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

INLG, 18-22 July 2022

Nazia Attari[*]

David Schlangen[†]

Martin Heckmann[‡]

Heiko Wersing[Ω]

Sina Zarrieß[*]

[*]Bielefeld University, [Ω]HRI-Europe, [†]University of Potsdam, [‡]Aalen University
Germany

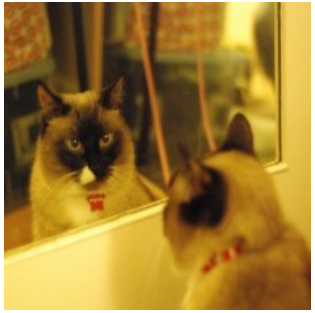
Outline

1. Background
2. Task
3. Group Decoding Approach
4. Data
5. Results
6. Limitations
7. Conclusion and Future Directions

Background

A range of research has explored improving generation of ***single image descriptions***.

Image instance(s)



(Anderson, P., et al. CVPR 2018)

Caption

A cat looking at its reflection in a mirror.

Background

- Context-aware captions using context-agnostic captions
- Inducing pragmatic reasoning during inference time

Context-aware captions at inference time

A *large passenger jet* flying through a blue sky.

(Vedantam, R., et al. CVPR 2017)

Group Decoding: Coherent

- Current image captioning approaches have not been utilized to describe a group (or set) of objects.
- A classical problem in referring expression generation (REG) [Stone, 2000; Gardent, 2002; Horacek, 2004; Gatt, 2007;Krahmer and van Deemter, 2011]

(true for all or majority of instances)
Coherence

Group of image instance(s)

T
A
R
G
E
T

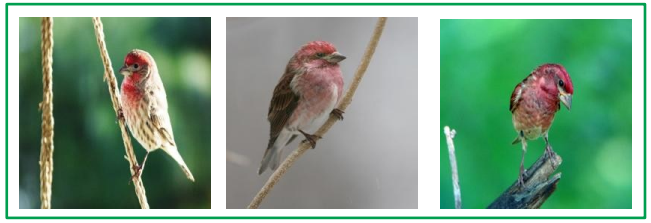
Group caption: this is a **blue bird** with a **short black and white bill** and **brown wingbars**

Group Decoding: Informative

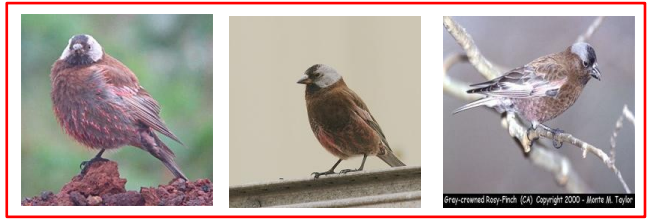
(distinctive in a particular context)
Informative

Group of image instance(s)

TARGET

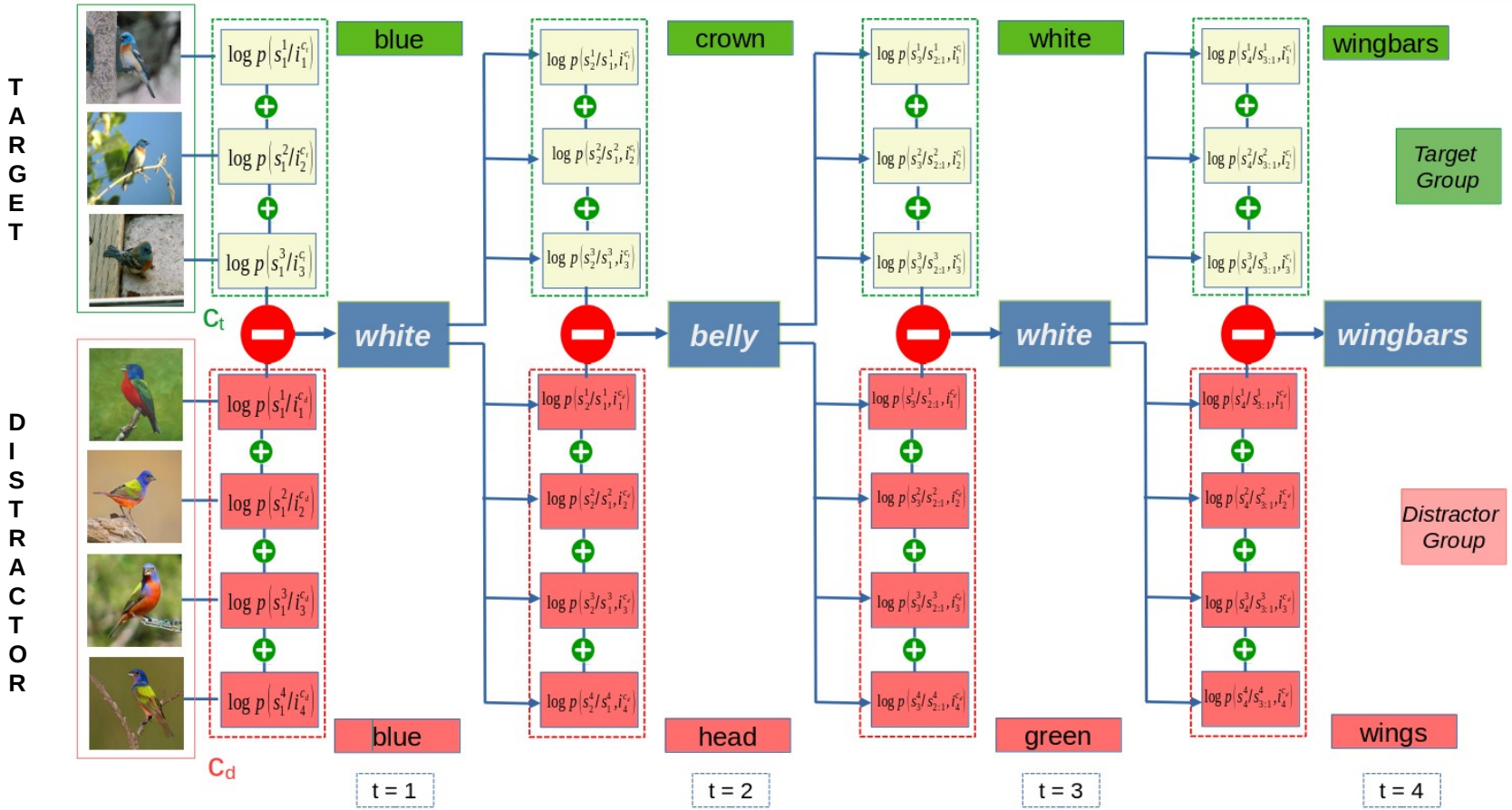


DISTRACTOR



Group caption: a small bird with **a red crown**

Block diagram: Decoding for coherent and informative Group Captioning



Data: Caltech Birds Dataset (CUB-200-2011)

Image Instances	Category	Instance Captions	Symbolic Attribute
	Horned Puffin	1. **this large black and white bird has orange feet, bill, and eyes.** 2. a large bird with an u ordinary shaped red and white bill, a white under belly, and orange webbed feet. 3. this large bird has a white face, breast, belly and vent, and black covering the rest of its body, except for the red patch next to its bill. 4. a large bird with bright orange feed and a short, wide bill. 5. a large bird with a white breast, black back, and orange feet.	has_bill_shape::triangle-shaped has_back_color::black has_belly_color::white has_crown_color::black has_wing_pattern::solid has_leg_color::orange . . and so on..
	Horned Lark	1. **this elongated bird has a soft tan coloring with yellow and black markings found on the face.** 2. bird is beige with a little beak and white thin legs 3. a bird with a yellow eyebrow, black cheek, small triangular bill and tan plumage 4. **this little beige bird is carrying some type of nut in his beak.** 5. this bird has a yellow throat, white breast, and brown body.	has_bill_length::short has_bill_shape::pointy has_belly_color::white has_eye_color::black has_wing_shape::tapered-wings . . and so on..
~60 instances per category	200	5 captions per instance	312 attribute-value pairs

Data: Sampling Groups

Using a shared attribute

Target Group: breast color - yellow

TARGET



Category-based Grouping

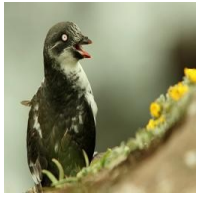
Target Group: Crested Auklet

TARGET



Distractor Group: Least Auklet (same hyper-category, using last name information)

DISTRACTOR


Data: Prototypical descriptions as *References*

***Problem*:** Missing group descriptions

Solution: Using top-5 descriptions for a category closest to the centroid based on cosine similarity.
(k-modes algorithm using instance-based descriptions)

Meaningful discriminative caption

this **dark grey bird** has a **orange bill** with **white eyes** and a **feather hanging over its bill**

Erroneous caption

the bird **looks up on the sky in search of its partner**

Experimental Set-Up

- Decoder (or Speaker) for instance-based image captioning
  - LSTM (Xu et al. ,2015)
  - Transformer (Vaswani et al., 2017)

- Evaluation for different types of groups:
  - Shared Attribute
    - *Phrase matching*

  - Category-based grouping
    - *Automatic metrics (BLEU-4, CIDEr) using:*
      - *References:*
        - *prototypical target description, prototypical distractor description*
      - Similarity measure:
        - target-target, target-distractor
    - *Text classification (BERT-based, Devlin et al., 2019)*
    - *Human Evaluation (using Mturk platform)*

Result: Group with a Shared Attribute

		Mention of shared attributes (%)			
Shared Attributes	Frequency (total)	LSTM		Transformer	
		group	instance	**group**	instance
breast color		50	35.40	25.95	12.30
crown color		31.57	20.57	38.61	19.59
belly color		47.67	34.85	25.00	14.62
eye color		14.86	10.06	19.08	16.22
bill length		61.63	44.61	56.08	41.63
bill shape		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.

Result: Category-based Grouping (Automatic Evaluation)

Model	Decoding	λ	Target-target similarity ↑		Target-distractor similarity ↓	
			BLEU-4	CIDEr	BLEU-4	CIDEr
LSTM	Target Group with Distractor Group	0.3	45.11	81.32	34.04	40.86
	Instance	-	42.41	68.89	36.56	44.97
Transf	Target Group with Distractor Group	0.5	43.69	**88.87**	31.27	41.54
	Instance	-	40.68	77.44	32.89	47.02

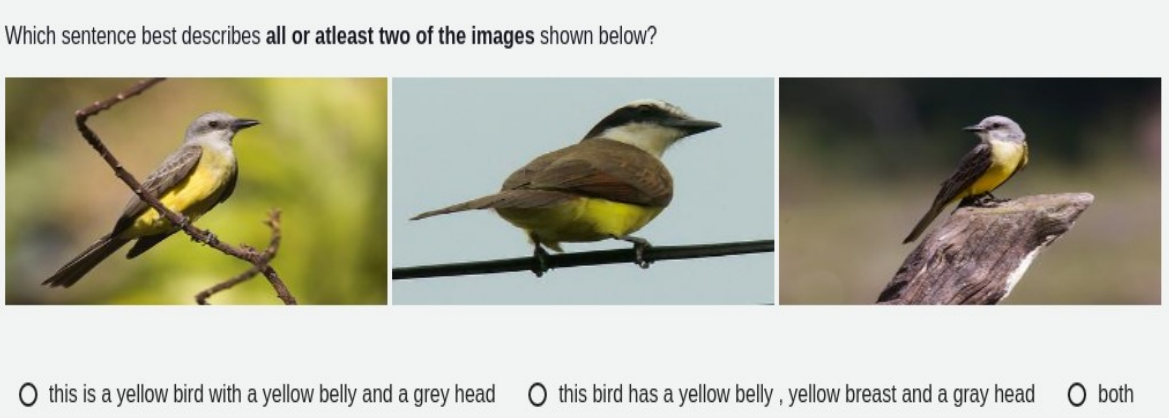
Table 2: Evaluation of category-level group captions for overlap with prototypical target and distractor references.

Result: Groups based on Category (Text Classification)

Translation Model	Decoding	λ	Accuracy
LSTM	Target Group with Distractor Group	0.3	33.14
	Instance	-	18.22
Transformer	Target Group with Distractor Group	0.5	**36.90**
	Instance	-	23.60

Table 3: Text classification performance for category identification.

Result: Human Evaluation



Amazon Mechanical Turk Task Set-Up

Choice of descriptions for participants:

- group caption
- one of the instances caption
- both (if both above two descriptions are true)

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

Model	Human Preference (%)		
	instance	**group**	both
LSTM	9.17	60	31.67
Transformer	17.5		23.33

Generated Group Caption: LSTM vs. Transformer



Generated group caption

LSTM: this bird has a yellow belly and breast with a short pointy bill

Transformer: this bird has a yellow belly and breast with a gray crown and wing

Feature Addition



Generated group caption

LSTM: this is a bird with a grey belly and a grey head

Transformer: this is a brown bird with a grey head

Feature Correction

Limitations

Property	Image Group	Group Caption
Describing discriminative details	   Baird Sparrow Field Sparrow Brewer Sparrow	this is a brown bird with black strips on the wings. non-distinctive
Disjunctive properties	   Male Female	This is a brown bird with a green head and yellow beak.
Completeness	     	This bird is blue with ~~black on its wings and~~ black and white beak.

Conclusion and Future Directions

- We have proposed a task, a set-up and a decoding procedure for generating group-level descriptions with an instance-level captioning model
- The classical problem of REG could be re-visited on a larger scale for sets of “real-world” object
- The use of group decoding in explanation scenarios as additional category label information could be explored
- Enhancing the decoding mechanism with deeper logical reasoning capabilities (e.g. on disjunctions) seem to be a promising direction

Thank you for your attention!