CoNAN: A Complementary Neighboring-based Attention Network for Referring Expression Generation

Jungjun Kim*
Korea University, Republic of Korea
jj_kim@korea.ac.kr

Hanbin Ko
Korea University, Republic of Korea
hb_ko@korea.ac.kr

Jialin Wu
University of Texas at Austin
jialinwu@utexas.edu

Abstract

Daily scenes are complex in the real world due to occlusion, undesired lighting conditions, etc. Although humans handle those complicated environments well, they evoke challenges for machine learning systems to identify and describe the target without ambiguity. Most previous research focuses on mining discriminating features within the same category for the target object. One the other hand, as the scene becomes more complicated, human frequently uses the neighbor objects as complementary information to describe the target one. Motivated by that, we propose a novel Complementary Neighboring-based Attention Network (CoNAN) that explicitly utilizes the visual differences between the target object and its highly-related neighbors. These highly-related neighbors are determined by an attentional ranking module, as complementary features, highlighting the discriminating aspects for the target object. The speaker module then takes the visual difference features as an additional input to generate the expression. Our qualitative and quantitative results on the dataset RefCOCO, RefCOCO+, and RefCOCOg demonstrate that our generated expressions outperform other state-of-the-art models by a clear margin.

1 Introduction

Generating referring expressions (Mao et al., 2016; Yu et al., 2016; Liu et al., 2017; Yu et al., 2017; Tanaka et al., 2019), which identify target objects with simple words and phrases in everyday discourse, has attracted attention from both computer vision (CV) and natural language processing (NLP) communities. With the rapid development of RNNs (Hochreiter and Schmidhuber, 1997; Cho et al., 2014; Bahdanau et al., 2014) and the emergence of transformers (Vaswani et al., 2017), machine learning systems can generate linguistically correct expressions in most cases. However, the remaining issue in the referring expression generation (REG) field is to avoid ambiguities ($i.e.$ the generated expression should refer to a unique target object). This issue becomes increasingly important when referring to an object in complex daily scenes where occlusion, undesired lighting conditions, complex formation of objects, occur regularly. This complex nature inhibits the system from mining the unique features automatically for the target object among its visually similar ones.

Previous works have mainly investigated model architectures to generate less ambiguous expressions. Speaker-Listener models (Mao et al., 2016; Liu et al., 2017) are widely adopted to encourage a speaker to generate expressions that can be comprehended by listener model. Further, Yu et al. (2017) employs a reinforcer module to reward the system if the generated expression is bonded to the target object. In order to find discriminative features for the target object and generate less ambiguous expression, plenty of research utilizes the visual differences between the target objects and the objects that belong to the same category determined by an object detector.

However, when multiple visually similar objects appear in the scene, mining the discriminative features for the target object becomes challenging. Instead, a human would use the surrounding objects

*Corresponding author

Proceedings of the 28th International Conference on Computational Linguistics, pages 1952–1962
Barcelona, Spain (Online), December 8-13, 2020

to help clarify the target one. Motivated by that, we propose a Complementary Neighboring-based Attention Network (CoNAN) that explicitly utilizes and highlights the visual difference between the target object and its neighbors, instead of mining discriminative feature within a class. CoNAN first finds and computes visual differences of the k spatial neighbors for each target object, and then uses the attentional ranking module to rank the potential contribution of each the neighbor object. Finally, the speaker (expression generator) in CoNAN additionally takes the top-M ranking visual differences together with the target object and the global representation as inputs to generate referring expressions.

Note that the CoNAN is compatible with most current learning-based expression generation systems. In particular, we adopt SLR(Yu et al., 2017) as our baseline system. Experimental evaluation shows a significant improvement for the generated expression compared to the state-of-the-art on the three RefCOCO datasets.

2 Related Work

2.1 Image Captioning

Image captioning (Vinyals et al., 2015; Anderson et al., 2018) is the task of generating textual sentences of the given image. Similarly, the referring expression generation task aims at describing a specific object in the daily environment unambiguously. Therefore, it requires machine learning system to figure out the key discriminating aspect of the target object for unambiguity while image captions only describe the general visual content. Most recent approaches use either recurrent models (Anderson et al., 2018) or visual transformers (Lu et al., 2019), on top of object-based bottom-up attention for speaker models. To achieve the unambiguity, REG models employ another comprehension module to check if the generated expressions can be grounded back to the target object.

2.2 Referring Expression Datasets

Referring expression generation (REG) has been studied for a long time using artificial dataset. The field has become more active with the appearance of RefCLEF (Kazemzadeh et al., 2014), a large-scale dataset with 20,000 real-world images. RefCLEF was collected in a two-player game, where one player clicks on the correct object with given the expression generated by another player. If the player correctly matches the object and the expression, both players get points and their roles switch. With the same idea, the authors collected RefCOCO and RefCOCO+ dataset from COCO images (Yu et al., 2016). The two datasets each contain about 50,000 objects. RefCOCO+ additionally uses location information for the expressions which are prohibited on RefCOCO. RefCOCOg (Mao et al., 2016) uses a non-interactive framework to build more complex expressions with further details that contain 54,822 objects with 85,474 referring expressions. Tanaka et al. (2019) proposes RefGTA that contains a complex composition of images from GTA V with sufficiently diverse appearances and locations.

2.3 Referring Expression Generation

Referring expression generation aims at generating unambiguous sentence given a specific region or object in a full image. Initial works have been studied on rule-based approaches (Gupta and Stent, 2005; Janarthanam and Lemon, 2010). Since large-scale datasets (RefCOCO, RefCOCO+, RefCOCOg, etc) were collected, many studies have tried to use the CNN-LSTM framework in the real world images (Mao et al., 2016; Yu et al., 2016; Liu et al., 2017; Yu et al., 2017; Tanaka et al., 2019) for automation.

To reduce the ambiguity of object descriptions, Mao et al. (2016) introduced Maximum Mutual Information (MMI) training which induces the speaker to generate more discriminative sentences based on the listener's response. In detail, the speaker is trained to generate more descriptive captions for the specific object so that the listener can easily localize the specific region. Yu et al. (2016) proposed to incorporate a better measure of visual context into the speaker to jointly generate expressions for all same category objects depicted in an image. Liu et al. (2017) introduced attribute embedding generation which improves the visual representation of the generation model. Yu et al. (2017) proposed a unified framework for the tasks of generation and comprehension where speaker-listener are trained complementarily by end-to-end learning with the reinforcer giving guidance to the speaker to generate a more discriminative

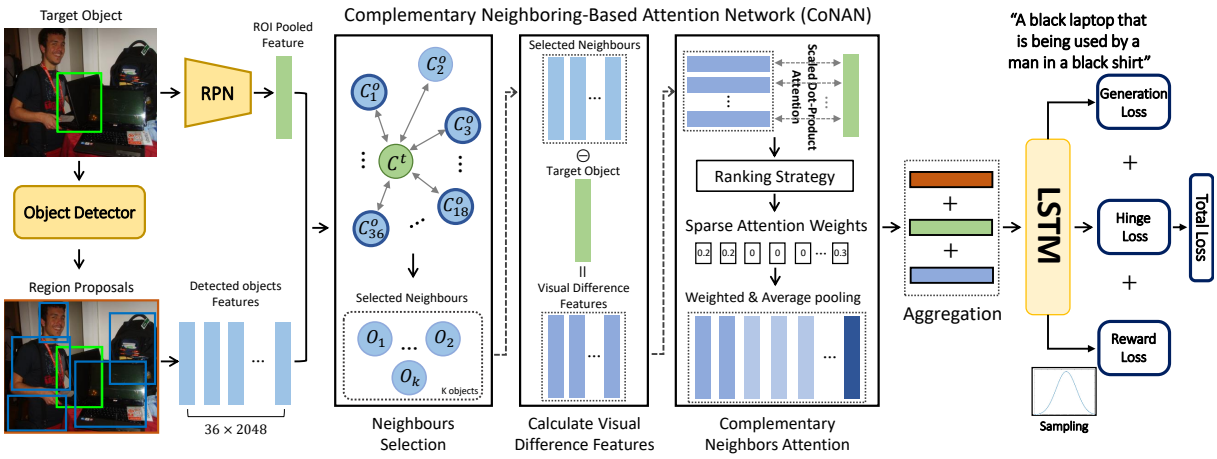

Figure 1: Framework of CoNAN: we extract the target feature and detected object features. We select the neighbors of the target based on euclidean distance metric. We calculate the visual difference between target and neighbor features. We perform a scaled dot-product attention function with ranking strategy. The aggregated features consist of global feature, target feature, weighted visual difference features and location/size difference features. We then train the expression generator using those aggregated features by minimizing the total loss.

sentence. Tanaka et al. (2019) focused on utilizing the environment around the target easy for a human to locate a target region.

3 Model

In this section, we present a Complementary Neighboring-based Attention Network (CoNAN) for generating unambiguous referring expressions. In particular, we first extract k neighbor objects for each target object, detailed in Section 3.1. To mine discriminating features for the target object, visual differences between it and its neighbors are utilized as complementary inputs. To better encode the local context, we also employ an attentional ranking strategy that weighs the neighbors to select meaningful ones in Section 3.2. Finally, we present an expression generating module in Section 3.3 that takes the attentional visual difference, target object feature, and the global features as inputs to generate high quality expressions.

3.1 Extracting Neighbor Objects

We present the approach of extracting the set of neighbor objects for the target object $o^\star$. To avoid duplication, we first perform non-maximum-suppression (NMS) to filter out the objects whose intersection-of-union (IoU) with the target box is over 0.5. Then, we extract k-nearest neighbors according to Euclidean distance between the center of the neighbors' bounding boxes and that of the target one. We denote the target object feature as $o^\star$ and the i-th object's as o_i.

3.2 Visual Difference as Complementary Features

Yu et al. (2016) emphasizes the importance of using the visual difference between the target object and the other from the same category to reduce the ambiguity. As a result, both unique attributes and spatial relationships to characterize the target object can be considered.

Instead of comparing to the objects from the same category, our visual differences compare the target object with all of the neighbors to mine the complementary aspects of the target in the local complex scene. To avoid overly complex contexts and preserve the briefness of the expression, we construct an attentional ranking module that ranks and selects interesting neighbor objects when generating the expression for the target object.

3.2.1 Computing Visual Differences

We adopted the bottom-up features as the representations for the target and its neighbor objects. In particular, following (Anderson et al., 2018), the Visual Genome (Krishna et al., 2017) pretrained Faster-RCNN (Ren et al., 2015) is used as the object feature extractor, resulting in a $2,048$-d vector for each object in the image. The visual difference δ_i^v between the target object, $o^\star$, and the i-th neighbor object, o_i, are calculated as $\delta_i^v = o_i - o^\star$.

3.2.2 Complementary Neighboring-based Attention

In contrast to (Yu et al., 2016), our system utilizes the visual differences between the target object and its neighbors for all categories to mine the complementary features.

Considering all the neighbors may introduce an overly complex local context. To address this issue, our system learns sparse attention for each neighbor, and only select top-M meaningful objects as the concise local context.

Technically, given the target feature $o^\star$ and visual difference features δ_i^v for the i-th object, we compute the scaled attention α_i as shown in Eq.1. In particular, both the target feature $o^\star$ and visual difference δ_i^v first go through a separate feed-forward network, which then we compute the attention logits by the inner product of the out projected features scaled by $\frac{1}{\sqrt{d}}$. Note that, f denote a linear transformation, where different f do not share parameters, d the dimension of the hidden feature vector.

$$\alpha_i = \frac{f(o^\star)^T f(\delta_i^v)}{\sqrt{d}} \tag{1}$$

To focus on helpful neighbors for generating unambiguous yet concise expressions, we only select top-M neighbors according to the learnt attention logits α_i to form a complementary neighbor object set $\mathcal{S} = \{i|\alpha_i \geq topM\}$, where $topM$ denotes the M largest attention logits.

Then, the final complementary visual difference features δ is computed as the weighted sum of the visual differences of the object in $\mathcal{S}$ as shown in Eq. 2

$$\delta^v = \sum_{i \in \mathcal{S}} \text{softmax}(\alpha_i)\delta_i^v \tag{2}$$

3.3 Referring Expression Generator

We employ five different types of features to generate referring expressions using CNN-LSTM framework. In particular, we consider the target object $o^\star$, global context g, target location/size l, target context δ^v, target location/size context δ^l.

Global context g is modeled as averaged feature vector of all the detected objects using the pretrained Faster-RCNN in the image. The location/size representation of target is modeled as a 5 dimension vectors $l = \left[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}\right]$, where w, h denote the width and height of the target bounding box and W, H are the width and height of the image, x_{tl}, y_{tl}, x_{br}, y_{br} are the coordinates of the top-left, top-right, bottom-left, bottom-right corner. This feature presents the relative position and the size of the object.

With the selected neighbors, we perform complementary neighboring attention to obtain fine-grained target context δ^v as described in the previous section. The final visual representation v is a combination of the above features followed by one linear layer, $v = W_m[o^\star, g, l, \delta^v, \delta^l]$.

We use v_i to denote the joint feature v that regards the i-th object as the target object, and use r_i to denote the human expression for the i-th object. To generate the expressions for each referred object, the joint feature v_i is fed into an LSTM and we minimize the negative log-likelihood with the parameters θ as shown in Eq. 3.

$$L_s^1(\theta) = -\sum_i logP(r_i|v_i;\theta) \tag{3}$$

3.4 Training Objectives

Following Mao et al. (2016), we use the Maximum Mutual Information (MMI) constraint to encourage the model to generate expression for the target object o_i that can be discriminated from the expression for another object. In particular, we consider two prior knowledge in advance, (1) ground-truth expression r_i should be more likely generated using the target object o_i than other randomly sampled objects o_k (2) the target object is more likely to generate the ground-truth expression r_i instead of other expression r_j for the positive pairs. Therefore, we adopt a margin loss as shown in Eq. 4. Note that, λ_1^s, λ_2^s, M_1, M_2 are hyper-parameters

$$L_s^2(\theta) = \sum_i \lambda_1^s max(0, M_1 + logP(r_i|v_k) - logP(r_i|v_i)) + \lambda_2^s max(0, M_2 + logP(r_j|v_i) - logP(r_i|v_i)) \quad (4)$$

We also use a reinforcer model (Yu et al., 2017) to generate a more precise and discriminative expression for the target object. Specifically, we build an MLP network to evaluate the consistency between the generated expression and visual features. Then, we use the evaluation score as a reward. In particular, we use the local-scene-aware target object feature v_i as the visual feature, and an LSTM to encode the generated expression as the sentence feature. We adopt the policy-gradient technique to optimize the reward function as shown in Eq. 5.

To achieve better performance, we adopt the re-ranking mechanism that selects the generated expression whose referred object by the listener module is the closest to the target one.

$$\bigtriangledown_\theta J = -E_{P(W_{1:T}|v_i)}[F(w_{1:T}, v_i) \bigtriangledown_\theta logP(w_{1:T}|v_i; \theta)] \quad (5)$$

The overall loss of our speaker model L_s is a summation of (Eqn. 3), (Eqn. 4) and (Eqn. 5) where λ^r is a hyper-parameter on the weight of reward loss term

$$L_s = L_1^s(\theta) + L_2^s(\theta) + \lambda^r J(\theta) \quad (6)$$

4 Experiments

4.1 Datasets

Our model is trained and evaluated on the three state-of-the-art referring expression datasets, RefCOCO, RefCOCO+ and RefCOCOg. Each dataset use the image data from COCO (Lin et al., 2014), where RefCOCO and RefCOCO+ are collected using ReferitGame (Kazemzadeh et al., 2014), and RefCOCOg is collected with a non-interactive setting. Further details of each dataset are listed in following sections:

RefCOCO(UNC RefExp). (Yu et al., 2016) is composed of 19,994 images with 142,209 referring expressions for 50,000 objects. The main characteristics of this dataset are that it contains a frequent amount of people compared to objects. Therefore, for testing, we split person vs objects: images with multiple people (Test A) and images with multiple objects (Test B).

RefCOCO+. (Yu et al., 2016) is composed of 19,992 images with 141,564 expressions for 49,856 objects. The main difference from the dataset RefCOCO is that the players were allowed to use location words to describe the objects which focus on the appearance-based description, e.g. left corner hat, top-right. Similar to RefCOCO, the splits are divided into tests for humans (Test A) and test for objects (Test B).

RefCOCOg(Google RefExp). (Mao et al., 2016) is composed of 26,711 images with 85,474 referring expressions for 54,822 objects. Compared to RefCOCO and RefCOCO+, the dataset contains longer sentences with more details.

	Features	RefCOCO				RefCOCO+				RefCOCOg	
		Test A		Test B		Test A		Test B		Val	
		Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr
SLR (Yu et al., 2017)	VGGNet	0.268	0.697	0.329	1.323	0.204	0.494	0.202	0.709	0.154	0.592
SLR+rerank	VGGNet	0.296	0.717	0.340	1.320	0.213	0.520	0.215	0.735	0.159	0.662
re-SLR (Tanaka et al., 2019)	VGGNet	0.279	0.729	0.334	1.315	0.201	0.491	0.211	0.757	0.146	0.679
re-SLR+rerank	VGGNet	0.278	0.717	0.332	1.262	0.198	0.476	0.206	0.721	0.150	0.676
baseline: re-SLR	ResNet	0.296	0.804	0.341	1.358	0.220	0.579	0.221	0.798	0.153	0.742
RefGTA-SR	ResNet	0.307	0.865	0.343	1.381	0.242	0.671	0.220	0.812	0.164	0.738
RefGTA-SR+rerank	ResNet	0.310	0.842	0.348	1.356	0.241	0.656	0.219	0.782	0.167	0.773
RefGTA-SLR	ResNet	0.310	0.859	0.342	1.375	0.241	0.663	0.225	0.812	0.164	0.763
RefGTA-SLR+rerank	ResNet	0.313	0.837	0.341	1.329	0.242	0.664	0.228	0.787	0.170	0.777
Our-SR	ResNet	0.119	0.230	0.100	0.162	0.152	0.316	0.114	0.398	0.086	0.223
Our-SR+rerank	ResNet	0.163	0.303	0.136	0.191	0.176	0.405	0.124	0.442	0.092	0.258
Our-SR+attn	ResNet	0.201	0.253	0.210	0.253	0.240	0.453	0.210	0.451	0.094	0.210
Our-SR+attn+rerank	ResNet	0.222	0.313	0.236	0.279	0.261	0.470	0.223	0.483	0.094	0.233
Our-SLR	ResNet	0.322	0.905	0.342	1.393	0.260	0.722	0.235	0.853	0.177	0.896
Our-SLR+rerank	ResNet	0.324	0.905	0.346	1.362	0.276	**0.769**	0.235	0.832	0.177	0.854
Our-SLR+attn	ResNet	0.328	0.912	0.351	**1.422**	0.281	0.750	0.243	0.860	0.180	0.905
Our-SLR+attn+rerank	ResNet	**0.330**	**0.915**	**0.354**	1.410	**0.288**	0.761	**0.250**	**0.876**	**0.183**	**0.910**

Table 1: Comparison of our results with state-of-the-art baseline methods on Referring Expression Dataset of RefCOCO, RefCOCO+, RefCOCOg. "+rerank" notes the reranking process for the generated expression according to the listener module. "+attn" indicates the addition of scaled dot-product attention with ranking strategy. SLR denotes the original SLR model, and the re-SLR is a reimplemented version that uses ResNet as the image feature extractor from (Tanaka et al., 2019).

4.2 Implementation and Training Details

4.2.1 Implementation

We optimize the speaker module using the Adam (Kingma and Ba, 2014) optimizer with a batch size of 128 with initializing the learning rate to 4e-4. The learning rate is set to decay by 0.5 every 500 iterations. The size of the hidden state and word embedding is set to 512. Also, we empirically found that taking 20 neighbors with 0.2 IoU in the NMS stage achieves optimal results. In the choice of ranking strategy, we set m to 8 for obtaining the sparse attention weights. For reinforcement learning, our model generates 3 sampled sentences to estimate the rewards. During test phase, we use a beam search with a beam size of 10. We set $\lambda_1^s = 1$, $\lambda_2^s = 0.1$ and $M_1 = 1$, $M_2 = 1$ for the hyper-parameters of the margin loss. We set the weight of the reward loss in the total loss function as $\lambda^r = 1$.

For the object representation, following (Anderson et al., 2018), we use object detection as bottom-up attention, which provides salient image regions with clear boundaries. In particular, a Faster R-CNN head (Ren et al., 2015) in conjunction with a ResNet-101 base network (He et al., 2016) is adopted as our detection module. The detection head is first pre-trained on the Visual Genome dataset (Krishna et al., 2017) and is capable of detecting $1,600$ objects categories and 400 attributes. To generate an output set of object features in the image, we take the final detection outputs and perform non-maximum suppression (NMS) for each object category using an IoU threshold of 0.7. Finally, a fixed number of 36 detected objects for each image are extracted as the image features ($2,048$ dimensional vector for each object)

4.2.2 Training Details

We trained our referring expression generator on three series of RefCOCO, RefCOCO+, RefCOCOg datasets following with the LSTM loss, reward loss, and hinge loss. In particular, we first train the reinforcer model by maximizing the reward for the consistency of image features and sentence features. We then jointly train the speaker and listener model with reinforcer's reward.

4.3 Comparison with State-Of-The-Arts Models

In this section, we perform both quantitative and qualitative experiments for SLR (Yu et al., 2017) and RefGTA (Tanaka et al., 2019). For quantitative analysis, we evaluate our generated referring expressions

	RefCOCO		RefCOCO+		RefCOCOg
	Test A	Test B	Test A	Test B	Val
SLR (ensemble) (Yu et al., 2017)	80.08%	81.73%	65.40%	60.73%	74.19%
re-SLR(ensemble) (Tanaka et al., 2019)	78.43%	81.33%	64.57%	60.48%	70.95%
baseline:re-SLR (Listener)	81.14%	80.80%	68.16%	59.69%	72.36%
RefGTA SLR (Listener) (Tanaka et al., 2019)	79.05%	80.31%	65.75%	62.18%	73.39%
Our SLR (Listener)	82.67%	78.83%	**75.80%**	63.69%	78.35%
Our SLR+attn (Listener)	**83.46%**	80.08%	74.01%	**64.37%**	**79.04%**
RefGTA SR (Reinforcer)	80.44%	**81.04%**	67.81%	58.97%	74.94%
Our SR (Reinforcer)	82.17%	79.04%	74.88%	62.81%	78.41%
Our SR+attn (Reinforcer)	80.06%	78.79%	74.20%	60.36%	74.56%
baseline:re-SLR (Speaker)	80.70%	81.71%	68.91%	60.77%	72.55%
RefGTA SLR (Speaker)	83.05%	81.84%	72.37%	59.13%	74.79%
Our SLR (Speaker)	83.21%	78.55%	**76.40%**	63.75%	78.31%
Our SLR+attn (Speaker)	**83.86%**	80.45%	74.56%	**65.52%**	**80.34%**
RefGTA SR (Speaker)	82.45%	**82.00%**	72.07%	61.06%	70.35%
Our SR (Speaker)	67.98%	64.83%	61.04%	48.40%	63.85%
Our SR+attn (Speaker)	70.22%	67.04%	62.15%	53.08%	66.78%

Table 2: Comprehension evaluation on the RefCOCO, RefCOCO+ and RefCOCOg. Ensemble refers to the use of both speaker and listener or reinforcer. Our three modules (speaker, listener, reinforcer) show better performance in most cases compared to the previous state-of-the-art models. "+attn" states that the model is applied to the scaled dot-product attention with ranking strategy. SLR denotes the original SLR model, and the re-SLR is a reimplemented version that uses ResNet as the image feature extractor from (Tanaka et al., 2019).

on RefCOCO, RefCOCO+, RefCOCOg datasets. To evaluate the quality of the expressions, we also adopt the METEOR and CIDEr automatic metrics commonly used in the field of image captioning. Our work confirms the effectiveness of our listener module in the following sections.

4.3.1 Quantitative Results

Evaluation on referring expression generation. We compare our generated expression with the recent models, including SLR (Yu et al., 2017), re-SLR (Tanaka et al., 2019) and RefGTA (Tanaka et al., 2019). We observed that using the reranking mechanism with the listener module generally improves the performance, although it was not quite helpful for the RefGTA model. For the well-generated expressions, the listener module contributed the most to enhancing the power of the model. In particular, the SR without the listener model performs much worse than using the listener as shown in the last three or four rows in Table 1. Since the reranking technique have a higher effect on our listener model compared to RefGTA, our model is able to outperform RefGTA on the comprehension evaluation as shown in Table 2. We found that our neighboring-based attention function helps to improve the performance of both speaker and listener module compared to the baseline of our model without attention. We analyze the effect that our proposed attention function results in that our model selects the meaningful neighboring objects to generate the referring expression as well as eliminating unnecessary neighbor objects that helps the listener model to focus on the target object by using the discrimination from surroundings.

Evaluation on referring expression comprehension. To find out the impact of each module for generation, we validate the performance of each speaker, listener, reinforcer module on comprehension evaluation. We compare following two models (Yu et al., 2017; Tanaka et al., 2019) based on speaker-listener-reinforcer for fair evaluation. We calculate the score of reinforcer, speaker by using ground truth bounding boxes for all the objects given r, $o^* = argmax_i F(r, o_i)$, $o^* = argmax_i P(r|o_i)$.

We report the expression comprehension results in Table 2. The listener module plays a crucial role

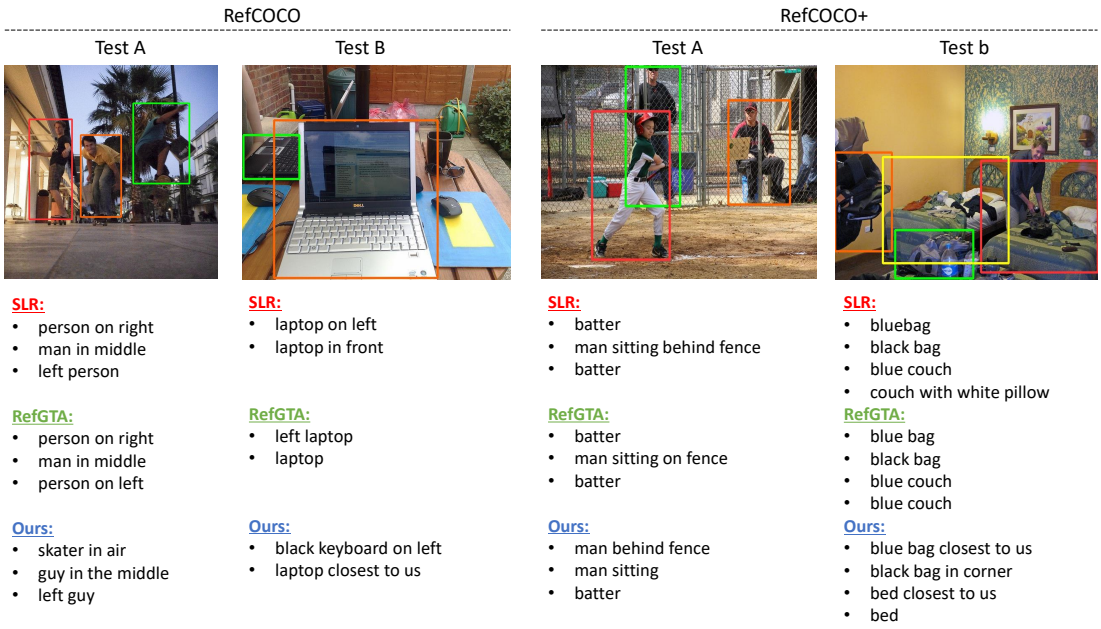

Figure 2: Qualitative comparisons for the generated referring expressions with (Yu et al., 2017), (Tanaka et al., 2019) on the RefCOCO, RefCOCO+. The order of expressions corresponds to green, orange, yellow red box, respectively.

in our model compared to others on improving the quality of expression. This is because our system additionally considers the neighbor objects' features for the target object. In particular, our speaker and listener module show better performance on the evaluation using the attended visual difference features compared to using simple visual difference features between target and neighbors. This is partly due to the scaled dot-product attention along with ranking strategy which not only reduces the complexity of the visual difference features to generate the referring expressions but also makes it easy to be referred back from the listener module.

As a result, our proposed method which considers the target's neighbors and performs the attention mechanism between target and visual difference features can improve the performance of the speaker and listener module by selecting the important context objects to identify the target surroundings itself.

4.3.2 Qualitative Results

In this section, we qualitatively analyze the tested data and results with comparison to SLR and RefGTA. Fig. 2 and Fig. 3 shows the generated sentences for each referring expression dataset: RefCOCO, RefCOCO+, and RefCOCOg. Particular objects are expressed with more detail as shown in Fig. 2 for RefCOCO dataset e.g. skater in air, black keyboard. Besides, some commonly mistaken objects are correctly spoken with additional descriptive location information for RefCOCO+ dataset e.g. man behind fence, bed closest to us. This indicates that our proposed method CoNAN has the potential to express the target object with a good description such as the location, attributes, and additional information covering the interaction with other objects. The consideration of the relationship with the target and neighbors along with base ideas effectively simplifies the listener's task of retrieving the object from the spoken expression without ambiguity, e.g. blue couch, the person on right.

Since RefCOCOg is known to contain longer and more complex expressions, the expectation of the performance boost with CoNAN is much more higher compared to other datasets. As the sentences were allowed to be long and complex, it is very important to contain as many details as possible. Fig. 3 shows our excellent and superior results compared to SLR and RefGTA. CoNAN generates a detailed expression for the baby along with the interaction information with other objects (e.g. holding a cell phone) for the first image. Also, CoNAN correctly generates the expression for the "arm behind" which

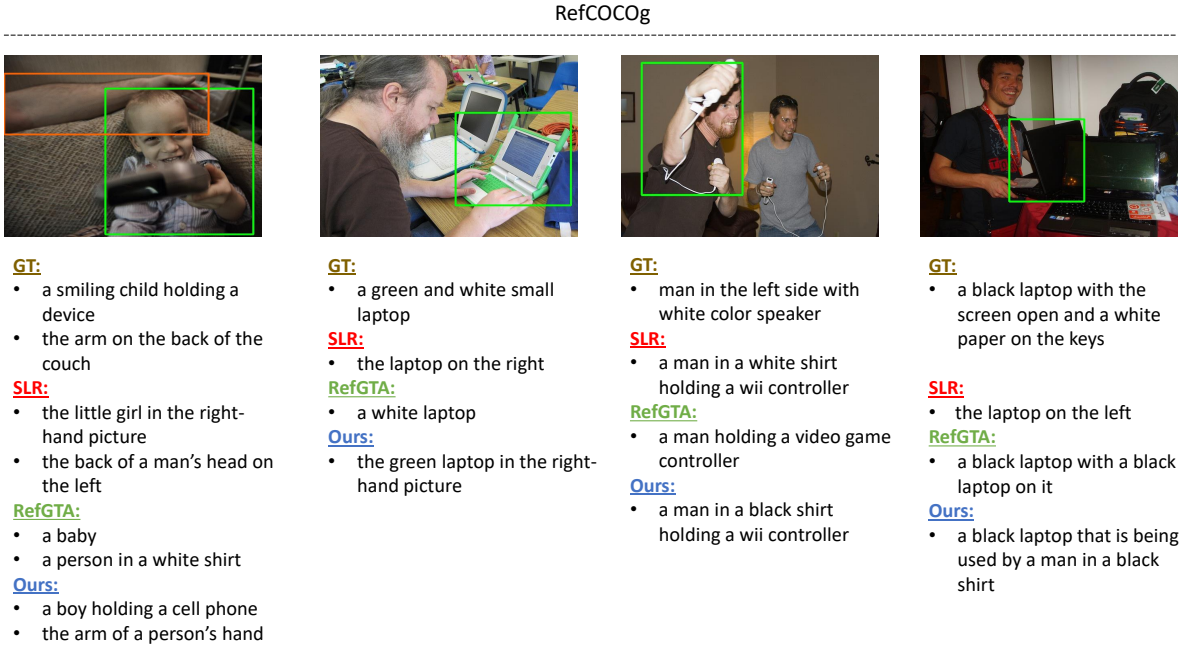

Figure 3: Qualitative comparisons of the generated referring expressions with (Yu et al., 2017), (Tanaka et al., 2019) and human annotation on RefCOCOg dataset. The order of expressions corresponds to green, orange, yellow red box, respectively.

are falsely assumed as a head or a person in a white shirt. Similarly, for the second and third image, CoNAN expresses the target object with far more details compared to other methods with having very low difference compared to the ground truth.

Interestingly, sometimes CoNAN can give out explanations that are far more clear compared to the ground truth as shown in the fourth image. While it is not clear to retrieve a black laptop with the screen open, it is more intuitive and easy to retrieve a black laptop which is being used by a man in a black shirt. This shows that taking the relationship with the neighbor object into account further helps the model to semantically understand the complex scene. The base generator and reinforcer are expected to have huge synergistic energy with additional object-level relation information, whereas, in the real world, humans tend to understand a given object along with the relationship to its surroundings.

5 Conclusion

In this work, we present an approach to explicitly mining complementary aspects for the target object in the local scene. In particular, the visual differences between the target and its neighbors are adopted. Instead of using all of the neighbors, we employ an attentional ranking module to filter out irrelevant neighbor objects. Finally, the speaker module is built upon the global features, target object features, and our complementary neighbor features to generate the expression. Our quantitative results show that CoNAN effectively enhances the performance for referring expression generation outperforming other state-of-the-art methods by a clear margin. Besides, our qualitative results state that CoNAN has the potential to give out the more descriptive expression for each target object sometimes even far superior to the ground truth.

Acknowledgement

We thank Taesan Kim, Kyungseo Min for constructive discussion and feedback on the draft.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Surabhi Gupta and Amanda Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, pages 1–6. Citeseer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Srinivasan Janarthanam and Oliver Lemon. 2010. Learning to adapt to unknown users: referring expression generation in spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 69–78. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. 2019. Generating easy-to-understand referring expressions for target identifications. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5794–5803.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7282–7290.