Generating Coherent and Informative Descriptions for Groups of Visual Objects and Categories: A Simple Decoding Approach

Nazia Attari
Research Institute for Cognition and Robotics
Bielefeld University, Germany
nattari@techkfak.uni-bielefeld.de

David Schlangen
Computational Linguistics
University of Potsdam, Germany

Martin Heckmann
Aalen University, Germany

Heiko Wersing
Honda Research Institute Europe, Germany

Sina Zarrieß
Faculty for Linguistics and Literature Studies
Bielefeld University, Germany

Abstract

State-of-the-art image captioning models achieve very good performance in generating descriptions for instances of visual categories and reasoning about them, e.g. imposing distinctiveness of the description in the context of distractors. In this work, we propose an inference mechanism that extends an instance-level captioning model to generate coherent and informative descriptions for groups of visual objects from the same or different categories. We test our model in the domain of bird descriptions. We show that group-level descriptions generated by our method are (i) coherent, pulling together properties that are true for all or majority of its instances, and (ii) informative, as they allow an external BERT-based text classifier to identify the target category more accurately in comparison to single-instance captions and are preferred by human evaluators.

1 Introduction

State-of-the-art image captioning models excel at generating semantically accurate descriptions of single images (Anderson et al., 2018; Cornia et al., 2020) and can be enhanced with communicative-pragmatic reasoning procedures that impose distinctiveness of the description in the context of distractors at inference time (Vedantam et al., 2017; Cohn-Gordon et al., 2018; Zarrieß and Schlangen, 2019). To date, however, discriminative image captioning has been restricted to informative *instance* descriptions and has not yet explored descriptions for *groups* (or sets) of objects – a classical problem in referring expression generation (REG) (Stone, 2000; Gardent, 2002; Horacek, 2004; Gatt, 2007; Krahmer and van Deemter, 2011). In this paper, we investigate whether an instance-level captioning model can be extended to generate coherent and informative descriptions for groups of visual instances, by integrating communicative-pragmatic reasoning at inference time.

Generating a description for a group of visual entities require optimizing two objectives: (i) coherence, i.e., the description should pull together properties that are true for all or most of the groups' instances and (ii) informativeness, i.e., it should mention those properties that are distinctive in a particular context (Gatt, 2007). Krahmer and van Deemter (2011) point out that the traditional Incremental Algorithm for symbolic REG directly applies to sets of entities, when they have certain properties in common. In this paper, we test whether this also holds for neural captioning models and propose a simple task, an inference scheme and experimental protocol for generating group-level descriptions. In particular, we extend the emitter-surpressor beam search by Vedantam et al. (2017) with an additional, simple coherence objective.

The ability to generate descriptions of groups is not only relevant for reference but also for explanation tasks, which become increasingly important in machine learning (Ribeiro et al., 2016; Lundberg and Lee, 2017). Here, systems commonly need to verbalize their knowledge about the shared properties of instances in a category, for instance, when learning to classify birds in images (Hendricks et al., 2016). However, an instance-based explanation might produce a rather idiosyncratic description of an image rather than a more representative description of the categories (that is true for majority of instances in a set). This becomes cru-

Group with a shared attribute (different categories)

(G) this is a bird with *yellow breast*

(a) this bird has a white belly with a short pointy bill

(b) this bird has a white belly with a short pointy bill

(c) this bird has a yellow belly and breast with a black cheek patch

Group with the same category

(G) this is a *blue bird* with a *short pointy bill* and black wings

(d) this a blue bird with brown wing-bars

(e) this is a blue bird with a very short beak.

(f) this bird has a speckled belly and breast with short pointy bill

Category group (left) with distractor group (right)

(G) a small bird with a *red crown*.

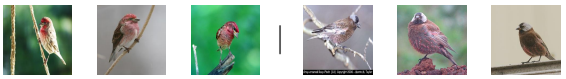


Figure 1: Examples of generated group (G) and instance (a-f) descriptions for types of bird groups.

cial in scenarios where a system needs to describe to a user the difference between two categories of birds. Here, it does not suffice to characterize only a single-instance, but, ideally, the system should preferably have a linguistic component explaining its knowledge about the category (Figure 1: middle section). Thus, for our study, we use the Caltech UCSD birds data (Wah et al., 2011) that provides fine-grained categories for bird images and descriptions of instances (Reed et al., 2016), and has been leveraged for instance-level explanation generation by Hendricks et al. (2016). In the context of captioning images of birds, we show that our approach to group-level decoding can be used for different types of groups and corresponding descriptions: (i) **objects with a shared attribute** but from different categories (i.e. bird species) and (ii) **objects of the same category**, sharing multiple visual properties, as shown in Figure 1. We assess the quality, coherence and informativeness of these group descriptions in human and automatic evaluation, including a set up for category prediction based on generated descriptions.

2 Related work

Research on language generation from visual inputs often builds upon generic image captioning models that are trained to produce "neutral" descriptions for images depicting instances of objects or scenes (Vinyals et al., 2015; Xu et al., 2015; You et al., 2016; Rennie et al., 2017; Anderson et al., 2018; Hossain et al., 2019; Cornia et al., 2020; Zhou et al., 2020; Luo et al., 2021). One line of work has extended captioning towards more complex visual inputs, e.g., sequences of images depicting events or stories (Yu et al., 2017; Mun et al., 2019; Gao et al., 2020). Other work has looked at enhancing captioning models towards generating more informative outputs that fulfill specific communicative goals, by leveraging contextual and contrasting information along with the target image at inference time (Vedantam et al., 2017; Cohn-Gordon et al., 2018; Zarrieß and Schlangen, 2019; Nie et al., 2020). Our work connects these two lines by extending Vedantam et al. (2017)'s discriminative instance-level decoding scheme for groups of image instances.

Our task and set-up is similar to Li et al. (2020)'s work on context-aware group captioning, where the goal is to build a model that captions a group of images with a matching scene graph (e.g. *women in chair*) in the context of a more general reference set of images (e.g. *women*). Their approach rests on a supervised model that is trained on a dataset of group captions (compiled from instance captions) and that performs group-wise visual feature aggregation with self-attention and contrastive visual feature construction. While Li et al. (2020) investigate rather short group descriptions for common objects (e.g., *women with hat*) with an average caption length of around 3, we test our approach on bird descriptions which involves a careful selection of properties for informative and coherent descriptions that have an average length greater than 10. Moreover, our work aims at describing groups by reasoning at the word level about which words can be used to refer to the group's instances, without retraining the underlying captioning model.

In comparison to earlier work on REG for sets, though, our approach targets rather simple descriptions of groups that essentially mention the properties that hold for the members of the set. Thus, we do not address more complex linguistic phenomena such as plurals, coordination, disjunction, or quantification. Gatt (2007), for instance, investigates conceptual coherence for the generation of sets whose entities cannot be referred to by the same head noun, triggering a competition between coordinations like *the chef and the engineer* and

the Italian and the Frenchman. As our approach assumes that group descriptions can be decoded from an instance-level captioning model, it will not be able to generate linguistic structures that do not appear in the training data of that captioning model. For instance, phenomena like coordination do not appear in our descriptions which typically enumerate properties of a single bird, named *bird*, see Figure 1. This is the case for our model as it uses bird description data where all captions refer to a single bird, which is named *bird*, see Figure 1.

3 Approach

This section defines the task and decoding procedures for coherent and informative group-level captioning.

3.1 Task Description

We assume to be given a dataset that pairs image instances i with verbal descriptions s and some category information c. We also assume that a captioning model of some sort, which we refer to as speaker $S(I)$, is trained on this data and predicts the probability of sequences of words given a single image *p(s|i)*.

We frame the task of generating group descriptions as a decoding or inference task, where the input to the model is a target group of n instances, $G_t = \{i_1, i_2, \ldots, i_n\}$ and the goal is to predict *p(s|G_t)* based on the speaker $S(I)$, without any further training or fine-tuning of the instance-level captioning.

This basic group description task can be extended towards a discriminative description task where the model receives an additional context, i.e. a distractor group of $G_d = \{i_1, i_2, \ldots, i_m\}$. In discriminative group description decoding, the goal is to predict a pragmatically informative sequence of words s such that a listener can distinguish the target from the distractor group or, more formally, such that $p(G_t|s) > p(G_d|s)$.

3.2 Coherent Group Decoding

The objective of the basic group-level speaker $S(G_\mathrm{t})$ is to maximize the probability of the output sequence given all images in the target group:

$$S(G_\mathrm{t}) = \underset{s}{\operatorname{argmax}} \frac{1}{n} \sum_{l=1}^{n} \log p(s|i_l) \qquad (1)$$

As the space over possible output sequences s cannot be searched exhaustively, we approximate this objective via beam search: at every time step, we (i) input all instances of the group to speaker $S(I)$ in parallel, (ii) compute the mean of log-probabilities over the entire vocabulary of all instances of the group and (iii) put the top-k words on the beam, as input to the next time-step. The stepwise averaging over log word probabilities directly implements the idea of coherence, i.e. the model should verbalize the common properties that are likely for all instances in the group.

3.3 Discriminative Group Decoding

We expect that $S(G_t)$ produces descriptions that summarize common properties of a group, but that it may not always select particularly informative properties that accurately discriminate a group in context. Thus, we define the discriminative group speaker $S(G_d)$ for instances in the distractor group, with the following objective:

$$S(G_\mathrm{d}) = \underset{s}{\operatorname{argmax}} \frac{1}{m} \sum_{k=1}^{m} \log p(s|i_k) \qquad (2)$$

We use $S(G_d)$ to induce discriminativeness of the output by combining it with $S(G_t)$ and reconstructing the emitter-suppressor beam objective by Vedantam et al. (2017) for groups:

$$S(G_\mathrm{t,d}) = S(G_\mathrm{t}) - (1-\lambda) \cdot S(G_\mathrm{d}) \qquad (3)$$

$S(G_{t,d})$ is the group speaker that maintains a trade-off between coherence and informativeness of the generated sequences, and can be pushed towards higher discriminativeness with appropriate values of the λ parameter.

The speakers in Equation 3 can be further factorized, incorporating word probabilities for the sequence as $\prod_{\tau=1}^{T} p(s_\tau|s_{1:\tau-1}, I)$, where T is the length of the sentence. Hence, we obtain the following objective for our inference mechanism:

$$S(G_\mathrm{t,d}) = \underset{s}{\operatorname{argmax}} \sum_{\tau=1}^{T} \frac{1}{n} \sum_{l=1}^{n} \log p(s_\tau|s_{1:\tau-1}, i_l) - (1-\lambda) \cdot \left(\frac{1}{m} \sum_{k=1}^{m} \log p(s_\tau|s_{1:\tau-1}, i_k)\right) \qquad (4)$$

Again, we approximate this objective via beam search. At every time-step, we subtract the average

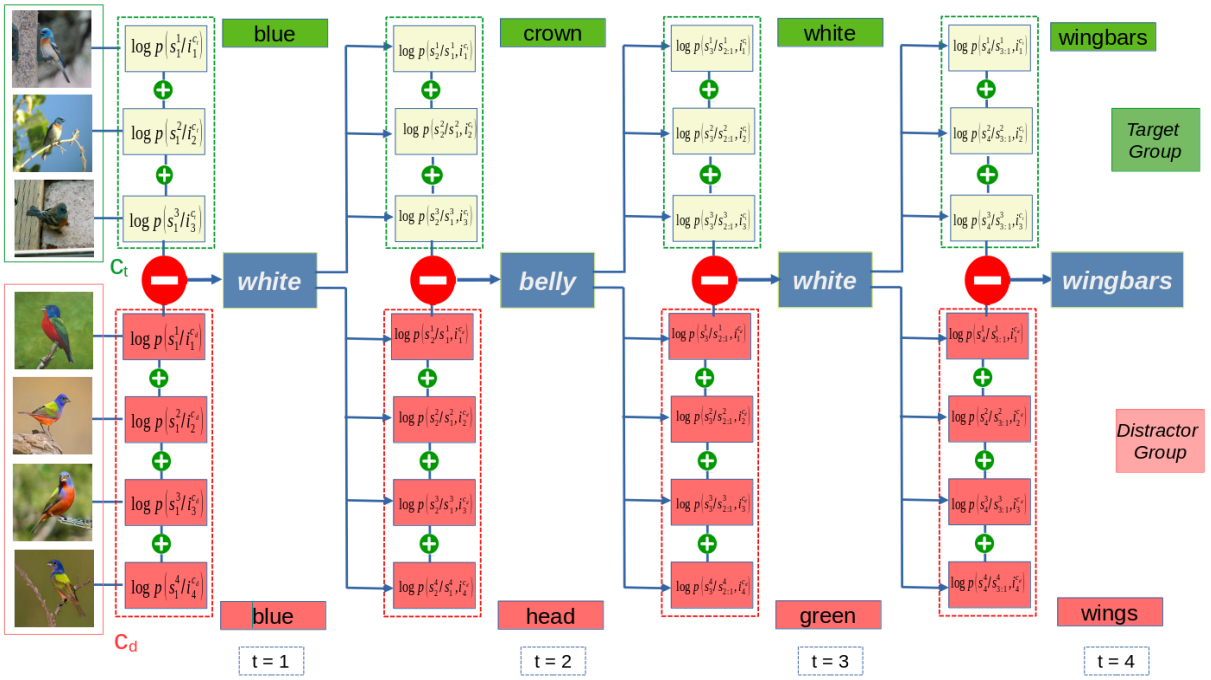


Figure 2: Illustration of an example phrase generated by discriminative group-level decoding with beam size 1 (*white belly white wingbars* in blue boxes). The decoding scheme favours coherent and discriminative properties over less discriminative ones predicted by (target only) group-level decoding (*blue crown* in top green boxes)

log probability of a word for the target instances by its log probability for the distractor instances. Words that have high probability for the in-group images and low probability for the out-group images will be more likely to be put on the beam than words that are equally probable for both, or even more probable for the out-group.

We demonstrate the mechanics of our decoding procedure in Figure 2, with a beam size of 1 for simplicity. It shows how the speaker S, at inference time, combines probability scores of the respective groups and produces the best possible output words. In our experiments, we used a beam size of 10.

4 Experimental Set-up

4.1 Data

We base our work on the CUB-200-2011 dataset (Wah et al., 2011), originally designed for subordinate category categorization, detection and part localization. It contains 11788 images of 200 North American bird species and every species has approximately 60 image instances. Each image instance is characterized by 28 symbolic attributes using an online tool for bird identification[1] curated by bird experts, further leading to an extensive set of human-annotated 312 binary attribute-value pairs (e.g. *beak-shape:hooked*, *belly-color:white*, *tail-pattern:spotted*). For our first experiment, we used this symbolic information to form groups of image instances from different bird categories.

We also have access to (five) textual descriptions for each image instance collected by Reed et al. (2016); the annotators were asked to mention the physical bird attributes (wing color, beak shape, body color and so on) visible in the image without any reference to the bird species and using basic vocabulary unlike sophisticated expert-level vocabulary. We note, however, that in some cases, the annotators also mentioned non-discriminating properties, for instance, where the bird is looking at, it's flying or sitting.

4.2 Sampling Groups

The target groups (G_t) in our experiments can be of two kinds: (i) groups of instances from different bird categories with a shared attribute, (ii) groups of instances from the same bird category. For the latter, we induce additional context from a similar distractor group (G_d), composed of instances from a distractor bird category. We use distractor categories that belong to the same bird family as the target, for instance, *Black-footed-Albatross* and *Laysan-Albatross* from the bird family *Albatross*, similar to Vedantam et al. (2017), to test whether

[1] www.whatbird.com

we can generate informative descriptions in challenging contexts.

For training the instance-level speaker (S), we used the split as provided by Hendricks et al. (2016) (train:4000, val:1994, test:5794). For our shared attribute grouping, we sample a target group of size 3 for every instance in each split, such that we obtain 1994 and 5794 groups for val and test. For category-level grouping, we sample target groups of size 3 and distractor groups of size 4 for each instance. There are 7 bird species that do not have distractors from the same family and we ignore these here. For discriminative group decoding, we obtain 3358, 1646, and 4833 groups for train, val and test respectively.

4.3 Model

We first train an image classifier by finetuning a pretrained resnet-101 architecture to predict bird categories from bird images. The training parameters were set to: batch size 16, (RGB) image size as 448, learning rate 0.001 for a total of 50 epochs with a decay factor of 0.1 after every 20 epochs. We use this image classifier as visual encoder for our speaker $S(I)$, the image captioning model.

We trained two versions of our speaker $S(I)$, (i) a basic recurrent **LSTM** model architecture from Xu et al. (2015) and (ii) a basic **Transformer** by (Vaswani et al., 2017). Generally, transformers are currently the more popular model due to their parallel processing and multi-head attention architecture (Devlin et al., 2019; Lan et al., 2020), but they may also be more data-intensive. We wanted to see how both architectures (LSTM and Transformer) perform given our dataset is quite small. Both models use the visual encoder described above. Both our captioning models, the LSTM and Transformer, achieve similar CIDEr-D validation scores of 49.4 and 49.5 respectively, similar to existing captioning models for the birds data (Vedantam et al., 2017).[2].

5 Experiment 1: Shared Attributes

In this experiment, we investigate whether our group decoding mechanism can be used to systematically include a shared visual attribute in a description for group of instances (which may belong to a different bird categories).

[2]Code and models can be found here

5.1 Attributes

We sample groups with shared attributes based on the symbolic attribute annotations in the birds data. We use the attribute-value pair as a reference pattern that needs to be included in an accurate, coherent group description (e.g. for *belly-color:white* we look for *white belly*). It is important to note that the symbolic attribute annotations are significantly more detailed and elaborate in terms of their vocabulary (Section 4.1) than the captions which were crowd-sourced with non-experts. This results in a mismatch between aspects of birds that are annotated and properties that are verbalized in the captions and that we can expect the captioning model to be able to pick up.To tackle this, we restricted our group sampling to attributes that can be detected in captions by simple pattern search. We ranked the symbolic attributes present in the captions by frequency and selected randomly four more frequent and two less frequent attributes for our experiment. For simplicity, we used only one shared attribute per group at a time and no distractors, as we expect to obtain rather noisy distractor sets due to above mentioned issues with the attribute annotations.

5.2 Results

We assess the accuracy of decoding for groups with a shared attribute, i.e. whether the output description contains the shared attribute as identified by pattern search. Table 1 shows that for 4 out of 6 selected attribute-value pairs, group captions are clearly more likely to mention the selected common property than the instance captions, with increase in accuracy of up to 17% for *bill length*. We also note that the instance-level captions generated by the LSTM and Transformer show differences in their attribute patterns, despite their overall similar performance. We will discuss differences between the two models further below. Figure 3 shows a qualitative example where the group descriptions mention the shared property (*blue wing*) in contrast to all the instance descriptions.

For the less frequent attribute *bill shape* in the ground-truth instance captions, and not so distinctive attribute *eye color* (as most of the times it's value is black), the accuracy is low for both instance and group-level decoding and the instance-level decoding outperforms the group-level decoding for the *bill shape* attribute. This suggests that achieving coherence in group descriptions in decoding is contingent on shared properties occurring with


(a) this bird has a white belly and breast with a blue crown


(b) this is a small bird with a white belly and a blue head


(c) this bird has a blue crown and a long bill

$S(G_t)$-LSTM: this is a bird with **blue wings**.
$S(G_{t,d})$-Transformer: this is a bird with **blue wing**.

Figure 3: Generated instance and group caption for a shared-attribute group.

a certain frequency in the instance caption data, or, vice versa, that the group decoding may not push the captioning models towards selecting rare attribute words and fine-grained visual details.

		Mentions of shared attribute(%)			
Shared Attributes	Frequency (total)	LSTM		Transformer	
		group	instance	**group**	instance
breast color	10158	50	35.40	25.95	12.30
crown color	9693	31.57	20.57	38.61	19.59
belly color	9379	47.67	34.85	25.00	14.62
eye color	8666	14.86	10.06	19.08	16.22
bill length	7372	61.63	44.61	56.08	41.63
bill shape	6882	7.61	11.54	15.76	23.86

Table 1: Accuracy of generated group captions and instance captions in terms of mentioning a shared attribute. Frequency shows occurrence of a shared attribute in original captions.

6 Experiment 2: Category-level Grouping

In the second experiment, we test whether our decoding mechanism generates coherent and informative descriptions of groups that correspond to categories, i.e. instances are sampled based on category-level annotation in the birds dataset and, optionally, paired with distractor groups/categories.

6.1 Evaluation

Evaluation is challenging as we do not have vision-oriented ground-truth category descriptions. Expert-level category definitions from, e.g., bird dictionaries would not help to objectively assess our group descriptions as they use a more sophisticated vocabulary and commonly mention non-visual properties that cannot be learned by a captioning model. Therefore, we combine automatic evaluation based on automatically selected, prototypical reference descriptions, automatic category inference and human evaluation on the most promising models.

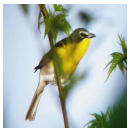

(a) this is a **black bird** with a **white eye** and a **large orange beak**.
(b) this is a **grey bird** with large feet, a white eye and an **orange beak**.
(c) this **dark grey** bird has a **orange bill** with **white eyes** and a **feather hanging over its bill**.
(d) this bird has an **all black body** with a large orange beak and a white eye.
(e) this is a **grey bird with black wings**,a white eye and an orange beak.

Figure 4: Five most similar instance descriptions for a bird category based on cosine similarity to centroid.

General quality of group captions We want to ensure that we do not lose lexical richness by moving from instance to group descriptions, as these problems have commonly been observed in neural NLG models. We computed average sentence length and Dist-k (Ippolito et al., 2019) (distinct unigrams and bigrams) to measure lexical diversity and repetitiveness in generated captions.

Prototypical reference captions We compile the reference set for a category by taking the top-5 prototypical descriptions for each bird category. We select these descriptions using kmodes clustering (de Vos, 2015) on pre-trained BERT sentence embeddings. We compute the centroid of the bird description embeddings and take the five most similar instance descriptions (according to cosine-similarity) as a stand-in for general, prototypical descriptions for the target category. Figure 4 shows an example of the top-5 descriptions determined by the clustering algorithm. It shows that they cover distinctive representative parts of a bird category, thereby getting rid of the erroneous (non-discriminating) descriptions.

Overlap with target and distractor references We use two standard overlap metrics, BLEU-4 (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015), to assess the similarity of generated group descriptions with reference sentence groups. We repurpose these for: (i) for **target-target similarity**, i.e. measuring the overlap of generated group descriptions to a set of references for the target group, (ii) for **target-distractor similarity**, i.e. measuring the overlap of generated group descriptions to a set of references for the *distractor group*. We expect the target-target similarity to go up for group captions and the target-distractor similarity to go down for group captions that are informative.

Category-level Inference In order to verify that the generated group descriptions indeed pull to-

gether properties relevant for the target category and make it distinct from the other distractor category, we learn an external text classifier based on BERT (Devlin et al., 2019). As we do not have ground-truth category descriptions, we use the generated group captions from our different group decoding methods and for a fair comparison, generated instance captions from the speaker $S(I)$ for training. The performance of these text classifiers give us some indication as to whether using group of instances during decoding leads to descriptions that make it easier to identify the target category, as compared to descriptions for single instances, in the absence of concrete visual instances. This resembles a setting where a speaker explains to a listener the properties that it has learned to detect for a given category. As for the training parameters of the text classifier, we set the batch size to 64 and learning rate to 0.00002 for a total of 60 epochs.

Human Evaluation We performed human evaluation on the most promising LSTM and Transformer speaker models using the Amazon Mechanical Turk (AMT) crowdsourcing platform, in order to analyze whether group descriptions were preferred over instance descriptions for describing a group of image instances. We showed the participants all images from the target group and two competing descriptions: (A) the generated discriminative group description and (B) the generated instance description (from a random instance in the group). We asked them to carefully observe the images and select the description(s) that best describe all or most of the images in the group in a forced-choice task with 3 options, (A), (B) or (C) both. We included the third choice as we observed that the instance descriptions in the birds data can be very similar to the prototypical description of the target category and we wanted to avoid random choices by participants for these cases. We randomly selected 2 groups out of 60 bird categories, having a total of 120 group and instances descriptions. More details on the set-up are provided in Appendix B.

6.2 Results

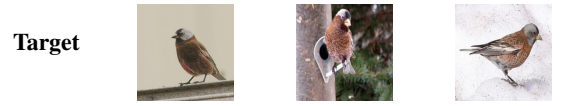
Figure 5 shows generated descriptions produced for instances and groups using category-level decoding with both LSTM and Transformer based speakers.

Quality and Overlap Metrics Table 2 reports the automatic overlap metrics for target-target similarity and for target-distractor similarity for different models and decoders. These results indicate that there is a general positive tendency towards higher target-target similarity and lower target-distractor similarity when using group-level instead of instance-level decoding. Another general tendency is that the difference between the instance-level decoding and the coherence-only group decoding (with distractors) is rather subtle and that the real gain comes from combining the coherence and discrimination objective, i.e. CIDEr scores for target-target similarity increase from 68 to 81 and 79 to 88 for the LSTM and Transformer when used with $S(G_{t,d})$ instead of $S(G_t)$ (the λ parameter needs to be set differently with the two captioning models). CIDER also predicts a rather sharp decrease of target-distractor similarity for the transformer-based decoding (47 to 36), but less of a decrease for the LSTM-based discriminative group decoding. This suggests that captions decoded on the group-level are more likely to mention properties that are both more coherent and informative for the target category. CIDEr scores show a big positive effect for using discriminative group-level decoding with the LSTM and the Transformer on target-target similarity, whereas the BLEU-4 score indicates a smaller increase. Furthermore, CIDER indicates a strong difference for instance-level decoding between LSTM und Transformer, whereas BLEU-4 favours instance-level captions generated by the LSTM (in terms of their similarity to the group reference). For this reason, we complement this type of evaluation with further assessments below. Finally, we find that the average sentence length and the dist-k scores are high for the instance and for category descriptions, as shown in Table 2. This shows our group-based decoding does not lead to negative effects regarding length or repetitiveness which have been observed for other decoding methods in neural NLG (Ippolito et al., 2019; Zarrieß and Schlangen, 2018).

Category-level Inference Table 3 shows accuracy results for text classifiers trained to identify the bird category based on generated captions. We find that coherent group decoding improves the prediction of target categories and discriminative decoding enhances the classifier further. Moreover, this evaluation indicates the superior performance of the Transformer over the LSTM speaker, in line with the CIDEr evaluation in Table 2. This suggests that the power of the underlying captioning model, which may not become apparent in instance-

Model	Decoding	λ	Target-target sim. (↑)		Target-distractor sim. (↓)		Diversity		
			BLEU-4	CIDEr	BLEU-4	CIDEr	Dist-1	Dist-2	avg. len
LSTM	$S(I)$	-	42.41	68.89	36.56	44.97	0.88	0.98	12.96
	$S(G_t)$	-	42.54	68.11	36.70	44.59	0.89	0.98	13.01
	$S(G_{t,d})$	0.3	45.11	**81.32**	34.04	**40.86**	0.86	0.97	12.91
	$S(G_{t,d})$	0.5	44.55	78.21	36.10	43.79	0.88	0.98	12.97
Transf	$S(I)$	-	40.68	77.44	32.89	47.02	0.89	0.98	13.29
	$S(G_t)$	-	41.16	79.45	32.45	44.46	0.90	0.99	13.27
	$S(G_{t,d})$	0.3	42.62	83.79	28.58	**36.96**	0.84	0.96	13.36
	$S(G_{t,d})$	0.5	43.69	**88.87**	31.27	41.54	0.88	0.98	13.31

Table 2: Evaluation of category-level group captions for overlap with prototypical target and distractor references. Decoding: $S(I)$ instance-level, $S(G_t)$ coherent group decoding, $S(G_{t,d})$ discriminative group decoding.



$S(I)$:this bird has a speckled belly and breast with a short pointy bill (same description for all instances)

$S(G_{t,d})$-LSTM: this is a bird with a grey belly and a grey head
$S(G_{t,d})$-Transformer: this is a brown bird with a grey head

Figure 5: Generated group caption for category.

level use, is important for high-quality group-level decoding. In future work, we plan to further analyze the interaction of the underlying captioning architecture with the decoding mechanism.

Model	Decoding	λ	Accuracy
LSTM	$S(I)$	-	18.22
	$S(G_{t,})$	-	19.70
	$S(G_{t,d})$	0.3	33.14
	$S(G_{t,d})$	0.5	25.59
Transformer	$S(I)$	-	23.60
	$S(G_t)$	-	29.48
	$S(G_{t,d})$	0.3	42.72
	$S(G_{t,d})$	0.5	36.90

Table 3: Text classification performance for category identification. Discriminative group decoding $S(G_{t,d})$ leads to best performance for LSTM and Transformer.

Human Evaluation Table 4 shows that participants prefer group over instances descriptions for the LSTM and Transformer model for 59% of the items. Again, we see that the instance-level Transformer outperforms the LSTM, i.e. there are fewer Transformer captions where participants rate the instance and group-level caption equally. Generally, this clearly supports our hypothesis that group-level decoding can pull together multiple distinctive properties common to a group or category.

	Selected by participants (%)		
Model	$S(I)$	$S(G_{t,d})$	*Both*
LSTM	9.17	59.17	31.67
Transformer	17.5	59.17	23.33

Table 4: Human evaluation with portion of items where participants selected generated instance-level, group-level or both captions as appropriate for a group.

6.3 Limitations

As our approach to decoding group-level descriptions is conceptually simple, it is not surprising that it has certain limitations in terms of the linguistic phenomena it is able to account for. Figure 6 shows examples for systematic limitations (and directions for future work): (i) **describing discriminative details**: for some bird families, the effect of group decoding is not significant and fixating fine-grained details is not yet possible, see the *sparrow* example in Figure 6's first row. (ii) **completeness**: group descriptions do not always mention all the properties that might be used to define a category because the distractor group has similar properties, in Figure 6 third row, *black on its wings* was ignored due to the distractor group. (iii) **disjunctive properties within a category**: different physical appearance of male and female instances of a bird species leads to incoherent captions as in Figure 6 second row.

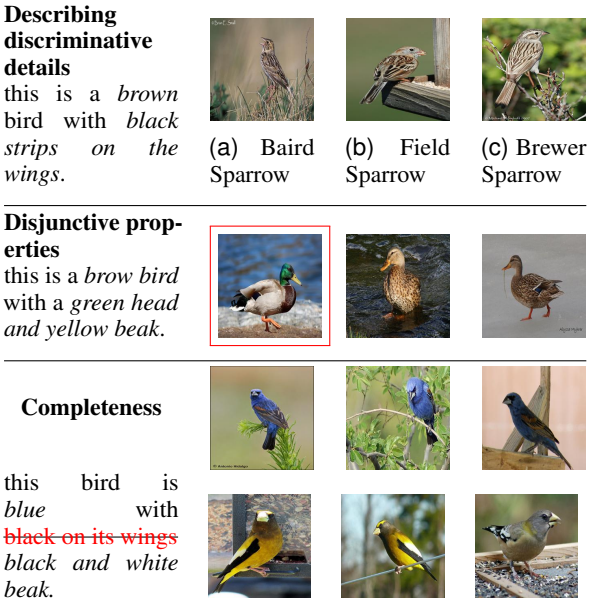


Figure 6: Examples for errors and limitations.

7 Conclusion

In this paper, we have proposed a task, a set-up and a decoding procedure for generating group-level descriptions with an instance-level captioning model. Despite our decoding approach being arguably simple, the results are encouraging and point into some interesting directions for future work. The classical problem of REG could be re-visited on a larger scale for sets of "real-world" objects or one could explore the use of group decoding in explanation scenarios where additional category label information or predicted attention maps could be integrated to provide post-hoc justifications. Finally, enhancing the decoding mechanism with deeper logical reasoning capabilities (e.g. on disjunctions) seems to be a promising direction.

References

Peter Anderson, X. 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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443.

Marcella Cornia, Matteo Stefanini, L. Baraldi, and R. Cucchiara. 2020. Meshed-memory transformer for image captioning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nelis J. de Vos. 2015. kmodes categorical clustering library. https://github.com/nicodv/kmodes.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

Lianli Gao, Xiangpeng Li, Jingkuan Song, and Heng Tao Shen. 2020. Hierarchical lstms with adaptive attention for visual captioning. *IEEE Trans. Pattern Anal. Mach. Intell.*

Claire Gardent. 2002. Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 96–103.

Albert Gatt. 2007. *Generating coherent references to multiple entities*. Ph.D. thesis, Citeseer.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating Visual Explanations. In *ECCV*.

Helmut Horacek. 2004. On referring to sets of objects naturally. In *International Conference on Natural Language Generation*, pages 70–79. Springer.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Zhuowan Li, Quan Hung Tran, Long Mai, Zhe Lin, and A. Yuille. 2020. Context-aware group captioning via self-attention and contrastive features. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3437–3447.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. 2021. Dual-level collaborative transformer for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jonghwan Mun, L. Yang, Zhou Ren, N. Xu, and Bohyung Han. 2019. Streamlined dense video captioning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning Deep Representations of Fine-Grained Visual Descriptions. In *CVPR*. IEEE Computer Society.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Matthew Stone. 2000. On identifying sets. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 116–123.

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

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*. IEEE Computer Society.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2017. Hierarchically-attentive rnn for album summarization and storytelling. In *EMNLP*.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512, Tilburg University, The Netherlands. Association for Computational Linguistics.

Sina Zarrieß and David Schlangen. 2019. Know what you don't know: Modeling a pragmatic speaker that refers to objects of unknown categories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 654–659, Florence, Italy. Association for Computational Linguistics.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34.

A Automatic Evaluation II

We used another reference set for automatic evaluation of category-level group descriptions which is union of ground-truth descriptions of all instances in a group. This amounts to 15 reference captions for the target group of size 3 in our case and 20 reference captions for the distractor group as distractor group size is 4. We observe that group of ground-truth instance descriptions are still to some extent composed of idiosyncratic properties of instances. This can be seen from small amount of increase in CIDEr scores from instance to discriminative group descriptions in Table 5 compared to that of in Table 2 using prototypical references.

B Crowdsourcing Details

In this section, we provide additional information on how we conducted the human evaluation using AMT crowdsourcing platform. We recruited

Model	Decoding	λ	Target-target sim. (↑)		Target-distractor sim. (↓)	
			BLEU-4	CIDEr	BLEU-4	CIDEr
LSTM	$S(I)$	-	62.85	42.73	61.27	28.08
	$S(G_t)$	-	64.01	44.38	63.15	29.61
	$S(G_{t,d})$	0.3	63.67	46.24	55.81	23.49
	$S(G_{t,d})$	0.5	64.57	46.84	60.18	26.91
Transf	$S(I)$	-	58.30	43.17	54.57	25.95
	$S(G_t)$	-	60.09	45.19	56.41	27.31
	$S(G_{t,d})$	0.3	57.30	43.58	46.03	19.58
	$S(G_{t,d})$	0.5	60.57	47.50	52.30	23.78

Table 5: Evaluation of category-level group captions for overlap with union of ground-truth instance descriptions from target and distractor groups.

participants who are native english speakers (e.g., from United Kingdom, United States) as our task requires English proficiency. We paid the participants 0.15$ for successfully completing the task based on a fair hourly wage. Figure 7 shows an example of how our task was presented to the participants. The participants were aware that the task is purely for research purposes and contains no form of controversial data.

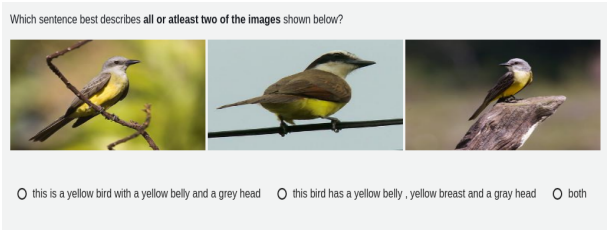


Figure 7: An example of the task seen by the participants on AMT platform.