Improve Image Captioning by Estimating the Gazing Patterns from the Caption

Rehab Alahmadi[1,2] and James Hahn[1]

[1]Department of Computer Science, The George Washington University
[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).

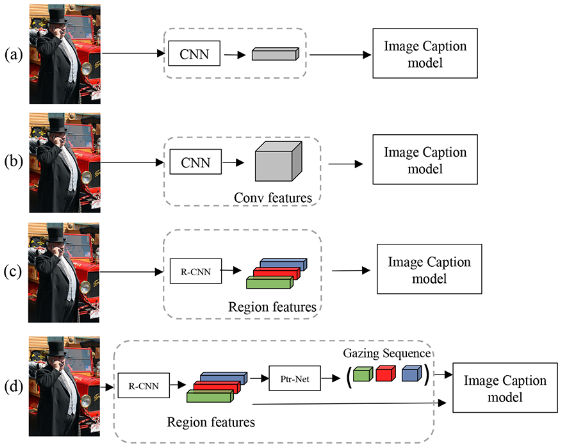


Figure 1. Traditionally, the visual feature is represented as a vector or convolutional map extracted from CNN (a and b), or as a bottom-up features from R-CNN (c). In our method, we add gazing sequence as an additional visual feature (d).

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

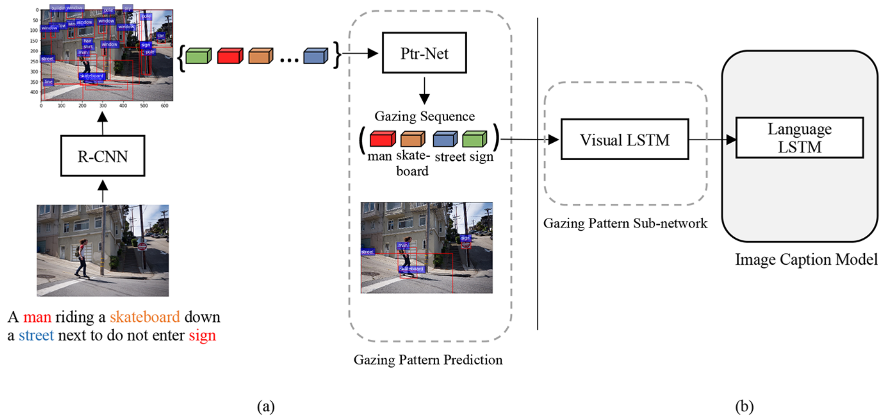

R-CNN
Ptr-Net
Gazing Sequence
man
skate-board
street
sign
A man riding a skateboard down a street next to do not enter sign
Gazing Pattern Prediction
Visual LSTM
Gazing Pattern Sub-network
Language LSTM
Image Caption Model
(a)
(b)

Figure 2. Overview of the model. (a) We predict the gazing sequence from a set of detected objects using Pointer Network (Ptr-Net). (b) The gazing sequence is then inputted to LSTM that encodes the sequence into a fixed size hidden vector, which then initializes the hidden states of the Language LSTM.

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

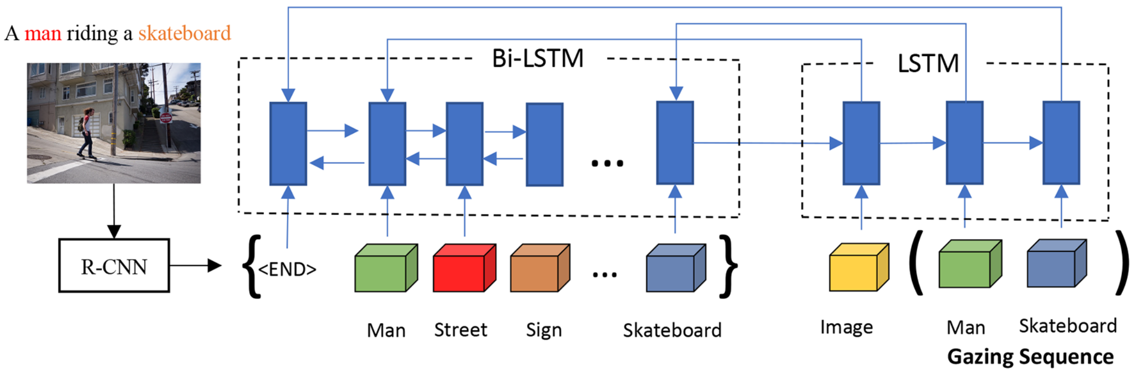


Figure 3. Pointer Network modeling gazing patterns prediction. The model takes set of objects detected by R-CNN as an input and learns to point to the objects that are part of the gazing sequence.

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 then integrated into the attention model. Some studies aggregate the gazing data into a static saliency map without considering their sequential nature [17, 37]. Other studies integrate sequential gazing information in the image caption to improve the attention [38].

Different than previous studies, we extract gazing pattern automatically from the captions rather than using 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

$$R = [r_1, r_2, r_3, ..., r_n] \tag{1}$$

where r_t is a region at position t 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

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

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

3.1.1 Model Features

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

$$r_t = [v_t; t_t; l_t] \tag{3}$$

where v_t, t_t, and l_t 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:

$$x_t = W(r_t) \tag{4}$$

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.[1]

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.

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

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

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

where $v \in R^d$, W_1 and $W_2 \in R^{dxd}$ 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.

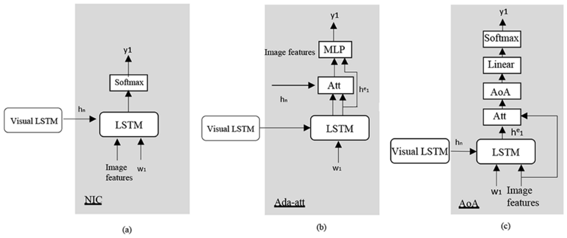


Figure 4. Integrating visual LSTM to (a) Neural Image Caption (NIC) (b) Adaptive attention model (Ada-att), and (c) Attention on Attention model (AoA).

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:

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

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:

$$h_1^e = LanguageLSTM(w_1, h_n^v) \tag{9}$$

where $LanguageLSTM$ is caption generator LSTM in image caption models, h_1^e is the hidden state of the first time step of the $LangaugeLSTM$, 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**[43] 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.

[1] 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.

	Bleu-1	**Bleu-4**	**METEOR**	**ROUGE-L**	**CIDEr**
NIC [43]	67.8	26.8	22.6	49.7	80.2
NIC+Att	70.5	29.7	24.2	51.8	90.7
NIC+GP (ours)	70.5	30.3	24.7	52.5	93.7
NIC+GT-Set	75.2	33.4	26.8	55.0	108.4
NIC+GT-Seq	**75.6**	**34.0**	**27.1**	**56.3**	**110.2**
Ada-att [28]	71.6	30.7	25.2	52.9	97.1
Ada-att+GP (ours)	72.7	31.6	25.4	53.6	100.3
Adat-att+GT-Set	74.8	33.5	**27.7**	55.6	110.9
Ada-att+GT-Seq	**75.6**	**34.2**	27.4	**56.4**	**111.3**
AoA [19]	76.8	36.6	28.3	57.1	117.3
AoA+GP (ours)	**77.8**	**37.8**	28.9	58.7	120.6
AoA+GT-Set	77.7	36.6	28.9	57.6	119.4
AoA+GT-Seq	77.7	37.7	**29.1**	**58.8**	**121.0**

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 attends 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.[2] 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,

[2] For 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[3]. 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 comparing 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) comparing 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 @xdefthefnmark3footnotemark 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; M^2Transformer[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.

[3] 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.

	Cross-Entropy Loss							CIDEr Score Optimization						
	B@1	**B@2**	**B@3**	**B@4**	**M**	**R**	**C**	**B@1**	**B@2**	**B@3**	**B@4**	**M**	**R**	**C**
SCST[34]	-	-	-	30.0	25.9	53.4	99.4	-	-	-	34.2	26.7	55.7	114.0
Up-Down[3]	77.2	-	-	36.2	27.0	56.4	113.5	79.8	-	-	36.6	27.7	56.9	120.1
RFNet[20]	76.4	60.4	46.6	35.8	27.4	56.5	112.5	79.1	63.1	48.4	36.5	27.7	57.3	121.9
GCN-LSTM[48]	77.3	-	-	36.8	27.9	57.0	116.3	80.5	-	-	38.2	28.5	58.3	127.6
SGAE[45]	77.6	-	-	36.9	27.7	57.2	116.7	80.8	-	-	38.4	28.4	58.6	127.8
AoA[19]	77.4	-	-	37.2	28.4	57.5	119.8	80.2	-	-	38.9	29.2	58.8	129.8
M^2 **Transformer**[9]	-	-	-	-	-	-	-	80.8	-	-	39.1	29.2	58.6	131.2
X-LAN[31]	**78.0**	**62.3**	**48.9**	**38.2**	28.8	58.0	**122.0**	80.8	65.6	51.4	39.5	**29.5**	59.2	**132.0**
AoA-BL	76.8	61.1	47.4	36.6	28.3	57.1	117.3	81.1	65.1	50.5	38.3	28.9	58.8	125.9
AoA+GP	77.8	62.0	48.5	37.8	**28.9**	**58.7**	120.6	**82.2**	**66.4**	**51.8**	**39.7**	**29.5**	**60.3**	128.4

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.

	Accuracy	τ	**PMR**
Sinkhorn Network [8]	62.2	0.633	-
PtrNet-GP (ours)	70.38	0.717	0.476
PtrNet-order	**83.15**	**0.863**	**0.720**

Table 3. Gazing sequence prediction model results. τ 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).

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 (τ) (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).

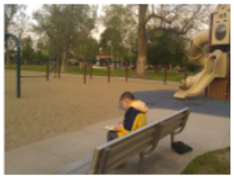
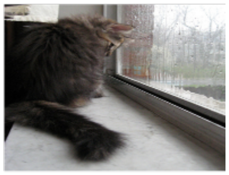
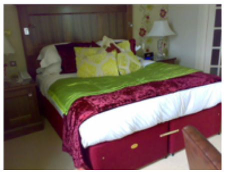
Image	**Captions**
	**AOA**: A man sitting on a bench next to a fire hydrant **AOA+GP**: A young boy sitting on a bench in a park **GT1**: A young man sitting on a park bench next to a playground **GT2**: A child sits on a bench at a playground **GT3**: A boy sits on a bench in a park, working on homework
	**AOA**: A cat looking out of a window **AOA+GP**: A cat is sitting on a window sill looking out the window **GT1**: A cat sitting by a window watching the rain **GT2**: A close up of a cat on a window sill looking the out window **GT3**: A cat sitting in a window watching the rain drip on the window glass
	**AOA**: A bedroom with a bed and a bed in it **AOA+GP**: A bed with a red comforter and pillows on it **GT1**:A bed with blankets, and pillows on it **GT2**:A made bed that has yellow pillows on it **GT3**:A bed set is made of up the colors fuschia, lime,and yellow

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).

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

[1] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. Sort story: Sorting jumbled images and captions into stories. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 925–931. The Association for Computational Linguistics, 2016.

[2] Rehab Alahmadi, Chung Hyuk Park, and James Hahn. Sequence-to-sequence image caption generator. In *Eleventh International Conference on Machine Vision, ICMV 2019, Munich, Germany, March 15, 2019*. SPIE Proceedings, 2019.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society, 2018.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[5] Alexander C. Berg, Tamara L. Berg, Hal Daumé III, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa C. Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, and Kota Yamaguchi. Understanding and predicting importance in images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3562–3569. IEEE Computer Society, 2012.

[6] Kathryn Bock, David E Irwin, Douglas J Davidson, and Willem JM Levelt. Minding the clock. *Journal of Memory and Language*, 48(4):653–685, 2003.

[7] Moreno I Coco and Frank Keller. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive science*, 36(7):1204–1223, 2012.

[8] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[9] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10575–10584. IEEE, 2020.

[10] Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. Deep attentive sentence ordering network. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4340–4349. Association for Computational Linguistics, 2018.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[12] Jingjing Gong, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. End-to-end neural sentence ordering using pointer network. *CoRR*, abs/1611.04953, 2016.

[13] Zenzi M Griffin and Kathryn Bock. What the eyes say about speaking. *Psychological science*, 11(4):274–279, 2000.

[14] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6837–6844. AAAI Press, 2018.

[15] Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. Aligning linguistic words and visual semantic units for image captioning. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 765–773. ACM, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[17] Sen He, Hamed Rezazadegan Tavakoli, Ali Borji, and Nicolas Pugeault. Human attention in image captioning: Dataset and analysis. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8528–8537. IEEE, 2019.

[18] John Henderson and Fernanda Ferreira. *The interface of language, vision, and action: Eye movements and the visual world*. Psychology Press, 2013.

[19] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *International Conference on Computer Vision*, 2019.

[20] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pages 510–526. Springer, 2018.

[21] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society, 2015.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

[24] Alon Lavie and Abhaya Agarwal. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In Chris Callison-Burch, Philipp Koehn, Cameron S. Fordyce, and Christof Monz, editors, *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, pages 228–231. Association for Computational Linguistics, 2007.

[25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10, 2004.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list.

[27] Lajanugen Logeswaran, Honglak Lee, and Dragomir R. Radev. Sentence ordering and coherence modeling using recurrent neural networks. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5285–5292. AAAI Press, 2018.

[28] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3242–3250. IEEE Computer Society, 2017.

[29] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7219–7228. IEEE Computer Society, 2018.

[30] Gonzalo E. Mena, David Belanger, Scott W. Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[31] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10968–10977. IEEE, 2020.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002.

[33] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015.

[34] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society, 2017.

[35] Edgar Rojas-Muñoz, Kyle Couperus, and Juan P. Wachs. DAISI: database for AI surgical instruction. *CoRR*, abs/2004.02809, 2020.

[36] Gautam Somappa and Sivaraman. Making a point with pointer networks : Arranging shuffled stories. 2017.

[37] Yusuke Sugano and Andreas Bulling. Seeing with humans: Gaze-assisted neural image captioning. *CoRR*, abs/1608.05203, 2016.

[38] Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. Generating image descriptions via sequential cross-modal alignment guided by human gaze. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4664–4677. Association for Computational Linguistics, 2020.

[39] Emiel van Miltenburg, Ákos Kádár, Ruud Koolen, and Emiel Krahmer. DIDEC: the dutch image description and eye-tracking corpus. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3658–3669. Association for Computational Linguistics, 2018.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

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

tern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 4566–4575. IEEE Computer Society, 2015.

[42] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in neural information processing systems*, pages 2692–2700, 2015.

[43] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society, 2015.

[44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org, 2015.

[45] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10685–10694. Computer Vision Foundation / IEEE, 2019.

[46] Zhilin Yang, Ye Yuan, Yuexin Wu, William W. Cohen, and Ruslan Salakhutdinov. Review networks for caption generation. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2361–2369, 2016.

[47] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pages 711–727. Springer, 2018.

[48] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pages 711–727. Springer, 2018.

[49] Alfred L Yarbus. *Eye movements and vision*. Springer, 2013.

[50] Yongjing Yin, Linfeng Song, Jinsong Su, Jiali Zeng, Chulun Zhou, and Jiebo Luo. Graph-based neural sentence ordering. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5387–5393. ijcai.org, 2019.

[51] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4651–4659. IEEE Computer Society, 2016.