Reasoning about Fine-grained Attribute Phrases using Reference Games

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

ICCV 2017
Expert-designed Attributes

- ✔ Modular - an instance can be described by a set of attributes
- ✗ A fixed set of attributes designed by experts before collecting the dataset (49 attributes from OID-Aircraft [1])

- Is military plane? No
- Is propellor plane? No

Image Captions

A large Air France jet sitting on top of a runway.

⇒ Usually a longer sentence describing many aspects
✔ Compositional language-based
✘ Not designed to describe differences between a pair of images
Image Captions

A large airplane on a runway.  

A large Air France jet sitting on top of a runway.

⇒ Usually a longer sentence describing many aspects

✅ Compositional language-based

❌ Not designed to describe differences between a pair of images
New Dataset - “Attribute Phrases”

- Short phrases describing **visual differences** within a pair of images sampled from different categories
- 9400 image pairs in total

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

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

- **Propeller**
  - Red and white body
  - Flat nose
  - In flight
  - Pilot visible

- **Jet engine**
  - Two-tone gray body
  - Pointed nose
  - Grounded
  - No pilot visible

✔ Modular like attributes
✔ Compositional and free-form like image captions
✔ More expressive and discriminative at fine-grained level
Attribute Phrases

• How to generate?
  
  “Blue plane vs. Red plane”

• How to evaluate?
  
  “Red plane”

• Use reference game
Reference Game

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

• Refer to a specific object in an image
• Usually focus on the category, spatial relationship etc.

• Our task focuses on attributes that enable fine-grained discrimination with instances of a category

Overview of Our Model

- Generation task - **speaker** model
- Comprehension task - **listener** model

1. Train the speaker and listener model separately
2. Use the listener model to evaluate the speaker model
3. Rerank phrases by the listener, then evaluate by human
Use Listener Model for Comprehension Task

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

- **Method**: Measure the similarity between the attribute phrase and images in a common embedded space.
Use Speaker Model for Generation Task

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

- **Method**: Use the image captioning model\(^1\) as the speaker model

\(^1\) Vinyals et al., Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge, *TPAMI*, 2016.
Variances of the Speaker Model

- **Simple Speaker (SS):** Given one image, generate one phrase
- **Discerning Speaker (DS):** Given two images, generate a pair of phrases

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Top</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>1</td>
<td>81.7</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>80.6</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>80.0</td>
</tr>
<tr>
<td>DS</td>
<td>1</td>
<td>92.8</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>91.4</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>90.5</td>
</tr>
</tbody>
</table>

- Use the listener model to evaluate the quality of the generated phrases
- DS generates better attribute phrases than SS
Discerning Speaker Generate Better Phrases

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

Some phrases are correct but not discriminative

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

1. Use speaker to generate attribute phrases
2. Re-rank the phrases by the scores from the listener model
   ⇒ More discriminative phrases on the top

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

SS + Re-ranking:
- 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
- t tail
- white and red
- facing right

DS + Re-ranking:
- commercial plane
- jet engine
- turbofan engine
- twin engine
- on concrete
- multi seater
- t tail
- white and red
- facing right

# Pragmatic Speaker Helps

- Use **human** listener for evaluation:
  - Given a attribute phrase, let users choose the image among two

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Top</th>
<th>Original Acc. (%)</th>
<th>After Re-ranking Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discerning Speaker</td>
<td>1</td>
<td>82.0</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>80.2</td>
<td>90.0</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>79.1</td>
<td>86.7</td>
</tr>
</tbody>
</table>

Re-ranking improves ~10% on top-5 accuracy
Are Attribute Phrases Better than Expert-designed Attributes?

- Use attribute as the feature for fine-grained classification task
- Use our listener model to get the scores between the image and the top-k most frequent attribute phrases
- Use expert-designed 46 attributes from OID dataset
- Test on FGVC-Aircraft dataset\(^1\) (100 classes)
- \(~20\%\) improvement

Generate Attribute for Sets

- Select two categories (A, B), generate attribute phrases for randomly selected image pairs (Im_1 ∈ A, Im_2 ∈ 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
Use the Listener Model for Image Retrieval

- **Query**: attribute phrase(s)

- Get scores of the query phrase and test images by the listener model

- We show top 18 images ranked by the scores
t-SNE Embeddings of Attribute Phrases from the Listener Model

Large commercial planes
- passenger plane
- side door
- commercial
- commercial plane
- more windows
- many windows
- multiple windows
- large plane
- bigger in size
- double engine
- twin engine

Military planes
- armed
- fighter jet
- fighter plane
- military jet
- military plane
- war plane
- air force plane
- camouflage color
- green
- grey
- gray color
- brown color

- passenger
- commercial plane
- large plane
- in the air
- outside
- red color
- civilian
- white color
- white and blue
- blue
- three wheels
- low wing
- facing right
- on ground
- inside
- no propeller
- glider
- wings on top
- single wing
- one window
- front engine
- small plane
- propeller engine
Thank you!

Dataset and Code are available at: