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.

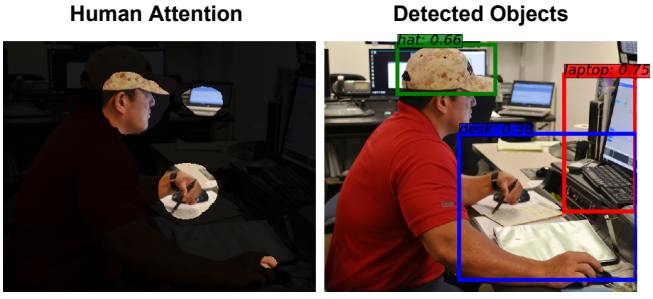

Figure 1: Existing methods use externally trained object detectors to generalize a pretrained image captioner to describe images with (color coded) novel objects. Our proposed ANOC introduces the human attention to complement novel object captioners.

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 including 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; Bujimalla *et al.*, 2020]. 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 $\mathcal{S} = \{\mathbf{\Pi}_1, \mathbf{\Pi}_2, \cdots, \mathbf{\Pi}_{N_s}\}$, where N_s is the length of the sentence and $\mathbf{\Pi}_i,\ i = 1, 2, \cdots, N_s$ are one-hot encoded word tokens. It outputs a sequence of probability distributions $\boldsymbol{y} = (y_0, y_1, \cdots, y_T)$ 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 $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m\}$, where $\boldsymbol{v}_i \in \mathbb{R}^D$ 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 $\boldsymbol{W}_e\mathbf{\Pi}_{t-1}$. The attended features are used as the input of a Language LSTM to generate the output probabilities $p_t^v(y_t|V, y_{0:t-1}; \theta)$ through a Softmax function, where θ is a set of trainable parameters.

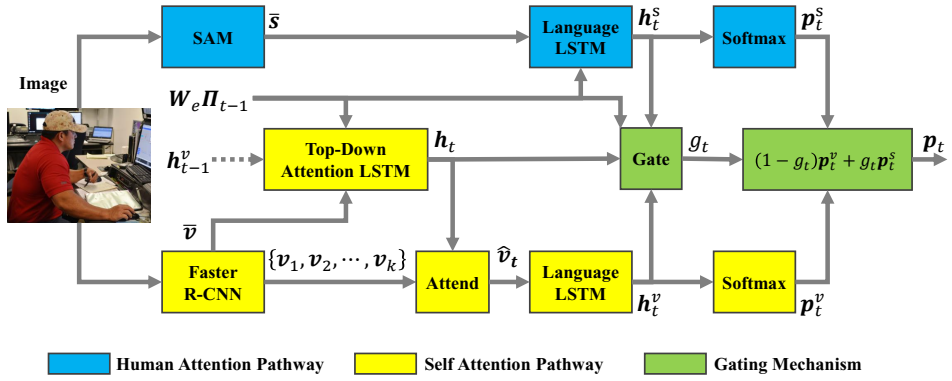

Figure 2: Overview of our Attention-based Novel Object Captioner (ANOC) architecture.

Human attention. The human attention pathway leverages a Saliency Attentive Model (SAM) [Cornia *et al.*, 2018b] pretrained 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 $S = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n\}$, where $\boldsymbol{s}_j \in \mathbb{R}^d$ 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 $\bar{\boldsymbol{s}} = \sum_{j=1}^{n} \boldsymbol{s}_j / n$ to represent image features that attract humans attention. The human attention pathway takes $\bar{\boldsymbol{s}}$ and the embedding of the previous word in the caption $\boldsymbol{W}_e\boldsymbol{\Pi}_{t-1}$ as the input of a Language LSTM. Finally, through a Softmax function, the human attention pathway outputs probabilities $p_t^s(y_t|\bar{\boldsymbol{s}}, y_{0:t-1}; \theta)$ where θ 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 $\boldsymbol{W}_e\boldsymbol{\Pi}_{t-1}$ and the hidden states of all the LSTM models (*i.e.,* $\boldsymbol{h}_t$, $\boldsymbol{h}_t^v$ and $\boldsymbol{h}_t^s$) 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:

$$\begin{aligned} &g_t(\boldsymbol{W}_e\boldsymbol{\Pi}_{t-1}, \boldsymbol{h}_t, \boldsymbol{h}_t^v, \boldsymbol{h}_t^s|\theta) = \\ &\sigma(\boldsymbol{G}_s\boldsymbol{W}_e\boldsymbol{\Pi}_{t-1} + \boldsymbol{G}_h\boldsymbol{h}_t + \boldsymbol{G}_{h_v}\boldsymbol{h}_t^v + \boldsymbol{G}_{h_s}\boldsymbol{h}_t^s + b), \end{aligned} \tag{1}$$

where σ is the Sigmoid function and $\boldsymbol{G}_s$, $\boldsymbol{G}_h$, $\boldsymbol{G}_{h_v}$, $\boldsymbol{G}_{h_s}$, 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 $p_t^v(y_t|V, y_{0:t-1}; \theta)$ and $p_t^s(y_t|\bar{\boldsymbol{s}}, y_{0:t-1}; \theta)$:

$$\begin{aligned} &p_t(y_t|V, \bar{\boldsymbol{s}}, y_{0:t-1}; \theta) = \\ &(1-g_t)p_t^v(y_t|V, y_{0:t-1}; \theta) + g_t p_t^s(y_t|\bar{\boldsymbol{s}}, y_{0:t-1}; \theta). \end{aligned} \tag{2}$$

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 θ with supervised training, based on the standard sequential cross-entropy loss:

$$\begin{aligned} L_{XE}(\theta) = &- \sum_{t=1}^{T} \overbrace{\log p_t^v(y_t|V, y_{0:t-1}; \theta)}^{\text{self attention}} \\ &- \sum_{t=1}^{T} \overbrace{\log p_t^s(y_t|\bar{\boldsymbol{s}}, y_{0:t-1}; \theta)}^{\text{human attention}} \\ &- \beta \sum_{t=1}^{T} \overbrace{\log p_t(y_t|V, \bar{\boldsymbol{s}}, y_{0:t-1}; \theta)}^{\text{gated attention}}, \end{aligned} \tag{3}$$

where the hyper-parameter β 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 γ 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-$N_{\min}$ 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 $0 \leq \phi(C_i) \leq \min\{N_o, N_{\min}\}$. 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_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. Therefore, 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 using the SCST approach [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 greedily 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

$$\nabla_\theta L(\theta) = -\nabla_\theta E_{w\sim p_\theta}[r(w)], \tag{4}$$

where p_θ 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_θ, and compute a baseline reward $b_i = \sum_{j=1}^{k} r(w^{i,j})/k$ 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:

$$\begin{aligned}\nabla_\theta L(\theta) &\approx -\frac{1}{lk}\sum_{\phi(C_i)\geq N_d}\sum_{j=1}^{k}(r(w^{i,j})-b_i)\nabla_\theta \log p(w^{i,j}) \\ &= -\frac{1}{lk}\sum_{\phi(C_i)\geq N_d}\sum_{j=1}^{k}\sum_{t=1}^{T}(r(w^{i,j})-b_i)\nabla_\theta \log\left(p_t^{i,j}(y_t|y_{0:t-1})\right),\end{aligned} \tag{5}$$

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

With this C-SCST method, we fine-tune the ANOC on the CBS-generated novel object captions. Although the fine-tuning is only performed on 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

In this section, 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 the state-of-the-art novel object captioners, along with extensive ablation studies for different model components. Finally, qualitative examples are presented.

4.1 Datasets and Evaluation Metrics

Datasets. We conduct experiments on two commonly used 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 selected 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. The dataset 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. On both datasets, we evaluate novel object captioners with five common metrics for image captioning: BLEU [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], ROUGE [Lin, 2004], CIDEr [Vedantam *et al.*, 2015], and SPICE [Anderson *et al.*, 2016]. In addition, on the Held-Out COCO dataset, we also use the F1 score [Hendricks *et al.*, 2016] to further evaluate the model's ability of describing novel objects. It indicates whether each novel object is addressed in the generated caption of a particular image containing the object. Since CIDEr [Vedantam *et al.*, 2015] is well-accepted to measure the information and smoothness of sentences, the hyper-parameters of our models are tuned on the validation sets for the best CIDEr scores.

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.

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 [Vedantam *et al.*, 2015] score as the reward function, since it agrees well with human judgement [Vedantam *et al.*, 2015].

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 $\mathcal{M}^2$ 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 eval-

Method	In-Domain		Near-Domain		Out-of-Domain		Overall					
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	BLEU-1	BLEU-4	Meteor	ROUGE_L	CIDEr	SPICE
NBT [Lu *et al.*, 2018]	62.3	10.3	61.2	9.9	63.7	9.1	-	-	-	-	61.9	9.8
Transformer [Cornia *et al.*, 2020]	74.3	11.0	66.7	10.5	62.5	9.2	-	-	-	-	66.9	10.3
UpDown [Anderson *et al.*, 2018]	80.8	12.3	73.7	11.5	68.6	9.8	76.5	18.7	24.0	51.7	73.7	11.3
UpDown + ELMo [Peters *et al.*, 2019]	79.3	**12.4**	73.8	11.4	71.7	9.9	-	-	-	-	74.3	11.2
$\mathcal{M}^2$ Transformer [Cornia *et al.*, 2020]	81.2	12.0	75.4	11.7	69.4	10.0	-	-	-	-	75.0	11.4
ANOC	**86.1**	12.0	**80.7**	**11.9**	**73.7**	**10.1**	**78.4**	**19.1**	**24.8**	**52.2**	**80.1**	**11.6**
Human	84.4	14.3	85.0	14.3	95.7	14.0	-	-	-	-	87.1	14.2

Table 1: Evaluation results on the nocaps validation set. All the compared methods adopt CBS for novel object captioning. The best results are highlighted in bold.

Method	In-Domain		Near-Domain		Out-of-Domain		Overall					
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	BLEU-1	BLEU-4	Meteor	ROUGE_L	CIDEr	SPICE
NBT [Lu *et al.*, 2018]	63.0	10.1	62.0	9.8	58.5	8.8	73.4	12.9	22.1	48.7	61.5	9.7
UpDown + ELMo [Peters *et al.*, 2019]	76.0	11.8	74.2	11.5	66.7	9.7	76.6	18.4	24.4	51.8	73.1	11.2
ANOC	**85.8**	**12.4**	**79.7**	**11.8**	**68.5**	**10.0**	**78.8**	**19.7**	**25.1**	**52.5**	**78.5**	**11.6**
Human	80.6	15.0	84.6	14.7	91.6	14.2	76.6	19.5	28.2	52.8	85.3	14.6

Table 2: Evaluation results on the nocaps test set. All the compared methods adopt CBS for novel object captioning. The best results are highlighted in bold.

Method	F1	SPICE	METEOR	CIDEr
DCC [Hendricks *et al.*, 2016]	39.8	13.4	21.0	59.1
NBT [Lu *et al.*, 2018]	48.5	15.7	22.8	77.0
NOC [Venugopalan *et al.*, 2017]	51.8	-	20.7	-
CBS [Anderson *et al.*, 2017]	54.0	15.9	23.3	79.9
KGA-CGM [Mogadala *et al.*, 2017]	54.5	14.6	22.2	-
LSTM-C [Yao *et al.*, 2017]	55.7	-	23.0	-
DNOC [Wu *et al.*, 2018]	57.9	-	21.6	-
LSTM-P [Li *et al.*, 2019]	60.9	16.6	23.4	88.3
CRN [Feng *et al.*, 2020]	64.1	-	21.3	-
ANOC	**64.3**	**18.2**	**25.2**	**94.7**

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

uation metrics by a substantial margin. In particular, compared with the state-of-the-art $\mathcal{M}^2$ 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 the UpDown [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], DNOC [Wu *et al.*, 2018] and CRN [Feng *et al.*, 2020]), models guided with object detectors (NBT [Lu *et al.*, 2018], DNOC [Wu *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.

Method	C-SCST	BLEU-1	BLEU-4	Meteor	ROUGE_L	CIDEr	SPICE
Self Attention (UpDown)		76.5	18.7	24.0	51.7	73.7	11.3
Human Attention		75.9	17.5	23.9	51.3	72.3	11.0
ANOC		76.6	18.6	24.2	51.9	75.0	11.3
Self Attention (UpDown)	✓	77.6	18.0	24.3	51.3	77.3	11.2
Human Attention	✓	77.0	17.9	24.6	51.6	77.0	11.2
ANOC	✓	**78.4**	**19.1**	**24.8**	**52.2**	**80.1**	**11.6**

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

Ablation study of the human attention. Table 4 compares ANOC with two baseline methods, each with only self attention or human attention, as well as their non-SCST variants. It is noteworthy that human attention itself can already provide sufficient information for describing novel objects, which leads to a comparable performance with the self attention as implemented by the UpDown [Anderson *et al.*, 2018] method. By integrating self attention with human attention, ANOC achieves a 75.0 CIDEr score without C-SCST, and an 80.1 CIDEr score with C-SCST. Both outperforming the existing UpDown-based methods (see Table 1). This promising result suggests that dynamically allocating the combination weights using the gating mechanism can effectively integrate 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-

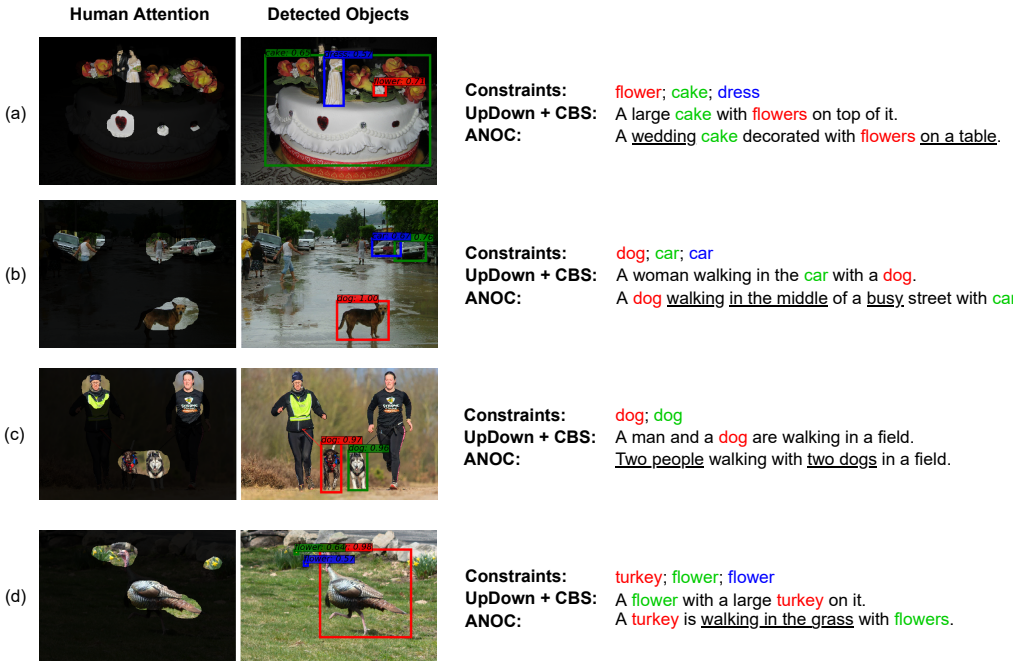

Figure 3: Qualitative results ANOC compared with the UpDown + CBS baseline on the nocaps validation set. The color-coded words indicate the objects detected by the external object detector and the underlines highlight improvements of ANOC over the baseline.

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 objection captioning and other related vision tasks.

Acknowledgements

This work is supported by NSF Grants 1908711.

References

[Agrawal *et al.*, 2019] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv

Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. *ICCV*, 2019.

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. *ECCV*, 2016.

[Anderson *et al.*, 2017] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *EMNLP*, 2017.

[Anderson *et al.*, 2018] 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. *CVPR*, 2018.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. *ACLW*, 2005.

[Bujimalla *et al.*, 2020] Shashank Bujimalla, Mahesh Subedar, and Omesh Tickoo. B-SCST: Bayesian self-critical sequence training for image captioning. *CoRR, abs/2004.02435*, 2020.

[Chen and Zhao, 2018] Shi Chen and Qi Zhao. Boosted attention: Leveraging human attention for image captioning. *ECCV*, 2018.

[Cornia *et al.*, 2018a] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Paying more attention to saliency: Image captioning with saliency and context attention. *TOMM*, 2018.

[Cornia *et al.*, 2018b] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an LSTM-based saliency attentive model. *TIP*, 2018.

[Cornia *et al.*, 2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. M^2: Meshed-memory transformer for image captioning. *CVPR*, 2020.

[Feng *et al.*, 2020] Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, and Yi Yang. Cascaded revision network for novel object captioning. *TCSVT*, 2020.

[Hendricks *et al.*, 2016] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. *CVPR*, 2016.

[Huang *et al.*, 2015] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. *ICCV*, 2015.

[Jiang *et al.*, 2015] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. *ICCV*, 2015.

[Kuznetsova *et al.*, 2018] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *CoRR, abs/1811.00982*, 2018.

[Li *et al.*, 2019] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. *CVPR*, 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *ECCV*, 2014.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. *ACLW*, 2004.

[Lu *et al.*, 2018] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. *CVPR*, 2018.

[Mogadala *et al.*, 2017] Aditya Mogadala, Umanga Bista, Lexing Xie, and Achim Rettinger. Describing natural images containing novel objects with knowledge guided assitance. *CoRR, abs/1710.06303*, 2017.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. *ACL*, 2002.

[Peters *et al.*, 2019] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *ACL*, 2019.

[Ren *et al.*, 2017] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *TPAMI*, 2017.

[Rennie *et al.*, 2017] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *CVPR*, 2017.

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

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. *CVPR*, 2015.

[Venugopalan *et al.*, 2017] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. *CVPR*, 2017.

[Wu *et al.*, 2018] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. *MM*, 2018.

[Yao *et al.*, 2017] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. *CVPR*, 2017.

[Zhou *et al.*, 2019] Lian Zhou, Yuejie Zhang, Yu-Gang Jiang, Tao Zhang, and Weiguo Fan. Re-caption: Saliency-enhanced image captioning through two-phase learning. *TIP*, 2019.