Describe Me an Auklet: Learning to Combine Visual and Conceptual Knowledge for Generation and Interpretation of Perceptual Categories

Nikolai Ilinykh Bill Noble
Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
{name.surname}@gu.se

CLASP Seminar, May 5, 2022

Outline

Outline

You have probably never seen a bird like this...

You have probably never seen a bird like this...

- **But what do we know about what is in this image?**
- Well, it *is* a bird.
- It is very likely that we do *not* know the type of the bird.
- This unknown bird would thus remind us of some other birds that "look like it".
- We can also say that this bird looks "goofy".
- How would you describe this bird to someone who has never seen one? It depends on (i) who is describing it, (ii) who are you describing it to.

No image, only text description: what can you learn?

GÖTEBORGS UNIVERSITET

A Crested Auklet has black wings, pointy orange bill,
a black thing on its head and looks goofy.

- Although image is not immediately available, from text alone you know which visual clues to look for once you see the bird.
- You are also very likely to *activate* knowledge of the bird domain and make your task easier by imagining how a Crested Auklet would look like.
- Grounded language learning is about grounding text in **perception and knowledge**.

Learning from Different Sources of Information

GÖTEBORGS UNIVERSITET

1. We can easily extract a lot of information about the world with our perception e.g., when seeing an unknown bird.
2. However, our "representational ability" heavily relies on previous knowledge, both visual and conceptual, because visual representations are not always immediately available to us and we tend to utilise other sources of knowledge.
3. We might ground what we comprehend into what we store and keep in mind.

Research Questions

GÖTEBORGS UNIVERSITET

Our **research questions** are as follows:

- To what extent can we employ both generation (NLG) and interpretation (NLU) in a more natural grounded language learning scenario?
- How can we synthesize perceptual and conceptual representational knowledge?
- What makes descriptions effective at teaching novel categories?

Outline

Learning from Descriptions

Bob: Do you know what a Crested Auklet looks like? I've never seen one.

Alice: It's a goofy looking, large bird that has a bright orange break with a musky gray body and charcoal wing feathers.

Bob: Hmm. Ok, I'll keep my eye out for one...





Outline

Grounded NLP I

GÖTEBORGS UNIVERSITET

Most work in grounding and NLP is (i) multi-modal (e.g., language-and-vision) and (ii) focuses on situations where there is an immediately available one-to-one correspondence between linguistic and perceptual input.

- Referring expression generation (Krahmer and van Deemter, 2012)
- Image captioning (Bernardi et al., 2017)
- Visually grounded dialogue games (De Vries et al., 2017; Haber et al., 2019; Ilinykh et al., 2019; Dobnik and Silfversparre, 2021)

What has been missing is the opposite: when language is used to describe situations that do not correspond to a shared visual scene.

Classifier-Based Perceptual Semantics

GÖTEBORGS UNIVERSITET

- Intuition: Part of what it is to know the meaning of a (perceptual) word is to be able to recognise instances of it in the world
- Two main approaches:
  - *functional approach* – classifier is a function $f : \mathit{PerceptualData} \rightarrow [0, 1]$, corresponding to $e \rightarrow t$ in classical type theoretic semantics (Larsson, 2013).
  - *distributed approach* – parameters of a classifier (e.g., weight matrix) are regarded as a representation of the word meaning (e.g., Schlangen et al., 2016).
- In this work, we take **the distributed approach**.

Zero-shot Classification

GÖTEBORGS UNIVERSITET

- In zero-shot classification, we split the classes according to those known at train, say Z time and those that are only shown at test time, say Z'.
- The knowledge gained by learning to differentiate between the classes in Z needs to be transferred to the task of differentiating between the classes in $Z \cup Z'$.
- Paz-Argaman et al. (2020): text-based zero-shot classification of categories of objects in images based on (i) visual similarities reflected in texts and (ii) visual features which are reflected in text.
- Hill et al. (2021): zero-shot learning of novel objects in more interactive scenarios, e.g. a robot that follows different commands and learns to interact with surrounding objects.

Outline

Data: Images (Wah et al., 2011)

GÖTEBORGS UNIVERSITET

CUB - Caltech-UCSB Birds 200

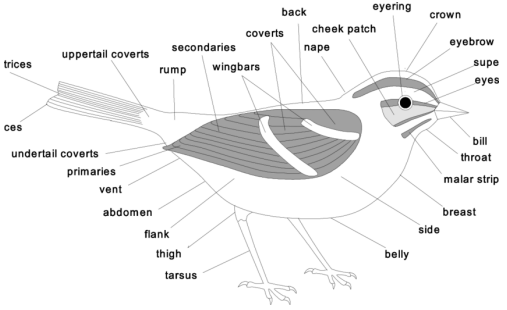
- 11K images of 200 different bird species, downloaded from Flickr
- bounding boxes and "attribute" values annotated by AMT workers (we don't use these currently)

Data: Descriptions (Reed et al., 2016)

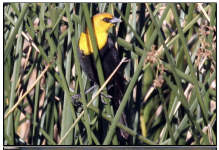
GÖTEBORGS UNIVERSITET

Descriptions

- 10 descriptions of each bird image collected from AMT
- Instructions:
  - *describe only visual appearance in at least 10 words, to avoid figures of speech, to avoid naming the species even if they knew it, and not to describe the background or any actions being taken*
  - *the prompt included three example sentences and a diagram labeling specific parts of a bird*

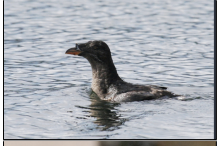


Data: Examples



the bird has a yellow breast and black belly as well as a small bill

this funny looking bird is black with white stripes and has a large white spot on its head



the ugly grey bird has a chicken like head but swims in the water .

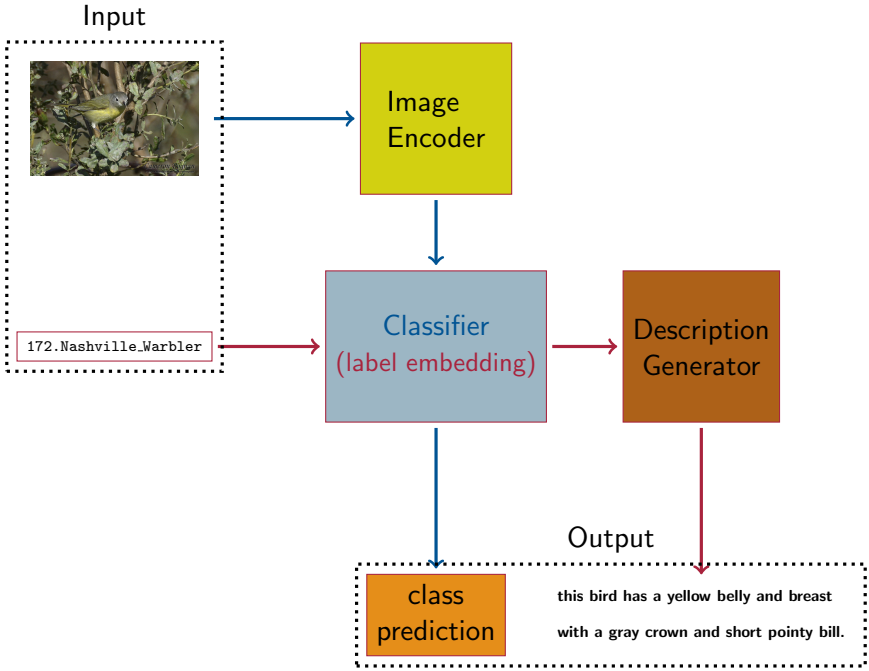
this bird is squat with a medium - sized dark bill , white head and breast , light brown abdomen , dark wings , and long tail that is twice the length of the bird ' s body .

Splits

- train/val/test split (instance-wise)
  - 80%/10%10%
  - 5-6 bird images per class in the val/test sets
- seen/un-seen split (category-wise)
  - 180 seen, 10 unseen (by Bob) bird categories
  - option for multiple folds

Outline

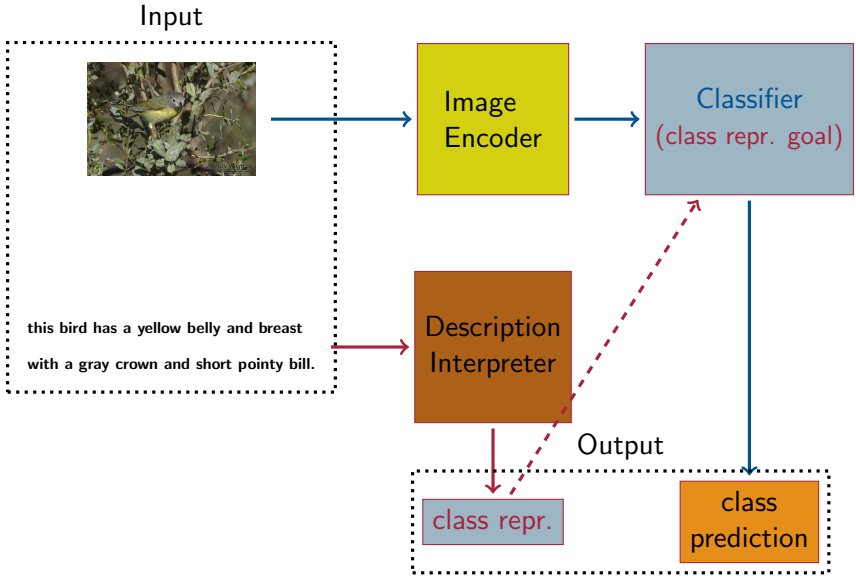
Generation model



Intepretation model

- Image encoder – VGG16 (pre-trained on Imagenet classification); convolutional layers + first two linear layers
- Classifier – Single fully-connected layer (with bias); softmax activation
- Description generator – LSTM decoder

Interpretation model



Intepretation model

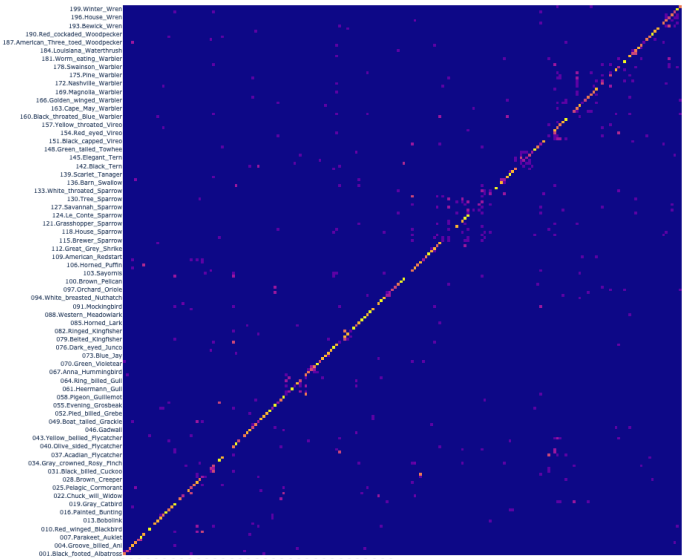
GÖTEBORGS UNIVERSITET

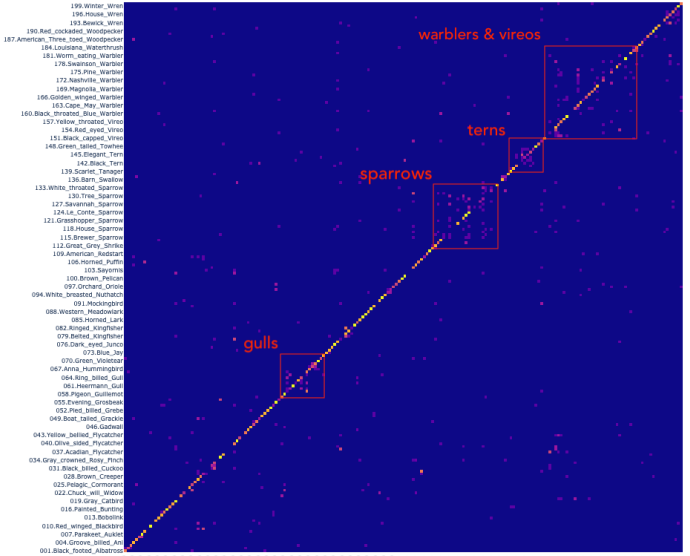
- Image encoder – VGG16 (pre-trained on Imagenet classification): convolutional layers + first two linear layers
- Classifier – Single fully-connected layer (no bias); softmax activation
- Description Interpreter – BERT [CLS] token pooler output; single linear layer with tanh activation (mean squared error loss function)

Outline

Classification results (not zero-shot)

	classifier loss	true rank	acc@1	acc@5	acc@10
without generation	4.786	5.7	0.58	0.84	0.91
with generation	4.756	6.4	0.61	0.83	0.89


199.Winter_Wren
196.House_Wren
193.Bewick_Wren
190.Red_cockaded_Woodpecker
187.American_Three_toed_Woodpecker
184.Louisiana_Waterthrush
181.Worm_eating_Warbler
178.Swainson_Warbler
175.Pine_Warbler
172.Nashville_Warbler
169.Magnolia_Warbler
166.Golden_winged_Warbler
163.Cape_May_Warbler
160.Black_throated_Blue_Warbler
157.Yellow_throated_Vireo
154.Red_eyed_Vireo
151.Black_capped_Vireo
148.Green_tailed_Towhee
145.Elegant_Tern
142.Black_Tern
139.Scarlet_Tanager
136.Barn_Swallow
133.White_throated_Sparrow
130.Tree_Sparrow
127.Savannah_Sparrow
124.Le_Conte_Sparrow
121.Grasshopper_Sparrow
118.House_Sparrow
115.Brewer_Sparrow
112.Great_Grey_Shrike
109.American_Kestrel
106.Horned_Puffin
103.Sayornis
100.Brown_Pelican
097.Orchard_Oriole
094.White_breasted_Nuthatch
091.Mockingbird
088.Western_Meadowlark
085.Horned_Lark
082.Ringed_Kingfisher
079.Belted_Kingfisher
076.Dark_eyed_Junco
073.Blue_Jay
070.Green_Violetear
067.Anna_Hummingbird
064.Ring_billed_Gull
061.Heermann_Gull
058.Pigeon_Guillemot
055.Evening_Grosbeak
052.Pied_billed_Grebe
049.Boat_tailed_Grackle
046.Gadwall
043.Yellow_bellied_Flycatcher
040.Olive_sided_Flycatcher
037.Acadian_Flycatcher
034.Gray_crowned_Rosy_Finch
031.Black_billed_Cuckoo
028.Brown_Creeper
025.Pelagic_Cormorant
022.Chuck_will_Widow
019.Gray_Catbird
016.Painted_Bunting
013.Bobolink
010.Red_winged_Blackbird
007.Parakeet_Auklet
004.Groove_billed_Ani
001.Black_footed_Albatross
199.Winter_Wren
196.House_Wren
193.Bewick_Wren
190.Red_cockaded_Woodpecker
187.American_Three_toed_Woodpecker
184.Louisiana_Waterthrush
181.Worm_eating_Warbler
178.Swainson_Warbler
175.Pine_Warbler
172.Nashville_Warbler
169.Magnolia_Warbler
166.Golden_winged_Warbler
163.Cape_May_Warbler
160.Black_throated_Blue_Warbler
157.Yellow_throated_Vireo
154.Red_eyed_Vireo
151.Black_capped_Vireo
148.Green_tailed_Towhee
145.Elegant_Tern
142.Black_Tern
139.Scarlet_Tanager
136.Barn_Swallow
133.White_throated_Sparrow
130.Tree_Sparrow
127.Savannah_Sparrow
124.Le_Conte_Sparrow
121.Grasshopper_Sparrow
118.House_Sparrow
115.Brewer_Sparrow
112.Great_Grey_Shrike
109.American_Kestrel
106.Horned_Puffin
103.Sayornis
100.Brown_Pelican
097.Orchard_Oriole
094.White_breasted_Nuthatch
091.Mockingbird
088.Western_Meadowlark
085.Horned_Lark
082.Ringed_Kingfisher
079.Belted_Kingfisher
076.Dark_eyed_Junco
073.Blue_Jay
070.Green_Violetear
067.Anna_Hummingbird
064.Ring_billed_Gull
061.Heermann_Gull
058.Pigeon_Guillemot
055.Evening_Grosbeak
052.Pied_billed_Grebe
049.Boat_tailed_Grackle
046.Gadwall
043.Yellow_bellied_Flycatcher
040.Olive_sided_Flycatcher
037.Acadian_Flycatcher
034.Gray_crowned_Rosy_Finch
031.Black_billed_Cuckoo
028.Brown_Creeper
025.Pelagic_Cormorant
022.Chuck_will_Widow
019.Gray_Catbird
016.Painted_Bunting
013.Bobolink
010.Red_winged_Blackbird
007.Parakeet_Auklet
004.Groove_billed_Ani
001.Black_footed_Albatross
1
0.8
0.6
0.4
0.2
0


warblers & vireos
terns
sparrows
gulls
1
0.8
0.6
0.4
0.2
0

Automatic Evaluation of Generation

Model Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
CLS vector + LSTM	76.12	60.87	46.74	35.04	29.78	20.06

- Decoding: greedy
- Inflated evaluation metrics due to the number of reference captions (Post, 2018).

Model Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
CLS vector + LSTM	46.12	28.54	18.24	12.07	19.47	19.98

- The scores are *still* high enough even when a single reference caption is used to evaluate the generated texts.
- Specifically, CIDEr score is affected the least.

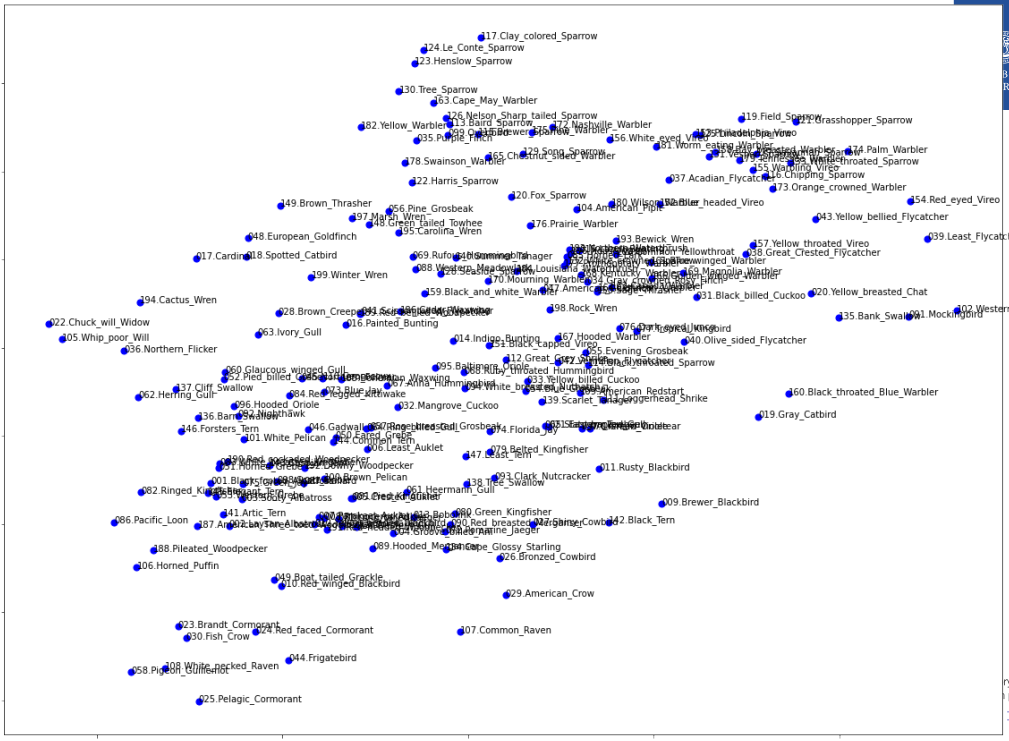
Generation: Examples

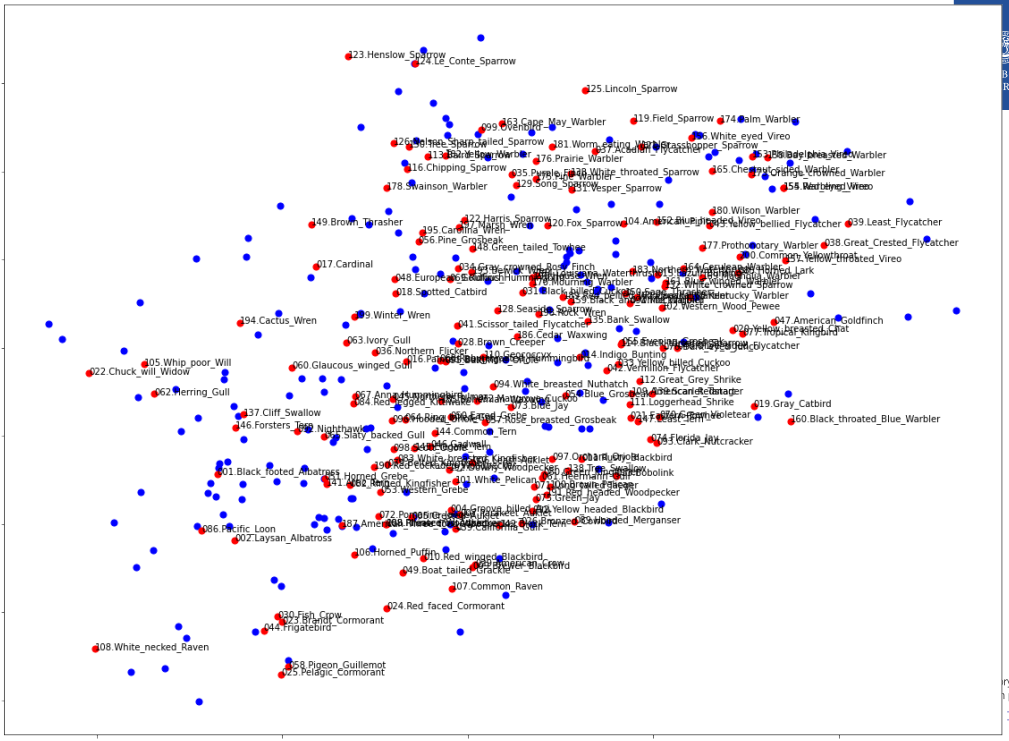
- **Reference**: this bird has large orange bill , a gray crown and nape , black and gray retrices and wings , and a white eye stripe .
- **Epoch 1**: this bird has a black crown , ~~a white breast~~ , and a black bill .
- **Epoch 4**: this bird has a black crown , a black bill , and ~~a white breast~~ .
- **Epoch 10:**: this bird has a black crown , **a short orange bill** , and a white eyering .
- **Epoch 20**: this bird has a black crown , a black breast , and **a short orange bill**

Generation: Analysis

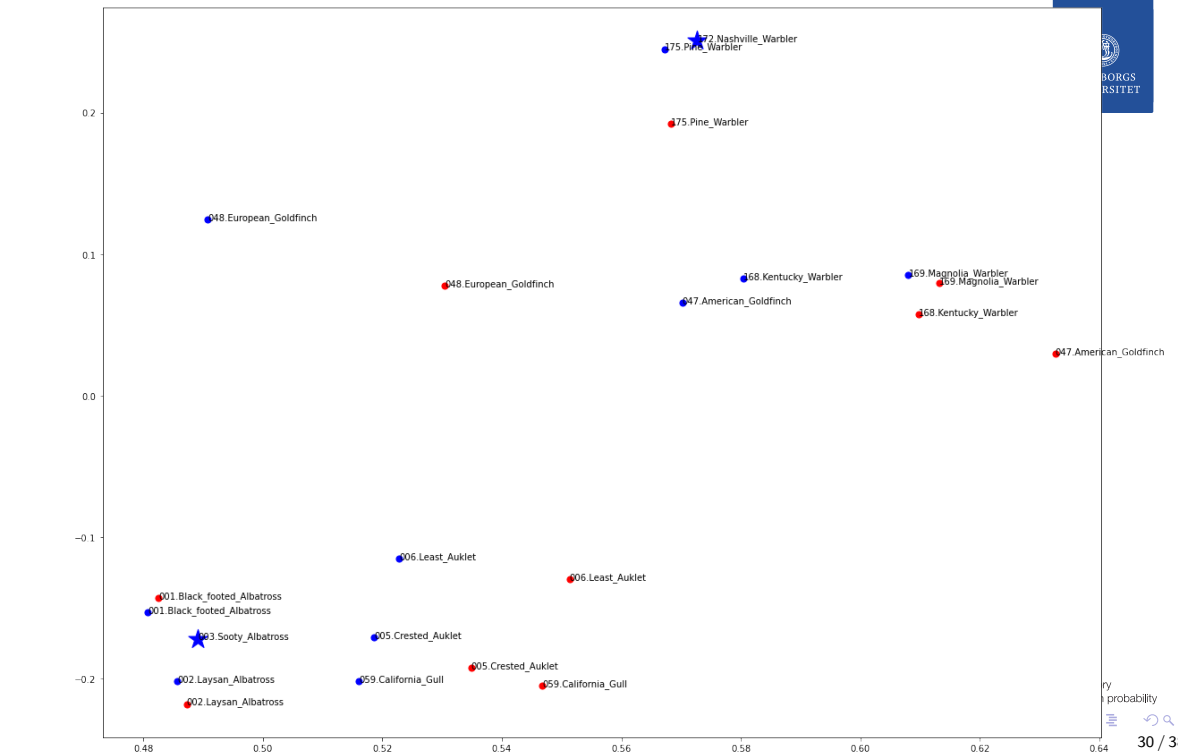
GÖTEBORGS UNIVERSITET

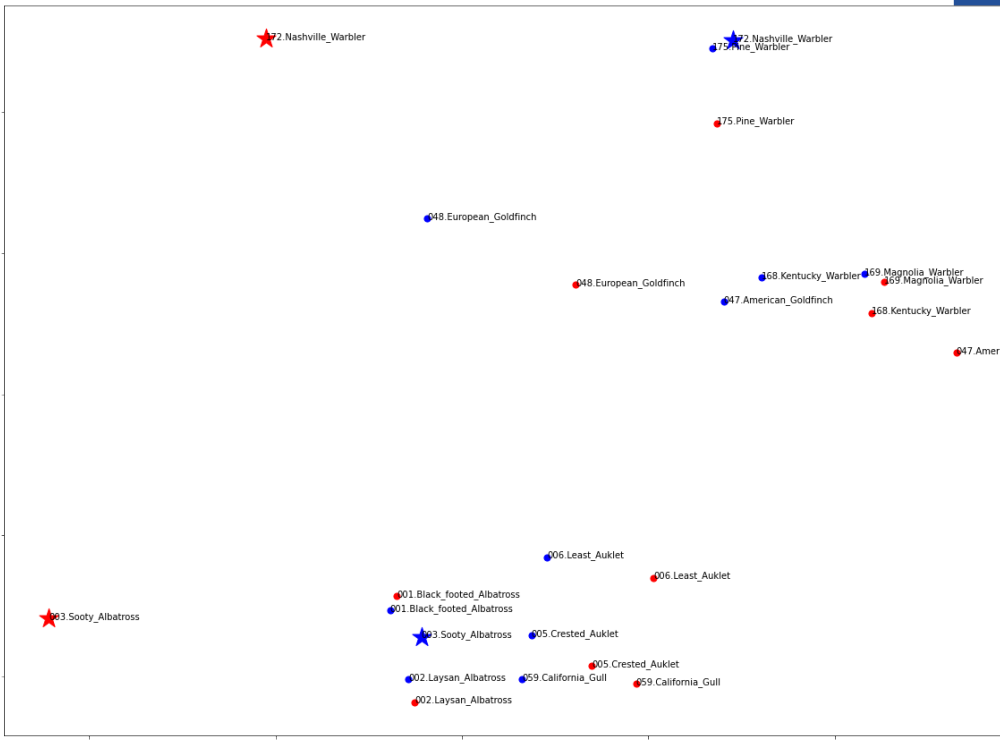
- Generated texts capture quite a lot of inter-class discriminative features in later stages of training; earlier stages of training capture more generic information (parts of birds and their attributes which appear very frequently between classes).
- We want to find a good trade-off between **discriminativeness** and **salience** of what is mentioned in texts.
- We propose to use interpretation task accuracy as means to evaluate the quality of generation (task-oriented) (will be examined in future work).


117.Clay_colored_Sparrow
124.Le_Conte_Sparrow
123.Henslow_Sparrow
130.Tree_Sparrow
63.Cape_May_Warrbler
126.Nelson_Sharp_tailed_Sparrow
119.Field_Sparrow
127.Grasshopper_Sparrow
113.Baird_Sparrow
182.Yellow_Warbler
122.Nashville_Warbler
56.White_eyed_Vireo
35.Purple_Finch
129.Song_Sparrow
74.Palm_Warbler
178.Swainson_Warbler
65.Chestnut_sided_Warbler
55.Warbling_Vireo
122.Harris_Sparrow
37.Acadian_Flycatcher
16.Chipping_Sparrow
73.Orange_crowned_Warbler
120.Fox_Sparrow
54.Red_eyed_Vireo
149.Brown_Thrasher
56.Pine_Grosbeak
104.American_Pipit
43.Yellow_bellied_Flycatcher
197.Marsh_Wren
48.Green_tailed_Towhee
95.Carolina_Wren
176.Prairie_Warbler
48.European_Goldfinch
93.Bewick_Wren
39.Least_Flycatcher
157.Yellow_throated_Vireo
17.Cardinal
18.Spotted_Catbird
69.Rufous_sided_Towhee
38.Great_Crested_Flycatcher
88.Western_Meadowlark
70.Mourning_Warbler
199.Winter_Wren
159.Black_and_white_Warbler
31.Black_billed_Cuckoo
20.Yellow_breasted_Chat
194.Cactus_Wren
28.Brown_Creeper
98.Rock_Wren
102.Western_Wood_Pewee
135.Bank_Swallow
91.Mockingbird
22.Chuck_will_Widow
16.Painted_Bunting
105.Whip_poor_Will
63.Ivory_Gull
114.Indigo_Bunting
167.Hooded_Warbler
40.Olive_sided_Flycatcher
36.Northern_Flicker
151.Black_Capped_Vireo
55.Evening_Grosbeak
112.Great_Grey_Shrike
60.Glaucous_winged_Gull
95.Baltimore_Oriole
33.Yellow_billed_Cuckoo
137.Cliff_Swallow
62.Herring_Gull
84.Red_legged_Kittiwake
39.Scarlet_Tanager
160.Black_throated_Blue_Warbler
96.Hooded_Oriole
192.Mangrove_Cuckoo
19.Gray_Catbird
36.Barn_Swallow
146.Forsters_Tern
46.Gadwall
74.Florida_Jay
101.White_Pelican
106.Least_Auklet
73.Belted_Kingfisher
147.Least_Tern
11.Rusty_Blackbird
93.Clark_Nutcracker
82.Ringed_Kingfisher
138.Tree_Swallow
121.Heermann_Gull
109.Brewer_Blackbird
141.Artic_Tern
86.Pacific_Loon
142.Black_Tern
188.Pileated_Woodpecker
89.Hooded_Merganser
126.Bronzed_Cowbird
106.Horned_Puffin
149.Boat_tailed_Grackle
110.Red_winged_Blackbird
29.American_Crow
123.Brandt_Cormorant
130.Fish_Crow
124.Red_faced_Cormorant
107.Common_Raven
58.Pigeon_Guillemot
108.White_necked_Raven
144.Frigatebird
125.Pelagic_Cormorant
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
0.45
0.50
0.55
0.60
0.65



probability


172.Nashville_Warbler
175.Pine_Warbler
175.Pine_Warbler
048.European_Goldfinch
168.Kentucky_Warbler
169.Magnolia_Warbler
048.European_Goldfinch
169.Magnolia_Warbler
047.American_Goldfinch
168.Kentucky_Warbler
047.American_Goldfinch
006.Least_Auklet
006.Least_Auklet
001.Black_footed_Albatross
001.Black_footed_Albatross
003.Sooty_Albatross
005.Crested_Auklet
005.Crested_Auklet
002.Laysan_Albatross
059.California_Gull
059.California_Gull
002.Laysan_Albatross
0.2
0.1
0.0
−0.1
−0.2
0.48
0.50
0.52
0.54
0.56
0.58
0.60
0.62
0.64


172.Nashville_Warbler
172.Nashville_Warbler
175.Pine_Warbler
175.Pine_Warbler
048.European_Goldfinch
168.Kentucky_Warbler
169.Magnolia_Warbler
048.European_Goldfinch
169.Magnolia_Warbler
047.American_Goldfinch
168.Kentucky_Warbler
047.American_Goldfinch
006.Least_Auklet
006.Least_Auklet
001.Black_footed_Albatross
001.Black_footed_Albatross
083.Sooty_Albatross
083.Sooty_Albatross
005.Crested_Auklet
005.Crested_Auklet
002.Laysan_Albatross
059.California_Gull
059.California_Gull
002.Laysan_Albatross
0.2
0.1
0.0
−0.1
−0.2
0.40
0.45
0.50
0.55
0.60

Interpretation: Results

	classifier loss	true rank	acc@1	acc@5	acc@10
seen	4.721	4.2	0.63	0.87	0.93
random embedding	5.305	163.4	0.0	0.0	0.0
from ground truth desc.	5.305	88.2	0.0	0.0	0.0
from generated desc.	5.305	91.3	0.0	0.0	0.0

Outline

Conclusion

GÖTEBORGS UNIVERSITET

- We demonstrate that categories learned through grounded language can be mapped to the same conceptual space as those learned by direct perception.
- We propose to use NLU model as a tool for automatic **"extrinsic"** evaluation of task-oriented generation (vs. intrinsic metrics such BLEU, etc.)
- Multi-task training including generation may be beneficial for learning perceptual classifiers.

Next Steps & Future Work

GÖTEBORGS UNIVERSITET

- Multiple seen/unseen folds
- LSTM with attention over class representation, Transformers for generation
- Adding more semantic information to the model, e.g. embedding of the class label (Liang et al., 2017; Ilinykh and Dobnik, 2020)
- Using networks that do not learn to ground, but to discriminate categories from each other (Cano Santín et al., 2020)
- Use other types of texts: captions vs class descriptions (the level of details and granularity of descriptions matter)

References I

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2017. Automatic description generation from images: A survey of models, datasets, and evaluation measures (extended abstract). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4970–4974.

José Miguel Cano Santín, Simon Dobnik, and Mehdi Ghanimifard. 2020. Fast visual grounding in interaction: bringing few-shot learning with neural networks to an interactive robot. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 53–61, Gothenburg. Association for Computational Linguistics.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition*, Honolulu, United States.

Simon Dobnik and Vera Silfversparre. 2021. The red cup on the left: Reference, coreference and attention in visual dialogue. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Potsdam, Germany. SEMDIAL.

References II

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. 2021. Grounded language learning fast and slow. In *International Conference on Learning Representations*.

Nikolai Ilinykh and Simon Dobnik. 2020. When an image tells a story: The role of visual and semantic information for generating paragraph descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meet up! a corpus of joint activity dialogues in a visual environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

References III

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P. Xing. 2017. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Tzuf Paz-Argaman, Reut Tsarfaty, Gal Chechik, and Yuval Atzmon. 2020. ZEST: Zero-shot learning from text descriptions using textual similarity and visual summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 569–579, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning Deep Representations of Fine-Grained Visual Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–58, Las Vegas, NV, USA. IEEE.

References IV

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1213–1223, Berlin, Germany. Association for Computational Linguistics.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. Caltech-ucsd birds 200. Technical Report CNS-TR-2011-001, California Institute of Technology.