Structured Probabilistic End-to-End Learning from Crowds

Zhijun Chen[1,2*], **Huimin Wang**[1,2*], **Hailong Sun**[1,2†], **Pengpeng Chen**[1,2],
Tao Han[1,2], **Xudong Liu**[1,2] and **Jie Yang**[3]
[1]SKLSDE Lab, School of Computer Science and Engineering, Beihang University, China
[2]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, China
[3]Web Information Systems, Delft University of Technology, Netherlands
{zhijunchen, whm2016, sunhl, chenpp, hantao, liuxd}@buaa.edu.cn, jie@exascale.info

Abstract

End-to-end learning from crowds has recently been introduced as an EM-free approach to training deep neural networks directly from noisy crowdsourced annotations. It models the relationship between true labels and annotations with a specific type of neural layer, termed as the crowd layer, which can be trained using pure backpropagation. Parameters of the crowd layer, however, can hardly be interpreted as annotator reliability, as compared with the more principled probabilistic approach. The lack of probabilistic interpretation further prevents extensions of the approach to account for important factors of annotation processes, *e.g.*, instance difficulty. This paper presents SpeeLFC, a structured probabilistic model that incorporates the constraints of probability axioms for parameters of the crowd layer, which allows to explicitly model annotator reliability while benefiting from the end-to-end training of neural networks. Moreover, we propose SpeeLFC-D, which further takes into account instance difficulty. Extensive validation on real-world datasets shows that our methods improve the state-of-the-art.

1 Introduction

The success of deep learning in many vision and language tasks heavily relies on the quantity and quality of labeled training data [Zhang *et al.*, 2016]. Crowdsourced data annotation offers a cost-effective means to acquire a large set of labeled data. However, crowdsourced labels are often of limited quality, which has become a key concern for training deep neural networks and other types of classifiers [Yang *et al.*, 2019].

To train classifiers from noisy crowdsourced labels, existing methods generally treat true labels as unknown variables, and infer the value by modeling a probabilistic relationship between true labels and the annotations (*e.g.*, the confusion matrix) [Zheng *et al.*, 2017]. Key considerations are factors in the annotation process such as annotator reliability and instance (data instance) difficulty, which are represented as parameters of the probabilistic relationship. Depending on how the inference is integrated with the training of classifier, existing methods fall into two broad categories:

- The two-stage approach [Dawid and Skene, 1979; Whitehill *et al.*, 2009; Welinder *et al.*, 2010; Han *et al.*, 2016] infers the true label for each instance and afterward, it trains classifiers using the inferred labels.
- The joint approach, also referred to as the learning-from-crowds approach [Raykar *et al.*, 2010; Kajino *et al.*, 2012; Yan *et al.*, 2012; Rodrigues *et al.*, 2013; Bi *et al.*, 2014; Albarqouni *et al.*, 2016; Rodrigues and Pereira, 2018], simultaneously infers the true labels while training the classifier, allowing the two processes to benefit from each other; consequently, this approach generally results in better performance.

Most of the methods in the joint approach, however, are computationally expensive when the classifier is a neural network, as the learning with generally carried out with expectation-maximization (EM) algorithm that infers the true labels and learns parameters of the classifier (especially neural networks) in an iterative manner [Rodrigues and Pereira, 2018].

Recently, Rodrigues and Pereira [2018] introduce an end-to-end learning-from-crowds approach where the relationship between true labels and annotations is modeled by a specific type of neural layers, *i.e.*, the crowd layer. Parameters of the crowd layer can be trained together with the parameters of the rest of the neural network using backpropagation, thus largely accelerating the learning process. Despite that, parameters of the crowd layer can hardly be interpreted as annotator reliability, which comes in contrast to the traditional probabilistic approach that always gives us effective and principled solutions. The lack of probabilistic interpretation also makes it challenging to extend the approach to account for other important factors of the annotation process, *e.g.*, instance difficulty, hindering the potential improvement of the learning process.

In this paper, we demonstrate that we can get the best of both worlds by presenting SpeeLFC (*Structured Probabilistic end-to-end Learning From Crowds*), a structured probabilistic model (*i.e.*, *probabilistic graphical models*) to end-to-end learning from noisy crowd annotations. SpeeLFC enforces

[*] Equal Contribution
[†] Corresponding Author

the parameters of the crowd layer to satisfy the constraints of probability axioms, thereby explicitly modeling annotator reliability in the network. To learn the parameters with such constraints, we introduce a reparameterization trick that allows the network to be trained with the standard gradient-based optimization algorithm. Our approach, therefore, benefits both from the expressiveness and reasonableness of probabilistic approaches for representing annotator reliability and from the effectiveness of the end-to-end approach for training neural networks. In addition, our approach can be easily extended to model other influential factors of the annotation process. Moreover, we propose SpeeLFC-D, which extends SpeeLFC to model instance difficulty as an additional set of parameters of the crowd layer.

In summary, we make the following key contributions:

- We propose SpeeLFC, a novel structured probabilistic model that learns interpretable parameters of the crowd layer in end-to-end learning from crowds;
- We propose SpeeLFC-D, an extension of SpeeLFC that further models instance difficulty in end-to-end learning from crowds;
- We conducted evaluation of our proposed approach on two real-world datasets, showing that both SpeeLFC and SpeeLFC-D are able to outperform the state-of-the-art models.

To the best of our knowledge, we are the first to consider the probabilistic interpretation of neural network parameters and the first to model instance difficulty in end-to-end learning from crowds. Experimental results show that the probabilistic constraints alone in SpeeLFC not only make our approach more interpretable and expressive, but also contribute to learning classifier with higher performance.

2 Preliminaries

To help motivate the two proposed models, here we first introduce the far-reaching approach Crowd-Layer [Rodrigues and Pereira, 2018], and then introduce the simple yet profound idea that underlies both of our proposed models.

LFC problem formulation. Let $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{Y}^{(i)}\}_{i=1}^{I}$ be an i.i.d. dataset, where for each instance $\mathbf{x}^{(i)} \in \mathbb{R}^D$ we are given a set of noisy crowdsourced annotations $\mathbf{Y}^{(i)} = \{\mathbf{y}^{(i,j)}\}_{j\in\mathcal{J}^{(i)}}$. $\mathcal{J}^{(i)}$ represents all the annotators who annotated the i^{th} instance, and $\mathbf{y}^{(i,j)}$ represents annotation by the j^{th} annotator (a total of J annotators) on instance $\mathbf{x}^{(i)}$, which is a $1-$of$-K$ encoded K dimensional vector. Each instance $\mathbf{x}^{(i)}$ has its corresponding *unobserved* ground truth $\mathbf{t}^{(i)}$. The goal of LFC is to train an accurate classifier for predicting $\mathbf{t}$ given new unknown instances $\mathbf{x}$ by using noisy data $\{\mathbf{x}^{(i)}, \mathbf{Y}^{(i)}\}_{i=1}^{I}$.

Crowd-Layer

The Crowd-Layer constructs a probabilistic discriminative model and optimizes the cross-entropy loss on crowdsourced annotations, which is shown in Figure 1 (a). Particularly, it adds functions $f_j(\mathbf{t}^{(i)}) = \mathbf{\Pi}^{(j)}\mathbf{t}^{(i)}$ to the ground truth $\mathbf{t}^{(i)}$ after the classifier to obtain a new vector $\mathbf{a}^{(i,j)}$, where $\mathbf{\Pi}^{(j)}$ is an annotator-specific matrix. The elements in the matrix are the parameters of neural network to be trained, whose value range is $(-\infty, +\infty)$. Then it uses a softmax function to map vector $\mathbf{a}^{(i,j)}$ to the new vector $\mathbf{y}^{(i,j)}$ which can satisfy a distribution form. Note that there are other variants in Rodrigues *et al.* [2018] , but the model presented here is the main model when faced with classification problems.

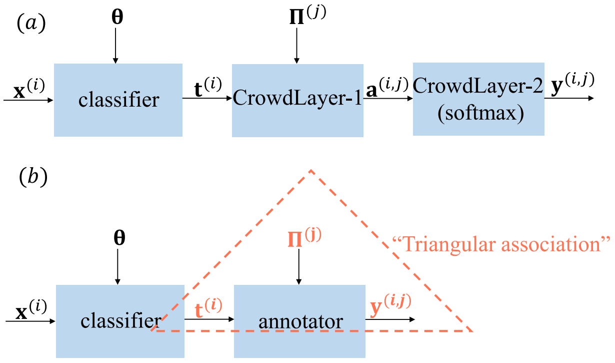


Figure 1: Schematics of the Crowd-Layer (a) and our models (b).

The Basic Idea in Our Models

Different from Crowd-Layer, in our two proposed models, the distribution of crowdsourced annotation $\mathbf{y}^{(i,j)}$ is obtained by the linear transformation of ground truth distribution $\mathbf{t}^{(i)}$ through the *annotator transition matrix* $\mathbf{\Pi}^{(j)}$ we construct. That is, assume in a binary classification scenario:

$$\mathbf{\Pi}^{(j)}\mathbf{t}^{(i)} = \mathbf{y}^{(i,j)}, \tag{1}$$

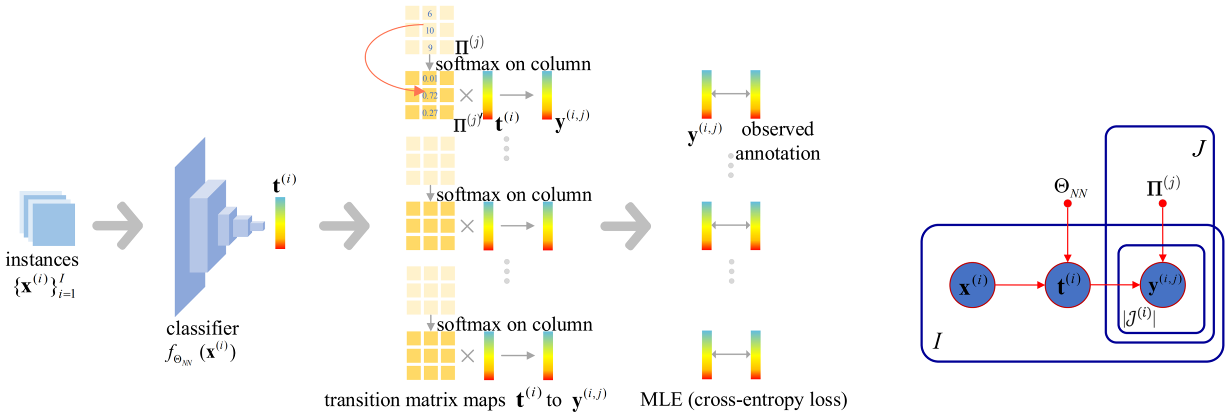
$$\begin{bmatrix} \pi_{11}^{(j)} & \pi_{12}^{(j)} \\ \pi_{21}^{(j)} & \pi_{22}^{(j)} \end{bmatrix} \begin{bmatrix} t_1^{(i)} \\ t_2^{(i)} \end{bmatrix} = \begin{bmatrix} \pi_{11}^{(j)} t_1^{(i)} + \pi_{12}^{(j)} t_2^{(i)} \\ \pi_{21}^{(j)} t_1^{(i)} + \pi_{22}^{(j)} t_2^{(i)} \end{bmatrix} = \begin{bmatrix} y_1^{(i,j)} \\ y_2^{(i,j)} \end{bmatrix}, \tag{2}$$

where element $\pi_{mn}^{(j)}$ represents the probability that the j^{th} annotator will annotate m given the ground truth is n. Note that the transition matrix we construct here is the transpose of the confusion matrix generally used in machine learning. Our annotator transition matrix $\mathbf{\Pi}^{(j)}$ can therefore express a probabilistic relationship between the annotation $\mathbf{y}^{(i,j)}$ and ground truth $\mathbf{t}^{(i)}$, and its meaning can thus be interpreted as worker reliability. We illustrate the relationship among the three variables as "triangular association" in Figure 1.

Recall that Crowd-Layer, like ours, uses a specific matrix to model an annotator. However, the linear transformation $\mathbf{a}^{(i,j)} = \mathbf{\Pi}^{(j)}\mathbf{t}^{(i)}$ without any probabilistic meaning and the nonlinear transformation of softmax in the model break the delicate "triangular association".

3 Method

In this section, we formally introduce two models, SpeeLFC and SpeeLFC-D, which exploit the power of structured probabilistic models for end-to-end learning from crowds.



(a) Framework of SpeeLFC

(b) Probabilistic graphical representation of SpeeLFC

Figure 2: SpeeLFC overview: (1) Given an input instance $\mathbf{x}^{(i)}$, the classifier $f_{\Theta_{NN}}(\mathbf{x}^{(i)})$ generates the ground truth $\mathbf{t}^{(i)}$. (2) Then, annotation transition matrix $\mathbf{\Pi}^{(j)'}$ converts ground truth $\mathbf{t}^{(i)}$ into a noisy version of it, *i.e.*, $\mathbf{y}^{(i,j)} = \mathbf{\Pi}^{(j)'}\mathbf{t}^{(i)}$. Here we use the reparameterization technique, that is, we construct an ancestral-transition matrix for each true-transition matrix, and the latter is obtained by performing "softmax on column" operation on the former. Thus, the model is parametrized by $\Theta = \{\Theta_{NN}, \mathbf{\Pi}^{(1)}, \ldots, \mathbf{\Pi}^{(J)}\}$. (3) Finally, given the observed annotations, the common optimization objective $\log p(\mathbf{Y}|\mathbf{X};\Theta)$ (equivalent to minimizing the corresponding cross-entropy loss) is built on the model, which can be done with standard stochastic optimization techniques.

3.1 SpeeLFC

Model

Here we introduce the probabilistic generative process we construct from instance features to noisy crowdsourced annotations. First, as shown in Figure 2, for each instance $\mathbf{x}^{(i)}$,

$$\mathbf{t}^{(i)}|\mathbf{x}^{(i)};\Theta_{NN} \sim \mathrm{Cat}(\mathbf{t}^{(i)}; f_{\Theta_{NN}}(\mathbf{x}^{(i)})), \tag{3}$$

where the distribution of its unobserved ground truth $\mathbf{t}^{(i)}$ comes from a conditional categorical distribution $\mathrm{Cat}(\mathbf{t}^{(i)}|f_{\Theta_{NN}}(\mathbf{x}^{(i)}))$, which can be a flexible neural network model parametrized by Θ_{NN}.

Then, we directly model each annotator with an annotator-specific transition matrix $\mathbf{\Pi}^{(j)'}$ as mentioned in Section 2, which represents her performance pattern on ground truth. Thus, the distribution of annotation $\mathbf{y}^{(ij)}$ is determined by:

$$p(\mathrm{ind}(\mathbf{y}^{(i,j)}) = m|\,\mathrm{ind}(\mathbf{t}^{(i)}) = n; \mathbf{\Pi}^{(j)'}) = \pi_{mn}^{(j)'}, \\ m, n, \mathrm{ind}(\cdot) \in \{1, \ldots, K\}, \forall j, \tag{4}$$

where $\pi_{mn}^{(j)'}$ represents the probability that the j^{th} annotator will annotate m given the ground truth is n, and $\mathrm{ind}(\cdot)$ is to take the location index of the value 1 in the one-hot vector (K denotes the number of categories). Here the generation process from ground truth to crowdsourced annotation corresponds to the basic idea introduced in Section 2.

Based on the probabilistic model constructed above, we unfortunately find an optimization dilemma when we try to optimize the log-conditional likelihood $\log p(\mathbf{Y}|\mathbf{X}; \Theta_{NN}, \{\mathbf{\Pi}^{(j)'}\}_{j=1}^{J})$ using gradient-based optimization algorithms. Because there are inevitable constraints of the parameters $\{\mathbf{\Pi}^{(j)'}\}_{j=1}^{J}$:

$$\sum_{m=1}^{K} \pi_{mn}^{(j)'} = 1, \quad m, n \in \{1, \ldots, K\}, \forall j, \tag{5}$$

$$0 \leq \pi_{mn}^{(j)'} \leq 1, \quad m, n \in \{1, \ldots, K\}, \forall j, \tag{6}$$

and the updates brought to the parameters by the backpropagation cannot make the parameters automatically obey the constraints of Eq. 5 and Eq. 6. Last, we break out of this optimization dilemma just by invoking the reparameterization technique, which also used in the "Mixture Density Network" in Bishop [2006] and variational autoencoder (VAE) [Kingma and Welling, 2013].

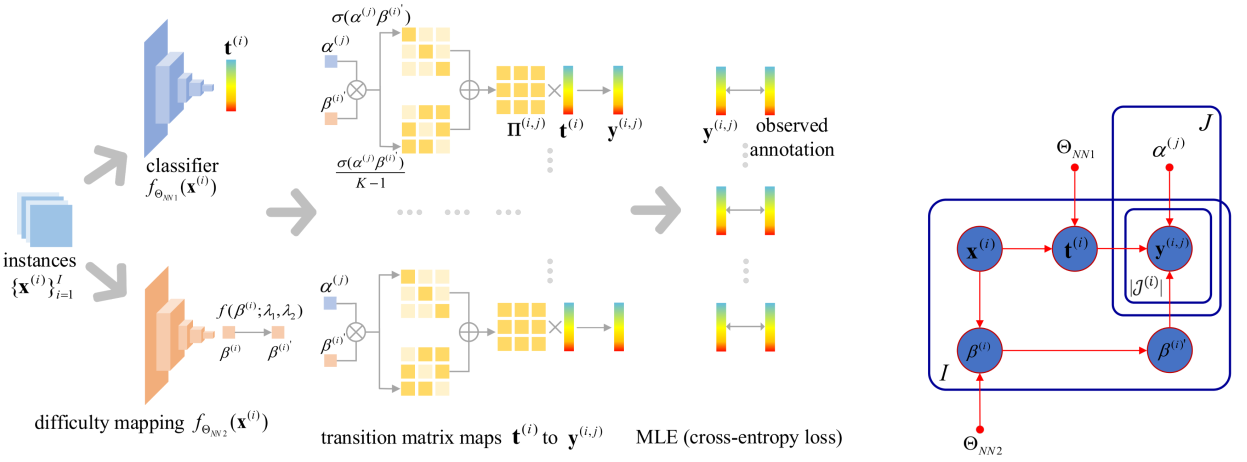
The reparameterization. The essence of the reparameterization is quite simple, *i.e.*, we posit that each column vector of each annotator's *true-transition matrix* $\boldsymbol{\pi}_{:,n}^{(j)'}$ is derived from the corresponding column vector of its *ancestral-transition matrix* $\boldsymbol{\pi}_{:,n}^{(j)}$ through softmax:

$$\boldsymbol{\pi}_{:,n}^{(j)'} = \mathrm{softmax}(\boldsymbol{\pi}_{:,n}^{(j)}), \tag{7}$$

i.e.,

$$\pi_{mn}^{(j)'} = \frac{\exp(\pi_{mn}^{(j)})}{\sum_{m=1}^{K} \exp(\pi_{mn}^{(j)})}, \quad m, n \in \{1, \ldots, K\}, \forall j. \tag{8}$$

Thus, after reparameterization, the whole network is parameterized by $\Theta = \{\Theta_{NN}, \mathbf{\Pi}^{(1)}, \ldots, \mathbf{\Pi}^{(J)}\}$ on which the Adam can be smoothly performed. Furthermore, the $\{\mathbf{\Pi}^{(j)'}\}_{j=1}^{J}$ can perfectly play the roles of annotator transition matrices which strictly satisfy Eq. 5 and Eq. 6.



(a) Framework of SpeeLFC-D

(b) Probabilistic graphical representation of SpeeLFC-D

Figure 3: SpeeLFC-D overview: (1) First, given an input instance $\mathbf{x}^{(i)}$, classifier $f_{\Theta_{NN1}}(\mathbf{x}^{(i)})$ and difficulty mapping function $f_{\Theta_{NN2}}(\mathbf{x}^{(i)})$ generate ground truth $\mathbf{t}^{(i)}$ and $\beta^{(i)}$, respectively. The $\beta^{(i)}$ is then mapped to a positive scalar $\beta^{(i)'} \in (\lambda_1, \lambda_1 + \lambda_2)$ with $f(\beta^{(i)}; \lambda_1, \lambda_2)$, and $\beta^{(i)'}$ represents the instance difficulty. (2) Then, annotation transition matrix $\mathbf{\Pi}^{(i,j)}$ converts ground truth $\mathbf{t}^{(i)}$ into a noisy version of it, *i.e.*, $\mathbf{y}^{(i,j)} = \mathbf{\Pi}^{(i,j)}\mathbf{t}^{(i)}$. Specifically, the transition matrix is simultaneously determined by annotator ability $\alpha^{(j)}$ and instance difficulty $\beta^{(i)'}$. $\sigma(\alpha^{(j)}\beta^{(i)'})$ constitutes each diagonal element of the transition matrix, and $\frac{\sigma(\alpha^{(j)}\beta^{(i)'})}{K-1}$ constitutes each element in other positions. (3) Finally, we just perform the gradient-based optimization algorithm on $\log p(\mathbf{Y}|\mathbf{X};\Theta)$.

Objective and Optimization

Based on the model we have constructed, our optimization objective is to maximize the log-conditional likelihood $\log p(\mathbf{Y}|\mathbf{X};\Theta)$ of the observed crowdsourced annotations w.r.t. $\Theta = \{\Theta_{NN}, \mathbf{\Pi}^{(1)}, \ldots, \mathbf{\Pi}^{(J)}\}$:

$$\begin{aligned}&\log p(\mathbf{Y}|\mathbf{X};\Theta)\\ =&\sum_{i=1}^{I}\sum_{j\in J^{(i)}} \log p(\mathbf{y}^{(i,j)}|\mathbf{x}^{(i)};\Theta)\\ =&\sum_{i=1}^{I}\sum_{j\in J^{(i)}} \log\Big\{\sum_{k=1}^{K}\Big[p(ind(\mathbf{t}^{(i)})=k|\mathbf{x}^{(i)};\Theta_{NN})\cdot\\ &p(\mathbf{y}^{(i,j)}|\,\mathrm{ind}(\mathbf{t}^{(i)})=k;\mathbf{\Pi}^{(j)})\Big]\Big\}\\ :=&\,\mathcal{U}(\Theta).\end{aligned} \tag{9}$$

Based on Eq. 3-8, the $\mathcal{U}(\Theta)$ provides a unified objective function for optimization in SpeeLFC, which can be done with standard stochastic optimization techniques, such as SGD or Adam [Kingma and Ba, 2014]. In fact, according to Eq. 9, we only need to use $\mathbf{\Pi}^{(i,j)}\mathbf{t}^{(i)}$ to obtain the distribution $\mathbf{y}^{(i,j)}$ as mentioned in Section 2, and then maximize the log-likelihood of the corresponding observed annotation.

3.2 SpeeLFC-D

Overall, the probabilistic generative process of crowdsourced annotations constructed in SpeeLFC-D is similar to SpeeLFC. As shown in Figure 3, for each instance $x^{(i)}$, its ground truth $\mathbf{t}^{(i)}$ is generated by:

$$\mathbf{t}^{(i)}|\mathbf{x}^{(i)};\Theta_{NN1} \sim \mathrm{Cat}(\mathbf{t}^{(i)}|f_{\Theta_{NN1}}(x^{(i)})). \tag{10}$$

Then, we construct the instance difficulty $\beta^{(i)'}$, representing how easily the instance can be annotated correctly by annotators. For images, it may be related to the resolution of the image. We assume that the instance difficulty is obtained by mapping the instance features through the instance difficulty mapping function $f_{\Theta_{NN2}}(\mathbf{x}^{(i)})$ and a subsequent auxiliary function. That is:

$$\beta^{(i)} = f_{\Theta_{NN2}}(\mathbf{x}^{(i)}), \tag{11}$$

$$\beta^{(i)'} = \lambda_1 + \lambda_2 \cdot \mathrm{sigmoid}(\beta^{(i)}), \tag{12}$$

where $f_{\Theta_{NN2}}(\mathbf{x}^{(i)})$ is a neural network function parametrized by Θ_{NN2}, and $\lambda_1, \lambda_2 \in (0, +\infty)$ are hyper-parameters. Eq. 12 helps us map the original $\beta^{(i)} \in (-\infty, +\infty)$ to a positive range $(\lambda_1, \lambda_1 + \lambda_2)$.

Last, we model each annotator ability with an annotator-specific scalar $\alpha^{(j)} \in (-\infty, +\infty)$, and we assume that the higher the annotator ability $\alpha^{(j)}$ and the less instance difficulty (meaning bigger $\beta^{(i)'}$ in our model), the greater the probability that the annotator will annotate the instance correctly, and vice versa. Thus, we have the following construction:

$$\begin{aligned}&p(\mathrm{ind}(\mathbf{y}^{(i,j)}) = k|\,\mathrm{ind}(\mathbf{t}^{(i)}) = k; \alpha^{(j)}, \beta^{(i)'})\\ =&\sigma(\alpha^{(j)}\beta^{(i)'}), \quad k \in \{1, 2, \ldots, K\},\end{aligned} \tag{13}$$

$$\sigma(\alpha^{(j)}\beta^{(i)'}) = \frac{1}{1 + \exp(-\alpha^{(j)}\beta^{(i)'})}, \tag{14}$$

where $\mathrm{ind}(\cdot)$ is to obtain the location index of the value 1 in the one-hot vector ($\mathrm{ind}(\cdot) \in \{1, 2, \ldots, K\}$). Correspond-

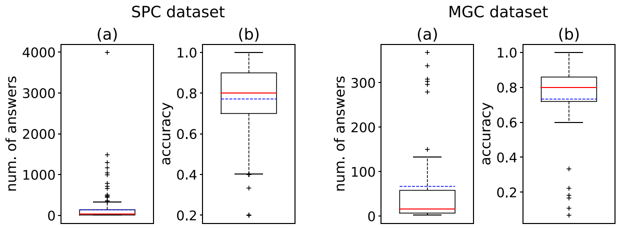


Figure 4: Boxplots for the number of annotations (a) and the accuracies (b) of the AMT annotators for the SPC and the MGC datasets.

Model	Accuracy	Model	Accuracy
MV + MLP	0.7226	PC-GP	0.7263
GLAD + MLP	0.7289	Raykar	0.4867
Crowd-Layer	0.7307	Raykar(w/prior)	0.7078
Semi-LFC	0.7290	MA-LR	0.7240
PC	0.7261	**SpeeLFC-D**	**0.7336**
SpeeLFC	**0.7314**		

Table 1: Accuracy on the SPC dataset.

ingly, the probability that the j^{th} annotator with $\alpha^{(j)}$ will annotate the instance $\mathbf{x}^{(i)}$ with $\beta^{(i)'}$ incorrectly is as follows:

$$\begin{aligned}&p(\mathrm{ind}(\mathbf{y}^{(i,j)}) = k' \,|\, \mathrm{ind}(\mathbf{t}^{(i)}) = k; \alpha^{(j)}, \beta^{(i)'})\\ =&\frac{1}{K-1}[1 - \sigma(\alpha^{(j)}\beta^{(i)'})], \qquad (15)\\ &k \in \{1, 2, \ldots, K\}, k' \in \{1, 2, \ldots, K\} - \{k\}.\end{aligned}$$

In fact, for each annotator, the above construction will separately models a transition matrix $\mathbf{\Pi}^{(i,j)}$ for her performance pattern on each instance $\mathbf{x}^{(i)}$. Thus, the diagonal elements of the matrix $\mathbf{\Pi}^{(i,j)}$ are obtained from Eq. 13, while the elements on other positions are obtained from Eq. 15.

Based on the model we construct, the objective and optimization in SpeeLFC-D are the same as in SpeeLFC.

4 Experiments

4.1 Settings

Datasets and Compared Methods

We performed experiments on two real-world datasets labeled from Amazon Mechanical Turk (AMT), *i.e.* the Sentiment Polarity Classification (SPC) dataset [Rodrigues *et al.*, 2013] and the Music Genre Classification (MGC) dataset [Rodrigues *et al.*, 2013]. The SPC dataset contains 5000 sentences (with crowdsourced annotations) from movie reviews extracted from the website RottenTomatoes.com and their sentiment polarity was classified as positive or negative, while the MGC dataset contains 700 samples (with crowdsourced annotations) of songs with 30 seconds in length and were divided into 10 different music genres (e.g., classical, country, disco). The two datasets received a total of 27747 and 2946 annotations from 203 and 44 distinct annotators on AMT, respectively. For both tasks, 5429 and 300 instances are provided as test sets respectively. Figure 4 shows the distributions of the number of annotations provided by each annotator and their accuracies.

Model	Accuracy	Model	Accuracy
MV + MLP	0.6267	Raykar	0.1200
GLAD + MLP	0.6603	Raykar(w/prior)	0.6300
Crowd-Layer	0.4287	MA-LR	0.6400
SpeeLFC	**0.6923**	**SpeeLFC-D**	**0.6833**

Table 2: Accuracy on the MGC dataset.

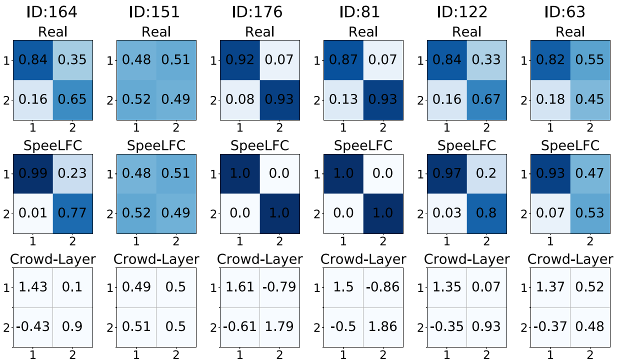


Figure 5: Comparison among the real transition matrices, the corresponding transition matrices estimated by SpeeLFC and the corresponding matrices estimated by Crowd-Layer on the SPC dataset. Note that the corresponding probability values are also shown in the transition matrices.

SpeeLFC and SpeeLFC-D are compared with: MV + MLP (Majority Voting + Multilayer Perceptron), first the ground truth was estimated by MV, and then a general MLP was trained using the estimated ground truth; GLAD + MLP, similar to MV + MLP, except that the truth inference algorithm is GLAD [Whitehill *et al.*, 2009]; Crowd-Layer [Rodrigues and Pereira, 2018]; Raykar and Raykar(w/prior) [Raykar *et al.*, 2010], the latter is the Raykar model with the prior for the annotator reliability; MA-LR [Rodrigues *et al.*, 2013].

Implementation Details

For the SPC dataset, we set the classifier in both SpeeLFC and SpeeLFC-D as an MLP with one hidden layer (with 1200 units, ReLU activations), using 50% dropout and Adam stochastic optimization [Kingma and Ba, 2014]. The learning rate is 0.0001, batch-size is 64, and epoch number is 200. In addition, the function $f_{\Theta_{NN2}}(\mathbf{x}^{(i)})$ in SpeeLFC-D is also an MLP with one hidden layer (with 128 units, ReLU activations, 50% dropout). And the values of the hyperparameters λ_1 and λ_2 in SpeeLFC-D are 0.001 and 100, respectively. In SpeeLFC, the values on the diagonal elements of $\mathbf{\Pi}^{(j)}$ $(j = 1, \ldots, J)$ were initially set to 1.4, and the other values were set to 1. $\alpha^{(j)}$ $(j = 1, \ldots, J)$ in SpeeLFC-D was initially set to 0.028. For the MGC dataset, Rodrigues and Pereira [2013] uses deep learning representation methods to extract 124 features for each instance. We directly set the classifier in both SpeeLFC and SpeeLFC-D as an multiclass Logic Regression just like Rodrigues and Pereira [2013], that is, the MLP without hidden layer. The learning rate is 0.001,

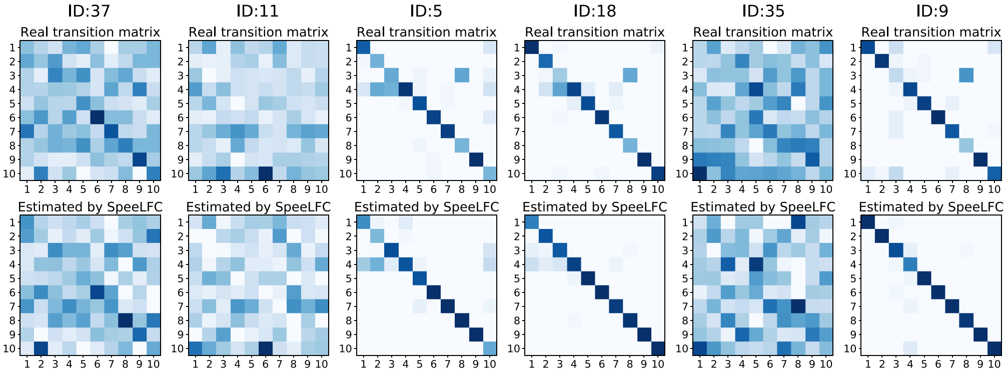


Figure 6: Comparison between the real transition matrices and the corresponding transition matrices estimated by SpeeLFC on the MGC dataset. Note that the color intensity of the cells increases with the relative magnitude of the value.

and epoch number is 5000. The function $f_{\Theta_{NN2}}(\mathbf{x}^{(i)})$ in SpeeLFC-D is an MLP with one hidden layer (with 32 units, ReLU activations, 50% dropout, the learning rate on θ_{NN2} is 0.0001). In SpeeLFC, the values on the diagonal elements of $\mathbf{\Pi}^{(j)}$ were initially set to 4.7, and the other values are 1. In addition, the other settings are the same as those in the SPC dataset. The settings of the compared methods are the same as ours.

4.2 Results

The accuracies of the classifiers on the two datasets are shown in Table 1 and Table 2, respectively. Among them, the results of Semi-LFC [Atarashi *et al.*, 2018], PC [Kajino *et al.*, 2012], PC-GP [Kajino *et al.*, 2012], Raykar [Raykar *et al.*, 2010], Raykar (w/prior) [Raykar *et al.*, 2010], and MA-LR [Rodrigues *et al.*, 2013], are taken from Atarashi *et al.* [2018] and Rodrigues and Pereira [2013]. We observe that the proposed SpeeLFC and SpeeLFC-D achieve better results on both datasets as compared to the baselines and state-of-the-art models. On the SPC dataset, because of the inherent task difficulty and the relatively more number of annotations received for each instance (5.55 on average), most methods exhibit similarly high accuracies. In particular, the highlight we can find is that our proposed models even surpass the recently released semi-supervised learning-from-crowds model called Semi-LFC [Atarashi *et al.*, 2018], even though the Semi-LFC uses more information (*i.e.*, unlabeled data) than ours.

In the case of the MGC dataset for the ten-class classification scenario, the performance of these models shows a larger difference. We observe that the proposed SpeeLFC and SpeeLFC-D show more obvious advantages over the compared methods; in particular, both of them significantly outperform Crowd-Layer by a large margin. We conducted an additional experiment to further optimize the network architecture of Crowd-Layer, by adding a hidden layer containing 128 units (50% dropout), and still found the performance (accuracy: 0.5850) to be largely below our proposed models. Those results clearly demonstrate the importance of probabilistically interpretable parameters in end-to-end learning from crowds. Additionally, we further observe in Table 2 that SpeeLFC outperforms SpeeLFC-D. Recall that SpeeLFC models each annotator with $K \times K$ parameters while SpeeLFC-D only models each annotator with one parameter. Such a comparison result shows that SpeeLFC is more robust for the ten-class classification task with more parameters to be learned.

We now analyze the annotator transition matrices estimated by the proposed SpeeLFC. Similar to Atarashi *et al.* [2018] and Rodrigues *et al.* [2017], the six annotators with the largest number of annotations were selected. The results are shown in Figure 5 and Figure 6. Note that the corresponding real annotator transition matrices are calculated based on their annotations and the ground truth. The accurate estimation of annotator performance pattern verifies the validity of the whole network we build and demonstrates that the structured probabilistic models provide us with an explainable and principled solution based on an end-to-end learning manner using pure backpropagation. For comparison, we also show in Figure 5 the corresponding matrices estimated with Crowd-Layer: among the six matrices, five contain negative numbers, thus cannot express the probability relationship between the annotation and ground truth.

5 Conclusion

This paper presents two novel models for end-to-end learning from crowds. We present SpeeLFC, a structured probabilistic model that benefits both from the expressiveness and reasonableness of probabilistic approaches for representing annotator reliability and from the effectiveness of the end-to-end approach for training neural networks. Moreover, we present SpeeLFC-D, which further models instance difficulty in end-to-end learning from crowds. Evaluation of our proposed methods on real-world datasets shows that both methods improve the state-of-the-art.

Acknowledgments

This work was supported partly by National Key Research and Development Program of China under Grant No.2019YFB1705902, partly by National Natural Science Foundation under Grant No.(61932007, 61972013, 61421003). We thank the authors of Rodrigues *et al.* [2018] for the sharing of the two datasets that greatly support our experiments.

References

[Albarqouni *et al.*, 2016] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5):1313–1321, 2016.

[Atarashi *et al.*, 2018] Kyohei Atarashi, Satoshi Oyama, and Masahito Kurihara. Semi-supervised learning from crowds using deep generative models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Bi *et al.*, 2014] Wei Bi, Liwei Wang, James T Kwok, and Zhuowen Tu. Learning to predict from crowdsourced data. In *UAI*, pages 82–91, 2014.

[Bishop, 2006] Christopher M Bishop. *Pattern recognition and machine learning*. Springer Science+ Business Media, 2006.

[Dawid and Skene, 1979] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

[Han *et al.*, 2016] Tao Han, Hailong Sun, Yangqiu Song, Yili Fang, and Xudong Liu. Incorporating external knowledge into crowd intelligence for more specific knowledge acquisition. In *IJCAI*, volume 2016, pages 1541–1547, 2016.

[Kajino *et al.*, 2012] Hiroshi Kajino, Yuta Tsuboi, and Hisashi Kashima. A convex formulation for learning from crowds. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Raykar *et al.*, 2010] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

[Rodrigues and Pereira, 2018] Filipe Rodrigues and Francisco C Pereira. Deep learning from crowds. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Rodrigues *et al.*, 2013] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12):1428–1436, 2013.

[Rodrigues *et al.*, 2017] Filipe Rodrigues, Mariana Lourenço, Bernardete Ribeiro, and Francisco C Pereira. Learning supervised topic models for classification and regression from crowds. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2409–2422, 2017.

[Welinder *et al.*, 2010] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pages 2424–2432, 2010.

[Whitehill *et al.*, 2009] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

[Yan *et al.*, 2012] Yan Yan, Rómer Rosales, Glenn Fung, and Jennifer Dy. Modeling multiple annotator expertise in the semi-supervised learning scenario. *arXiv preprint arXiv:1203.3529*, 2012.

[Yang *et al.*, 2019] Jie Yang, Alisa Smirnova, Dingqi Yang, Gianluca Demartini, Yuan Lu, and Philippe Cudré-Mauroux. Scalpel-cd: leveraging crowdsourcing and deep probabilistic modeling for debugging noisy training data. In *The World Wide Web Conference*, pages 2158–2168. ACM, 2019.

[Zhang *et al.*, 2016] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[Zheng *et al.*, 2017] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017.