Leveraging Human Attention in Novel Object Captioning

Xianyu Chen, Ming Jiang, Qi Zhao
Department of Computer Science and Engineering, University of Minnesota
{chen6582, mjiang}@umn.edu, qzhao@cs.umn.edu

Abstract
Image captioning models depend on training with paired image-text corpora, which poses various challenges in describing images containing novel objects absent from the training data. While previous novel object captioning methods rely on external image taggers or object detectors to describe novel objects, we present the Attention-based Novel Object Captioner (ANOC) that complements novel object captioners with human attention features that characterize generally important information independent of tasks. It introduces a gating mechanism that adaptively incorporates human attention with self-learned machine attention, with a Constrained Self-Critical Sequence Training method to address the exposure bias while maintaining constraints of novel object descriptions. Extensive experiments conducted on the nocaps and Held-Out COCO datasets demonstrate that our method considerably outperforms the state-of-the-art novel object captioners. Our source code is available at https://github.com/chenxy99/ANOC.

1 Introduction
Image captioning, a task aiming at generating a natural and concrete sentence that describes an image, has received increasing research attention [Anderson et al., 2018; Cornia et al., 2020]. Driven by the success of deep neural networks, image captioners typically operate on the visual features extracted by image classifiers or object detectors and learn from large datasets to generate sentences with flexible syntactical structures. Although these data-driven methods have demonstrated promising results, their success has been inherently limited by the training images typically focusing on a small number of object categories. Due to this limitation, image captioners fail to describe novel objects that they have not seen in the training images [Hendricks et al., 2016; Agrawal et al., 2019]. Therefore, novel object captioning methods have been proposed to generalize image captioners by exploiting the knowledge of novel objects from externally trained image taggers or object detectors [Yao et al., 2017; Li et al., 2019]. The external knowledge of novel objects is either explicitly injected into the generated captions [Yao et al., 2017; Li et al., 2019] or used as constraints at the test time [Anderson et al., 2017]. Although these methods can help image captioners generalize, the object detectors used for detecting novel objects may still not be general enough, and forcing the inclusion of novel objects into the generated captions may also be suboptimal.

To address these challenges, we present Attention-based Novel Object Captioner (ANOC), the first method to leverage human attention in the task of novel object captioning. Attention is an information selection mechanism that allows humans and machines to focus their visual perception and cognition on the most important visual input. Many computational models apply the self-learned attention mechanism that learns to prioritize inputs based on the training data, yet they cannot generalize to process out-of-domain information. Differently, human attention models that are externally trained to predict people’s eye movements during image-viewing [Huang et al., 2015; Jiang et al., 2015] can offer task-free prior knowledge to characterize generally important information. Our method incorporates both types of attention mechanism in image captioning and adaptively allocates weights between human attention and self attention. We further propose a variant of the Self-Critical Sequence Training (SCST) [Rennie et al., 2017] approach, namely the Constrained SCST (C-SCST), to fine-tune image captioners with test-time metrics, while simultaneously forcing and optimiz-
ing the inclusion of novel objects in the generated captions. It addresses the contextual discrepancy between the training and test settings (i.e., the exposure bias) using a novel baseline reward that ensures the inclusion of novel objects into the generated captions. We demonstrate that our ANOC can generate more specific and fluent captions about novel objects with extensive experiments on two public image captioning datasets: nocaps and Held-Out COCO, with promising quantitative and qualitative results. As shown in Figure 1, incorporating human attention highlights important objects, resulting in better descriptions of novel objects and more natural captions (e.g., the hat is with the man but not the desk).

In sum, the main contributions of this work include: 1. the first novel object captioner that incorporates human attention to describe images with unseen objects, and 2. a novel C-SCST method that addresses the exposure bias while considering the inclusion of novel objects.

2 Related Work

Novel object captioning. Our work is related to novel object captioning methods that leverage unpaired visual and semantic data to describe novel objects [Hendricks et al., 2016; Venugopalan et al., 2017]. Previous studies typically address this problem with sentence templates [Wu et al., 2018; Lu et al., 2018; Feng et al., 2020], copying mechanisms [Mogadala et al., 2017; Yao et al., 2017; Li et al., 2019], or object constraints [Anderson et al., 2017]. In particular, the Constrained Beam Search (CBS) [Anderson et al., 2017] forces the inclusion of multiple novel objects in the captions, facilitating novel object captioning at the test time. These methods are based on the self-attention mechanism that cannot generalize to process out-of-domain information. They also suffer from the contextual discrepancy between the training and test settings (i.e., the exposure bias). Compared with these studies, our work proposes two novel techniques: 1. it leverages human attention models in the novel object captioning task, and 2. it incorporates the CBS approach with SCST to address the exposure bias while satisfying object constraints. Both techniques demonstrate their effectiveness in generating natural and fluent image captions containing novel objects.

Human attention for image captioning. Recent studies have incorporated human attention to guide image captioning [Sugano and Bulling, 2016; Chen and Zhao, 2018; Cornia et al., 2018a; Zhou et al., 2019], yet the incorporation methods are wildly different. Some use human eye-tracking data as an auxiliary input [Sugano and Bulling, 2016], while others acquire gaze prediction results [Cornia et al., 2018a; Zhou et al., 2019] or intermediate features [Chen and Zhao, 2018] from human attention models. While these methods target the conventional image captioning task, our proposed ANOC specifically focuses on using human attention in the description of novel objects. Different from these methods incorporate human attention through early fusion, our proposed method adaptively selects the outputs from two language models, one guided by human attention and the other by the captioner’s self-attention mechanism. This late fusion strategy allows the ANOC to be adaptively trained with the proposed C-SCST, to maximize the test-time performance when describing novel object categories.

Self-critical sequence training. Image captioners are typically trained with maximum likelihood estimation based on the sequential cross-entropy loss. Due to the contextual discrepancy between the training and evaluation settings, this supervised training strategy commonly leads to the exposure bias [Rennie et al., 2017]. To address this problem, SCST directly optimizes the non-differentiable test-time evaluation metrics and demonstrates significant performance improvements in image captioning [Rennie et al., 2017]. In spite of its effectiveness, SCST cannot be directly applied in novel object captioning, because it depends on a greedy sampling approach that does not guarantee the inclusion of novel objects. In this work, we design a novel baseline reward so that SCST can work with the CBS to consider the constraints from external object detectors. This C-SCST method allows us to fine-tune novel object captioners on the in-domain data using test-time metrics, while taking novel objects into account.

3 Approach

This section presents the architecture of the proposed ANOC and the novel object captioning method to adapt it to describe images with novel objects.

3.1 Network Design

The proposed ANOC is an attention-driven image captioner aiming at describing the input image \( I \) with a natural-language sentence \( S = \{\Pi_1, \Pi_2, \cdots, \Pi_{N_s}\} \), where \( N_s \) is the length of the sentence and \( \Pi_i, i = 1, 2, \cdots, N_s \) are one-hot encoded word tokens. It outputs a sequence of probability distributions \( y = (y_0, y_1, \cdots, y_T) \) representing the generated image caption. As illustrated in Figure 2, the ANOC consists of two parallel image captioning pathways, each driven by an attention model. The self-attention pathway is designed following the UpDown captioner [Anderson et al., 2018], using a top-down attention LSTM to prioritize a set of object features extracted using object detectors. The human attention pathway extracts features from a pretrained gaze prediction model [Cornia et al., 2018b] that simulates the image-viewing behavior of humans. Each pathway uses a language model directly conditioned on the features attended by the self-attention or the human attention, respectively. A gating mechanism computes the weights to fuse the predictions from both pathways and generate the final output.

Self attention. The self-attention pathway is designed following the UpDown [Anderson et al., 2018] captioner. It adopts a pretrained object detector [Ren et al., 2017] to extract features from objects detected in the input image. These object features are denoted as \( V = \{v_1, v_2, \cdots, v_m\} \), where \( v_i \in \mathbb{R}^D \) is a \( D \)-dimensional vector pooled over the \( i \)-th region proposal, and \( m \) is the number of detected objects. At each step \( t \), a Top-Down Attention LSTM computes the attention weight of each object. It is conditioned on the object features \( V \) and the embedding of the previous word in the caption \( W_t, \Pi_{t-1} \). The attended features are used as the input of a Language LSTM to generate the output probabilities \( p_i^\theta(y_t|V, y_{0:t-1}; \theta) \) through a Softmax function, where \( \theta \) is a set of trainable parameters.
**Human attention.** The human attention pathway leverages a Saliency Attentive Model (SAM) [Cornia et al., 2018b] pre-trained on a large human attention dataset [Jiang et al., 2015] to predict where people look in images. We extract features from the penultimate convolutional layer of this network, denoted as $S = \{s_1, s_2, \cdots, s_n\}$, where $s_j \in \mathbb{R}^d$ is the $d$-dimensional feature vector at the $j$-th pixel and $n$ is the number of pixels. We use the globally-averaged feature vector $\bar{s} = \sum_{j=1}^n s_j / n$ to represent image features that attract human attention. The human attention pathway takes $\bar{s}$ and the embedding of the previous word in the caption $W_i \Pi_{t-1}$ as the input of a Language LSTM. Finally, through a Softmax function, the human attention pathway outputs probabilities $p_t^s(y_t|\bar{s}, y_{0:t-1}; \theta)$ where $\theta$ is a set of trainable parameters.

**Gating mechanism.** The proposed ANOC maintains a gate function to generate a weight $g_t$ to dynamically combine the output distributions from the self attention pathway and human attention pathway. Specifically, we compute a gate $g_t$ from the input word embedding $W_i \Pi_{t-1}$ and the hidden states of all the LSTM models (i.e., $h_t$, $h^s_t$, and $h^v_t$) encoding potential word distribution. This allows the gate to adaptively decide the weight of human attention. Taking all the information into consideration, the gate is able to accurately balance the outputs from both attention pathways and integrate them to generate the final predictions:

$$g_t(W_i \Pi_{t-1}, h_t, h^s_t, h^v_t|\theta) = \sigma(G_s W_e \Pi_{t-1} + G_h h_t + G_{h^s} h^s_t + G_{h^v} h^v_t + b),$$

where $\sigma$ is the Sigmoid function and $G_s$, $G_h$, $G_{h^s}$, $G_{h^v}$, $b$ are learnable parameters. Given the gate $g_t$, the final word distribution at time step $t$ is a combination of the two probability distributions $p_t^s(y_t|V, y_{0:t-1}; \theta)$ and $p_t^v(y_t|\bar{s}, y_{0:t-1}; \theta)$:

$$p_t(y_t|V, \bar{s}, y_{0:t-1}; \theta) = (1 - g_t) p_t^s(y_t|V, y_{0:t-1}; \theta) + g_t p_t^v(y_t|\bar{s}, y_{0:t-1}; \theta).$$

Thus, the proposed ANOC adaptively integrates the information from self attention pathway and human attention pathway and outputs the final predictions.

### 3.2 CBS-Based Novel Object Captioning

Novel object captioning aims to adapt a pretrained image captioner to describe images with unseen objects. The pretraining of our ANOC requires optimizing the model parameters $\theta$ with supervised training, based on the standard sequential cross-entropy loss:

$$L_{XE}(\theta) = -\sum_{t=1}^T \log p_t^s(y_t|V, y_{0:t-1}; \theta) - \sum_{t=1}^T \log p_t^v(y_t|\bar{s}, y_{0:t-1}; \theta) - \beta \sum_{t=1}^T \log p_t(y_t|V, \bar{s}, y_{0:t-1}; \theta),$$

where the hyper-parameter $\beta$ determines the weight of its corresponding loss term.

To generalize this pretrained image captioner for novel object captioning, we apply an externally trained object detector to detect novel objects in the input images. The detector returns a large number of candidate objects, in which only a few need to be included in the caption. Therefore, we select the objects with their classification probabilities above a minimum threshold $\gamma$ and resolve overlapping bounding boxes based on the class hierarchy of the Open Images dataset [Agrawal et al., 2019; Kuznetsova et al., 2018], which results in a total of $N_o$ selected objects. We apply CBS [Anderson et al., 2017] to force the inclusion of the selected objects in the captions. We take the top-$N_{\text{min}}$ objects based on their classification probabilities as constraints. The constraint sets are defined as $C_i, i = 1, \cdots, f$, where $f$ is the number of Finite State Machines (FSMs) in the beam search. The number of objects in a given constraint set $C_i$ is denoted as $\phi(C_i)$, and $0 \leq \phi(C_i) \leq \min\{N_o, N_{\text{min}}\}$. For each time step $t$ and each constraint set $C_i$, we maintain the top-$k$ captions with the highest probabilities. Finally, we select the first caption that satisfies at least $N_{\text{d}}$ constraints as the final output.
3.3 Constrained SCST

Due to the additional constraints on the out-of-domain semantics (i.e., the novel objects) at test time, CBS-based captioners may not preserve important in-domain semantics. Hence, to naturally and smoothly describe both the in-domain and out-of-domain semantics with a human-like sentence, we propose to fine-tune the image captioner with SCST [Rennie et al., 2017], which also addresses the exposure bias, i.e., the discrepancy between the training- and test-time contexts.

The standard SCST [Rennie et al., 2017] is a policy-gradient method using a greedyly sampled caption to estimate the baseline reward. Because it does not consider the constraints from the CBS, the greedy sampling cannot guarantee the inclusion of novel objects. To address this issue, we propose the Constrained SCST (C-SCST) by taking into account the constraint sets of CBS. Given a reward function \( r ( \cdot ) \), the gradient expression for the expectation of the reward is

\[
\nabla_{\theta} L (\theta) = - \nabla_{\theta} E_{w \sim p_w} [r(w)],
\]

where \( p_w \) is the probability distribution of the generated captions. To apply constraints that force the inclusion of novel objects, based on the constraint sets of the CBS [Anderson et al., 2017], we sample the probability distribution \( p_w \), and compute a baseline reward \( b_j = \sum_{j=1}^{k} r(w^{i,j}) / k \) for each constraint set \( C_i \), to reduce the variance of the gradient, where \( w^{i,j} \) is the \( j \)-th caption in the beam from the constraint set \( C_i \). Formally, the expression of the gradient is denoted as:

\[
\nabla_{\theta} L (\theta) \approx - \frac{1}{lk} \sum_{\phi(C_i) \geq N_d} \sum_{j=1}^{k} (r(w^{i,j}) - b_j) \nabla_{\theta} \log p(w^{i,j})
\]

\[
= - \frac{1}{lk} \sum_{\phi(C_i) \geq N_d} \sum_{j=1}^{k} \sum_{t=1}^{T} (r(w^{i,j}) - b_j) \nabla_{\theta} \log (p_t^{i,j}(y_t | y_{t-1}))
\]

(5)

where \( \theta \) is the model parameters and \( l \) is the number of constraint sets that satisfy \( \phi(C_i) \geq N_d \).

With C-SCST, we fine-tune the ANOC on the CBS-generated captions. Though we fine-tune the training images without novel objects, it helps the image captioner describe both in-domain and out-of-domain semantics more naturally by jointly optimizing them in the same sentence.

4 Experiments

We report experimental details and results to demonstrate the effectiveness of the proposed method. We first present datasets, evaluation metrics, and implementation details. We then present quantitative results in comparison with state of the arts, along with extensive ablation studies for different model components. Finally, we present qualitative examples.

4.1 Datasets and Evaluation Metrics

Datasets. We conduct experiments on two test-beds for novel object captioning: the nocaps [Agrawal et al., 2019] and Held-Out COCO [Hendricks et al., 2016] datasets. The nocaps dataset consists of 15,100 images from the Open Images [Kuznetsova et al., 2018] validation and test sets. Each image has 11 human-annotated captions, of which 10 are used as reference captions for automatic evaluation and the other is used as the human baseline. It is split into a validation set of 4,500 images and a test of 10,600 images. We train our model using the MS COCO training set and evaluate it on the nocaps validation and test sets. The Held-Out COCO dataset [Hendricks et al., 2016] is a subset of MS COCO [Lin et al., 2014] where the following eight object categories are excluded from the training set: bottle, bus, couch, microwave, pizza, racket, suitcase and zebra. We randomly split the COCO validation set and use half of it for validation and the other half for testing, each with 20,252 images.

Evaluation metrics. We evaluate novel object captioners with BLEU [Papineni et al., 2002], METEOR [Banerjee and Lavie, 2005], ROUGE [Lin, 2004], CIDEr [Vedantam et al., 2015], and SPICE [Anderson et al., 2016]. On the Held-Out COCO dataset, we also use the F1 score [Hendricks et al., 2016] to evaluate the model’s ability of describing novel objects. It indicates if each novel object is addressed in the generated caption of a particular image containing the object.

4.2 Implementation Details

We set the hyperparameters \( \beta = 1 \) and \( \gamma = 0.45 \) based on a grid search, which consistently lead to the optimal performance across different settings with the best CIDEr scores. We implement the CBS following the nocaps baseline in [Agrawal et al., 2019]: we set beam size \( k = 5 \) and initialize the FSM with \( f = 24 \) states. We incorporate up to \( N_{\min} = 3 \) selected objects as constraints including two- or three-word phrases. We select the highest log-probability caption that satisfies at least \( N_d = 2 \) constraints.

On both datasets, we train the image captioner for 70,000 iterations with a batch size of 150 [Agrawal et al., 2019] samples and then fine-tune it for 210,000 iterations with a 0.00005 learning rate and a batch size of 1 using the proposed C-SCST. In the C-SCST, we use the CIDEr-D score as the reward function, since it agrees well with human judgement.

4.3 Quantitative Evaluation

Results on the nocaps dataset. We compare our method with the state-of-the-art methods on the nocaps dataset. These methods are NBT [Lu et al., 2018], UpDown [Anderson et al., 2018], Transformer and \( M^2 \) Transformer [Cornia et al., 2020], as well as a variant of UpDown based on the pre-trained language model ELMo [Peters et al., 2019]. For a fair comparison, we apply the same CBS [Anderson et al., 2017] method on all the compared models, based on object detection results of the Faster R-CNN [Ren et al., 2017] trained on the Open Images dataset [Kuznetsova et al., 2018]. We also compare the models with a human baseline. Table 1 and Table 2 report the quantitative results on the nocaps validation and test sets, respectively. The state-of-the-art test-set performances are from the official nocaps leaderboard. As shown in Table 1, our proposed ANOC model significantly outperforms the state-of-the-art approaches on 11/12 of the evaluation metrics by a substantial margin. In particular, compared with the state-of-the-art \( M^2 \) Transformer [Cornia et al., 2020], ANOC achieves a significant improvement of 6.8% in the overall CIDEr (from 75.0 to 80.1) and 6.2% in the out-of-domain CIDEr (from 69.4 to 73.7). It also outperforms
the compared models on the nocaps test set (see Table 2). It is noteworthy that all the compared model are based on the same CBS method, so the performance gains in novel object captioning are mostly attributed to the contributions of human attention and the C-SCST. Moreover, compared with ELMo [Peters et al., 2019], an externally pretrained language model, our method is considerably more effective in boosting the performance of novel object captioning over theUpDown [Anderson et al., 2018] baseline.

### Quantitative results on the Held-Out COCO dataset

We further evaluate ANOC on the Held-Out COCO dataset, in comparison with three types of state-of-the-art methods: models leveraging unannotated text corpora (DCC [Hendricks et al., 2016] and NOC [Venugopalan et al., 2017]), template-based models (NBT [Lu et al., 2018], DNOPC [Hu et al., 2018] and CRN [Feng et al., 2020]), models guided with object detectors (NBT [Lu et al., 2018], DNOCL [Hu et al., 2018], CRN [Feng et al., 2020], and CBS [Anderson et al., 2017]). In Table 3, ANOC significantly outperforms the state-of-the-art methods with substantial margins in the F1, SPICE, METEOR, and CIDEr metrics. In particular, compared with LSTM-P, the existing top-performer on the Held-Out COCO dataset, ANOC achieves a 7.2% improvement in CIDEr (from 88.3 to 94.7) and a 7.7% improvement in METEOR (from 23.4 to 25.2), and a 9.6% improvement in SPICE (from 16.6 to 18.2). The 5.6% improvement in the F1 score also suggests that ANOC performs better in describing the novel objects.

#### Ablation study of the human attention

Table 4 compares ANOC with two baseline methods, each with only self attention or human attention, and the best results are highlighted in bold.

### Table 3: Evaluation results on the Held-Out COCO test set. The best results are highlighted in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>In-Domain</th>
<th>Near-Domain</th>
<th>Out-of-Domain</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B-1</td>
</tr>
<tr>
<td>NBT [Lu et al., 2018]</td>
<td>62.3</td>
<td>10.3</td>
<td>61.2</td>
<td>9.9</td>
</tr>
<tr>
<td>Transformer [Cornia et al., 2020]</td>
<td>74.3</td>
<td>11.0</td>
<td>66.7</td>
<td>10.5</td>
</tr>
<tr>
<td>UpDown [Anderson et al., 2018]</td>
<td>80.8</td>
<td>12.3</td>
<td>73.7</td>
<td>11.5</td>
</tr>
<tr>
<td>UpDown + ELMo [Peters et al., 2019]</td>
<td>79.3</td>
<td>12.4</td>
<td>73.8</td>
<td>11.4</td>
</tr>
<tr>
<td>M\textsuperscript{2} Transformer [Cornia et al., 2020]</td>
<td>81.2</td>
<td>12.0</td>
<td>75.4</td>
<td>11.7</td>
</tr>
<tr>
<td>ANOC</td>
<td>86.1</td>
<td>12.0</td>
<td>80.7</td>
<td>11.9</td>
</tr>
<tr>
<td>Human</td>
<td>84.4</td>
<td>14.3</td>
<td>85.0</td>
<td>14.3</td>
</tr>
</tbody>
</table>

#### Table 4: Ablation study of the attention methods and C-SCST on the nocaps validation set. All the compared methods adopt CBS and the best results are highlighted in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>S</th>
<th>M</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCC [Hendricks et al., 2016]</td>
<td>39.8</td>
<td>13.4</td>
<td>21.0</td>
<td>59.1</td>
</tr>
<tr>
<td>NBT [Lu et al., 2018]</td>
<td>48.5</td>
<td>15.7</td>
<td>22.8</td>
<td>77.0</td>
</tr>
<tr>
<td>NOC [Venugopalan et al., 2017]</td>
<td>51.8</td>
<td>-</td>
<td>20.7</td>
<td>-</td>
</tr>
<tr>
<td>CBS [Anderson et al., 2017]</td>
<td>54.0</td>
<td>15.9</td>
<td>23.3</td>
<td>79.9</td>
</tr>
<tr>
<td>KGA [Mogadala et al., 2017]</td>
<td>54.5</td>
<td>14.6</td>
<td>22.2</td>
<td>-</td>
</tr>
<tr>
<td>LSTM-C [Yao et al., 2017]</td>
<td>55.7</td>
<td>-</td>
<td>23.0</td>
<td>-</td>
</tr>
<tr>
<td>DNOPC [Hu et al., 2018]</td>
<td>57.9</td>
<td>-</td>
<td>21.6</td>
<td>-</td>
</tr>
<tr>
<td>LSTM-P [Li et al., 2019]</td>
<td>60.9</td>
<td>16.6</td>
<td>23.4</td>
<td>88.3</td>
</tr>
<tr>
<td>CRN [Feng et al., 2020]</td>
<td>64.1</td>
<td>-</td>
<td>21.3</td>
<td>-</td>
</tr>
<tr>
<td>ANOC</td>
<td>64.3</td>
<td>18.2</td>
<td>25.2</td>
<td>94.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>B-1</th>
<th>B-4</th>
<th>M</th>
<th>R</th>
<th>C</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self Attention</td>
<td>76.5</td>
<td>18.7</td>
<td>24.0</td>
<td>51.7</td>
<td>73.7</td>
<td>11.3</td>
</tr>
<tr>
<td>+ C-SCST</td>
<td>77.6</td>
<td>18.0</td>
<td>24.3</td>
<td>51.3</td>
<td>77.3</td>
<td>11.2</td>
</tr>
<tr>
<td>Human Attention</td>
<td>75.9</td>
<td>17.5</td>
<td>23.9</td>
<td>51.3</td>
<td>72.3</td>
<td>11.0</td>
</tr>
<tr>
<td>+ C-SCST</td>
<td>77.0</td>
<td>17.9</td>
<td>24.6</td>
<td>51.6</td>
<td>77.0</td>
<td>11.2</td>
</tr>
<tr>
<td>ANOC</td>
<td>76.6</td>
<td>18.6</td>
<td>24.2</td>
<td>51.9</td>
<td>75.0</td>
<td>11.3</td>
</tr>
<tr>
<td>+ C-SCST</td>
<td>78.4</td>
<td>19.1</td>
<td>24.8</td>
<td>52.2</td>
<td>80.1</td>
<td>11.6</td>
</tr>
</tbody>
</table>
the two attention pathways to generating better captions.

**Ablation study of the C-SCST.** We also evaluate the contribution of C-SCST to the performance of CBS-based novel object captioning. In Table 4, applying the C-SCST fine-tuning to the ANOC demonstrates consistent improvements in all the evaluation metrics. In particular, it achieves a remarkable 6.8% improvement in CIDEr (from 75.0 to 80.1), suggesting the effectiveness of the C-SCST. Notably, the application of C-SCST allows ANOC to make better use of human attention: Not only the CIDEr score increases from 77.3 to 80.1, which is a 3.6% performance gain, but also consistent improvements are observed in other metrics. These improvements are higher than non-SCST counterparts. The results indicate that C-SCST can effectively boost the performance of novel object captioning by jointly fine-tuning the self attention and human attention pathways with test-time metrics.

### 4.4 Qualitative Analysis

In addition to the quantitative results, we further demonstrate the effectiveness of our method with qualitative examples. Figure 3 presents a comparison between our ANOC model and the UpDown + CBS baseline. These examples illustrate the images, detected objects, regions attended by the human attention model, and the generated captions. From these examples we can observe that human attention 1. complements the object detector by highlighting important objects missed by the object detector (e.g., *table* in Figure 3a), 2. allows our ANOC to describe the scene with more details (e.g., *dog* in Figure 3b), 3. allows our ANOC to count better in complex scenes (e.g., *two people and two dogs* in Figure 3c), and 4. prioritizes the objects being described to generate more natural descriptions (e.g., *turkey* before *flower* in Figure 3d). These improvements suggest that human attention encodes important image features that object detectors may fail to detect, such as small but salient objects, parts of objects, salient spots in the background, etc.). It also simultaneously highlights multiple regions, allowing the model to capture their relationships and contexts. These characteristics of human attention features significantly improve the diversity and specificity of the generated captions, leading to more fluent and natural descriptions of the novel objects.

### 5 Conclusion

In this paper, we have introduced the ANOC that leverages human attention in the novel image captioning task. We have shown that human attention naturally offers prior knowledge of important visual features, and acts as a complementary modality for prioritizing the description of various objects in the image. Our method introduces a gating mechanism to dynamically combine the information from human attention and self attention. We have also developed a Constrained SCST strategy that addresses the exposure bias while constraining the inclusion of novel objects in the generated captions. Our method has outperformed the state-of-the-art methods on the nocaps and Held-Out COCO datasets. Future efforts will be focused on the exploration of different fusion methods to make the best use of human attention in novel object captioning and other related vision tasks.

**Acknowledgements**

This work is supported by NSF Grants 1908711.
References


