

Learning from Ambiguously Labeled Images

Timothee Cour, Benjamin Sapp, Chris Jordan, Ben Tasker

Presented by:
Aneesh Sood

Motivation

- Who is in the picture.
- Access to only weakly/Ambiguously labeled data.
- Learn from this Ambiguously labeled data.

Motivation

The image displays a 3x3 grid of video frames, each with a red bounding box around a face. Below each frame is a caption. At the bottom, three pairs of face images are shown with labels and question marks.

| Frame 1 (Top Left) | Frame 2 (Top Middle) | Frame 3 (Top Right) |
|---|---|--|
|  |  |  |
|  |  | |
| <p>Mark? Mark?</p> | <p>Eric? Eric?</p> | <p>Eric? Eric?</p> |
|  |  |  |
| <p>Eric? Eric?</p> | <p>Kate? Kate?</p> | <p>Kate? Kate?</p> |

Dataset

- TV Videos: 100 episodes of LOST and CSI.
- Labeled Faces in the Wild.

What is Ambiguous Labeling?

- Each image/example is supplied with multiple potential labels.
- Only one of the labels is correct. (true label)

Formulation

- For supervised multiclass setting:

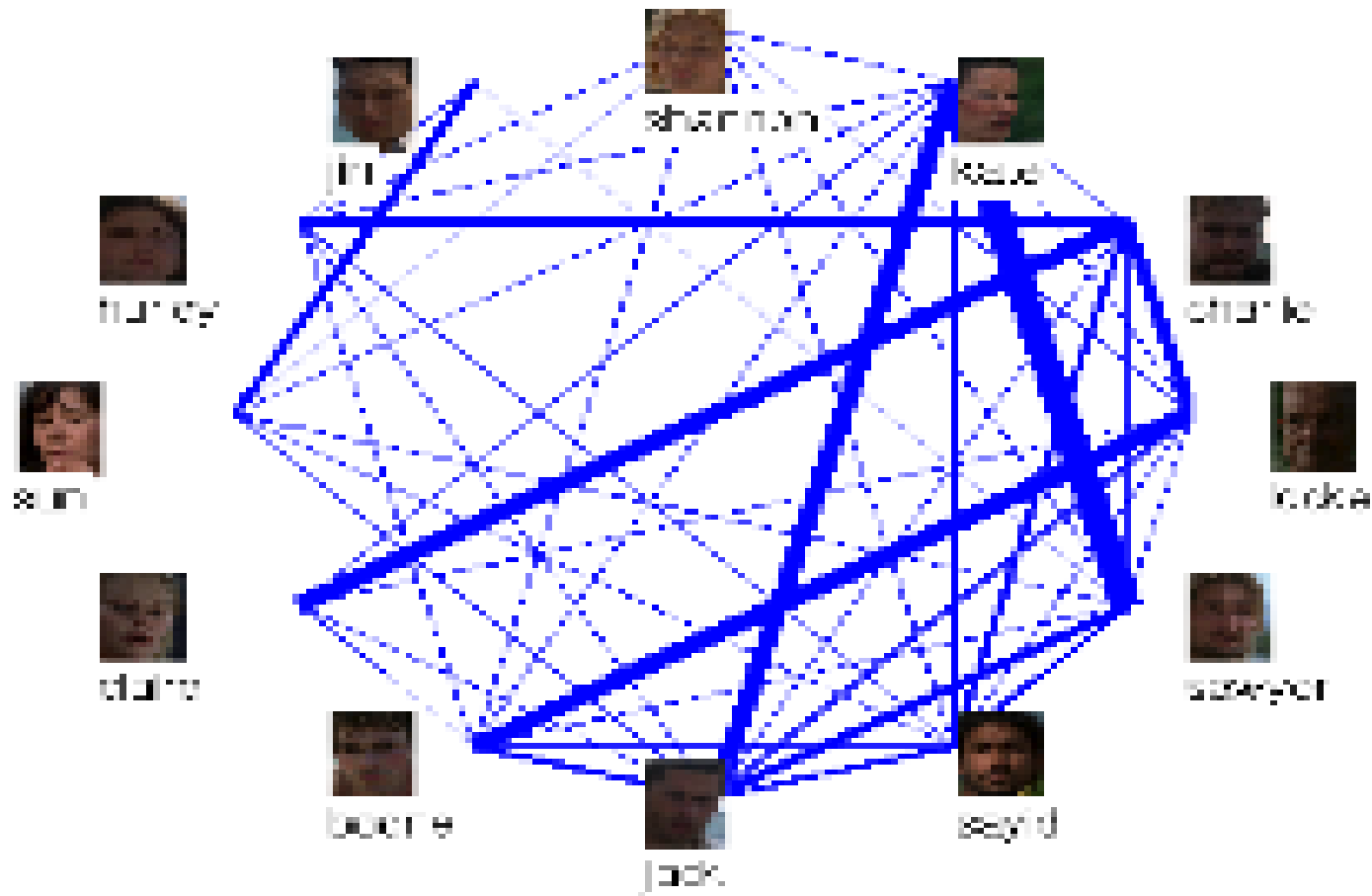
$$S = \{(x_i, y_i)\} \quad i = 1 \dots m.$$

- $x \in X$, the set of input and $y \in \{1, \dots, L\}$, the set of labels
- For the partially supervised setting:

$$S = \{(x_i, y_i, Z_i)\} \quad i = 1 \dots m.$$

- $Y_i = y_i \cup Z_i \quad Y_i \in \{1, \dots, L\}$

Co-occurrence



The Model

- A mapping from inputs to d -real valued features.

$$f(x) : X \rightarrow \mathbb{R}^d$$

- multi-linear classifier $g(x)$, with L components, one for each label.

$$g^a(x) = w^a \cdot f(x), \quad a \in \{1, \dots, L\}$$

- $g^a(x)$ is the class score and w^a are $L \times d$ weights.
- The prediction of the classifier is determined by:

$$g^*(x) = \arg \max_a g^a(x)$$

Loss Functions

- The usual 0/1 loss:

$$L_{01}(g(x), y) = 1 (g^*(x) \neq y).$$

- In our case, the ambiguous 0/1 loss:

$$L_{01}(g(x), Y_i) = 1 (g^*(x) \notin Y_i).$$

- Need a way to upperbound the 0/1 loss with the ambiguous loss.

Ambiguity Degree

- ambiguity degree $\varepsilon(P)$ of a distribution $P(x,y,Z)$:

$$\varepsilon(P) = \sup_{x \in X, y, a \in \{1, \dots, L\}} P(a \in Z \mid x, y).$$

Proposition 1

- For any classifier g and distribution P with $\epsilon(P) < 1$

$$\begin{aligned}\mathbf{E}_P[\mathcal{L}_{01}(g(x), Y)] &\leq \mathbf{E}_P[\mathcal{L}_{01}(g(x), y)] \\ &\leq \frac{1}{1 - \epsilon(P)} \mathbf{E}_P[\mathcal{L}_{01}(g(x), Y)]\end{aligned}$$

- We can bound the unobserved 0/1 loss by the observed ambiguous 0/1 loss.

Problem!

- unlikely pairs (x, y) might force ε to be large, making the bound very loose.

Solution

- (ϵ, δ) -ambiguous distribution.
- A distribution P containing a subset of the space $A \subseteq X \times \{1, \dots, L\}$ with probability mass at least $1 - \delta$, (i.e. $P((x, y) \in A) \geq 1 - \delta$), where

$$\sup_{(x,y) \in A, a \in \{1, \dots, L\}} P(a \in Z \mid x, y) \leq \epsilon$$

- Rewrite proposition 1 in terms of ϵ and δ .

Proposition 2

- For any classifier g and (ϵ, δ) -ambiguous $P(Z | x, y)$,

$$\mathbf{E}_P[\mathcal{L}_{01}(g(x), y)] \leq \frac{1}{1 - \epsilon} \mathbf{E}_P[\mathcal{L}_{01}(g(x), Y)] + \delta.$$

- The bound now depends on ϵ instead of $\epsilon(P)$.

Label Specific bounds

- Further tighten the bounds by considering label specific information.
- Define a label specific ambiguity degree $\varepsilon^a(P)$

$$\varepsilon^a(P) = \sup_{x \in \mathcal{X}; a' \in \{1, \dots, L\}} P(a' \in Z \mid x, y = a).$$

Proposition 3

- For any classifier g and distribution P with $\varepsilon^a(P) < 1$,

$$\mathbf{E}_P[\mathcal{L}_{01}(g(x), y) \mid y = a] \leq \frac{1}{1 - \varepsilon^a} \mathbf{E}_P[\mathcal{L}_{01}(g(x), Y) \mid y = a]$$

Convex Learning Formulation

- Binary loss function $\psi(\cdot) : \mathbb{R} \rightarrow \mathbb{R}_+$
- Combine binary losses to create a multiclass loss function.
- Example of a binary loss function: Square hinge loss.

$$l(g^a(x)) = \max(0, 1 - g^a(x))^2$$

Proposed Loss function

$$\mathcal{L}_\psi(g(\mathbf{x}), Y) = \psi \left(\frac{1}{|Y|} \sum_{a \in Y} g^a(\mathbf{x}) \right) + \sum_{a \notin Y} \psi(-g^a(\mathbf{x}))$$

The Algorithm

- Select a classifier g which minimizes the loss.
- An optimization problem.
- Solved using L2 regularization.

$$\begin{aligned} & \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \|\boldsymbol{\xi}\|_2^2 \\ \text{s.t.} \quad & \frac{1}{|Y_i|} \sum_{a \in Y_i} \mathbf{w}^a \cdot \mathbf{f}(x_i) \geq 1 - \xi_i \\ & -\mathbf{w}^a \cdot \mathbf{f}(x_i) \geq 1 - \xi_{ia} \quad \forall a \notin Y_i \end{aligned}$$

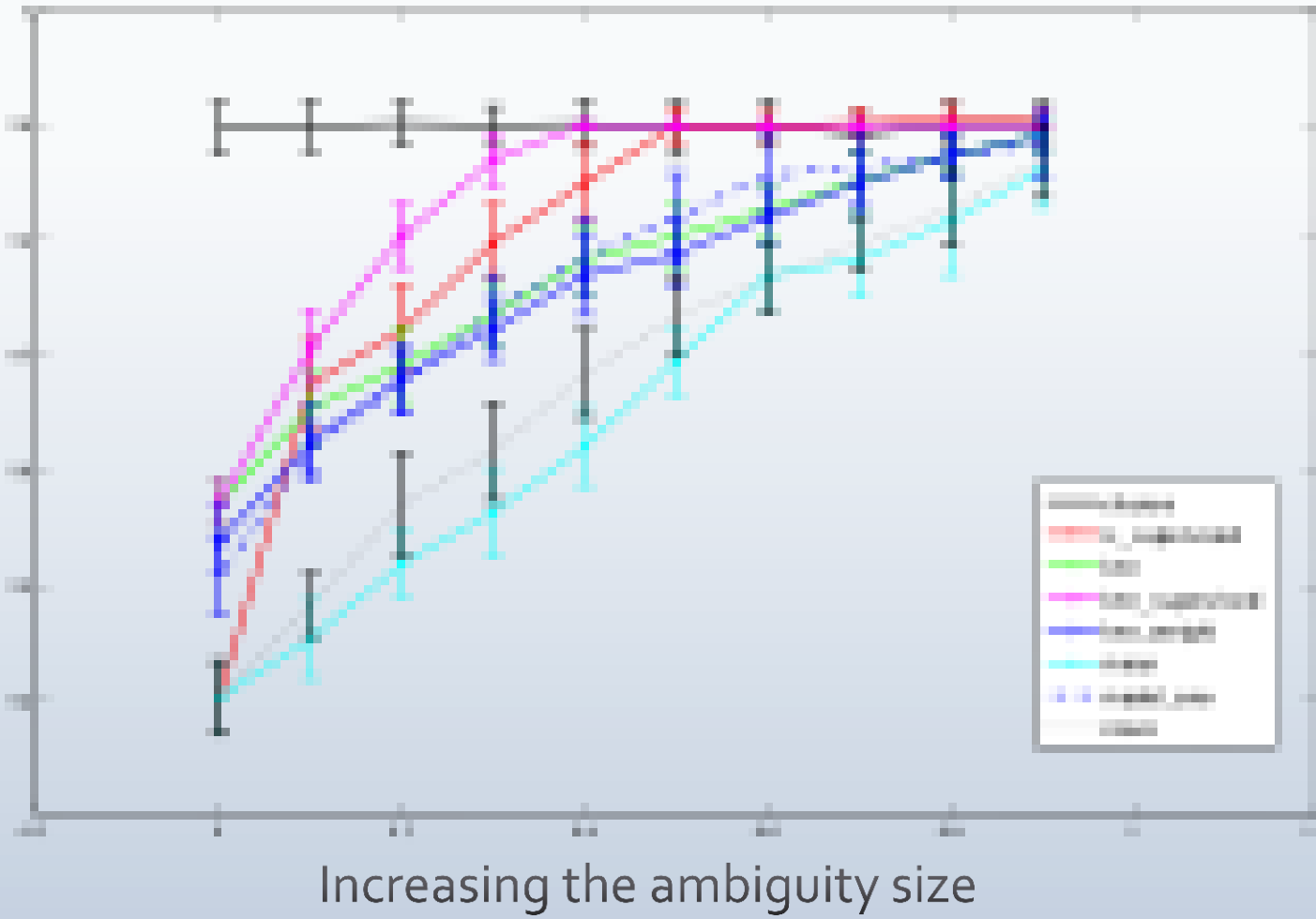
Experiments

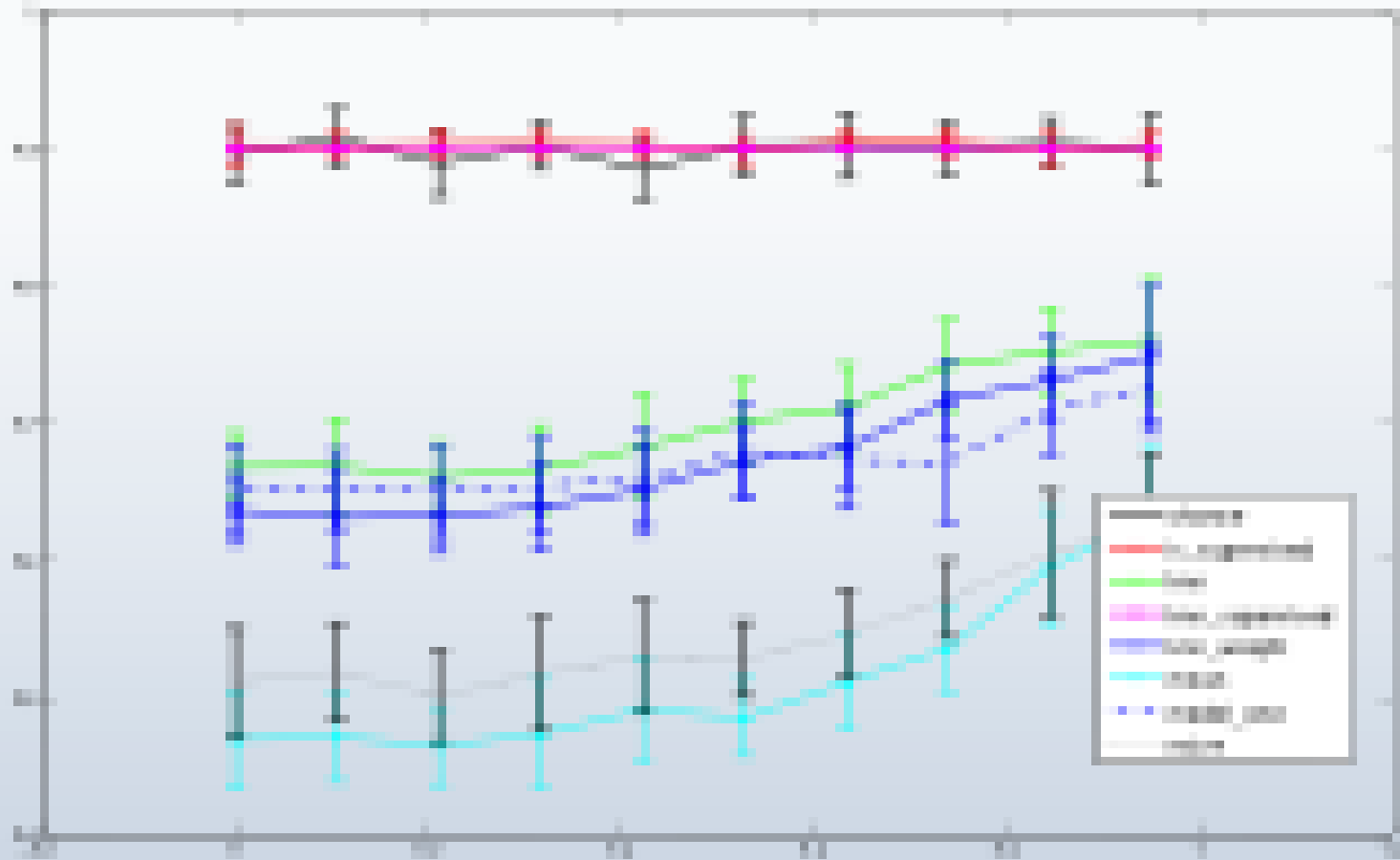
- Dataset: Labeled Faces in the Wild.
- Take first 50 images of 10 most frequent people.
- Goal: correctly label faces from examples that have multiple potential labels (transductive case), as well as learn a model from ambiguous data that generalizes to other unlabeled examples (inductive case).
- Results report the average test error rate over 20 trials.

Baselines

- Random model.
- IBM Model 1
- Discriminative EM
- k-nearest neighbor (uniform weights)
- k-nearest neighbor (linearly decreasing weights)
- Naïve model
- Supervised models.

Transductive setting





Increasing the Ambiguity degree.

Original Problem

- Labeling people in TV shows.
- Dataset : 100 episodes of LOST and CSI.
- Extract ambiguously labeled faces to learn models of frequently occurring characters.
- Experiment with top {8,16,32} characters.
- Ambiguous bag size restricted to maximum of 3.
- Same algorithm as used for Labeled faces in the wild dataset.

Adding constraints

- Mouth motion

$$Y := \begin{cases} \{\alpha\} & \text{if mouth motion} \\ Y & \text{if refuse to predict or } |Y| = \{\alpha\} \\ Y - \{\alpha\} & \text{if absence of mouth motion} \end{cases}$$

- Gender constraints

$$Y := \begin{cases} Y & \text{if gender uncertain} \\ Y - \{\alpha : \alpha \text{ is male}\} & \text{if gender predicts female} \\ Y - \{\alpha : \alpha \text{ is female}\} & \text{if gender predicts male} \end{cases}$$

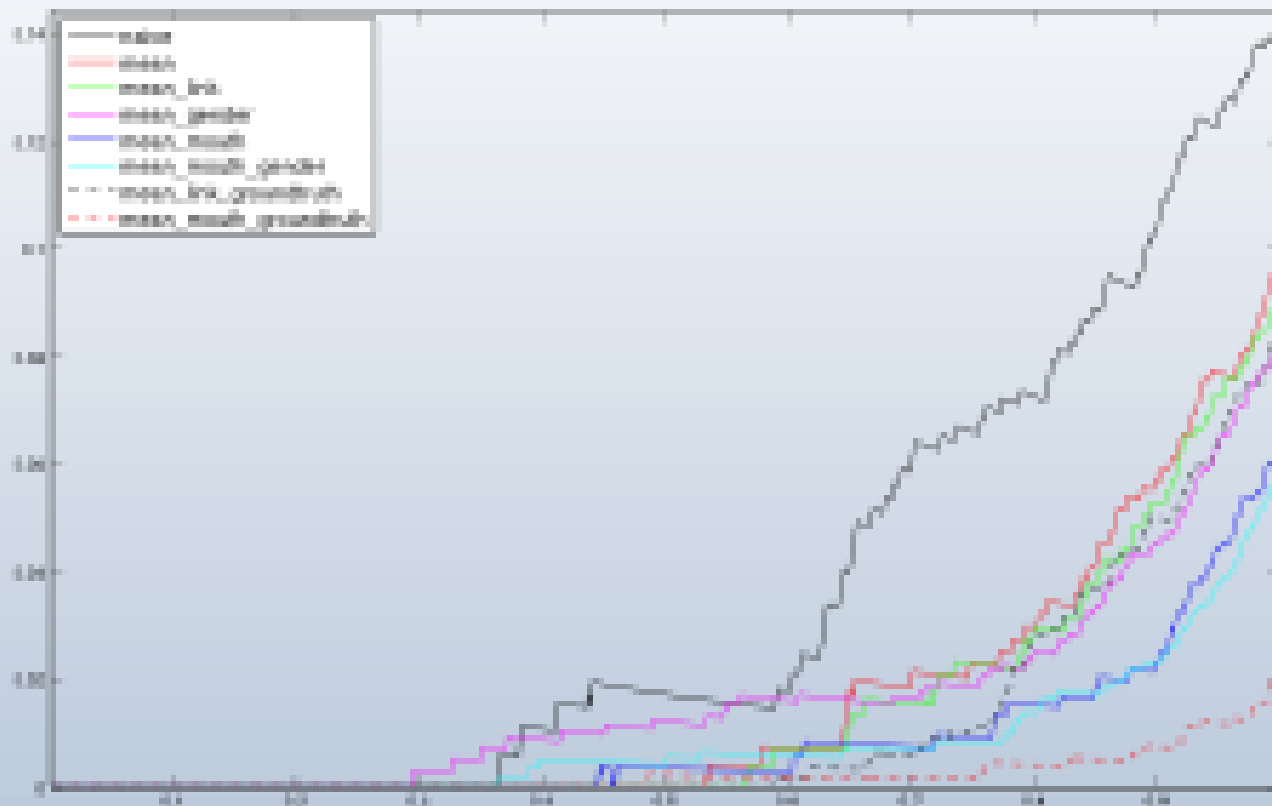
Adding constraints

- Grouping constraints.
- $y_i \neq y_j$ if face tracks x_i, x_j are in two consecutive shots

| LOST (#labels, #eps.) | (8,16) | (16,16) | (32,16) |
|-----------------------|-----------|------------|------------|
| Naive | 14% | 16.5% | 18.5% |
| ours ("mean") | 10% | 14% | 17% |
| ours+constraints | 6% | 11% | 13% |

Ablative analysis

- For a given recall rate $r \in [0, 1]$, extract the $r \cdot m$ most confident predictions and compute error rate on those examples.





Examples classified as Kate in LOST. The precision is 97.5%.

Thank you!