

Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise

Jacob Whitehill, Paul Ruvolo, Ting-fan Wu, Jacob Bergsma, and Javier Movellan – Machine Perception Laboratory, University of California San Diego (UCSD)

{jake,paul,ting,jbergsma,movellan}@mplab.ucsd.edu

Abstract

Modern machine learning-based approaches to computer vision require very large databases of hand labeled images. Some contemporary vision systems already require on the order of millions of images for training (e.g., Omron face detector). New Internet-based services allow for a large number of labelers to collaborate around the world at very low cost. However, using these services brings interesting theoretical and practical challenges: (1) The labelers may have wide ranging levels of expertise which are unknown a priori, and in some cases may be adversarial; (2) images may vary in their level of difficulty; and (3) multiple labels for the same image must be combined to provide an estimate of the actual label of the image. Probabilistic approaches provide a principled way to approach these problems. In this paper we present a probabilistic model and use it to simultaneously infer the label of each image, the expertise of each labeler, and the difficulty of each image. On both simulated and real data, we demonstrate that the model outperforms the commonly used “Majority Vote” heuristic for inferring image labels, and is robust to both noisy and adversarial labelers.

Quality Control of Dataset Labels

Machine learning systems require many thousands, if not millions, of labeled data. Classifier accuracy may be highly dependent on the accuracy of the labels.

Traditional approach: Assess labeler accuracy on a set of *pre-labeled* instances, and reject labelers whose accuracy is below threshold. Problems:

1. Re-labeling data with known labels is wasteful.
2. No clear rule to select a threshold.
3. How to weight the opinions of multiple labelers who are all above threshold?

Proposed approach: We present a probabilistic model for optimal inference of data labels from multiple labelers which overcomes these problems. The model simultaneously infers:

- The **labels** of the data.
- The **abilities** of the labelers.
- The **difficulty** of the data instances.
- The model also implicitly detects **adversarial labelers** and flips their labels accordingly.

The model is **GLAD** (Generative Model of Labels, Abilities, and Difficulties).

Assumptions

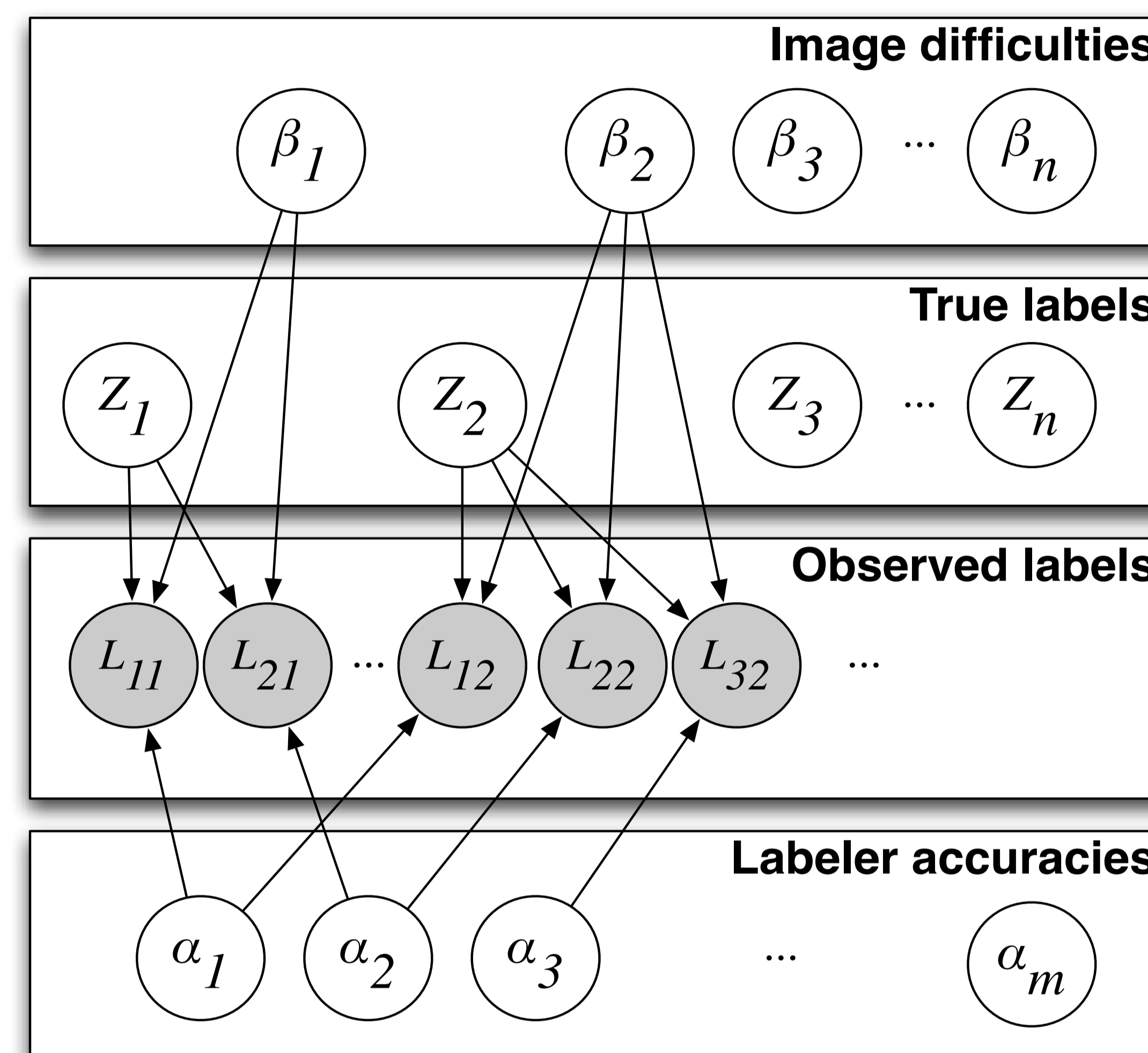
We focus on binary classification, but the model can be straightforwardly extended to handle multiple (1 of K) classes. We assume that the probability of labeler i correctly labeling instance j is given by a logistic function

$$p(\text{correct}) = \sigma(\alpha_i \beta_j) = \frac{1}{1 + e^{-\alpha_i \beta_j}}$$

where α_i is the **ability** of labeler i , and β_j is the **difficulty** of instance j .

- **Large** α means the labeler has a high ability (discriminates with high accuracy).
- **Large negative** α means the labeler has a high ability but is **adversarial**.
- **Small** β means the instance is very difficult.

Graphical Model



Maximum Likelihood Solution

We use Expectation-Maximization (EM) to compute the maximum-likelihood solution of $p(\mathbf{l} | \boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\alpha}, \boldsymbol{\beta}$. For the \mathbf{Z} (instance labels), we take the values of \mathbf{Z} at the last E-Step.

E-Step:

$$\begin{aligned} p(z_j | \mathbf{l}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= p(z_j | \mathbf{l}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}_j) \\ &\propto p(z_j | \boldsymbol{\alpha}, \boldsymbol{\beta}_j) p(\mathbf{l}_j | z_j, \boldsymbol{\alpha}, \boldsymbol{\beta}_j) \\ &\propto p(z_j) \prod_i p(l_{ij} | z_j, \alpha_i, \beta_j) \end{aligned}$$

M-Step:

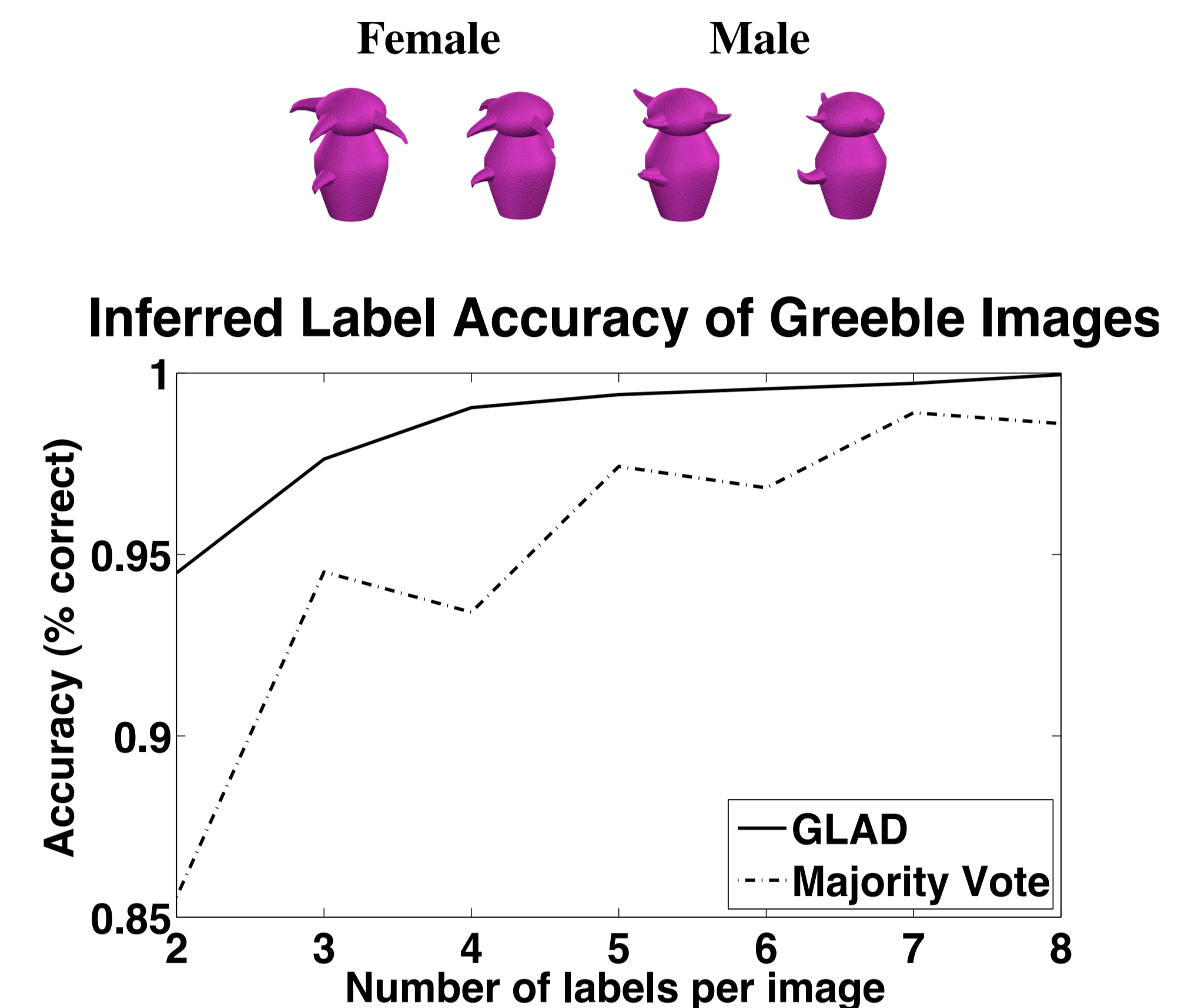
$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= E[\ln p(\mathbf{l}, \mathbf{z} | \boldsymbol{\alpha}, \boldsymbol{\beta})] \\ &= E\left[\ln \prod_j \left(p(z_j) \prod_i p(l_{ij} | z_j, \alpha_i, \beta_j) \right)\right] \\ &\quad \text{since } l_{ij} \text{ are conditionally independent given } \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta} \\ &= \sum_j E[\ln p(z_j)] + \sum_{ij} E[\ln p(l_{ij} | z_j, \alpha_i, \beta_j)] \\ \frac{\partial Q}{\partial \alpha_i} &= \sum_j (p^1 l_{ij} + p^0 (1 - l_{ij}) - \sigma) \beta_j \\ \frac{\partial Q}{\partial \beta_j} &= \sum_i (p^1 l_{ij} + p^0 (1 - l_{ij}) - \sigma) \alpha_i \end{aligned}$$

where $p^k = p(z_j = k | \mathbf{l}, \boldsymbol{\alpha}^{old}, \boldsymbol{\beta}^{old})$.

Empirical Study I: Greebles Gender Labeling

We posted 100 “greebles” (Gauthier & Tarr, 1997) images onto the Amazon Mechanical Turk. The task was to label the gender of each greeble.

We compared **GLAD** to the **Majority Vote heuristic** in terms of accuracy of inferred gender labels with respect to ground-truth. **Accuracy** was measured as a function of number of labels per image.



Empirical Study II: Duchenne Smile Labeling

We posted 160 smiling face images onto the Mechanical Turk. The task was to distinguish Duchenne from Non-Duchenne smiles.

We compared GLAD to Majority Vote using ground-truth labels obtained from facial expression experts. We simulated **noisy and adversarial labelers** by randomly flipping labels of some labelers.

