Explain and Improve: LRP-Inference Fine Tuning for Image Captioning Models

Jiamei Sun[a], Sebastian Lapuschkin[b], Wojciech Samek[b,*], Alexander Binder[c,*]

[a]*Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore*
[b]*Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany*
[c]*Department of Informatics, University of Oslo, Oslo, Norway*

Abstract

This paper analyzes the predictions of image captioning models with attention mechanisms beyond visualizing the attention itself. We develop variants of layer-wise relevance propagation (LRP) and gradient-based explanation methods, tailored to image captioning models with attention mechanisms. We compare the interpretability of attention heatmaps systematically against the explanations provided by explanation methods such as LRP, Grad-CAM, and Guided Grad-CAM. We show that explanation methods provide simultaneously pixel-wise image explanations (supporting and opposing pixels of the input image) and linguistic explanations (supporting and opposing words of the preceding sequence) for each word in the predicted captions. We demonstrate with extensive experiments that explanation methods 1) can reveal additional evidence used by the model to make decisions compared to attention; 2) correlate to object locations with high precision; 3) are helpful to "debug" the model, e.g. by analyzing the reasons for hallucinated object words. With the observed properties of explanations, we further design an LRP-inference fine-tuning strategy that reduces the issue of object hallucination in image captioning models, and meanwhile, maintains the sentence fluency. We conduct experiments with two widely used attention mechanisms: the adaptive attention mechanism calculated with the additive attention and the multi-head attention mechanism calculated with the scaled dot product.

[*]corresponding author

Email addresses: jiamei_sun@mymail.sutd.edu.sg (Jiamei Sun), sebastian.lapuschkin@hhi.fraunhofer.de (Sebastian Lapuschkin), wojciech.samek@hhi.fraunhofer.de (Wojciech Samek), alexabin@ifi.uio.no (Alexander Binder)

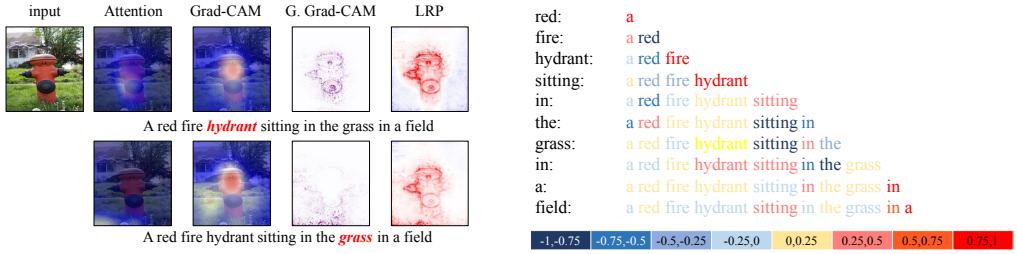
Keywords:
Explainable AI, Image captioning, Attention, Neural networks

1. Introduction

Image captioning is a setup that aims at generating text descriptions from image representations. This task requires a comprehensive understanding of the image content and a well-performing decoder which translates image features into sentences. The combination of a convolutional neural network (CNN) and a recurrent neural network (RNN) is a commonly used structure in image captioning models, with CNN as the image encoder and RNN as the sentence decoder [1, 2, 3]. An established feature of image captioning is the attention mechanism that enables the decoder to focus on a sub-region of the image when predicting the next word of the caption [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. Attentions are usually visualized as attention heatmaps, indicating 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. Furthermore, attention heatmaps are usually considered as the qualitative evaluations of image captioning models in addition to the quantitative evaluation metrics such as BLEU [16], METEOR [17], ROUGE-L [18], CIDEr [19], SPICE [20].

Attention heatmaps provide a certain level of interpretability for image captioning models since they can reflect the locations of objects. However, the outputs of image captioning models rely on not only the image input but also the previously generated word sequence. Attention heatmaps alone meet difficulties in disentangling the contributions of the image input and the text input.

To gain more insights into the image captioning models, we adapt layer-wise relevance propagation (LRP) and gradient-based explanation methods (Grad-CAM, Guided Grad-CAM [21], and GuidedBackpropagation [22]) to explain image captioning predictions with respect to the image content and the words of the sentence generated so far. These approaches provide high-resolution image explanations for CNN models [22, 23]. LRP also provides plausible explanations for LSTM architectures [24, 25]. Figure 1 shows an example of the explanation results of attention-guided image captioning models. Taking LRP as an example, both positive and negative evidence is shown in two aspects: 1) for image explanations, the contribution of the image input is visualized as heatmaps; 2) for linguistic explanations, the contribution of the previously generated words to the latest predicted word is shown.



(a) image explanations

(b) linguistic explanations of LRP

Figure 1: (a): Image explanations of the word *hydrant* (the first row) and *grass* (the second row) with attention, Grad-CAM, Guided Grad-CAM (G.Grad-CAM) and LRP. (b): The linguistic explanations of LRP for each word in the predicted caption. Blue and red colors indicate negative and positive relevance scores, respectively.

The explanation results in Figure 1 exhibit intuitive correspondence of the explained word to the image content and the related sequential input. However, to our best knowledge, few works quantitatively analyze how accurate the image explanations are grounded to the relevant image content and whether the highlighted inputs are used as evidence by the model to make decisions. We study the two questions by quantifying the grounding property of attention and explanation methods and by designing an ablation experiment for both the image explanations and linguistic explanations. We will demonstrate that explanation methods can generate image explanations with accurate spatial grounding property, meanwhile, reveal more related inputs (pixels of the image input and words of the linguistic sequence input) that are used as evidence for the model decisions. Also, explanation methods can disentangle the contributions of the image and text inputs and provide more interpretable information than purely image-centered attention.

With explanation methods [26], we have a deeper understanding of image captioning models beyond visualizing the attention. We also observe that image captioning models sometimes hallucinate words from the learned sentence correlations without looking at the images and sometimes use irrelevant evidence to make predictions. The hallucination problem is also discussed in [27], where the authors state that it is possibly caused by language priors or visual mis-classification, which could be partially due to the biases present in the dataset. The image captioning models tend to generate those words and sentence patterns that appear more frequently during training. The language priors are helpful, though, in some cases. [28] incorporates the inductive bias of natural language with scene graphs to facilitate image captioning. However, language bias is not always correct, for

example, not only men ride snowboards [29] and bananas are not always yellow [30, 31]. To this end, [29] and [31] attempted to generate more grounded captions by guiding the model to make the right decisions using the right reasons. They adopted additional annotations, such as the instance segmentation annotation and the human-annotated rank of the relevant image patches, to design new losses for training.

In this paper, we reduce object hallucination by a simple ***LRP-inference fine-tuning*** (LRP-IFT) strategy, without any additional annotations. We firstly show that the explanations, especially LRP, can weakly differentiate the grounded (true-positive) and hallucinated (false-positive) words. Secondly, based on the findings that LRP reveals the related features of the explained words and that the sign of its relevance scores indicates supporting versus opposing evidence (as shown in Figure 1), we utilize LRP explanations to design a re-weighting mechanism for the context representation. During fine-tuning, we up-scale the supporting features and down-scale the opposing ones using a weight calculated from LRP relevance scores. Finally, we use the re-weighted context representation to predict the next word for fine-tuning.

LRP-IFT is different from standard fine-tuning which weights the gradients of parameters with small learning rates to gradually adapt the model parameters. Instead, it pinpoints the related features/evidence for a decision and guides the model to tune more on those related features. This fine-tuning strategy resembles how we correct our cognition bias. For example, when we see a green banana, we will update the color feature of bananas and keep the other features such as the shape.

We will demonstrate that LRP-IFT can help to de-bias image captioning models from frequently occurring object words. Though language bias is intrinsic, we can guide the model to be more precise when generating frequent object words rather than hallucinate them. We implement the LRP-IFT on top of pre-trained image captioning models trained with Flickr30K [32] and MSCOCO2017 [33] datasets and effectively improve the mean average precision (mAP) of predicted frequent object words evaluated across the test set. At the same time, the overall performance in terms of sentence-level evaluation metrics is maintained.

The contributions of this paper are as follows:

- We establish explanation methods that disentangle the contributions of the image and text inputs and explain image captioning models beyond visualizing attention.
- We quantitatively measure and compare the properties of explanation meth-

ods and attention mechanisms, including tasks of finding the related features/evidence for model decisions, grounding to image content, and the capability of debugging the models (in terms of providing possible reasons for object hallucination and differentiating hallucinated words).

- We propose an LRP-inference fine-tuning strategy that reduces object hallucination and guides the models to be more precise and grounded on image evidence when predicting frequent object words. Our proposed fine-tuning strategy requires no additional annotations and successfully improves the mean average precision of predicted frequent object words.

In the rest of this paper, Section 2 introduces recent image captioning models, the state-of-the-art explanation methods for neural networks, and other related works. In Section 3, we will introduce the image captioning model structures applied in this paper. The adaptations of explanation methods to attention-guided image captioning models are summarized in Section 4. The analyses of attention and explanations and our proposed LRP-inference fine-tuning strategy are introduced in Section 5.

2. Related Work

2.1. Image Captioning

Many models adopt the encoder-decoder approach to bridge the gap between image and text, usually with a CNN as the image encoder and an RNN as the sentence decoder [1, 2, 3]. Considering that it might be helpful to focus on a sub-region of the image when generating a word of the caption, various attention mechanisms have been developed, guiding the model to focus on the relevant parts of the image when predicting a word. Some representative works include hard or soft attention [4], semantic attention [6], adaptive attention [7], bottom-up and top-down attention [8], adaptive attention time [9], hierarchical attention [10], X-Linear attention [34], and spatio-temporal memory attention [35]. Recently, many works build the attention mechanism with the multi-head attention originated from Transformer models [11], such as attention on attention [12], entangled transformer [13], multi-modal transformer [14], meshed-memory transformer [15]. These attention mechanisms effectively facilitate image captioning models to better recognize and locate the objects in an image. We will analyze the adaptive attention mechanism [7, 8, 9, 10] and the multi-head attention mechanism [11, 36, 12, 13, 14, 15]. Both attention mechanisms are employed as a sub-module in a number of works.

Recognizing and locating the objects in an image is often not sufficient to generate fine-grained captions. In addition to studying attention mechanisms, a branch of research explores the relations of objects (e.g. *playing with* balls) and object attributes (e.g. a *wooden* desk). Many of these works build a graph to capture the relation and attribute representations of objects, such as the scene graph [28, 37, 38, 39, 40] and visual relation graph [41]. Some other works aim to generate more fine-grained captions by learning global and local representations in a distilling fashion [42], by gradually learning the representation via context-aware visual policy [43], by parsing and utilizing the noun chunks in the reference captions [44]. The unified VLP [45] learns unified image-text representations in the spirit of the BERT embedding [46]. VIVO [47] and OSCAR [48] further enhance the unified representation by incorporating external image-tag pairs for training. These unified representations can be used in various visual-language tasks. [49] uses additional rank annotations of the referenced captions.

There are also other challenging directions of image captioning like novel object captioning (NOC) and captioning with different styles. NOC tries to predict novel objects that are not in the image-caption training pairs, which overcomes the limitation of fixed training vocabulary and achieves better generalization [50, 51, 52, 53, 54, 47, 55]. [56] and [57] attempt to generate captions with controlled sentiments and styles.

2.2. Towards de-biasing visual-language models

The intrinsic composition of the training data can lead to biased visual-language models. To this end, many works aim to reduce model bias and improve the grounding property of visual-language models. For visual-question-answering (VQA) models, [30] learns the language bias in advance by using the textual question-answer pairs in order to increase the loss computation for biased answers during training. [58] proposes a grounded visual question answering model that disentangles the yes/no questions and visual concept-related questions. Both show an effective reduction of the bias for the VQA models. As for image captioning models, [29] designs an appearance confusion loss and a confidence loss using segmentation annotations to reduce the gender bias of the captioning models. [31] adopts external human-annotated attention maps to guide the model to generate more grounded captions. Different from the above methods, we propose an LRP-inference fine-tuning strategy that requires no additional annotations to mitigate the influence of language bias for image captioning models. The guidance comes from the explanation scores obtained from explanation methods.

2.3. Explanation methods for image captioning models.

Many explanation methods explain the predictions of DNNs such as gradient-based methods [59, 22, 60, 21], decomposition-based methods [61, 23, 24, 62, 63, 64, 65], and sampling-based methods [66, 67, 68, 69, 70]. These explanation methods have provided plausible explanations for various DNN architectures including CNNs [23, 21, 62, 66, 63, 64, 67, 68, 71], RNNs [24, 65, 66, 64], graph neural networks (GNNs) [72, 73, 74, 75, 76], and clustering models [77], making it practical to derive the explanation methods for image captioning models. However, to our best knowledge, only a few works have studied the interpretability of image captioning models so far. In principle, gradient-based methods can be directly applied to image captioning models. Grad-CAM and Guided Grad-CAM have been used to explain non-attention image captioning models [21]. [78] introduces an explanation method for video captioning models. They further adapt the method to image captioning models by slicing an image with grids to form a sequence of image patches, treated as video frames, however, the slicing operation may cut through object structures. Attention heatmaps are usually considered as explanations of image captioning models. The question to what extent attention is suitable as an explanation has been raised in the natural language processing context [79, 80, 81]. For the image captioning task, although attention heatmaps can show the locations of object words, they cannot disentangle the contributions of the image and text inputs. Furthermore, attention heatmaps meet difficulties to provide pixel-wise explanations that reflect the positive and negative contributions of pixels and regions. These issues can be addressed by several explanation methods. For the sake of keeping the scope of analyses within reasonable limits, we will adapt exemplarily LRP, Grad-CAM, and Guided Grad-CAM to image captioning models.

2.4. Explanation-guided training

Recently, some studies observe that explainable AI is not limited to providing post-hoc insights into neural networks but can also be applied to train a model. [82] utilizes the saliency maps of Grad-CAM and Guided Grad-CAM to design a pixel-wise cross-entropy loss for transfer learning. They show that the pixel-wise cross-entropy loss can guide the model to make the right decisions using the right reasons, meanwhile, improve image classification accuracy. [31] also uses Grad-CAM saliency maps together with additional human-annotated attention maps to design a ranking loss for image captioning models. They show that the ranking loss can help to generate more grounded captions and maintain sentence fluency. [83] adopts LRP explanations to guide few-shot classification models. They

demonstrate that explanation-guided training can improve the model generalization and classification accuracy for cross-domain datasets. We will show that LRP explanations can also help to mitigate the influence of language bias for image captioning models.

3. Backgrounds of Image Captioning Models

3.1. Notations for image captioning models

In this section, we recapitulate common structures of image captioning models, which consist of an image encoder, a sentence decoder, and a word predictor module. To caption a given image, we first encode the image with pre-trained CNNs or detection modules such as a Faster RCNN and extract a visual feature $\boldsymbol{I} \in \mathbb{R}^{n_v \times d_v}$, where n_v and d_v are the number and dimension of the visual feature. For $\boldsymbol{I}$ from a Faster R-CNN, n_v would be the number of regions of interest (ROIs), and for $\boldsymbol{I}$ from a CNN, n_v would be the number of spatial elements in a feature map. Then, the visual feature $\boldsymbol{I}$ is decoded by an LSTM augmented with an attention mechanism to generate a context representation. Finally, the word predictor takes the context representation and the hidden state of the decoder as inputs to predict the next word.

During training, there is a reference sentence as the ground truth, $\mathcal{S} = (w_t)_{t=1}^{l}$, where w_t is a word token, and l is the sentence length. At each time step t, the LSTM updates the hidden state $\boldsymbol{h}_t$ and memory cell $\boldsymbol{m}_t$ as follows.

$$\boldsymbol{x}_t = [\boldsymbol{E}_{\boldsymbol{m}}(w_{t-1}), \boldsymbol{I}_g] \tag{1}$$

$$\boldsymbol{h}_t, \boldsymbol{m}_t = LSTM(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{m}_{t-1}) \tag{2}$$

where $[\cdot]$ denotes concatenation, $\boldsymbol{E}_{\boldsymbol{m}}$ is a word embedding layer that encodes words to vectors, $\boldsymbol{E}_{\boldsymbol{m}}(w_{t-1}) \in \mathbb{R}^{d_w}$. $\boldsymbol{I}_g = 1/n_v \sum_{k=1}^{n_v} \boldsymbol{I}_{(k)}$ represents an averaged global visual feature. During inference, the w_{t-1} is the predicted word from the last step. Then, an attention mechanism $ATT(\cdot)$ uses $\boldsymbol{h}_t$ and $\boldsymbol{I}$ to generate a context representation $\boldsymbol{c}_t$ for word prediction.

$$\boldsymbol{c}_t = ATT(\boldsymbol{h}_t, \boldsymbol{I}) \tag{3}$$

$$\boldsymbol{p}_t = Predictor(\boldsymbol{h}_t, \boldsymbol{c}_t) \tag{4}$$

where $\boldsymbol{p}_t$ is the predicted score over the vocabulary. The concrete implementations of $\boldsymbol{E}_{\boldsymbol{m}}$, $\boldsymbol{I}_g$, $ATT(\cdot)$, and $Predictor$ may vary across different models.

3.2. Attention mechanisms used in this study

We choose two representative attention mechanisms, adaptive attention [7] and a modified multi-head attention [11, 12]. They are employed in variants by several image captioning models, thus aiming at generalizability for our studies.

3.2.1. Adaptive attention mechanism

The adaptive attention mechanism generates a context representation by calculating a set of weights over the visual feature $\boldsymbol{I}$ and a sentinel feature $\boldsymbol{s}_t$ that represents the textual information. At time step t:

$$\boldsymbol{s}_t = \sigma(\boldsymbol{W}_x \boldsymbol{x}_t + \boldsymbol{W}_h \boldsymbol{h}_{t-1}) \odot \tanh(\boldsymbol{m}_t) \tag{5}$$

$\boldsymbol{W}_x \in \mathbb{R}^{d_h \times d_x}$ and $\boldsymbol{W}_h \in \mathbb{R}^{d_h \times d_h}$ are trainable parameters. d_h and d_x denote the dimension of the hidden state and $\boldsymbol{x}_t$, respectively. σ denotes the *sigmoid* function. The weights for $\boldsymbol{I}$ and $\boldsymbol{s}_t$ are calculated as follows:

$$\boldsymbol{a} = \boldsymbol{w}_a \tanh(\boldsymbol{I}\boldsymbol{W}_I + \boldsymbol{W}_g \boldsymbol{h}_t) \tag{6}$$

$$b = \boldsymbol{w}_a \tanh(\boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{W}_g \boldsymbol{h}_t)) \tag{7}$$

$$\boldsymbol{\alpha}_t = \mathrm{softmax}(\boldsymbol{a}) \in \mathbb{R}^{n_v} \tag{8}$$

$$\beta_t = \mathrm{softmax}([\boldsymbol{a}, b])_{(n_v+1)} \in \mathbb{R}^1 \tag{9}$$

$$\boldsymbol{c}_t = (1 - \beta_t) \sum\nolimits_{k=1}^{n_v} \alpha_{t_{(k)}} \boldsymbol{I}_{(k)} + \beta_t \boldsymbol{s}_t \tag{10}$$

where $\boldsymbol{W}_I \in \mathbb{R}^{d_h \times n_v}$, $\boldsymbol{W}_s$ and $\boldsymbol{W}_g \in \mathbb{R}^{n_v \times d_h}$, $\boldsymbol{w}_a \in \mathbb{R}^{n_v}$ are trainable parameters[1]. $\boldsymbol{\alpha}_t \in \mathbb{R}^{n_v}$ is the attention weight for $\boldsymbol{I}$. It tells the model which regions within the image to use for generating the next word. β_t is the $(n_v + 1)^{th}$ element of the *softmax* over $[\boldsymbol{a}, b]$, corresponding to the weight for the component b. It balances the visual and textual information used to predict the next word. We use the following expression to summarize the adaptive attention mechanism.

$$\boldsymbol{c}_t = ATT_{ada}(\boldsymbol{h}_t, \boldsymbol{s}_t, \boldsymbol{I}) \tag{11}$$

3.2.2. Multi-head attention mechanism

The multi-head attention is defined with a triplet of query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$). To apply the multi-head attention to the sentence decoder, we adopt $\boldsymbol{h}_t$ as the query and two distinct linear projections of $\boldsymbol{I}$ as $\boldsymbol{K}$ and $\boldsymbol{V}$.

$$\boldsymbol{Q} = \boldsymbol{h}_t, \boldsymbol{K} = \boldsymbol{I}\boldsymbol{W}_k, \boldsymbol{V} = \boldsymbol{I}\boldsymbol{W}_v \tag{12}$$

[1] Adaptive attention mechanism encodes the visual feature $\boldsymbol{I}$ with the same dimension as the hidden state, $d_v = d_h$, $\boldsymbol{I} \in \mathbb{R}^{n_v \times d_h}$.

where $\boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{d_v \times d_h}$. We evenly split the hidden dimension d_h to obtain multiple triplets of ($\boldsymbol{Q}^{(i)}$, $\boldsymbol{K}^{(i)}$, $\boldsymbol{V}^{(i)}$), denoted as multiple heads. For each head, the attention weight over $\boldsymbol{V}^{(i)}$ is the scaled dot product of $\boldsymbol{Q}^{(i)}$ and $\boldsymbol{K}^{(i)}$ and we can obtain a weighted feature $\boldsymbol{v}^{(i)}$ as follows.

$$\begin{aligned} \boldsymbol{\alpha}^{(i)} &= \mathrm{softmax}(\frac{\boldsymbol{Q}^{(i)}\boldsymbol{K}^{(i)\boldsymbol{T}}}{\sqrt{d_h/n_h}}) \in \mathbb{R}^{n_v} \\ \boldsymbol{v}^{(i)} &= \sum_{k=1}^{n_v} \boldsymbol{\alpha}^{(i)}_{(k)} \boldsymbol{V}^{(i)}_{(k)} \in \mathbb{R}^{d_h/n_h} \end{aligned} \tag{13}$$

where n_h is the number of head[2]. By concatenating the weighted feature of each head, we can obtain the integral attended feature $\boldsymbol{v}$, which is further fed to a linear layer to generate the visual representation.

$$\begin{aligned} \boldsymbol{v} &= [\boldsymbol{v}^{(0)}, \ldots, \boldsymbol{v}^{(n_h)}] \in \mathbb{R}^{d_h} \\ \hat{\boldsymbol{v}} &= \boldsymbol{W}_{\boldsymbol{v}}\boldsymbol{v} + \boldsymbol{b}_v \end{aligned} \tag{14}$$

where $\boldsymbol{W}_v \in \mathbb{R}^{d_h \times d_h}$ and $\boldsymbol{b}_v \in \mathbb{R}^{d_h}$ are trainable parameters.

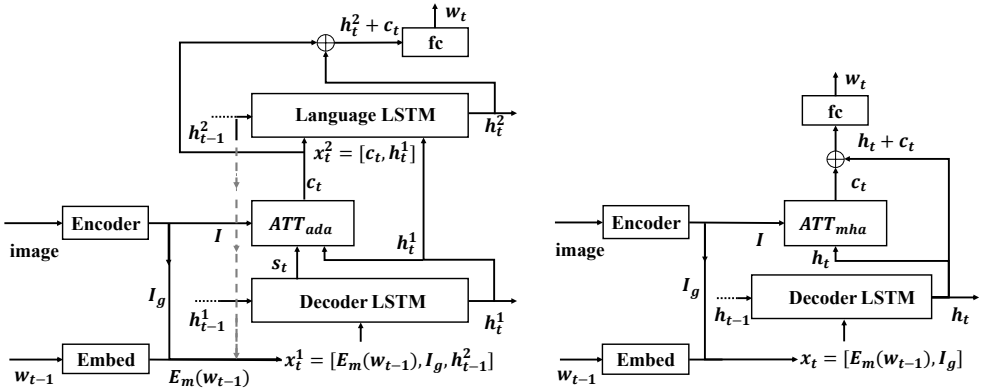
Under the image captioning setup, there are cases where the visual feature is less relevant to the predicted word, e.g. "a" and "the". Thus, we add another gate to control the visual information, which is consistent with many recent image captioning models using the multi-head attention module [12, 13, 15]. This also shares the same spirit of β_t in the adaptive attention mechanism, which controls the proportion of image and textual information. Specifically, we generate the gate using the hidden state and the gated output $\boldsymbol{c}_t$ is the context representation for prediction.

$$\boldsymbol{c}_t = \sigma(\boldsymbol{W}_{mh}\boldsymbol{h}_t + \boldsymbol{b}_{mh}) \odot \hat{\boldsymbol{v}} \tag{15}$$

where $\boldsymbol{W}_{mh} \in \mathbb{R}^{d_h \times d_h}$ and $\boldsymbol{b}_{mh} \in \mathbb{R}^{d_h}$ are trainable parameters and σ is the *sigmoid* function. We briefly summarize the multi-head attention mechanism as follows.

$$\boldsymbol{c}_t = ATT_{mha}(\boldsymbol{h}_t, \boldsymbol{I}) \tag{16}$$

[2]In most of the works using multi-head attention, d_h is divisible by n_h.



(a) Ada-LSTM (b) MH-FC

Figure 2: The model structures of two image captioning models. (a): The Ada-LSTM model with the adaptive attention mechanism and an *LSTM + fc* module as the word predictor. (b): the MH-FC model with the multi-head attention mechanism and an *fc* layer as the word predictor.

3.2.3. Image captioning models with adaptive attention and multi-head attention

We build two image captioning models in this paper. The details of the two models are illustrated in Figure 2. The left of Figure 2 is the **Ada-LSTM** model that consists of an adaptive attention module and an LSTM followed by a fully connected (*fc*) layer as the word predictor. Note that the $\boldsymbol{x}_t$ is adjusted accordingly to incorporate the predictor. On the right is the **MH-FC** model that adopts a multi-head attention module followed by an *fc* layer as the word predictor. Both model structures are commonly used [7, 8, 12, 43, 44].

The image captioning models are usually trained with cross-entropy loss in the first stage:

$$\mathcal{L} = \mathcal{L}_{ce}(\boldsymbol{p}, \boldsymbol{y}) \tag{17}$$

where $\boldsymbol{p} = (\boldsymbol{p}_t)_{t=0}^{l}$ is the predicted scores over vocabulary, l is the sentence length, and $\boldsymbol{y}$ is the ground truth label of a referenced caption. Then, the models are further optimized with the *SCST* algorithm from [84]. SCST optimizes non-differentiable evaluation metrics, e.g. CIDEr score [19], using reinforcement learning:

$$R = \mathbb{E}_{S^s, S^{greedy} \backsim \boldsymbol{p}}[\text{metric}(S^s, S^{gt}) - \text{metric}(S^{greedy}, S^{gt})] \tag{18}$$

where R is the reward, S^s is the sampled sentence from the predicted distribution $\boldsymbol{p} = (\boldsymbol{p}_t)_{t=0}^{l}$, S^{greedy} is the predicted sentence with greedy search, and S^{gt} is the referenced caption. The training objective is to obtain higher reward R. CIDEr is

usually adopted to calculate the reward and some papers also call this algorithm as CIDEr optimization [8, 12].

4. Explanation methods for image captioning models

In this section, we will explain how to adapt LRP [23], Grad-CAM, and Guided Grad-CAM [21] for use in attention-guided image captioning models. For brevity, we will use Grad* to denote Grad-CAM and Guided Grad-CAM.

Grad* methods are based on gradient backpropagation and can be directly applied to the attention-guided image captioning models. Grad* methods first backpropagate the gradient of a prediction till the visual feature $\boldsymbol{I}$, denoted as $g(\boldsymbol{I}) \in \mathbb{R}^{n_v \times d_v}$. Then, we can obtain a channel-wise weight from $g(\boldsymbol{I})$ for the visual feature $\boldsymbol{I}$, which is $\boldsymbol{w}_I = \sum_{k=1}^{n_v} g(\boldsymbol{I})_{(k)} \in \mathbb{R}^{d_v}$. $\boldsymbol{I}$ is further summed up over the feature dimension, weighted by $\boldsymbol{w}_I$, to generate the class activation map, $CAM = ReLU(\sum_{k=1}^{d_v} w_{I_{(k)}} \boldsymbol{I}_{(k)}) \in \mathbb{R}^{n_v}$, which reflects the importance of each pixel in the feature map. Grad-CAM reshapes and up-samples the class activation map to generate the image explanations. To obtain fine-grained and high-resolution explanations, Grad-CAM is fused with GuidedBackpropagation [22] by element-wise multiplication. GuidedBackpropagation can be easily implemented in pytorch by writing a custom `torch.autograd.Function` wrapping the stateless `ReLU` layers. This fused method is Guided Grad-CAM. The linguistic explanations of Grad* methods are obtained by summing up the gradients of the word embeddings. Next, we will elaborate on LRP for image captioning models.

We briefly introduce the basics of LRP. For an in-depth introduction, we refer to a book chapter like [85]. LRP explains neural networks by assigning a *relevance score* to every neuron within the network. The relevance assignment is achieved by backpropagating the relevance score of a target prediction along the network topology until the inputs according to LRP rules.

Consider the basic component of neural networks as a linear transformation followed by an activation $f(\cdot)$.

$$\begin{aligned} z_j &= \sum_i w_{ij} y_i + b_j \\ \hat{z}_j &= f(z_j) \end{aligned} \tag{19}$$

where y_i is the input neuron, z_j is the linear output, and $\hat{z}_j$ is the activation output. We use $R(\cdot)$ to denote the relevance score of a neuron. Suppose $R(\hat{z}_j)$ is known, we would like to distribute $R(\hat{z}_j)$ to all of its input neurons y_i, denoted as relevance

attribution $R_{i \leftarrow j}$. We refer to two LRP rules for relevance backpropagation that are frequently applied [23, 86, 87, 88]:

1. ϵ-rule
$$R_{i \leftarrow j} = R(\hat{z}_j) \frac{y_i w_{ij}}{z_j + \epsilon \odot \text{sign}(z_j)} \tag{20}$$
where ϵ is a small positive number. The stabilizer term $\epsilon \odot \text{sign}(z_j)$ guarantees that the denominator is non-zero.
2. α-rule
$$R_{i \leftarrow j} = R(\hat{z}_j) \left((1+\alpha) \frac{(y_i w_{ij})^+}{z_j^+} - \alpha \frac{(y_i w_{ij})^-}{z_j^-} \right) \tag{21}$$
where $\alpha \geqslant 0$, $(\cdot)^+ = \max(\cdot, 0)$, and $(\cdot)^- = \min(\cdot, 0)$. By separating $y_i w_{ij}$ and z_j into positive and negative parts, the α-rule ensures a boundedness of relevance terms. The parameter α determines the ratio of focus on positive and negative contribution during relevance backpropagation, from the output $\hat{z}_j$ to all of its inputs y_i.

The relevance of neuron y_i is the summation of all its incoming relevance attribution flows.
$$R(y_i) = \sum_j R_{i \leftarrow j} \tag{22}$$

LRP has provided plausible explanations for CNNs [23], RNNs such as LSTM [24], and also GNNs [72]. These modules are commonly used in image captioning models. To explain image captioning models with LRP, we define next how to apply LRP to the attention mechanisms.

From Section 3, we have seen that attention mechanisms involve non-linear interactions of the visual features and the hidden states of the decoder. However, the attention mechanisms mainly serve as weighting operations for features. Thus, we consider an attention mechanism as a linear combination over a set of features with weights such that LRP relevance scores are not backpropagated through the weights. This is consistent with the "signal-take-all" redistribution explored in [89]. In this way, we can directly apply LRP rules to distribute the relevance score of the context representation to the visual features according to the attention weights and bypass the computations within the attention mechanisms.

To give an overview of LRP for image captioning models, we take the Ada-LSTM model as an example and elaborate on each step of the explanation in Figure 3 and Algorithm 1. It is important to realize here, that LRP follows topologically the same flow as the gradient backpropagation (except the attention mechanisms) along the edges of a directed acyclic graph. The difference lies in replacing

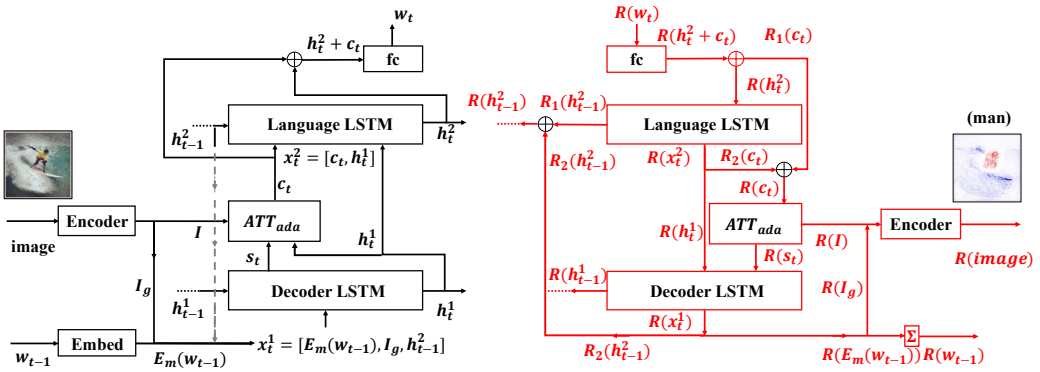


Figure 3: The LRP relevance backpropagation flow path through the Ada-LSTM model.

the partial derivatives on the edges by LRP redistribution rules motivated by the deep Taylor framework [61].

We initialize the relevance score of a target word, $R(w_T)$, from the output of the last *fc* layer (the logits). Then, as illustrated in Figure 3, LRP-type operations for computing $R(\cdot)$ are applied to the layers *fc*, $\oplus$, *Language LSTM*, ATT_{ada}, *Decoder LSTM*, and *Encoder*. The LRP operations used for these layers are shown as the $\Longrightarrow$ in Algorithm 1. For each word to be explained, LRP assigns a relevance score to every pixel of the input image ($R(image)$) and every word of the sequence input ($R(w_{T-1}), \ldots, R(w_1)$). We can visualize the image explanation as a heatmap after averaging $R(image)$ over the channel dimension. The relevance score of each preceding word is the summation of the relevance scores over the word embedding. In the experiments, we will also use the *relevance score* to denote the explanation scores of gradient-based methods.

5. Experiments

5.1. Model preparation and implementation details

We train the Ada-LSTM model and the MH-FC model on Flickr30K [32] and MSCOCO2017 [33] datasets for the following experiments[3].

Dataset: We prepare the Flickr30K dataset as per the Karpathy split [2]. 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

[3] `https://github.com/SunJiamei/LRP-imagecaptioning-pytorch.git`

Algorithm 1: LRP for Ada-LSTM model to explain w_T. The appearing symbols correspond to those in Figure 3. Notations: $\boldsymbol{\alpha}_t$ (Eq. (8)), β_t (Eq. (9)), and $\boldsymbol{s}_t$ (Eq. (5)), ϵ-rule (Eq. (20)), α-rule (Eq. (21)) , $[\cdot]$ denotes concatenation.

Require: $R(w_T), \boldsymbol{\alpha}_t, \beta_t$
Ensure: $R(image), R(w_{T-1}), \ldots, R(w_1)$

$R(w_T), \text{fc} \xRightarrow{\epsilon\text{-rule}} R(\boldsymbol{c}_T + \boldsymbol{h}_T^2)$

$R(\boldsymbol{c}_T + \boldsymbol{h}_T^2), \oplus \xRightarrow{\epsilon\text{-rule}} R_1(\boldsymbol{c}_T), R(\boldsymbol{h}_T^2)$

for $t \in [T, \ldots, 0, start]$ **do**

  $R(\boldsymbol{h}_t^2), \text{Language-LSTM} \xRightarrow{\epsilon\text{-rule}} R_2(\boldsymbol{c}_t), R(\boldsymbol{h}_t^1), R_1(\boldsymbol{h}_{t-1}^2)$

  $R_1(\boldsymbol{c}_t) + R_2(\boldsymbol{c}_t), ATT_{ada} \xRightarrow{\epsilon\text{-rule}} R(\boldsymbol{s}_t), R_t(\boldsymbol{I})$

  $R(\boldsymbol{h}_t^1), R(\boldsymbol{s}_t), \text{Decoder LSTM} \xRightarrow{\epsilon\text{-rule}} \underbrace{R(\boldsymbol{E_m}(w_{t-1})), R_t(\boldsymbol{I}_g), R_2(\boldsymbol{h}_{t-1}^2)}_{=R(\boldsymbol{x}_t^1)}, R(\boldsymbol{h}_{t-1}^1)$

  $R(\boldsymbol{E_m}(w_{t-1})) \xRightarrow{\sum} R(w_{t-1})$

end for

$\sum_t R_t(\boldsymbol{I}), \sum_t R_t(\boldsymbol{I}_g), \text{CNN} \xRightarrow{\epsilon\text{-rule},\alpha\text{-rule}} R(image)$

return $R(image), R(w_{T-1}), \ldots, R(w_1)$

sets are with 110000/5000/5000 images. Vocabularies are built only on the training set. We encode the words that appear less than 3 and 4 times as an unknown token <**unk**> for Flickr30K and MSCOCO2017, respectively, resulting in 9585 and 11026 vocabularies for the two datasets.

Encoder: We experiment with CNN and FasterRCNN as the image encoder. The CNN features are extracted from the pre-trained VGG16 [90] on ImageNet, specifically, we use the output of "*block5_conv3*" with a shape of $14 \times 14 \times 512$. The Faster RCNN encoder provides bottom-up image features corresponding to the candidate regions for object detection. We refer to **Detectron2** [91][4] to extract $n_v = 36$ features per image with 2048 channels each. Both the CNN features and the bottom-up features are further processed by a linear layer to generate the visual feature $\boldsymbol{I} \in \mathbb{R}^{n_v \times d_v}$.

Decoder and predictor: We train the Ada-LSTM and MH-FC models in Figure 2, with $d_v, d_h = 512$ for CNN features and $d_v, d_h = 1024$ for bottom-up features. The word embedding dimension is 512 for all the models. The number of multiple heads for MH-FC model is 8.

LRP parameters: We follow the suggestions of [86] on the best practice for LRP rules. We use α-rule for convolutional layers with $\alpha = 0$ and ϵ-rule for fully connected layers and LSTM layers with $\epsilon = 0.01$.

[4] `https://github.com/airsplay/py-bottom-up-attention.git`

Table 1: The performance of the Ada-LSTM model and the MH-FC model on the test set of Flickr30K and MSCOCO2017 datasets. The performance of AdaATT, SCST, BUTD models are from the corresponding papers. *BU* and *CNN* denote bottom-up features and CNN features, respectively.

Flickr30K	F_{BERT}	CIDEr	SPICE	ROUGE-L	METEOR
Ada-LSTM-CNN	90.56	51.54	13.87	46.79	20.18
Ada-LSTM-BU	90.04	63.03	16.52	49.32	21.94
MH-FC-CNN	90.54	53.65	14.85	46.92	20.71
MH-FC-BU	90.14	63.22	16.90	49.22	22.37
AdaATT [7] *(CNN+fc)*	–	53.10	14.50	46.70	20.40
MSCOCO2017	F_{BERT}	CIDEr	SPICE	ROUGE-L	METEOR
Ada-LSTM-CNN	91.83	107.03	19.49	54.34	26.10
Ada-LSTM-BU	91.01	111.87	19.17	55.04	25.93
MH-FC-CNN	91.85	108.16	20.10	54.42	26.45
MH-FC-BU	91.29	120.31	21.80	56.52	28.02
MSCOCO2014	F_{BERT}	CIDEr	SPICE	ROUGE-L	METEOR
AdaATT [7] *(CNN+fc)*	–	108.50	19.40	54.90	26.60
SCST:att2all [84] *(CNN+fc)*	–	114.00	–	55.70	26.70
BUTD [8] *(BU + LSTM)*	–	120.10	21.40	56.90	27.70

Training details: We adopt the Adam optimizer for training, with $\beta_1 = 0.8$, $\beta_2 = 0.999$, and a learning rate $lr = 0.0005$. We anneal *lr* by 20% when the CIDEr score does not improve for the last 3 epochs and stop the training when the CIDEr score does not improve for 6 epochs. We further optimize the models with the SCST optimization [84] using CIDEr score with $lr = 0.0001$. For the models using CNN features, we also fine-tune the CNN encoder with $lr = 0.0001$ before applying the SCST optimization.

Table 1 lists the performance of the Ada-LSTM model and the MH-FC model. We generate the captions with beam search (beam size=3) and report five evaluation metrics of image captioning task: METEOR [17], ROUGE-L [18], SPICE [20], CIDEr [19], and the F_{BERT}(idf) metric of BERTScore [92]. To validate our models, we include the performance of some benchmark image captioning models with similar model structures. **AdaATT** [7] is the first paper that proposes the adaptive attention mechanism. **SCST** [84] adapts reinforcement learning to image captioning and optimizes non-differentiable evaluation metrics. **BUTD** [8] adopts the bottom-up features and uses an LSTM as the word predictor. We can see that our models are properly trained and achieve comparable performance.

5.2. Explanation results and evaluation

Section 1 has shown some examples of the explanation results generated by LRP, Grad-CAM, Guided Grad-CAM. In comparison to attention heatmaps, we observe the following.

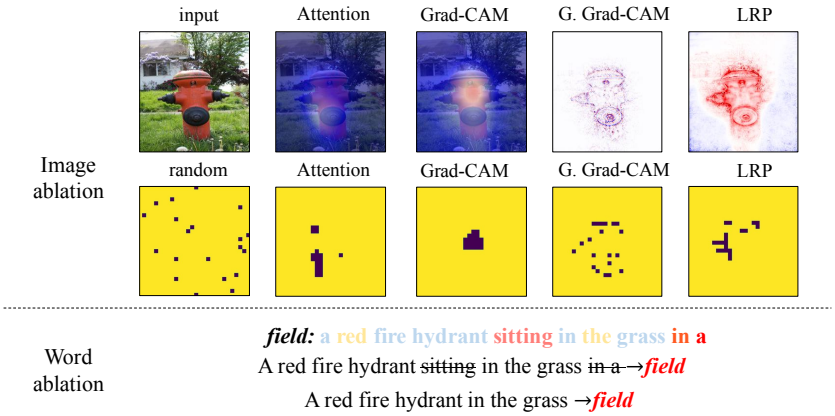


Figure 4: The image ablation (upper) and word ablation (lower) experiment. For image ablation, the first row shows the image explanations of the word *hydrant*, the second row shows the masked patches with high relevance scores.

Firstly, explanation methods can disentangle the contributions of the image input and the textual input, which is beyond the interpretability that attention mechanisms can provide. Secondly, some explanation methods provide high-resolution, pixel-wise image explanations, such as LRP and Guided Grad-CAM. Thirdly, LRP explicitly shows the positive and negative evidence used by the model to make decisions. In the following experiments, we will quantitatively evaluate the information content of attention, LRP, Grad-CAM, and Guided Grad-CAM with two ablation experiments and one object localization experiment. The ablation experiment aims to measure the information in the visual domain and the text domain, expressed by the relevance scores assigned to pixels and words. The object localization experiment evaluates the visual grounding property of relevance scores for image regions.

5.2.1. Ablation experiment

We conduct the ablation experiment for both the image explanations and the linguistic explanations, as illustrated in Figure 4. We demonstrate the approach using the same example in Section 1 based on the caption: *A red fire hydrant sitting in the grass in a field*.

The first row of Figure 4 shows the image explanations of the word *hydrant*, which highlight parts of the image related to the *hydrant*. To assess whether the highlighted areas contribute to the prediction, we firstly segment the image into non-overlapping 8×8 patches. Secondly, we sum the relevance scores within each

patch as the patch relevance. Thirdly, we mask the top-20 high-relevance patches with the training data mean, to eliminate the contributions of these patches. The top-20 high-relevance patches found by different explanation methods are shown in the second row of Figure 4. Finally, we predict a caption on the masked image. If the masked areas are important to the prediction, the model will be less confident to predict the target word or will not generate the target word at all from the masked image.

The linguistic explanations reflect the contributions of the previously generated sequence. For example, when generating the word *field*, the model perhaps uses the words “”*sitting*, “*in*”, and “*a*” as related evidence. Similar to the idea of the image ablation experiment, we remove the top-3 relevant words in the preceding sequence and forward the modified sequence to the model in a teacher-forcing manner. Finally, we observe the new probability of the target word. We do not modify the image for the word ablation experiment. If the removed words are strongly related to the prediction, the new probability of the target word will drop considerably compared to its original value.

We conduct the ablation experiment using image captioning models trained on the MSCOCO2017 dataset and CNN features. We report the results on the test set. For the word ablation experiment, we consider the predicted words with a sequence index greater than 6 so that there is a sufficiently long preceding word sequence to avoid evaluating purely frequency-based predictions in the experiment. For the image ablation experiment, we consider all the predicted object words. A random ablation is included as a baseline.

Figure 5 shows the results of word-ablation experiments. The words we explain are split into object words and stop-words. We show the frequency of probability drop, and the difference between the original word probability and the new word probability after the word deletion (denoted as an average score of probability drop). A higher average score of probability drop means the model is less confident to make the original prediction after ablation, therefore, the ablated words are more strongly related to the prediction. LRP and gradient-based explanation methods achieve a decrease in prediction probability more often and with greater impact than the random ablation, indicating that the words found by explanation methods are used by the model as important evidence to predict the target word. LRP achieves both the highest frequency and the highest average score of probability drop.

In our word ablation experiment, we use 8 heads for the multi-head attention mechanism of the MH-FC model, resulting in 8 sets of attention weights. This is computationally too heavy for use in the image ablation experiment. We, there-

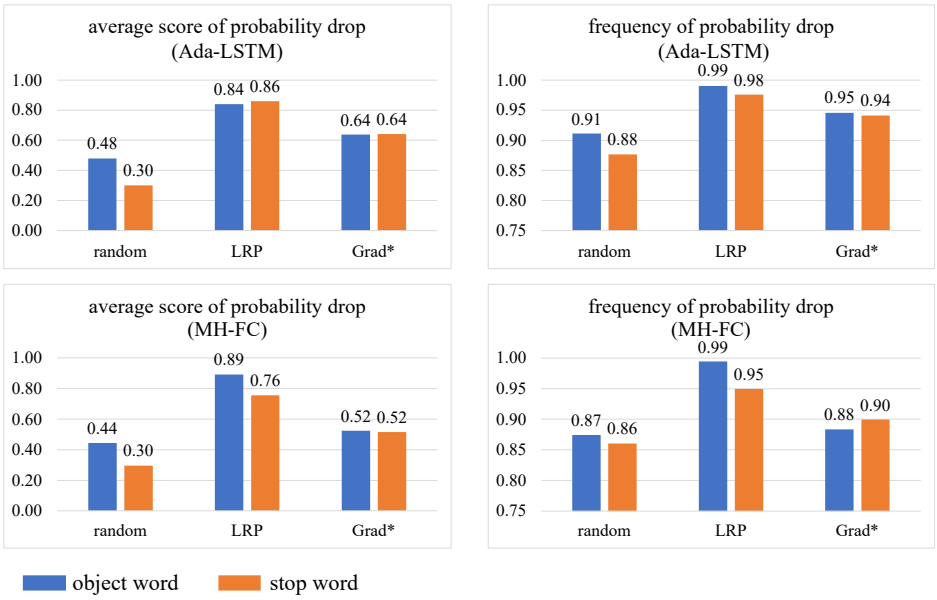


Figure 5: The results of the word ablation experiment on MSCOCO2017 test set. The numbers of evaluated object words and stop-words are 3,710 and 11,686 for the Ada-LSTM model, and 3,359 and 11,512 for the MH-FC model. Higher average scores and higher frequencies are better.

fore, implement the image ablation experiment with the Ada-LSTM model and show how often the model *fails to* generate the target word after the image ablation, as shown in Figure 6 (left). We can see that high-resolution explanations from the evaluated explanation methods LRP, Guided Grad-CAM, and Guided-Backpropagation achieve a higher frequency of object words vanishing, indicating that the highlighted areas are related to the evidence for model decisions.

With the above experiment results, we verify that using explanation methods adds information compared to relying on attention heatmaps alone.

5.2.2. Measuring the correlation of explanations to object locations

Many studies employ attention heatmaps as a tool to verify the visual grounding property qualitatively [4, 7, 12, 10, 35]. In this part, we will quantify the correlation of explanation results to object locations and show that high-resolution explanations can also achieve a high correlation to the object locations.

To assess the correlation of explanations to object locations, we utilize the bounding box annotations of the MSCOCO2017 dataset and extend the *correctness* measure from [93], which evaluates the grounding property of attention heatmaps, to the explanation results. For a correctly predicted object word, we

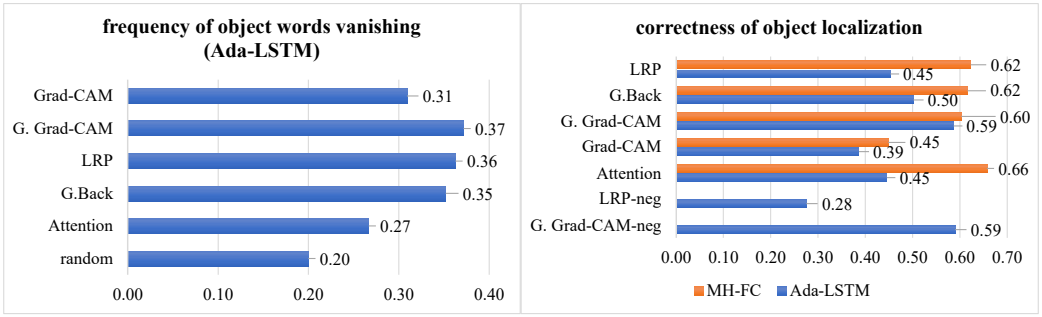


Figure 6: Left: the results of the image ablation experiment using the Ada-LSTM model. There are 9,645 evaluated object words. Higher frequency is better. Right: the average *correctness* of object localization. There are 4,691/4,649 correctly predicted words for the Ada-LSTM/MH-FC model using the MSCOCO2017 test set. Higher *correctness* scores mean better localization. G. denotes Guided. G.Back denotes GuidedBackpropagation.

first obtain the relevance scores of the image input, $R(image)$, with explanation methods and average $R(image)$ over the channel dimension, resulting in a spatial explanation $\boldsymbol{E} \in \mathbb{R}^{h \times w}$, where h and w are the height and width of the image. We keep the positive scores of $\boldsymbol{E}$ for object localization. The $correctness$ is the proportion of the relevance scores within the bounding box.

$$\boldsymbol{E}_p = norm(\max(\boldsymbol{E}, 0)) \tag{23}$$

$$correctness = \frac{\sum_{ij \in bbox} \boldsymbol{E}_p[i,j]}{\sum_{ij} \boldsymbol{E}_p[i,j]} \in [0,1] \tag{24}$$

where the $norm(\cdot)$ is the normalization with the maximal absolute value. For the MH-FC model with the multi-head attention mechanism, we generate the explanations for each head, $R(image)^{(i)}$, by only backpropagating the relevance scores or gradients through head i. The *correctness* of the MH-FC model is the maximum across the $correctness^{(i)}$ of all the heads, i.e.

$$correctness_{MH-FC} = \max_i(correctness^{(i)}) \tag{25}$$

Higher *correctness* means the relevance scores concentrate more within the bounding box, indicating a better grounding property. Figure 6 (right) shows the average $correctness$ of all the correctly predicted object words across the MSCOCO2017 test set, evaluated with image captioning models trained using CNN features.

First of all, the MH-FC model achieves consistently higher *correctness* than the Ada-LSTM model, indicating that there is at least one head of the MH-FC

model that accurately locates the object, especially for attention and LRP where there is a large discrepancy of the *correctness* between the Ada-LSTM and the MH-FC models.

Secondly, high-resolution explanations provided by LRP, Guided Grad-CAM, and GuidedBackpropagation achieve comparable or higher *correctness* than attention. The notable exception is due to the spatial localization property of the multiple heads in the MH-FC model. Combining the results of the ablation experiments, explanation methods tend to find parts of objects which correlate well to the prediction.

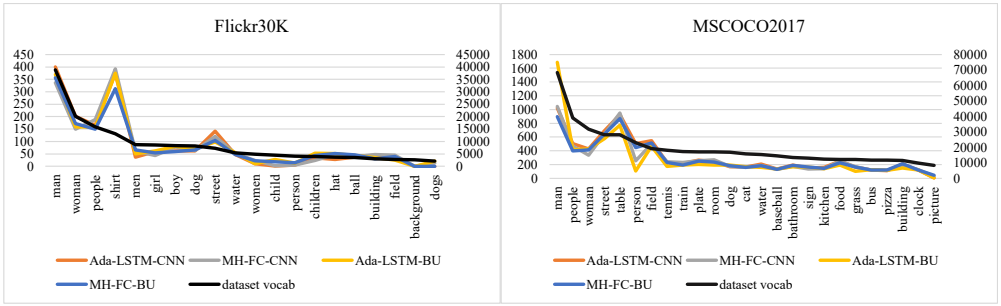
Thirdly, to further get insights into the role of the sign of the relevance scores, we calculate the *correctness* using the absolute value of the negative relevance scores, $\boldsymbol{E}_n = norm(\max(-\boldsymbol{E}, 0))$. As shown in Figure 6 (right), the low *correctness* of "LRP-neg" and the high $correctness$ of "G. Grad-CAM-neg" verifies that the positive/negative sign of LRP relevance scores reveals the support/opposition of a pixel to the predictions, while for Guided Grad-CAM, *both* positive and negative relevance scores are related to the predictions and irrelevant pixels have low absolute relevance scores.

Last but not least, our *correctness* evaluation results over various explanation methods under the image captioning scenario are consistent with some prior works. GuidedBackpropagation and LRP generate more coherent explanations for MRI data than other gradient-based methods [94], despite failing certain sanity checks postulated in [95]. This underlines the importance of considering multiple criteria in contrast to decisions based on selected axiomatic requirements. Furthermore, the sign of LRP relevance scores is meaningful [86]. Both properties can be helpful for *model debugging* [95, 96, 97]. In the next section, we will show how we use LRP to *"debug"* and improve image captioning models.

5.3. Reducing object hallucination with explanation

In our experiment, we observe the common hallucination problem of image captioning models. Image captioning models sometimes generate object words that are not related to the image content, which is possibly caused by the learned language priors. The vocabulary and sentence patterns of the image-caption pairs are intrinsically biased toward frequent occurrences. As illustrated in Figure 7, the vocabulary count distribution of the predicted words is close to that of the training vocabulary.

A language bias can be helpful for image captioning models. [28] learns the inductive language bias to guide the model to deduce the object relations and attributions. However, it can also cause mistakes. For example, the models could be



(a) Top-20 frequent object words (b) Top-25 frequent object words

Figure 7: The counts of top-k frequently appearing object words in Flick30K and MSCOCO2017 training set (right ordinate) and the counts of the predicted object words in the test set (left ordinate).

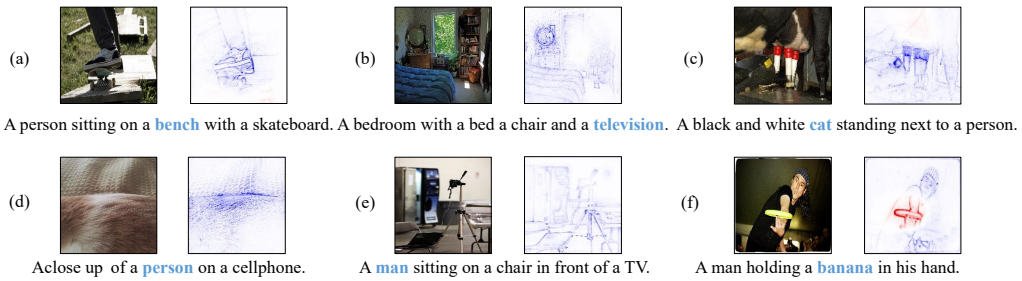


Figure 8: The LRP image explanations of hallucinated words (blue) in the generated image captions. Blue and red pixels indicate negative and positive relevance scores, respectively.

flawed when predicting gender [29] or always paint *bananas yellow* irrespective of their actual color [30, 31]. To this end, we explore the explanations of hallucinated words and investigate using approaches from explainability to reduce object hallucination.

5.3.1. Exploring the explanations of hallucinated words

Based on the findings in Section 5.2 that high-resolution explanations obtained by LRP and Guided Grad-CAM correlate to the object locations and reflect well the related evidence for predictions, we explore the difference of image explanations between grounded (true-positive) and hallucinated (false-positive) object words. Figure 8 illustrates some examples of LRP image explanations for hallucinated words.

In Figure 8 (a) to (e), the LRP image explanations show more negative scores,

Table 2: The AUC scores calculated with different statistics and explanation methods. G. denotes Guided and G.Back denotes GuidedBackpropagation. Higher AUC means the statistic can better differentiate the hallucinated words. The AUC score calculated with $1 - \beta$ is **0.6005**.

AUC	LRP	G.Grad-CAM	Grad-CAM	Attention	G.Back
quantile-5%($\boldsymbol{E}$)	**0.6022**	0.4392	0.5936	0.5598	0.4621
quantile-50%($\boldsymbol{E}$)	0.5821	0.5358	0.5730	0.5136	0.5168
max($\boldsymbol{E}$)	0.5168	0.5743	0.5580	0.5169	0.5575
mean($\boldsymbol{E}$)	0.5798	0.4319	0.5857	0.5308	0.4648

implying that the model generates hallucinated words mainly with the linguistic information rather than the image information. In Figure 8 (f), the model mistakes the yellow frisbee for a banana, evidenced by red pixels (positive scores).

We now quantify the difference in image explanations between true-positive and false-positive object words. Specifically, we use the statistics of image explanations (the $\boldsymbol{E}$ mentioned in Section 5.2.2) to differentiate the hallucinated words.

We assign a label 1/0 to the true-positive/false-positive predicted words, respectively. Each word is also assigned with a statistic calculated from the image explanation $\boldsymbol{E}$, such as the maximum value (*max(*$\boldsymbol{E}$*)*), the 5% and 50% quantiles (*quantile-5%/50%(*$\boldsymbol{E}$*)*), and the mean (*mean(*$\boldsymbol{E}$*)*). We also evaluate $1 - \beta$ from Eq. (10) of the adaptive attention mechanism. We remind that the adaptive attention mechanism contains a sentinel feature $\boldsymbol{s}_t$ that represents the text-dominant information. It then learns a weight, β_t, which controls the proportion of linguistic information used for predictions. Thus, it is a model-intrinsic baseline to show differences between grounded and hallucinated object words.

We calculate the AUC scores, using the labels and statistics of true-positive and false-positive words. A higher AUC score indicates a better differentiation between hallucinated and grounded words. Table 2 lists the AUC scores computed with various explanation methods. We conduct the experiment with the Ada-LSTM model trained on Flickr30K dataset, because its vocabularies are more imbalanced than that of the MSCOCO2017 dataset. The results are reported on the test set of Flickr30K. The evaluated words are the top-20 frequent object words[5] with 715 false-positive and 1,027 true-positive cases.

The LRP *quantile-5%(*$\boldsymbol{E}$*)* achieves a slightly higher AUC score than $1-\beta$ and can weakly recognize the hallucinated words, which indicates that true-positive words are usually with higher LRP *quantile-5%(*$\boldsymbol{E}$*)* and false-positive words are

[5]These most frequent object words are: dogs, building, person, background, field, women, hat, ball, children, child, water, street, boy, dog, girl, men, shirt, people, woman, man.

with lower LRP *quantile-5%(**E**)*. The statistics of LRP all obtain AUC scores greater than 0.5, which verifies that the LRP image explanations consist of lower relevance scores for false-positive words, and thus, reflect less supporting evidence for the hallucinated words.

In the next section, we will introduce a fine-tuning strategy that builds upon LRP-based explanations to reduce object hallucination.

5.3.2. Using LRP explanations to reduce object hallucination

We introduce an LRP-inference fine-tuning (LRP-IFT) strategy that can help to de-bias a pre-trained image captioning model and reduce object hallucination. We design a re-weighting mechanism inspired by two properties of LRP explanations: 1) meaningfulness of the positive and negative sign of LRP relevance scores, indicating the support and opposition to the predictions; 2) the property of finding the regions and evidence in the image used by the model to make predictions. In particular, we design weights for the input features of the last *fc* layer using the LRP relevance scores and embed the re-weighted features into the model for fine-tuning. We elaborate on each step of the fine-tuning strategy with Algorithm 2 and detail the underlying idea as follows.

To fine-tune an image captioning model $\mathcal{M}$, we generate an initial caption first.

$$(\boldsymbol{p}_t)_{t=0}^{l} = \mathcal{M}(\mathcal{I}) \tag{26}$$

$$h(w_t) = \text{argmax}(\boldsymbol{p}_t) \tag{27}$$

where $\mathcal{I}$ is the image , $\boldsymbol{p}_t \in \mathbb{R}^V$ is the probability distribution over the vocabulary at time step t, V is the vocabulary size, and $h(w_t)$ is the label of the word w_t.

If w_t is ***not*** a stop-word, we will explain the predicted label $h(w_t)$ through the last *fc* layer using LRP and obtain the relevance scores of the context representation and the hidden state, $R(\boldsymbol{c}_t)$ and $R(\boldsymbol{h}_t)$. (Remember that $\boldsymbol{c}_t + \boldsymbol{h}_t$ is the input of the last *fc* layer.)

We then normalize $R(\boldsymbol{c}_t)$ and $R(\boldsymbol{h}_t)$ with the maximal absolute value, so that their values are in $[-1, +1]$, and generate a new word probability distribution $\hat{\boldsymbol{p}}_t$ as follows.

$$\boldsymbol{\omega}_{\boldsymbol{c}_t} = norm(R(\boldsymbol{c}_t)) + 1 \in [0, 2] \tag{28}$$

$$\boldsymbol{\omega}_{\boldsymbol{h}_t} = norm(R(\boldsymbol{h}_t)) + 1 \in [0, 2] \tag{29}$$

$$\hat{\boldsymbol{p}}_t = fc(\boldsymbol{\omega}_{\boldsymbol{c}_t} \odot \boldsymbol{c}_t + \boldsymbol{\omega}_{\boldsymbol{h}_t} \odot \boldsymbol{h}_t) \tag{30}$$

Algorithm 2: LRP-inference fine-tuning

Require: predicted sequence:$(w_t)_{t=0}^{l}$, predicted probability:$(\boldsymbol{p}_t)_{t=0}^{l}$
Ensure: LRP-inference prediction:$(\hat{\boldsymbol{p}}_t)_{t=0}^{l}$
for $t \in [0, \ldots, l]$ **do**
  if w_t not in stop-words **then**
    $\boldsymbol{p}_t^{h(w_t)}, fc \xrightarrow{\text{LRP}} R(\boldsymbol{c}_t), R(\boldsymbol{h}_t)$
    $R(\boldsymbol{c}_t) \xrightarrow{\text{Eq. (28)}} \boldsymbol{\omega}_{\boldsymbol{c}_t}$
    $R(\boldsymbol{h}_t) \xrightarrow{\text{Eq. (29)}} \boldsymbol{\omega}_{\boldsymbol{h}_t}$
    $\hat{\boldsymbol{p}}_t = fc(\boldsymbol{\omega}_{\boldsymbol{c}_t} \odot \boldsymbol{c}_t + \boldsymbol{\omega}_{\boldsymbol{h}_t} \odot \boldsymbol{h}_t)$
  else
    $\hat{\boldsymbol{p}}_t = \boldsymbol{p}_t$
  end if
end for
return $(\hat{\boldsymbol{p}}_t)_{t=0}^{l}$

In LRP explanations, positive relevance is attributed to features supporting the prediction of the target class and negative relevance is attributed to contradicting features. The operations performed in Eqs. (28) and (29) construct a weight ω such that $\omega < 1$ for the opposing features and $\omega > 1$ for the supporting features. The re-weighting mechanism will thus up-scale the supporting features and down-scale the opposing ones.

During fine-tuning, we use the LRP-inference prediction $\hat{\boldsymbol{p}} = (\hat{\boldsymbol{p}}_t)_{t=0}^{l}$ to calculate the loss. For the cross-entropy loss function, we can combine both the original loss and the new loss with a parameter $\lambda \in [0, 1]$. The loss function from Eq. (17) is updated as follows.

$$\mathcal{L} = \lambda \mathcal{L}_{ce}(\boldsymbol{p}, \boldsymbol{y}) + (1 - \lambda)\mathcal{L}_{ce}(\hat{\boldsymbol{p}}, \boldsymbol{y}) \tag{31}$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss and $\boldsymbol{y}$ is the ground truth label. We can also use $\hat{\boldsymbol{p}}$ for the SCST optimization and the reward formula from Eq. (18) is re-written as follows.

$$R = \mathbb{E}_{S^s, S^{greedy} \backsim \hat{\boldsymbol{p}}}[\text{metric}(S^s, S^{gt}) - \text{metric}(S^{greedy}, S^{gt})] \tag{32}$$

where we replace the original probability distribution $\boldsymbol{p}$ with the LRP-inference one $\hat{\boldsymbol{p}}$. R is the reward, S^s is the sampled sentence, S^{greedy} is the greedily sampled sentence, and S^{gt} is the referenced caption.

Different from standard fine-tuning, LRP-IFT disentangles the contributions of the visual information, $R(\boldsymbol{c}_t)$, and the hidden state, $R(\boldsymbol{h}_t)$. It selects and fine-tunes the more related features rather than training all the features generally.

Table 3: The mean average precision (mAP) of the predicted frequent object words. *(ce)* denotes that the models are trained only with cross-entropy loss and the other models are optimized with SCST. *BU* and *CNN* denote bottom-up features and CNN features. Bold numbers indicate better results. Higher mAP means less object hallucination.

dataset	Flickr30K		MSCOCO2017	
mAP	baseline	LRP-IFT	baseline	LRP-IFT
Ada-LSTM-CNN	52.95	**54.47**	72.29	**73.85**
Ada-LSTM-BU	63.84	**64.61**	78.57	**80.55**
MH-FC-CNN	55.98	**57.71**	**73.74**	73.42
MH-FC-BU	64.46	**64.98**	**78.10**	77.71
Ada-LSTM-CNN (ce)	58.53	**60.80**	73.65	**74.00**
Ada-LSTM-BU (ce)	60.70	**65.01**	79.06	**79.80**
MH-FC-CNN (ce)	55.50	**59.23**	**77.15**	76.87
MH-FC-BU (ce)	64.08	**66.10**	81.02	**81.16**

To evaluate the performance of the LRP-IFT, we observe the mean average precision (mAP) of the frequent object words[6]. The motivation of LRP-IFT is to guide the model to make more grounded captions rather than thoroughly enumerate all objects within an image. Therefore, we do not use the recall and F1 score.

Table 3 lists the mAP of the models with or without LRP-IFT. We implement the LRP-IFT on two sets of pre-trained models. The first set of models are from Table 1 that are optimized with SCST optimization, and we refer to Eq. (32) to fine-tune the models for one epoch. The second set of models are trained only with cross-entropy loss, denoted as *(ce)* in the table and we refer to Eq. (31) with $\lambda = 0.5$ to fine-tune the models for one epoch. For the baseline models, we fine-tune the two sets of models with standard SCST optimization or cross-entropy loss with the same training hyperparameters.

As shown in Table 3, the mAP is effectively improved after LRP-IFT for both sets of models except the MH-FC models trained on the MSCOCO2017 dataset. We discuss the mAP results from three aspects: 1) the MSCOCO2017 dataset has a more balanced vocabulary and more training data than the Flickr30K dataset, which results in less biased models. This also explains the more pronounced improvement of mAP on the Flickr30K dataset; 2) the multi-head attention mechanism has better grounding property as discussed in Section 5.2.2, which is the

[6]The frequent object words of Flickr30K are the same as Section5.3.1. The top-25 frequent object words of MSCOCO2017 datasets include: clock, kitchen, picture, water, food, pizza, grass, building, bus, sign, bathroom, baseball, dog, room, cat, plate, train, field, tennis, person, table, street, woman, people, man.

Table 4: The performance of the Ada-LSTM model and the MH-FC model with or without LRP-IFT on the test set of Flickr30K and MSCOCO2017 datasets. L. denotes LRP-inference fine-tuned models. *(ce)* denotes that the models are trained only with cross-entropy loss and the other models are further optimized with SCST. *BU* and *CNN* denote bottom-up features and CNN features. F_B: F_{BERT}, C: CIDEr, S: SPICE, R: ROUGE-L, M: METEOR.

dataset	Flickr30K					MSCOCO2017				
	F_B	C	S	R	M	F_B	C	S	R	M
Ada-LSTM-CNN	90.6	51.1	13.9	46.4	20.0	91.7	107.1	19.5	54.2	26.1
L.Ada-LSTM-CNN	90.6	50.9	14.0	46.7	20.1	91.2	106.8	19.2	54.0	26.0
Ada-LSTM-BU	90.0	63.8	16.4	49.3	22.1	91.0	111.6	19.2	55.3	25.9
L.Ada-LSTM-BU	90.0	61.9	16.5	49.2	21.9	91.0	111.1	19.3	55.2	25.9
MH-FC-CNN	89.9	53.3	14.5	46.5	20.5	91.1	108.8	20.1	54.6	26.5
L.MH-FC-CNN	89.7	52.7	14.2	46.0	20.0	91.0	107.5	20.1	54.3	26.3
MH-FC-BU	90.1	63.5	17.1	49.4	22.5	91.3	120.9	21.8	56.6	28.1
L.MH-FC-BU	90.1	63.5	17.0	49.1	22.4	91.3	120.8	21.9	56.7	28.1
Ada-LSTM-CNN(ce)	89.7	44.6	13.3	44.4	19.0	91.7	96.2	18.1	52.9	25.1
L.Ada-LSTM-CNN(ce)	89.6	43.5	13.1	44.0	18.9	91.5	92.3	18.0	52.1	24.9
Ada-LSTM-BU(ce)	90.0	53.3	15.6	47.3	21.2	91.9	107.4	19.8	54.9	26.9
L.Ada-LSTM-BU(ce)	89.9	52.2	15.6	47.0	21.2	91.7	103.0	19.6	54.1	26.3
MH-FC-CNN (ce)	89.7	46.5	13.7	44.8	19.4	90.7	97.1	18.8	53.1	25.5
L.MH-FC-CNN(ce)	89.6	47.3	14.1	45.6	19.8	90.7	97.2	18.8	53.0	25.4
MH-FC-BU(ce)	90.0	52.3	15.2	46.2	20.9	91.8	105.8	19.9	54.7	26.7
L.MH-FC-BU(ce)	89.8	52.7	15.3	46.5	21.0	91.8	105.7	19.9	54.6	26.6

possible reason why LRP-IFT obtains similar mAP for the MH-FC model trained on the MSCOCO2017 dataset; 3) as expected, the image captioning models with bottom-up features consistently obtain higher mAP than those with CNN features, demonstrating the potential of better feature representation for visual-language models such as VIVO [47] and OSCAR [48].

Furthermore, LRP-IFT maintains the overall performance on the sentence level, as shown in Table 4. Figure 9 illustrates some example captions of the baseline models and the LRP-inference fine-tuned models. LRP-IFT is conducted on the non-stop words and can improve the precision of the frequent object words. As shown in Figure 9, with LRP-IFT, the model can correct or remove the hallucinated words and maintain the sentence structure. This can partially explain why the sentence-level performance is very close to that of the baseline models. We will provide more detailed analyses of the sentence-level performance in Section 5.4.

From the above analyses, the LRP-IFT can effectively de-bias and reduce object hallucination for a biased image captioning model, meanwhile, maintain the sentence-level performance in terms of F_{BERT}, CIDEr, SPICE, METEOR, and ROUGE-L. On the other hand, this fine-tuning strategy does not degrade the performance of a less biased image captioning model notably. We remark that the

Figure 9: Examples of the captioning results with or without LRP-IFT. Blue color marks the hallucinated words and red color marks the words corrected by LRP-IFT.

LRP-IFT requires no additional training parameters and human annotations. The fine-tuning procedure is also analogous to the human's recognition process that we first build prior knowledge by learning the objects, relations, and attributes and update related features when facing new shifts in distributions.

5.4. Discussion and outlook

In the experiments of LRP-IFT, we have observed that LRP-IFT alleviates the object hallucination issue of image captioning models measurably. However, we can also see that LRP-IFT does not effectively improve sentence-level performance. In this part, we will further analyze the effects of the LRP re-weighting mechanism and we will take a closer look at the samples where LRP-IFT improves the sentence-level performance. We conclude by proposing a potential future direction where the LRP-inference training can be helpful.

5.4.1. On limitations of the LRP re-weighting mechanism

We performed an analysis on the samples where LRP-IFT improves or degrades sentence-level performance. At first, for each word in a *ground truth* caption, we computed the count of that word within the *training set*. Then, for each

Table 5: The average $c(S^{gt})$ over the ground truth captions from two sets of samples: the LRP-IFT-improved set, where LRP-IFT increases the CIDEr scores, and the LRP-IFT-degraded set, where LRP-IFT decreases the CIDEr scores. *(ce)* denotes that the models are trained only with cross-entropy loss. The other models are further optimized with SCST. *BU* and *CNN* denote bottom-up features and CNN features. Bold numbers indicate lower counts of the ground truth words in the training set. This statistic can be interpreted as a heuristic for training data density.

dataset	Flickr30K		MSCOCO2017	
average counts	LRP-IFT-improved	LRP-IFT-degraded	LRP-IFT-improved	LRP-IFT-degraded
Ada-LSTM-CNN	**26.1**	35.2	**123.7**	134.0
Ada-LSTM-BU	**30.1**	31.4	**130.8**	134.7
MH-FC-CNN	**29.3**	31.4	**124.4**	132.8
MH-FC-BU	**29.3**	29.7	**118.7**	139.0
Ada-LSTM-CNN (ce)	34.4	**28.5**	**124.4**	137.0
Ada-LSTM-BU (ce)	31.7	**28.1**	**119.0**	150.6
MH-FC-CNN (ce)	**29.4**	30.6	**128.0**	142.6
MH-FC-BU (ce)	**22.6**	35.9	**124.7**	148.5

ground truth caption in the test set, we find the minimum of the word counts, denoted as $c(S^{gt})$, over the *non-stop* words in the caption S^{gt}:

$$c(S^{gt}) = \min_{w_t \in S^{gt}} count(w_t) \tag{33}$$

where $count(w_t)$ returns the counts of the word w_t in the training set.

This statistic $c(S^{gt})$ for test set captions can be viewed as a heuristic 1-gram estimate of the training data density for the linguistic modality of image captioning. For images with multiple ground truth captions, we take the minimum of $c(S^{gt})$ over all the captions of one image. We verified that taking the average yields the same qualitative results.

Finally, we compute the average of this heuristic $c(S^{gt})$ for two sets of images: 1) the images on which LRP-IFT improves the predictions compared to the baseline model and 2) the images for which LRP-IFT degrades the predictions compared to the baseline model. We refer to the sentence-level evaluation metrics, such as the CIDEr score, to separate the two sets of image samples.

Table 5 lists the results of average $c(S^{gt})$ using CIDEr scores for performance comparison. We observe a clear correlation across most of the models (except only one): The LRP-IFT-improved set exhibits a lower average $c(S^{gt})$, while the LRP-IFT-degraded set shows a higher average $c(S^{gt})$. In summary, LRP-IFT achieves a tradeoff. It performs worse on those test images with a higher estimate of the sample density, where the base model seemingly generalizes sufficiently well. On the other hand, it achieves an improvement on images with lower training data density. The results using other metric scores for comparison lead to the

same finding. This makes intuitively sense as one can expect that captions supported by a higher amount of training data would profit less from learning with explanations. A similar correlation for using explanations to improve age prediction models using image data is reported in [98]. The authors observe that using explanations improves predictions on the poorly performing age subset 48-53 years, which has a small sample size, while slightly degrades the performance on age subsets with larger sample sizes.

There are further possible reasons for the non-improved sentence-level performance. [27] points out that hallucinating less does not necessarily render higher sentence-level evaluation metrics [29, 31, 99], which is also in line with our observations in Table 4. Furthermore, LRP-IFT implements the re-weighting mechanism on top of pre-trained models as a fine-tuning step, making it challenging to achieve larger changes over pre-trained models.

5.4.2. An outlook for re-weighting mechanisms based on explanations

Based on the above analyses, we surmise that the LRP re-weighting mechanism could be helpful for novel object captioning (NOC). NOC aims to predict those object words that are unseen by the model during training. It also faces the challenge of unbalanced training data, in an even more extreme case where some object words are not shown in the training data. For example, [52] proposed a pointing mechanism to combine the sentence correlation representation and object representation, which dynamically decides whether to include an object word from a detection model. The LRP re-weighting mechanism could be helpful here to better guide the model when and where to include the detected objects in the caption.

6. Conclusion

We adapt LRP and gradient-based explanation methods to explain the attention-guided image captioning models beyond visualizing attention. With extensive qualitative and quantitative experiments, we demonstrate that explanation methods provide more interpretable information than attention, disentangle the contributions of the visual and linguistic information, help to debug the image captioning models such as mining the reasons for the hallucination problem. With the properties of LRP explanations, we propose an LRP-inference fine-tuning strategy that can successfully de-bias image captioning models and alleviate object hallucination. The proposed fine-tuning strategy requires no additional annotations and training parameters.

Acknowledgements

This work was supported by the Singaporean Ministry of Education of Singapore (MoE) Tier 2 grant MOE2016-T2-2-154. This work was also partly supported by the German Ministry for Education and Research as BIFOLD (ref. 01IS18025A and ref. 01IS18037A), and TraMeExCo (ref. 01IS18056A), by the European Union's Horizon 2020 programme (grant no. 965221) and by the Research Council of Norway, via the SFI Visual Intelligence, project number 309439.

References

[1] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.

[2] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.

[3] M. Soh, Learning cnn-lstm architectures for image caption generation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep (2016).

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.

[5] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, R. R. Salakhutdinov, Review networks for caption generation, in: Advances in Neural Information Processing Systems 29, 2016, pp. 2361–2369.

[6] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4651–4659.

[7] J. Lu, C. Xiong, D. Parikh, R. Socher, 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, 2017, pp. 375–383.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077–6086.

[9] L. Huang, W. Wang, Y. Xia, J. Chen, Adaptively aligned image captioning via adaptive attention time, in: Advances in Neural Information Processing Systems, 2019, pp. 8942–8951.

[10] W. Wang, Z. Chen, H. Hu, Hierarchical attention network for image captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8957–8964.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[12] L. Huang, W. Wang, J. Chen, X.-Y. Wei, Attention on attention for image captioning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4634–4643.

[13] G. Li, L. Zhu, P. Liu, Y. Yang, Entangled transformer for image captioning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8928–8937.

[14] J. Yu, J. Li, Z. Yu, Q. Huang, Multimodal transformer with multi-view visual representation for image captioning, IEEE Transactions on Circuits and Systems for Video Technology (2019).

[15] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10578–10587.

[16] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[17] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.

[18] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[19] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.

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

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE ICCV, 2017, pp. 618–626.

[22] J. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, in: ICLR (workshop track), 2015.

[23] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE 10 (7) (2015) e0130140.

[24] L. Arras, G. Montavon, K.-R. Müller, W. Samek, Explaining recurrent neural network predictions in sentiment analysis, in: Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2017, pp. 159–168.

[25] J. Sun, A. Binder, Generalized patternattribution for neural networks with sigmoid activations, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–9.

[26] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applications, Proceedings of the IEEE 109 (3) (2021) 247–278.

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

[28] X. Yang, K. Tang, H. Zhang, J. Cai, Auto-encoding scene graphs for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10685–10694.

[29] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, A. Rohrbach, Women also snowboard: Overcoming bias in captioning models, in: European Conference on Computer Vision, Springer, 2018, pp. 793–811.

[30] R. Cadene, C. Dancette, H. Ben younes, M. Cord, D. Parikh, Rubi: Reducing unimodal biases for visual question answering, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019, pp. 839–850.

[31] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, D. Parikh, Taking a hint: Leveraging explanations to make vision and language models more grounded, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2591–2600.

[32] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[34] Y. Pan, T. Yao, Y. Li, T. Mei, X-linear attention networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10971–10980.

[35] J. Ji, C. Xu, X. Zhang, B. Wang, X. Song, Spatio-temporal memory attention for image captioning, IEEE Transactions on Image Processing 29 (2020) 7615–7628.

[36] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.

[37] S. Herdade, A. Kappeler, K. Boakye, J. Soares, Image captioning: Transforming objects into words, in: Advances in Neural Information Processing Systems, 2019, pp. 11137–11147.

[38] S. Chen, Q. Jin, P. Wang, Q. Wu, Say as you wish: Fine-grained control of image caption generation with abstract scene graphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 9962–9971.

[39] T. Yao, Y. Pan, Y. Li, T. Mei, Hierarchy parsing for image captioning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2621–2629.

[40] T. Yao, Y. Pan, Y. Li, T. Mei, Exploring visual relationship for image captioning, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 684–699.

[41] Z. Shi, X. Zhou, X. Qiu, X. Zhu, Improving image captioning with better use of caption, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7454–7464.

[42] F. Liu, X. Ren, Y. Liu, K. Lei, X. Sun, Exploring and distilling cross-modal information for image captioning, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, pp. 5095–5101.

[43] Z. Zha, D. Liu, H. Zhang, Y. Zhang, F. Wu, Context-aware visual policy network for fine-grained image captioning., IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).

[44] M. Cornia, L. Baraldi, R. Cucchiara, Show, control and tell: A framework for generating controllable and grounded captions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8307–8316.

[45] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, J. Gao, Unified vision-language pre-training for image captioning and vqa., in: AAAI, 2020, pp. 13041–13049.

[46] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[47] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, Z. Liu, Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training, in: AAAI, 2021.

[48] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, in: European Conference on Computer Vision, Springer, 2020, pp. 121–137.

[49] Z. Wang, Z. Huang, Y. Luo, Human consensus-oriented image captioning, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2020, pp. 659–665.

[50] J. Lu, J. Yang, D. Batra, D. Parikh, Neural baby talk, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7219–7228.

[51] S. Venugopalan, L. Anne Hendricks, M. Rohrbach, R. Mooney, T. Darrell, K. Saenko, Captioning images with diverse objects, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5753–5761.

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

[53] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, T. Darrell, Deep compositional captioning: Describing novel object categories without paired training data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1–10.

[54] Q. Feng, Y. Wu, H. Fan, C. Yan, M. Xu, Y. Yang, Cascaded revision network for novel object captioning, IEEE Transactions on Circuits and Systems for Video Technology (2020).

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

[56] L. Guo, J. Liu, P. Yao, J. Li, H. Lu, Mscap: Multi-style image captioning with unpaired stylized text, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4204–4213.

[57] A. F. Biten, L. Gomez, M. Rusinol, D. Karatzas, Good news, everyone! context driven entity-aware captioning for news images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12466–12475.

[58] A. Agrawal, D. Batra, D. Parikh, A. Kembhavi, Don't just assume; look and answer: Overcoming priors for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4971–4980.

[59] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: ICLR (workshop track), 2014.

[60] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th ICML Volume 70, JMLR. org, 2017, pp. 3319–3328.

[61] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognition 65 (2017) 211–222.

[62] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: ICML, 2017, pp. 3145–3153.

[63] P. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, S. Dähne, Learning how to explain neural networks: Patternnet and patternattribution, in: ICLR, 2018.

[64] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: NIPS, 2017, pp. 4765–4774.

[65] W. J. Murdoch, P. J. Liu, B. Yu, Beyond word importance: Contextual decomposition to extract interactions from LSTMs, in: ICLR, 2018.

[66] M. T. Ribeiro, S. Singh, C. Guestrin, 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, ACM, 2016, pp. 1135–1144.

[67] L. M. Zintgraf, T. S. Cohen, T. Adel, M. Welling, Visualizing deep neural network decisions: Prediction difference analysis, in: ICLR, 2017.

[68] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.

[69] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, in: BMVC, 2018.

[70] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, C.-C. Tu, Generating contrastive explanations with monotonic attribute functions, arXiv preprint arXiv:1905.12698 (2019).

[71] R. Fergus, M. D. Zeiler, G. W. Taylor, D. Krishnan, Deconvolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2528–2535.

[72] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, G. Montavon, XAI for Graphs: Explaining graph neural network predictions by identifying relevant walks, arXiv preprint arXiv:2006.03589 (2020).

[73] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, Gnnexplainer: Generating explanations for graph neural networks, in: Advances in Neural Information Processing Systems, 2019, pp. 9244–9255.

[74] X. Li, et al., Explain graph neural networks to understand weighted graph features in node classification, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2020, pp. 57–76.

[75] Y. Zhang, D. Defazio, A. Ramesh, Relex: A model-agnostic relational model explainer, arXiv preprint arXiv:2006.00305 (2020).

[76] Q. Huang, M. Yamada, Y. Tian, D. Singh, D. Yin, Y. Chang, Graphlime: Local interpretable model explanations for graph neural networks, arXiv preprint arXiv:2001.06216 (2020).

[77] J. Kauffmann, M. Esders, G. Montavon, W. Samek, K.-R. Müller, From clustering to cluster explanations via neural networks, arXiv preprint arXiv:1906.07633 (2019).

[78] V. Ramanishka, A. Das, J. Zhang, K. Saenko, Top-down visual saliency guided by captions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7206–7215.

[79] S. Jain, B. C. Wallace, Attention is not explanation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 3543–3556.

[80] S. Wiegreffe, Y. Pinter, Attention is not not explanation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 11–20.

[81] S. Serrano, N. A. Smith, Is attention interpretable?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2931–2951.

[82] N. Halliwell, F. Lecue, Trustworthy convolutional neural networks: A gradient penalized-based approach, arXiv preprint arXiv:2009.14260 (2020).

[83] J. Sun, S. Lapuschkin, W. Samek, Y. Zhao, N.-M. Cheung, A. Binder, Explanation-guided training for cross-domain few-shot classification, in: Proceedings of the 25th International Conference on Pattern Recognition, 2020, pp. 7609–7616.

[84] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008–7024.

[85] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: an overview, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 193–209.

[86] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, S. Lapuschkin, Towards best practice in explaining neural network decisions with

LRP, in: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–7.

[87] S. Houidi, D. Fourer, F. Auger, On the use of concentrated time–frequency representations as input to a deep convolutional neural network: Application to non intrusive load monitoring, Entropy 22 (9) (2020) 911.

[88] M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Müller, A. Binder, Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, Scientific Reports 10 (2020) 6423.

[89] L. Arras, A. Osman, K.-R. Müller, W. Samek, Evaluating recurrent neural network explanations, in: Proceedings of the ACL 2019 BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 2019, pp. 113–126.

[90] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: ICLR, 2015.

[91] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, `https://github.com/facebookresearch/detectron2` (2019).

[92] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: ICLR, 2019.

[93] C. Liu, J. Mao, F. Sha, A. Yuille, Attention correctness in neural image captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017.

[94] F. Eitel, K. Ritter, Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification, in: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham, 2019, pp. 3–11.

[95] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems 31, 2018, pp. 9505–9515.

[96] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nature Communications 10 (1) (2019) 1096.

[97] C. J. Anders, T. Marinč, D. Neumann, W. Samek, K.-R. Müller, S. Lapuschkin, Analyzing imagenet with spectral relevance analysis: Towards imagenet un-hans' ed, arXiv preprint arXiv:1912.11425 (2019).

[98] L. Weber, Towards a more refined training process for neural networks: Applying layer-wise relevance propagation to understand and improve classification performance on imbalanced datasets, Master's thesis, Technische Universität Berlin (2020).
URL https://zenodo.org/record/4700661#.YJU7FCaxXmF

[99] R. Tang, M. Du, Y. Li, Z. Liu, N. Zou, X. Hu, Mitigating gender bias in captioning systems, arXiv preprint arXiv:2006.08315 (2020).