Overview

Problem
- Parts and attributes exhibit weaknesses
  - Scalability issues; costly; reliance on experts, but experts are scarce

Proposed Solution
- Use relative similarity comparisons to reduce dependence on expert-derived part and attribute vocabularies

 Contributions
- We present an efficient, flexible, and scalable system for interactive fine-grained visual categorization
  - Based on perceptual similarity
  - Combines similarity metrics and computer vision methods in a unified framework
- Outperforms state-of-the-art relevance feedback-based and part/attribute-based approaches

Approach

INTERACTIVE CATEGORIZATION
- Compute per-class probabilities as:
  \[ p(c|x, u_j) = \int p(c|x, u_j|x) \, dx \]
  where
  \[ w^*_j = p(c|x, u_j|x) = p(0|u_j|x) p(c|x) \]

Efficient computation
- Approximate per-class probabilities as:
  \[ p(c|x, u_j) = \sum_{i=1}^k w_i \]
  \[ i.e. \text{sum of weights of examples of class } c \text{ where } k \text{ enumerates training examples} \]
- Weight \( w_i \) represents how likely \( x_i \) is true location \( x \) such that
  \[ w_i = \int p(c|x, u_j)|x = x_i| \, dx \]
  where
  \[ w_i^* = \sum_{i=1}^k w_i \]

Learning a Metric
- Given set of triplet comparisons \( \mathcal{T} \), learn embedding \( Z \) of \( \mathbb{N} \) training images with stochastic triplet embedding (van der Maaten & Hinton 2009)
- From \( Z \), generate similarity matrix \( S \in \mathbb{N} \times \mathbb{N} \)

Computer Vision
- Easy to map off the shelf CV algorithms into framework, e.g., multiclass classification scores
  \[ p(c|x) = \frac{1}{\sum_{i=1}^k w_i} \]

Incorporating Users
- \( D \) is grid of images for each question
- Incorporate independent user response as:
  \[ p(c|x) = \frac{1}{\sum_{i=1}^k w_i} \]

Selecting the Display
- Approximate solution: maximizes expected information gain in terms of entropy of \( p(c|x, u_j|x) \)

Multiple Metrics
- System supports multiple similarity metrics as different types of questions
- Simulate perceptual spaces using CUB-200-2011 attribute annotations

Qualitative Results
- With computer vision
- No computer vision

Results

Learned Embedding
- Learn category-level embedding of \( \mathbb{N} \approx 200 \) nodes
- Category-level embedding requires much fewer comparisons compared to the instance-level

Interactive Categorization
- Similarity comparisons are advantageous compared to part/attribute questions
- Using computer vision reduces the burden on the user
- Intelligently selecting image displays reduces effort
- The system is robust to user noise