

Exploiting Heterogeneous Graph Neural Networks with Latent Worker/Task Correlation Information for Label Aggregation in Crowdsourcing

HANLU WU, Zhejiang University, China

TENGFEE MA, IBM T. J. Watson Research Center, USA

LINGFEI WU, JD.COM Silicon Valley Research Center, USA

FANGLI XU, Squirrel AI Learning, USA

SHOULING JI*, Zhejiang University, China

Crowdsourcing has attracted much attention for its convenience to collect labels from non-expert workers instead of experts. However, due to the high level of noise from the non-experts, a label aggregation model that infers the true label from noisy crowdsourced labels is required. In this paper, we propose a novel framework based on graph neural networks for aggregating crowd labels. We construct a heterogeneous graph between workers and tasks and derive a new graph neural network to learn the representations of nodes and the true labels. Besides, we exploit the unknown latent interaction between the same type of nodes (workers or tasks) by adding a homogeneous attention layer in the graph neural networks. Experimental results on 13 real-world datasets show superior performance over state-of-the-art models.

CCS Concepts: • **Information systems** → **Crowdsourcing**; *Social tagging*; *Data analytics*.

Additional Key Words and Phrases: crowdsourcing, graph neural network, label aggregation

ACM Reference Format:

Hanlu Wu, Tengfei Ma, Lingfei Wu, Fangli Xu, and Shouling Ji. 2018. Exploiting Heterogeneous Graph Neural Networks with Latent Worker/Task Correlation Information for Label Aggregation in Crowdsourcing. *J. ACM* 37, 4, Article 111 (August 2018), 18 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Recent years have witnessed the successful applications of machine learning in many fields, such as computer vision and natural language processing. Most of the machine learning tasks require large amounts of labeled data, however, obtaining labeled data from experts is quite expensive and time-consuming. Therefore, crowdsourcing has flourished as

*Shouling Ji is the corresponding author.

This work was partly supported by NSFC under No.61772466, U1936215, and U1836202, the National Key Research and Development Program of China under No.2020YFB2103802, 2018YFB0804102, and 2020AAA0140004, the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under No.LR19F020003, and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform).

Authors' addresses: Hanlu Wu, Zhejiang University, China, wuhanlu@zju.edu.cn; Tengfei Ma, IBM T. J. Watson Research Center, USA, tengfei.ma1@ibm.com; Lingfei Wu, JD.COM Silicon Valley Research Center, USA, lwu@email.wm.edu; Fangli Xu, Squirrel AI Learning, USA, lili@yixue.us; Shouling Ji, Zhejiang University, China, sji@zju.edu.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

one of the most important tools for data annotation and labeling. With online platforms such as Amazon Mechanical Turk (AMT) ¹ and CrowdFlower ², one can get efficient and inexpensive access to crowdsourced resources.

Crowdsourcing systems generally break down a time-consuming task into more manageable microtasks, which can then be accomplished by distributed workforce independently. For example, to obtain enough labels for training a classifier, one can break down the labeling task into microtasks by assigning non-overlapping items to different workers for annotation. However, this way of task assignment suffers from the unprofessional nature of the workers, which leads to highly noisy data. A common practice is to increase the overlap of assignments between workers, i.e., collecting multiple labels from different workers for each single item. Then the ground-truth label can be induced from the noisy crowdsourced labels. By aggregating the wisdom of crowds, one can reduce the error rates and thereby improve the quality of the labeled data [42].

An intuitive strategy for crowdsourcing label aggregation is majority voting [34]. However, this simple strategy is deficient for it ignores some important factors, such as worker ability. In a crowdsourcing system, workers usually show different expertise or reliability within a certain task, and a worker may be acquainted with some sort of tasks but fail when facing some others. What’s worse is that a malicious worker may even submit wrong answers intentionally. As a consequence, majority voting, which equally treats each worker, can hardly make a reliable enough inference.

If the exact ability of each worker is known, the ground-truth label can be better inferred by weighted majority voting [29]. Based on this assumption, Tao et al. managed to learn the weight of each worker for weighted majority voting [43]. However, we believe that the abilities of workers cannot be simply quantified by a single value. A worker may have a relatively strong ability in labeling one type of items correctly, but not good at another type of items (e.g. some workers are familiar with road signs and thus more professional in labeling related items, but they know little about animals and are easy to make mistakes in distinguishing animals.). Findings from existing work on crowdsourcing illustrate that it is crucial to model multi-dimensional latent features of workers [10, 43, 54, 55], which indicates different aspects of the workers. Meanwhile, the latent features of items (in the following, an item to be labeled is also called as a *task*) also count for a lot. The difficulty of a task impacts the average rating as well as each worker’s ability. Various probabilistic models have been proposed under the assumption that worker abilities and task difficulties are both key factors for inferring true labels [55, 56] and obtained performance superior to majority voting. However, most of them require a delicate design for a sophisticated generative process and complex inference algorithms, and they are difficult to be generalized to large-scale datasets. Besides, there are also some deep learning models that jointly learn a classifier together with the label aggregation model [1, 4, 9, 37]. However, they usually require given features for each task, and different feature extraction strategies or different model structures of the classifier are needed according to the labeling tasks. Hence, the necessity of input task features will reduce the applicability of the model to some extent. In contrast, in this paper we are assumed to know only the task assignments of workers and their labels, and the workers and tasks are simply identified by ID numbers.

In order to model the relationship between workers and tasks, we propose to apply heterogeneous graph neural networks to crowdsourced label aggregation. To construct the graph, we model workers and tasks as two different types of nodes. If a worker and a task is connected by an edge, it indicates that the task was labeled by the worker. The main idea of graph neural networks is to iteratively aggregate information from their local neighborhoods, thus the graph neural network can naturally model the mutual interaction between tasks and workers and learn a good

¹Amazon Mechanical Turk (AMT) can be found in www.mturk.com

²CrowdFlower can be found in www.crowdflower.com

representation for them. We then infer the true label of a task from its representation. In this way, the crowdsourcing label inference problem is turned into a node classification problem in graph neural networks.

Despite the representation power of graph neural networks, in our constructed graphs they can only utilize the assignment relationship between workers and tasks, while ignoring the workers' or tasks' latent relationship. Workers' correlation has been identified as another important factor for increasing the truth label inference in crowdsourcing [27]. Motivated by this observation, we further take into account the latent worker correlation, as well as task correlation, in our model and develop a new heterogeneous graph neural network based framework for crowdsourcing. In addition to the message passing between worker nodes and task nodes, we build an extra layer to implicitly propagate information among the same type of nodes, which has never been explored by previous heterogeneous graph neural networks to the best of our knowledge.

Our contributions are summarized as follows:

- We provide a new perspective for crowd label aggregation in the context of graph representation learning. To the best of our knowledge, it is the first model utilizing graph neural network to solve the crowdsourcing problem.
- Different from existing heterogeneous graph neural networks and most crowd label aggregation methods, our model learns a latent interaction among the same type of nodes to implicitly integrate the worker correlation and task correlation.
- We experiment on 13 real-world crowdsourcing datasets and demonstrate advantageous performance over state-of-the-art models. We also conduct ablation studies to explain the effectiveness of different components.

2 RELATED WORK

2.1 Crowdsourcing

The increasing popularity of crowdsourcing as a labeling tool has led to a lot of attention to solve the issues of noisy crowdsourced labels. The early work of label aggregation can be traced back to [10], which firstly proposed an Expectation-Maximization(EM)-based model to estimate the error rate of patients' answers to clinical problems. This model can be naturally transferred to the label aggregation problem. It utilizes workers' latent aspects by using a confusion matrix indicating the probability of a worker to choose each label for a task given the true label of it.

Many follow-up studies can be viewed as extensions of the Dawid & Skene model [22, 30, 44, 49, 55, 56, 61]. Some work introduced task heterogeneity. In [61], the authors incorporated both abilities and difficulties for workers and tasks respectively and inferred the truth using a min-max entropy principle. Venanzi et al. modeled workers in community clusters to make workers share similar confusion matrices within the community [49]. Khetan and Oh also introduced task difficulty into the Dawid & Skene model and designed an adaptive task assignment scheme to provide more budget for tasks with more difficulty [22]. The GLAD model (Generative model of Labels, Abilities, and Difficulties) considered both the abilities of workers and the difficulties of tasks and can simultaneously infer true labels as well as worker ability and task difficulty [55]. LAA (Label-Aware Autoencoders) trains a classifier and a reconstructor, and the truth is inferred by the classifier as latent features [56]. They also provided two extended models in their paper by considering object ambiguity (LAA-O) or latent aspects (LAA-L). From the above-mentioned work, we can safely draw a conclusion that it's necessary to model the heterogeneity of both workers and tasks. Table 1 compares a few methods in task modeling, worker modeling and correlation modeling (part of this table is quoted from [59]). Different from previous methods, EBCC (enhanced Bayesian classifier combination) additionally captures worker-worker correlations by dividing each true class into several subtypes and modeling the correlations between workers in the subtype level. Their approach

infers true labels using a mean-field variational approach [27]. Inspired by this work, our model also incorporates inner-worker correlation. However, we also model the inner-task correlation in addition.

Other methods have been explored to select workers who can produce high-quality labels. Based on the assumptions that some workers may assign labels casually (these workers are called *spammer*), Raykar and Yu defined a spammer score to rank the workers and proposed an empirical Bayesian algorithm to iteratively eliminate the workers with high spammer score and estimate the ground-truth labels based only on those with low spammer score [35]. Ipeirotis et al. tried to evaluate the score of workers before task assignment and only assign tasks to workers with higher scores [20]. CrowdDQS dynamically issues golden standard questions and estimate the accuracies of workers in real-time, then it can select workers with higher accuracies for task assignment [21]. Tu et al. suggest that the attention of workers changes over time, thus the accuracy of workers can not be kept constant, therefore, they proposed a probabilistic model that takes into account workers’ attention [46]. Compared to these models, this paper focuses on a different scenario and our assumption is that the ability of a worker is diverse but constant (i.e. a worker will always give the same label to the same task).

Table 1. Comparisons of Existing Methods. "×" indicates the model does not consider this aspect.

Method	Task	Worker	Worker-Worker Corr	Task-Task Corr	Worker-Task Corr
MV	×	×	×	×	×
D&S [10]	×	✓	×	×	×
ZC [11]	×	✓	×	×	×
Minimax [61]	×	✓	×	×	✓
GLAD [55]	✓	✓	×	×	×
BCC [23]	×	✓	×	×	×
LFC [36]	×	✓	×	×	×
iBCC-MF [27]	×	✓	×	×	×
EBCC [27]	×	✓	✓	×	×
LAA [56]	✓	✓	×	×	×
CATD [25]	×	✓	×	×	×
PM [2, 26]	×	✓	×	×	×
The proposed	✓	✓	✓	✓	✓

2.2 Graph Neural Networks and General Frameworks

A graph is a structured data consisting of nodes and edges connecting them. Data in many application scenarios has a natural graph structure, such as social networks, molecular structures, etc. In these scenarios, traditional deep learning methods are difficult to apply to the graph data. Therefore, in recent years, there is increasing interest in extending deep learning algorithms to the field of graphs as Graph Neural Networks (GNNs) [7, 19, 24, 38]. GNNs are capable of dealing with non-Euclidean structured data such as protein interaction networks [63], citation networks [24], traffic networks [32], social networks, knowledge graphs [15, 18], device-sharing network [28, 31], and text graph in natural language processing [6] etc.

Some of these scenarios have various types of entities and relations (i.e. nodes and edges in the graph), hence called heterogeneous graphs. Several heterogeneous graph neural networks have been proposed and applied to various domains recently [5, 8, 53, 58]. To illustrate some, Zitnik et al. developed a heterogeneous graph neural network for drug side effect detection [63]; Fan et al. used heterogeneous graph neural networks for product recommendation [13]; Wang et al.

Table 2. Notation and Explanation

Notation	Definitions and Description
u_i	worker node i
v_j	task node j
n	number of workers
m	number of tasks
g_j	the label of task j inferred using majority voting
l_{ij}	crowd label given to task j by worker i
e_{ij}	a one-hot vector indicating the crowd label given to task j by worker i
$\mathcal{N}(i)$	neighborhood of node i
$\mathcal{N}(u_i)$	the set of tasks labeled by worker u_i
$\mathcal{C}(v_j)$	the set of workers assigning labels to task v_j
h_i^t	hidden state of worker or task i
$h^t(u_i)$	hidden state of worker u_i
$h^t(v_j)$	hidden state of task v_j
c_i	a constant coefficient
W_r	weight parameter used in MP1
W^u, W^v	weight parameters used in MP2
W_e^u, W_e^v	weight parameters used in MP2
W_1, W_2	weight parameters used in MP2
b_1, b_2	biases used in MP2
α_{ij}, β_{ij}	attention weights in MP2
W_c^u, W_c^v	weight parameters used in COR
W_3, W_4	weight parameters used in COR
γ_{ij}, δ_{ij}	attention weights in COR

proposed a heterogeneous graph neural network with hierarchical attention mechanism that aggregates information from meta-path based neighbors [53]. To the best of our knowledge, our work is the first trial to combine graph neural networks with the label aggregation problem in crowdsourcing. Moreover, different from previous heterogeneous graph neural networks, our work is the first one modeling the **implicit correlation** among the same type of nodes in a heterogeneous graph.

Some studies on general frameworks for graph neural networks have also emerged [3, 17, 52, 62]. Gilmer et al. proposed message passing neural network (MPNN) which unified various graph neural network approaches [17]. MPNN abstracts these graph neural networks into two phases, message passing phase and readout phase. The message passing phase aggregates information from the neighborhood based on a message function and an update function, and the readout phase is to obtain a representation of the whole graph based on the hidden states of each node. Our model is designed under MPNN framework. Wang et al. proposed non-local neural network (NLNN) to capture the non-local dependencies of nodes [52]. Battaglia et al. unified most of the graph neural networks including MPNN and NLNN by a graph networks (GN) framework [3].

3 PROBLEM STATEMENT AND NOTATIONS

In this paper, we study the crowdsourcing label aggregation problem. To formulate it, assume we have n workers and m tasks. The tasks can be classified into K categories. For each task, a worker needs to select a single label out of K

candidate labels (we only consider the scenario of single-choice tasks, while a multi-choice task can be transformed into a set of single-choice tasks [59, 60]). We denote the label that worker i assigns to task j as $l_{ij} \in \{1, \dots, K\}$. The goal of label aggregation in crowdsourcing is to infer the ground-truth label y_j of each task j . In this work, we assume that we already have ground-truth labels for some tasks, and the task is to predict the remaining unknown labels for other tasks. Note that our method is applicable to both the case that each worker only assigns labels to part of the tasks and the case that each worker assigns labels to all of the tasks.

4 METHOD

In this section, we describe how our method is designed in detail.

We first construct a graph to connect all the workers and tasks as shown in Fig. 1. Then we develop a new heterogeneous graph neural network to encode the worker nodes and task nodes into vector representations. Our new heterogeneous graph neural network contains two types of message passing layers [17]: the layer passing messages between workers and tasks, which captures the worker-task interactions; and the layer passing message among the same types of nodes, which captures the worker-worker correlation and task-task correlation. After we get the node embeddings from the heterogeneous graph neural network, we add a prediction layer to predict the true label of each task.

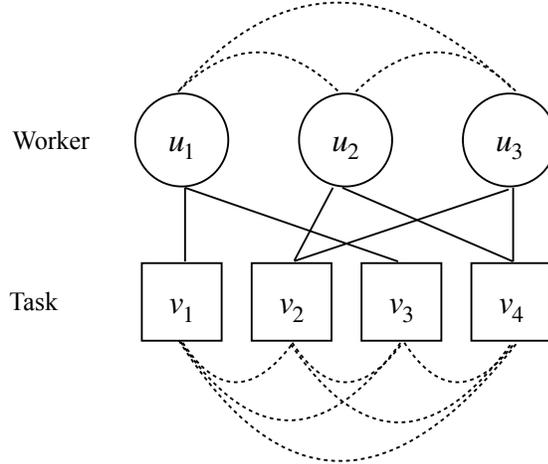


Fig. 1. A worker-task assignment graph and the latent interaction between workers/tasks. u_i indicates the feature of the i^{th} worker, and v_j indicates the feature of the j^{th} task. Solid lines represent that a worker assigns a label to a task, while dashed lines represent the latent correlation between workers or between tasks. For simplicity, on the solid-line edges we omit the crowdsourced labels l_{ij} that workers assign to tasks.

4.1 Motivation and Graph Construction

Most previous methods for crowdsourcing formulate the label aggregation process as a complex generative process that is dependent on either worker ability or task difficulty. For example, in [55],

$$p(l_{ij} = z_j | \alpha_i, \beta_j) = \frac{1}{1 + \exp(-\alpha_i \beta_j)}$$

where l_{ij} is the label that worker i assigned to task j , z_j is the ground truth label of task j , α_i denotes the ability variable of worker i , and β_j denotes the difficulty variable of task j . However, in these models we need to make delicate assumptions for the priors of these variables (e.g. Dirichlet priors) and carefully design a generative process, in order to make the inference tractable. In addition, the latent variables are generally scalars. This largely limits the modeling capacity because the worker’s ability and task’s difficulty may contain different aspects.

Inspired by the recent success of deep learning, we aim at using a deep neural network to explicitly learn the embeddings of worker features and task features which can determine the true labels. Considering that the labeling process can be represented as a graph, a graph neural network is a natural solution to the embedding problem.

We show an example of the worker-task assignment graph in Fig. 1. In the graph, the nodes are either workers or tasks. If a worker u_i assigns a label to v_j , there will be an edge connecting u_i and v_j , and the edge feature is the one-hot crowdsourced label vector $e_{ij} \in \{0, 1\}^K$ which is derived from the label l_{ij} .

To initialize the features of nodes, we followed the feature representation method in [16]. We denote g_j as the label of the task u_i inferred by majority voting. For a worker node u_i , we calculate its features as below:

$$f(u_i) = \frac{|\{j \in \mathcal{N}(u_i) | l_{ij} = g_j\}|}{|\mathcal{N}(u_i)|} \quad (1)$$

For a task node v_j ,

$$f(v_j) = \frac{|\{i \in C(v_j) | l_{ij} \neq g_j\}|}{|C(v_j)|} \quad (2)$$

where $\mathcal{N}(u_i)$ denotes the set of tasks labeled by worker u_i and $C(v_j)$ is the set of workers that assigned labels to task v_j . $|\cdot|$ denotes the cardinality of a set. This is based on an assumption that if the labels given by a worker is the same as the majority of people most of the time, he/she should have good labeling ability; for a task, the more worker who assigned different labels from the majority voting label to it, the more difficult the task can be. We fill the d -dimensional feature vector with the same value of $f(u_i)$ for worker u_i and the same way for tasks. We also tried random initialization, the results can be found in Table 3.

Table 3. Comparison of Feature Initialization Methods.

Datasets	Our Initialization Method	Random Initialization
bird	0.8610±0.0508	0.8517±0.0376
flowers	0.8638±0.0133	0.8600±0.0169
web	0.9734±0.0215	0.9284±0.0124
dog	0.8243±0.0088	0.8175±0.0098
rte	0.9269±0.0104	0.9259±0.0103
SP	0.9149±0.0091	0.9044±0.0045
SP*	0.9445±0.0025	0.9425±0.0040
ZC _{all}	0.9076±0.0162	0.9012±0.0184
ZC _{in}	0.7942±0.0071	0.7828±0.0071
ZC _{us}	0.9130±0.0069	0.9034±0.0078
face	0.6635±0.0118	0.6682±0.0126
product	0.9363±0.0019	0.9365±0.0023
sentiment	0.9608±0.0076	0.9560±0.0060

4.2 Message Passing Between Workers and Tasks

Given the worker-task assignment graph, we cast the label aggregation problem as a node prediction problem in a heterogeneous graph neural network. To this aim, we develop a non-linear multi-layer message passing scheme for the graph node embedding. Message passing has been a key operation for many graph neural networks [17, 63]. The key idea is to propagate the information across all the edges of the graph in each layer. To illustrate, in the case of the worker-task graph, a worker's embedding is obviously impacted by its assigning labels and the corresponding tasks; and a task's embedding can also be inferred by the interaction with the workers who assign labels to it. In this paper, we implement two versions of message passing schemes between workers and tasks, denoted as MP1 and MP2 separately.

4.2.1 MP1. Following RGCN [39], one intuitive idea of message passing to update the hidden states of worker nodes and task nodes is the following formula, which we call MP1:

$$\mathbf{h}_i^{t+1} = \mathbf{h}_i^t + \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{W}_r \mathbf{h}_j^t \quad (3)$$

where $\mathcal{N}(i)$ represents the neighborhood of node i . When i is a task node, $\mathcal{N}(i)$ denotes a set of workers that have assigned labels to it; When i is a worker node, $\mathcal{N}(i)$ stands for a set of tasks that worker i has assigned labels to. \mathbf{W}_r is a matrix parameter for the edge label $l_{ij} = r$. In this way, we pass the message from workers to tasks and from tasks to workers.

4.2.2 MP2. The above message passing scheme assumes the neighbors have the same weight in the update function. This may lose importance information of different nodes. We can also employ the attention mechanism to re-weight the messages and derive another message passing scheme, MP2.

For a worker \mathbf{u}_i with a hidden state $\mathbf{h}^t(\mathbf{u}_i)$, we update $\mathbf{h}^t(\mathbf{u}_i)$ by the following formula:

$$\begin{aligned} \mathbf{h}^{t+1}(\mathbf{u}_i) &= \phi \left(c_i \mathbf{h}^t(\mathbf{u}_i) + (1 - c_i) \right. \\ &\quad \left. \sum_{\mathbf{v}_j \in \mathcal{N}(\mathbf{u}_i)} \alpha_{ij} M_t^u(\mathbf{h}^t(\mathbf{u}_i), \mathbf{h}^t(\mathbf{v}_j), \mathbf{e}_{ij}) \right) \end{aligned} \quad (4)$$

where M_t^u is the message function, ϕ is a nonlinear activation function (in this work we use ReLU), and $c_i \in [0, 1]$ is a weight. In our case, the interaction between a worker and a task not only contains the worker/task node features, but also include the information of the crowdsourced labels. So our message function is calculated by taking into account both the node and the edge features. We first use a learnable matrix \mathbf{W}_e^u to embed the edge vector \mathbf{e}_{ij} into an embedding vector and then concatenate it with the node features. Then we use an attention mechanism to re-weight the messages from different edges.

$$M_t^u(\mathbf{h}^t(\mathbf{u}_i), \mathbf{h}^t(\mathbf{v}_j), \mathbf{e}_{ij}) = \phi \left(\mathbf{W}^u \left(\mathbf{h}^t(\mathbf{v}_j) \oplus (\mathbf{W}_e^u \mathbf{e}_{ij}) \right) \right) \quad (5)$$

where \mathbf{W}^u is a parameter matrix, α_{ij} is the attention weight calculated by

$$\alpha_{ij} = \frac{\exp(\mathbf{W}_1(M_{ij} \oplus \mathbf{h}^t(\mathbf{u}_i) + b_1))}{\sum_k \exp(\mathbf{W}_1(M_{ik} \oplus \mathbf{h}^t(\mathbf{u}_i) + b_1))} \quad (6)$$

here we simplify $M_t^u(\mathbf{h}^t(\mathbf{u}_i), \mathbf{h}^t(\mathbf{v}_j), \mathbf{e}_{ij})$ as M_{ij} .

Then we pass the messages from the workers to tasks. Similar to the above message passing phase, for each task \mathbf{v}_j , we also receive the messages from its connected edges and workers:

$$M_i^v(\mathbf{h}^t(\mathbf{v}_j), \mathbf{h}^t(\mathbf{u}_i), \mathbf{e}_{ij}) = \phi\left(\mathbf{W}^v \left(\mathbf{h}^t(\mathbf{u}_i) \oplus (\mathbf{W}_e^v \mathbf{e}_{ij})\right)\right) \quad (7)$$

We use a different matrix \mathbf{W}_e^v for edge embedding, and a different parameter matrix \mathbf{W}^v . The attention weights are derived similarly by

$$\beta_{ji} = \frac{\exp(\mathbf{W}_2(M_{ji} \oplus \mathbf{h}^t(\mathbf{v}_j) + b_2))}{\sum_k \exp(\mathbf{W}_2(M_{jk} \oplus \mathbf{h}^t(\mathbf{v}_j) + b_2))} \quad (8)$$

Here we simplify $M_i^v(\mathbf{h}^t(\mathbf{v}_j), \mathbf{h}^t(\mathbf{u}_i), \mathbf{e}_{ij})$ as M_{ji} . Aggregating the messages from all the edges, we obtain the updated task embedding

$$\begin{aligned} \mathbf{h}^{t+1}(\mathbf{v}_j) &= \phi\left(c_i \mathbf{h}^t(\mathbf{v}_j) + (1 - c_i) \right. \\ &\quad \left. \sum_{\mathbf{v}_j \in \mathcal{N}(\mathbf{u}_i)} \beta_{ji} M_i^v(\mathbf{h}^t(\mathbf{v}_j), \mathbf{h}^t(\mathbf{u}_i), \mathbf{e}_{ij})\right) \end{aligned} \quad (9)$$

4.3 COR: Latent Correlation Between Workers/Tasks

The above message passing layer (either MP1 or MP2) explores the interaction between workers and tasks along the explicit edges which represent the assignment relationship. In practice, there may be also latent interaction/correlation among the same type of nodes (i.e. workers or tasks). For example, if two workers belong to the same community [49], or they are close friends in a social network, they may have highly correlated preference or make similar mistakes in the labeling process. As to tasks, if their content is similar or they belong to the same category, it is highly possible that their labels have correlations. However, in a crowdsourcing platform, the explicit relationship among the workers or the tasks is often unknown. In this work, we develop a new layer to model the implicit inner-worker correlation and inner-task correlation and integrate the information into our new heterogeneous graph neural network. We denote this layer as COR.

Implicit worker correlation has been exploited in some Bayesian models before and demonstrated useful [4, 27, 49]. However, it is never explored in previous heterogeneous graph neural networks. Our model is also related to non-local neural networks [52] and self-attention models [47], which utilize long-range dependency of the inputs and improves the performance a lot.

Generally, a (heterogeneous) graph neural network requires to know the complete graph structures, i.e. all the edges. To utilize the correlation between the same type of nodes, we essentially add implicit edges among workers/tasks (based on some correlation function), as shown in Fig. 1 (dashed lines).

Specifically in our model, for worker nodes, we assume that each node can be implicitly correlated to each of the other worker nodes. This is based on the assumption that even though two workers are not connected in the worker-task assignment graph (i.e. the two workers do not assign labels to the same task), they can still have some kind of implicit correlation between them. But when we are faced with a quite large dataset, we can approximately reduce the number of neighbor nodes in the correlation layer to accelerate the message passing process. Two simple strategies are suggested, one is uniform sampling, the other is to select the 2-hop neighborhood in the worker-task assignment graph, i.e. only to capture the relations between two workers who share at least one task and between two tasks that are assigned to at least one same worker. Table 4 shows the performance of our final model that using different neighborhood sampling strategies in the COR layer, both of the strategies have quite close performance to the original fully connected network. Inspired by [48], we update the worker embeddings as follows:

Table 4. A comparison between different neighbourhood sampling strategy.

Datasets	Fully Connected	Uniform Sampling	2-Hop Neighbourhood
bird	0.8610±0.0508	0.8402±0.0306	0.8449±0.0337
flowers	0.8638±0.0133	0.8688±0.0153	0.8638±0.0143
web	0.9734±0.0215	0.9703±0.0272	0.9852±0.0069
dog	0.8243±0.0088	0.8299±0.0101	0.8169±0.0138
rte	0.9269±0.0104	0.9263±0.0068	0.9284±0.0074
SP	0.9149±0.0091	0.9073±0.0081	0.9116±0.0080
SP*	0.9445±0.0025	0.9420±0.0033	0.9425±0.0025
ZC _{all}	0.9076±0.0162	0.9006±0.0088	0.9050±0.0081
ZC _{in}	0.7942±0.0071	0.7852±0.0031	0.7832±0.0096
ZC _{us}	0.9130±0.0069	0.9062±0.0090	0.9022±0.0150
face	0.6635±0.0118	0.6665±0.0226	0.6670±0.0136
product	0.9363±0.0019	0.9354±0.0023	0.9351±0.0014
sentiment	0.9608±0.0076	0.9588±0.0099	0.9583±0.0082

$$\mathbf{h}^{t+1}(\mathbf{u}_i) = \sigma \left(\sum_{\mathbf{u}_j \in \mathcal{N}} \gamma_{ij} \mathbf{W}_c^u \mathbf{h}^t(\mathbf{u}_j) \right) \quad (10)$$

where σ is a non-linear activation function which is ReLU in our experiment. \mathcal{N} denotes the set of all worker nodes including \mathbf{u}_i . \mathbf{W}_c^u represents a parameter matrix. γ_{ij} is the attention weight calculated by

$$\gamma_{ij} = \frac{\exp \left(\sigma \left(\mathbf{a}^T (\mathbf{W}_3 \mathbf{h}^t(\mathbf{u}_i) \oplus \mathbf{W}_3 \mathbf{h}^t(\mathbf{u}_j)) \right) \right)}{\sum_{\mathbf{u}_k \in \mathcal{N}} \exp \left(\sigma \left(\mathbf{a}^T (\mathbf{W}_3 \mathbf{h}^t(\mathbf{u}_i) \oplus \mathbf{W}_3 \mathbf{h}^t(\mathbf{u}_k)) \right) \right)} \quad (11)$$

where \mathbf{a} is a weight vector. We update the embeddings of task nodes in the same way as worker nodes, see the following equations. In our experiment, we found that only one head attention is enough for our task.

$$\mathbf{h}^{t+1}(\mathbf{v}_j) = \sigma \left(\sum_{\mathbf{v}_i \in \mathcal{C}} \delta_{ij} \mathbf{W}_c^v \mathbf{h}^t(\mathbf{v}_i) \right) \quad (12)$$

$$\delta_{ij} = \frac{\exp \left(\sigma \left(\mathbf{b}^T (\mathbf{W}_4 \mathbf{h}^t(\mathbf{v}_j) \oplus \mathbf{W}_4 \mathbf{h}^t(\mathbf{v}_i)) \right) \right)}{\sum_{\mathbf{v}_k \in \mathcal{C}} \exp \left(\sigma \left(\mathbf{b}^T (\mathbf{W}_4 \mathbf{h}^t(\mathbf{v}_j) \oplus \mathbf{W}_4 \mathbf{h}^t(\mathbf{v}_k)) \right) \right)} \quad (13)$$

We analyze the complexity of our model in terms of each layer. We can split the edges into three categories: worker-worker, worker-task, task-task. Assume the worker-task edge set is \mathcal{E} , since we pass the messages from all these edges in MP1 layer, the complexity of MP1 layer is $O(|\mathcal{E}|d_t d_{t+1})$ where d_t is the dimension of node embeddings at the t^{th} -layer. The complexity of MP2 layer is $O(|\mathcal{E}|(d_t + d_e)d_{t+1})$ where d_e is the dimension of the edge vector. The complexity of the correlation layer will be $O((n^2 + m^2)d_t d_{t+1})$. To reduce the complexity, we can use random sampling to sample only a subset of nodes as neighborhoods, or we can only use 2-hop neighborhoods in the correlation layer. As shown in table 4, these approximations do not comprise much performance.

4.4 Prediction and Training

In previous sections, we introduced the message passing layer between workers and tasks, and the message passing layer between the same type of nodes. These layers can be stacked multiple times to get the final embeddings of workers

and tasks. Then we can use the final task embeddings to predict their true labels. For a task v_j with the final embedding $\mathbf{h}(v_j)$, we predict its label by:

$$\hat{\mathbf{y}}_j = \text{softmax}(\mathbf{W}_3 \mathbf{h}(v_j) + b_3) \quad (14)$$

We use the cross-entropy loss between the prediction $\hat{\mathbf{y}}_j (1 \leq j \leq m)$ and the true labels $\mathbf{y}_j (1 \leq j \leq m)$ as the loss function,

$$L = \sum_{v_j \in V_{train}, 1 \leq k \leq K} y_{jk} \log \hat{y}_{jk} + (1 - y_{jk}) \log(1 - \hat{y}_{jk}) \quad (15)$$

where y_{jk} and \hat{y}_{jk} are the k^{th} elements of \mathbf{y}_j and $\hat{\mathbf{y}}_j$ separately. The model is then trained on the training tasks V_{train} with known true labels with Adam and early stopping. The whole algorithm of our model MP2+COR+MP2 (i.e. stacked by an MP2 layer, a COR layer and another MP2 layer) can be expressed as below:

Algorithm 1 MP2+COR+MP2

Input: the worker-task assignment graph $G = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} consists of worker nodes $\mathcal{N}(u_i)$ and task nodes $\mathcal{C}(v_j)$. \mathcal{E} is the set of edges between worker nodes and task nodes.

Output: the predicted true labels y_j of each task nodes v_j .

- 1: Initialize the features of worker nodes h_i^0 by Equation (1) and the features of task nodes h_j^0 by Equation (2); Initialize the edge features as the one-hot label vector; $t=1$;
 - 2: **while** not converge and $t < t_{max}$ **do**
 - 3: Update h_i^t by Equation (4) and h_j^t by Equation (9);
 - 4: Update worker features h_i^t by Equation (10), h_j^t by Equation (12);
 - 5: Update h_i^t by Equation (4) and h_j^t by Equation (9);
 - 6: Predict the label $\hat{\mathbf{y}}$ for tasks by Equation (14).
 - 7: Obtain the loss by Equation (15) and update model parameters.
 - 8: $t = t+1$;
 - 9: **end while**
-

5 EXPERIMENT

5.1 Datasets

We ran our experiment on 13 widely-used real-world datasets. These datasets are from four crowdsourcing dataset collections. Among them, bird, dog, rte and web are from [57]³, flowers is obtained from [45]⁴, SP, SP*, ZC_{all}, ZC_{in} and ZC_{us} are from [50]⁵, face, product and sentiment are from [59]⁶. Among them, ten datasets are binary tasks including bird to determine whether an image contains any bird [54], flowers to distinguish whether the flower in an image is peach flower [45], rte to recognize textual entailment [41], SP and SP* to perform sentiment analysis for movie reviews [50], ZC_{all}, ZC_{in}, and ZC_{us} to judge whether a URI is relevant to a named entity extracted from news [50], product to tell whether two products are the same given their descriptions [51], sentiment to perform sentiment analysis for companies mentioned in tweets [59]. There are also three multi-class tasks include web judging the relevance of web search results [61], dog determining the breed of a dog from ImageNet [12], and face distinguishing the facial expressions [33].

³<https://github.com/zhangyuc/SpectralMethodsMeetEM>

⁴https://github.com/coverdark/deep_laa

⁵<https://github.com/orchidproject/active-crowd-toolkit>

⁶https://zhydhkcws.github.io/crowd_truth_inference/index.html

Table 5. Datasets statistics

Dataset	#Tasks	#Workers	#Categories	#Labels
bird	108	39	2	4,212
flowers	200	36	2	2,366
web	2,653	177	5	15,539
dog	807	109	4	8,070
rte	800	164	2	8,000
SP	4,999	203	2	27,746
SP*	500	143	2	10,000
ZC _{all}	2,040	78	2	20,125
ZC _{in}	2,040	25	2	10,495
ZC _{us}	2,040	74	2	11,155
face	584	27	4	5,242
product	8,315	176	2	24,945
sentiment	1,000	85	2	20,000

The statistics of datasets are shown in Table 5, these datasets vary considerably in the number of tasks (from 108 to 8,315) and labels (from 2,366 to 27,746). The results of our experiments suggest that our method is adaptive to different scales of datasets. Our model has proven its capability of handling multi-label crowdsourcing problem by superior performance on these datasets. See Table 6.

5.2 Baselines

We use these following methods for comparison:

- **MV**: the MV is an abbreviation of majority voting, it is a basic model, which considers workers equally and selects the label that received most votes from workers as the true label.
- **GLAD**: the GLAD is an abbreviation of Generative model of Labels, Abilities, and Difficulties. This a probabilistic model that jointly infers the true label of each task, the expertise of workers, and the difficulty of tasks [55].
- **MLP**: a three-layer MLP (Multi-Layer Perception) trained in a similar way as our method.
- **iBCC-MF**: Bayesian Classifier Combination (BCC) was proposed by [23] for ensemble learning purpose. BCC has several variants, iBCC-MF is a mean-field variational inference implementation of independent BCC (iBCC) [14, 27, 40] and performs slightly better than iBCC [27]. Hence we include iBCC-MF as a baseline.
- **EBCC**: an enhanced Bayesian classifier combination model proposed by Li et al. [27]. This method models worker reliability at a subtype level, where each class is considered as a mixture of subtypes and worker performance at per subtype induces inter-worker correlations.

5.3 Implementation Details

Our model⁷ is implemented based on Pytorch⁸ and Deep Graph Library (DGL)⁹. We perform cross-validation to evaluate the performance of each model. Each dataset is separated into n splits. We use one split for training and the rest for testing, and obtain the mean accuracy as the evaluation result. n is set to 5, 10, and 20. Note that we randomly split the datasets and fix the splits afterward when evaluating all methods for a fair comparison.

⁷https://github.com/whl97/Crowdsourcing_Label_Inference

⁸<https://pytorch.org>

⁹<http://dgl.ai>

Table 6. Accuracy comparison on 5-fold cross validation.

Dataset	MV	GLAD	MLP	iBCC-MF	EBCC	MP2+COR+MP2
bird	0.7592 ±0.0235	0.7593 ±0.0149	0.9074±0.0218	0.8889 ±0.0177	0.8610±0.0225	0.8610±0.0508
flowers	0.7600±0.0114	0.7950±0.0120	0.8213±0.0264	0.8700±0.0149	0.7200±0.0093	0.8638±0.0133
web	0.7765±0.0030	0.7252±0.0025	0.7982±0.0088	0.7508±0.0033	0.7437± 0.0045	0.9734±0.0215
dog	0.8178± 0.0052	0.8092±0.0054	0.6366±0.0117	0.8389±0.0050	0.8401±0.0057	0.8243±0.0088
rte	0.9188±0.0053	0.9050±0.0060	0.8463±0.0248	0.9275±0.0053	0.9313±0.0048	0.9269±0.0104
SP	0.8896±0.0018	0.8872±0.0013	0.8833±0.0114	0.9150±0.0019	0.9152±0.0017	0.9149±0.0091
SP*	0.9440±0.0034	0.9360±0.0034	0.9300±0.0132	0.9440±0.0034	0.9460±0.0022	0.9445±0.0025
ZC _{all}	0.8348±0.0069	0.8294±0.0042	0.7936±0.0610	0.7951±0.0032	0.8632±0.0039	0.9076±0.0162
ZC _{in}	0.7441±0.0013	0.7304±0.0020	0.7933±0.0154	0.7696±0.0034	0.7784±0.0039	0.7942±0.0071
ZC _{us}	0.8696±0.0038	0.8221±0.0019	0.7830±0.0596	0.8270±0.0005	0.9123±0.0023	0.9130±0.0069
face	0.6301±0.0102	0.6336±0.0086	0.6015±0.0156	0.6404±0.0082	0.6336±0.0062	0.6635±0.0118
product	0.8966±0.0020	0.9040±0.0016	0.8784±0.0017	0.9383±0.0012	0.9349±0.0016	0.9363±0.0019
sentiment	0.9320±0.0038	0.9510±0.0046	0.9517±0.0048	0.9600±0.0055	0.9610±0.0045	0.9608±0.0076

5.4 Results

We compare our method with the aforementioned baselines on different real-world datasets. Table 6 compares the accuracy on different datasets under the 5-fold cross validation settings. The results demonstrate that our method outperforms others in most of the datasets. Due to the various natures of different datasets, it is hard for one crowdsourcing model to beat all others on all datasets (as shown in previous papers [27]). Among all 13 datasets, our method achieves the best accuracy on 5 datasets and is also comparable to the best performance on the other 8 datasets. The result on the dataset web is extremely remarkable, probably due to its good graph structure. When looking into detailed statistics of datasets, we notice that there are 7 datasets that have no less than 1000 tasks while other datasets are relatively small. Among the 7 larger datasets, our method achieves the highest accuracy on 4 of them and is less than 0.2% worse than the best on the other 3 datasets. From another perspective, among 5 datasets on which we obtained the best results, 4 datasets are relatively larger. This suggests that our method is more superior on large datasets.

EBCC, another model with worker correlation in consideration, achieves the best results on 5 datasets (dog, rte, SP, SP*, and sentiment). Compared to EBCC, our method uses a different methodology from deep learning and graph neural networks, and achieves much more stable results across all datasets. Specifically, our model obtains the same accuracy on bird, and is better on 7 datasets (flowers, web, ZC_{all}, ZC_{in}, ZC_{us}, face, product, and only slightly inferior on 5 datasets (dog, rte, SP, SP* and sentiment).

It is worth noting that the MLP method has the same setting as our method, but the results are much worse than ours. That may be explained by the advantage of iterative message passing between workers and tasks in graph neural networks. Another reason may be that MLP can only utilize information from those tasks with ground truth during the training phase. Other tasks without ground-truth labels, however, have a lot of hidden information as well. Our method, as a semi-supervised graph neural network, is trained on the whole worker-task assignment graph, thus we can fully capture the hidden states of all tasks and workers and the structural information among them.

5.5 Ablation Studies

We study the effect of model components by comparing the prediction accuracy of different ablation models. Comparison of MP1, MP2 and their variants are shown in Table 7. MP n ($n = 1, 2$) denotes a single message passing layer, MP n +MP n indicates that we stack two message passing layers, MP n +COR+MP n means that we put a latent correlation layer

between two message passing layers. The results show that on most of the datasets $MPn+COR+MPn$ almost constantly outperforms $MPn+MPn$ as well as the single layer MPn , regardless the selection of message passing method MPn . This demonstrates the effectiveness of capturing inter-worker and inter-task latent correlations. The COR layer brings in possible dependency between distant nodes, which the 2-hop model ($MPn+MPn$) cannot provide.

Table 7. Prediction accuracy of different ablation model on 5-fold cross validation.

Dataset	MP1	MP1+MP1	MP1+COR+MP1	MP2	MP2+MP2	MP2+COR+MP2
bird	0.8472±0.0312	0.8219±0.0563	0.8658±0.0355	0.8841±0.0218	0.8609±0.0174	0.8610±0.0508
flowers	0.8163±0.0140	0.8287±0.0191	0.8438±0.0288	0.8475±0.0230	0.8575±0.0163	0.8638±0.0133
web	0.8585±0.0076	0.8606±0.0062	0.9428±0.0232	0.9509±0.0065	0.9710±0.0056	0.9734±0.0215
dog	0.8324±0.0077	0.8271±0.0069	0.8278±0.0074	0.8206±0.0161	0.8042±0.0146	0.8243±0.0088
rte	0.9256±0.0062	0.9256±0.0085	0.9272±0.0099	0.9284±0.0050	0.9269±0.0072	0.9269±0.0104
SP	0.8971±0.0037	0.9019±0.0059	0.9113±0.0032	0.9130±0.0026	0.9138±0.0032	0.9149±0.0091
SP*	0.9455±0.0052	0.9440±0.0057	0.9435±0.0021	0.9425±0.0029	0.9420±0.0021	0.9445±0.0025
ZC _{all}	0.8456±0.0093	0.8513±0.0078	0.8739±0.0062	0.8989±0.0042	0.9083±0.0034	0.9076±0.0162
ZC _{in}	0.7828±0.0071	0.7828±0.0071	0.7828±0.0071	0.7875±0.0023	0.7904±0.0059	0.7942±0.0071
ZC _{us}	0.8757±0.0085	0.8795±0.0042	0.8819±0.0076	0.8968±0.0047	0.9062±0.0057	0.9130±0.0069
face	0.6678±0.0127	0.6712±0.0151	0.6742±0.0182	0.6675±0.0102	0.6618±0.0205	0.6635±0.0118
product	0.9232±0.0009	0.9295±0.0020	0.9314±0.0015	0.9336±0.0026	0.9338±0.0024	0.9363±0.0019
sentiment	0.9535±0.0067	0.9500±0.0091	0.9530±0.0095	0.9563±0.0059	0.9545±0.0084	0.9608±0.0076

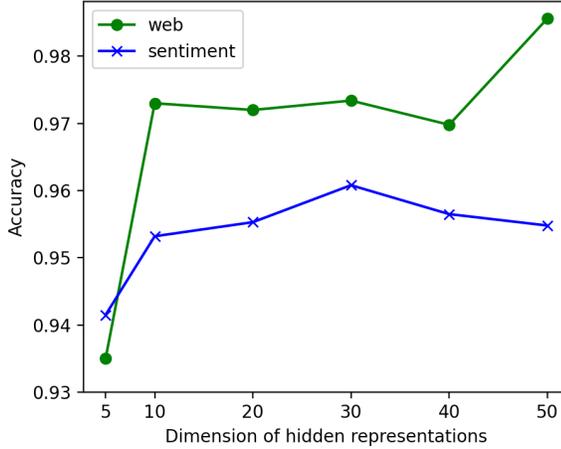


Fig. 2. Effect of different dimensions of hidden representations. We display the accuracies of our final model (MP2-COR-MP2) on two datasets, web and sentiment, along the change of dimensionality.

5.6 Effect of Dimensionality

We also study the impact of the dimensions of hidden representations. We experiment on the proposed $MP2+COR+MP2$ model. As shown in Fig. 2, the best dimension for each dataset to obtain the highest accuracy are not always the same. When faced with a new dataset, it is difficult for us to know the best dimension. Thus we fix this hyperparameter to 30 for all datasets to present the final results.

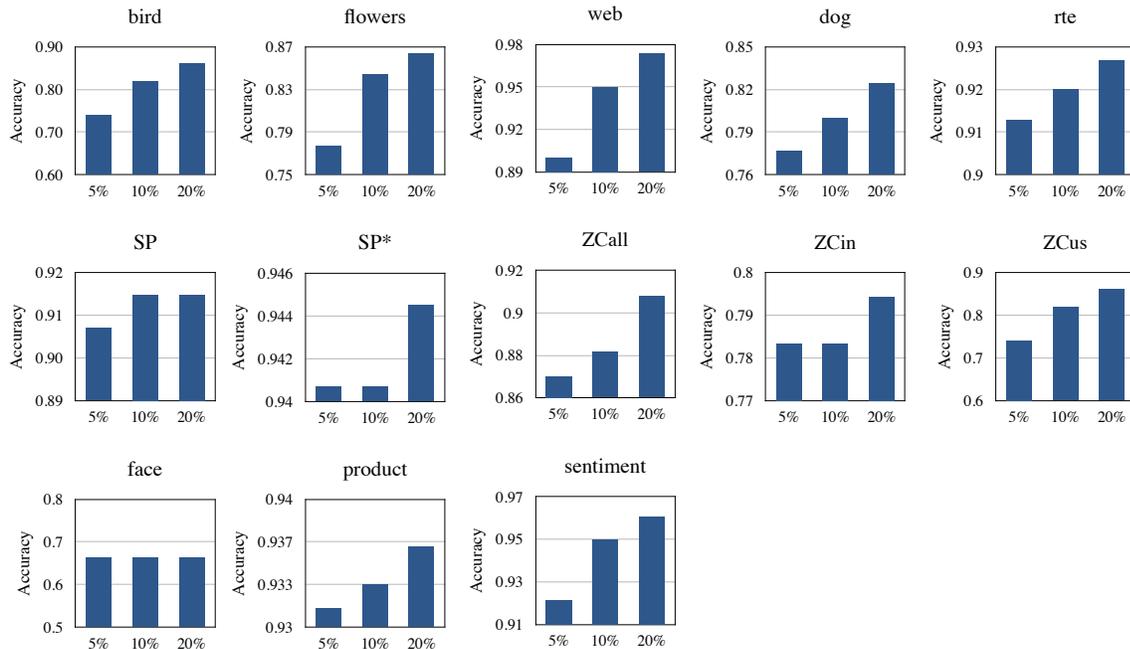


Fig. 3. Effect of training proportion on different datasets. Each subgraph shows the accuracies of our final model (MP2-COR-MP2) on different datasets along the change of training proportion (5%, 10%, and 20%).

5.7 Effect of Training Proportion

To study the effect of different training proportions, Fig. 3 demonstrates how the performance of our model varies with the training proportion on each dataset. On all datasets, the accuracy increases as the training proportion becomes larger. But the trends of some datasets are barely noticeable, which indicates that on these datasets our method can achieve quite good performance with very little training data (e.g. 5%). Some other datasets increase obviously with the proportion of training data, we find that our model can fully utilize the training data and achieve quite remarkable performance compared to other methods (e.g. on web and ZC_{a11}).

6 CONCLUSION AND FUTURE WORK

We present a novel Heterogeneous Graph Neural Network for label aggregation in crowdsourcing. Constructing a graph to represent the worker-task interactions, we utilize the power of graph neural networks to learn a better representation for workers and tasks. Moreover, our heterogeneous graph neural network differs from previous works by adding new latent correlations among the same type of nodes (i.e. worker nodes and task nodes), which captures the worker-worker and task-task correlation in the crowdsourcing problem. Comparing with state-of-the-art label aggregation models and our own ablation models, we demonstrated the effectiveness of heterogeneous graph neural networks on real-world crowdsourcing datasets, as well as the usefulness of modeling the latent correlation of workers/tasks. Future work includes exploring the generative models for crowdsourcing graphs and extends our model to the unsupervised setting (without the requirement of ground-truth labels).

REFERENCES

- [1] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. 2016. AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Transactions on Medical Imaging* 35, 5 (2016), 1313–1321. <https://doi.org/10.1109/TMI.2016.2528120>
- [2] Bahadir Ismail Aydin, Yavuz Selim Yilmaz, Yaliang Li, Qi Li, Jing Gao, and Murat Demirbas. 2014. Crowdsourcing for multiple-choice question answering. In *Twenty-Sixth IAAI Conference*.
- [3] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).
- [4] Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. 2019. Max-MIG: an Information Theoretic Approach for Joint Learning from Crowds. *International Conference on Learning Representations (ICLR)* (2019).
- [5] Liang Chen, Yang Liu, Zibin Zheng, and Philip Yu. 2018. Heterogeneous Neural Attentive Factorization Machine for Rating Prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 833–842.
- [6] Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2019. Reinforcement learning based graph-to-sequence model for natural question generation. In *The Eighth International Conference on Learning Representations (ICLR 2020)*.
- [7] Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2020. Iterative Deep Graph Learning for Graph Neural Networks: Better and Robust Node Embeddings. In *Thirty-Fourth annual conference on Neural Information Processing Systems (NeurIPS 2020)*.
- [8] Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2020. Toward subgraph guided knowledge graph question generation with graph neural networks. *arXiv preprint arXiv:2004.06015* (2020).
- [9] Zhendong Chu, Jing Ma, and Hongning Wang. 2020. Learning from Crowds by Modeling Common Confusions. *arXiv preprint arXiv:2012.13052* (2020).
- [10] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.
- [11] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*. 469–478.
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [13] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph Neural Networks for Social Recommendation. In *The World Wide Web Conference*. ACM, 417–426.
- [14] Paul Felt, Kevin Black, Eric Ringger, Kevin Seppi, and Robbie Haertel. 2015. Early gains matter: A case for preferring generative over discriminative crowdsourcing models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 882–891.
- [15] Hanning Gao, Lingfei Wu, Po Hu, and Fangli Xu. 2020. RDF-to-Text Generation with Graph-augmented Structural Neural Encoders. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*. 3030–3036.
- [16] Alex Gaunt, Diana Borsa, and Yoram Bachrach. 2016. Training deep neural nets to aggregate crowdsourced responses. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 242–251.
- [17] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1263–1272.
- [18] Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto. 2017. Knowledge transfer for out-of-knowledge-base entities: a graph neural network approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 1802–1808.
- [19] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*. 1024–1034.
- [20] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. 64–67.
- [21] Asif R Khan and Hector Garcia-Molina. 2017. Crowddqs: Dynamic question selection in crowdsourcing systems. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 1447–1462.
- [22] Ashish Khetan and Sewoong Oh. 2016. Achieving budget-optimality with adaptive schemes in crowdsourcing. In *Advances in Neural Information Processing Systems*. 4844–4852.
- [23] Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian classifier combination. In *Artificial Intelligence and Statistics*. 619–627.
- [24] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [25] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. 2014. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment* 8, 4 (2014), 425–436.
- [26] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 1187–1198.
- [27] Yuan Li, Benjamin Rubinstein, and Trevor Cohn. 2019. Exploiting Worker Correlation for Label Aggregation in Crowdsourcing. In *International Conference on Machine Learning*. 3886–3895.

- [28] Chen Liang, Ziqi Liu, Bin Liu, Jun Zhou, and Xiaolong Li. [n.d.]. Who Stole the Postage? Fraud Detection in Return-Freight Insurance Claims. ([n. d.]).
- [29] Nick Littlestone, Manfred K Warmuth, et al. 1989. *The weighted majority algorithm*. University of California, Santa Cruz, Computer Research Laboratory.
- [30] Qiang Liu, Jian Peng, and Alexander T Ihler. 2012. Variational inference for crowdsourcing. In *Advances in neural information processing systems*. 692–700.
- [31] Ziqi Liu, Chaochao Chen, Longfei Li, Jun Zhou, Xiaolong Li, Le Song, and Yuan Qi. 2019. Geniepath: Graph neural networks with adaptive receptive paths. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4424–4431.
- [32] Mingqi Lv, Zhaoxiong Hong, Ling Chen, Tieming Chen, Tiantian Zhu, and Shouling Ji. 2020. Temporal multi-graph convolutional network for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems* (2020).
- [33] Barzan Mozafari, Purna Sarkar, Michael Franklin, Michael Jordan, and Samuel Madden. 2014. Scaling up crowd-sourcing to very large datasets: a case for active learning. *Proceedings of the VLDB Endowment* 8, 2 (2014), 125–136.
- [34] Shmuel Nitzan and Jacob Paroush. 1982. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review* (1982), 289–297.
- [35] Vikas C. Raykar and Shipeng Yu. 2012. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *J. Mach. Learn. Res.* 13, null (Feb. 2012), 491–518.
- [36] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research* 11, 4 (2010).
- [37] Filipe Rodrigues and Francisco C Pereira. 2018. Deep learning from crowds. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [38] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2008), 61–80.
- [39] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*. Springer, 593–607.
- [40] Edwin Simpson, Stephen Roberts, Ioannis Psorakis, and Arfon Smith. 2013. Dynamic bayesian combination of multiple imperfect classifiers. In *Decision making and imperfection*. Springer, 1–35.
- [41] Rion Snow, Brendan O’connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.
- [42] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.
- [43] Dapeng Tao, Jun Cheng, Zhengtao Yu, Kun Yue, and Lizhen Wang. 2018. Domain-weighted majority voting for crowdsourcing. *IEEE transactions on neural networks and learning systems* 30, 1 (2018), 163–174.
- [44] Tian Tian and Jun Zhu. 2015. Max-margin majority voting for learning from crowds. In *Advances in neural information processing systems*. 1621–1629.
- [45] Tian Tian and Jun Zhu. 2015. Uncovering the latent structures of crowd labeling. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 392–404.
- [46] Jingzheng Tu, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. 2020. Attention-aware answers of the crowd. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 451–459.
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [48] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [49] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*. ACM, 155–164.
- [50] Matteo Venanzi, Oliver Parson, Alex Rogers, and Nick Jennings. 2015. The ActiveCrowdToolkit: An open-source tool for benchmarking active learning algorithms for crowdsourcing research. In *Third AAAI Conference on Human Computation and Crowdsourcing*.
- [51] Jiannan Wang, Tim Kraska, Michael J Franklin, and Jianhua Feng. 2012. Crowder: Crowdsourcing entity resolution. *arXiv preprint arXiv:1208.1927* (2012).
- [52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.
- [53] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous Graph Attention Network. In *The World Wide Web Conference*. ACM, 2022–2032.
- [54] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. 2010. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*. 2424–2432.
- [55] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*. 2035–2043.
- [56] Li’ang Yin, Jianhua Han, Weinan Zhang, and Yong Yu. 2017. Aggregating crowd wisdoms with label-aware autoencoders. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 1325–1331.

- [57] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. 2014. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems*. 1260–1268.
- [58] Yizhou Zhang, Yun Xiong, Xiangnan Kong, Shanshan Li, Jinhong Mi, and Yangyong Zhu. 2018. Deep collective classification in heterogeneous information networks. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 399–408.
- [59] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552.
- [60] Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, and Jianhua Feng. 2015. QASCA: A Quality-Aware Task Assignment System for Crowdsourcing Applications. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (Melbourne, Victoria, Australia) (SIGMOD '15)*. Association for Computing Machinery, New York, NY, USA, 1031–1046. <https://doi.org/10.1145/2723372.2749430>
- [61] Dengyong Zhou, Sumit Basu, Yi Mao, and John C Platt. 2012. Learning from the wisdom of crowds by minimax entropy. In *Advances in neural information processing systems*. 2195–2203.
- [62] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434* (2018).
- [63] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34, 13 (2018), i457–i466.