Reasoning about Fine-grained Attribute Phrases using Reference Games

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

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Motivation

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

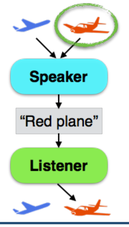
Facing right
In the air
Closed cockpit
White and green
Propeller spinning

vs.
vs.
vs.
vs.
vs.

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

We collect a dataset based on OID[1] dataset, containing 9,400 image pairs and 5 descriptions per pair.

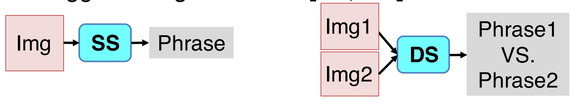
Reference Game: The speaker model describes the attribute of an image. The listener model guesses which images it is referring to.



Approach

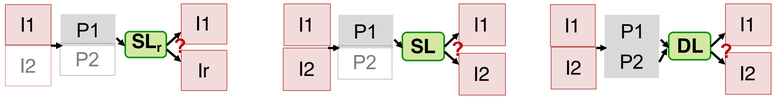
Speaker:

We follow the Show-and-tell model[2]: a LSTM image captioning model conditioned on vgg-16 image features. **[SS, DS]:**

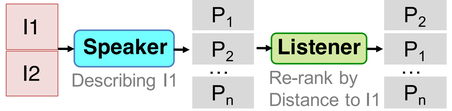


Listener:

We train a joint embedding model for images and descriptions with vgg-16 and LSTM, with triplet loss on top. **[SL$_r$, SL, DL]**:



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, report the accuracy from 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 the result!

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

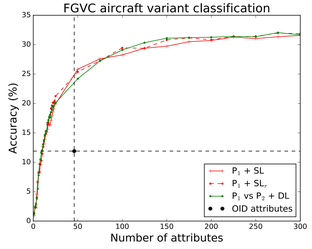
SS + SL$_r$:
✔ commercial plane
✔ large
✔ large size
✔ jet engine
✔ on runway
✔ passenger plane
? on the ground
? _UNK
? white
? facing right

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

We compare the FGVC-aircraft[3] classification result using two attributes:
Use our listener model to get scores between the top-k frequent attribute phrases and the image;
Expert-designed 46 attributes from OID[1] dataset.
Our attribute phrases outperform by **20%**.

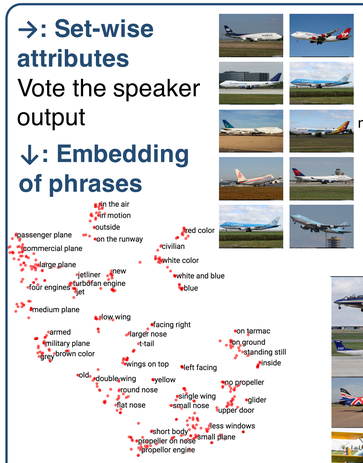


Visualization & Application

→: Set-wise attributes

Vote the speaker output

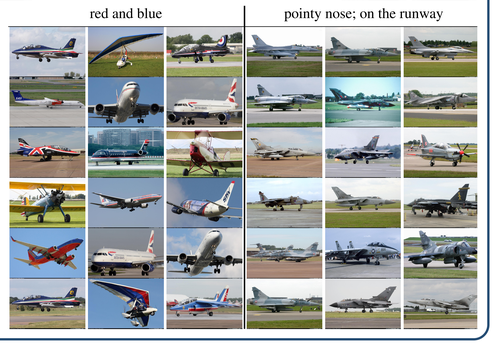
↓: Embedding of phrases



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

→: Image retrieval

Top-18 retrieved images in the test set, ranked by the listener



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