Integrating Rule-based Entity Masking into Image Captioning

Anonymous authors

Abstract

Given an image, generating its natural language description i.e., caption is a well studied problem in the recent years. Approaches proposed to address this problem usually relying on image features whose behavior is hardly interpretable from outside. Particularly, these image features are subdivided into global and local features, where global features are extracted from the global representation of the image, while local features are extracted from the objects detected locally in an image. Although, local features extract rich visual information from the image, existing models generate captions in a blackbox manner and humans can hardly interpret which local objects the caption is aimed to represent. Hence in this paper, we propose a novel framework for the image caption generation with an explicit object (e.g., knowledge graph entity) selection process while still maintaining its end-to-end training ability. The model first explicitly selects which local entities to include in the caption according to a human-interpretable mask, then generate proper captions by attending to selected entities. Experiments conducted on MSCOCO dataset demonstrate that our method achieves good performance in terms of the caption quality and diversity with a more interpretable generating process than previous counterparts.

1. Introduction

Over the past few years, the task of generating descriptions for images (i.e., image captioning) [Kiros et al., 2014, Vinyals et al., 2015, Wang et al., 2016, Anderson et al., 2017, Gu et al., 2018, Zhou et al., 2019] has become popular as it effectively brings together vision and natural language to serve various real-world applications. Most of the existing approaches are efficient in learning a correspondence between image and sequence of words with different techniques that either improve how visual information is captured with attention [Xu et al., 2015, Lu et al., 2016, Anderson et al., 2017] or language model interactions [Lu et al., 2018].

Careful analysis of methods that aim to effectively capture visual information reveal that either utilize global image features or attend to regions for local image features to generate captions. However, this makes it hard to interpret, as they do not select or control objects in an image which may be prominent for caption generation. It is especially important for easy understanding of the caption generation process in case of failures in those systems that cater real-world applications such as autonomous driving, medical imaging and surveillance. Also, observed previously [Yin and Ordonez, 2017, Wang et al., 2018] that rich entities and their interactions in some kind of a layout can help to better understand image captioning.
Therefore, in this paper, we introduce our interpretable image caption generation model (henceforth, Interpret-IC) to address the limitations of previous approaches as shown in the Figure 1. Our proposed approach work with a human-interpretable mask which selects the set of local objects observed in an image based on human proposed rules. These rules ensure that only those desirable objects are selected which human wants to observe in the caption. For this to work, the local objects need to be represented with semantically enriched labels so that humans can comprehend. As none of the current approaches provide such local object information. We leveraged relational knowledge provided by the knowledge graph [Ehrlinger and Wöß, 2016] entities to attain semantic labels by building a multi-label image classifier and replace local object visual features with entity distributed representations [Bordes et al., 2011]. We show that these entity labels and its features are superior w.r.t detected local object features in terms of interpreting knowledge from the image. Very close to our approach by Cornia et al. [2019], who considers the decomposition of a sentence into noun chunks and models the relationship between image regions and textual chunks. However, we dynamically select the number of objects prior learning the model. Our main contributions are as follows:

- We proposed a novel end-to-end caption model for interpretable image captioning.
- We used knowledge graph entities as image labels for grounding visual and factual knowledge.
- We show that interpretable image captioning can attain diversity in the captions generated with simple visual object masking.

2. Related Work

In the related work, we explore deep neural network based approaches which generate sentence-level natural language description for images.

Diverse Image Captioning   In the recent years, monolingual image caption generation is explored to incorporate diversity in the generated captions. Approaches [Dai et al., 2017,
Li et al., 2018 has leveraged adversarial training using either generative adversarial networks [Shetty et al., 2017] or variational auto-encoder [Wang et al., 2017, Shen et al., 2019]. While, Vijayakumar et al. [2016] used diverse beam search to decode diverse image captions in English. Approaches were also proposed to describe images from cross-domain [Chen et al., 2017]. However, our goal in this research is to provide better selection procedure for identifying preferable objects in images. Nevertheless, we show that interpretability can also assist diversity.

Controllable Image Captioning Approach that is closer to interpretable image captioning is a procedure to control local objects in images. Cornia et al. [2019] used either a sequence or a set of local objects by explicitly grounding them with noun chunks observed in the captions to generate diverse captions. Further, instead of making captions only diverse, Deshpande et al. [2019] made the captioning more accurate. Our work falls into this space, however understanding the important entities that represent the image and controlling them is what we aim to achieve.

3. Interpretable Image Captioning

3.1 Base-IC Model

The base image caption model (Base-IC) is built without masking. Given an image \( I \), its global representation \( I_v \in \mathbb{R}^V \) denote the encoding of the full image, while the spatial objects \( a_v = \{a_{v_1}, \ldots, a_{v_L}\} \) encode local regions of the image provided as \( a_{v_j} \in \mathbb{R}^D \). Similar to previous works [Lu et al., 2016, Anderson et al., 2017], our proposed image description model also leverages soft attention mechanism to weight spatial objects during description generation using the partial output sequence as context. Figure 2 illustrates the architecture.

Initially, L-1 of the model receives input from the global visual context provided by \( I_v \) and textual sequence, where each word \( (w_t \in \mathbb{R}^T) \) at time step \( t \) in the textual sequence is initialized with the pretrained word embeddings to produce hidden vectors \( h^1_t \in \mathbb{R}^{H_1} \). Furthermore, \( h^1_t \) is used in combination with \( a_v \) to compute soft attention. Later, \( h^1_t \) and attended spatial features are added and provided as input to L-2 for attaining \( h^2_t \in \mathbb{R}^{H_2} \). For convenience and to reduce many parameter names, we use \( \Theta \) as the reference for the parameters of the LSTM.

To calculate attended spatial features \( (\hat{a}_t) \) we leverage \( a_v \). Hidden sequences \( h^1_t \) at each time step \( t \) is used to generate a normalized attention weight \( \alpha_t \) for each of the spatial object features \( (a_{v_j}) \) given by Equation 1 and Equation 2.

\[
\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{L} exp(e_{tk})} \tag{1}
\]

\[
e_{tj} = \tanh(W_{ae}a_{v_j} + W_{he}h^1_t) \tag{2}
\]

where \( L \) represent the cardinality of set \( a_v \), \( W_{ae} \in \mathbb{R}^{M \times D} \) and \( W_{he} \in \mathbb{R}^{M \times H_1} \) are learnable parameters. Further, \( \hat{a}_t \) is calculated with Equation 3 and is used as input along with \( h^1_t \) to the L-2 at every time step \( t \).

\[
\hat{a}_t = \sum_{j=1}^{L} \alpha_{tj}a_{v_j} \tag{3}
\]
Figure 2: Illustration of Base-IC Model

The final Base-IC using $w_t$ and $I_v$ as input to L-1 is given by Equation 4 and $h_t^1$ is given by Equation 5. Further, $\hat{a}_t$ and $h_t^1$ are added using Equation 6 to provide as input for L-2 for generating $h_t^2$ as given by Equation 7. It is then used to predict next words in the sequence as given in the Equation 8.

\[
x_t = I_v \oplus w_t
\]  
\[
h_t^1 = L-1(x_t, h_{t-1}^1; \Theta)
\]  
\[
x_t' = \hat{a}_t + h_t^1
\]  
\[
h_t^2 = L-2(x_t', h_{t-1}^2; \Theta)
\]  
\[
p_{t+1} = \text{softmax}(W_{vocab}h_t^2)
\]

where $W_{vocab} \in \mathbb{R}^{vocab \times (V + H_2)}$, $\oplus$ represents concatenation and $vocab$ refers to vocabulary of the caption dataset.

### 3.2 Interpret-IC Model

Main aim of the Interpret-IC model is to select objects present in the spatial objects set $a_v$ with human-interpretable masking. This is in contrast with earlier approaches [Xu et al., 2015, Anderson et al., 2017], who decoded the caption by attending to spatial objects only by ranking them according to their importance at each time step. Also, these approaches provide no control for humans to select their desirable objects. It clearly sets expectation from Interpret-IC model that the selected objects should provide more prominence in caption generation by discarding those objects that are not selected.

Hence, we introduce masked attention to select those objects that human wants to see in the generated captions. To achieve it, we leverage ground truth mask i.e., $\text{mask}_{gt}$ where each object in the $a_v$ is masked with a binary parameter $\beta_1, \beta_2, \ldots, \beta_n$. We set $\beta_i = 1$ if selected and 0 otherwise. Also, $\beta_i$ is assumed to be independent from each other and is sampled from a bernoulli distribution. Prediction mask i.e., $\text{mask}_{pred}$ is estimated during training with a multi-layer perceptron (MLP).
Further, attention weights computed in the Equation 1 is modified with the estimated mask\textsubscript{pred} as shown in Equation 9.

\[
\alpha_{tj}^{\text{mask}} = \frac{\exp(e_{tj})\text{mask}_{\text{pred}}}{\sum_{k=1}^{L} \exp(e_{tk})\text{mask}_{\text{pred}}}
\]  

(9)

It is then used to calculate \(\hat{a}_t^{\text{mask}}\) given by Equation 10, which is further used as input along with \(h_t^1\) to the L-2 at every time step \(t\). Figure 3 illustrates the overall architecture.

\[
\hat{a}_t^{\text{mask}} = \sum_{j=1}^{L} \alpha_{tj}^{\text{mask}} a_{vj}
\]  

(10)

Note that our selection strategy is very different from Cornia et al. [2019], who control spatial objects using the fixed noun-chunks extracted from captions which are not available during testing phase. While, we use human designed rules to change our mask, so that we control the mask as we aim to use it.

### 3.3 Ground Truth Mask Selection

In the Interpret-IC model, mask\textsubscript{pred} needs to be optimized during training phase closer to the ground truth binary mask i.e., mask\textsubscript{gt} such that it can be utilized during the testing phase. However, first we need to create such mask\textsubscript{gt} based on human-interpretable rules to influence the caption generation process.

There can be several ways to create mask\textsubscript{gt} by changing the rules. In this paper, we apply visual entities to caption noun matching approach to build the mask\textsubscript{gt}. Our rule here states that for each noun identified\(^1\) in the caption, we need to find the closest visual entity by computing cosine distance between the noun and visual entity vectors attained using pretrained fastText\(^2\) vectors. For all nouns identified, closest visual entities are set to 1, while rest are set to 0. This rule ensures that the nouns observed in the caption representing some kind of objects present in images have to be given higher preference.

\(^1\) https://spacy.io/
\(^2\) https://fasttext.cc/
during caption generation. While, rest of the visual entities (e.g., actions) are put on back burner. Algorithm 1 presents the overview of selection process.

| Nouns (N), Visual Entities (VE), fastText Embeddings (FTE) mask$_{gt}$ for each caption Initialize $N_{emb} = \text{FTE}(N)$; Initialize $VE_{emb} = \text{FTE}(VE)$; Initialize Image$_{velist}$ as $I_{velist}$; Initialize Caption$_{list}$ as $C_{list}$; Function mask$_{gt}$ Selection for $C, VE$ in $C_{list}, I_{velist}$ do Extract N from caption; Initialize mask$_{gt} = \text{zeros}[\text{len}(VE)]$; for $n$ in $N$ do if $n$ not EMPTY then dist = CosineDistance($n_{emb}, VE_{emb}$) close$_{index}$ = (dist) mask$_{gt}[$close$_{index}$] = 1 end end return mask$_{gt}$; end |

**Algorithm 1:** mask$_{gt}$ selection process

4. Training and Inference

**Base-IC** The parameters ($\theta$) of the Base-IC model are trained for optimizing the cost function ($C$) to minimize the sentence-level categorical cross-entropy loss by finding negative log likelihood of the appropriate ground truth word ($y_t^*$) at each time step $t$ as shown in Equation 11. Here, we leverage teacher forcing [Sutskever et al., 2014], where ground truth ($y_t^*$) is fed to next step in the layer $L-1$, instead of the predicted word in previous step.

$$C(\theta) = - \sum_{t=0}^{T^{(n)}} \log p_{\theta}(y_t^*)$$  \hspace{1cm} (11)

The $T^{(n)}$ represents the length of the sentence at $n$-th training sample. During inference, we leverage beam search with beam size is set to 5 in our experiments.

**Interpret-IC** Similar to Base-IC model, parameters ($\theta'$) of the Interpret-IC are trained for optimizing the cost function ($C'$) which minimizes both the sentence-level categorical cross-entropy loss along with binary cross-entropy loss that approximate (mask$_{pred}$) closer to the ground truth mask (mask$_{gt}$) as shown in Equation 12.

$$C'(\theta') = - \left( \sum_{t=0}^{T^{(n)}} \log p_{\theta}(y_t^*) \right) + \text{mask}_{gt}\log(\text{mask}_{pred}) + (1 - \text{mask}_{gt})\log(1 - \text{mask}_{pred})$$  \hspace{1cm} (12)
During inference, similar to Base-IC model, we leverage beam search by setting beam size to 5 in our experiments.

5. Evaluation Setup

Datasets  For experimental evaluation, we use MSCOCO dataset with splits of Karpathy and Fei-Fei [2015].

Local and Global Image Features  Spatial object \(a_v\) features are extracted in two different ways.

- Faster R-CNN [Ren et al., 2015] in conjunction with the ResNet-101 [He et al., 2016] trained on visual genome data by Anderson et al. [2017] is used to extract top 36 local object features \(a_vj\) of dimension 2048. There are pure visual features and we refer to this set as Obj→RCNN.

- Since, Obj→RCNN represent pure visual features without label information. Following Mogadala et al. [2018], we extracted semantically enriched labels denoting entities from captions aligned to an image in training set of MSCOCO with a knowledge graph annotation tool such as DBpedia spotlight\(^3\). In total, 812 unique human-interpretable already disambiguated labels are extracted. Further, a multi-label image classifier is trained with sigmoid cross-entropy loss by fine-tuning VGG-16 [Simonyan and Zisserman, 2014] pre-trained on the training part of the ILSVRC12 with training images in MSCOCO. After training, we use the classifier to acquire Top-15 entity labels for each image present in the training, validation and testing set of MSCOCO. Now, to use entity labels similar to Obj→RCNN features. We use knowledge graph embeddings [Ristoski and Paulheim, 2016] and generate 500 dimensional vectors\(^4\) for each entity-label. We refer to this set as Obj→VisualEntity.

- The global visual features \(I_v\) of dimension 2048 is extracted using the average pooling of Obj→RCNN features.

Caption Model  Both Base-IC and Interpret-IC models are built by initializing the model with input \(w_t\) word embeddings pretrained using Glove [Pennington et al., 2014] on the MSCOCO training captions corpora. The dimensions of the hidden units \(h_{1t}\), \(h_{2t}\) in L-1 and L-2 of models are set to 512. Also, the hidden units of shared layer \(h_{s}^{(s)}\) are set to 512. All models are then trained with Adam optimizer [Kingma and Ba, 2014] with gradient clipping having maximum norm of 1.0 and mini-batch size of 50 for 25 epochs. Initially, the learning is set to 0.001 and is reduced by a factor of 10 if there is no improvement in the validation loss for 3 continuous epochs.

Evaluation Measures  We first evaluate the generated captions based on correctness which guarantee the generation quality based on standard captioning metrics. Further, we check if our proposed model with human-interpretable masking can generate diverse and

---

4. Please note that these embeddings are different from fastText Vectors used to build mask\(_{gt}\). These embeddings are analogous to pure visual features, however, learned from knowledge graph structure.
interesting captions. For this, we leverage earlier proposed [Shetty et al., 2017, Deshpande et al., 2019] metrics such as vocabulary size and novel caption with best (i.e., Top-1) generated caption. Vocabulary Size (VS) find unique words in generated captions and Novel captions (NC) identify the percentage of generated captions that are not seen in the training set.

6. Results

6.1 Quantitative Results

We compared our proposed Base-IC and Interpret-IC along with other recent baselines. Table 1 shows the results obtained. It can be observed that the Interpret-IC model was able improve over recent approaches by allowing better control over the caption generation process.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adv-bs [Shetty et al., 2017]</td>
<td>-</td>
<td>23.9</td>
<td>-</td>
<td>-</td>
<td>16.7</td>
</tr>
<tr>
<td>CNN+CNN [Wang and Chan, 2018]</td>
<td>26.7</td>
<td>23.4</td>
<td>51.0</td>
<td>84.4</td>
<td>-</td>
</tr>
<tr>
<td>Convolutional-IC [Aneja et al., 2018]</td>
<td>31.6</td>
<td>25.0</td>
<td>53.1</td>
<td>95.2</td>
<td>17.9</td>
</tr>
<tr>
<td>POS+Joint [Deshpande et al., 2019]</td>
<td>-</td>
<td>24.7</td>
<td>-</td>
<td>-</td>
<td>18.0</td>
</tr>
<tr>
<td>Base-IC +Obj→RCNN</td>
<td>31.8</td>
<td>24.9</td>
<td>52.9</td>
<td>96.7</td>
<td>18.2</td>
</tr>
<tr>
<td>+Obj→VisualEntity</td>
<td>32.1</td>
<td>24.8</td>
<td>53.6</td>
<td>96.9</td>
<td>18.0</td>
</tr>
<tr>
<td>Interpret-IC +Obj→VisualEntity</td>
<td>32.4</td>
<td>24.9</td>
<td>53.7</td>
<td>97.8</td>
<td>18.1</td>
</tr>
</tbody>
</table>

Table 1: Results achieved with our models in comparison with baseline approaches.

6.2 Qualitative Results

To understand the contribution made by human-interpretable mask to caption generation. We explored qualitatively the captions generated by both Base-IC and Interpret-IC models with visual entities from two different perspectives. First, we observed the quality of the predicted mask in selecting required visual entities for better coverage. Second, we checked if Interpret-IC model could overcome or correct mistakes made by the Base-IC model. In the following, we discuss each of these cases briefly by showing some examples.

Caption Coverage We use visual entities such that they represent local objects in images to be incorporated them in the caption. However, this cannot be simply achieved with a Base-IC model. As seen in Figure 4, the Interpret-IC model which weighs each of these objects differently based on the predicted mask, when compared with the Base-IC model giving equal importance to each of them. Although the Base-IC model generated partially relevant caption, masking has shown to improve coverage of local objects in the image. The
attained. We observe that, our Interpret-IC generated caption using diversity measures described earlier. Table 2 shows the results We compared our models with other diverse caption generation baselines that compare best Base-IC and Interpret-IC.

6.3 Diversity

Although our aim is not to achieve diverse captions, to comprehend whether our proposed Base-IC and Interpret-IC models generate best (i.e., Top-1) diverse and interesting caption. We compared our models with other diverse caption generation baselines that compare best generated caption using diversity measures described earlier. Table 2 shows the results attained. We observe that, our Interpret-IC model cannot exceed scores of the baseline

Figure 4: Caption Coverage Example (Entities with mask_{pred} > 0.5 are highlighted in blue): (a) Missing local object (Dog) in the caption generated by Base-IC, while “White Dog” is included by Interpret-IC providing better coverage. (b) Missing details about the birthday cake, Interpret-IC generated better and interesting caption by highlighting objects that need to be focused on.

Figure 5: Caption Correction Example (Entities with mask_{pred} > 0.5 are highlighted in blue): (a) Base-IC generate a caption by including wrong objects i.e., sheep, while “cattle” is included by Interpret-IC because a lower weight (0.2) is assigned to “sheep” hence filters out the wrongly detection. (b) Although Base-IC covers the correct object (Birds), it is too general and fails to provide more informative caption. Interpret-IC replaces it with the exact object by giving a large weight to emphasize the detected entity “duck”. selector is able to assign higher scores to prominent objects in the image which increases the probability of covering them in the generated caption.

Caption Correction We also observe that, apart from providing better coverage of visual entities in the generated captions. Masking also plays a prominent role in the caption correction. That is, as seen in the Figure 5, although the Base-IC model generated a partially relevant caption, Interpret-IC generated the most accurate caption with effective selection of relevant visual entities. The selector is expected to assign lower scores to inappropriate (bird in Figure 5b) or wrongly detected objects (sheep in Figure 5a) thus encouraging the decoder to attend to more plausible entities.
trained to generate diverse captions in an adversarial setting (i.e., Adv-bs). However, with less effort and simple masking we could see a significant jump on the standard caption model (i.e., Base-bs).

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Base-bs</th>
<th>Adv-bs</th>
<th>Base-IC</th>
<th>Interpret-IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VS</td>
<td>756</td>
<td><strong>1508</strong></td>
<td>443</td>
<td>862</td>
</tr>
<tr>
<td>NC</td>
<td>34.18</td>
<td><strong>68.62</strong></td>
<td>36.23</td>
<td>51.54</td>
</tr>
</tbody>
</table>

Table 2: Diversity: Comparison of vocab size (VS) and novel captions (NC) using Top-1 generated caption with Base-bs [Shetty et al., 2017] and Adv-bs [Shetty et al., 2017]. Base-IC and Interpret-IC use +Obj→VisualEntity features.

Also, in Figure 6, we plot unique unigrams and bigrams predicted at every word position. The plot shows that the Interpret-IC have higher unique unigrams at different word positions and is consistently higher for the bigrams when compared against Base-IC with visual entities as features. This supports our hypothesis that Interpret-IC can produce more diverse captions as it can alter caption generation process.

7. Conclusion and Future Work

In this paper, we aimed to address the problem of interpretable image captioning by leveraging knowledge graph entity features. Initially, we obtained local objects as visual entities in the image by grounding knowledge graph entities. Further, the human-interpretable masking rules are developed to select those visual entities for generating desirable captions. Experimental results show that interpretability in caption generation can help to alter caption generation process hence allowing control and selection. In Future, we aim to improve caption generation process by trying different masks and better sampling.
References


