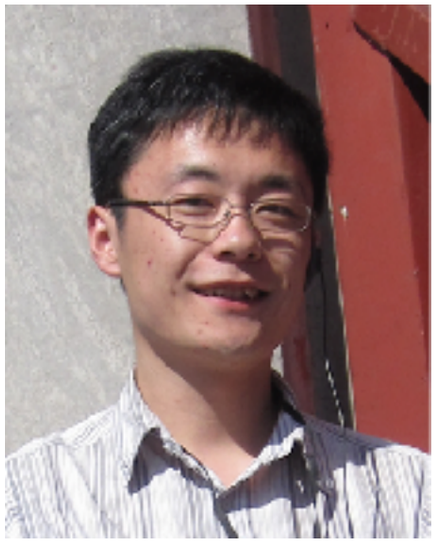
Reasoning about Fine-grained Attribute Phrases using Reference Games

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

to appear in ICCV 2017

UMASS AMHERST






Attribute

- Facing left
- Propellor plane
- Low-wing
- Small plane
- Red and white

Attribute

- Why attribute?
  - Can be used as an intermediate representation
  - Language-based
    - Communication between human and machine
    - Compositional
    - Semantically alignment

- Applications
  - Image retrieval
  - Interactive tasks

Expert-designed Attribute

- 49 attributes from OID-Aircraft [1]
  - Facing direction? west
  - Is airliner? no
  - Is propellor plane? yes
  - ……

- Are there other attributes?
- How to generalize to other domain?

[1] A. Vedaldi et al., Understanding Objects in Detail with Fine-grained Attributes, CVPR, 2014.

New Dataset - “Attribute Phrases”

- Attribute should be - **easy to describe**, but **discriminative**
- Ask **MTurkers** to describe 5 **visual differences within a pair**
  - Handle open-ended descriptions
  - Better generalize to new domain
  - More fine-grained and more discriminative


Facing right	vs.	Facing left
In the air	vs.	On the ground
Closed cockpit	vs.	Open cockpit
White and green	vs.	White and blue color
Propeller spinning	vs.	Propeller stopped

Propeller	vs.	Jet engine
Red and white body	vs.	Two-tone gray body
Flat nose	vs.	Pointed nose
In flight	vs.	Grounded
Pilot visible	vs.	No pilot visible

How to Predict Attribute Phrases?

- Train classifiers to predict expert-designed attributes
- However, attribute phrases are open-ended and not fixed
- How to better train a generative model?

Reference Game

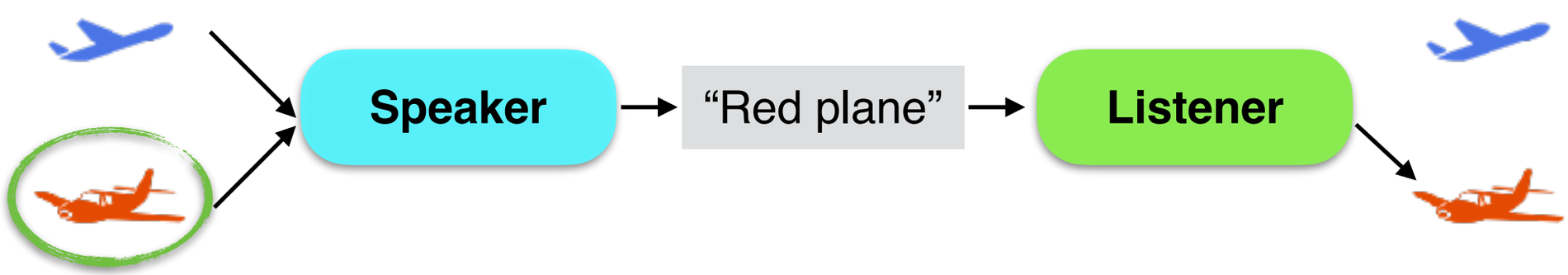
- Refer It Game[1]
- RefCOCO[2]

Generation

guy in yellow dirbbling ball
yellow shirt and black shorts
yellow shirt in focus

Comprehension

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



[1] Kazemzadeh et al. "ReferItGame: Referring to Objects in Photographs of Natural Scenes” EMNLP 2014.
[2] Yu et al. "Modeling Context in Referring Expressions" ECCV 2016.

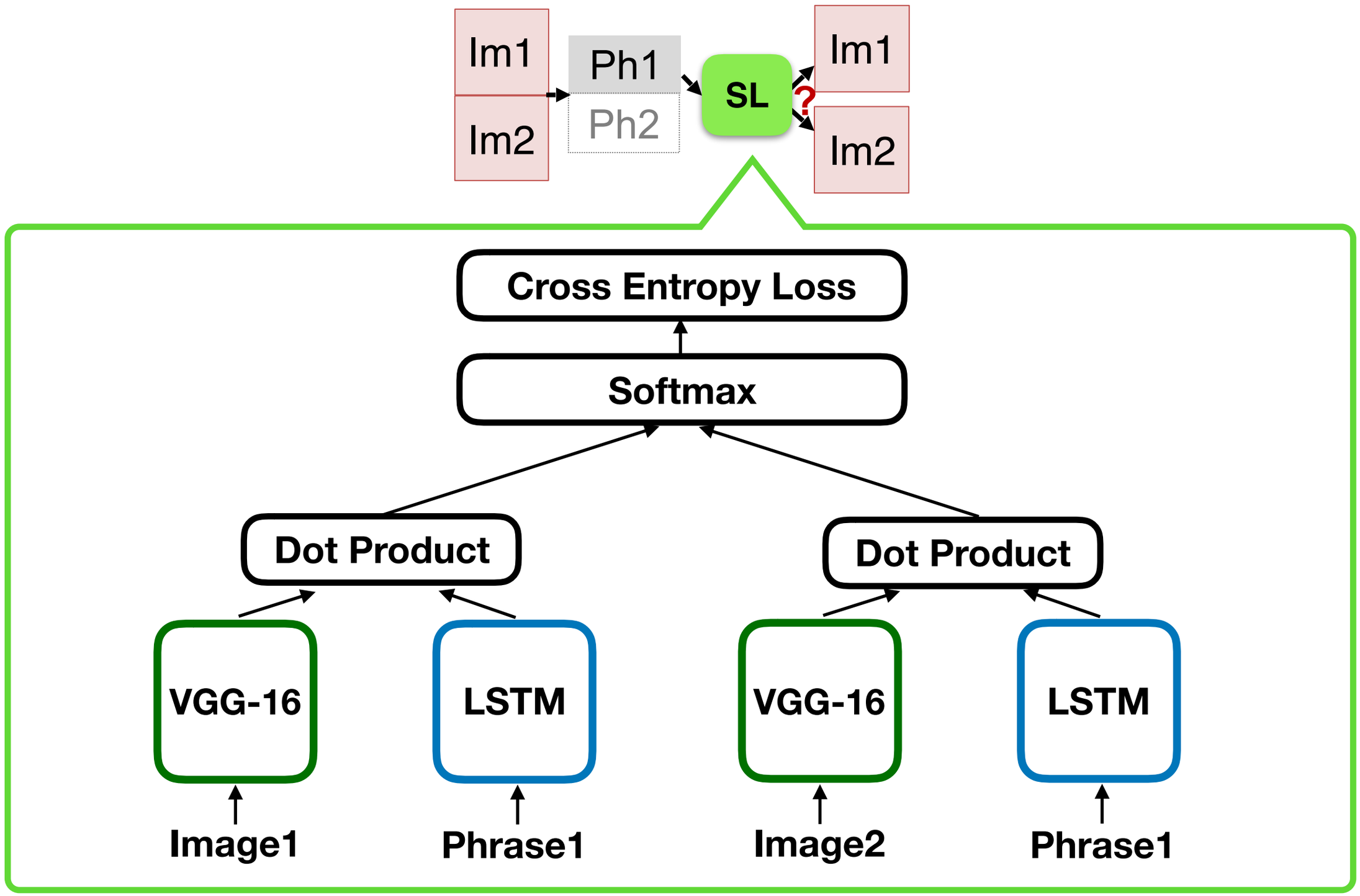
Use Listener for Comprehension Task

- *Task*: Given an attribute phrase and two images, decide which image it is referring to.

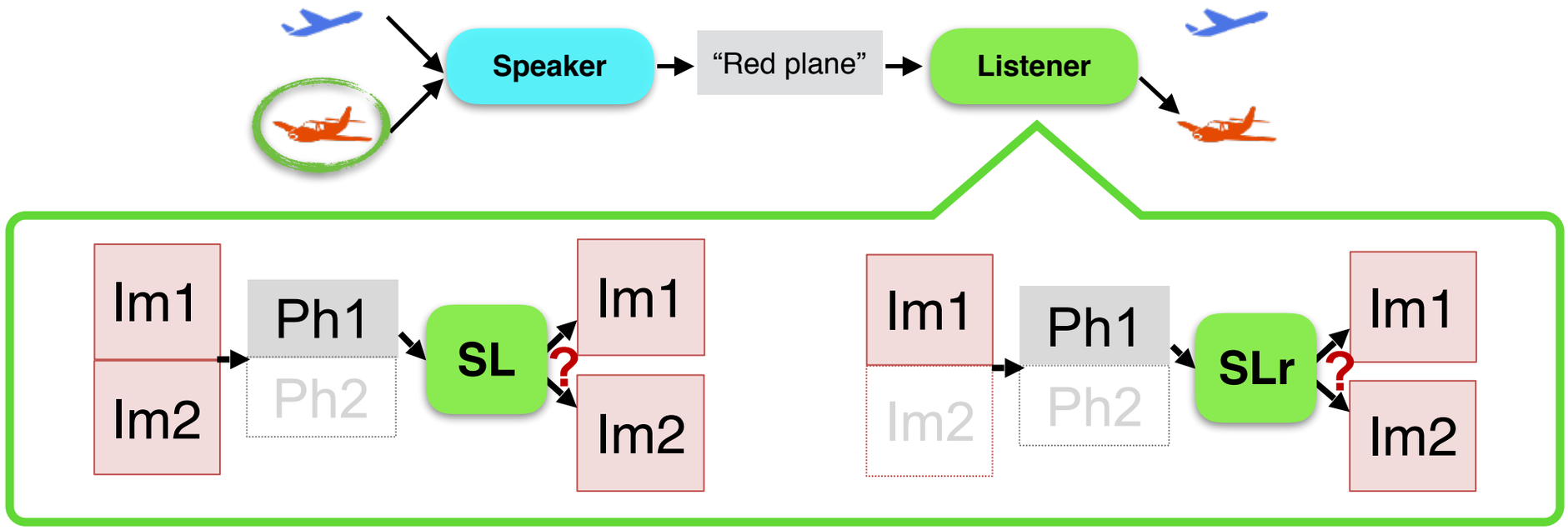
- *Goal*: Ground attribute phrases to images.

- *Method*: Measure the similarity between the phrase(s) and images in a common embedded space.

Listener Model



Listener Model



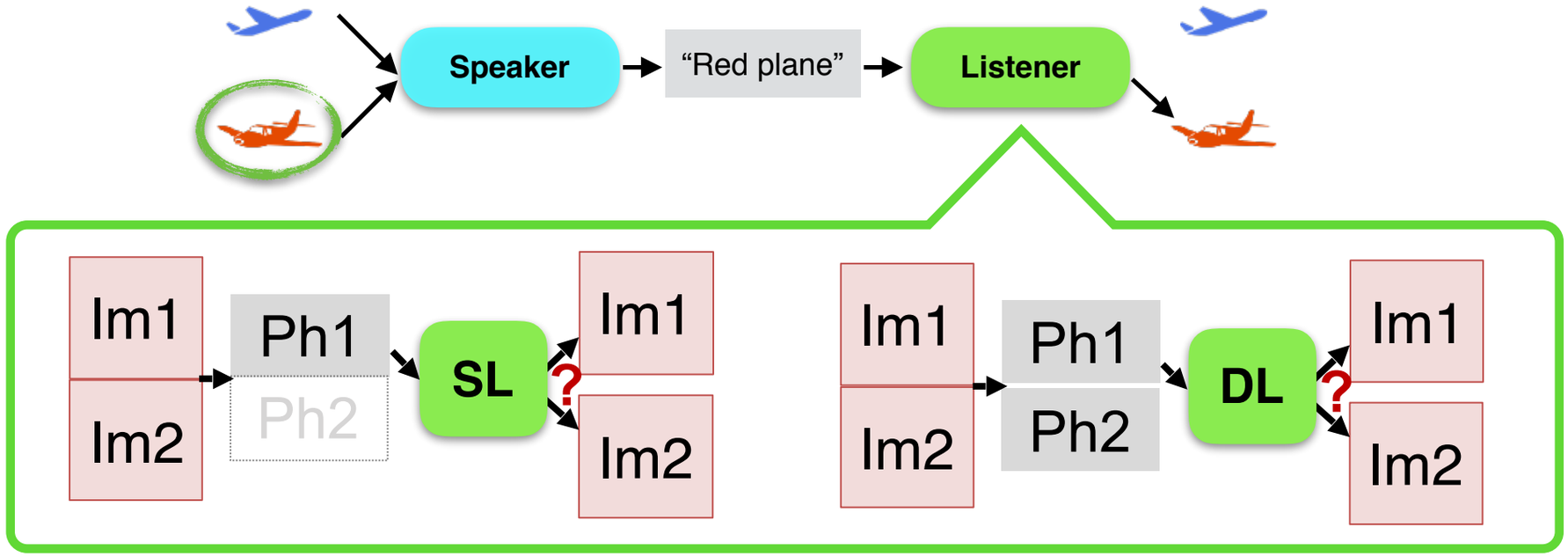
- Simple Listener (**SL**)
- Simple Listener (**SLr**) - trained on non-contrastive data

Input	Speaker	Listener	Test
P_1	Human	SL_r	84.2
		SL	86.3

Evaluate listeners using human-generated phrases

Contrastive data helps!

Listener Model

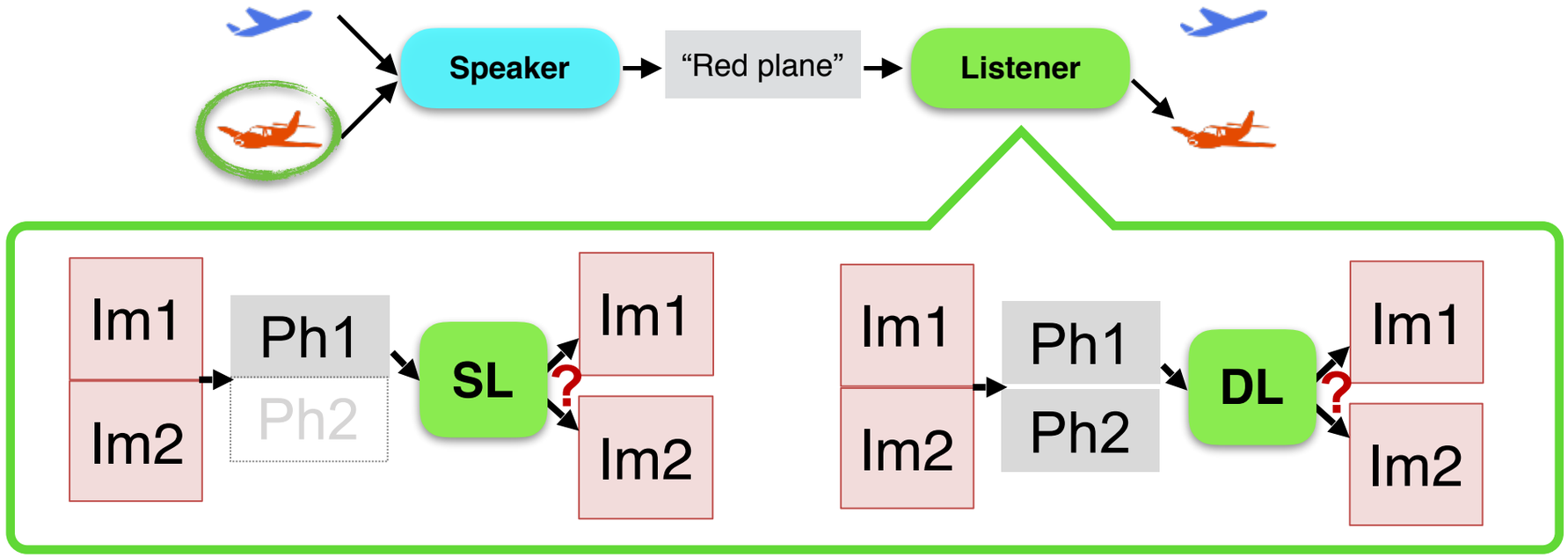


- Discerning Listener (**DL**) - given two phrases

Input	Speaker	Listener	Test
P_1	Human	SL_r	84.2
		SL	86.3
P_1 vs. P_2	Human	DL	88.9

2.6% better

Listener Model

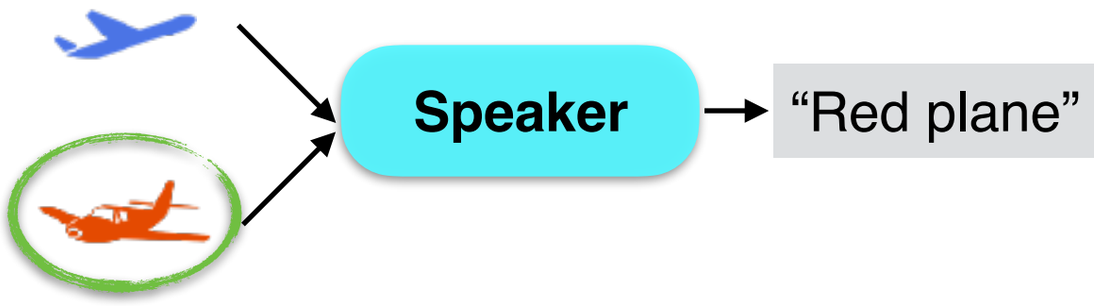


- Discerning Listener (**DL**) - given two phrases

Input	Speaker	Listener	Test
P_1	Human		
P_1 vs. P_2	Human	DL	88.9
		2×SL	89.3

DL is similar to averaging the predictions from two SL (2 x SL)

Use Speaker for Generation Task

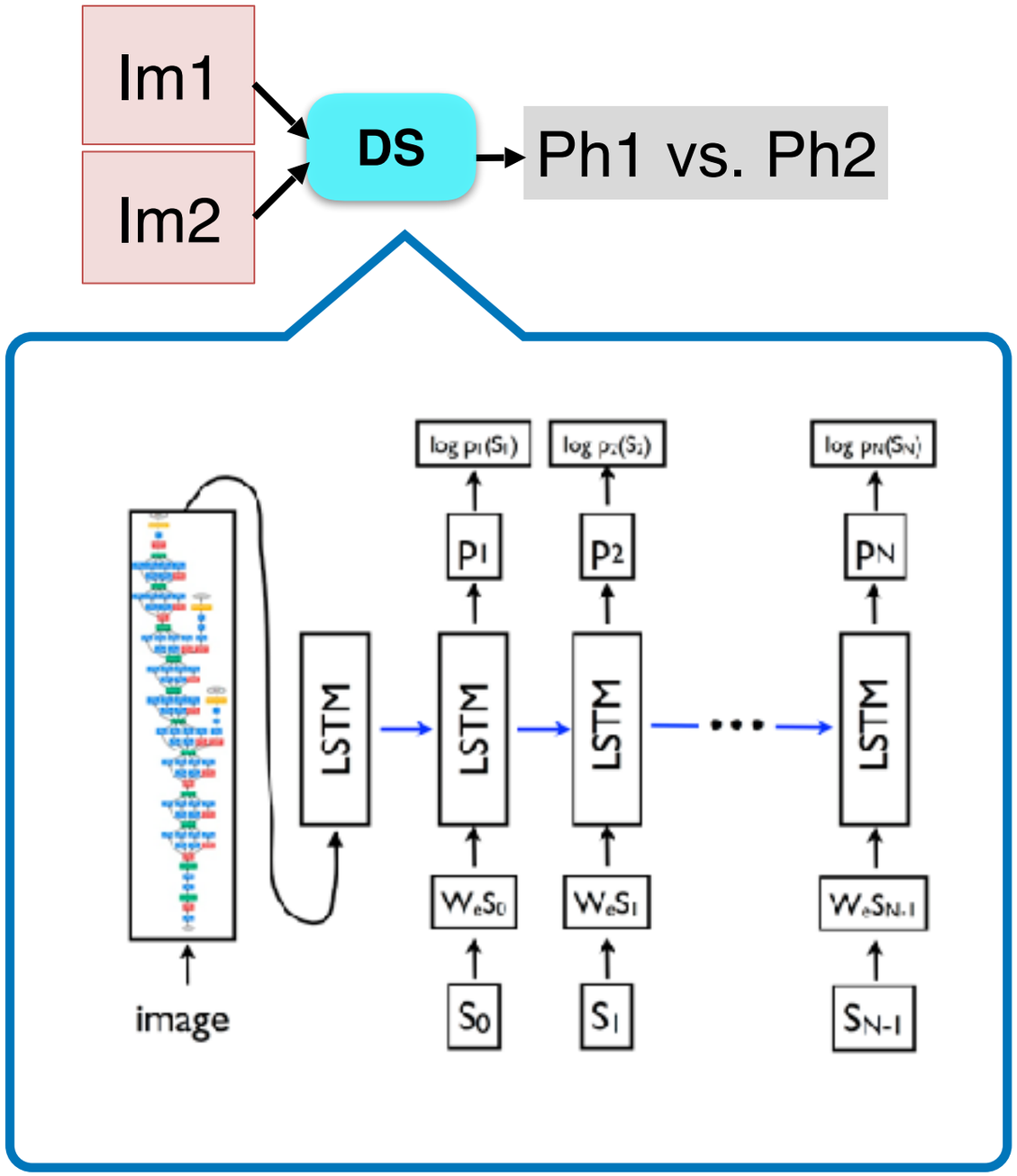


- *Task*: Given two images, generate attributes.

- *Goal*: The description must be **discriminative**.

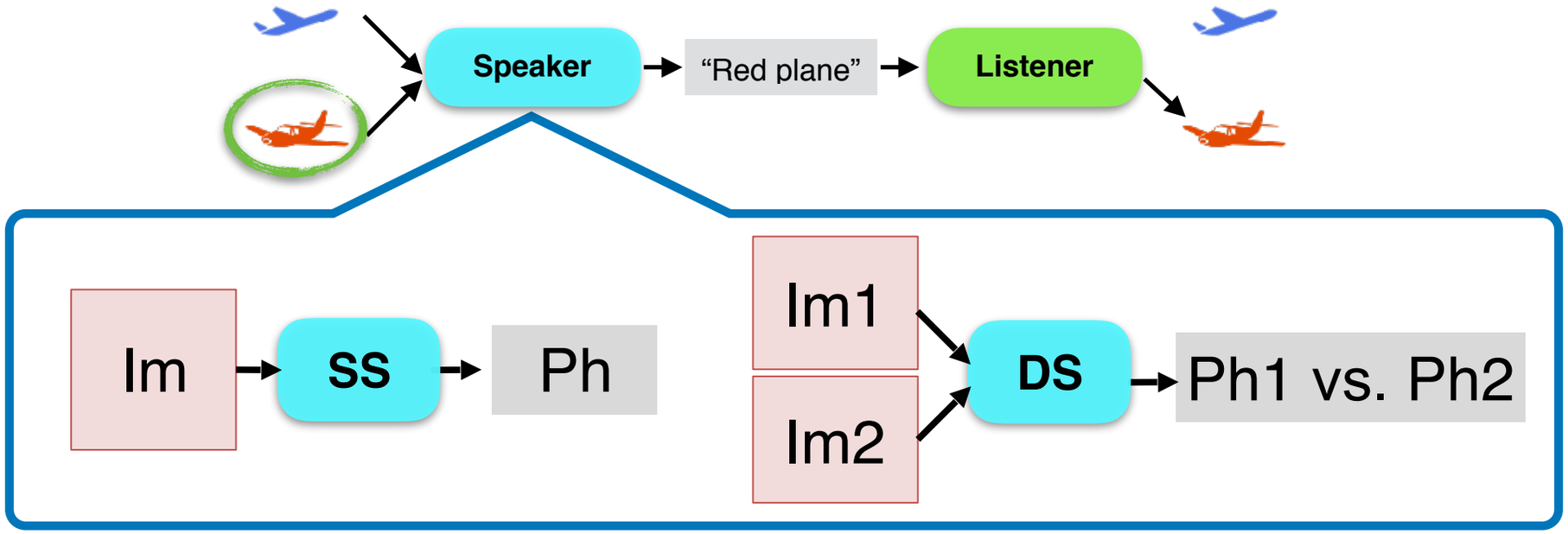
- *Method*: Use image captioning model as the speaker.

Image Captioning Model



O. Vinyals et al., Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge, TPAMI, 2016.

Use Speaker to Generate Attribute Phrases



- Simple Speaker (**SS**)
- Discerning Speaker (**DS**)

	Top	Accuracy (%) SL_r	Accuracy (%) SL
SS	1		
	5		
	10		
DS	1		
	5		
	10		

- Use our listener models to evaluate the quality of the descriptions.
- DS is ~10% better than SS.
- Use human to evaluate.
  - Given a phrase, user pick the image among two, or say “not sure”
  - More strict.

Attribute Phrases Generated by Speakers

Ground Truth: (Human generated)
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

SS:
1) no engine
2) small
3) private plane
4) on the ground
5) propellor engine
6) on ground
7) glider
8) white color
9) small plane
10) no propeller

DS:
1) private plane **VS** commercial plane
2) private **VS** commercial
3) small plane **VS** large plane
4) facing left **VS** facing right
5) short **VS** long
6) white **VS** red
7) high wing **VS** low wing
8) small **VS** large
9) glider **VS** jetliner
10) white and blue color **VS** white red and blue color

Pragmatic Speaker Helps

- Pragmatic speaker[1]:
  - Use speaker to generate attribute phrases
  - Use the score from listener to re-rank them

SS:	SS + SL_r:	DS:	DS + SL_r:
✔ passenger plane	✔ commercial plane	✔ commercial plane	✔ commercial plane
? white	✔ large	? facing right	✔ jet engine
✔ jet engine	✔ large size	✔ turbofan engine	✔ turbofan engine
? facing right	✔ jet engine	✔ on concrete	✔ twin engine
✔ commercial plane	✔ on runway	✔ t tail	✔ on concrete
? _UNK	✔ passenger plane	✔ jet engine	✔ multi seater
? on the ground	? on the ground	✔ twin engine	✔ t tail
✔ large	? _UNK	✔ multi seater	✔ white and red
✔ large size	? white	✔ white and red	? facing right
✔ on runway	? facing right	✔ white colour with red stripes	✔ white colour with red stripes

[1] Andreas et al., “Reasoning About Pragmatics with Neural Listeners and Speakers”, EMNLP, 2016

Pragmatic Speaker Helps

- Use human listener for evaluation:
  - Given a phrase, user pick the image among two, or say “not sure”.

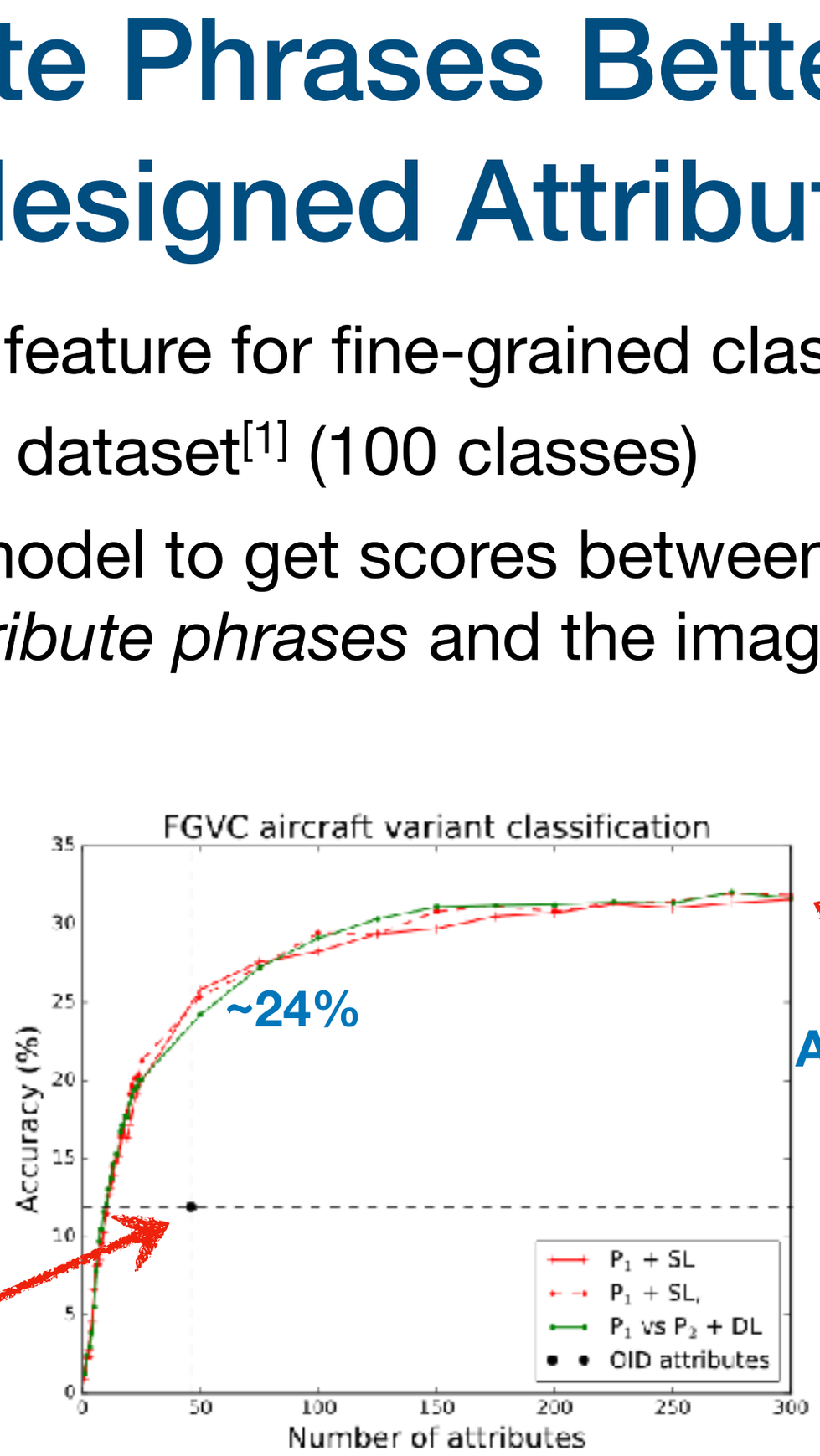
		Human listener accuracy (%)		
		Reranker listener		
	Top	None	SL_τ	SL
DS	1	82.0 (88.5)	95.0 (96.5)	95.0 (97.0)
	5	80.2 (86.7)	90.0 (93.3)	88.6 (92.8)
	7	79.1 (85.6)	86.7 (91.5)	86.1 (91.1)

DS improves ~10%

Pragmatic improves ~10% on top-5 accuracy

Are Attribute Phrases Better than Expert-designed Attributes?

- Use attribute as the feature for fine-grained classification
  - On FGVC-Aircraft dataset[1] (100 classes)
  - Use our listener model to get scores between the *top-k most frequent attribute phrases* and the image
    - 46 OID attributes
- ~20% improvement



[1] S, Maji et al., Fine-grained Visual Classification of Aircraft. arXiv:1306.5151, 2013.

Image Retrieval Using Listener

- *Query*: attribute phrases
- For each test image and the query phrase, get scores by listener model
- Top 18 images ranked by the score

Image Retrieval Using Listener

- *Query*: attribute phrases
- For each test image and the query phrase, get scores by listener model
- Top 18 images ranked by the score

Generate Attribute for Sets

- Select two categories (A,B), generate attributes for randomly selected image pairs ($Im_1 \in A$, $Im_2 \in B$)
- Sort them by frequency

747-400

large plane
more windows
commercial plane
more windows on body
big plane
commercial
jet engine
turbofan engine
engines under wings
on ground

ATR-42

private plane
less windows
medium plane
propellor engine
fewer windows on body
small plane
private
propeller engine
stabilizer on top of tail
british airways

Embedding Space of Attribute Phrases

- Visualize the space of attribute phrases using the embedding of the listener model
- Projected to 2 dimensions using t-SNE

Thank you!