Understanding Image Captioning Models beyond Visualizing Attention

Jiamei Sun [1] **Sebastian Lapuschkin** [2] **Wojciech Samek** [2] **Alexander Binder** [1]

Abstract

This paper explains predictions of image captioning attention models beyond visualizing the attention itself. In this paper, we develop variants of layer-wise relevance backpropagation (LRP) tailored to image captioning models with attention mechanisms. We show that the explanations, firstly, correlate to object locations with higher precision than attention, secondly, identify object words that are unsupported by image content, and thirdly, provide guidance to improve the model. Results are reported using two different image captioning attention models trained with Flickr30K and MSCOCO2017 datasets. Experimental analyses show the strength of explanation methods for understanding image captioning attention models.

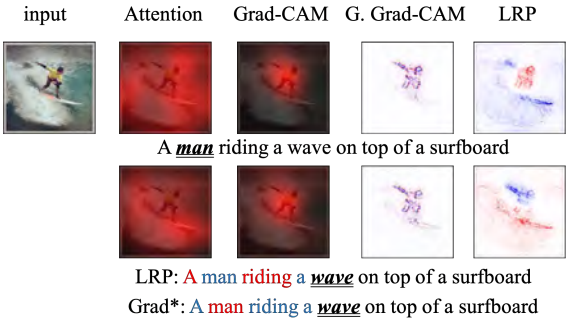

Figure 1. Image explanations of the word *man* (the first row) and *wave*(the second row) with attention, Grad-CAM, Guided Grad-CAM(G.Grad-CAM) and LRP. The latter three also provide linguistic explanations and the texts on the bottom show the linguistic explanations of the word *wave*. Red pixels and words indicate positive explanation scores and blue indicates negative explanation scores. Grad* denotes both Grad-CAM and Guided Grad-CAM.

1. Introduction

Image captioning is a task which gained interest along with the revival of neural networks. It aims at generating text descriptions from the image content. It requires a comprehensive understanding of the image and a well-performing decoder which translates the image features into sentences. Attention layers are an established component of image captioning models, particularly those for the image component. They enable the decoder to focus on a sub-region of the image when predicting the next word in the caption (Yang et al., 2016; Xu et al., 2015; Yao et al., 2017; You et al., 2016; Lu et al., 2017; Anderson et al., 2018; Vaswani et al., 2017; Huang et al., 2019). Attention heatmaps for the image part reflect which parts of the image are related to the generated words. As such they are a natural resource to explain the prediction of a word in a caption. However, for a multi-input model, the outputs of image captioning models rely on not only the image input but also the previously generated word sequence. Attention heatmaps for the image part will meet difficulties to disentangle the contribution of the image input and the text input.

To this end, we pose two questions here. Firstly, how suitable are attention heatmaps for the image part to explain the decision for a word when creating a caption? Secondly, to what extent is the image content actually used when predicting a caption word? These questions correspond to two desirable properties in image captioning: grounding (good localization from the attention model) and the consistency of the predicted caption to the image content.

To address the above questions, we adapt LRP and Grad-CAM to attention-guided models and explain image captioning predictions. Figure 1 shows an example of the explanations. Both positive and negative evidence is shown in LRP and Grad-CAM explanations in two aspects: the contribution of the image input visualized as heatmaps and the contribution of previously generated words to the latest predicted word. The contribution of previously generated words to the latest predicted word we will refer to as the linguistic explanation. It reveals those among the previously generated words which contribute strongly to the prediction of the explained word.

The contributions of this paper are 1). We establish explanation methods which reveal the contribution of both the

[1]Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore [2]Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Berlin, Germany. Correspondence to: Jiamei Sun <jiamei_sun@mymail.sutd.edu.sg>.

image and text inputs for image captioning attention models; 2). A comparison of the grounding property between attention and explanations; 3). We show that LRP can measure to what extent image content is used to predict the next word and can identify the object words that are hallucinated by the model. 3). We develop a fine-tuning strategy using LRP explanations to tackle the hallucination problem of image captioning, at the same time maintain the overall performance.

2. Related Work

Image Captioning Image captioning task has gained significant progress with deep neural networks(DNN). Many models adopt the encoder-decoder fashion to bridge image and text (Vinyals et al., 2015; Karpathy & Fei-Fei, 2015; Soh, 2016). Attention mechanisms are introduced to image captioning models to tell the decoder where to look to generate the words (Xu et al., 2015; Lu et al., 2017; You et al., 2016; Vaswani et al., 2017; Huang et al., 2019). Some works boost image captioning with enhanced image features using object detection models (Anderson et al., 2018) or graph convolutional networks(GCN) (Yao et al., 2019). A recent trend of image captioning is to include novel objects into the generated captions, which overcomes the limitation of fixed training vocabulary and achieves better generalization (Lu et al., 2018; Venugopalan et al., 2017; Li et al., 2019; Agrawal et al., 2019). This paper focuses on explaining the predictions of image captioning models and analyzing the interpretability of attention and explanation methods. We experiment with the fundamental CNN-RNN based attention models.

Explanation methods for neural networks. A number of methods explain DNNs such as gradient based methods and decomposition-based methods. Gradient based methods like gradient, gradient∗input (Simonyan et al., 2013), guided backpropagation (Springenberg et al., 2015), integrated gradient (Sundararajan et al., 2017), Grad-CAM, and Guided Grad-CAM (Selvaraju et al., 2017), process and visualize the backpropagated gradient in different ways as explanations. Decomposition based methods often rely on variants of neuron-wise Taylor decomposition (Montavon et al., 2017), resulting in different decomposition rules such as ϵ-rule, $\alpha\beta$-rule (Bach et al., 2015). Relevance scores may also be obtained by DEEPLIFT (Shrikumar et al., 2017) and PatternAttribution (Kindermans et al., 2018), and the generically applicable LIME (Ribeiro et al., 2016). SHARP (Lundberg & Lee, 2017) explains many of the above methods in a general framework of Shapley values.

As for explaining image captioning models, Grad-CAM and Guided Grad-CAM has been used to explain non-attention image captioning models (Selvaraju et al., 2017). Attention is often visualized to verify the correctness of the attention-guided image captioning models.

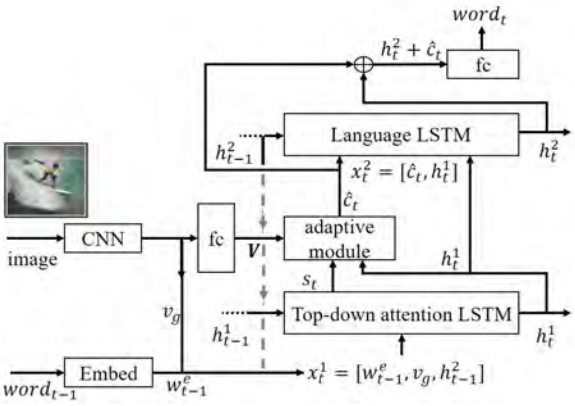

Figure 2. Image captioning model with grid-TD attention.

3. Methodology

In this section, we introduce the image captioning attention and explanation models used in our experiments. While many explanation methods are available, for showing selected qualitative differences to attention, it is sufficient to focus on a few. We will provide details of the extension of LRP, Grad-CAM, and Guided Grad-CAM to our two image captioning models. For clarity, Grad* denotes both Grad-CAM and Guided Grad-CAM in the following.

3.1. Attention mechanisms applied in this study

We introduce a grid-TD attention mechanism based on two well-performing ones here: the adaptive attention mechanism (Lu et al., 2017) and the bottom-up and top-down attention mechanism (BUTD) (Anderson et al., 2018). As illustrated in Figure 2, the image is first encoded by a CNN into image features $\boldsymbol{V}$. Derived from the output of the CNN encoder, a global image feature vector $\boldsymbol{v}_g$ is concatenated with each word embedding to generate the sequential input of an LSTM decoder, which is then augmented with the grid-TD attention module.

The grid-TD attention mechanism contains the *top-down attention LSTM* module, the *adaptive module*, and the *language LSTM*. The *top-down attention LSTM* generates a visual sentinel $\boldsymbol{s}_t$ using the memory cell $\boldsymbol{m}_t^1$ and the sequential input $\boldsymbol{x}_t^1$. $\boldsymbol{s}_t$ contains the text-only information.

$$\boldsymbol{s}_t = \sigma(\boldsymbol{W}_x \boldsymbol{x}_t^1 + \boldsymbol{W}_h \boldsymbol{h}_{t-1}^1) \odot tanh(\boldsymbol{m}_t^1) \qquad (1)$$

The *adaptive module* takes $\boldsymbol{s}_t$ and $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L\}$ as the input and calculates the context $\hat{\boldsymbol{c}}_t$. Let $\boldsymbol{\alpha}^{(t)}$ denote the attention weight for the image features $\boldsymbol{V}$, which is computed from both $\boldsymbol{V}$ and the hidden state $\boldsymbol{h}_t^1$ of the *top-down*

attention LSTM(eq.(2)(3)), resulting in a visual context $\boldsymbol{c}_t$.

$$\boldsymbol{a} = \boldsymbol{w}_a \tanh(\boldsymbol{W}_v \boldsymbol{V} + \boldsymbol{W}_g \boldsymbol{h}_t^1 \mathbb{1}^T) \quad (2)$$

$$\boldsymbol{\alpha}^{(t)} = \text{softmax}(\boldsymbol{a}) \quad (3)$$

$$\boldsymbol{c}_t = \sum_i^L \alpha_i^{(t)} \boldsymbol{v}_i \quad (4)$$

$\boldsymbol{s}_t$ is further concatenated to the spatial image features $\boldsymbol{V}$ and the attention is redistributed as $\hat{\boldsymbol{\alpha}}^{(t)}$. We can obtain the attention weight of $\boldsymbol{s}_t$ (the last element of $\hat{\boldsymbol{\alpha}}_t$), denoted as β_t. The final context $\hat{\boldsymbol{c}}_t$ is a linear combination of $\boldsymbol{c}_t$ and $\boldsymbol{s}_t$ weighted with β_t.

$$\boldsymbol{b} = \boldsymbol{w}_a \tanh(\boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{W}_g \boldsymbol{h}_t^1)) \quad (5)$$

$$\hat{\boldsymbol{\alpha}}^{(t)} = \text{softmax}([\boldsymbol{a} : \boldsymbol{b}]), \beta_t = \hat{\boldsymbol{\alpha}}_{L+1}^{(t)} \quad (6)$$

$$\hat{\boldsymbol{c}}_t = (1 - \beta_t)\boldsymbol{c}_t + \beta_t \boldsymbol{s}_t \quad (7)$$

$[\cdot]$ indicates concatenation. $\hat{\boldsymbol{c}}_t$ is passed to the *language LSTM* module and the *fc* layer to predict the next word. In the experiment, we apply explanation methods to both the adaptive attention mechanism in (Lu et al., 2017),which is composed of only the *top-down LSTM* and the *adaptive module*, and the more complex grid-TD attention mechanism.

3.2. Extending LRP to attention mechanisms

This section takes the grid-TD attention mechanism as an example to elaborate each step of applying LRP to image captioning attention models, as summarized in Algorithm 1. Firstly, we initialize the relevance of the T-th word, $R(word_T)$, from the output score of the *fc* layer. LRP-type operations for computing relevance $R(\cdot)$ are then applied to the layers *fc, $\oplus$, Language LSTM, attention-module, Top-down attention LSTM, and CNN*. The LRP operations used for these layers are shown as the $\Longrightarrow$ in Algorithm 1.

For the *fc* and $\oplus$ layers, we backpropagate the LRP relevance according to the ϵ-rule introduced in (Bach et al., 2015). For the convolutional layers in the image encoder, we apply $\alpha\beta$-rule (Bach et al., 2015). As for the *adaptive module*, we interpret

$$\hat{\boldsymbol{c}}_t = (1 - \beta_t) \sum_i \alpha_i^{(t)} \boldsymbol{v}_i + \beta_t \boldsymbol{s}_t \quad (8)$$

as a linear combination over $\{\boldsymbol{V} = (\boldsymbol{v}_i)_{i=1}^L, \boldsymbol{s}_t\}$, while treating the coefficients $((1-\beta_t)\alpha_i^{(t)})_{i=1}^L, \beta_t$ as the weights of the linear combination. Thus, we can apply the ϵ-rule to it to obtain $R(\boldsymbol{V})$ and $R(\boldsymbol{s}_t)$.

For each word to be explained, LRP generates an image explanation and the relevance scores for all the preceding words.

Algorithm 1 LRP for grid-TD attention model to explain $word_T$. For the appearing symbols consider Figure 2. Notations: $\boldsymbol{\alpha}^{(t)}$ (eq.(2)(3)), β_t (eq.(6)), and $\boldsymbol{s}_t$ (eq.(1)); LRP-LSTM (Arras et al., 2017); ϵ-rule, LRP-CNN (Bach et al., 2015).

Require: $R(word_T), \boldsymbol{\alpha}^{(t)}, \beta_t$
Ensure: $R(image), R(word_{T-1}), \ldots, R(word_0)$

$R(word_T)$, fc $\xrightarrow{\epsilon\text{-rule}}$ $R(\hat{\boldsymbol{c}}_T + \boldsymbol{h}_T^2)$
$R(\hat{\boldsymbol{c}}_T + \boldsymbol{h}_T^2), \oplus \xrightarrow{\epsilon\text{-rule}} R_1(\hat{\boldsymbol{c}}_T), R(\boldsymbol{h}_T^2)$
for $t \in [T, \ldots, 0, start]$ **do**
  $R(\boldsymbol{h}_t^2)$, Language-LSTM $\xrightarrow{\text{LRP-LSTM}}$ $R_2(\hat{\boldsymbol{c}}_t), R(\boldsymbol{h}_t^1), R_1(\boldsymbol{h}_{t-1}^2)$
  $R_1(\hat{\boldsymbol{c}}_t) + R_2(\hat{\boldsymbol{c}}_t)$, adaptive module $\xrightarrow{\epsilon\text{-rule}}$ $R(\boldsymbol{s}_t), R_t(\boldsymbol{V})$
  $R(\boldsymbol{h}_t^1), R(\boldsymbol{s}_t)$, Top-down Attention LSTM $\xrightarrow{\text{LRP-LSTM}}$ $\underbrace{R(\boldsymbol{W}_{t-1}^e), R_t(\boldsymbol{v}_g), R_2(\boldsymbol{h}_{t-1}^2)}_{=R(\boldsymbol{x}_t^1)}, R(\boldsymbol{h}_{t-1}^1)$
  $R(\boldsymbol{W}_{t-1}^e) \overset{\sum}{\Longrightarrow} R(word_{t-1})$
end for
$\sum_t R_t(\boldsymbol{V}), \sum_t R_t(\boldsymbol{v}_g)$, CNN $\xrightarrow{\epsilon\text{-rule,LRP-CNN}}$ $R(image)$

3.3. Extending Grad* methods to attention mechanisms

Besides LRP, Grad-CAM and Guided Grad-CAM are also adapted to the above attention mechanisms. Both methods backpropagate the gradient of a prediction to the image feature maps of the CNN encoder. The gradients of each feature map are summed up as the weight of the image feature (Selvaraju et al., 2017). Grad-CAM reshapes and up-samples the weight vector derived from the gradient to generate the image explanation. To obtain fine-grained and high-resolution explanations, Grad-CAM is fused with guided backpropagation (Springenberg et al., 2015) by pixel-wise multiplication and this fused method is Guided Grad-CAM. The linguistic explanation of Grad* methods is obtained by summing up the gradients of the word embedding.

4. Experiments

We train the adaptive attention model and the grid-TD attention model on Flickr30K and MSCOCO2017(Lin et al., 2014) datasets for the experiments. **Dataset**: We prepare the Flickr30K dataset as per the Karpathy split (Karpathy & Fei-Fei, 2015). For MSCOCO2017, we use the original validation set as the offline test set and extract 5000 images from the training set as the validation set. The train/validation/test sets are with 110000/5000/5000 images. Vocabulary is built only on the training set. The words that appear less than 3 times in the training set are not considered in the vocabulary for Flickr30K, and 5 times for MSCOCO2017. **CNN encoder**: We adopt the pre-trained VGG16 (Simonyan & Zisserman, 2015) on ImageNet as the image encoder and

Table 1. The performance of the adaptive attention model and the grid-TD attention model on the test set of Flickr30K and MSCOCO2017 datasets. The evaluation metrics are F_B: F_{BERT}(idf)(Zhang et al., 2019), C: CIDEr(Vedantam et al., 2015), S: SPICE(Anderson et al., 2016),R-L: ROUGE-L(Lin, 2004), M: METEOR(Banerjee & Lavie, 2005)

Flickr	F_B	C	S	R-L	M
adaptive	90.10	42.28	15.65	48.68	21.25
grid-TD	90.24	44.95	16.23	49.71	21.75
COCO	F_B	C	S	R-L	M
adaptive	91.19	87.75	20.28	55.79	26.61
grid-TD	91.25	90.67	20.49	56.41	27.04

Table 2. The percentage of the words that receive probability drops in the ablation experiment. The numbers within the brackets indicate the total number of experimented words. A higher percentage means the model relies more on the linguistic information to generate the target words. Experimented with the MSCOCO dataset.

adaptive	LRP	Grad*	random
stop words (12902)	90.32%	90.25%	78.32%
category words (2188)	92.06%	92.97%	84.30%
grid-TD	LRP	Grad*	random
stop words (12943)	71.21%	73.65 %	17.49%
category words (1888)	54.87%	48.04 %	11.76%

extract the output of '*block5_conv3*' layer as the raw image features. The raw image features are expanded to form the grid spatial image features $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_L\}$. $\boldsymbol{A}$ is further encoded with a time distributed fully connected layer to obtain $\boldsymbol{V}$, $\boldsymbol{v}_i = ReLU(\boldsymbol{W}_a \boldsymbol{a}_i)$ and the global image feature $\boldsymbol{v}_g = ReLU(\boldsymbol{W}_b \frac{1}{L} \sum \boldsymbol{a}_i)$. $\boldsymbol{W}_a$ and $\boldsymbol{W}_b$ are trainable parameters. **LSTM decoder**: See Figure 2. The dimension of word embedding and the hidden state is set as 512. **LRP parameters**: We use $\alpha\beta$-rule for convolutional layers and ϵ-rule for fully connected layers. We set $\alpha = 1$, $\epsilon = 0.01$. As for LRP-LSTM, we adopt ϵ-rule with $\epsilon = 0.01$. The performance of the adaptive attention and grid-TD attention model with beam size 3 is listed in Table 1.

4.1. Explanation results and evaluation

Figure 1 shows an example of the explanations, where LRP and Guided Grad-CAM both provide high-resolution image explanations and linguistic explanations. We here quantitatively analyze the linguistic explanation by a word ablation experiment and analyze the localization property of the image explanation by an object detection experiment.

Word ablation experiment for linguistic explanation The word ablation experiment is designed to prove that the related words found by explanation methods contribute to the predictions. The related words are with higher LRP scores or higher absolute value of Grad* scores. We experiment with two kinds of words in the generated caption, the stop words and the object words. The first 5 words in the generated captions are not considered. For each target word, we calculate the linguistic explanation using LRP and Grad* methods and delete the top-3 related words. The modified word sequence and the test image are then fed into the same captioning model and we observe how the probability of the target word changes. If the deleted words contribute to the prediction, the probability of the target word will drop. Table 2 shows how often the probability drops in this ablation experiment. A random ablation is included as the baseline. LRP and Grad* methods can find the relevant words more often than the random baseline, which proves that explanation methods can find the words that contribute to the prediction. From the results of the grid-TD model, we find that the stop words receive probability drops more often than the category words. This can be explained that the object words depend more on the image information and the stop words depend more on the linguistic information. However, the adaptive attention model is much more sensitive to the linguistic information.

Measuring the correlation of explanation scores to object locations In this part, we show that the image explanation of LRP and Grad* methods provide better localization property than attention, moreover, the sign of LRP explanation scores reflects the support for or the opposition to the prediction.

This experiment is conducted with the MSCOCO dataset, which provides referenced objects and the corresponding bounding boxes for each image. We evaluate the localization correctness of image explanations with the attention correctness measure (Liu et al., 2017), which is designed for attention models. Specifically, we explain the object words that appear in both the predicted captions and the referenced ground truth captions and obtain the image explanation E. We only keep the positive scores of E, $E_p = \max(E, 0)$ and normalize it to $[0, 1]$. The correctness of the localization is defined as the summation of E_p scores within the bounding box divided by the total score:

$$Correctness = \frac{\sum_{ij \in bbox} E_p[i,j]}{\sum_{ij} E_p[i,j]} \in [0,1] \qquad (9)$$

Figure 3 shows the average $correctness$ of all the correctly predicted object words of the grid-TD model (the results of the adaptive attention model are similar). The curve is generated by counting the normalized E_p scores that are larger than varying thresholds. We can see all the explanation methods achieve higher $correctness$ than attention. To further get insights into the role of the sign for LRP and Guided Grad-CAM explanations, we locate the objects using the absolute value of the negative image explanation scores, $E_n = \max(-E, 0)$, shown as the dashed line 'N-

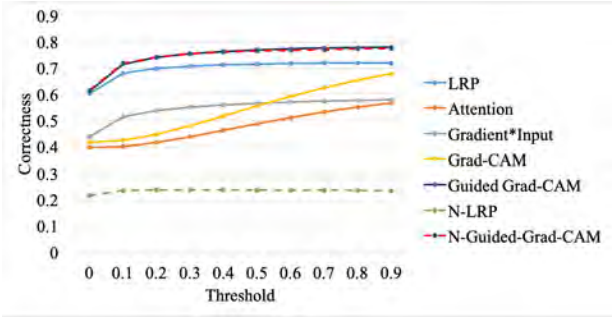

Figure 3. The average *correctness* of all the 5786 correctly detected objects using the grid-TD attention model on the test set of MSCOCO2017 dataset. Higher is better.

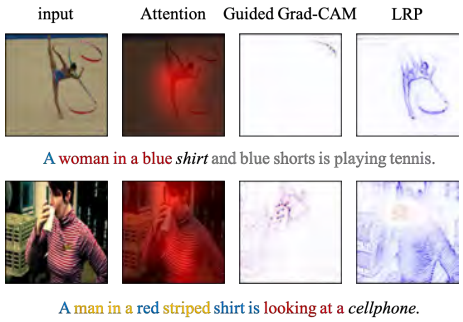

Figure 4. Explanation of LRP, Grad* and attention for unsupported words *shirt* and *cellphone*. Colors of the words represent different ranges of the normalized LRP relevance scores, red: $[0.3, 1]$, yellow: $(0, 0.3]$, blue:$(-1, 0]$, gray: not related.

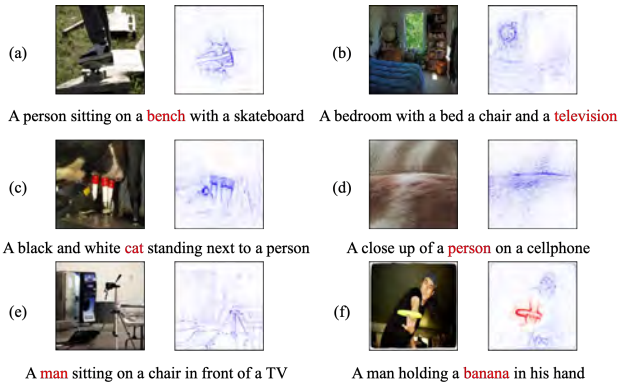

Figure 5. LRP explanations for hallucinated words. LRP explanation shows nearly all negative scores for (a)-(e). For (f), the model may mistake the yellow ring as a banana.

LRP' and 'N-Guided-Grad-CAM'. The low *correctness* of 'N-LRP' and the high *correctness* of 'N-Guided Grad-CAM' verifies that the sign of LRP explanations reveals the support('+') or the opposition('-') of a pixel to the predictions, while for Guided Grad-CAM both positive and negative pixels are related to the predictions and irrelevant pixels have low absolute scores.

4.2. Using explanation to improve the model

In this section, we show that explanation methods such as LRP and guided-Grad-CAM have the capability to identify the caption words that are not supported by the image content. In a second step, starting from this property, we propose an LRP inference fine-tuning strategy to improve the image captioning model.

Identifying hallucinated words using explanations We observe that some predicted words are not supported by the image content and constitute false-positive. This is called object hallucination in image captioning (Rohrbach et al., 2018). These words are generated from the learned sentence correlation, without looking at the image, and could be inferred from the frequent words in the training set.

Figure 4 illustrates the explanation of LRP, Grad*, and attention for two hallucinated words *shirt* and *cellphone*. Attention heatmaps hardly tell the reason why the model emits such words, while explanation methods cast light on the details. *shirt* is likely to be generated from the linguistic information because LRP shows negative and Guided Grad-CAM shows zero scores in the heatmaps. LRP finds supporting words in the linguistic explanation shown in red. For the second example, the model seems to mistake the cup for a cellphone since the edges of the hand and the cup are highlighted in the LRP explanation heatmap, which resembles a cellphone. Guided Grad-CAM also shows intensive blue and red pixels near the cup and hand, while attentions are dispersed to the person. We include more examples in Figure 5 where LRP heatmaps exhibit almost all negative evidence or very similar features for hallucinated words.

To quantitatively analyze whether the explanations are capable of identifying hallucinated words, we use the image explanation scores to detect the hallucinated word and evaluate this property by calculating the AUC value. Specifically, object words that appear more than 50 times in the predictions of the Flickr30K test set are used in this experiment [1], resulting in 889 true-positive and 771 false-positive words using grid-TD attention model and 816 true-positive and 766 false-positive words using adaptive attention model. True-positive words are labeled 1 and false-positive words receive a label 0. Each word is assigned with a score as the mean of the image explanation. We also evaluate the maximal value of LRP explanation scores (L-max), mean of the absolute value of Guided Grad-CAM (Guided Grad-CAM-abs), and the intrinsic model parameter $1 - \beta_t$, which weights the image information and linguistic information.

[1]These object words include *man, shirt, woman, people, group, street, dog, bench, boy*

Table 3. The AUC scores of different explanation methods. L.: LRP, G.: Guided Grad-CAM, Att.: Attention. Higher is better.

AUC	L.-max	L.	G.-abs	Att.	$1-\beta_t$	G.
grid-TD	0.61	**0.64**	0.54	0.53	0.54	0.45
adaptive	**0.66**	0.64	0.58	0.51	0.53	0.42

Table 3 summarizes the AUC value of explanation methods, attention, and the model parameter β. We find that the intrinsic model parameter $1-\beta_t$ performs better than attention maps, the absolute value of guided-Grad-CAM performs better than $1-\beta_t$, and the preferable ability of LRP to identify the hallucinated words.

Alleviating hallucination problem with LRP explanations With the above finding, we introduce an LRP inference fine-tuning strategy to alleviate the object hallucination problem. The idea is to re-weight the model prediction scores (or "logits") $\boldsymbol{p} \in \mathbb{R}^{l_v}$ from the *fc* layer by a weight $\boldsymbol{m}$ during training :

$$\hat{p}^{(r)} = m^{(r)} p^{(r)},\ r \in \{1, \ldots, l_v\} \tag{10}$$

$\boldsymbol{m}$, $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$ are vectors with length as the vocabulary size, l_v and r is the index. Given a predicted caption $\{w_t\}_{t=0}^{T}$, we explain each of its words w_t (excluding stop words) by LRP to obtain image explanation $R(image)$. $R(image)$ is normalized with its maximal absolute value resulting in $E_t \in [-1, 1]$. Let $h(w_t)$ be the one-hot mapping of word w_t onto its vocabulary index in $\{1, \ldots, l_v\}$. The weights $\boldsymbol{m} \in \mathbb{R}^{l_v}$ are defined as:

$$m^{(h(w_t))} = \begin{cases} 1 + \text{mean}(E_t) & w_t \notin \text{stop words} \\ 1 & \text{otherwise} \end{cases} \tag{11}$$

If a word is frequently predicted by the model but has negative mean(E_t), then $\hat{p}^{(r)}$ will decrease its probability. In other words, $\hat{p}^{(r)}$ guides the model to look more at the image.

During training, we calculate the cross entropy loss with both the original predicted scores $\boldsymbol{p}$ and LRP-inference score $\hat{\boldsymbol{p}}$, and combine both loss with a parameter $\lambda \in [0, 1]$.

$$L = \lambda CE(\boldsymbol{p}) + (1-\lambda) CE(\hat{\boldsymbol{p}}) \tag{12}$$

where CE denotes the cross entropy loss.

We fine-tune our captioning models on Flickr30K and MSCOCO2017 datasets using LRP inference with learning rate $1e-6$, $\lambda = 0.5$. The batch size is 32 and we train 320 iterations for Flickr30K dataset and 500 iterations for MSCOCO2017 dataset. As a baseline, we also fine-tune our pre-trained models without LRP inference using the same training samples and parameters. The overall performance after LRP inference fine-tuning maintains or even slightly better as shown in Table 4.

Table 4. The performance of the models tuned w/wo LRP inference (L.-) on the test set. F_B: F_{BERT}(idf), C: CIDEr, S: SPICE, R-L: ROUGE-L, M: METEOR.

Flickr	F_B	C	S	R-L	M
adaptive	90.06	41.11	15.68	48.58	21.14
L.-adaptive	90.08	40.35	15.76	48.70	21.30
grid-TD	90.21	45.26	16.27	49.66	21.86
L.-grid-TD	90.25	45.73	16.38	49.92	22.95
MSCOCO	F_B	C	S	R-L	M
adaptive	91.22	87.68	20.30	55.99	26.47
L.-adaptive	91.23	88.22	20.31	55.99	26.49
grid-TD	91.20	89.07	20.41	56.18	26.89
L.-grid-TD	91.20	89.21	20.40	56.18	26.90

Table 5. The mean average precision (mAP) of the frequent object words. Results of models tuned w/wo LRP inference (L.-). True positive words are supported by the image content and false positive words are hallucinated. Higher mAP means less object hallucination.

mAP	ada.	L.-ada.	grid-TD	L.-grid-TD
Flickr30K	51.14	**54.57**	54.66	**55.22**
MSCOCO	63.53	**64.58**	64.13	**64.19**

To evaluate whether the captioning models hallucinate less, we calculate the mean average precision (mAP) for the frequent object words, which appear over 50 times for Flickr30K dataset and over 100 times for MSCOCO2017 dataset. The mAP results are listed in Table 5. We can observe an improvement in mAP with the proposed LRP-inference fine-tuning. Furthermore, the improvement of the adaptive model is larger than for the grid-TD model. This matches our finding in the word ablation experiment (cf. Table 2) that the adaptive attention model relies more on the sentence correlation while the grid-TD model looks more at the images, and thus is less prone to word hallucination.

5. Conclusion

We have applied a variant of LRP, Grad-CAM, and Guided Grad-CAM to explain the attention-guided image captioning models beyond visualizing the attentions. With the qualitative explanation results and the quantitative evaluations, we show that explanation methods provide more interpretable information than attention including high-resolution image explanations, improved localization, and the capability to identify supporting words in the generated caption for targeted explained words. Explanations methods are shown to identify hallucinated words and help to reduce object hallucination meanwhile maintain the performance. The comparison of explanation types shows a diversified picture. Guided Grad-CAM performs best for localization, LRP best for identifying words unsupported by image content.

References

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8948–8957, 2019.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pp. 382–398. Springer, 2016.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Arras, L., Montavon, G., Müller, K., and Samek, W. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 159–168, 2017.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4634–4643, 2019.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Kindermans, P., Schütt, K. T., Alber, M., Müller, K., Erhan, D., Kim, B., and Dähne, S. Learning how to explain neural networks: Patternnet and patternattribution. In *International Conference on Learning Representations*, 2018.

Li, Y., Yao, T., Pan, Y., Chao, H., and Mei, T. Pointing novel objects in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12497–12506, 2019.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, C., Mao, J., Sha, F., and Yuille, A. Attention correctness in neural image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Lu, J., Xiong, C., Parikh, D., and Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 375–383, 2017.

Lu, J., Yang, J., Batra, D., and Parikh, D. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7219–7228, 2018.

Lundberg, S. M. and Lee, S. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65: 211–222, 2017.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

Soh, M. Learning cnn-lstm architectures for image caption generation. *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep*, 2016.

Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (workshop track)*, 2015.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328. JMLR. org, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., and Saenko, K. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5753–5761, 2017.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.

Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., and Salakhutdinov, R. R. Review networks for caption generation. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2361–2369. Curran Associates, Inc., 2016.

Yao, T., Pan, Y., Li, Y., Qiu, Z., and Mei, T. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4894–4902, 2017.

Yao, T., Pan, Y., Li, Y., and Mei, T. Hierarchy parsing for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2621–2629, 2019.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, 2016.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.