Learning Effective Embeddings From Crowdsourced Labels: An Educational Case Study

Guowei Xu[†], Wenbiao Ding[†], Jiliang Tang[*]
Songfan Yang[†], Gale Yan Huang[†] and Zitao Liu[†]
[†]TAL AI Lab, Beijing, China
[*]Michigan State University, MI, USA

Introduction

Representation learning is important. Most existing learning approaches are based on neural networks, which rely on massive data and labels. However, there are cases where labelled data is limited. One practice to obtain labels is by crowdsourcing.

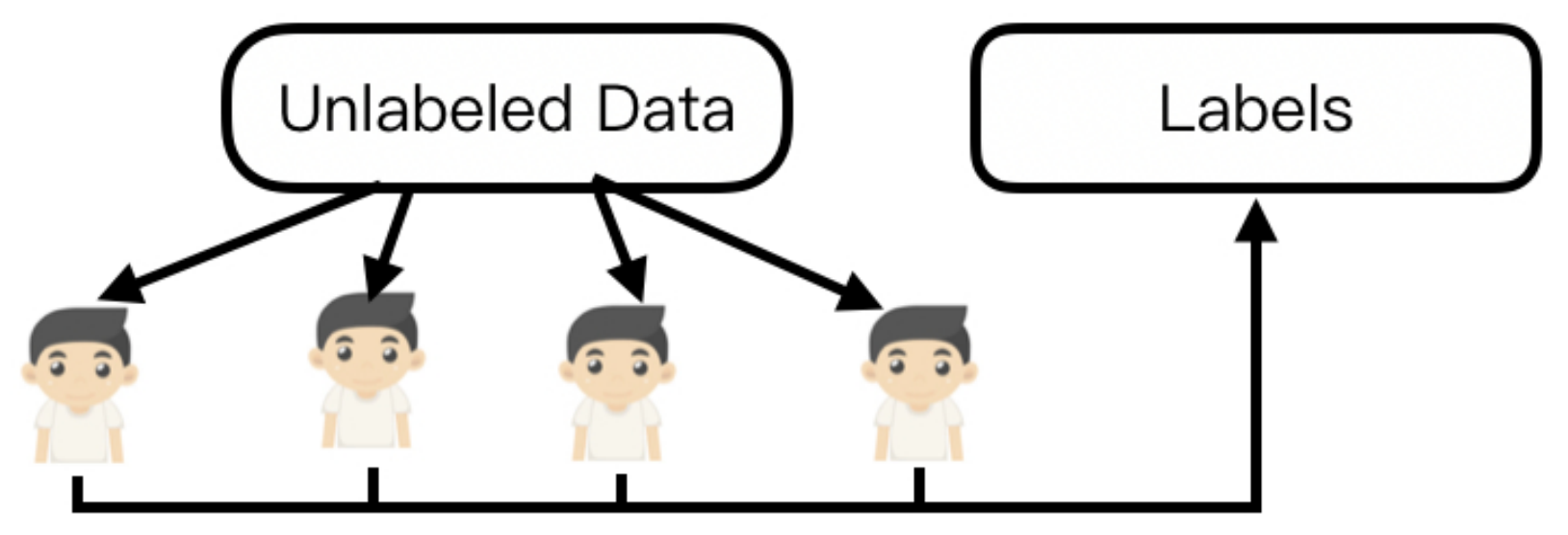

Figure 1: Crowdsourcing.

Considering limited budgets, task difficulty and diverse worker quality, there are still two problems with crowdsourcing:

- **limited labelled data**
- **inconsistent crowdsourced labels**

Existing neuralNet based approaches produce suboptimal performance with limited noisy labels. To joinly solve these two problems, we propose “RLL”, which learns representation of data with crowdsourced labels by jointly and coherently solving the challenges introduced by limited and inconsistent labels.

The Representation Learning Framework

Step1: Grouping Based Deep Architecture

The idea is to pair positive and nagative examples together to form groups. The strategy substantially generates more training groups.

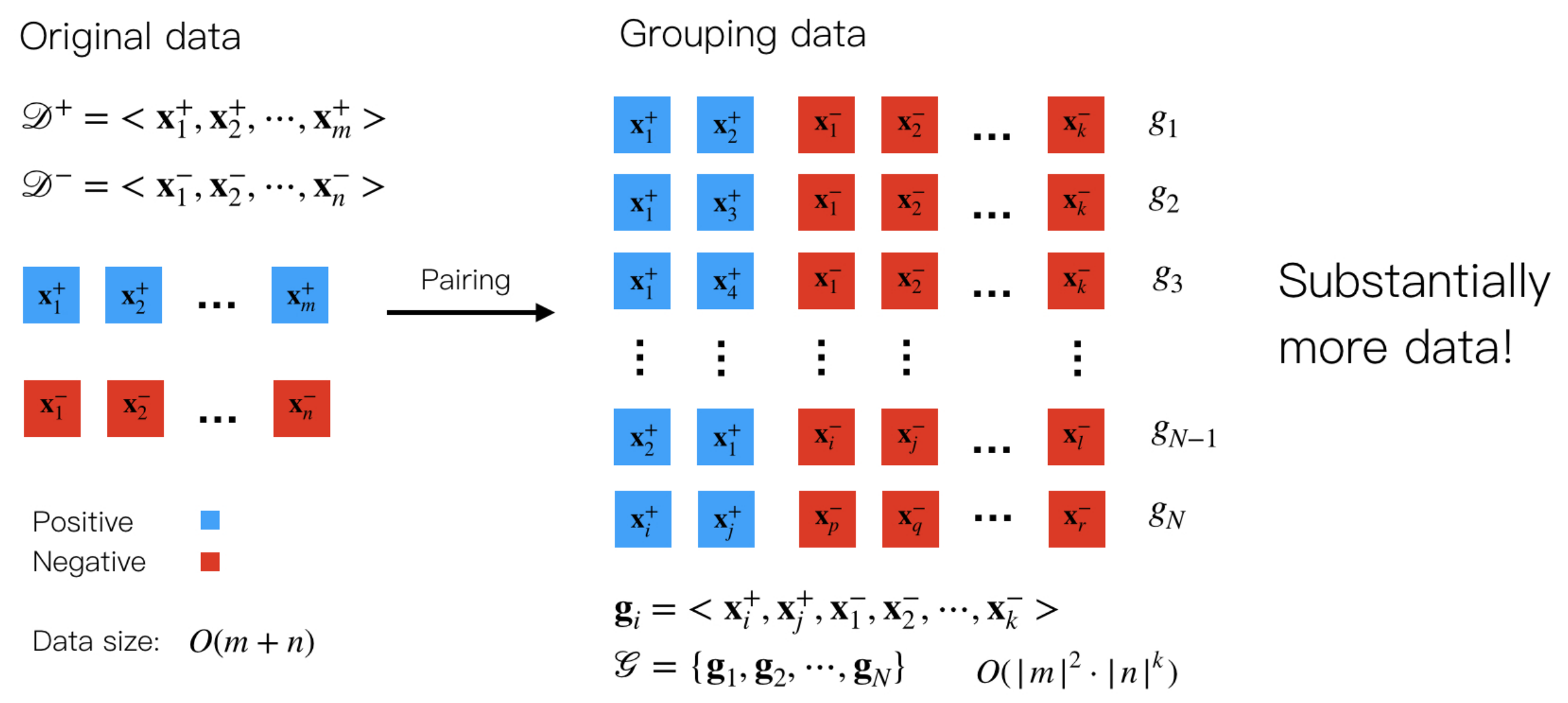
Original data
$\mathscr{D}^+ = < \mathbf{x}_1^+, \mathbf{x}_2^+, \cdots, \mathbf{x}_m^+ >$
$\mathscr{D}^- = < \mathbf{x}_1^-, \mathbf{x}_2^-, \cdots, \mathbf{x}_n^- >$
Pairing
Positive
Negative
Data size: $O(m+n)$
Grouping data
g_1
g_2
g_3
g_{N-1}
g_N
Substantially more data!
$\mathbf{g}_i = < \mathbf{x}_i^+, \mathbf{x}_j^+, \mathbf{x}_1^-, \mathbf{x}_2^-, \cdots, \mathbf{x}_k^- >$
$\mathscr{G} = \{\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_N\}$
$O(|m|^2 \cdot |n|^k)$

Figure 2: Grouping based architecture.

Let $\mathbf{x}_i^+$ and $\mathbf{x}_i^-$ denote positive and negative examples. A group $\mathbf{g}_i =< \mathbf{x}_i^+, \mathbf{x}_j^+, \mathbf{x}_1^-, \mathbf{x}_2^-, \cdots, \mathbf{x}_k^- >$. Using the grouping strategy, we can create $O(m^2 \cdot n^k)$ groups for training theoretically, where m and n are the number of positive and negative examples in the original labeled data.

Step2: Bayesian Confidence Estimator

To deal with inconsistent crowdsourced labels, each example will have a confidence score based on its crowdsourced labels. In step3, the score is integrated into loss function for learning representation.

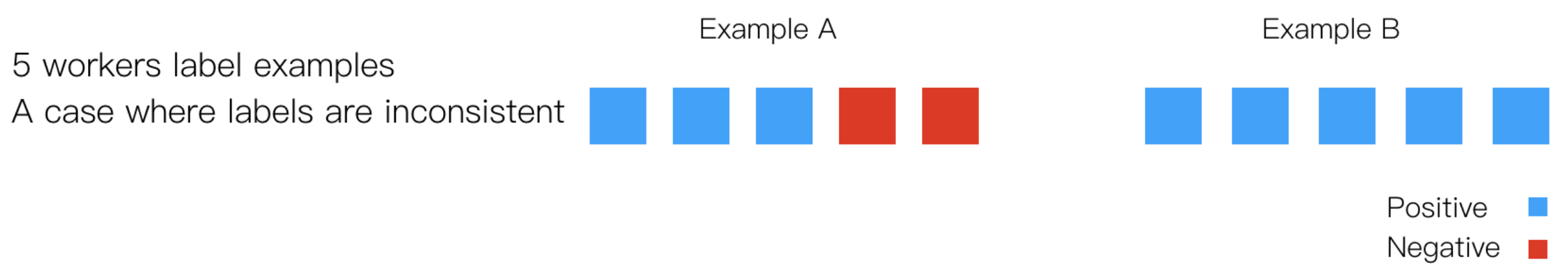

Figure 3: Confidence Score Example.

We treat confidence score as a random variable and use Bayesian inference to estimate it. We assign a Beta prior to δ_i, i.e., $\delta_i \sim \text{Beta}(\alpha, \beta)$. Therefore, the posterior estimation of the crowdsourced label confidence is

$$\delta_i^{\text{Bayesian}} = \frac{\alpha + \sum_{j=1}^{d} y_{i,j}}{\alpha + \beta + d} \quad (1)$$

where d is the number of crowd workers, $y_{i,j}$ is the label for example i from worker j

Step3: Model Learning

We integrate the confidence score into our representation learning. The confidence weighted conditional probability is defined as follows:

$$\hat{p}(\mathbf{x}_j^+ | \mathbf{x}_i^+) = \frac{\exp\left(\eta \cdot \delta_j \cdot r(\mathbf{x}_i^+, \mathbf{x}_j^+)\right)}{\sum_{\mathbf{x}_* \in \mathbf{g}_i, \mathbf{x}_* \neq \mathbf{x}_i^+} \exp\left(\eta \cdot \delta_* \cdot r(\mathbf{x}_i^+, \mathbf{x}_*)\right)} \quad (2)$$

where δ_j and δ_* are confidence scores of $\mathbf{x}_j^+$ and $\mathbf{x}_*$.

Model Summary

RLL = Grouping based strategy + Confidence estimator + Model learning

In our RLL framework, given the limited data with crowdsourced labels, we

- generate a fair large amount of groups of training examples by pairing both positive and negative examples;
- estimate the label confidence for each crowdsourced data by a Bayesian estimator;
- feed all groups into a DNN which maximizes the confidence-weighted conditional likelihood of retrieving the positive examples in Equation 2.

The entire RLL framework is illustrated in Figure 4.

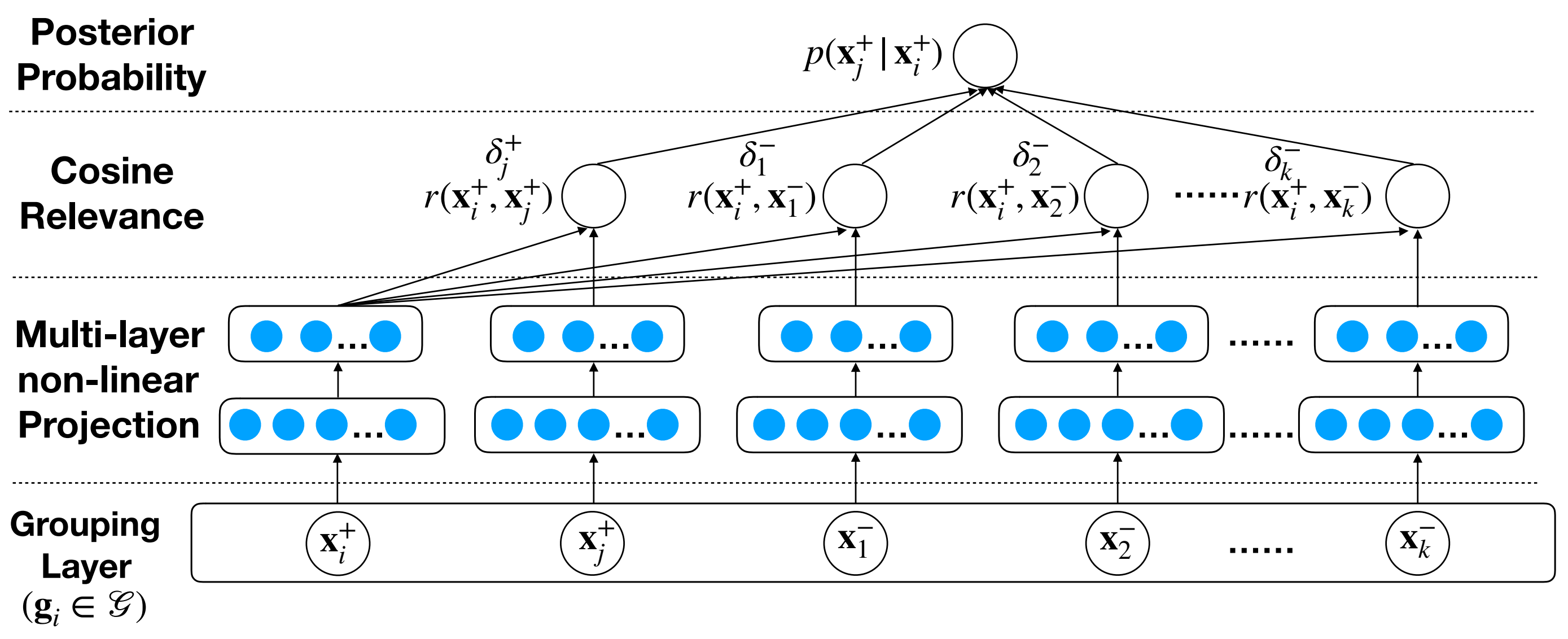
Posterior Probability
$p(\mathbf{x}_j^+ | \mathbf{x}_i^+)$
Cosine Relevance
δ_j^+
$r(\mathbf{x}_i^+, \mathbf{x}_j^+)$
δ_1^-
$r(\mathbf{x}_i^+, \mathbf{x}_1^-)$
δ_2^-
$r(\mathbf{x}_i^+, \mathbf{x}_2^-)$
δ_k^-
$r(\mathbf{x}_i^+, \mathbf{x}_k^-)$
Multi-layer non-linear Projection
Grouping Layer
$(\mathbf{g}_i \in \mathscr{G})$
$\mathbf{x}_i^+$
$\mathbf{x}_j^+$
$\mathbf{x}_1^-$
$\mathbf{x}_2^-$
$\mathbf{x}_k^-$

Figure 4: The overview of the RLL framework.

Experiments

Dataset

- **Oral Math Questions**(“oral”) 880 audio files about oral math questions, each labelled as fluent or disfluent in terms of speaking fluency.
- **Online 1v1 Class Qualities** (“class”) 472 videos from online 1v1 classes, each labelled as good or bad in terms of class quality.

Baselines

Group 1: Logistic Regression + True Label Inference

- Logistic regression with every pair, i.e. SoftProb.
- Logistic regression with EM labels, i.e. EM.
- Logistic regression with GLAD labels, i.e. GLAD.

Group 2: Representation Learning with Limited Labels

- Siamese network, i.e. SiameseNet.
- Triplet network, i.e. TripletNet.
- Relation network for few shot learning, i.e. RelationNet.

Group 3: Two-stage Models

This group combines group 1 and group 2. They solve the problems of the limited and inconsistent labels in two stages.

Group 4: Our Methods

RLL framework and its variants. They solve the problems of the limited and inconsistent labels jointly.

- RLL without Bayesian confidence score, i.e., RLL.
- RLL with confidence score estimated by MLE, i.e., RLL-MLE.
- RLL with confidence score estimated by Bayesian approach, i.e., RLL-Bayesian.

Result

RLL-Bayesian demonstrates the best performance on both datasets.

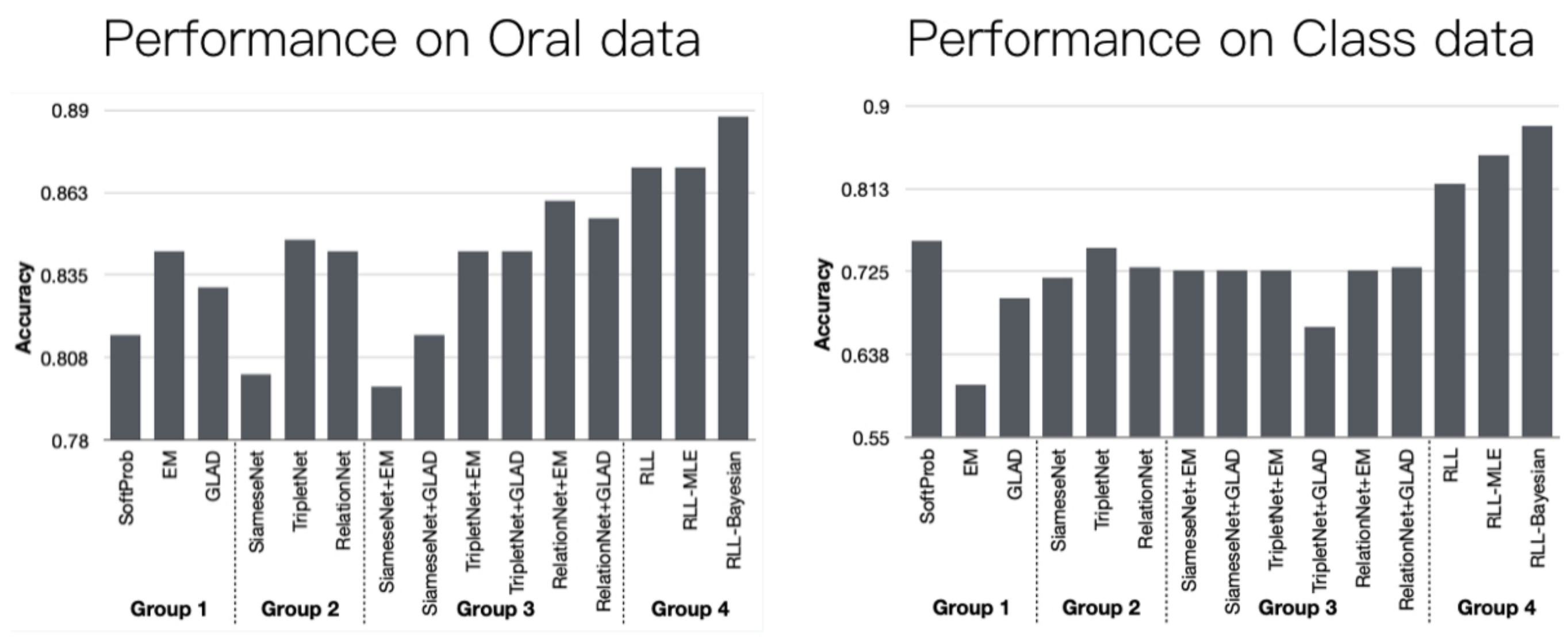

Figure 5: Performance Results on Both Datasets.

Conclusion

We design a novel representation learning framework RLL for crowdsourced labels under the limited and inconsistent settings. Experimental results on two real-world applications demonstrate (1) the proposed framework outperforms the representative baselines; and (2) it is necessary to address the limited and inconsistent label problems simultaneously. Our current model does not make use of any information about individual crowd worker and we want to extend the proposed framework to incorporate such information in the future.

Contact: liuzitao@100tal.com