 1: a truck parked on the side of the road in front of a house. Top 2: a truck parked on the side of the road in a residential area. Top 3: a truck parked on the side of the road in front of a building.
	No	Top 1: a large white truck parked in a parking lot. Top 2: a large white truck parked next to a white truck. Top 3: a large white truck parked in front of a white truck.
	Yes	Top 1: a train on a track with a sky background. Top 2: a train on the tracks in the country side. Top 3: a train on the tracks in the middle of a rural area.
	No	Top 1: a blue and yellow train traveling down train tracks. Top 2: a blue and white train traveling down train tracks. Top 3: a blue and yellow train traveling down the tracks.
	Yes	Top 1: a group of people riding on the back of a horse drawn carriage. Top 2: a group of people riding on the back of a carriage. Top 3: a group of horses pulling a carriage with people in it.
	No	Top 1: a group of people riding on the backs of horses. Top 2: a group of people riding on the back of a horse. Top 3: a large group of people riding on the back of a horse.
	Yes	Top 1: a yellow fire hydrant on sidewalk next to parked cars. Top 2: a yellow fire hydrant on sidewalk next to cars and buildings. Top 3: a yellow fire hydrant on sidewalk next to cars and sidewalk.
	No	Top 1: a yellow fire hydrant sitting on the side of a street. Top 2: a yellow fire hydrant on the side of a street. Top 3: a yellow fire hydrant on the side of the street.
	Yes	Top 1: a clock on a pole on a city street. Top 2: a clock on a pole in front of a tree. Top 3: a clock on a pole in front of a building.
	No	Top 1: a large white clock on a pole. Top 2: a white and green clock on a pole. Top 3: a large clock on a pole on a street.
	Yes	Top 1: a man sitting on a bench in a park. Top 2: a man sitting on a bench with a dog on a leash. Top 3: a man sitting on a bench with a dog in a park.
	No	Top 1: a couple of people sitting on a bench in a park. Top 2: a couple of people sitting on a bench in the park. Top 3: a couple of people sitting on a bench in the woods.
	Yes	Top 1: a skate boarder doing a trick in the air. Top 2: a skate boarder doing a trick on a cement wall. Top 3: a skate boarder doing a trick on a cement block.
	No	Top 1: a skate boarder doing tricks on a skateboard ramp. Top 2: a skate boarder doing a trick on a skateboard ramp. Top 3: a skate boarder doing a trick on a skate board.

[48], our model still is the best. Although the improvements in some metrics are modest, the most authoritative CIDEr has significant boosts.

E. Experiment 3 - Evaluation of Caption Diversity

Semantic guidance can also enhance the caption diversity. In Table V, top-3 captions are generated for models with and without semantic guidance. Besides higher accuracy, the captions have high diversity thanks to more meaningful words. For image 1, the "parked" location of the "truck" has three different correct descriptions: "in front of a house", "in a residential area", "in front of a building". For image 2, the surroundings of "train" are described with "a sky background", "the country side", "a rural area". The third caption for image 3 has a totally different structure, compared with the first one. In contrast, those top-3 captions generated without semantic guidance have low diversity and even mistakes. Therefore, powered by semantic segmentation, our attention mechanism can generate more diversified captions than the traditional one, which is powered by saliency.

TABLE VI
PROCESSING TIME OF OUR ABLATED MODELS

Attention Resolution	Semantic Guidance	Processing Time (ms)
14 x 14	Case 1 & 2	21.1
14 x 14	Case 3	32.1
14 x 14	Case 4	39.7
27 x 27	Case 1 & 2	44.1
27 x 27	Case 3	67.7
27 x 27	Case 4	84.3

F. Computational Costs

We use Nvidia GTX1080Ti to train and test all our models. Finetuning the DeepLab model takes 48 hours. For our best model with 27×27 resolution and full semantic guidance, it takes around 240 hours for training 30 epochs. At the testing phase, the per-image processing time on all our models is displayed in Table VI. Case 1 and Case 2 have almost the same computation costs, as they are only different in feature pooling and have the same captioning model. All processing times are below 100 ms. Increasing the attention resolution from 14×14 to 27×27 doubles the processing time in all cases. The

comparison between Case 1&2 and Case 4 shows that adding full attention guidance nearly doubles the processing time.

V. Conclusion

In this paper, we proposed a fine-grained and semantic-guided attention mechanism over a novel end-to-end FCN-LSTM framework for image captioning for the first time. Our model achieves state-of-the-art performance on the MSCOCO dataset compared with models using the VGG-based encoder. Moreover, the framework of our model can be easily adapted to all approaches that are based on soft-attention. The results show that our model has huge potential for a comprehensive attention method on the abstract visual relationship. Moreover, our framework could have a broad application in other tasks, like Image QA.

References

[1] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, pp. 32–73, 2016.

[2] Y. Yao *et al.*, "Extracting multiple visual senses for web learning," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 184–196, Jan. 2019.

[3] L. Pang, S. Zhu, and C. W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

[4] L. Baraldi, C. Grana, and R. Cucchiara, "Recognizing and presenting the storytelling video structure with deep multimodal networks," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 955–968, May 2017.

[5] J. Li *et al.*, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.

[6] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "Fine-grained and semantic-guided visual attention for image captioning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Lake Tahoe, NV, USA, 2018, pp. 1709–1717.

[7] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp 6298–6306.

[8] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp 3242–3250.

[9] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[10] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 21–29.

[11] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4651–4659.

[12] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3135–3144.

[13] L. Yang, K. Tang, J. Yang, and L. J. Li, "Dense captioning with joint inference and visual context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1978–1987.

[14] S. Venugopalan *et al.*, "Captioning images with diverse objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1170–1178.

[15] C. C. Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6432–6440.

[16] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating copying mechanism in image captioning for learning novel objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jul. 2017, pp. 5263–5271.

[17] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 873–881.

[18] H. R. Tavakoliy, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2506–2515.

[19] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, "GLA: Global-local attention for image description," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 726–737, Mar. 2018.

[20] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[21] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.

[22] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1209–1218.

[23] Y. Yao *et al.*, "Exploiting web images for dataset construction: A domain robust approach," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1771–1784, Aug. 2017.

[24] T. Zhang, C. Xu, G. Zhu, S. Liu, and H. Lu, "A generic framework for video annotation via semi-supervised learning," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1206–1219, Aug. 2012.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[27] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4565–4574.

[28] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[29] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 824–848, Apr. 2017.

[30] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[31] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[33] S. Banerjee and A. Lavie, "Meteor: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Meas. Mach. Transl. Summarization*, 2005, pp. 65–72.

[34] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[36] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 940–1000.

[37] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2422–2431.

[38] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1473–1482.

[39] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 289–297.

[40] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1–9.

[41] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v. d. Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 203–212.

[42] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen, "Review networks for caption generation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2361–2369.

[43] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 4904–4912.

[44] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Watch what you just said: Image captioning with text-conditional attention," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 305–313.

[45] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.
[46] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the longshort term memory model for image caption generation," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 2407–2415.
[47] J. Mao *et al.*, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1000–1020.
[48] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017. [Online]. Available: doi.ieeecomputersociety.org/10.1109/TPAMI.2016.2587640

Zongjian Zhang received the B.S. and M.S. degrees from Beijing Jiaotong University, Beijing, China, in 2007 and 2011, respectively. He is currently working toward the Ph.D. degree at the University of Technology Sydney, Ultimo, NSW, Australia. His research interests include computer vision, image captioning, deep learning, scene understanding, image retrieval, pattern recognition, and data analysis.

Qiang Wu received the B.Eng. and M.Eng. degrees from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree from the University of Technology Sydney, Ultimo, NSW, Australia, in 2004.

He is currently an Associate Professor and a Core Member of the Global Big Data Technologies Centre, University of Technology Sydney. His research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. His application fields where the research outcomes are applied span over video security surveillance, biometrics, video data analysis, and human-computer interaction. His research outcomes have been published in many premier international conferences, including European Conference on Computer Vision, Computer Vision and Pattern Recognition, International Conference on Image Processing, International Conference on Pattern Recognition, and Workshop on Applications of Computer Vision, and the major international journals, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *Public Relations*, *Physical Review Letters*, and *Signal Processing*.

Yang Wang received the Ph.D. degree in computer science from the National University of Singapore, Singapore, in 2004. He is currently a Principal Researcher and a Team Leader with Analytics Research Group, Data61, Commonwealth Scientific and Industrial Research Organisation, Eveleigh, NSW, Australia. Before Joining Data61 (formerly NICTA) in 2006, he was with the Institute for Infocomm Research, Rensselaer Polytechnic Institute, and Nanyang Technological University. He has authored or coauthored more than 100 international conference and journal papers and filed 5 patents. His research interests include machine learning and information fusion techniques, and their applications to asset management, intelligent infrastructure, cognitive and emotive computing, medical imaging, image and video analysis.

Fang Chen is the Executive Director Data Science and a Distinguished Professor with the University of Technology Sydney, Ultimo, NSW, Australia. She is a Thought Leader in AI and data science. She has created many world-class AI innovations while working with Beijing Jiaotong University, Intel, Motorola, NICTA, and CSIRO, and helped governments and industries utilising data and significantly increasing productivity, safety, and customer satisfaction. Through impactful successes, she gained many recognitions, such as the ITS Australia National Award 2014 and 2015, and NSW iAwards 2017. She is the NSW Water Professional of the Year 2016, the National and NSW Research, and the Innovation Award by Australian Water association. She was the recipient of the "Brian Shackle Award" 2017 for the most outstanding contribution with international impact in the field of human interaction with computers and information technology. She is the recipient of the Oscar Prize in Australian science-Australian Museum Eureka Prize 2018 for Excellence in Data Science. She has 280 publications and 30 patents in 8 countries.