End-to-End Weak Supervision

Anonymous Author(s)
Affiliation
Address
email

Abstract

Aggregating multiple sources of weak supervision (WS) can ease the data-labeling bottleneck prevalent in many machine learning applications, by replacing the tedious manual collection of ground truth labels. Current state of the art approaches that do not use any labeled training data, however, require two separate modeling steps: Learning a probabilistic latent variable model based on the WS sources – making assumptions that rarely hold in practice – followed by downstream model training. Importantly, the first step of modeling does not consider the performance of the downstream model. To address these caveats we propose an end-to-end approach for directly learning the downstream model by maximizing its agreement with probabilistic labels generated by reparameterizing previous probabilistic posteriors with a neural network. Our results show improved performance over prior work in terms of end model performance on downstream test sets, as well as in terms of improved robustness to dependencies among weak supervision sources.

1 Introduction

The success of supervised machine learning methods relies on the availability of large amounts of labeled data. The common process of manual data annotation by humans, especially when domain experts need to be involved, is expensive, both in terms of effort and cost, and as such presents a major bottleneck for deploying supervised learning methods to new domains and applications.

Recently, data programming, a paradigm that makes use of multiple sources of noisy labels, has emerged as a promising alternative to manual data annotation [30]. It encompasses previous paradigms such as distant supervision from external knowledge bases [28, 33], crowdsourcing [14, 25, 13, 41], and general heuristic and rule-based labeling of data [22, 19]. In the data programming framework, users encode domain knowledge into so called labeling functions (LFs), which are functions (e.g. domain heuristics or knowledge base derived rules) that noisily label subsets of data. The main task for learning from multiple sources of weak supervision is then to recover the sources' accuracies in order to estimate the latent true label, without access to ground truth data. In previous work [30, 31, 18], this is achieved by first learning a generative probabilistic graphical model (PGM) over the weak supervision sources and the latent true label to estimate *probabilistic labels*, which are then used in the second step to train a *downstream model* via a noise-aware loss function.

Data programming has led to a wide variety of success stories in domains such as healthcare [17, 16] and e-commerce [5], but the existing PGM based frameworks still come with a number of drawbacks. The separate PGM does not take the predictions of the downstream model into account, and indeed this model is trained independently of the PGM. In addition, current approaches for estimating the unknown class label via a PGM need to rely on computationally expensive approximate sampling methods [30], estimation of the full inverse of the LFs covariance matrix [31], or they need to make strong independence assumptions [18]. Furthermore, existing prior work and the associated theoretical analyses make assumptions that may not hold in practice [30, 31, 18], such as availability of a well-specified generative model structure (i.e. that the dependencies and correlations between

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

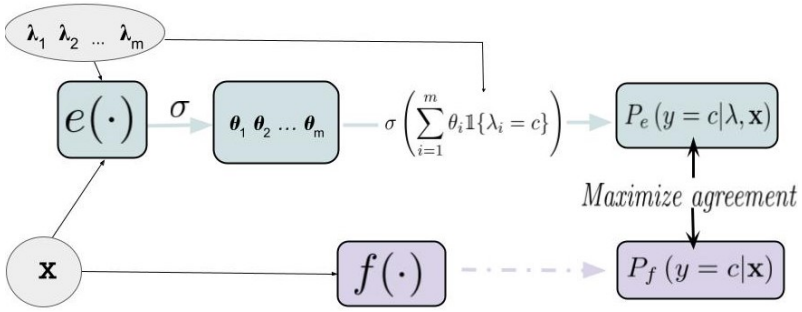


Figure 1: For a task with unobserved ground truth labels y, given m sources of weak supervision λ_i and training features $\mathbf{x}$, we train a downstream model f by maximizing the agreement of its predictions $P_f(y = c|\mathbf{x})$ with probabilistic labels $P_e(y = c|\,\boldsymbol{\lambda}, \mathbf{x})$ generated by reparameterizing the posterior of prior work with sample-dependent accuracy scores θ produced by an encoder network e.

the weak sources have been correctly specified by the user), that LF errors are randomly distributed across samples, and that the latent label is independent of the features given the weak labels (i.e. only the joint distribution between the sources and labels needs to be modeled).

We present an end-to-end method for training neural networks with, exclusively, multiple sources of weak supervision as noisy signals for the latent labels. The proposed approach is based on 1) *reparameterizing previous PGM based posteriors* with a neural encoder network that produces *attention scores for each weak supervision source*; and 2) training the encoder and downstream model on *the same target loss, using the other model's predictions as constant targets*, to maximize the agreement between both models. The proposed method needs no labeled training data, and neither assumes sample-independent source accuracies nor redundant features for latent label modeling. We show empirically that it is not susceptible to highly correlated LFs. In addition, the proposed approach can learn from multiple probabilistic sources of weak supervision.

Our contributions include:

- We introduce a flexible, end-to-end method for learning models from multiple sources of weak supervision.
- We empirically demonstrate that the method is naturally robust to adversarial sources as well as highly correlated weak supervision sources.
- We release an open-source, end-to-end system for arbitrary PyTorch downstream models that will allow practitioners to take advantage of our approach[1].
- We show that our method outperforms, by as much as 6.1 F1 points, state-of-the-art latent label modeling approaches on 4 out of 5 relevant benchmark datasets, and achieves state-of-the-art performance on a crowdsourcing dataset against methods specifically designed for this setting.

2 Related Work

Multi-source Weak Supervision The data programming paradigm [30] allows users to programmatically label data through multiple noisy sources of labels, by treating the true label as a latent variable of a generative PGM. Several approaches for learning the parameters of the generative model have been introduced [31, 18, 10] to address computational complexity issues. Existing methods are susceptible to misspecification of the dependencies and correlations between the LFs, which can lead to substantial losses in performance [35]. Indeed, it is common practice to assume a conditionally independent model – without any dependencies between the sources – in popular libraries [5, 29] and related research [14, 2, 39, 7], even though methods to learn the intra-LF structure have been proposed [4, 37, 38]. As in the approach proposed in this paper, the aforementioned methods do not assume any labeled training data, i.e. the downstream model is learned based solely on outputs of multiple LFs on unlabeled data. The traditional co-training paradigm [6] on the other hand is similar in spirit but requires some labeled data to be available. Recent methods that study the co-training

[1]To be released on GitHub upon acceptance. Research code can be found in the supplementary material.

setup where labeled training data supplements multiple WS sources, include [3, 24]. Note that the experiments in [3, 24] rely on large pre-trained language models, making the applicability of the approach without such models or to non-text domains unclear.

Crowdsourcing Aggregating multiple noisy labels is also a core problem studied in the crowdsourcing literature. Common approaches model worker performance and the unknown label jointly [14, 13, 41] using expectation maximization (EM) or similar approaches. Some core differences to learning from weak supervision sources are that errors by crowdworkers are usually assumed to be random, and that task assignment is not always fixed but can be optimized for. Multiple end-to-end methods have been proposed in the crowdsourcing problem setting [32, 21, 40, 26, 34, 8], often focusing on image labeling and EM-like algorithms for modeling the workers. Importantly, our proposed approach can be used in general applications with weak supervision from multiple sources without any restrictive assumptions specific to crowdsourcing, and we show that our approach outperforms the aforementioned methods on a crowdsourcing benchmark task.

3 End-to-End Weak Supervision

In this section we present our flexible base algorithm that we call E2E, which can be extended to probabilistic sources and other network architectures (Section 7). See Algorithm 1 for the pseudocode of E2E.

Algorithm 1 The proposed end-to-end (E2E) algorithm for learning from multiple weak supervision sources.

input: batch size n, networks e, f, inverse temperatures τ_1, τ_2, noise-aware loss function L, class balance $P(y)$.
for sampled minibatch $\{z^{(k)} = (\mathbf{x}^{(k)}, \boldsymbol{\lambda}^{(k)})\}_{k=1}^{n}$ **do**
  for all $k \in \{1, \ldots, n\}$ **do**
    # Produce accuracy scores for all weak sources
    $\theta\left(z^{(k)}\right) = \mathrm{softmax}\left(e(z^{(k)})\tau_1\right)$
    # Generate probabilistic labels
    define $\mathbf{s}^{(k)}$ **as** $\mathbf{s}^{(k)} = \theta(z^{(k)})^T \bar{\boldsymbol{\lambda}}^{(k)}$
    $y_e^{(k)} = P_e(y|z^{(k)}) = \mathrm{softmax}\left(\mathbf{s}^{(k)}\tau_2\right) \odot P(y)$
    # Downstream model forward pass
    $y_f^{(k)} = P_f(y|\mathbf{x}^{(k)}) = f(\mathbf{x}^{(k)})$
  end for
  $\mathcal{L}_f = \frac{1}{n}\sum_{k=1}^{n} L\left(y_f^{(k)}, \texttt{stop-grad}\left(y_e^{(k)}\right)\right)$
  $\mathcal{L}_e = \frac{1}{n}\sum_{k=1}^{n} L\left(y_e^{(k)}, \texttt{stop-grad}\left(y_f^{(k)}\right)\right)$
  update e to minimize $\mathcal{L}_e$, and f to minimize $\mathcal{L}_f$
end for
return downstream network $f(\cdot)$

3.1 Problem Setup

Let $(\mathbf{x}, y) \sim \mathcal{D}$ be the data generating distribution, where the unknown labels belong to one of C classes: $y \in \mathcal{Y} = \{1, ..., C\}$. As in [30], users provide an unlabeled training set $X = \{\mathbf{x}_i\}_{i=1}^{N}$, and m labeling functions $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\mathbf{x}) \in \{0, 1, ..., C\}^m$, where 0 means that the LF abstained from labeling for any class. We write $\bar{\boldsymbol{\lambda}} = (\mathbb{1}\{\boldsymbol{\lambda} = 1\}, \ldots, \mathbb{1}\{\boldsymbol{\lambda} = C\}) \in \{0, 1\}^{m \times C}$ for the one-hot representation of the LF votes provided by the m LFs for C classes. Our goal is to train a downstream model $f : \mathcal{X} \to \mathcal{Y}$ on a *noise-aware* loss $L(y_f, y_e)$ that operates on the model's predictions $y_f = f(\mathbf{x})$ and *probabilistic labels* $y_e = P_e(y|\,\boldsymbol{\lambda}, \mathbf{x})$ generated by an encoder model e that has access to LF votes and features. Note that prior work restricts the probabilistic labels to only being estimated from the LFs $\boldsymbol{\lambda}$.

3.2 Posterior Reparameterization

Previous PGM based approaches assume that the joint distribution $p(\boldsymbol{\lambda}, y)$ of the LFs and the latent true label can be modeled as a Markov Random Field (MRF) with pairwise dependencies between weak supervision sources [30, 29, 31, 18, 10]. These models are parameterized by a set of LF accuracy and intra-LF correlation parameters and in some cases by additional parameters to model LF and class label propensity. Note however, that the aforementioned models ignore features X when modeling the latent labels and therefore disregard that LFs may differ in their accuracy across samples and data slices.

We relax these assumptions, and instead view the latent label as an *aggregation of the LF votes that is a function of the entire set of LF votes and features, on a sample-by-sample basis*. That is, we model

the probability of a particular sample $\mathbf{x}$ having the class label $c \in \mathcal{Y}$ as

$$P_e(y = c|\, \boldsymbol{\lambda}, \mathbf{x}) = \mathrm{softmax}\,(\mathbf{s})_c\, P(y = c), \tag{1}$$

$$\mathbf{s} = \theta(\boldsymbol{\lambda}, \mathbf{x})^T \bar{\boldsymbol{\lambda}} \in \mathbb{R}^C\,. \tag{2}$$

where $\theta(\boldsymbol{\lambda}, \mathbf{x}) \in \mathbb{R}^m$ weighs the LF votes on a sample-by-sample basis and the softmax for class c on s is defined as

$$\mathrm{softmax}\,(\mathbf{s})_c = \frac{\exp\left(\theta(\boldsymbol{\lambda}, \mathbf{x})^T \mathbb{1}\{\boldsymbol{\lambda} = c\}\right)}{\sum_{j=1}^{C} \exp\left(\theta(\boldsymbol{\lambda}, \mathbf{x})^T \mathbb{1}\{\boldsymbol{\lambda} = j\}\right)}.$$

While we do not use the class balance $P(y)$ in our experiments for our own model, E2E, it is frequently assumed to be known [31, 18, 10], and can be estimated from a small validation set, or from unlabeled data as described in [31]. Our formulation can be seen as a reparameterization of the posterior of the pairwise MRFs in [29, 31, 18], where θ corresponds to the LF accuracies that are fixed across the dataset and are solely learned via LF agreement and disagreement signals, ignoring the informative features. We further motivate this formulation and expand upon this connection in the appendix A.

3.3 Neural Encoder

Based on the setup introduced in the previous section and captured in Eq. (1), our goal is to estimate latent labels by means of learning sample-dependent accuracy scores $\theta(\boldsymbol{\lambda}, \mathbf{x})$, which we propose to parameterize by a neural encoder e. This network takes as input the features $\mathbf{x}$ and the corresponding LF outputs $\boldsymbol{\lambda}(\mathbf{x})$ for a data point, and outputs unnormalized scores $e(\boldsymbol{\lambda}, \mathbf{x}) \in \mathbb{R}^m$. Specifically, we define

$$\theta(\boldsymbol{\lambda}, \mathbf{x}) = \tau_2 \cdot \mathrm{softmax}\left(e(\boldsymbol{\lambda}, \mathbf{x})\tau_1\right), \tag{3}$$

where τ_2 is a constant factor that scales the final softmax transformation in relation to the number of LFs m, and is equivalent to an inverse temperature for the output softmax in Eq. 1. It is motivated by the fact that most LFs are sparse in practice, and especially when the number of LFs is large this leads to small accuracy magnitudes without scaling (since, without scaling, the accuracies after the softmax sum up to one)[2]. τ_1 is an inverse temperature hyperparameter that controls the smoothness of the predicted accuracy scores: The lower τ_1 is, the less emphasis is given to a small number of LFs – as $\tau_1 \to 0$, the model aggregates according to the equal weighted vote. The $\mathrm{softmax}$ transformation naturally encodes our understanding of wanting to aggregate the weak sources to generate the latent label.

3.4 Training the Encoder

The key question now is how to train e, i.e. how can we learn an accurate mapping of the sample-by-sample accuracies, given that we do not observe any labels?

We hypothesize that in most practical cases, features, latent label, and labeling function aggregations are intrinsically correlated due to the design decisions made by the users defining the features and LFs. Thus, we can jointly optimize e and f by maximizing their agreement with respect to the target downstream loss L in an end-to-end manner.

See Algorithm 1 for pseudocode of the resulting E2E algorithm. The natural classification loss is the cross-entropy, which we use in our experiments, but in order to encode our desire of maximizing the agreement of the two separate models that predict based on different views of the data, we adapt it[3] in the following form: The loss is symmetrized in order to compute the gradient of both models using the other model's predictions as targets. To that end, it is crucial to use the `stop-grad` operation on the targets (the second argument of L), i.e. to treat them as though they were ground truth labels. This choice is supported by our synthetic experiment and ablations. This operation has also been shown to be crucial in siamese, non-contrastive, self-supervised learning, both empirically [20, 11] and theoretically [36]. By minimizing simultaneously, both, $L(y_e, y_f)$ and $L(y_f, y_e)$ to jointly learn the network parameters for e and the downstream model f respectively, we learn the accuracies of the noisy sources $\boldsymbol{\lambda}$ that best explain the patterns observed in the data, and vice versa the feature-based predictions that are best explained by aggregations of LF voting patterns.

[2] In our main experiments we set $\tau_2 = \sqrt{m}$.

[3] This holds for any asymmetric loss, while for symmetric losses this is not needed.

Table 1: Test F1 performance of various label models over seven runs using different random seeds, are averaged out $\pm$ standard deviation. The top 2 performance scores are highlighted as **First**, **Second**. Triplet-median [10] is not listed as it only converged for IMDB with 12 LFs (F1 = 73.0 ± 0.22), and Spouses (F1 = 48.7 ± 1.0). Sup. (Val. set) is the performance of the downstream model trained in a supervised manner on the labeled validation set. The rest are state-of-the-art latent label models. For reference, we also report the *Ground truth* performance of a fully supervised model trained on true training labels (which are unused by all other models, and not available for Spouses).

Model	**Spouses** (9 LFs)	**ProfTeacher** (99 LFs)	**IMDB** (136 LFs)	**IMDB** (12 LFs)	**Amazon** (175 LFs)
Ground truth	–	90.65 ± 0.29	86.72 ± 0.40	86.72 ± 0.40	92.93 ± 0.68
Sup. (Val. set)	20.4 ± 0.2	73.34 ± 0.00	68.76 ± 0.00	68.76 ± 0.00	84.18 ± 0.00
Snorkel	48.79 ± 2.69	85.12 ± 0.54	**82.22 ± 0.18**	**74.45 ± 0.58**	80.54 ± 0.41
Triplet	45.88 ± 3.64	74.43 ± 10.59	75.36 ± 1.92	73.15 ± 0.95	75.44 ± 3.21
Triplet-Mean	**49.94 ± 1.47**	82.58 ± 0.32	79.03 ± 0.26	73.18 ± 0.23	79.44 ± 0.68
Majority vote	40.67 ± 2.01	**85.44 ± 0.37**	80.86 ± 0.28	74.13 ± 0.31	**84.20 ± 0.52**
E2E	**51.98 ± 1.60**	**86.98 ± 0.45**	**82.10 ± 0.45**	**77.22 ± 1.02**	**86.60 ± 0.71**

3.5 E2E Design Choices

Note that it is necessary to encode the inductive bias that the unobserved ground truth label y is a (normalized) linear combination of LF votes – weighted by sample- and feature-dependent accuracy scores. Otherwise, if the encoder network directly predicts $P_e(y|\,\boldsymbol{\lambda}, \mathbf{x})$ instead of the accuracies $\theta(\boldsymbol{\lambda}, \mathbf{x})$, the pair of networks e, f have no incentive to output the desired latent label, without observed labels. We do acknowledge that this two-player cooperation game with strong inductive biases could still allow for degenerate solutions. However, we empirically show that our simple E2E model that goes beyond multiple earlier WS assumptions is 1) competitive and frequently outperforms state-of-the-art PGM-based and crowdsourcing models (see Tables 1 and 2); and 2) is robust against massive LF correlations and able to recover the performance of a fully supervised model on a synthetic example, while all other models break in this setting (see section 4.3 and appendix F).

4 Experiments

Datasets As in related work on label models for weak supervision [30, 31, 18, 10], we focus for simplicity on the binary classification case with unobserved ground truth labels $y \in \{-1, 1\}$. See Table 3 for details about dataset sizes and the number of LFs used. We also run an experiment on a multi-class, crowdsourcing dataset (see subsection 4.2). We evaluate the proposed end-to-end system for learning a downstream model from multiple weak supervision sources on previously used benchmark datasets in weak supervision work [29, 7, 10]. Specifically, we evaluate test set performance on the following classification datasets:

The IMDB movie review dataset [27] contains movie reviews to be classified into positive and negative sentiment. We run two separate experiments, where in one we use the same 12 labeling functions as in [10], and for the other we choose 136 text-pattern based LFs. More details on the LFs can be found in the appendix C.

We use a subset of the *Amazon review dataset* [23], where the task is to classify product reviews into positive and negative sentiment.

We use the *BiasBios biographies* [15] to distinguish between binary categories of frequently occurring occupations and use the same subset of professor vs teacher classification as in [7].

Finally, we use the highly unbalanced *Spouses dataset* (90% negative class labels), where the task is to identify mentions of spouse relationships amongst a set of news articles from the Signal Media Dataset [12]. For the Spouses datasets, the same data split, and LFs as in [18] are used, while for the rest we take a small subset of the – official, or as used in previous work – test set as validation set. This is common practice in the related work [29, 31, 18, 7] for tuning hyperparameters, and allows for a fair comparison of models.

4.1 Benchmarking Weak Supervision Label Models

To evaluate the proposed system, we benchmark it against state-of-the-art systems that aggregate multiple weak supervision sources for classification problems, without any labeled training data. We compare our proposed approach with the following systems: 1) *Snorkel*, a popular system proposed in

[29, 31]; 2) *Triplet*, exploits a closed-form solution for binary classification under certain assumptions [18]; and 3) *Triplet-mean* and *Triplet-median* [10], which are follow-up methods based on *Triplet* with the aim of making the method more robust.

We report the held-out test set performance of E2E's downstream model f. Note that in many settings it is often not possible to apply the encoder model to make predictions at test time, since the LFs usually do not cover all data points (e.g. in Spouses only 25.8% of training samples get at least one LF vote), and can be difficult to apply to new samples (e.g. when the LFs are crowdsourced annotations). In contrast, the downstream model is expected to generalize to arbitrary unseen data points.

We observe strong results for our model, with 4 out of 5 top scores, and a lift of 6.1 F1 points over the next best label model in the Amazon dataset. Our results are summarized in Table 1. Since our model is based on a neural network, we hypothesize that the large relative lift in performance on the Amazon review dataset is due to it being the largest dataset size on which we evaluate on – we expect this lift to hold or become larger as the training set size increases. To obtain the comparisons shown in Table 1, we run Snorkel over six different label model hyperparameter configurations, and report the best score based on the validation set AUC. We do not report Triplet-median in the main table, since it only converged for the two tasks with very small numbers of labeling functions. Interestingly, we observed that training the downstream model on the hard labels induced by majority vote leads to a competitive performance, better than triplet methods in four out of five datasets. This baseline is not reported in previous papers (only the raw majority vote is usually reported, without training a classifier). Our own model, E2E, on the other hand consistently improves over the majority vote baseline (which in Table 4, in the appendix, can be seen to lead to similar performance as an untrained encoder network, e, that is left at its random initialization).

4.2 Crowdsourcing dataset

Data programming and crowdsourcing methods have been rarely compared against each other, even though the problem setup is quite similar. Indeed, end-to-end systems specifically for crowdsourcing have been proposed [32, 26, 34, 8]. These methods follow crowdsourcing-specific assumptions and modeling choices (e.g. independent crowdworkers, a confusion matrix model for each worker, and in general build upon [14]). Still, since crowdworkers can be seen as a specific type of labeling functions, the performance of general WS methods on crowdsourcing datasets is of interest, but has so far not been studied. We therefore choose to also evaluate our method on the multi-class LabelMe image classification dataset that was previously used in the core related crowdsourcing literature [34, 8]. The results are reported in Table 2, and more details on this experiment can be found in Appendix E.

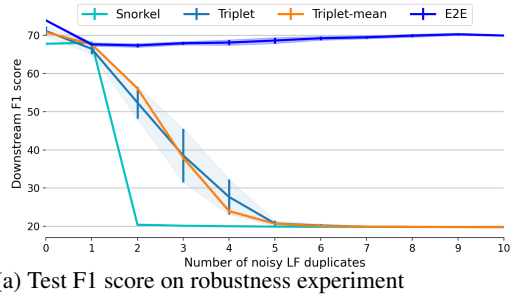
Table 2: Test accuracy scores on the crowdsourced, multi-class LabelMe image classification dataset.

Model	Accuracy
Majority vote	79.23 ± 0.5
MBEM [26]	76.84 ± 0.4
DoctorNet [21]	81.31 ± 0.4
CrowdLayer [34]	82.83 ± 0.4
AggNet [1]	84.35 ± 0.4
MaxMIG [8]	$\mathbf{85.45 \pm 1.0}$
Snorkel+CE	82.89 ± 0.7
E2E+CE	82.46 ± 0.8
Snorkel+MIG	85.15 ± 0.8
E2E+MIG	$\mathbf{86.36 \pm 0.3}$

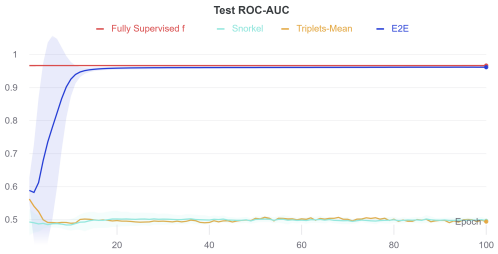
Note that the evaluation procedure in [8] reports the best test set performance for all models, while we follow the more standard practice of reporting results obtained by tuning based on a small validation set – as in our main experiments. We find that our model, E2E is able to outperform Snorkel as well as multiple state-of-the-art methods that were specifically designed for crowdsourcing (including several end-to-end approaches). Interestingly, this is achieved by using the mutual information gain loss (MIG) function introduced in [8], which significantly boosts performance of both Snorkel (the end-model, f, trained on the MIG loss with respect to soft labels generated by the first Snorkel label model step) and E2E that use the cross-entropy (CE) loss. This suggests that the MIG loss is a great choice for the special case of crowdsourcing, due to its strong assumptions common to crowdsourcing which are much less likely to hold for general LFs. This is reflected in our ablations too, where using the MIG loss leads to a consistently worse performance on our main multi-source weak supervision datasets.

4.3 Robustness to Adversarial LFs and LF correlations

Users will sometimes generate sources they mistakenly think are accurate. This also encompasses the 'Spammer' crowdworker-type studied in the crowdsourcing literature. Therefore, it is desirable



(a) Test F1 score on robustness experiment as a function of the number of adversarial LFs.



(b) Test AUC by epoch in an experiment where one LF corresponds to the true class label and others are random.

Figure 2: E2E is significantly more robust against correlated adversarial (left) or random (right) LFs than prior work whose assumptions make them equivalent to a Naive Bayes model. For subfigure (a), we duplicate a fake adversarial LF up to 10 times, and observe that our end-to-end system is robust against the adversarial LF, while other systems quickly degrade in performance (over ten random seeds). In (b), we let one LF be the true labels y^* and then duplicate a LF that votes according to a coin flip 2, 5, ..., 2000 times. We plot the test AUC performance curve as a function of the epochs, averaged out over the different number of duplications (and five random seeds). E2E consistently recovers the test performance of the supervised end-model f trained directly on the true labels y^*, whose end performance (AUC $= 0.967$) is shown in red.

to build models that are robust against such sources. We argue that our system that is trained by maximizing the agreement between an aggregation of the sources and the downstream model's predictions should be able to distinguish the adversarial sources. In Fig. 2a we show that our system does not degrade in its initial performance, even after duplicating an adversarial LF ten times. Prior latent label models, on the other hand, rapidly degrade, given that they often assume the weak label sources to be conditionally independent given the latent label, equivalent to a Naive Bayes generative model. Note that the popular open-source implementation of [29, 31] does not support user-provided LF dependencies modeling, while [18, 10] did not converge in our experiments when modeling dependencies, and as such we were not able to test their performance when the correlation dependencies between the duplicates are provided (which in practice, of course, are not known).
We also run a synthetic experiment inspired by [8], where one LF is set to the true labels of the ProfTeacher dataset, i.e. $\lambda_1 = y^*$, while the other LF simply votes according to a coin flip, i.e. $\lambda_2 \sim P(y)$, and we then duplicate this latter LF, i.e. $\lambda_3 = \cdots = \lambda_m = \lambda_2$. Under this setting, our E2E model is able to consistently *recover the fully supervised performance* of the same downstream model directly trained on the true labels y^*, *even when we duplicate the random LF up to* 2000 *times* ($m = 2001$). Snorkel and triplet methods, on the other hand, were unable to recover the true label (AUC ≈ 0.5). Importantly, we find that the design choices for E2E are to a large extent key in order to recover the true labels in a stable manner as in 2b. Various other choices either collapse similarly to the baselines, are not able to fully recover the supervised performance, or lead to unstable test performance curves, see Fig. 3 in the appendix. More details on the experimental design and an extensive discussion, ablation, and figures based on the synthetic experiment can be found in the appendix F.

4.4 Implementation Details

Here we provide a high-level overview over the used encoder architecture, the LF sets, and the features. More details, especially hyperparameter and architecture details, are provided in Appendix C. All downstream models are trained with the (binary) cross-entropy loss, and our model with the symmetric version of it that uses `stop-grad` on the targets.

Encoder network The encoder network e does not need to follow a specific neural network architecture and we therefore use a simple multi-layer perceptron (MLP) in our benchmark experiments.
Features for the encoder A big advantage of our model is that it is able to take into account the features $\mathbf{x}$ for generating the sample-by-sample source accuracies. For all datasets, we concatenate the LF outputs with the same features that are used by the downstream model as input of our encoder model (for Spouses we use smaller embeddings than the ones used by the downstream LSTM).

Table 3: Dataset details, where training, validation and test set sizes are N_{train}, N_{val}, N_{test} respectively, and f denotes the downstream model type. We also report the total coverage Cov. of all LFs, which refers to the percentage of training samples which are labeled by at least one LF (the rest is not used). For IMDB we used two different sets of labeling functions of sizes 12 and 136.

Dataset	#LFs	N_{train}	Cov. (in %)	N_{val}	N_{test}	f
Spouses	9	$22,254$	25.8	2811	2701	LSTM
BiasBios	99	$12,294$	81.8	250	$12,044$	MLP
IMDB	12	$25k$	88.0	250	$24,750$	MLP
IMDB	136	$25k$	83.1	250	$24,750$	MLP
Amazon	175	$160k$	65.5	500	$39,500$	MLP

Weak supervision sources For the Spouses dataset, and the IMDB variant with 12 LFs, we use the same LFs as in [18, 10] respectively. The remaining three LF sets were selected by us prior to running experiments. These LFs are all pattern- and regex-based heuristics, while the Spouses experiments also contain LFs that are distant supervision sources based on DBPedia.

5 Ablations

In this section we demonstrate the strength of the E2E model design decisions. We perform extensive ablations on all four main datasets but Spouses for twenty configurations of E2E with different encoder architectures, hyperparameters, and loss functions. The tabular results and a more detailed discussion than in the following can be found in Appendix D.

We observe that ignoring the features when modeling the sample-dependent accuracies, i.e. $\theta(\boldsymbol{\lambda}, \mathbf{x}) = \theta(\boldsymbol{\lambda})$, usually underperforms by up to 1.2 F1 points. A more drastic drop in performance, up to 4.9 points, occurs when the encoder network is linear, i.e. without hidden layers, as in [8]. It also proves helpful to scale the $\mathrm{softmax}$ in Eq. 3 by $\sqrt{m}$ via the inverse temperature parameter τ_2. Further, while the MIG loss proved important for E2E to achieve state-of-the-art performance on the crowdsourcing dataset (with a similar lift in performance observable for Snorkel using MIG for downstream model training), this does not hold for the main datasets. This indicates that the MIG loss is a good choice for crowdsourcing, but not for more general WS settings.

Our ablations also show that it is important to restrict the accuracies to a positive interval (e.g. (0, 1), with the sigmoid function being a good alternative to the softmax we use). On the one hand, this encodes the inductive bias that LFs are not adversarial, i.e. can not have negative accuracies, (using tanh to output accuracy scores does not perform well), and on the other hand does not give the encoder network too much freedom in the scale of the scores (using ReLU underperforms significantly as well).

Additionally, we find that our choice of using the symmetric cross-entropy loss with `stop-grad` applied to the targets is crucial for the obtained strong performance of E2E. Removing the `stop-grad` operation, or using the standard cross-entropy (without `stop-grad` on the target) leads to significantly worse scores and a very brittle model. Losses that already are symmetric (e.g. L1 or Squared Hellinger loss) neither need to be symmetrized nor use `stop-grad`. While the L1 loss consistently underperforms, we find that the Squared Hellinger loss can lead to better performance on two of the four datasets.

However, only the symmetric cross-entropy loss with `stop-grad` on the targets is shown to be robust and able to recover the true labels in our synthetic experiment in Section 4.3. Thus, to complement the above ablation on real datasets, we additionally run extensive ablations on this synthetic setup in Appendix F. This synthetic ablation gives interesting insights, and strongly supports the proposed design of E2E. Indeed, many choices for E2E that perform well enough on the real datasets, such as no features for the encoder, $\tau_2 = 1$, sigmoid parameterized accuracies, and all other losses that we evaluated, lead to significantly worse performance and less robust learning on the synthetic adversarial setups.

6 Practical Aspects and Limitations

On why it works & degenerate solutions Overall, E2E avoids trivial overfitting and degenerate solutions by hard-coding the encoder generated labels as a (normalized) linear combination of the m

LF outputs, weighted by m sample-dependent accuracy scores. This design choice also ensures that the randomly initialized e will lead the downstream model f that is trained on soft labels generated by the random encoder, to obtain performance similar to when f is trained on majority vote labels. In fact, the random-encoder-E2E variant itself often outperforms other baselines, and triplet methods in particular (see appendix B).

Empirically, we only observed degenerate solutions when training for too many epochs. Early-stopping on a small validation set ensures that a strong final solution is returned, and should be done whenever such a set exists or is easy to create. When no validation set is available, we find that choosing the temperature hyperparameter in Eq. 3 such that $\tau_1 \leq 1/3$ avoids collapsed solutions on all our datasets. This can be explained by the fact that a lower inverse temperature forces the encoder-predicted label to always depend on multiple LF votes when available, rather than a single one (which happens when the $\mathrm{softmax}$ in Eq. 3 becomes a `max` as $\tau_1 \to \infty$). This makes it harder for the encoder to overfit to individual LFs. Our ablations indicate that this temperature parameter setting comes at a small cost in terms of loss in downstream performance, compared to when using a validation set for early stopping. Thus, when no validation set is available, we advise to lower τ_1.

Complex downstream models We have shown that E2E achieves competitive or state-of-the-art performance on all datasets we tried it on, for a given set of LFs. In practice, however, this LF set needs to first be defined by users. This can be done via an iterative process, where the feedback is sourced from the quality of the probabilistic labels generated by the label model. A limitation of our model, is that each such iteration would require training the downstream model, f. When f is slow to train, this may slow down the LF development cycle and lead to unnecessary energy consumption. A practical solution to this can be to a) do the iteration cycle with a less complex downstream model; or b) use the fast to train PGM-based label models to choose a good LF set, and then move to E2E in order to achieve incrased downstream performance.

7 Extensions

Probabilistic labeling functions Our learning method can easily support labeling functions that output continuous scores instead of discrete labels as in [9]. In particular, this includes probabilistic sources that output a distribution over the potential class labels. This can be encoded in our model by changing the one-hot representation of our base model to a continuous representation $\bar{\boldsymbol{\lambda}} \in [0,1]^{m \times C}$.

Modeling more structure While we use a simple multi-layer perceptron (MLP) as our encoder e in our benchmark experiments, our formulation is flexible to support arbitrarily complex networks. In particular, we can naturally model dependencies amongst weak sources via edges in a Graph Neural Network (GNN), where each LF is represented by a node that is given the LF outputs as features. Furthermore, while we only explicitly reparameterized the accuracy parameters of the sources in our base model, it is straightforward to augment $\bar{\boldsymbol{\lambda}}$ with additional sufficient statistics, e.g. the fixing or priority dependencies from [30, 35] that encode that one source fixes (i.e. should be given priority over) the other whenever both vote.

8 Conclusion

We proposed a new approach for end-to-end learning of neural network models for classification from, exclusively, multiple sources of weak supervision that streamlines prior latent variable models. We evaluated the proposed approach on benchmark datasets and observe that the downstream models outperform state-of-the-art data programming approaches in 4 out of 5 cases while remaining highly competitive on the remaining task, and outperforming several state-of-the-art crowdsourcing methods on a crowdsourcing task. We also demonstrated that our integrated approach can be more robust to dependencies between the labeling functions as well as to adversarial labeling scenarios. The proposed method works with discrete and probabilistic labeling functions and can utilize various neural network designs for probabilistic label generation. This end-to-end approach can simplify the process of developing effective machine learning models using weak supervision as the primary source of training signal, and help adoption of this form of learning in a wide range of practical applications.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Potential negative societal impacts are not specific to our new method, but to the overall studied area of (multi-source) weak supervision. We provide a brief discussion of this in Appendix G.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplementary code and instructions therein (README.md)
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Dataset details are in section 4 and see section C in the appendix for hyperparameter details, as well as appendix E and F for training & baseline setup details specific to the crowdsourcing and model robustness experiments, respectively.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For the main experiments/benchmarking seven random seeds were used, for all other experiments at least five random seeds were used (and the standard deviations across the seeds are always reported).
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See dataset paragraph in section 4, and README.md in the data- and LF- set link in the supplementary material (`https://drive.google.com/drive/u/2/folders/1v7IzA3Ab5zDEsRpLBWmJnXo5841tSOlh`)
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Yes, instructions to download the main datasets and LF sets (including the newly introduced sets) used are included in the source code README.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All data was public and no data agreements were needed.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We state that all datasets were publicly released, commonly used benchmark datasets, none of which contain sensitive personally identifiable information.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We used publicly released crowdsourcing benchmark data but did not conduct crowdsourcing exercises
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

[1] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321, 2016.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[3] Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, and Sunita Sarawagi. Learning from rules generalizing labeled exemplars. In *ICLR*, 2020.

[4] Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 273–282, 2017.

[5] Stephen H. Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alex Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Chris Ré, and Rob Malkin. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 362–375, New York, NY, USA, 2019. Association for Computing Machinery.

[6] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA, 1998. Association for Computing Machinery.

[7] Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. Interactive weak supervision: Learning useful heuristics for data labeling. *ICLR*, 2021.

[8] Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. Max-mig: an information theoretic approach for joint learning from crowds. *ICLR*, 2019.

[9] Oishik Chatterjee, Ganesh Ramakrishnan, and Sunita Sarawagi. Data programming using continuous and quality-guided labeling functions. *AAAI*, 2020.

[10] Mayee F. Chen, Benjamin Cohen-Wang, Steve Mussmann, Frederic Sala, and Christopher Ré. Comparing the value of labeled and unlabeled data in method-of-moments latent variable estimation. *AISTATS*, 2021.

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

[12] David Corney, Dyaa Albakour, Miguel Martinez-Alvarez, and Samir Moussa. What do a million news articles look like? *Workshop on Recent Trends in News Information Retrieval*, pages 42–47, 2016.

[13] Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pages 285–294, 2013.

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

[15] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

[16] Jared A Dunnmon, Alexander J Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew P Lungren, Daniel L Rubin, et al. Cross-modal data programming enables rapid medical machine learning. *Patterns*, 1(2):100019, 2020.

[17] Jason A Fries, Paroma Varma, Vincent S Chen, Ke Xiao, Heliodoro Tejeda, Priyanka Saha, Jared Dunnmon, Henry Chubb, Shiraz Maskatia, Madalina Fiterau, et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences. *Nature Communications*, 10(1):1–10, 2019.

[18] Daniel Y. Fu, Mayee F. Chen, Frederic Sala, Sarah Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. *ICML*, 2020.

[19] Garrett B Goh, Charles Siegel, Abhinav Vishnu, and Nathan Hodas. Using rule-based labels for weak supervised learning: a chemnet for transferable chemical property prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 302–310, 2018.

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

[21] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[22] Sonal Gupta and Christopher D Manning. Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 98–108, 2014.

[23] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, pages 507–517, 2016.

[24] Giannis Karamanolakis, Subhabrata (Subho) Mukherjee, Guoqing Zheng, and Ahmed H. Awadallah. Self-training with weak supervision. In *NAACL*, 2021.

[25] David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems*, volume 24, pages 1953–1961. Curran Associates, Inc., 2011.

[26] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *ICLR*, 2018.

[27] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[28] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[29] Alexander Ratner, Stephen Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29, 07 2019.

[30] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29, 05 2016.

[31] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:4763–4771, 07 2019.

[32] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322, 2010.

[33] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[34] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[35] Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. Dependency structure misspecification in multi-source weak supervision models. *ICLR Workshop on Weakly Supervised Learning*, 2021.

[36] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. *ICML*, 2021.

[37] P. Varma, F. Sala, A. He, A. Ratner, and C. Ré. Learning dependency structures for weak supervision models. *ICML*, 2019.

[38] Paroma Varma, Bryan D He, Payal Bajaj, Nishith Khandwala, Imon Banerjee, Daniel Rubin, and Christopher Ré. Inferring generative model structure with static analysis. In *Advances in neural information processing systems*, pages 240–250, 2017.

[39] Paroma Varma and Christopher Ré. Snuba: Automating weak supervision to label training data. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, page 223. NIH Public Access, 2018.

[40] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, volume 23, pages 2424–2432. Curran Associates, Inc., 2010.

[41] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *The Journal of Machine Learning Research*, 17(1):3537–3580, 2016.