Reasoning about Fine-grained Attribute Phrases using Reference Games

Jong-Chyi Su* Chenyun Wu* Huaizu Jiang Subhransu Maji
University of Massachusetts, Amherst

ICCV17 International Conference on Computer Vision 2017

Motivation

Different from fixed attributes designed by experts, we collect **attribute phrases** describing differences between two images. They can be better generalized to new domains, and handle open-ended descriptions from non-expert users.

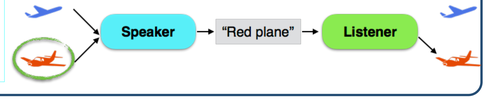
Reference Game:
We learn to **describe** and **ground** attribute phrases to images using reference game between a **speaker** and a **listener**.

Dataset:
We collect a dataset based on OID[1] dataset, containing 9400 image pairs and 5 descriptions per pair.

Facing right
In the air
Closed cockpit
White and green
Propeller spinning

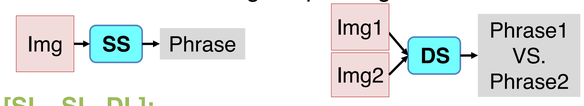
vs.
vs.
vs.
vs.
vs.

Facing left
On the ground
Open cockpit
White and blue color
Propeller stopped

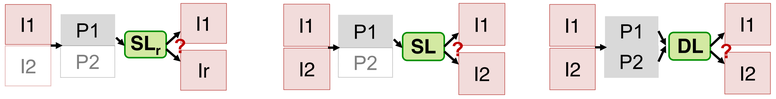


Approach

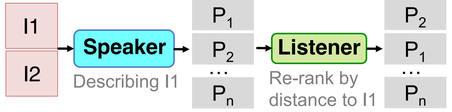
Speaker [SS, DS]:
We follow the Show-and-tell image captioning model[2].



Listener [SL_r, SL, DL]:
We train model to measure the similarity between the phrases and images in a common embedded space.



Pragmatic speaker:



Quantitative Result

→: Listener
The accuracy of picking the target image out of a pair

Input	Speaker	Listener	Val	Test
P_1	Human	SL_r	82.7	84.2
		SL	85.3	86.3
P_1 vs. P_2	Human	DL	88.7	88.9
		2×SL	89.6	89.3

		Accuracy (%)				
		SL_r		SL		Human
	Top	Test*	Test	Test*	Test	Test*
SS	1	84.0	79.8	83.0	81.7	68.0 (77.0)
	5	80.0	79.2	78.0	80.6	64.2 (74.1)
	10	78.0	78.9	76.6	80.0	61.6 (72.4)
DS	1	94.0	92.8	92.0	92.8	82.0 (88.5)
	5	91.2	90.3	91.2	91.4	80.2 (86.7)
	10	88.6	88.8	90.0	90.5	77.9 (85.0)

↑: Speaker
Use speaker models to generate phrases, evaluated by different listeners

		Human listener accuracy (%)		
		Reranker listener		
	Top	None	SL_r	SL
SS	1	68.0 (77.0)	94.0 (96.0)	87.0 (92.0)
	5	64.2 (74.1)	82.6 (88.3)	80.8 (87.1)
	7	63.1 (72.8)	74.3 (82.0)	74.3 (82.4)
DS	1	82.0 (88.5)	95.0 (96.5)	95.0 (97.0)
	5	80.2 (86.7)	90.0 (93.3)	88.6 (92.8)
	7	79.1 (85.6)	86.7 (91.5)	86.1 (91.1)

↑: Pragmatic speaker
The accuracy from human listener of the top-k phrases generated by speakers and re-ranked by listeners

Example Output

→: Speaker
DS sees two images
SS sees only the image in the green box
DS is better than SS!

Ground Truth:
1) small size **VS** large size
2) single seat **VS** more seated
3) facing left **VS** facing right
4) private **VS** commercial
5) wings at the top **VS** wings at the bottom

DS:
1) private plane **VS** commercial plane (p=0.3338)
2) private **VS** commercial (p=0.1648)
3) small plane **VS** large plane (p=0.0701)
4) facing left **VS** facing right (p=0.0355)
5) short **VS** long (p=0.0250)
6) white **VS** red (p=0.0228)
7) high wing **VS** low wing (p=0.0184)
8) small **VS** large (p=0.01775)
9) glider **VS** jetliner (p=0.0170)
10) white and blue color **VS** white red and blue color (p=0.0159)

SS:
1) no engine (p=0.2963)
2) small (p=0.1800)
3) private plane (p=0.0650)
4) on the ground (p=0.0519)
5) propellor engine (p=0.0322)
6) on ground (p=0.0250)
7) glider (p=0.0228)
8) white color (p=0.0163)
9) small plane (p=0.0151)
10) no propeller (p=0.0124)

↓: Pragmatic speaker
Re-ranking improves!

SS:
✔ passenger plane
? white
✔ jet engine
? facing right
✔ commercial plane
? _UNK
? on the ground
✔ large
✔ large size
✔ on runway

SS + SL_r:
✔ commercial plane
✔ large
✔ large size
✔ jet engine
✔ on runway
✔ passenger plane
? on the ground
? _UNK
? white
? facing right

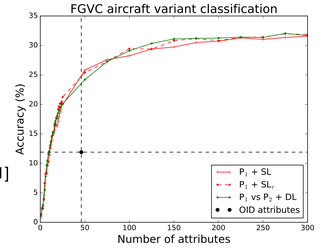
DS:
✔ commercial plane
? facing right
✔ turbofan engine
✔ on concrete
✔ t tail
✔ jet engine
✔ twin engine
✔ multi seater
✔ white and red
✔ white colour with red stripes

DS + SL_r:
✔ commercial plane
✔ jet engine
✔ turbofan engine
✔ twin engine
✔ on concrete
✔ multi seater
✔ t tail
✔ white and red
? facing right
✔ white colour with red stripes

Fine-grained Classification

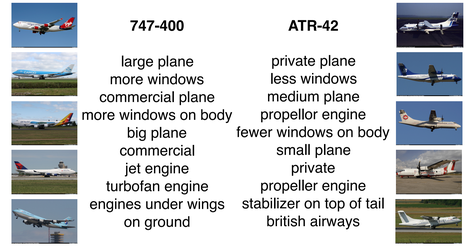
Use attributes as features for fine-grained classification on FGVC-aircraft[3] dataset:
1. Use our listener model to get scores between the *top-k most frequent attribute phrases* and the image
2. Expert-designed 46 attributes from OID[1]

Our attribute phrases outperform by **20%**.



Visualization & Application

→: Set-wise attributes



747-400	ATR-42
large plane	private plane
more windows	less windows
commercial plane	medium plane
more windows on body	propellor engine
big plane	fewer windows on body
commercial	small plane
jet engine	private
turbofan engine	propeller engine
engines under wings	stabilizer on top of tail
on ground	british airways

↓: t-SNE Embedding of attribute phrases

→: Image retrieval
Top-18 retrieved images in the test set, ranked by the listener

[1] Vedaldi et al., Understanding objects in detail with fine-grained attributes, *CVPR*, 2014.
[2] Vinyals et al., Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge, *TPAMI*, 2016.
[3] Maji et al., Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013.

School Logo here

Your title here: Maybe add some pictures and/or school logo on the left

authors and affiliation (First names of authors increase interaction potential)

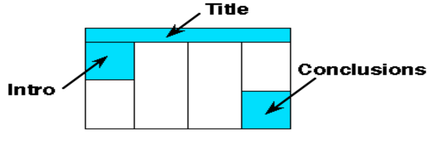
ICCV17
International Conference on Computer Vision 2017

Section 1 (sizes):

- Posters boards are 48” tall and 96” wide, but we recommend you leave a little border since you may not be able to pin at the vertical edge. Since PowerPoint does not let one define such a large paper size, this template is designed to be printed at 200%, yielding a 46” x94” poster. You can scale it up or down a bit (e.g. 42” is a common paper size at FexEd). Note there is no direct international A0.. A1 equivalent. The poster size is approximately three A0 boards next to each other, i.e., each column in this example is about one A0 board.

- Ideally you want to keep it very readable: this is not your paper, it is a poster. 32pt here (64 final printing) is good for most text:
  - Sub-bullets are 28 here (56 final)
  - Don’t use smaller than 24pt in this template (which is 48pt in final printing at 200%)
  - Insert plenty of graphics and any math you need

- When inserting graphics or equations, keep the resolution high (remember this will be printed at 200%). If you can see blocking artifacts at 400% magnification in PowerPoint, consider finding better graphics. This is an example of BAD/LOW RES GRAPHICS



- Leave enough margin for pushpin and remember many big plotters cannot get within .5” of the actual paper edge.

- You are free to use colored backgrounds and such but they generally reduce readability.

- You are free to use what ever fonts you like.
  - San Serif fonts like Arial are more readable from a distance,
  - Serif fonts like times may look more consistent with your mathematics

Section 2 (layout):

- Remember the poster session will be crowded so design the poster to be read in columns so people can read what is in front of them and move left to right to get the whole story.

- The poster should use photos, figures, and tables to tell the story of the study. For clarity, present the information in a sequence that is easy to follow.

- There is often way too much text in a poster - there definitely is in this template! Posters primarily are visual presentations; the text should support the graphics. Look critically at the layout. Some poster 'experts' suggest that if there is about 20-25% text, 40-45% graphics and 30-40% empty space, you are doing well.

Section 3:

- Include more figures than are in the paper so you can talk to them. Include things that are not in the paper and then encourage them to read the paper. Don’t try to just put all the paper here.

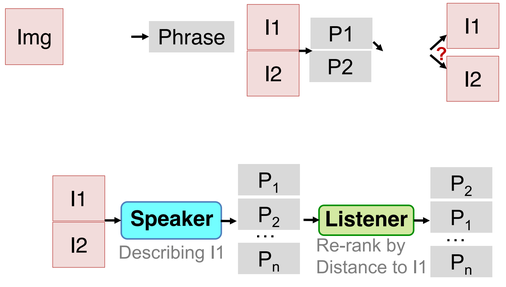
- If it looks like a cut/paste of the paper, people skip that poster since they can read the papers after the conference. Many people find it better to spend time talking with poster presenters that have more to offer than just redoing the paper content paper in big fonts.

- Remember Poster boards look like this.. This is your canvas. Paint us a picture of your work.

Maybe a QrCode to the website with your code.

ICCV17
International Conference on
Computer Vision 2017

- **SS**: sees and describes one image
- **DS**: sees and describes an image pair



- **SL$_r$**: trained on random negative images
- **SL**: trained on original image pairs as collecting the phrases
- **DL**: sees paired phrases

Embedding
We show the 1024-dimensional feature of the 500 most frequent phrases, projected into two dimensions using t-SNE.

Pragmatic speaker
We use our speaker to generate several phrases, and re-rank them with our listener model.