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
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?

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

<table>
<thead>
<tr>
<th>Facing right</th>
<th>Facing left</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the air</td>
<td>On the ground</td>
</tr>
<tr>
<td>Closed cockpit</td>
<td>Open cockpit</td>
</tr>
<tr>
<td>White and green</td>
<td>White and blue color</td>
</tr>
<tr>
<td>Propeller spinning</td>
<td>Propeller stopped</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Propeller spinning</th>
<th>Jet engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red and white body</td>
<td>Two-tone gray body</td>
</tr>
<tr>
<td>Flat nose</td>
<td>Pointed nose</td>
</tr>
<tr>
<td>In flight</td>
<td>Grounded</td>
</tr>
<tr>
<td>Pilot visible</td>
<td>No pilot visible</td>
</tr>
</tbody>
</table>
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\textsuperscript{[1]}
- RefCOCO\textsuperscript{[2]}

We learn to \textbf{describe} and \textbf{ground} attribute phrases to images using reference game between a \textbf{speaker} and a \textbf{listener}.

\textsuperscript{[1]} Kazemzadeh et al. "ReferItGame: Referring to Objects in Photographs of Natural Scenes" EMNLP 2014.
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

Cross Entropy Loss

Dot Product

VGG-16
Image1

LSTM
Phrase1

VGG-16
Image2

LSTM
Phrase1
Listener Model

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

<table>
<thead>
<tr>
<th>Input</th>
<th>Speaker</th>
<th>Listener</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>Human</td>
<td>$SL_r$</td>
<td>84.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$SL$</td>
<td>86.3</td>
</tr>
</tbody>
</table>

Evaluate listeners using human-generated phrases

Contrastive data helps!
**Listener Model**

- Discerning Listener (DL) - given two phrases

---

<table>
<thead>
<tr>
<th>Input</th>
<th>Speaker</th>
<th>Listener</th>
<th>Test</th>
<th>2.6% better</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>Human</td>
<td>$SL_r$</td>
<td>84.2</td>
<td></td>
</tr>
<tr>
<td>$P_1$ vs. $P_2$</td>
<td>Human</td>
<td>DL</td>
<td>88.9</td>
<td>2.6% better</td>
</tr>
</tbody>
</table>
 Listener Model

- Discerning Listener (DL) - given two phrases

<table>
<thead>
<tr>
<th>Input</th>
<th>Speaker</th>
<th>Listener</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>Human</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_1$ vs. $P_2$</td>
<td>Human</td>
<td>DL 2×SL</td>
<td>88.9</td>
</tr>
</tbody>
</table>

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

Use Speaker to Generate Attribute Phrases

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

Use our listener models to evaluate the quality of the descriptions.

DS is ~10% better than SS.
Use Speaker to Generate Attribute Phrases

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

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SL&lt;sub&gt;T&lt;/sub&gt;</td>
</tr>
<tr>
<td>Top</td>
<td>Test*</td>
</tr>
<tr>
<td>SS</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>DS</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>

- Use human to evaluate.
  - Given a phrase, user pick the image among two, or say “not sure”
  - More strict.
- DS is ~10% better than SS.
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\textsuperscript{[1]}:
  • Use speaker to generate attribute phrases
  • Use the score from listener to re-rank them

\textbf{SS:}
- passenger plane
- \textcolor[HTML]{000000}{white}
- \textcolor[HTML]{000000}{jet engine}
- \textcolor[HTML]{000000}{facing right}
- \textcolor[HTML]{000000}{commercial plane}
- ? _UNK
- ? on the ground
- \textcolor[HTML]{000000}{large}
- \textcolor[HTML]{000000}{large size}
- \textcolor[HTML]{000000}{on runway}

\textbf{SS + SLr:}
- \textcolor[HTML]{000000}{commercial plane}
- \textcolor[HTML]{000000}{large}
- \textcolor[HTML]{000000}{large size}
- \textcolor[HTML]{000000}{jet engine}
- \textcolor[HTML]{000000}{on runway}
- \textcolor[HTML]{000000}{passenger plane}
- ? on the ground
- ? _UNK
- ? \textcolor[HTML]{000000}{white}
- ? \textcolor[HTML]{000000}{facing right}

\textbf{DS:}
- \textcolor[HTML]{000000}{commercial plane}
- ? \textcolor[HTML]{000000}{facing right}
- \textcolor[HTML]{000000}{turbofan engine}
- \textcolor[HTML]{000000}{on concrete}
- \textcolor[HTML]{000000}{t tail}
- \textcolor[HTML]{000000}{jet engine}
- \textcolor[HTML]{000000}{twin engine}
- \textcolor[HTML]{000000}{multi seater}
- \textcolor[HTML]{000000}{white and red}

\textbf{DS + SLr:}
- \textcolor[HTML]{000000}{commercial plane}
- \textcolor[HTML]{000000}{jet engine}
- \textcolor[HTML]{000000}{turbofan engine}
- \textcolor[HTML]{000000}{twin engine}
- \textcolor[HTML]{000000}{on concrete}
- \textcolor[HTML]{000000}{multi seater}
- \textcolor[HTML]{000000}{t tail}
- \textcolor[HTML]{000000}{white and red}

\textsuperscript{[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”.

<table>
<thead>
<tr>
<th></th>
<th>Human listener accuracy (%)</th>
<th>Reranker listener</th>
<th>SL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Top</td>
<td>None</td>
</tr>
<tr>
<td>Top 1</td>
<td>82.0 (88.5)</td>
<td>95.0 (96.5)</td>
<td>95.0 (97.0)</td>
</tr>
<tr>
<td>DS 5</td>
<td>80.2 (86.7)</td>
<td>90.0 (93.3)</td>
<td>88.6 (92.8)</td>
</tr>
<tr>
<td>DS 7</td>
<td>79.1 (85.6)</td>
<td>86.7 (91.5)</td>
<td>86.1 (91.1)</td>
</tr>
</tbody>
</table>

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

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 (Im₁ ∈ A, Im₂ ∈ 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!