Size-Invariant Attention Accuracy Metric for Image Captioning with High-Resolution Residual Attention

Zongjian Zhang, Qiang Wu
School of Electrical and Data Engineering
University of Technology Sydney
Sydney, Australia
Zongjian.Zhang@student.uts.edu.au, Qiang.Wu@uts.edu.au

Yang Wang, Fang Chen
Data61
CSIRO
Sydney, Australia
{Yang.Wang, Fang.Chen}@data61.csiro.au

***Abstract*—Spatial visual attention mechanisms have achieved significant performance improvements for image captioning. To quantitatively evaluate the performances of attention mechanisms, the "attention correctness" metric has been proposed to calculate the sum of attention weights generated for ground truth regions. However, this metric cannot consistently measure the attention accuracy among the element regions with large size variance. Moreover, its evaluations are inconsistent with captioning performances across different fine-grained attention resolutions. To address these problems, this paper proposes a size-invariant evaluation metric by normalizing the "attention correctness" metric with the size percentage of the attended region. To demonstrate the efficiency of our size-invariant metric, this paper further proposes a high-resolution residual attention model that uses RefineNet as the Fully Convolutional Network (FCN) encoder. By using the COCO-Stuff dataset, we can achieve pixel-level evaluations on both object and "stuff" regions. We use our metric to evaluate the proposed attention model across four high fine-grained resolutions (i.e., 27×27, 40×40, 60×60, 80×80). The results demonstrate that, compared with the "attention correctness" metric, our size-invariant metric is more consistent with the captioning performances and is more efficient for evaluating the attention accuracy.**

***Index Terms*—image captioning, size-invariant attention correctness, high-resolution residual attention, attention accuracy, quantitative evaluation metric**

I. Introduction

Image captioning automatically generates captions based on a comprehensive understanding of the real-world scene [1]–[16]. It is a challenging multi-modal scene understanding task, requiring a deep understanding of two completely different types of media data, i.e., vision and language. Particularly, the joint modelling of vision and language is the key challenging part. As a method to address this issue, the spatial visual attention mechanism has attracted a great deal of research interests, leading to significant performance improvement for image captioning [1]–[9], [14], [15]. Generally, the spatial visual attention mechanism has two roles in bridging the image encoder and the caption decoder together. The first one is to map language feature and visual feature into a shared feature space for joint learning. The second one is to seek the semantic alignment between words/phrases and relevant visual regions for extracting a fine-grained visual context feature. In this way, an accurate attention model can extract a precise visual information for the language decoder to generate high-quality captions.

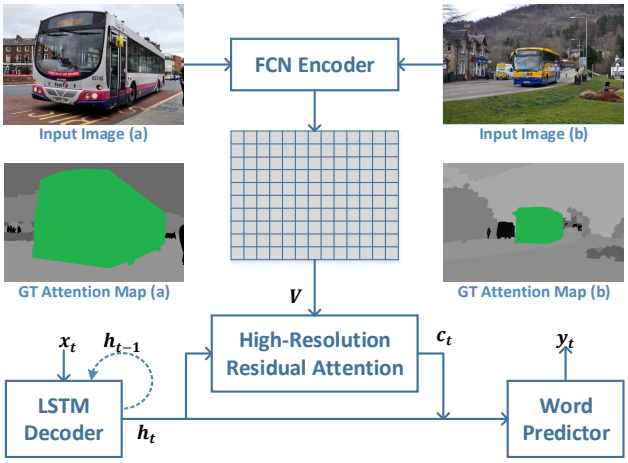

Fig. 1. The overview of our proposed framework. Image (a) and image (b) are two input examples having the "bus" element of different size. Ground Truth (GT) attention maps are provided for both images. The "bus" regions are highlighed with green color.

However, a proper metric or approach for the quantitative evaluation on spatial attention mechanism had always been absent until [15] proposed the first evaluation metric "attention correctness". This metric can quantitively evaluate the attention accuracy by measuring the consistency between the generated attention map and human-annotated region mask for the relevant caption word. Specifically, the "attention correctness" metric is defined as the sum of attention weights generated for ground truth regions. Using this metric, [15] has successfully evaluated the improvement of attention accuracy brought by adding explicit supervision on the attention model.

However, this "attention correctness" metric has three limitations. 1) It cannot consistently measure the attention accuracy among elements of the same class but having different region sizes in the image. Specifically, the numeric value of this metric tends to be larger for elements of larger size, and vice versa. Therefore, for the element with large size variance, this metric would have large evaluation error. Fig. 1 uses a simple example to illustrate this problem. Image (a) and image

(b) have the same size, but the "bus" element (occupying 83% area) in the image (a) is much larger than the one (occupying 15% area) in the image (b). For simplicity, lets assume the attention model can generate the same uniform weight map for all regions in the image during the caption decoding process. Accordingly, the "attention correctness" value for the larger "bus" in the image (a) is 0.83, which is much larger than the value 0.15 for the smaller one in the image (b). Different evaluation scores are obtained by the same attention model just because of the size difference. However, such large numeric difference is conflicting with the same attention accuracy. In other words, the numeric value of "attention correctness" is not equivalent to the ability of attention model across different element sizes. 2) Moreover, it shows the limited ability to evaluate attention models in different fine-grained resolutions either, as our experiment results demonstrate that it cannot reflect the improvement boosted by higher fine-grained resolution. 3) Moreover, [15] uses the imprecise bounding box labels to evaluate the attention accuracy on object regions via the Flicker30 dataset, which leads to inaccurate evaluations.

To overcome these limitations, we propose an improved evaluation metric "normalised attention correctness" by normalizing the "attention correctness" metric with the size percentage of the attended region. By calculating the ratio of attention correctness to the size percentage, our metric can capture the density of attention weights in the ground truth region R_{GT}. This size-invariant feature enables our metric to perform consistent evaluations on elements across different sizes and attention resolutions. In particular, it can be applied to the fine-grained attention model based on the FCN(Fully Convolutional Network)-LSTM(Long Short-Term Memory) framework proposed by [8].

To demonstrate the efficiency of our novel metric, we propose a novel high-resolution residual attention model by applying the residual attention model [3] into the FCN-LSTM framework (Fig. 1). To enable high resolutions, we design the FCN encoder based on the RefineNet [17] model, which is the state-of-the-art model for high-resolution semantic segmentation. Specifically, we utilize four high fine-grained resolutions (i.e., 27×27, 40×40, 60×60, 80×80). Furthermore, by using the COCO-Stuff [18] dataset, we can achieve pixel-level evaluations on both object and "stuff" regions. Object regions (e.g., bus, car, bike) have well-defined shapes and identifiable parts, whereas stuff regions (e.g., sky, grass, water) are amorphous backgrounds with a strong texture. Moreover, the size variance of stuff regions is much larger than that of object regions. Therefore, the size-invariant merit of our metric can be fully demonstrated by evaluations on stuff regions.

Therefore, targeting the novel high-resolution residual attention model, we use the novel "normalised attention correctness" metric to perform pixel-level evaluations of attention accuracy improved by residual attention and fine-grained resolution respectively. Compared with the "attention correctness", our metric is more rational and has higher consistency with the captioning performances. To sum up, this paper has four major contributions:

1) We propose a high-resolution residual attention model for image captioning using the RefineNet [17] (Fig. 1) as the FCN encoder. Compared with [3], our model is based on the FCN-LSTM framework and supports fine-grained attention in multiple high-resolutions. Our attention model is the first one using such high resolution 80×80.
2) We propose a size-invariant "normalized attention correctness" metric that can rationally and consistently evaluate the attention accuracy across different fine-grained resolutions. Besides the novel metric, we also use the COCO-Stuff [18] dataset to achieve pixel-level evaluations on both object and stuff regions.
3) We perform a detailed analysis of the improvement, for the first time, that the residual attention model has contributed to the captioning performance, by jointly analyzing the quantitative evaluations on both attention model and captioning model.
4) We further demonstrate that our "normalized attention correctness" metric can more effectively evaluate the improvements that higher fine-grained resolutions (i.e., 27×27, 40×40, 60×60, 80×80) have contributed to both attention accuracy and captioning performance.

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 methodology 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

Most state-of-the-art spatial visual attention models are based on the encoder-decoder framework in an end-to-end trainable way [1]–[5], [8], [14], [15], [19], [20]. The attention mechanism serves as an agent between the image encoder and the caption decoder. In generating each word, the mechanism makes joint inferences and adaptively attends to those semantically relevant image regions by generating a distinct attention weight for each region. Based on this weight map, a visual context feature is summarized through the weighted sum of all-region features encoded by the image encoder. Then, it is sent into the caption decoder for language inference and generation.

Specifically, an accurate attention model should achieve not only spatial region accuracy but also semantic weight accuracy. 1) The spatial region accuracy means that the attention model can capture relevant visual regions at a fine-grained level. It is generally contributed by a powerful image encoder, which has an accurate way of capturing image regions and extracting features. 2) The semantic weight accuracy means that the attention model should accurately assign importance to relevant regions via a spatial weight map. It needs a well-designed structure that can strongly seek the semantic alignment between visual region features and language state features. 3) However, it is not enough to use captioning performances to compare the performances of attention models. The quantitative comparison among all attention models needs

an evaluation metric or approach. Therefore, all related works are discussed based on below three categories:

A. Spatial Region Accuracy Based on Different Image Encoders

Grid-level attention based on CNN encoder. This attention model splits the image into equally sized grid regions based on the grid structure of CNN's last convolutional layer. [4] firstly proposed a 14×14 grid-resolution soft attention model based on VGG19 for image captioning. [3] 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, [2] proposed a novel channel-wise and multi-layer spatial attention model, which additionally attend to related channels among the multi-layer feature maps. However, all these attention models have a fixed low grid resolution, which is difficult to convert to high resolution to capture fine-grained attention regions. Specifically, a large object is usually split into different grid regions, and one grid region usually contains portions of several objects. This damages the semantic correspondence of region features.

Object-level attention based on R-CNN encoder. This attention model can capture object-level regions via the bounding box. [20] 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, [19] 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. However, the bounding box is still not fine-grained enough due to the irregular shape of the object boundary. It inevitably includes some portions of other objects at its corners. Moreover, it ignores amorphous stuff regions [18] that does not have a well-defined shape (e.g., sky, grass, water).

Pixel-level attention based on FCN encoder. This attention model segments the image into all semantic regions at pixel level. As the only pioneer, [20] proposed the fine-grained and semantic-guided attention model based on a novel FCN-LSTM framework. However, constrained by the GPU computation power, the practical attention resolution was only increased to 27×27. Fine-grained attention in higher resolutions are not studied so far. In this paper, our attention model uses the RefineNet [17] as the FCN encoder, which enables the fine-grained attention model to support a high resolution up to 80×80.

B. Semantic Weight Accuracy Based on Different Structures

To improve the structure for semantic alignment, most attention models focus on how to encode the language state feature. Soft-attention mechanism [4] is the first spatial visual attention model proposed for image captioning. It is designed as a two-layer perception based on the CNN-LSTM framework, which is used by most state-of-the-art attention models [3], [8], [9], [15], [19], [20]. They use the hidden state h_{t-1} of LSTM decoder at previous time $t-1$ as the language state feature. For each time t, it predicts attention weights by aligning each visual region feature with the hidden states h_{t-1} at previous timestep. This model is a weights predictor because the hidden states h_{t-1} only contains the knowledge of previously generated words. The underlying idea is to use such previous knowledge until $t-1$ to attend to relevant visual regions that are responsible for generating a word at the next time t. However, the knowledge in h_{t-1} is limited for inferring relevant regions for a future word.

To resolve this limitation, [3] proposed a novel state-of-the-art spatial attention model by integrating the idea of the residual network [21]. Different from the soft-attention model, it uses the hidden states h_t at current timestep t as the language state feature to generate the attention weights. Here, h_t is generated by an independent LSTM language decoder and represents the language knowledge until the current time t. Then, the summarized context feature c_t serves as the residual visual information to rectify the language information included in h_t for generating the next word y_t. As a visual residual model, this attention model plays a role of "corrector" and rectifies the output of the LSTM language model to generate an accurate caption. However, no direct evaluations are provided to demonstrate the improved accuracy of these attention models in either qualitative or quantitative manner.

C. Quantitative Evaluation on Attention Accuracy

So far, only [15] proposed a metric "attention correctness to quantitatively evaluate the accuracy of attention model. However, it only studied whether adding explicit supervision on attention model can boost the attention accuracy and captioning performances based on bounding-box labels. More importantly, its metric is limited for consistently evaluating the attention accuracy across varying object size and attention resolutions. In this paper, we improve this metric with size normalization and propose a novel metric "normalized attention correctness for evaluating our fine-grained residual attention model based on pixel-level labels.

To the best of our knowledge, our model is the first work to propose a high-resolution residual attention up to 80×80 and perform quantitative evaluations of attention accuracy with a size-invariant metric across different resolutions.

III. METHOD

We firstly describe our captioning model with high-resolution residual attention model in Section A, and then further introduce our novel "normalized acctention correctness" metric in Section B.

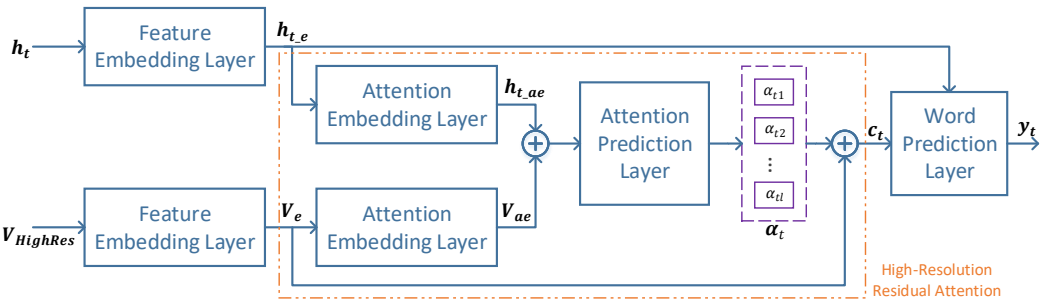

Fig. 2. The detailed structure of our high-resolution residual attention model. (Best viewed in color.)

A. Image Captioning Model with High-Resolution Residual attention

Our image captioning model is based on the FCN-LSTM framework. Given an image and its corresponding caption, our image captioning 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 $\mathbf{y} = \{\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_t}\}$ 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:

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

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 captioning model consists of three components: the FCN encoder, the LSTM decoder, and the high-resolution residual attention model (Fig. 1). It firstly uses the FCN encoder to extract spatial visual features from the image at the pixel level. Then, the high-resolution residual attention model summarizes all relevant region features as the visual context feature c_t 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 model employs the FCN to directly extract visual feature for each different pixel in the image. Specifically, the N $\times$ N sized image $\boldsymbol{I}$ can be represented by the spatial visual features:

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

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. Our FCN encoder uses the RefineNet [17], which is a multi-path refinement network based on the encoder-decoder structure. It is a state-of-the-art model for high-resolution semantic segmentation. Therefore, we use the output of the last decoder block RefineNet-1 as the pixel-level features.

To allow joint modelling of the spatial visual feature V and the language state feature h_t under shared feature space, the spatial visual features V is mapped to the embedded visual feature V_e through below feature embedding layer:

$$V_e = ReLu(W_{ve}V + b_{ve}) \tag{4}$$

where $ReLu$ stands for the rectified linear unit.

2) LSTM Decoder: We use the LSTM as the language decoder to model each conditional probability in (2). At time t, the previous conditional variable-length word sequence $\{y_1, y_2, \cdots, y_{t-1}\}$ and image I are represented by the fixed-length hidden state h_t of LSTM as following:

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

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

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. Note that the LSTM is an independent language decoder that does not use the visual context feature c_t summarized by our attention model. The hidden state h_t is used as the language state feature for attention model to jointly model both vision and language.

To work in the same shared space with the embedded visual feature V_e, the language state feature h_t is mapped to the embedded language state feature h_{t_e} through below feature embedding layer:

$$h_{t\_e} = ReLu(W_{he}h_t + b_{he}) \tag{7}$$

Finally, the probability of generating word y_t at time t is modeled based on the embedded language state feature h_{t_e} and the visual context feature c_t as follow:

$$p(y_t|y_{t-1}, \cdots, y_1, \boldsymbol{I}) = \text{softmax}(W_p Z_f + b_p) \tag{8}$$

$$Z_f = \tanh(W_{hc}(h_{t\_e} + c_t) + b_{hc}) \tag{9}$$

3) High-Resolution Residual Attention Model:

High-Resolution Attention. Based on the FCN-LSTM framework, our attention model can employ the RefineNet-based FCN encoder to capture semantic regions at the pixel level. Specifically, the FCN encoder can extract visual features for all pixels, which enables our attention model to attend to relevant regions with pixel-level accuracy. Practically, constrained by limited GPU memory and computation power, both RefineNet and attention model cannot achieve the full resolution of the input image. However, by designing an efficient structure, our attention model can still achieve the super-pixel-level accuracy at a relatively low resolution. This is equivalent to a fine-grained grid-wise resolution, where the super-pixel is a small patch, which can distinguish the object/stuff boundary more precisely.

Residual Attention. As an independent language model, the LSTM decoder output the hidden state h_t to represent the language state. The word prediction layer can generate rational word using only language state feature h_t. However, the residual model can accurately attend to relevant regions and summarize their visual features as the context feature c_t. This context feature can function as a residual visual feature and rectify the error of h_t. In this way, our high-resolution residual attention model can help the word prediction layer to generate high-quality captions. The detailed structure is illustrated in Fig. 2.

Specifically, our attention model aims to generate an accurate attention weight map based on the embedded spatial visual feature V_e and embedded language state feature h_{t_e}, both of which share the same feature space. The model is defined as below:

$$c_t = f_{att}(h_{t\_e}, V_e) \tag{10}$$

Specifically, this attention generation model is specifically designed as a two-layer perception. The first layer is mainly responsible for features embedding (11 and 12) and fusion (13). The overall process can be illustrated in Fig. 2. The fused feature z_t is then fed into the second layer with a softmax function to generate the attention weights over k grid regions.

$$V_{ae} = W_{v\_ae}(V_e + b_{V\_ae}) \tag{11}$$

$$h_{ae} = W_{h\_ae}(h_{t\_e} + b_{h\_ae}) \tag{12}$$

$$z_t = \tanh(V_{ae} + h_{ae}\hat{\mathbf{1}}) \tag{13}$$

$$\boldsymbol{\alpha_t} = \text{softmax}(W_{att} z_t + b_{att}) \tag{14}$$

where α_{ti} represents the attention distribution for the region location $i = 1, 2, ..., k$ at the time t.

Based on the attention weight map α_t, the visual context feature c_t is computed as the weighted sum of embedded visual features V_e:

$$c_t = \boldsymbol{\alpha_t} \cdot V_e = \sum_{i=1}^{k} \alpha_{ti} v_{ei} \tag{15}$$

B. "Normalized Attention Correctness" Metric

For the evaluation on attention models in different fine-grained resolutions, the attention map is rescaled back to the original resolution of input image with needed weight normalization. At time t, the binary mask of generated word is extracted from the ground-truth semantic label map as the ground truth region for calculating the metric. The "attention correctness" metric proposed by [15] is defined as:

$$AC_t = \sum_{i \in R_{GT}} \alpha_{ti} \tag{16}$$

where α_{ti} is the attention weight at location i at time t, and R_{GT} is the ground truth attention region for the generated word y_t. The metric value ranges from 0 to 1. The value 0 means that the attention model is not working at all and the value 1 indicates complete correctness. Between 0 and 1, there is no consistent value point indicating that the attention model start to function normally.

we propose an improved evaluation metric by normalizing the "attention correctness" metric with the size percentage of the attended element in the image. By calculating the ratio of "attention correctness" to size percentage, our metric can capture the density of attention weights in the ground truth region R_{GT} as follow:

$$AC_{Nt} = \frac{A_{R_{GT}}}{A_{Img}} \sum_{i \in R_{GT}} \alpha_{ti} \tag{17}$$

where $A_{R_{GT}}$ is the area of the ground truth region and A_{Img} is the image area. The metric value ranges from 0 to ∞ with the value 1 as the working point. A value larger than 1 means that the attention model is working and vice versa.

IV. Experiment

This section firstly specifies datasets, evaluation metrics, and experiment settings. Then, we discuss the results of two experiments that evaluate image captioning performances and the efficiency of "normalized attention correctness". They can demonstrate that overall improvements of captioning performance are more consistent with this size-invariant attention accuracy metric.

A. Datasets and Metrics

Our experiments use three datasets. **MSCOCO** [22] is the largest dataset for image captioning, with 82,783 training images, 40,504 validation images, and 40,775 testing images. This dataset is used for training and testing our captioning model. For the offline evaluation, we use the same data split as [4], [6], 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 least 5 times in the training caption set, containing about 8443 words. **COCO-Stuff** [18] is a semantic-complete dataset for semantic segmentation. In total, it provides full annotations for all MSCOCO images, including 80 objects, 91 stuff, and 1 unknown background. This dataset is used for the quantitative evaluation on attention models. Our

TABLE I
PERFORMANCES OF DIFFERENT MODELS ON MSCOCO TEST SPLIT ON ALL METRICS

Attention Model	Fine-Grained Resolution	B@1	B@2	B@3	B@4	METEOR	CIDEr
Soft-Attention	27×27	0.688	0.482	0.336	0.238	0.232	0.767
	40×40	0.692	0.485	0.341	0.241	0.232	0.772
	60×60	0.692	0.486	0.342	0.243	0.233	0.778
	80×80	0.693	0.489	0.346	0.247	0.235	0.791
Residual Attention	27×27	0.698	0.496	0.357	0.261	0.236	0.810
	40×40	0.703	0.500	0.360	0.264	0.237	0.815
	60×60	0.703	0.500	0.361	0.265	0.239	0.821
	80×80	**0.706**	**0.505**	**0.366**	**0.269**	**0.241**	**0.838**

Fig. 3. Qualitative Analysis of the Advantages Provided by Higher-Resolution Residual Attention. The attention maps of differently colored boundaries (i.e., blue, green, orange and red) correspond to the different words (highlighted by blue, green, orange, or red) in the captions. The three attention maps in the first row are generated by the soft-attention model in resolution 27×27. The four in the second row are generated by our residual attention mode in 80×80. The attention-word pairs highlighted blue, green and orange colors are generated by both attention models. The red color pair is only generated by our model. (Best viewed in color and high resolution.)

RefineNet encoder is pre-trained on the **ADE20K** dataset [23] to extract high-quality visual feature, which is a dataset for scene parsing and includes 150 scene classes.

For **image captioning**, we use BLEU@N (B@1, B@2, B@3, B@4) [24], METEOR [25], and CIDEr [26] as the evaluation metrics. Our performance comparison mainly focuses on CIDEr, METEOR, and BLEU@4. For **attention accuracy**, we use "Attention Correctness" and our novel "normalized attention correctness" as the evalution metrics to analyze our high-resolution residual attention model.

B. Experiment Settings

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

Captioning model: The high-resolution RefineNet [17], designed based on the ResNet-152, is used as the FCN encoder. The output of the last decoder block RefineNet-1 is extracted as the pixel-level spatial visual features with the dimension of 256d. The resolution of the extracted feature map is 1/4 of the original image resolution and is generally larger than 100×100. It is still too high for the 15G memory of Nvidia P5000 GPU. Therefore, we down-scale our highest resolution to 80×80 by average pooling. A single-layer LSTM with the hidden size of 1024 is used as the LSTM decoder in our model. The dimension of word embedding or visual feature is 512.

Attention model: We use four fine-grained attention resolutions (i.e., 27×27, 40×40, 60×60, 80×80) to demonstrate the improvement of captioning performances that is contributed by increasing fine-grained resolution based on both soft-attention model and our residual attention model. In addition, the improvements of attention accuracy is quantitatively evaluated via two metrics: "attention correctness" and our novel "normalized attention correctness". Furthermore, attention accuracy is analyzed for object and stuff regions.

Training details: We use the Adam optimizer with a base learning rate of 0.0001 and dropout ratio 0.5 for training our image captioning model. The network is trained for up to 30 epochs with early stopping if the CIDEr [26] score had not improved over the last 4 epochs. We use the beam size of 3 when sampling the caption for MSCOCO.

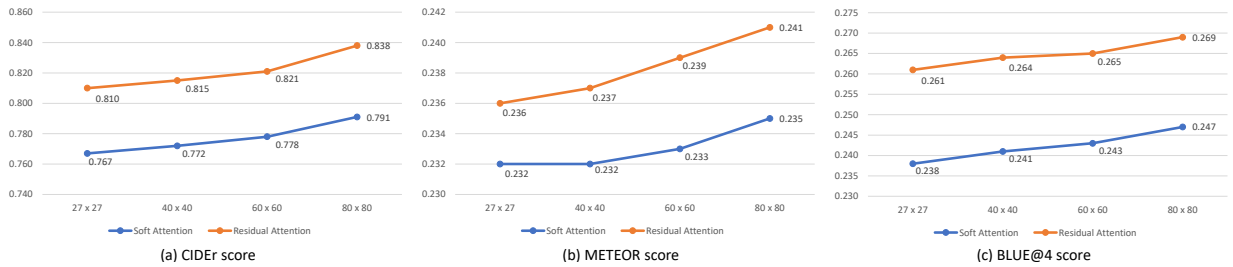

Fig. 4. The comparisons of captioning performances between soft-attention model and our residual attention model across four fine-grained resolutions (i.e., 27×27, 40×40, 60×60, 80×80). (Best viewed in color.)

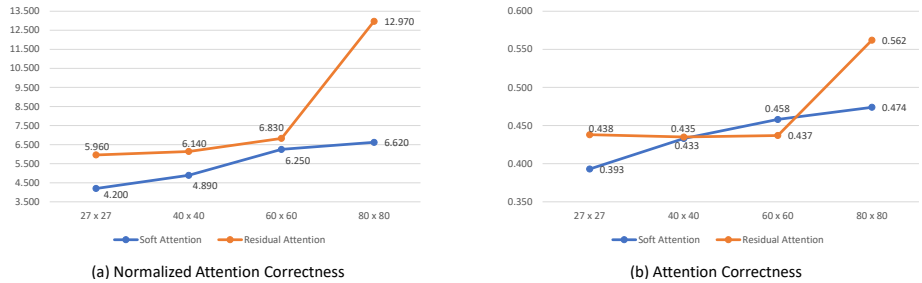

Fig. 5. The evaluation effects of "normalized attention correctness" and "attention correctness" between soft-attention model and our residual attention model across four fine-grained resolutions (i.e., 27×27, 40×40, 60×60, 80×80). (Best viewed in color.)

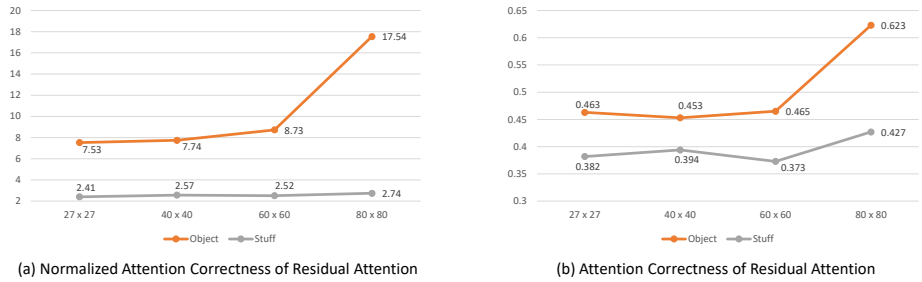

Fig. 6. The comparisons of attention accuracy between soft-attention model and our residual attention model across four fine-grained resolutions (i.e., 27×27, 40×40, 60×60, 80×80). (Best viewed in color.)

C. Experiment-1: Evaluation on Image Captioning Performance by Increasing Resolution and Using New Residual Attention Model

This experiment aims to evaluate the captioning performance that is improved by two dimensions: 1) high fine-grained resolution and 2) novel residual attention model. The captioning performances of both soft-attention model and residual attention model are all evaluated in four fine-grained resolutions (i.e., 27×27, 40×40, 60×60, 80×80). From Table I, all metric scores consistently demonstrate that 1) higher fine-grained resolution lead to a better captioning performance for both attention models and 2) the residual attention model is significantly better than the soft-attention model in all fine-grained resolutions. By comparing the performances of the soft-attention model in resolution 80×80 and the residual attention model in 27×27, we can see that the contribution of residual attention is larger than that of higher resolution. Also, these improvements are clearly shown by Fig. 4. The residual attention model in resolution 80×80 has the best performance. This ablated study uses captioning results to demonstrate that the high-resolution residual attention is a better attention model.

The qualitative analysis of attention maps and captions are shown in Fig. 3. For words "sheep" and "field", our residual attention model (80×80) can accurately attend to relevant regions , particularly in boundary regions. For word "standing", our model can attend to those leg regions that are most relevant to the action "standing". Moreover, our model can attend to regions of "grass" field.

D. Experiment-2: Evaluation on the Efficiency of the Proposed "normalized attention correctness" Metric for Measuring Attention Accuracy

This experiment aims to evaluate the efficiency of our "normalized attention correctness" metric for measuring the extent, to which attention accuracy is improved by 1) high fine-grained resolution and 2) novel residual attention model.

Our intuitive expectation is that the region accuracy of attention model would be increased by high fine-grained resolution

because of that boundary regions can be distinguished more precisely. At the object/stuff boundary, the grid patch contains pixels of both this object/stuff and its neighbors (including other object/stuff). Higher attention resolution means smaller grid patch, which could decrease noisy pixels from neighbor objects/stuff. As object is usually small in size and have identified parts, the improvement to boundary regions would be significant. For stuff with large size and repeated texture, this improvement would be minor.

Firstly, we use both metrics to evaluate the accuracy performances of residual attention and soft-attention models across four fine-grained resolutions. In Fig. 5 chart (b), the performances of residual attention model evaluated by "attention correctness" is conflicting with both our expectation and captioning performances in experiment 1. The metric values of resolution 40×40 and 60×60 are unexpectedly lower than that of resolution 27×27. For resolution 60×60, the performance of residual attention is significantly lower than that of soft-attention model. However, the performances of our "normalized attention correctness" metric shown in Fig.4 chart (a) are quite consistent with our expectation and captioning performances. Therefore, our metric is more rational and efficient than the "attention correctness".

Then, for our residual attention model, we further compare two metrics' evaluations on both object and stuff regions. In Fig. 6 chart (a), the accuracy performances of our metric are quite consistent with our expectation. Increasing attention resolution can significantly boost attention accuracy for object regions, but rather slightly for stuff regions. However, the results in Fig. 6 chart (b) shows inconsistent improvements from resolution 27×27 to 60×60 for object and stuff regions.

V. Conclusion

In this paper, we proposed a rational metric "normalized attention correctness" for the quantitative evaluation on attention accuracy. To demonstrate the efficiency of our metric, we also proposed the high-resolution residual attention model for image captioning based on the FCN-LSTM encoder. By using the MSCOCO and COCO-Stuff datasets, our experiments demonstrate that both high fine-grained resolution and residual attention can boost the attention accuracy and hence captioning performances. Moreover, our "normalized attention correctness" metric is more consistent with intuitive expectations and captioning performances for the quantitative evaluation on the attention accuracy. Future researches can focus on the object-size analysis of high-resolution attention or the explicit supervision on attention model.

References

[1] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," in *CVPR*, 2016.

[2] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *CVPR*, 2017.

[3] J. Lu, C. Xiong, D.Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[5] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016.

[6] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016.

[7] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Watch what you just said: Image captioning with text-conditional attention," in *arXiv preprint arXiv:1606.04621*, 2016.

[8] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "Fine-grained and semantic-guided visual attention for image captioning," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[9] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3135–3144.

[10] L. Yang, K. Tang, J. Yang, and L. J. Li, "Dense captioning with joint inference and visual context," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1978–1987.

[11] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1170–1178.

[12] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating copying mechanism in image captioning for learning novel objects," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5263–5271.

[13] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 873–881.

[14] H. R. Tavakoliy, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2506–2515.

[15] C. Liu, J. Mao, F. Sha, and L. Y. Alan, "Attention correctness in neural image captioning." in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2017, pp. 4176–4182.

[16] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1179–1195.

[17] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5168–5177.

[18] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *arXiv preprint arXiv:1612.03716*, 2016.

[19] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *CVPR*, 2016.

[20] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[23] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," 2016.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[25] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL*, 2005.

[26] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.