Referring Expression Generation and Comprehension

Bereket Frezgiy [*1] **Jing Wen** [*2] **Parth Shah** [*3] **Yansen Wang** [*2]

Abstract

Referring expression generation and comprehension is a challenging task where the objective is to generate unambiguous descriptions for referred objects and recognize the objects that the sentences referring to. Different from image captioning and image-sentence retrieval, this task particularly requires the model to have an unambiguous understanding of the referred object either by identifying the unique property or by capturing the relationship between different regions in an image. Based on this assumption, we propose the Adapted Speaker-Listener, Speaker-Listener-Discriminator, Co-ordinated Autoencoder. We conducted the experiments on RefCOCO, RefCOCO+ and RefCOCOg dataset and our methods show significant improvements in both generation and comprehension tasks. Further analysis on the results proves the ability of our models to distinguish the unique properties and capture the relationship information.

1. Introduction

Referring expressions, which aims at identifying particular objects within a scene in natural language, is rather an important and critical task. People use referring expressions such as *the man wearing red t-shirt* and *the black car on left* in everyday speech to explicitly point out the person or the object they are talking about. This task is closely related to the connection between images and descriptive language and can be considered as both comprehending and generating natural language referring to objects in an image.

A successful referring expression should be informative, succinct and unambiguous (Yu et al., 2016). Concluding from daily experiences, people mainly use two kinds of information to make a referring expression unambiguous:

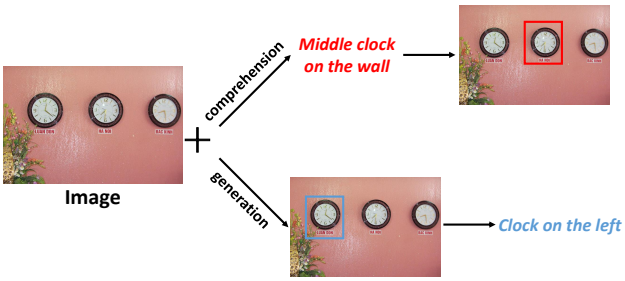

Figure 1. Two different subtasks in Refcoco. Spacial relationship is used to make the expression unambiguous.

- Pointing out the unique property or the referred object. People will use expressions like *the guy wearing red sweater* or *the green apple* to distinguish a person/object among a group when nobody/nothing else shares the feature.
- Describe the relationship between the referred object and other objects. When there is no obvious feature, people will use expressions like *the boy next to his mother* or *the sandwich in the basket* to make it clear who or what is being referred.

Referring expressions research has been tackled in the context of two subtasks that are dual to each other: **expression generation** which aims at generating unambiguous natural language expression for a proposed area in an image, and **expression comprehension** to retrieve object from an image referred by natural language. Fig. 1 shows an example for these two tasks. In this case, there are three clocks in the image that are almost the same except for the time they're pointing at. Therefore, *clock pointing at 12:20* will be a good expression but is hard for a model to capture. An easier way showed in the figure to avoid ambiguity is to use words that indicate relative positions (middle, left), which requires a model to understand the spacial relationship between the objects within an image.

While significant attempts have been made on individual subtasks, there are papers by Mao et al. (2016b) and Yu et al. (2017) that tries to solve the referring expression tasks in a joint manner. However, due to the difficulty of this task, even the current SOTA model (Yu et al., 2017)

[*]Equal contribution [1]Electrical and Computer Engineering Department, Carnegie Mellon University [2]School of Computer Science Department, Carnegie Mellon University [3]Heinz College, Carnegie Mellon University. Correspondence to: Bereket Frezgiy <bfrezgi@andrew.cmu.edu>.

indicates certain shortcomings with generating texts that is distinct, discriminative and less ambiguous expressions text sampling.

In this paper, we propose three different frameworks to address this problem. The adapted Speaker-Listener model is improved in the aspect of image representations, which automatically learns and encodes the relationships between the referred object and others. The Speaker-Listener-Discriminator model uses discriminators to add supervision based on the difference between positive and negative samples to help the speaker and listener module capture discriminative information in an adversarial fashion. We also tried to leverage Co-ordinated Auto-Encoder(Rumelhart et al., 1985) which learns a shared latent representation to reconstruct cross-modal information. Experiments on both generation and comprehension tasks proved the effectiveness of the first two methods. Further analysis on some cases showed that our model have gained the ability to capture unique properties of the referred objects and relationships between different regions.

We claim our contributions as follows:

- To assess the machinery ability of referring expression generation/comprehension, we conclude two possible angles from daily experience. One is to capture the unique features and the other is to find the relationship between regions.
- We propose several neural models to add relationship features using negative samples in the image and force model to generate unambiguous results by adversarial training. Experiments show that these two techniques are helpful to generate informative expressions and get better comprehension results.

2. Related Works

Referring expression problems involves two related sub-tasks: image captioning and grounding. While there is plenty of research in each, fewer works explore jointly training of two sub-tasks. We conclude the related works in three parts: image captioning, visual grounding, and joint referring expression generation and comprehension.

Image Captioning One of research area that is closely-related to our task is image captioning, as in the work from Karpathy & Fei-Fei (2017) and Vinyals et al. (2015), follows to encode an image using convolutional networks (CNN) then fed it as an input to recurrent networks (RNN). This CNN-LSTM architecture generates sequence of words conditioned on the visual features obtained from the image. Jia et al. (2015) uses additional extracted semantic information of the image as input to the LSTM. Xu et al. (2015) attends to the most relevant regions of the images, using attention mechanism improving the performance of the model to relate image features and text features.

Visual Grounding Another highly researched field that is related to our task is the grounding (natural language object retrieval tasks). The basics for grounding referring expression is to use the context of a text to distinguish the reference from other objects in an image. Guadarrama et al. (2014) leverages given regions of interest for objects in an image, generates texts for those regions as bag-of-words then compares the generated bag-of-words with the given text bag-of-words. Other research works uses holistic context such as the entire image (Hu et al., 2016; Mao et al., 2016b) or visual feature difference between regions(Yu et al., 2016; 2018).

Referring Expression Generation and Comprehension Image captioning generates sequence of words that best describe the whole image, which is different from our research of referring expression comprehension and generation. Referring expressions related literature have attracted attention mechanisms and related work after the release of the standard datasets(Kazemzadeh et al., 2014; Mao et al., 2016b). For the referring expression generation that is object grounding in an image conditioned on a text Mao et al.(2016b), Yu et al.(2016) and Nagaraja et al.(2016) models the probability of the region of interest features given an object then pinpoints an object that maximizes this probability. For the generation task Johnson et al.(2016) models the probability of a sentence given an image as the matching score. (Rohrbach et al., 2016) (2016) addresses the comprehension task as a classification problem by proposing multi-modal embedding. For combining both the tasks (Mao et al., 2016b) introduces Maximum Mutual Information (MMI), which is basically a constraint encouraging generated expression that describe the target object better than the other objects that are found in the image. The main idea is to model context regions, location/size and whole image features. The Speaker-Listener-Reinforcer (SLR) generalises the idea of MMI by incorporating two triplet loss composed of a positive match and negative match. On top of that the reinforcer helps the the speaker to produce discriminative sequence of words. During inference, the listener re-ranks the captions sampled from the speaker. The reinforcer performs a non-differentiable policy gradient update to the speaker which is an issue since the reinforcer has to wait for the speaker to finish the expression generation.
In all of the above papers visual and textual representations are not aligned or co-ordinated in a way to learn a hidden latent representation. Gu et al.(2018) proposed learning appropriate representation for the cross-modal visual-textual representation. Luo & Shakhnarovich(2017) proposed different approach that attempted Generative Adversarial Net-

work collaborative framework where the generator generates the expression and the discriminator directs the generator to improve the generation of expression.
Our paper will replace the reinforcer module with a adversarial framework as proposed in (Luo & Shakhnarovich, 2017), the reinforcer effect is not tangible to reduce the ambiguous natural language expressions. Section 4 explains in detail the modifications we made to the standard SLR to overcome the shortcomings of current state-of-the-art(SOTA).

3. Proposed Approaches

In this section, we present three approaches for generation and comprehension tasks.

3.1. Task Definitions

The referring expression generation task can be formulated as following: given an image X and several bounding boxes o_i in this image, the system should generate natural and meaningful expressions $\hat{r}_i$ which unambiguously describe the object within every bounding box, formally as:

$$\hat{r}_i = \underset{r_i}{argmax}\, \mathcal{P}(r_i|X_i, o_i). \tag{1}$$

Similarly, the comprehension task is to find bounding boxes $\hat{o}_i$ in the given image X corresponding to every expressions r_i as:

$$\hat{o}_i = \underset{o_i}{argmax}\, \mathcal{P}(o_i|X_i, r_i). \tag{2}$$

It is noteworthy that in most circumstances, there are more than one pairs of (r_i, o_i) in a given image X. Also, following (Yu et al., 2017), we leverage unmatched pairs (r_i, o_j) where $i \neq j$ during training to gain a better ability to generate discriminated results on the same image.

3.2. Adapted Speaker-Listener

3.2.1. Speaker-Listener

We start with the Speaker-Listener model which is presented by Yu et al.. The Speaker-Listener model has two part: the speaker module for generating the description and the listener module for joint embeddings between two modalities.

Speaker The Speaker takes the image features as inputs and generates the description for the referred object. To get the unambiguous representation for the referred object, the input image feature is the concatenation of five components: (1) The representation $\mathbf{v}_i$ for the referred region o_i; (2) The representation for the whole image $\mathbf{v}_I$; (3) The representation for difference $\mathbf{v}_{\text{diff}} = \frac{1}{n-1}\sum_{j\neq i, 1\leq j\leq n}(\mathbf{v}_i - \mathbf{v}_j)$, where $\mathbf{v}_j$ represents the feature of another region in the same image; (4) The bounding box of the referred object b_i; (5) The distance between the referred object and other objects in the same image b_{diff}. The image feature is then forwarded to the fully connected layer and LSTM, and generates the sentence.

The speaker is trained with some contrastive pairs, i.e., given a pair of positive match (r_i, o_i) for a image X, they also sampled the contrastive pairs (r_j, o_i) and (r_i, o_k) from the same image and define the generation loss as:

$$\begin{aligned}
L_1^s(\theta) &= -\sum_i \log \mathcal{P}(r_i|o_i;\theta),\\
L_2^s(\theta) &= \sum_i max(0, M + \log \mathcal{P}(r_i|o_k) - \log \mathcal{P}(r_i|o_i)),\\
L_3^s(\theta) &= \sum_i max(0, M + \log \mathcal{P}(r_j|o_i) - \log \mathcal{P}(r_i|o_i)),\\
L^s(\theta) &= \lambda_1^s L_1^s(\theta) + \lambda_2^s L_2^s(\theta) + \lambda_3^s L_3^s(\theta).
\end{aligned} \tag{3}$$

Here, L_1 is a generation loss provided by cross entropy with ground truth labels while L_2 and L_3 are hinge losses to from negative samples.

Listener Listener is trained to encode visual and text information in a joint space. The loss for the listener module is defined as:

$$\begin{aligned}
L_1^l(\theta) &= \sum_i max(0, M + \log \mathcal{P}(r_i|o_k) - \log \mathcal{P}(r_i|o_i)),\\
L_2^l(\theta) &= \sum_i max(0, M + \log \mathcal{P}(r_j|o_i) - \log \mathcal{P}(r_i|o_i)),\\
L^l(\theta) &= \lambda_1^l L_1^l(\theta) + \lambda_2^l L_2^l(\theta).
\end{aligned} \tag{4}$$

3.2.2. Model Overview

The Speaker-Listener (SL) has two main shortcomings. On one hand, it manually constructs the image feature to represent the difference between the referred object and others. The feature, which is the subtraction of the features of the referred object and the other regions, can hardly capture the differences between properties and categories. On the other hand, features for all regions should pass through the model one by one when grounding, which makes it less efficient.

As shown in 2, the adapted Speaker-Listener (adapted SL) improves the image encoder in the Speaker. Instead a single feature vector for the referred object, the adapted model takes the representations for all regions in the same image as inputs, which is $\mathbf{v} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. The features go through the transformer based image encoder, where the contexts (relations, differences, etc.) are automatically encoded into the output feature for each image region. We define $\mathbf{v}_i^o$ as the output feature vector for region o_i from the image encoder.

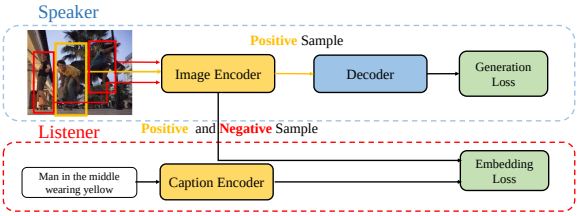

Figure 2. The adapted Speaker-Listener model. The features for all regions are generated from single pass while both the positive and negative samples can be sampled from the image encoder.

During training, both the Speaker and Listener require to sample positive and negative region-description pairs. With the adapted model, we can directly sample from the output of image features.

3.3. Speaker-Listener-Discriminator

3.3.1. Model Description

An extension of SLR model (Yu et al., 2017) where the reinforcer module is replaced by discriminator is presented in fig. 3. One of the flaws of the reinforcer is the fact it is trained in an end-to-end way together with other modules, it is hardly accurate to judge whether an expression is good (less ambiguous) and is unable to give a very helpful reward.

After (Goodfellow et al., 2014) (2014) proposed adversarial train routine, community have seen a massive development in applying adversarial training and Generative Adversarial Networks (GANs) to different generation tasks. As a typical framework of GAN, the **generator** and the **discriminator** are trained in an adversarial way to improve both performances. The function of the discriminator is very similar to the reinforcer but is explicitly trained with golden truths.

On the other hand, since we want to capture the discriminative features of referred objects, a well-designed discriminator(Wang et al., 2018) will not only be able to make discrimination between ground-truth labels and generated ones, but also can figure out whether a label is unique enough to describe the referred object by discriminate between positive samples and negative samples, which will help the generator focus mostly on informative features.

We thus devise the Speaker-Listener-Discriminator model (SLD), which replaces the reinforcer module with two discriminator modules to better allocate the reward loss. As a modification of the original discriminator which is essentially a 2-category classifier between generated expressions and ground truth, we want to add a 3rd category to bring in negative samples provided by expressions and embeddings provided by other bounding boxes from the same picture, just as (Wang et al., 2018) does in the visual storytelling task, to help the speaker capture unique properties.

The SLD model consists of four parts: Speaker, Listener, Generation Discriminator and Embedding Discriminator, which is shown in fig. 3.

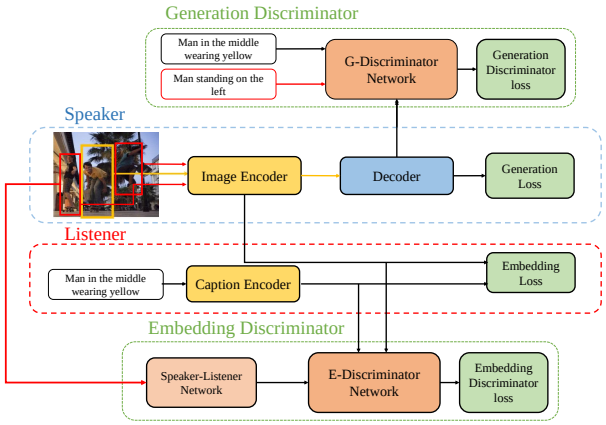

Figure 3. The Speaker-Listener-Discriminator model. Bounding box and arrows in red illustrate how the negative sample from other object/expressions in the same picture can be used to train the two discriminators.

Speaker-Listener The Speaker module and the Listener module are essentially the same as that in the SLR model by Yu et al.. However, here they serve as two generators in GAN and are trained end-to-end with a small caveat that we pre-train Speaker and Listener on a supervised maximum likelihood loss. This method is fairly common and is incorporated by Yang et al., 2017; Wu et al., 2017;Li et al., 2017; Zhang et al., 2016 and many more for generating text sequence. To keep our model updated, we substitute the encoder in the listener and the decoder in the speaker with two transformers.

Generation Discriminator In order to estimate the quality of the generated expressions, we devised the Generation Discriminator D^g to classify an expression $r_{candidate}$ into three category $\{paired, unpaired, generated\}$, respectively standing for the ground truth expression, other expressions in the same image, and generated images. This part will provide a generation discriminator loss L^{gd} based on the probability for a generated sentence y_i to be classified to the $paired$ category, i.e. $P(paired|(\mathbf{x_i}, \mathbf{y_i}))$.

Embedding Discriminator Similar to the Generation Discriminator, the Embedding Discriminator is devised to classify the input-embedding pairs $\mathbf{e}_{candidate}$ into $\{paired, unpaired, generated\}$. This part will generate an embedding discriminator loss L^{ed} based on $P(paired|(\mathbf{x_i}, \mathbf{e_i}))$.

3.3.2. Adversarial Training

We train our model in a typical adversarial training method. The Speaker-Listener is pretrained in an end-to-end way for a few epochs without the two discriminators, and the loss function in this stage only contains the generation loss and embedding loss, as follows,

$$L_{pre} = L^s + L^l.$$

Then we pretrained the discriminator with fixed generator with losses as follows,

$$L_{gd} = \sum -\log \mathcal{P}(c_r|X, o, r),$$
$$L_{ed} = \sum -\log \mathcal{P}(c_e|X, o, r),$$

where c_r and c_e are the right category of a expression r and embedding e.

After that, we train the generator and discriminator simultaneously. At this stage, the loss function contains the term from discriminators.

$$L_s^{gd} = \sum -\log \mathcal{P}_g(paired|X, o, r),$$
$$L_l^{ed} = \sum -\log \mathcal{P}_e(paired|X, o, r),$$
$$L = L^s + L^l + \lambda^{gd} L_s^{gd} + \lambda^{ed} L_l^{ed},$$

where the possibility $\mathcal{P}_g$ and $\mathcal{P}_e$ are given by the generation discriminator and the embedding discriminator respectively.

To make it differentiable, we use the gumbel-softmax technique(Jang et al., 2016) to sample the predicted sentences by the speaker.

3.4. Co-ordinated AutoEncoder

3.4.1. Model Description

The task of both referring expression generation and comprehension can be recognized as multimodal translation problem, i.e. extracting information from one modality and present it in another modality. In this translation process, what changes is how the information is presented explicitly rather than the implicit information itself. Since this information is shared by both modality, a well-trained model should be able to get this information from either modality, and reconstruct the other modality based on it.

We thus devised the coordinated AutoEncoder to capture this kind of information and get a coordinated representation of it. In this model, two modality-specific encoders are trained to generate a similar vector representation z_i for a box-expression pair (o_i, r_i) in the image X. Two decoders are applied to reconstruct the results of different modalities. When doing inference, only the encoder for the given modality will be used to get the representation and used by the decoder from the other modality to generate the result.

3.4.2. Model Overview

The coordinated auto-encoder model consists of two pairs of encoder and decoder for language and visual modality respectively, as shown in fig. 4.

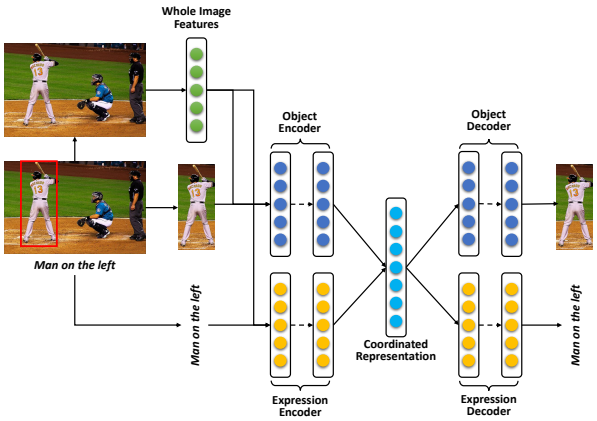

Figure 4. Coordinated AutoEncoder

During training, the expression encoder tries to get the representation from the language modality via:

$$z_i^{(l)} = s_m^{(l)},$$
$$s_t^{(l)} = f_t^{(l)}(s_{t-1}^{(l)}),$$
$$s_0^{(l)} = [X : r_i],$$

where $f_t^{(l)}(\cdot)$ is the t-th trainable linear layer with proper activation function, $[\cdot : \cdot]$ represents for the concatenation of two vectors(matrices, tensors).

Then the expression decoder is applied to the decode a expression via:

$$\hat{r}_i = Decoder^{(l)}(h_m^{(l)}),$$
$$h_t^{(l)} = g_t^{(l)}(g_{t-1}^{(l)}),$$
$$h_0^{(l)} = z_i^{(l)}.$$

The decoder for the language modality is a vanilla decoder based on Gated Recurrent Unit(GRU) layer. In the same way, a representation $z_i^{(v)}$ was generated by the object encoder and used to reconstruct $\hat{o}_i$ by object decoder.

In this way, the loss function will be a combination of two reconstruction losses and a similarity loss between $z_i^{(l)}$ and

$z_i^{(v)}$:

$$L_z = \sum_i distance(z_i^{(l)}, z_i^{(v)}), \tag{5}$$
$$L_l = \sum_i -\log \mathcal{P}(r_i|X, r_i),$$
$$L_v = \sum_i -\log \mathcal{P}(o_i|X, o_i),$$
$$L = \lambda_z L_z + \lambda_v L_v + \lambda_l L_l.$$

In eq. 5, the distance is measured by as simple as L2-norm, or as complex as another model mentioned in (Zhang et al., 2019).

When doing inference on the referring expression generation task, we only get X and o_i as inputs. We then generate the visual representation $z_i^{(v)}$ with the object encoder to decode it with a expression decoder. Similarly, when doing the comprehension task, we used the expression encoder and object decoder to get the comprehension result.

4. Experiments

4.1. Datasets

Over the years, researchers of referring expression generation leveraged on the datasets that are extensions to the available datasets on the task of image captioning. For example, Flickr30K dataset (Plummer et al., 2017; Young et al., 2014) has seen some benchmark results for generating expressions given an image and bounding box. As for the comprehension task, dataset like Visual Gnome (Krishna et al., 2016) has also been made public based on the MSCOCO(Lin et al., 2014) dataset specifically for grounding objective.

However, most researches addressing the joint task, i.e. generation and comprehension in the same time, utilized RefCOCO, RefCOCO+, and RefCOCOg datasets (Kazemzadeh et al., 2014). These datasets are also extensions of the MSCOCO2014 (Lin et al., 2014) train data. RefCOCO and RefCOCO+ were generated in an interactive setting where Amazon Mechanical Turk (AMT) workers interacted on annotating given the Region of Interest (RoI) and coming up with the RoIs given the annotation. Expressions in these two datasets are mostly short descriptions for the object. RefCOCOg dataset was developed in a non-interactive setting for the work in (Mao et al., 2016b) by AMT workers generating expressions for given RoIs, resulting in more complex expressions. Fig. 5 (Yu et al., 2016) represents expressions generated on the same image by different RefCOCO datasets and the quantitative information of each dataset is shown in Table 2.

We followed some conventional splits for these three datasets. For Refcoco and Refcoco+ datasets, we're using the *unc* split which divides the datasets into train, validation, test A and test B. Referred objects in Test A set are only the human objects while in Test B they're all the other objects. And we're using the *google* split for Refcocog which only divides the datasets into train and validation sets. Table 1 gives the information about the splits for all the three datasets.

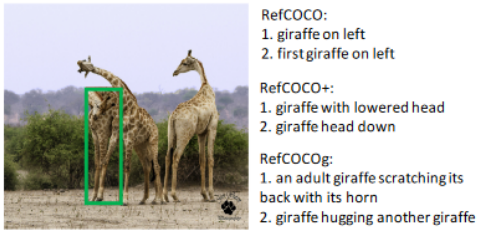

Figure 5. Expression generated for the outlined giraffe from 3 datasets (Yu et al., 2016).

4.2. Evaluation Metrics

Comprehension: The evaluation of referring expression comprehension is similar to that of visual grounding: given a set of regions, pick out the ones which are corresponding to the expression. In our experiments, The set of regions for an image is made up by all the possible bounding boxes contained in the dataset. Under this experiment setting, accuracy@n / recall@n / precision@n(Mao et al., 2016a) are calculated, in which @n refers to the accuracy / recall / precision among top-n predictions. Practically, we set $n = 1$ following the convention by previous works and only accuracy score needs calculating.

It's noteworthy that in real applications, the set of regions is not necessarily required. Randomly generated regions of the given image as the set will enable the model to do reranking.

Generation: For the generation task, we're using two widely-used method in image captioning to judge the model: Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee & Lavie, 2005) and Consensus-based Image Description Evaluation (CIDEr) (Vedantam et al., 2015) score. METEOR score is simply a Harmonic Mean (HM) of precision and recall of the unigrams of generated expression to the reference expression unigrams, with recall weighted 9 times more than the precision. CIDEr is an extension to METEOR where we look at the weighted HM and are not restricting the evaluation to just unigrams but extending to n-grams.

	RefCOCO				RefCOCO+				RefCOCOg	
	Train	Val	Test A	Test B	Train	Val	Test A	Test B	Val	Train
Images	16,994	1,500	750	750	16,992	1,500	750	750	4,650	24,698
Expressions	120,624	10,834	5,657	5,095	120,191	10,758	5,726	4,889	9,536	85,474
Bounding Boxes	42,404	3,811	1,975	1,810	42,244	3,805	1,975	1,798	5,000	44,822

Table 1. Split information for all datasets.

Dataset	#Expressions	#Bounding Boxes	#Images
RefCOCO	142,210	50,000	19,994
RefCOCO+	141,564	49,856	19,992
RefCOCOg	95,010	49,822	25,799

Table 2. Quantitative values for datasets.

Figure 6. Error Analysis for SLR.

4.3. Baselines and Experiment Settings

We compared our model to the SOTA baseline SLR(Yu et al., 2017) which can do the referring expression generation and comprehension tasks jointly. To make the result comparable, we reran their model in the same model/evaluation framework so the results reported are not exactly the same as those in the original paper.

We used ResNet-152 (He et al.,2016) pretrained on ImageNet (Deng et al.,2009) to encode the whole image and the bounding box. This part was fixed except the last layer of ResNet-152. The SLR baseline used a 1-layer LSTM to encode the sentence in listener and another 1-layer LSTM in speaker to decode the sentence. The Adapted SL and SLD were using 2-layer transformer encoder (Vaswani et al., 2017) to encode sentences and another 2-layer transformer decoder is used to generate the sentence. Also, the Adapted SL used a 2-layer transformer encoder to encode the features from all the regions of an image.

When doing the generation task, we used beam search with beam width 2 for every model and chose sentence with higher generation probability as the predicted sentence.

4.4. Results

We present the results for the generation and comprehension tasks in Table 3 and Table 4.

Generation: Adapted SL and SLD outperformed the SOTA baseline significantly on most of datasets and evaluation metrics. As for the Adapted SL model, incorporating information from all the negative regions helps the model capture the relationship between the proposed region and other parts of the image, while for the SLD model, the discriminators adding supervision on the difference between positive and negative samples help the speak capture the unique property of the bounding box. Both methods are proved to be effective: Adapted SL got better scores on Refcoco+ and better METEOR scores on Refcoco, and SLD achieved the best on Refcocog and for CIDEr on Refcoco.

However, the results from the Autoencoder seem terrible, indicating that it's not a good idea to force two essentially different embedding spaces (image and language) to be coordinated in a hard way.

Comprehension[1]**:** For comprehension we use both speaker and listener by assembling them together and picking the most probable object given a referring expression. Datasets has pre-defined regions where we pick the region for which the expression has the highest probablity. As the number of regions are limited from table 4, even AutoEncoder gives reasonable outputs.

4.5. Analysis

We present the analysis of how the proposed SLD performs when compared to the SLR baseline. We analyse images where we have multiple count of similar objects. Thus we check for the ambiguity of the generated expression.

4.5.1. Error Analysis for SLR

As seen on the left of fig. 6 the ground truth expression focuses on Woman in the middle. Our generated expression *"White Shirt"* leads us to two sort of error.

- If more than one object of similar category is present then it becomes difficult for the model to correctly predict the expression. One reason could be model's

[1]We haven't got the results from SLD but we'll report it in the conference version.

	RefCOCO				RefCOCO+				RefCOCOg	
	Test A		Test B		Test A		Test B		Val	
	MET	CIDEr	MET	CIDEr	MET	CIDEr	MET	CIDEr	MET	CIDEr
SLR	0.187	0.773	0.252	1.323	0.145	0.618	0.144	0.709	0.154	0.754
Autoencoder	0.133	0.251	0.151	0.434	0.096	0.206	0.078	0.250	0.010	0.252
Adapted SL	**0.203**	0.834	**0.261**	1.385	**0.165**	**0.648**	0.155	**0.781**	0.153	0.716
SLD	0.199	**0.837**	0.257	**1.401**	0.160	0.641	**0.157**	0.755	**0.160**	**0.758**

Table 3. Generation results for all ideas mentioned. All results are generated for beam width 2. Ablation studies showed that SLR baseline had a similar performance even without the non-differential policy gradient.

	RefCOCO		RefCOCO+		RefCOCOg
	Test A	Test B	Test A	Test B	Val
SLR	78.95	80.22	64.60	59.62	72.63
AutoEncoder	78.1	78.4	65.1	57.1	71.2
Adapted SL	**81.05**	**80.22**	**68.6**	**60.1**	**73.2**

Table 4. Comprehension results for SLR, Adapted SL and AutoEncoder.

inability to detect multiple objects from the same image.

- Even if the speaker is attending the wrong object of the same category it focuses on inessential objects (in our case t-shirt) rather than the entire bounding box.

We can see some similar sort of results for Test B example.

4.5.2. Comparing SLR, Adapted SL and SLD

We check on several images where ambiguity is possible in fig.7.

- The qualitative analysis shows that the adapted SL model tends to generate the sentences with more semantic information including the categories and properties of the referred objects. For example, in the first image of fig. 7, the adapted SL model generates the *"man in white shirt"* instead of *"woman"* as the SLR model does. Also, the adapted SL model can encode the spatial relationships between objects especially in complex scenes. In the left-bottom example in fig. 7, the adapted SL model is able to encode the spatial relationships between the motorcycle behind and other motorcycles which results in unambiguous description.
- SLD successfully distinguish the referred object and others, especially when these objects have shared properties. In the second image of the first line in fig. 7, the SLD can tell the difference between adults and children and generate the unambiguous description. In the upright example in fig. 7, the SLD is able to distinguish the standing man with the sitting man and capture the major difference between two people while SLR model only captures the category.

Figure 7. Comparing SLR, Adapted SL and SLD expressions.

5. Conclusions and Future Works

In this paper, we emphasized the importance to use unique features and relationship between objects in the referring expression generation and comprehension task. Based on it, we propose three new methods, i.e. Adapted SL, SLD and Co-ordinated Auto-Encoder. Both quantitative experiments and case studies have shown their ability to generate informative and unambiguous results.

In the future, we'll try to merge Adapted SL and SLD into a model to benefit from both unique features and relationship and try to get better results. Also we'll try to test our model on other datasets to see whether doing the joint task will help get improvement on individual tasks.

References

Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gu, J., Cai, J., Joty, S. R., Niu, L., and Wang, G. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7181–7189, 2018.

Guadarrama, S., Rodner, E., Saenko, K., Zhang, N., Farrell, R., Donahue, J., and Darrell, T. Open-vocabulary object retrieval. In *Robotics: science and systems*, volume 2, pp. 6, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., and Darrell, T. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4555–4564, 2016.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2407–2415, 2015.

Johnson, J., Karpathy, A., and Fei-Fei, L. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4565–4574, 2016.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, Apr 2017. ISSN 2160-9292. doi: 10.1109/tpami.2016.2598339. URL http://dx.doi.org/10.1109/TPAMI.2016.2598339.

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalanditis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

Li, J., Monroe, W., Shi, T., Ritter, A., and Jurafsky, D. Adversarial learning for neural dialogue generation. *CoRR*, abs/1701.06547, 2017. URL http://arxiv.org/abs/1701.06547.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Luo, R. and Shakhnarovich, G. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7102–7111, 2017.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., and Murphy, K. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016a.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. Generation and comprehension of unambiguous object descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016b.

Nagaraja, V. K., Morariu, V. I., and Davis, L. S. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pp. 792–807. Springer, 2016.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., and Schiele, B. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pp. 817–834. Springer, 2016.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Wang, J., Fu, J., Tang, J., Li, Z., and Mei, T. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Wu, L., Xia, Y., Zhao, L., Tian, F., Qin, T., Lai, J., and Liu, T. Adversarial neural machine translation. *CoRR*, abs/1704.06933, 2017. URL http://arxiv.org/abs/1704.06933.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.

Yang, Z., Chen, W., Wang, F., and Xu, B. Improving neural machine translation with conditional sequence generative adversarial nets. *CoRR*, abs/1703.04887, 2017. URL http://arxiv.org/abs/1703.04887.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85. Springer, 2016.

Yu, L., Tan, H., Bansal, M., and Berg, T. L. A joint speaker-listener-reinforcer model for referring expressions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., and Berg, T. L. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315, 2018.

Zhang, C., Liu, Y., and Fu, H. Ae2-nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2577–2585, 2019.

Zhang, Y., Gan, Z., and Carin, L. Generating text via adversarial training. In *NIPS workshop on Adversarial Training*, volume 21, 2016.