**Spoken Attributes: Mixing Binary and Relative Attributes to Say the Right Thing**

Amir Sadovnik\(^1\)
Andrew Gallagher\(^1\)
Devi Parikh\(^2\)
Tsuhan Chen\(^1\)

\(^1\)Cornell University  \(^2\)Virginia Tech

---

### Motivation

- Most previous work has treated attributes as either binary or relative.
- However, this does not model the actual way humans use these attributes in descriptive statements of an image.
- Each attribute might be used as relative or binary depending on situational context.
- We model this context for each attribute and learn when it is appropriate to use binary vs. relative.

Examples of two different attributes. Smiling is typically used in relative form when both people are smiling, but "has glasses" attribute is never used in relative form.

### Understanding Spoken Attributes

For 600 images of pairs of people we collect the following ground truth data:

- **Binary ground truth for each face:**
  - A has/does not have attribute
  - B has/does not have attribute

- **Relative ground truth for each pair:**
  - A has attribute more than B
  - B has attribute more than A
  - A and B have attribute equally

- **Spoken Attribute ground truth for pair:**
  - A and B have attribute
  - A and B do not have attribute
  - A has attribute and B does not
  - B has attribute and A does not
  - A has attribute more than B
  - B has attribute more than A

- Binary expressions cannot express relative relationships when appropriate
- Relative expression "force" users to select a relative statement inappropriately
- Appropriate usage of relative and binary statement are attribute-dependent

### Our Approach

- We build a classifier, that selects the **most accurate statement** to describe a pair of faces.
- We model this as a multiclass classifier whose output is one of 6 possible statements (4 binary, 2 relative).
- We use a 2-layer classifier, where the first layer consists of a binary SVM and a rank SVM trained on binary and relative data.

### Reading Between the Lines

Given a spoken attribute, can we infer the correct binary statement?

- Some attributes are only used in relative form when both people have the attribute.

### Searching for the Right Image

Given a description, can we return better search results using our spoken attributes classifier score?

### Prediction Results

We use leave-one-out cross validation and compare with the following baselines:

- **Raw Feature:** Skip first layer classifiers, and concatenate raw feature vectors of pair.
- **Difference:** skip first layer classifiers, and take difference of raw feature vectors of pair.
- **Relative:** can only output relative statement (or majority binary statement if below optimal threshold).
- **Binary:** Can only output a binary statement

---

**Examples**

- **Examples of two different attributes.** Smiling is typically used in relative form when both people are smiling, but "has glasses" attribute is never used in relative form.

---

\[^{1}\] 21504 feature vector composed of densely sampled SIFT descriptors at different grid cells and scales.