Learning with Group Noise

Qizhou Wang[1,2,*], **Jiangchao Yao**[3,*], **Chen Gong**[2,4,†], **Tongliang Liu**[5], **Mingming Gong**[6], **Hongxia Yang**[3], **Bo Han** [1,†]

[1] Department of Computer Science, Hong Kong Baptist University
[2] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of MoE, School of Computer Science and Engineering, Nanjing University of Science and Technology
[3] Data Analytics and Intelligence Lab, Alibaba Group
[4] Department of Computing, Hong Kong Polytechnic University
[5] Trustworthy Machine Learning Lab, School of Computer Science, Faculty of Engineering, The University of Sydney
[6] School of Mathematics and Statistics, The University of Melbourne

Abstract

Machine learning in the context of noise is a challenging but practical setting to plenty of real-world applications. Most of the previous approaches in this area focus on the pairwise relation (casual or correlational relationship) with noise, such as learning with noisy labels. However, the group noise, which is parasitic on the coarse-grained accurate relation with the fine-grained uncertainty, is also universal and has not been well investigated. The challenge under this setting is how to discover true pairwise connections concealed by the group relation with its fine-grained noise. To overcome this issue, we propose a novel Max-Matching method for learning with group noise. Specifically, it utilizes a matching mechanism to evaluate the relation confidence of each object (*cf.* Figure 1) w.r.t. the target, meanwhile considering the Non-IID characteristics among objects in the group. Only the most confident object is considered to learn the model, so that the fine-grained noise is mostly dropped. The performance on a range of real-world datasets in the area of several learning paradigms demonstrates the effectiveness of Max-Matching.

Introduction

The success of machine learning is closely related to the availability of data with accurate relation descriptions. However, the data quality usually cannot be guaranteed in many real-world applications, *e.g.*, image classification (Li et al. 2017), machine translation (Belinkov and Bisk 2017), and object recognition (Yang et al. 2020). To overcome this issue, learning from cheap but noisy assignments has attracted intensive attention. Especially, in the recent years, lots of works have contributed to learning with label noise (Xia et al. 2020; Han et al. 2020; Chen et al. 2020).

Nevertheless, most of the previous works focus on the pairwise relation with noise as characterized in Figure 1(a). For notion simplicity, we call it *pairwise noise*. Another type of noise, which is implicitly parasitic on the weak relations as illustrated in Figure 1(b)-(d), is also general but has not been well investigated. We specially term it *group noise* based on the two following characteristics: 1) it occurs in the group whose coarse-grained relation to the target is correct, while the fine-grained relation of each object in the group to the target might be inaccurate; 2) it is not proper to independently consider fine-grained noisy relations like Figure 1(a), since objects in one group exhibit strong Non-IID characteristics. Correlation analysis for each group of objects can help us discover better evidences to this type of noise. In the following, we enumerate some examples about group noise.

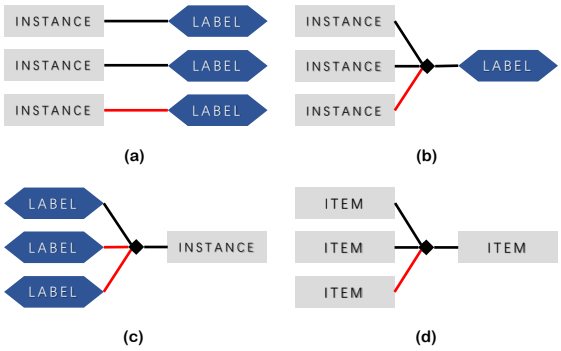

Figure 1: Illustration of the supervised learning with pairwise noise (a) and three settings with group noise (b)-(d), in which the objects are realized by instances, labels, and items respectively. In the figure, black lines represent the correct relations, while red lines mean the incorrect relations.

- Figure 1(b): In region-proposal-based object localization, each group is a set of regions of one image. Given the image-level label, we aim to find its corresponding regions and remove the irrelevant and background parts. Here, a group of instances (or regions) are weakly supervised by a category label, while some instances are mismatched to this category. This is like the multiple-instance learning problem in the case of the instance-level classi-

[*] Equal contribution.
[†] Corresponding authors. Emails: chen.gong@njust.edu.cn, bhanml@comp.hkbu.edu.hk.

fication (Liu, Wu, and Zhou 2012), not the more popular bag-level classification (Maron and Lozano-Pérez 1998).

- Figure 1(c): In face naming, characters' faces may appear simultaneously in a screenshot from the TV serials, and each face is assigned with a set of candidate names in the script or in dialogue. Under this setting, only one name (or label) in the group is correct to the face and all the candidates are correlated due to relationship among characters. From these data, we are to determine the true name of each face, which has also been viewed as a partial-label learning problem (Gong et al. 2017).
- Figure 1(d): In recommender system, item-based collaborative filtering (Sarwar et al. 2001) is a classical method. It builds upon that the co-occurrence information of item pairs is relatively reliable. However, due to the uncertainty of user behavior in e-commerce, it exists that the historical items are irrelevant to the subsequent items. This introduces the group noise when we consider the sequence of each user as a group for the next-item, leading to the deterioration of applying the NeuralCF model with the fine-grained pairwise relations (He et al. 2017).

Although several works more or less explore this type of scenarios, they are usually tailored to their ultimate goals and may distort the characteristics of group noise. For example, previous multiple-instance learning, which considers the instance-level modeling (Settles, Craven, and Ray 2008; Pao et al. 2008), may make the strong IID assumption about the instances in the group. Partial-label learning methods (Zhang, Zhou, and Liu 2016) suppose the equal confidence of the candidate labels or model the ground-truth as a latent variable, which might not be very effective. Besides, all these works do not explicitly construct the denoising mechanism to avoid the influence of group noise.

In this paper, we investigate the problem of learning with group noise, and introduce a novel Max-Matching approach. Specifically, it consists of two parts, a matching mechanism and a selection procedure. The matching mechanism leverages the *pair matching* to evaluate the confidence of relation between the object and the target, meanwhile adopts the *group weighting* to further consider the Non-IID property of objects in the group. The final matching scores are achieved by combining the pair matching and the group weighting, of which the results evaluate both each fine-grained relation pair and the object importance in the group. Then, the selection procedure chooses the most confident relation pair to train the model, which at utmost avoids the influence of the irrelevant and mismatched relations. The whole model is end-to-end trained and Figure 2 illustrates the structure of Max-Matching. We conduct a range of experiments, and the results indicate that the proposed method can achieve superior performance over baselines from three different learning paradigms with group noise in Figure 1.

Related Works

Learning with Pairwise Noise

For learning with pairwise noise, researchers mainly focus on instances with error-prone labels (Frénay and Verleysen 2014; Algan and Ulusoy 2019), where the noise occurs in pairwise relations between individual instances to their assigned labels. By making assumptions on label assignment, robust loss functions (Manwani and Sastry 2013; Ghosh, Kumar, and Sastry 2017; Han et al. 2018b; Yao et al. 2019) and various consistent learning algorithms are proposed (Liu and Tao 2016; Han et al. 2018a; Xia et al. 2019; Yao et al. 2019; Yao et al. 2020b).

Learning with Group Noise

For learning with group noise, we have a group of objects collectively connected to the target with the coarse-grained guarantees but the fine-grained uncertainty. Several previous methods, in Multiple-Instance Learning (MIL), Partial-Label Learning (PLL), and Recommender System (RS), have mediately investigated this problem.

MIL probably is one of the most illustrative paradigms about group noise, of which the supervision is provided for a bag of instances. In MIL, prediction can either be made for bags or individuals, respectively termed as the bag-level prediction and instance-level prediction. For bag-level prediction, many works estimate instance labels as an intermediate step (Ray and Craven 2005; Settles, Craven, and Ray 2008; Wang et al. 2018).

However, as suggested by (Vanwinckelen et al. 2016), the MIL methods designed for bag classification are not optimal for the instance-level tasks. The methods for instance-level prediction are only studied in the minority but close to the problem of our paper. Existing methods are devised based on key instance detection (Liu, Wu, and Zhou 2012), label propagation (Kotzias et al. 2015), or unbiased estimation (Peng and Zhang 2019) with IID assumptions.

PLL also relates to the problem of learning with group noise, where each instance is assigned with a group of noisy labels, and only one of them is correct. To avoid the influence of the group noise, two general methodologies, namely, the average-based strategy and the detection-based approach, are proposed. The average-based strategy usually treats candidate labels equally, and then adapts PLL to the general supervision techniques (Hüllermeier and Beringer 2006; Cour, Sapp, and Taskar 2011; Wu and Zhang 2018). The detection-based methods aim at revealing the true label among the candidates, mainly through label confidence learning (Zhang, Zhou, and Liu 2016), maximum margin (Yu and Zhang 2016), or alternating optimization (Zhang and Yu 2015; Feng and An 2019; Yao et al. 2020a). Above methods do not explicitly build the denoising mechanism, which might not be effective in learning with group noise.

RS targets to recommend the points of interest for users given their historical behaviors. In e-commerce, item-based collaborative filtering (Sarwar et al. 2001; Linden, Smith, and York 2003) has been used as a popular technique, which discovers new items based on the similar ones. It builds upon that the item relation is relatively reliable, so that the unseen true correlations between items can be learned via matrix factorization (Mnih and Salakhutdinov 2008), autodecoders (Sedhain et al. 2015), or deep models (Huang et al. 2013; Xue et al. 2017; He et al. 2017; Cui et al. 2018). Unfortunately, in practice, it is not very easy to accurately construct the such pairwise relation for training, especially

in the interest-varying user click sequences. Although more advanced studies mine the multiple interests of users and sequential behavior analysis (Hidasi and Karatzoglou 2018; Wu et al. 2019) to acquire benefits, the effect of group noise has not been well studied yet. Our experiments reveal that eliminating the group noise from the user click sequences for the next-item can effectively improve the performance.

Preliminary

Assume that we have a source set $\mathcal{X}$ and a target set $\mathcal{Y}$. For example, in classification tasks, $\mathcal{X}$ and $\mathcal{Y}$ can be considered as the sample set and the label set respectively. Ideally, we have the collection $S = \{(x_i, y_i)\}_{i=1}^n$ (n is the sample size) for training, where the source object $x_i \in \mathcal{X}$ connects to the target $y_i \in \mathcal{Y}$ via the true pairwise relation. For generality, we use $f : \mathcal{X} \to \mathbb{R}^d$ and $g : \mathcal{Y} \to \mathbb{R}^d$ to map both the objects in $\mathcal{X}$ and $\mathcal{Y}$ into the embedding space. Then, the solution is formulated as the following problem:

$$f^*, g^* \leftarrow \arg\min_{f,g} \sum_{i=1}^{n} \ell(f(x_i), g(y_i)), \tag{1}$$

where $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is a proper loss function. After training, the optimal mapping functions are used to make various prediction tasks, such as classification or retrieval.

However, in many real-world situations, group-level data acquisition is cheaper, in which a group of source objects are collectively connected to a target. Unfortunately, as shown in Figure 1, some objects in the group can be irrelevant to the target regarding the pairwise relations. This forms the problem of learning with group noise, where we have to handle the noise that is parasitic on $S_{\text{group}} = \left\{(\bar{X}_i, y_i)\right\}_{i=1}^n$. Here, $\bar{X}_i = \{\bar{x}_{i1}, \ldots, \bar{x}_{iK}\} \in \mathcal{X}^K$ contains a set of source objects collectively related to a target object $y_i \in \mathcal{Y}$. Note that $\bar{x}_{ik}$ is different from x_{ik} regarding the notation, indicating there may exist $\bar{x}_{ik} \in \bar{X}_i$, such that $(\bar{x}_{ik}, y_i) \notin S$, *i.e.*, $\bar{x}_{ik}$ is mismatched to the target y_i in terms of the pairwise relation. In this setting, we aim at devising a novel objective function $\ell_{\text{group}} : \mathbb{R}^{d\times K} \times \mathbb{R}^d \to \mathbb{R}^+$ such that

$$f^*, g^* \leftarrow \arg\min_{f,g} \sum_{i=1}^{n} \ell_{\text{group}}(F(\bar{X}_i), g(y_i)) \tag{2}$$

can find the same optimal mapping functions f^*, g^* as in Eq. (1), where $F(\bar{X}_i) = \{f(\bar{x}_{i1}), \ldots, f(\bar{x}_{iK})\}$ denotes the set of embedding features. After training, the evaluation is still implemented on the individual pairwise connections between the source object and the target object.

Max-Matching

We need to modify the original loss functions such that the classifier learned with group noise can converge to the optimal one learned without any noise.

In this section, we introduce a novel method, namely, Max-Matching, for learning with group noise. It consists of two parts, a matching mechanism and a selection procedure. The matching mechanism jointly considers the following two aspects of relation: 1) the pairwise relation of the source objects to the target; 2) the relation among the source objects in the group. Accordingly, the correctness of the pairwise relations as well as object correlations in the group are revealed by each matching score. Subsequently, based on the results given by the matching mechanism, the selection procedure chooses the best matched object to optimize the model. The group noise can mostly be removed, since the selected object is at utmost guaranteed to be correct regarding its pairwise relation to the target object, and other less confident objects in the group are not considered. Formally, the objective function ℓ_{group} of Max-Matching is,

$$-\max_{\bar{x}_{ik}\in\bar{X}_i} \{\log \underbrace{\hat{P}\left(y_i|\bar{x}_{ik}; f, g\right)}_{\text{Pair Matching}} + \log \underbrace{\hat{P}\left(\bar{x}_{ik}|\bar{X}_i; f, g\right)}_{\text{Group Weighting}}\}, \tag{3}$$

where $\hat{P}(\cdot)$ denotes the estimated probability. Note that, the two terms in Eq. (3) are equally combined[1] and they are interdependent in the training phase. The second term helps the reliable pairwise relation to be identified, and the first term also boosts the weighting measure to be learned.

In the following, we explain the intuition behind Eq. (3). In learning with group noise, we have no explicitly clean pairwise relations that can be directly used for training. Therefore, we inevitably build a weighting schema to measure the importance of the data in the group, which we assume is $\hat{P}(\bar{x}_{ik}|\bar{X}_i; f, g)$. Then, following the law of the total probability, it might be possible to decompose $\hat{P}(y_i|\bar{X}_i; f, g)$ into a probabilistic term *w.r.t.* the target for each object $\bar{x}_{ik} \in \bar{X}_i$ combined with $\hat{P}(\bar{x}_{ik}|\bar{X}_i; f, g)$. However, the optimization obstacle caused by the integral will prohibit this choice. In this case, Eq. (3) is an alternative approximation to this goal, which we use the following theorem to formulate.

Theorem 1. *Assume $\bar{X}_i = \{\bar{x}_{i1}, \ldots, \bar{x}_{iK}\} \in \mathcal{X}^K$ collectively connects to the target y_i, where there is at least one true pairwise relation $(\bar{x}_{ik}, y_i)$ and some possible pairwise relation noise. Then, optimizing Eq.* (3) *is approximately optimizing all pairwise relations with weights to learn the optimal mapping functions f^* and g^*.*

Proof. According to the law of total probability, the log-likelihood on the coarse-grained relation $(\bar{X}_i, y_i)$ has a following decomposition and the lower-bound approximation,

$$\begin{aligned}
&\log \sum_{\bar{x}_{ik}\in\bar{X}_i} \hat{P}(y_i|\bar{x}_{ik}; f, g)\hat{P}(\bar{x}_{ik}|\bar{X}_i; f, g) \\
&\geq \log \max_{\bar{x}_{ik}\in\bar{X}_i} \left\{\hat{P}(y_i|\bar{x}_{ik}; f, g)\hat{P}(\bar{x}_{ik}|\bar{X}_i; f, g)\right\} \\
&= \max_{\bar{x}_{ik}\in\bar{X}_i} \left\{\log \hat{P}(y_i|\bar{x}_{ik}; f, g) + \log \hat{P}(\bar{x}_{ik}|\bar{X}_i; f, g)\right\}
\end{aligned} \tag{4}$$

The first line in the above deduction can be considered as a weighted counterpart of Eq (1) in the setting of group noise. The last line, *i.e.*, Eq. (3), is its lower bound, which alleviates the optimization obstacle caused by integral. Optimizing such a lower bound yields the optimization of the

[1]Non-equal combination with proper tradeoff may lead to better performance, which is left for future exploration in our work.

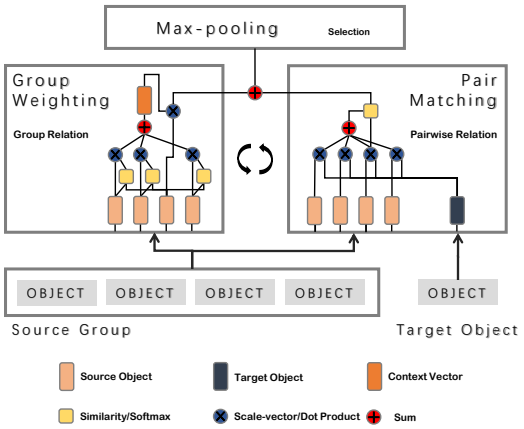

Figure 2: Max-Matching. The *pair matching* evaluates the confidence of individual connections between source objects and the target. The *group weighting* captures the object correlations in the group by measuring their importance. They are combined to form the final matching scores, followed by a max-pooling selection to choose the trustworthy object regarding the target. Group weighting and pair matching are interdependent and benefit from each other during training.

first line, and progressively makes the learning procedure approach the optimal mapping functions f^* and g^*. □

Due to the adverse impact of group noise, $\hat{P}(y_i|\bar{x}_{ik}; f, g)$ may still memorize some pairwise relation noise. In this case, the second term $\hat{P}(\bar{x}_{ik}|\bar{X}_i; f, g)$ can leverage the non-IID characteristics of the objects in the group to sufficiently capture their correlation, and distinguish the irregular noise by measuring their importance regarding the group. Besides, the max-pooling operation in Eq. (3) guarantees that only the most confident object is used, reducing the risk of group noise as much as possible.

Implementation

In this section, we give our implementation of Eq. (3) in detail. First, the term $\hat{P}(y_j|\bar{x}_{ik}; f, g)$ is named as the *pair matching*, as it is the probability of matching between the source object $\bar{x}_{ik}$ and the target object y_i. It is estimated by the Softmax on the inner product of their embedding vectors, constructed as follows:

$$\hat{P}(y_i|\bar{x}_{ik}; f, g) = \frac{\exp\{f(\bar{x}_{ik})^\top g(y_i)\}}{\sum_{y\in\mathcal{Y}} \exp\{f(\bar{x}_{ik})^\top g(y)\}}. \quad (5)$$

The second term $\hat{P}(\bar{x}_{ik}|\bar{X}_i; f, g)$ aims to capture the object correlation by measuring the importance of $\bar{x}_{ik}$ regarding the group. It is termed as a *group weighting* mechanism, as it assigns different weights for pair matching regarding the group. Accordingly, the Non-IID property in the group is considered, since the group weighting is essentially designed as a cluster-aware weighting method. Note that, the weights can either be calculated based on the embedding features $f(\bar{x})$ or the probabilistic features $\hat{P}(y|\bar{x}; f, g)$. To unify these two operations together, we denote the mapping function $h(\cdot)$ for the input features of the group weighting with the similarity measurement $\mathcal{S}(\cdot,\cdot)$. Then, the group weighting $\hat{P}(\bar{x}_{ik}|\bar{X}_i; f, g)$ is calculated by following steps:

- a) Measuring the similarity of the object $\bar{x}_{ik}$ with all other objects in the group (denoted by $\bar{x}'_{i1}, \ldots, \bar{x}'_{i,K-1}$):

$$s_{\bar{x}_{ik}} = \left[\mathcal{S}(h(\bar{x}_{ik}), h(\bar{x}'_{i1})), \ldots, \mathcal{S}(h(\bar{x}_{ik}), h(\bar{x}'_{i,K-1}))\right]^\top,$$

  and normalizing by Softmax $\tilde{s}_{\bar{x}_{ik}} = \mathrm{Softmax}(s_{\bar{x}_{ik}})$;
- b) Calculating the final weight of the object $\bar{x}_{ik}$ in the group with Sigmoid:

$$\hat{P}(\bar{x}_{ik}|\bar{X}_i; f) = \mathrm{Sigmoid}\left(\mathcal{S}(c_{\bar{x}_{ik}}, h(\bar{x}_{ik}))\right), \quad (6)$$

  where $c_{\bar{x}_{ik}}$ is the context vector w.r.t. the object $\bar{x}_{ik}$, calculated by $c_{\bar{x}_{ik}} = \sum_{l=1}^{K-1} \tilde{s}_{\bar{x}_{ik},l} h(\bar{x}'_{il})$.

The context vector $c_{\bar{x}_{ik}}$ is constructed by the weighted sum of all the other objects in the group, in which the weights $\tilde{s}_{\bar{x}_{ik}}$ assign the higher values for those objects similar to $\bar{x}_{ik}$. Intuitively, the context vector resembles the original $\bar{x}_{ik}$ if there exists plenty of objects in the group that are similar to the object $\bar{x}_{ik}$. A large value of group weighting (or a large $\mathcal{S}(c_{\bar{x}_{ik}}, h(\bar{x}_{ik}))$) indicates that the object $\bar{x}_{ik}$ deserves more attention regarding its owning group $\bar{X}_i$.

By mixing the pair matching and the group weighting, we have the final matching score that evaluates the object confidence regarding the target as well as the group. A large value of the matching score generally indicates the corresponding object is trustworthy in its fine-grained relation to the target. The selection procedure is then deployed upon the matching mechanism via a simple max-pooling operation. It selects the object that is the most confident in terms of the pairwise relationship, and the irrelevant objects can be dropped. The model structure is summarized in Figure 2.

	Object		Function		Weighting	
	$\mathcal{X}$	$\mathcal{Y}$	$f(\cdot)$	$g(\cdot)$	$h(\cdot)$	$\mathcal{S}(\cdot,\cdot)$
MIL	ins	lab	ide	emb	Eq. (5)	neg-KL
PLL	lab	ins	emb	lin	$f(\cdot)$	dot
RS	item	item	emb	emb	$f(\cdot)$	dot

Table 1: The Specification of Max-Matching on three types of learning settings with group noise.

Experiments

Experimental Settings

To demonstrate the effectiveness of Max-Matching, we conduct extensive experiments in three representative learning settings with group noise, including MIL, PLL, and RS. Table 1 summarizes their specifications regarding sample sets (*i.e.,* $\mathcal{X}, \mathcal{Y}$), mapping functions (*i.e.,* f, g), and group weighting (*i.e.,* $\mathcal{S}, h$). Therein, "ins", "lab", and "item" respectively denotes the instance with features, the label, and the item ID. Moreover, "emb" represents the embedding function that maps discrete category labels or item IDs to the embedding space; "lin" is a linear function for the instances with normalized features; and "ide" is the identity function.

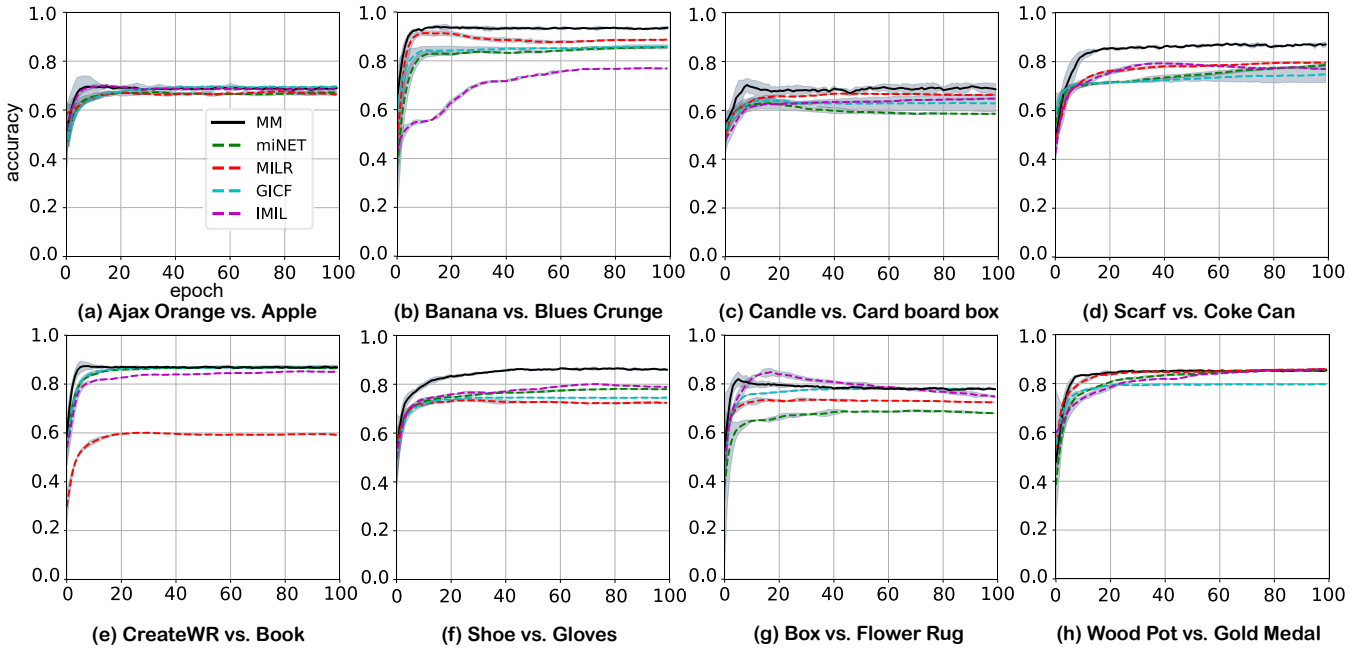

Figure 3: The test accuracy curves on *SIVAL* for learning with group noise. Colored curves show the mean accuracy of 5 trials and shaded bars denote standard deviation.

Our specification in MIL degenerates Eq. (5) into a linear function with Softmax, and its outputs are the inputs of group weighting with negative KL-divergence (neg-KL) as the similarity metric. By contrast, in PLL, instances and labels are both mapped, where Max-Matching can explore the non-IID characteristics of labels in the embedding space, and dot product (dot) is adopted as a proper metric. Similar deliberation holds for RS to measure the confidence of matching in the embedding space. Moreover, we implement Max-Matching using PyTorch, the Adam (Kingma and Ba 2015) is adopted with the learning rate selected from $\{10^{-1}, \cdots, 10^{-4}\}$, and the methods are run for 50 epochs.

Application to Multiple-Instance Learning

In this section, we focus on the MIL setting, where we aim to learn an instance classifier given instances with only bag labels. Here, instances in the bag that may deviate from their bag labels introduce group noise.

The experiments are conducted on an object localization dataset *SIVAL* (Rahmani et al. 2005) in the literature of MIL, as it provides instance-level annotations for evaluation. We compare Max-Matching with two state-of-the-arts that focus on the instance classification, IMIL (Peng and Zhang 2019) and GICF (Kotzias et al. 2015); two strong baselines that estimate instance labels in an intermediate step for bag classification, MILR (Ray and Craven 2005) and miNET (Wang et al. 2018). Since the baselines only focus on binary classification, we use the data of each adjacent classes to construct the binary classification datasets. Each dataset is then partitioned into 8:1:1 for training, validation, and test.

The experimental curves in terms of the test accuracy are illustrated in Figure 3 with 5 individual trials. From them, we find Max-Matching achieves superior performance over the baselines in most cases. For two bag-level prediction methods, the test accuracy is not very competitive since they implicitly consider the instance labels in the bag. As suggested by (Carbonneau et al. 2018), the instance-level performance cannot be guaranteed for MIL methods that only focus on the coarse-grained bag labels. For two instance-level methods, although they generally show better performance than MILR and miNET, they are still inferior to Max-Matching, since they fail to sufficiently leverage the correlation among objects in the group. The results demonstrate the effectiveness of our method in learning with group noise.

	Accuracy	Selection	non-IID
Pairwise	0.302±0.005	×	×
Matching	0.324±0.002	×	✓
Maximizing	0.315±0.006	✓	×
Max-Matching	**0.368±0.005**	✓	✓

Table 2: Average test accuracy and standard deviation in learning with group noise on *SIVAL*.

Furthermore, we conduct multi-class classification experiments on *SIVAL*, since our method is not restricted to the binary classification. To show the advantages of the matching mechanism and the selection procedure in Max-Matching, we leverage following three baselines for the ablation study:

- **Pairwise**: Taking the group label y_i as the label for each instance $\bar{x}_{ik}$ in the group $\bar{X}_i$, the objective can be written as $\sum_{i=1}^{n}\sum_{k=1}^{K} -\log \hat{P}(y_i|\bar{x}_{ik}; f, g)$.
- **Matching**: Taking the matching scores of individuals

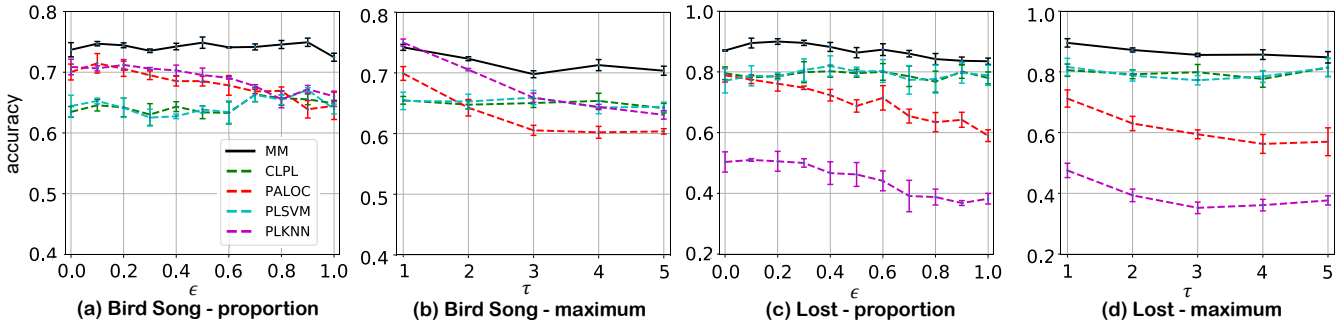

Figure 4: Test accuracy curves on PLL datasets for learning with group noise. Colored curves show the mean accuracy of 5 trials and error bars denote standard deviation. Therein, ϵ denotes the proportion of instances that are partially labeled, and τ is the maximum numbers of partial labels.

	FG-NET	*MSRC v2*	*Bird Song*	*Yahoo! News*	*Lost*
PLKNN	0.059± 0.005	0.446± 0.007	0.613± 0.004	0.426± 0.004	0.428± 0.003
PLSVM	0.064± 0.013	0.475± 0.008	0.625± 0.019	0.629± 0.012	0.801± 0.025
CLPL	0.065± 0.029	0.480± 0.015	0.628± 0.012	0.537± 0.017	0.793± 0.022
PALOC	0.054± 0.005	0.463± 0.011	0.598± 0.020	0.434± 0.0013	0.546± 0.007
Max-Matching	**0.110± 0.021**	**0.517± 0.007**	**0.642± 0.010**	**0.647± 0.005**	**0.823± 0.025**

Table 3: The average test accuracy and its standard deviation on the PLL datasets in learning with group noise.

in the group as the objective directly, which means $\sum_{i=1}^{n}\sum_{k=1}^{K} -\log \hat{P}(y_i|\bar{x}_{ik}; f, g) - \log \hat{P}(\bar{x}_{ik}|\bar{X}_i; f)$.

- **Maximizing**: Selecting the most confident instance $\bar{x}_{ik}$ only in terms of the pairwise relation to the target, namely, $\sum_{i=1}^{n} -\max_{\bar{x}_{ik}\in\bar{X}_i} \log \hat{P}(y_i|\bar{x}_{ik}; f, g)$.

The test accuracy with 5 individual trials for Max-Matching and three baselines are summarized in Table 2. Accordingly, "Pairwise" achieves the worst test accuracy, sine the model directly fits the group noise and the Non-IID property of the group is simply ignored. Plugging the selection mechanism ("Maximizing") can generally perform better, and the similar result occurs in "Matching" that explores the non-IID property in the group. In comparison, Max-Matching, which both considers the object correlation and the pairwise relation, can significantly outperform all these baselines. Actually, we also find that tuning the trade-off between Maximizing and Matching can achieve further improvement. Therefore, it is possible to acquire a better performance to select a proper weight for two terms in Max-Matching.

Application to Partial Label Learning

In this section, we validate Max-Matching in the setting of PLL, in which each instance is assigned with a set of candidate labels and only one of them is correct.

The experiments are conducted on five PLL datasets from various domains: *FG-NET* (Panis and Lanitis 2014) aims at facial age estimation; *MSRCv2* (Liu and Dietterich 2012) and *Bird Song* (Briggs, Fern, and Raich 2012) focus on object classification; *Yahoo! News* (Guillaumin, Verbeek, and Schmid 2010) and *Lost* (Cour, Sapp, and Taskar 2011) deal with face naming tasks. Each dataset is partitioned randomly into 8:1:1 for training, validation, and test. We compare Max-Matching with four popular PLL methods, including a non-parametric learning approach PLKNN (Hüllermeier and Beringer 2006); a maximum margin based method PLSVM (Nguyen and Caruana 2008); a statistical consistent method CLPL (Cour, Sapp, and Taskar 2011); and a decomposition based approach PALOC (Wu and Zhang 2018).

The test accuracy of 5 individual trials for our method and baselines are reported in Table 3. According to the results, PALOC shows extremely poor performance on datasets like *Bird Song* and *Lost*. This is because it has no explicit denoising mechanism to avoid the influence of group noise. PLKNN also achieves relatively inferior results due to its strong assumption on the data distribution. Although PLSVM and CLPL can generally perform better, they still fail to explore the non-IID characteristics of candidate labels. In comparison, Max-Matching have the best performance among all these methods, as it further considers the correlations among the candidate labels. Notably, on *FG-NET*, a challenging PLL dataset with a great many of strongly correlated candidate labels (7.48 partial labels per instance on average), Max-Matching is 4.37% better than the second best method CLPL on average.

To study the robustness of these methods in learning with different levels of group noise, we further conduct experiments on *Lost* and *Bird Song* with controlled proportion ϵ of partial labeled instances and controlled maximum numbers τ of partial labels. The test accuracy for varying ϵ and τ is summarized in Figure 4. Similar to the above results, PLKNN is unstable across these two datasets due to its assumption on data distribution. PALOC is also vulnerable to the group noise, and its accuracy drops quickly with the growth of ϵ and τ. Although the performances are relatively stable for CLPL and PLSVM, their test accuracy is consistently inferior to Max-Matching. These results further

demonstrate the effectiveness of Max-Matching in PLL.

Application to Recommender System

Finally, we conduct experiments of recommendation, which aims at recommending points of interest to the users, *e.g.*, item recommendation in e-commerce. The classical item-based collaborative filtering (Sarwar et al. 2001) critically depends on the trustworthy pairwise relationship, which is not practical on e-commercial websites. Generally, due to the varying interests of the user, his/her historically visited items are not always relevant to the subsequent items. Then, taking the user click sequence as a group and the next item as the target, we have the coarse relation as Figure 1(d). As a result, we face the problem of learning with group noise when applying the item-based collaborative filtering.

The offline experiments are implemented on a range of datasets from Amazon: *Video*, *Beauty*, and *Game*. In each dataset, the visited items of each user are segmented into subsets with at most 6 items, where the last item of each subset is taken as the target, and the others are taken as the group with noise. For each user, we randomly take two subsets for validation and test, and the remaining data are used for training. In the experiments, we consider several classical and advanced baselines, including a simple method that ranks items according to their popularity and recommends new items regarding the co-occurrence, Pop; a popular collaborative filtering method, I-CF (Linden, Smith, and York 2003); and two deep model based approaches that exploit the sequential behavior in the group, Caser (Tang and Wang 2018) and Att (Zhang et al. 2019). For Max-Matching, we recommend new items by ranking the probabilities $\hat{P}(y|x; f, g)$, where x can be the second last visited item (MM), or any item in the considered group (MM+)[2]. Following (Zhang et al. 2019), we report the performance on two widely used metrics, HIT@10 and NDCG@10. HIT@10 counts the fraction of times that the true next item is in the top-10 items, while NDCG@10 further assigns weights to the rank.

The average results of 5 individual trials in terms of the HIT@10 and NDCG@10 are summarized in Table 4. First, we compare MM with Pop and I-CF, which all recommend new items according to the last visited ones of users. Pop always shows extremely poor performance, as it is based on the popularity and cannot learn the correlations between items. While I-CF performs much better, it relies on the reliable pairwise relations without considering the group noise. By contrast, MM is robust to the fine-grained uncertain relations, which achieves the significant improvements. Second, we compare MM+ with Caser and Att, which are two recommendation approaches that can implicitly model the group relation. However, they mainly focus on the temporal behavior of users, making them fail to explicitly distinguish true relations from the irrelevant noise. By contrast, MM+ considers both the group relation and the denoising mechanism, and the experimental results on average demonstrate its effectiveness and rationality.

[2]Note that, MM+ is to compare the sequence-based recommendation methods Caser and Att which use all items in the group.

	Video		*Beauty*		*Game*	
	HIT	NDCG	HIT	NDCG	HIT	NDCG
Pop	0.515	0.397	0.401	0.258	0.402	0.252
I-CF	0.622	0.420	0.429	0.285	0.405	0.298
MM	**0.692**	**0.471**	**0.543**	**0.381**	**0.495**	**0.332**
Caser	0.643	0.425	0.523	0.345	0.493	0.311
Att	0.624	0.429	0.445	0.335	0.427	0.292
MM+	**0.694**	**0.473**	**0.561**	**0.389**	**0.518**	**0.345**

Table 4: Average HIT@10 (HIT for short) and NDCG@10 (NDCG for short) with standard deviation on Amazon.

Figure 5: The top-5 recall examples from NeuralCF and Max-Matching given the user clicked dress items. The 5 candidates in the first row are recalled by NeuralCF and the 5 candidates in the second row are recalled by Max-Matching.

We also conduct online experiments by deploying Max-Matching to the recommender system on one e-commerce platform. Like many large-scale recommenders, it consists of two stages, the recall stage and the ranking stage. The recall stage generates the most relevant candidate items that are related to the visited items of users in the middle-scale. The ranking stage scores the candidates in a fine-grained granularity for the top-k recommendation. We deploy Max-Matching to the recall stage and compare our method with the online item-based NeuralCF (He et al. 2017). Here, NeuralCF is supervised by the pairwise relations manually extracted from the user-click sequence. After one-week experiments, we achieved about 10% improvement on click-through rate (CTR). Figure 5 illustrates one example of the top-5 recommendation from NeuralCF and Max-Matching. According to the results, we can find the dress recommendation from NeuralCF is mixed with the shorts, which in fact, origins from the training with non-ideal pairwise relations.

Conclusion

In this paper, we focus on the learning paradigm with group noise, where a group of correlated objects are collectively related to the target with fine-grained uncertainty. To handle the group noise, we propose a novel Max-Matching mechanism in selecting the most confident objects in the group for training, which considers both the correlation among the group as well as the pairwise matching to the target. The experimental results in three different learning settings demonstrate its effectiveness. In the future, we will generalize Max-Matching to handle the independent pairwise relations, e.g., learning with label noise, and explore a better trade-off between two terms in our objective.

Acknowledgments

JCY and HXY was supported by NSFC No. U20A20222. CG was supported by NSFC No. 61973162, the Fundamental Research Funds for the Central Universities No. 30920032202, CCF-Tencent Open Fund No. RAGR20200101, the "Young Elite Scientists Sponsorship Program" by CAST No. 2018QNRC001, and Hong Kong Scholars Program No. XJ2019036. TLL was supported by Australian Research Council Project DE-190101473. BH was supported by the RGC Early Career Scheme No. 22200720, NSFC Young Scientists Fund No. 62006202, HKBU Tier-1 Start-up Grant, HKBU CSD Start-up Grant, and HKBU CSD Departmental Incentive Grant.

References

Algan, G.; and Ulusoy, I. 2019. Image classification with deep learning in the presence of noisy labels: A survey. *arXiv preprint arXiv:1912.05170* .

Belinkov, Y.; and Bisk, Y. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173* .

Briggs, F.; Fern, X. Z.; and Raich, R. 2012. Rank-loss support instance machines for MIML instance annotation. In *KDD*.

Carbonneau, M.; Cheplygina, V.; Granger, E.; and Gagnon, G. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.* 77: 329–353. doi:10.1016/j.patcog.2017.10.009.

Chen, P.; Ye, J.; Chen, G.; Zhao, J.; and Heng, P.-A. 2020. Robustness of Accuracy Metric and its Inspirations in Learning with Noisy Labels. *arXiv preprint arXiv:2012.04193* .

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from Partial Labels. *J. Mach. Learn. Res.* 12: 1501–1536.

Cui, K.; Chen, X.; Yao, J.; and Zhang, Y. 2018. Variational collaborative learning for user probabilistic representation. In *IJCAI*.

Feng, L.; and An, B. 2019. Partial label learning with self-guided retraining. In *AAAI*.

Frénay, B.; and Verleysen, M. 2014. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Networks Learn. Syst.* 25(5): 845–869. doi:10.1109/TNNLS.2013.2292894.

Ghosh, A.; Kumar, H.; and Sastry, P. 2017. Robust loss functions under label noise for deep neural networks. In *AAAI*.

Gong, C.; Liu, T.; Tang, Y.; Yang, J.; Yang, J.; and Tao, D. 2017. A regularization approach for instance-based superset label learning. *IEEE transactions on cybernetics* .

Guillaumin, M.; Verbeek, J.; and Schmid, C. 2010. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*.

Han, B.; Niu, G.; Yu, X.; Yao, Q.; Xu, M.; Tsang, I.; and Sugiyama, M. 2020. Sigua: Forgetting may make learning with noisy labels more robust. In *ICML*.

Han, B.; Yao, J.; Niu, G.; Zhou, M.; Tsang, I.; Zhang, Y.; and Sugiyama, M. 2018a. Masking: A new perspective of noisy supervision. In *NeurIPS*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018b. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *WWW*.

Hidasi, B.; and Karatzoglou, A. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *CIKM*.

Huang, P.-S.; He, X.; Gao, J.; Deng, L.; Acero, A.; and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*.

Hüllermeier, E.; and Beringer, J. 2006. Learning from ambiguously labeled examples. *Intell. Data Anal.* 10(5): 419–439.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kotzias, D.; Denil, M.; De Freitas, N.; and Smyth, P. 2015. From group to individual labels using deep features. In *KDD*.

Li, W.; Wang, L.; Li, W.; Agustsson, E.; and Van Gool, L. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862* .

Linden, G.; Smith, B.; and York, J. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.* .

Liu, G.; Wu, J.; and Zhou, Z. 2012. Key Instance Detection in Multi-Instance Learning. In *ACML*.

Liu, L.; and Dietterich, T. G. 2012. A conditional multinomial mixture model for superset label learning. In *NeurIPS*.

Liu, T.; and Tao, D. 2016. Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(3): 447–461. doi:10.1109/TPAMI.2015.2456899.

Manwani, N.; and Sastry, P. S. 2013. Noise tolerance under risk minimization. *IEEE Trans. Cybern.* 43(3): 1146–1151. doi:10.1109/TSMCB.2012.2223460.

Maron, O.; and Lozano-Pérez, T. 1998. A framework for multiple-instance learning. In *NeurIPS*.

Mnih, A.; and Salakhutdinov, R. R. 2008. Probabilistic matrix factorization. In *NeurIPS*.

Nguyen, N.; and Caruana, R. 2008. Classification with partial labels. In *KDD*.

Panis, G.; and Lanitis, A. 2014. An overview of research activities in facial age estimation using the FG-NET aging database. In *ECCV*.

Pao, H.; Chuang, S. C.; Xu, Y.; and Fu, H. 2008. An EM based multiple instance learning method for image classification. *Expert Syst. Appl.* 35(3): 1468–1472. doi:10.1016/j.eswa.2007.08.055.

Peng, M.; and Zhang, Q. 2019. Address instance-level label prediction in multiple instance learning. *arXiv preprint arXiv:1905.12226* .

Rahmani, R.; Goldman, S. A.; Zhang, H.; Krettek, J.; and Fritts, J. E. 2005. Localized content based image retrieval. In *ACM MM*.

Ray, S.; and Craven, M. 2005. Supervised versus multiple instance learning: An empirical comparison. In *ICML*.

Sarwar, B.; Karypis, G.; Konstan, J.; and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*.

Sedhain, S.; Menon, A. K.; Sanner, S.; and Xie, L. 2015. Autorec: Autoencoders meet collaborative filtering. In *WWW*.

Settles, B.; Craven, M.; and Ray, S. 2008. Multiple-instance active learning. In *NeurIPS*.

Tang, J.; and Wang, K. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *WSDM*.

Vanwinckelen, G.; do Ó, V. T.; Fierens, D.; and Blockeel, H. 2016. Instance-level accuracy versus bag-level accuracy in multi-instance learning. *Data Min. Knowl. Discov.* 30(2): 313–341. doi:10.1007/s10618-015-0416-z.

Wang, X.; Yan, Y.; Tang, P.; Bai, X.; and Liu, W. 2018. Revisiting multiple instance neural networks. *Pattern Recognit.* 74: 15–24. doi:10.1016/j.patcog.2017.08.026.

Wu, S.; Tang, Y.; Zhu, Y.; Wang, L.; Xie, X.; and Tan, T. 2019. Session-based recommendation with graph neural networks. In *AAAI*.

Wu, X.; and Zhang, M.-L. 2018. Towards enabling binary decomposition for partial label learning. In *IJCAI*.

Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020. Parts-dependent label noise: Towards instance-dependent label noise. In *NeurIPS*.

Xia, X.; Liu, T.; Wang, N.; Han, B.; Gong, C.; Niu, G.; and Sugiyama, M. 2019. Are anchor points really indispensable in label-noise learning? In *NeurIPS*.

Xue, H.-J.; Dai, X.; Zhang, J.; Huang, S.; and Chen, J. 2017. Deep matrix factorization models for recommender systems. In *IJCAI*.

Yang, Y.; Qiu, J.; Song, M.; Tao, D.; and Wang, X. 2020. Distilling knowledge from graph convolutional networks. In *CVPR*.

Yao, J.; Wang, J.; Tsang, I. W.; Zhang, Y.; Sun, J.; Zhang, C.; and Zhang, R. 2019. Deep learning from noisy image labels with quality embedding. *IEEE Trans. Image Process.* 28(4): 1909–1922. doi:10.1109/TIP.2018.2877939.

Yao, J.; Wu, H.; Zhang, Y.; Ivor W., T.; and Sun, J. 2019. Safeguarded dynamic label regression for noisy supervision. In *AAAI*.

Yao, Y.; Chen, G.; Jiehui, D.; Xiuhua, C.; Jianxin, W.; and Yang, J. 2020a. Deep discriminative CNN with temporal ensembling for ambiguously-labeled image classification. In *AAAI*.

Yao, Y.; Liu, T.; Han, B.; Gong, M.; Deng, J.; Niu, G.; and Sugiyama, M. 2020b. Dual T: Reducing estimation error for transition matrix in label-noise learning. *NeurIPS* .

Yu, F.; and Zhang, M.-L. 2016. Maximum margin partial label learning. In *ACML*.

Zhang, M.-L.; and Yu, F. 2015. Solving the partial label learning problem: An instance-based approach. In *IJCAI*.

Zhang, M.-L.; Zhou, B.-B.; and Liu, X.-Y. 2016. Partial label learning via feature-aware disambiguation. In *KDD*.

Zhang, S.; Tay, Y.; Yao, L.; Sun, A.; and An, J. 2019. Next item recommendation with self-attentive metric learning. In *AAAI*.