Improve Image Captioning by Estimating the Gazing Patterns from the Caption

Rehab Alahmadi\textsuperscript{1,2} and James Hahn\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, The George Washington University
\textsuperscript{2}Department of Information Technology, King Saud University
\{raa53, hahn\}@gwu.edu

Abstract

Recently, there has been much interest in developing image captioning models. State-of-the-art models reached a good performance in producing human-like descriptions from image features that are extracted from neural network models such as CNN and R-CNN. However, none of the previous methods have encapsulated explicit features that reflect a human perception of the images such as gazing patterns without the use of the eye-tracking systems. In this paper, we hypothesize that the nouns (i.e. entities) and their orders in the image description reflect human gazing patterns and perception. To this end, we estimate the sequence of the gazed objects from the words in the captions and then train a pointer network to learn to produce such sequence automatically given a set of objects in new images. We incorporate the suggested sequence by pointer network in existing image caption models and investigate its performance. Our experiments show a significant increase in the performance of the image captioning models when the sequence of the gazed objects are utilized as additional features (up to 13 points improvement in CIDEr score when combined with Neural Image Caption model).

1. Introduction

Image captioning is the process of automatically generating a human-like natural language description of an image [44]. This is beneficial for applications like providing guidance to medical practitioners [35] and helping visually impaired people to understand visual contents. However, image captioning is not a trivial task and is highly inspired by human cognition – mainly image perception (understanding image contents including objects and their relationships) and sentence planning and generation (describing the image with a natural language).

To generate an image caption, current state-of-the-art models heavily depend on CNN and R-CNN to extract the visual features as an input to their model [43, 44, 28, 46]. These studies rely solely on such features without explicitly modeling the relations between image nuances and captions. More recent studies suggested different mechanisms to address this issue with attention mechanism which can implicitly learns the relation between entities and regions in an image [44, 51, 28, 19]. Other studies [47, 15] integrated the graph convolutional network to the image encoder to learn the relationship between objects in the image. Although these models could successfully generate human-like captions, the images are not perceived for the caption purpose. Specifically, CNN and R-CNN are built and trained for recognition and detection purposes respectively, but not for image captioning. Research shows that human perception, specifically gazing patterns, during detection and description tasks are different [13, 49, 17, 39]. Therefore, the image encoder could be boosted by explicit
visual features learnt from the captions that reflect the gazing behavior when the image initially described.

Several works [17, 38, 37] use the gazing information to improve the attention in the image captioning models. These studies showed the effectiveness of integrating gazing information to the image captions. However, gazing information is extracted from eye tracking systems which is expensive and not available for all researchers; in such studies, they did not propose a method to generate the gazing information for unseen images without the use of the eye tracking system. In this paper, we assume that entities mentioned in the captions can reflect human perception of an image when human initially generated the captions motivated by psycholinguistics studies which suggest that there is a relationship between word production and eye movements [13, 7, 6, 18]. These studies aim to understand human perception (e.g. which regions speakers look at and in what order [18]). For example, many researchers study the sentence production of speakers describing images while tracking their eye movements [13, 7, 6].

Griff and Bock [13] and Coco and Keller [7] found that similar eye scan patterns, when looking at the image as an attempt to describe it, lead to similar sentences. Specifically, the order of gazing at objects is correlated with the order of mentioning these objects in a sentence (description). Similarly, Spain and Peron [5] found out that human choose to describe certain objects (e.g. dog) and ignore the other (e.g. sidewalk).

Guided by aforementioned studies, we hypothesize that the entities that mentioned in the captions are the important objects that human choose to describe more than the other, and their order reflects the human perception of an image (gazing patterns) when human initially generated the captions. In this paper, we develop a model based on pointer network [42] that learns from the entities in the captions to generate gazing sequences automatically and then integrate the learnt gazing sequence to image caption models. Pointer network has been widely used to order sentences [12, 10, 27, 50] or stories [1, 36]. We integrate this gazing pattern as a sub-network to current image caption models. We show up to 13 points improvements in CIDEr score with explicit modeling for gazing patterns in image captioning models.

In this paper, our contributions are as following:

1. We estimate gazing patterns from entities mentioned in captions directly rather than using expensive eye tracking system.

2. We propose gazing pattern prediction model based on pointer network that can automatically generate gazing pattern for unseen images.

3. We propose a model-agnostic gazing pattern sub-network that can be integrated to image caption models as an additional visual features.

2. Related Work

Image Captions. Most image caption methods utilize CNN to encode images and the recurrent neural network as a language model [43, 21, 44, 51, 28, 3]. To further boost the image captioning models, attention mechanism [4] has been introduced to image captioning models to allow more interaction between image encoder and the language model [44, 51, 28]. More specifically, the importance score for each region is computed while generating a specific word then normalized with softmax function; these scores are then applied to regions to reflect its importance in
generating that word. Regions of the image are either represented as vector extracted from fixed size grid CNN features [44, 28], semantic attributes [51], or as bottom-up features extracted from R-CNN detected regions [3]. To produce rich description of images, [14] introduced coarse-to-fine multiple stage model with multiple LSTMs, such that early stage produce coarse description while the later stages add details to descriptions. Additionally, transformer [40] has been utilized for image captioning either as image encoder [19] or as a language modeling [9]. In this work, we focus on boosting LSTM based image captioning models by exploiting the visual features that can be learned from the captions.

Several works are introduced to boost visual features by integrating the graph-convolutional networks to the image encoder in image captioning models. Yao et al. [47] integrated the semantic and spatial relationships between objects in an image, while Guo et al. [15] explicitly model the relationship as an additional node in the convolutional graph.

Similar to our work, Cornia et al. [8] and Alahmadi et al. [2] integrated entities found in the captions with their corresponding sequence or set of regions to their image caption model. We instead learn the gazing sequence that is more likely produced by humans describing the image automatically. Our proposed gazing prediction model can be plugged into different image caption models to boost its performance.

**Image Captions with Gazing Information.** Few studies integrated gazing information in combination with attention mechanism to improve captioning models [17, 38, 37]. The gazing data are acquired from eye-tracking systems that are expensive to obtain. We use the learned gazing data as an input to the model rather than replacing or enhancing the attention mechanism.

3. Model

An overview of the model architecture is shown in figure 2. Our model consists of two parts: 1) gazing pattern prediction network (section 3.1) which learns to produce gazing sequence of an image; 2) integration of the gazing pattern to image caption models (section 3.2).

3.1. Gazing Pattern Prediction Model

The image descriptions consist of sequences of words that can be visually grounded into image regions. Given these descriptions and their grounding regions in an image, we can construct a sequence of image regions that reflects their order in the caption. We call it **gazing sequence** $R$, such that

$$R = [r_1, r_2, r_3, ..., r_n]$$  \hspace{1cm} (1)

where $r_i$ is a region at position $i$ that is associated with an entity in the caption, and $n$ is the number of gazed regions.

More formally, Given $m$ regions $r = [r_{o_1}, r_{o_2}, ..., r_{o_m}]$ with an arbitrary order $o = [o_1, o_2, ..., o_m]$, where $r$ is set of regions detected by R-CNN, $o$ is their arbitrary order, and $m$ is the number of regions detected by R-CNN.

The goal is to find the gazing sequence $R$ by finding the order $\hat{o} = [\hat{o}_1, \hat{o}_2, ..., \hat{o}_n]$ that is closest to the gold order $o^* = [o^*_1, o^*_2, ..., o^*_n]$ by maximizing $P(o^*|r)$ such that

$$P(o^*|r) > P(o|r) \quad \forall o \in \psi$$  \hspace{1cm} (2)

where $\psi$ is the set of all permutation of $o$, and $n$ is the number of regions mentioned in the captions, such that $n \leq m$. 

Figure 3. Pointer Network modeling gazing patterns prediction. The model takes set of objects detected by R-CNN as an input and learns to point to the objects that are part of the gazing sequence.
3.1.1 Model Features

Following the same approach in [3, 29, 8], a region $r_i$ is represented by concatenating three features: visual, textual, and spatial features, such that:

$$r_i = [v_i; t_i; l_i]$$ (3)

where $v_i$, $t_i$, and $l_i$ are the visual, textual, and spatial features, respectively.

Following [8], the visual feature is extracted from Faster R-CNN and then processed by two fully connected layers, and the textual feature is the Glove embedding of the class label of the region and processed by one fully connected layer. The spatial features are the normalized position and size of the bounding box of regions.

The concatenated features are then encoded through a fully connected layer, such that:

$$x_t = W(r_t)$$ (4)

where $W$ is a learnable parameter and $t$ is the position of the region in the gazing sequence.

3.1.2 Pointer Network

We model the gazing pattern prediction using pointer network [42] as shown in figure 3. Pointer network consists of Bi-LSTM as an encoder which learns features from a sequence of regions in an arbitrary order and an LSTM as a decoder which learns to point to the regions in order generating the gazing sequence.\footnote{The decoder keeps pointing to the regions until the gazed-sequence reaches a maximum length or the model points to the $<END>$ token that is inserted as the first element in the encoder.}

We initialize the first time step in the decoder with features learned from the encoder. Remaining time steps use information learned from the previous step. The decoder also utilizes information from attention mechanism which computes the probability distribution over all the encoder’s input; the region with the highest probability is chosen to be the region in the position $i$.

$$h_i^d, c_i = LSTM(h_{i-1}^d, c_{i-1}, r_i)$$ (5)

$$u_j^i = v^T \tanh(W_1 h_j^v + W_2 h_i^d)$$ (6)

$$P(r_i | r_{i-1}, ..., r_0) = \text{Softmax}(u_j^i)$$ (7)

where $v \in R^d$, $W_1$ and $W_2 \in R^{d \times d}$ are learnable parameters and $j \leq m$ and $P(r_i | r_{i-1}, ..., r_0)$ is the probability of the chosen region at time $i$, and $u_j^i$ is the output distribution over the input which acts as a pointer to the input elements.

The input to the decoder at the time step $i$ is the previously predicted region. In the first time step, the input is the feature of the image, so the model can have contextual information before pointing to the first region.

3.2. Image Caption with Gazing Sequence

To incorporate the gazing pattern into image caption models, we propose a model-agnostic gazing pattern sub-network (section 3.2.1) that can be plugged as an additional feature to image caption models (section 3.2.2).

3.2.1 Model-agnostic Gazing Pattern Sub-network

To model the gazing mechanism, we utilize LSTM (VisualLSTM, henceforth) that takes as an input the sequence of the gazed regions generated by pointer network $[x_0, x_1, x_2, ..., x_n]$, such that $x_t$ is a region feature obtained from equation 4. The VisualLSTM then encodes the gazed sequence into a fixed size hidden vector $h_n^v$, such that:

$$h_n^v = \text{VisualLSTM}(x_n, h_{n-1}^v)$$ (8)

where $n$ is the number of regions.

Equation 8 is optimized simultaneously with an image captioning model such that the LSTM in the language model is initialized with the learned gazing pattern in equation 8 as the following:

$$h_1^c = \text{LanguageLSTM}(w_1, h_n^v)$$ (9)

where LanguageLSTM is caption generator LSTM in image caption models, $h_1^c$ is the hidden state of the first time step of the LanguageLSTM, and $w_1$ is the word embedding of the first word of the caption.

3.2.2 Integrating Image Captioning Models with Gazing Patterns

We integrate the gazing pattern sub-network to three baseline image captions models: Neural Image Captioning (NIC)\footnote{Adaptive Attention model (Ada-att)[28] that apply attention on fixed grid CNN features and Attention on attention model (AoA) [19] that apply attention on a bottom-up features.} that does not apply any attention, Adaptive attention model (Ada-att)[28] that apply attention on fixed grid CNN features and Attention on attention model (AoA) [19] that apply attention on a bottom-up features.
<table>
<thead>
<tr>
<th>Model</th>
<th>Bleu-1</th>
<th>Bleu-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIC [43]</td>
<td>67.8</td>
<td>26.8</td>
<td>22.6</td>
<td>49.7</td>
<td>80.2</td>
</tr>
<tr>
<td>NIC+Att</td>
<td>70.5</td>
<td>29.7</td>
<td>24.2</td>
<td>51.8</td>
<td>90.7</td>
</tr>
<tr>
<td>NIC+GP (ours)</td>
<td>70.5</td>
<td>30.3</td>
<td>24.7</td>
<td>52.5</td>
<td>93.7</td>
</tr>
<tr>
<td>NIC+GT-Set</td>
<td>75.2</td>
<td>33.4</td>
<td>26.8</td>
<td>55.0</td>
<td>108.4</td>
</tr>
<tr>
<td>NIC+GT-Seq</td>
<td>75.6</td>
<td>34.0</td>
<td>27.1</td>
<td>56.3</td>
<td>110.2</td>
</tr>
<tr>
<td>Ada-att [28]</td>
<td>71.6</td>
<td>30.7</td>
<td>25.2</td>
<td>52.9</td>
<td>97.1</td>
</tr>
<tr>
<td>Ada-att+GP (ours)</td>
<td>72.7</td>
<td>31.6</td>
<td>25.4</td>
<td>53.6</td>
<td>100.3</td>
</tr>
<tr>
<td>Ada-att+GT-Set</td>
<td>74.8</td>
<td>33.5</td>
<td>27.7</td>
<td>55.6</td>
<td>110.9</td>
</tr>
<tr>
<td>Ada-att+GT-Seq</td>
<td>75.6</td>
<td>34.2</td>
<td>27.4</td>
<td>56.4</td>
<td>111.3</td>
</tr>
<tr>
<td>AoA [19]</td>
<td>76.8</td>
<td>36.6</td>
<td>28.3</td>
<td>57.1</td>
<td>117.3</td>
</tr>
<tr>
<td>AoA+GP (ours)</td>
<td>77.8</td>
<td>37.8</td>
<td>28.9</td>
<td>58.7</td>
<td>120.6</td>
</tr>
<tr>
<td>AoA+GT-Set</td>
<td>77.7</td>
<td>36.6</td>
<td>28.9</td>
<td>57.6</td>
<td>119.4</td>
</tr>
<tr>
<td>AoA+GT-Seq</td>
<td>77.7</td>
<td>37.7</td>
<td>29.1</td>
<td>58.8</td>
<td>121.0</td>
</tr>
</tbody>
</table>

Table 1. Performance comparison of the baseline models and the integrated gazing pattern model, where NIC, Ada-Attention, AoA, GP, and GT are short of Neural Image Caption, Adaptive attention, Attention on Attention, Gazed Pattern, and Ground Truth, respectively. All values are reported as a percentage (%). The bold value represents the highest value.

All three models use LSTM as an encoder to generate the captions. In NIC, the image is encoded to a feature vector extracted from the last layer of CNN. We choose NIC as one of the baselines to show the effectiveness of integrating the gazing pattern to a simple model, and to compare its performance with attention models. On the other hand, Ada-att and AoA are both attention-based models. Ada-att dynamically decides whether to attend a region in a convolutional map or not when generating a word. AoA extends the attention operator by adding an attention gate that weight the final attention information.

As shown in Figure 4, the gazing pattern is integrated to NIC, Ada-att, and AoA by initializing LSTM that model language with the visual LSTM. Therefore, besides the image features that are extracted from CNN (NIC and Ada-att) or R-CNN (AoA), the model also receive the gazing pattern as visual features.

4. Experiments

4.1. Dataset

For the gazing pattern prediction model, we use the publicly available COCO entities release [8]. This release provides a region with a bounding box and a class label for each entity mentioned in the captions provided by the COCO image captioning dataset [26]. The regions are linked to regions detected by pre-trained Faster R-CNN model [33] with ResNet-101[16] trained on ImageNet [11] and Visual Genome dataset [23] to provide the bottom-up features [3], such that for each region we obtain 2048 dimensional vector. The input to the gaze pattern prediction model is a shuffled set of the detected regions. We find out that a shuffled set of objects with the top 15 detection scores give the best results.

We evaluate the image caption models with the popular MS COCO dataset [26]. The dataset contains 123,287 images. We follow the same widely adopted split - 113,287 training images and 5,000 images for each validation and testing [21]. We lower case sentences and eliminate words that occur less than 5 times in the overall training corpus.

4.2. Experimental setting

Gazing Sequence Modeling. We set the hidden size to 256 for both encoder and decoder. We optimize the model with ADAM optimization [22] and we set the learning rate to 1e-4. We train it with cross-entropy loss.

Image captions. We set the hidden size of the visual LSTM to 512 for NIC, 512 for Ada-att, and 1024 for AoA, such that the hidden size of the visual LSTM equals the hidden size of the language LSTM in all models. We use ADAM [22] for optimization for the three models with learning rate 2e-4 for AoA and 1e-4 and 5e-4 for encoder and decoder in NIC and Ada-att. We train all models with cross-entropy loss. For AoA, we also optimize the CIDEr-D score [34].

Experimental Conditions. We experiment the image caption models with different gazing patterns: gazing pattern generated by pointer network (*+GP), ground truth gazing sequence (*+GT-Seq), and the ground truth gazing set (*+GT-Set) extracted from the captions. Specifically,

---

2For implementation, we use the open-source codes that re-implement Ada-att and AoA architectures: https://github.com/fawazsammani and https://github.com/husthuaan/AoANet
*+GT-Seq preserve the order of the objects in the captions while *+GT-Set has the objects in arbitrary order.

5. Results

5.1. Image Captions Results

Metrics. We used the following metrics to evaluate the image captioning performance: BLEU [32], METEOR [24], ROUGE-L [25], and CIDEr [41].

Gazing Pattern Sub-network Performance. Table 1 shows the results for the investigated image captioning models: NIC, Ada-att, and AoA with and without the additional gazed features. All image caption models benefit from the gazing pattern sub-network to some extent; some models improved significantly than others. We observe significant improvement across all metrics (up to 13.5 improvements in CIDEr) when we integrate gazing sub-network in NIC (NIC+GP), which is expected given that it does not include objects information and it does not apply any attention mechanism comparing to other methods. It is important to note that applying attention to image captions will give the model the benefit of relating between words in the caption and objects in the image; therefore, integrating gazing features to NIC leads to significant boosting in performance compared to Ada-att and AoA.

Attention vs. Gazing Pattern. We add attention to the NIC model (NIC+Att) to compare its performance with NIC+GP. We note that NIC+GP slightly outperforms the NIC with attention (Table 1). However, we notice a significant difference in the performance when the ground truth gazing sequence is integrated to NIC (NIC+GT-Seq) compared to NIC+Att. We also note that the NIC+GT-Seq performance is comparable to Ada-att+GT-Seq, which shows that NIC can perform as well as Ada-Att with a perfect gazing sequence. However, NIC+GT-Seq is still below AoA that has a stronger attention mechanism comparing to Adaptive attention.

Ground Truth Gazing Sequence and Set. We evaluate the image caption models with ground truth gazing patterns during training and evaluation as an upper bound of performance. We evaluate the performance of image caption models with two gazing patterns: the ground truth gazing sequence and the ground truth gazing set. Both patterns contain the same objects. The ground truth gazing sequence (*+GT-Seq) contains the objects that are ordered based on their position in the caption. The ground truth gazing set (*+GT-Set) is a set of objects that are mentioned in the caption. The goal of this experiment is to test if the order of the objects in the gazed sequence will make a difference in the performance. Table 1 shows the results of the models with the ground truth gazing patterns. Training the model with a ground truth set significantly increased the performance of the baselines, because the model has information about the objects that are more likely to be described regardless of their order. This shows the importance of including the gazed objects in the gazing pattern. Preserving the order in the gazing pattern further boosts the performance. It is important to note that the GT-Seq is inclusive of the GT-Set; hence, it has the benefit of the GT-Set with the extra advantage of keeping the objects’ orders.

We also can note in Table 1 that AoA+GP performance is comparable to its performance with the ground truth gazed pattern. However, this is not the case for NIC and Ada-att. We can see a significant difference in the performance of these two models when the ground truth gazing patterns are integrated compared to the predicted gazing pattern. We notice that these two models are very sensitive to the objects in the gazing pattern, such that different objects produce different captions; therefore, inaccurate predicted gazing pattern will generate captions that different from the ground truth captions; hence, lower evaluation scores.

Comparison with state-of-the-art model. In Table 2, we compare the performance of AoA+GP with state-of-the-art models on offline COCO Karpathy test split. We report the results for two optimizations: cross-entropy loss and CIDEr optimizations. We include the results of the AoA baseline (AoA-BL) @xdefthefnmark2footnotemark that we used to integrate the gazing pattern. We compare our results to the following models: SCST [34], which uses modified visual attention; Up-Down[3], which uses attention over bottom-up features extracted from Faster-RCNN; RFNet [20], which employ a recurrent fusion encoder to fuse features from multiple CNN networks; GCN-LSTM [48], which utilize the relationship between regions in the image through Graph CNN; SGAE [45], which utilizes auto-encoding scene graphs; M2Transformer[9], which employ transformer as a decoder instead of LSTM; and X-LAN[31], which introduced X-Linear attention block. Currently, X-LAN achieved state-of-the-art performance. Comparing our method with X-LAN, our method achieved comparable performance with X-LAN when both models trained with cross-entropy; however, the performance of our method was slightly better in most metrics when optimized by the CIDEr score.

---

*We use use the open-source implementations of the authors’ architectures in both Ada-att and AoA and reproduced the scores. We obtained different scores than reported in their papers.
Table 2. Performance comparison of AoA with gazing pattern (AoA+GP) with state-of-the-art methods on COCO Dataset, where B@N, M, R, and C are short for BLEU, METEOR, ROUGE-L, and CIDEr respectively. All values are reported as percentage(%). The bold value represents the highest value, while underlined one represents the second highest.

<table>
<thead>
<tr>
<th></th>
<th>Cross-Entropy Loss</th>
<th>CIDEr Score Optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B@1</td>
<td>B@2</td>
</tr>
<tr>
<td>SCST[34]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Up-Down[3]</td>
<td>77.2</td>
<td>-</td>
</tr>
<tr>
<td>RFNet[20]</td>
<td>76.4</td>
<td>60.4</td>
</tr>
<tr>
<td>GCN-LSTM[48]</td>
<td>77.3</td>
<td>-</td>
</tr>
<tr>
<td>SGAE[45]</td>
<td>77.6</td>
<td>-</td>
</tr>
<tr>
<td>AoA[19]</td>
<td>77.4</td>
<td>-</td>
</tr>
<tr>
<td>M² Transformer[9]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>X-LAN[31]</td>
<td>78.0</td>
<td>62.3</td>
</tr>
<tr>
<td>AoA-BL</td>
<td>76.8</td>
<td>61.1</td>
</tr>
<tr>
<td>AoA+GP</td>
<td>77.8</td>
<td>62.0</td>
</tr>
</tbody>
</table>

Table 3. Gazing sequence prediction model results. \( \tau \) is Kendall’s tau that computes the correlation between the predicted order and the ground truth order, and its value ranges from -1 (worst) to 1 (the best).

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>( \tau )</th>
<th>PMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinkhorn Network [8]</td>
<td>62.2</td>
<td>0.633</td>
<td>-</td>
</tr>
<tr>
<td>PtrNet-GP (ours)</td>
<td>70.38</td>
<td>0.717</td>
<td>0.476</td>
</tr>
<tr>
<td>PtrNet-order</td>
<td>83.15</td>
<td>0.863</td>
<td>0.720</td>
</tr>
</tbody>
</table>

Table 4. Example of captions generated by the Attention on Attention model (AoA) and AoA with Gazing Pattern (AoA+GP), as well as the ground truth captions (GT).

5.2. Gazing Sequence Model results

In Table 3, we compare our gazing pattern prediction model to the previous work by [8] which uses Sinkhorn Network [30] to order a set of pre-defined regions. Similar to [10], we use accuracy (the percentage of regions that are located in the correct position), Kendall’s tau (\( \tau \)) (correlation between the predicted order and the ground truth order), and Perfect Match Ratio (PMR) (ratio of the exact matching orders) for evaluation.

(PtrNet-GP) outperformed Sinkhorn network significantly (8.18% absolute improvements in accuracy). Pointer network selects the important objects and then learns to order them based on the entity order in the caption as opposed to Sinkhorn network which only order pre-selected objects. For a fair comparison, we trained the pointer network model to order the regions that are already mentioned in the captions **PtrNet-order** similar to that in [8]. We observe that PtrNet-order outperforms the Sinkhorn network by a significant margin.

5.3. Qualitative Results

Table 4 shows some selected captions of the AoA, AoA+GP and the ground truth captions (GT). Although AoA can generate captions relevant to the image, adding gazing patterns can produce more accurate descriptions. For example, the first image in Table 4 shows that AoA generates an object that is not in the image (e.g. fire hydrant) while AoA+GP described it more accurately. We notice similar observations in the second and third images (e.g. window sill and red comforter are generated when the model is boosted with gazing patterns).

6. Conclusion

We present a gazing sub-network that models human perception as additional visual features to the image captioning models. We first estimate the gazing sequence from the
entities in the caption and then adopt the pointer network that automatically produces a similar sequence. Our experiments show that adding a gazing pattern as an additional feature to the image encoder has enhanced the performance of image captioning models.

References


[41] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In IEEE Conference on Computer Vision and Pat-


