CML: A Contrastive Meta Learning Method to Estimate Human Label Confidence Scores and Reduce Data Collection Cost

Bo Dong, Yiyi Wang, Hanbo Sun, Yunji Wang, Alireza Hashemi, Zheng Du
Amazon Alexa AI
{dongbd,yiyiwang,sunhanbo,yunjiwan,arhash,zhengdu}@amazon.com

Abstract

Deep neural network models are especially susceptible to noise in annotated labels. In the real world, annotated data typically contains noise caused by a variety of factors such as task difficulty, annotator experience, and annotator bias. Label quality is critical for label validation tasks; however, correcting for noise by collecting more data is often costly. In this paper, we propose a contrastive meta-learning framework (CML) to address the challenges introduced by noisy annotated data, specifically in the context of natural language processing. CML combines contrastive and meta learning to improve the quality of text feature representations. Meta-learning is also used to generate confidence scores to assess label quality. We demonstrate that a model built on CML-filtered data outperforms a model built on clean data. Furthermore, we perform experiments on de-identified commercial voice assistant datasets and demonstrate that our model outperforms several SOTA approaches.

1 Introduction

Deep neural networks' remarkable capacity for representation learning has resulted in performance gains across a wide range of applications. The majority of these gains are dependent on having high-fidelity data; however, in practice, large-scale datasets are frequently corrupted by noise caused by a variety of factors such as task difficulty, annotator experience, and annotator bias. This is a concern because training data with corrupted labels can adversely affect the performance of deep learning models. Collecting ground truth data to alleviate this problem, on the other hand, has proven both time consuming and costly.

We can broadly enhance model performance on corrupted labels by applying two approaches: improving model robustness and improving quality of feature representation. A number of existing methods such as robust loss function based methods (Wang et al., 2019a; Zhang and Sabuncu, 2018; Ghosh et al., 2017), loss adjustment based methods (Ren et al., 2018; Zheng et al., 2021; Shu et al., 2019), and sample selection based methods (Jiang et al., 2018; Malach and Shalev-Shwartz, 2018; Han et al., 2018), have been proposed to enhance model robustness. In spite of these advances, these methods are primarily concerned with computer vision and lack quality assessment for feature representation.

There are also several approaches to improving the quality of existing feature representations. They do not, however, specifically address the adverse effects of noisy (corrupted) labels in the context of natural language processing (NLP). (Chen et al., 2020; Ghosh and Lan, 2021) proposed a contrastive learning-based approach for improving feature representation quality for computer vision (not NLP) applications through augmentation steps in the feature learning process. (Gao et al., 2021) proposed a contrastive learning-based approach to improving the quality of sentence embedding. However, this approach does not address the problems caused by noise-corrupted labels.

In this paper, we propose a contrastive meta-learning (CML) framework for simultaneously addressing the challenges introduced by corrupted labels, in the context of NLP. Importantly, our framework learns a confidence score for evaluating the quality of annotations. In the real world, the work of data annotators is typically evaluated against ground truth data. The term "ground truth data" refers to validated data labels that have been subjected to multiple passes by data annotators and labelled using majority vote. (Namazifar et al., 2021). We will hereafter refer to a dataset consisting of ground truth labels as a "gold" dataset, contrasting it with a "standard" dataset annotated by a single data associate.

Collecting ground truth data ("gold" datasets) is time consuming and expensive, and sometimes in-

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 35 - 43

volves heavy engineering efforts (Sun et al., 2020). The confidence score generated by our model offers the potential to perform large-scale evaluations of annotation tasks. Another application of the confidence score generated by our model is to select high-quality data from a noisy dataset.

To address the issue of corrupted labels, we employ meta learning, which combines meta data and training data. Our meta data is comprised of a relatively small number of ground truth labels. The training data consists of unverified annotations (i.e. corrupted labels). Our model also learns an explicit loss-weight function while performing classification, which predicts label confidence scores.

We utilized a meta learning approach (Shu et al., 2019) as our meta module. The meta module is concatenated with the classifier. The meta module learns an explicit loss-weight function in a meta-learning manner. We use the output loss of the classifier as the input to the meta module. The meta module learns confidence scores using a multilayer perceptron on the output loss of the classifier, which reflects the quality of labels and can also improve classifier performance. Each of our training processes contains two steps. The first step consists of using our training data to update the parameters of the classifier, and the second step consists of using our meta data to update the parameters of the meta weight net. To improve the quality of feature representations, we apply contrastive learning to a pretrained BERT model (Gao et al., 2021). Contrastive learning aims to learn effective representations by pulling semantically close neighbors together and pushing apart non-neighbors. In order to have semantically close neighbors, the same sentence pass the pretrained encoder twice to predict the sentence itself with noise introduced by standard dropout layer. In such way, we have two embeddings generated from the same sentence but with slight difference. These two embeddings are "positive pair".

As we mentioned above, current methods (Zheng et al., 2021; Shu et al., 2019) designed to address corrupted labels are mainly geared towards computer vision applications. In addition, their motivation is largely centered around label correction or improving the performance of a classifier. In contrast, in the NLP domain, researchers mainly focus on improving the quality of embeddings or learning representations of text data. Our proposed framework unifies the advantages of current methods and is suitable for NLP. We demonstrate that CML not only addresses concerns stemming from having corrupted labels, but also learns feature representations from raw text data effectively.

Additionally, in real-world applications, we are concerned with the quality of annotations. As such, evaluating the work of annotators presents an important challenge. During the training process, CML learns a confidence score for labels, which can be used to evaluate annotators' work online. In other words, CML offers the potential for scaling the work of data annotators. Furthermore, because CML predicts a confidence score for labels, it can be applied to a wide range of use cases. In many instances only incorrectly labelled data is available and collecting ground truth data is time-consuming and costly. We can therefore use CML to filter high-quality data for such problems, saving both time and money.

To summarize, the main contributions of our work are listed as follows:

- CML combines meta learning and contrastive learning to address the corrupted label issue and feature representation quality issue in tandem.
- CML predicts confidence score for annotated labels which solves the problem of evaluating annotators' work at scale.
- CML can be applied to filter high quality data from raw annotations, which proves to be of the same level of quality as the ground truth data. It reduces costs associated with collecting massive amounts of ground truth data for downstream model development.

The remainder of this paper is organized as follows: We introduce related work in section 2. We then formalize the problem and present our proposed approach in section 3. Next, we present applications and corresponding results of our experiments in section 4 and 5. Finally, we present our conclusion and propose future research directions.

2 Related Work

In this section, we discuss some of the related work in contrastive learning and learning from corrupted labels.

2.1 Learning from Corrupted Labels

Machine learning techniques (Liu et al., 2021, 2019; Dong et al., 2017, 2018; Wang et al., 2020,

2019b; Li et al., 2021; Dong et al., 2019) have been widely applied on labeling tasks. With respect to learning from corrupted labels, a variety of methods have been proposed. Namely, (Zheng et al., 2021) proposes a meta-learning framework for re-weighting and correcting corrupted labels. This method requires a clean dataset (without corrupted labels) alongside a dataset with corrupted labels. The focus of above paper is label correction. It provides an approach to predict a set of weights in label space for each instance. We cannot get a single weight score for each instance directly by using this method. (Li et al., 2017) proposes a distillation framework which uses metadata, including a small clean dataset and label relations in a knowledge graph to learn from corrupted labels. However, in the real world setting, it is difficult to collect sufficient and useful metadata. (Dong et al., 2020) proposes a new loss function which includes an importance weight for training instance. This importance embedding serves the function of finding important training instances. The importance embedding is trained during model training. However, this method is not originally designed for corrupted labels and can not make use of ground truth data and corrupted labels in tandem, in the training process. (Shu et al., 2019) proposes a meta learning method to learn weight score to evaluate label quality. Their approach focuses on corrupted labels and addresses class imbalance issues. It also learns an explicit loss-weight function, parameterized through a multi-layer perceptron during meta-learning.

2.2 Learning Feature Representation

Feature representation quality is a critical factor affecting deep neural network performance. One research direction is contrastive learning. (Chen et al., 2020) proposes a contrastive learning framework which can improve feature representation via contrastive loss with augmented data for computer vision applications. (Ghosh and Lan, 2021) demonstrates that initializing supervised robust methods using representations learned through a contrastive learning framework leads to significantly improved performance with noisy labels. (Kim et al., 2021) proposes a contrastive learning method that uses self-guidance to fine tune BERT, which does not rely on sentence augmentation. (Fang et al., 2020) also fine tune a pretrained language encoder like BERT. This approach uses back-translation as augmentation of input sentence. (van den Oord et al., 2019) designs a contrastive predictive coding method to extract representations from high-dimensional data in a universal unsupervised manner.

3 Approach

3.1 Problem Setting

In this paper, we propose a contrastive meta learning framework(CML). Basically, in learning with corrupted labels, we assume that a small set of data with clean labels and a large set of data with noisy (corrupted) labels are needed (Zheng et al., 2021). Usually due to scarcity and high cost of generating ground truth labels, the relative size of the clean dataset is much smaller than the noisy one. Since a small training set tends to cause overfitting, utilizing a clean dataset alone may lead to creating a model that does not generalize well. On the other hand, training with noisy data is also not a very desirable option since large high-capacity models will fit and memorize the noise (Zhang et al., 2017). Therefore, an effective way to overcome the aforementioned challenges is to build a framework which utilizes both noisy/corrupted data and clean data. Our framework consists of two modules: the main module and meta module. The main module learns feature representations in a contrastive manner and builds a predictive model. At the same time, we also learn a meta module which is a loss weight function. The meta module tries to learn confidence scores for corresponding labels. Our framework allows the main module and meta module to learn from each other.

3.2 Framework

The CML framework (Figure 1) consists of two modules: the main module and the meta module. The main module adopts pre-trained BERT-base in a contrastive manner followed by a dropout layer, a hidden layer, and several fully connected layers to map the input data into a semantic representation. The last layer in the main module is a linear output layer. The meta module is an MLP(multilayer perceptron) network with only one hidden layer. The activation function for all hidden layers is ReLU. The meta module utilizes a small set of clean data to guide the training of all of its parameters.

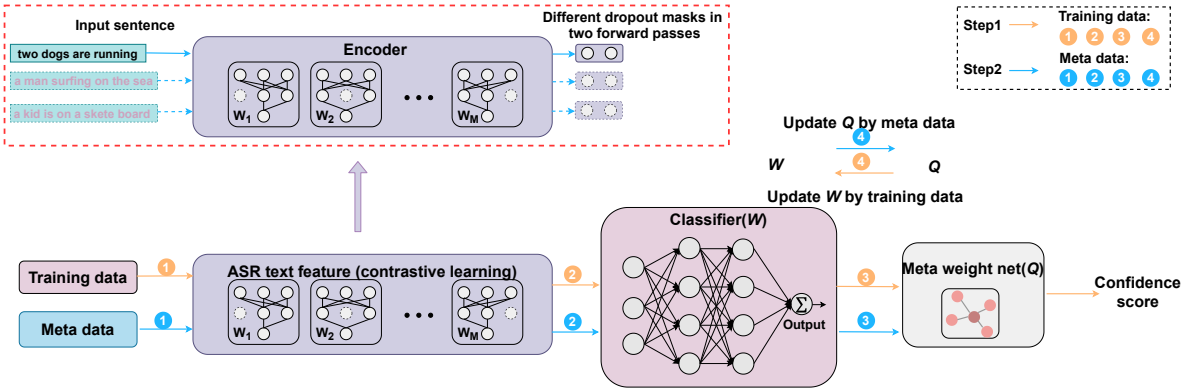


Figure 1: CML framework: one iteration consists of two steps. Both start with contrastive learning that is a pretrained dropout masked BERT, followed by fully connected layers and meta weight net. In the first step, we keep the meta weight net unchanged and only update weight of fine tuned layers. The second step is to feed meta data (ground truth) to update meta weight net with main model's predicted probability as input to meta weight net. In addition, meta weight net learns an explicit loss-weight function to predict label confidence. We demonstrate an example in the figure where contrastive learning takes automatic speech recognition(ASR text) to predict if the recognition is capable to represent the speaker's goal.

3.2.1 Main Module

We fine-tune simCSE framework (Gao et al., 2021) for learning the text feature representation in a contrastive manner. A commonly used contrastive learning setting is as follows, assume we have a collection of paired sentences $S = (\mathbf{x}_i, \mathbf{x}_i^+)$, where $\mathbf{x}_i$ and $\mathbf{x}_i^+$ are semantically related. Then we can use a base encoder $\mathcal{F}(\cdot)$(pre-trained BERT-base) to encode each sentence $\mathbf{x}_i$ as follows

$$\mathbf{e}_i = \mathcal{F}(\mathbf{x}_i) \tag{1}$$

Let $\mathbf{e}_i$ and $\mathbf{e}_i^+$ represent the feature representation of $\mathbf{x}_i$ and $\mathbf{x}_i^+$. The contrastive learning loss function is designed as:

$$l = \sum_{i=1}^{N} log \frac{exp(sim(\mathbf{e}_i, \mathbf{e}_i^+)/\tau)}{\sum_{j=1}^{N} exp(sim(\mathbf{e}_i, \mathbf{e}_j^+)/\tau)} \tag{2}$$

where τ is the temperature parameter and sim is the cosine similarity $\frac{\mathbf{e}_i^\top \mathbf{e}_i^+}{\|\mathbf{e}_i\| \cdot \|\mathbf{e}_i^+\|}$.

In above setting, simCSE let $\mathbf{x}_i^+ = \mathbf{x}_i$. Then use $\mathbf{e}_i^m = \mathcal{F}(\mathbf{x}_i^m)$ to represent the feature representation of $\mathbf{x_i}$ with random mask m for dropout. The same sentence pass to the encoder twice with different dropout masks m, n. The loss function is defined as follows:

$$l = \sum_{i=1}^{N} log \frac{exp(sim(\mathbf{e}_i^{m_i}, \mathbf{e}_i^{n_i})/\tau)}{\sum_{j=1}^{N} exp(sim(\mathbf{e}_i^{m_i}, \mathbf{e}_j^{n_j})/\tau)} \tag{3}$$

We utilize pre-trained simCSE to learn representations of the input sentence in our main module for predicting labels.

3.2.2 Meta Module

Inspired by (Shu et al., 2019), we incorporate a loss-weight function(a multilayer perceptron) into our meta module. This module learns confidence scores which can be used to evaluate the quality of labels. When an input sentence passes the main module, we have a loss computed using the predicted label and the original label. In the meta module, we utilize a small set of clean data to learn a confidence score from the output of the main module. We initialize the parameters $\mathbf{w}$ of the main module and parameters θ of the meta module. In general, our framework is an iterative procedure. For each iteration, it mainly contains two steps. The first step is to update the parameters $\mathbf{w}$ of the main module as equation 4 indicated by feeding biased training data.

$$\hat{\mathbf{w}}^{t+1}(\theta) = \mathbf{w}^t - \alpha \frac{1}{n} \times \sum_{i=1}^{n} G(L_i^{train}(\mathbf{w}^t), \theta^{t+1}) \nabla_{\mathbf{w}} \tag{4}$$

where $\nabla_{\mathbf{w}}$ is computed as follows

$$\nabla_{\mathbf{w}} = \frac{\partial L_i^{train}(\mathbf{w})}{\partial \mathbf{w}} |_{\mathbf{w}=\mathbf{w}^t} \tag{5}$$

The second step is to pass the clean data to update the parameters θ of meta module as equation 6 indicated.

$$\theta^{t+1} = \theta^t - \beta \frac{1}{m} \sum_{i=1}^{m} \frac{\partial L_i^{meta}(\hat{\mathbf{w}}^t(\theta))}{\partial \theta} |_{\theta=\theta^t} \tag{6}$$

After the learning process, we can predict confidence scores for annotated labels by inference from our trained model.

Algorithm 1 CML Learning Algorithm

Input: Biased training data S with batch size m, Unbiased meta data $\hat{S}$ with batch size n, max iterations I

Parameter: Main module parameters $\mathbf{w}$ and meta module parameters θ

Output: Main module parameters $\mathbf{w}$ and meta module parameters θ

Let $t = 0$.
while t in range $[0, I)$ **do**
  $(x, y) \leftarrow$ Sample Mini Batch (S,m)
  $(x^{meta}, y^{meta}) \leftarrow$ Sample Mini Batch ($\hat{S}$,n)
  if Data comes from S **then**
    Fine-tune main module
    update main module parameters $\mathbf{w}$ with equation 4
  else
    update meta module parameters θ with equation 6
  end if
  $t = t + 1$
end while
return $\mathbf{w}$, θ

4 Application

The confidence score obtained from CML's output has an important application for data labeling services: measuring label quality at scale. In the industrial setting, the quality of each annotator's work is measured by ground truth reference, which is usually of limited quantity. Small volumes of gold reference data could cause high variance in assessing annotator's performance. As such, it often requires complex procedures to find root cause of quality issue. Error detection model is broadly used in industry but remains a challenge due to limited ground truth labels.

The confidence score from CML is a promising attempt to solve the aforementioned challenges. In particular we implement the following two applications:

4.1 Application 1

Use confidence scores to generate a quality metric for each label, and shows that these scores manage to distinguish the labels in different levels of quality.

4.2 Application 2

Use the data filtered by the confidence score to build an error detection model and demonstrate that it will produce better results.

Figure	Statistics	P value
Figure 2 (a)	0.834	0.000
Figure 2 (b)	0.752	0.000

Table 1: Kolmogorov–Smirnov test results

5 Evaluation

This section shows the experiment results. We also compare our method to state-of-the-art methods.

5.1 Dataset

In our experiment, we process and de-identify commercial voice assistant dataset that assess goal success rate(GSR). We evaluate the goal category task, i.e. labeling a given utterance to a fixed taxonomy of categories. Text ASR (automatic speech recognition) is used as the input feature. Data is collected from both the standard and gold data. Because the standard data only performs one pass on each task, it contains some corrupted labels. Data collected from the gold data can be conceived as "ground truth" data. It will be the data source from which we will generate synthetic data.

5.2 Application 1

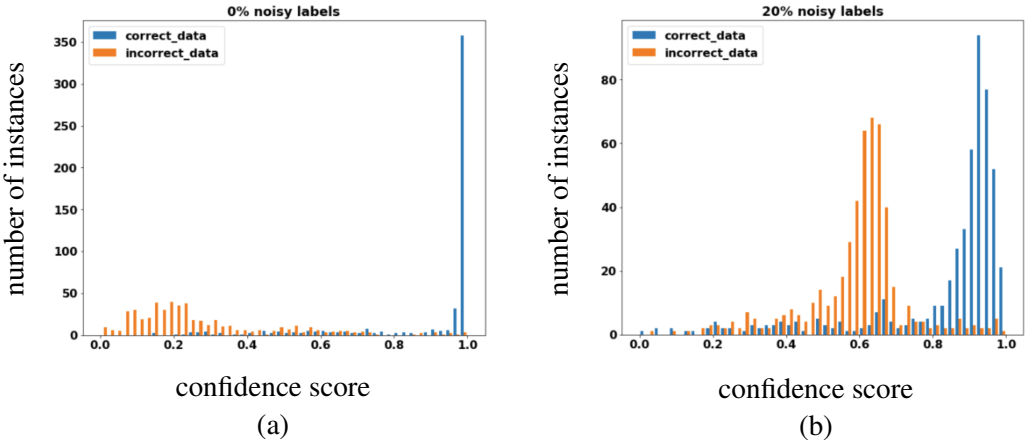
In this experiment, we use synthetic data to show that the confidence score produced by CML is capable to separate the incorrect label from correct label at various noise level.

5.2.1 Synthetic data generation process

We generate synthetic data by flipping labels of gold data with different noise ratios for the training set with corrupted labels. Ambiguous labels are generated by flipping label based on assuming that the gold dataset is "correct". The level of noise is also varied between 0% and 20%. We generate synthetic training data by flipping X% labels to incorrect labels. When the flipping rate is 0, the training data are all ground truth. For the test set, we synthetically flip 50% data to incorrect labels.

5.2.2 Metrics and graph explanations

Figure 2 illustrates the confidence scores for the test dataset. The left figure represents the confidence score learned by CML model with training data containing 0% corrupted labels. The right figure represents the confidence score result learned by CML with training data containing 20% corrupted labels. In Table 1, a Kolmogorov–Smirnov test (Massey, 1951) shows the confidence scores from



(a) (b)

Figure 2: Confidence score distribution on test set learned from training set with 0%, 20% corrupted labels experiments

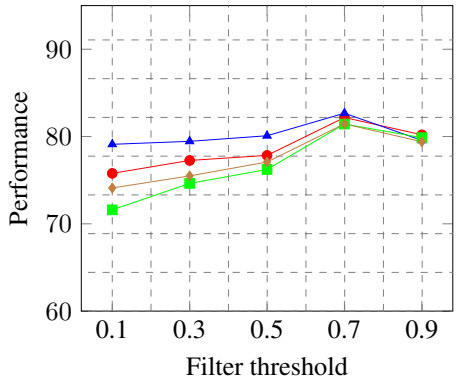


Figure 3: Parameter sensitivity on training data set contains 30% noise (Accuracy Precision Recall F1 score)

the correct labels and the incorrect labels are from different distributions in both scenarios, with the test statistic showing a more extreme value in the 0% noisy label case. From Figure 2 we can see, our learned confidence scores can differentiate correct labels and incorrect labels clearly. In addition, the noise ratio of training set negatively correlate with the level of difference of the confidence scores.

5.3 Application 2

In this experiment, we use both synthetic data and commercial voice assistant data to show that the data filtered by the confidence score works better in error prediction than the same model trained with either the raw training data or the clean data alone.

5.3.1 Synthetic data generation

For the training set, we generate synthetic data similarly as in application 1 by flipping labels of gold data with different noise ratios. We generate training data with 10%, 20%, 30%, 40% and 50% noisy labels. For test set, we synthetically flip 50% labels to incorrect labels.

5.3.2 Experiment setup

The experiment is conducted with the following steps: 1) In a classification task, we train CML to learn the confidence score from the training set, i.e. to predict the correct goal category. 2) We set a threshold to filter good quality data based on the learned confidence score (the threshold is treated as a hyperparameter). 3) Using the filtered training data set, we train a separate BERT-based error detection model. 4) We train the same error detection model using only biased training data. 5) We run the above two models on the same test set and compare their performance. For model evaluation, commonly used metrics such as accuracy, precision, recall, and F1 score are used.

5.3.3 Explain results

Table 2 shows the model performance. The result obtained after CML filtered data outperforms the result obtained using biased training data for all metrics. In addition, based on the last two columns, with the data filtering setup, we select most correct labeled instances and very few incorrect instances. Figure 3 indicates the sensitivity experiment result of threshold for filtering. Figure 3 illustrates that threshold is sensitive for all evaluation metrics.

5.3.4 Real data

We utilize data collected from standard dataset and gold dataset for evaluation. Data collected from standard dataset contains corrupted labels. We design two sets of experiments.

Training set	Accuracy	Precision	Recall	F1 score	#Correct	#Incorrect
Biased training data with 50% noise	40.04	57.54	40.51	43.53	5000	5000
CML Filtered data	**77.38**	**79.57**	**74.81**	**76.33**	4664	902
Biased training data with 40% noise	68.60	73.27	65.01	67.13	6000	4000
CML Filtered data	**79.76**	**81.13**	**77.50**	**77.98**	5590	714
Biased training data with 30% noise	77.36	78.33	72.09	74.51	7000	3000
CML Filtered data	**82.18**	**82.67**	**81.43**	**81.45**	6509	537
Biased training data with 20% noise	80.64	82.61	78.68	80.01	8000	2000
CML Filtered data	**82.98**	**83.30**	**82.26**	**82.17**	7432	350
Biased training data with 10% noise	83.70	**83.55**	82.82	**83.10**	9000	1000
CML Filtered data	**83.72**	83.14	**83.32**	82.77	8353	162

Table 2: Experiment result for synthetic data: compare model built on full data with model built on selected data that are filtered by CML. The number of correct and incorrect in the table stand for the volume of examples with correct and incorrect labels respectively. For instance, 50% noise data contain 5000 correct labeled examples and 5000 incorrect labeled examples. Filtered by CML, we obtain 4664 correct labeled and 902 incorrect labeled examples respectively.

5.3.5 Experiment setup

1) We sample data from the gold dataset (gold-1 of size 3k, and gold-2 of size 6k) to be the unbiased meta data, and sample data from standard dataset (of size 100k) to be the noisy training data. We utilize both of these datasets to train the CML model. We filter the noisy training data by the confidence score learned from CML model, and then build two error detection models with gold-1 data and the filtered data separately. At last we compare the performance on a hold-out gold data set of size 5k. This hold-out data set is used for all the evaluation cases. In a variant of this experiment, we replicate the same process for gold-2. 2) This set of experiment is designed to verify that CML achieves better performance by ingesting small proportion of gold data, compared to model trained on noisy data alone. We sample data from gold dataset of size 3k and sample standard(noisy) data of size 20k.

5.3.6 Experiment results

Comparing row 1 and 3 of Table 3, we demonstrate that the model trained with filtered standard data outperforms the model trained with gold data. This is achieved when the size of noisy training data is 30 folds larger than that of gold data. Comparing rowd 2 and 3, even though the size of the gold data is doubled, the model's performance is still worse than the model trained with filtered standard data. Comparing rowd 4 and 5, we demonstrate that using CML with a noisy training set and small meta data outperforms using noisy training set alone.

In the commercial setting, we hold a large amount of data with corrupted labels. Collecting ground truth data is time consuming and expensive. Based on the above experiment results, CML and its applications provide a economic way to building label error detection model.

Training set	Accuracy	Precision	Recall	F1 score
gold1(3k)	79.42	78.78	78.64	78.70
gold2(6k)	80.72	80.02	**79.93**	79.97
Filtered data	**81.28**	**81.66**	79.44	**80.53**
Biased data	80.20	80.71	77.67	78.02
Biased+meta data	**81.82**	**81.16**	**81.82**	**81.25**

Table 3: Experiment result for real data.
Filtered data: filter from the biased training data(100k).

5.4 Model Evaluation

5.4.1 Data set

For this set of experiment we want to verify the performance of CML. We sample 3k examples from gold dataset as meta data. We also sample 20k examples from standard dataset as noisy training data.

5.4.2 Experiment setting

We evaluate our proposed approach CML against state-of-the-art methods(Shu et al., 2019; Han et al., 2018) for learning with noisy labels. As we mentioned in Section 1, current state-of-the-art methods mainly focus on computer vision domain. We revise the architecture of these two baselines by using pre-trained BERT-base as their main classifiers. We also compare our approach with contrastive learning benchmarks (Gao et al., 2021).

Method	Accuracy	Precision	Recall	F1 score
MetaNet	82.54	82.25	80.79	81.51
simCSE	79.66	80.31	79.65	77.34
Co-teaching	77.00	69.78	76.74	72.73
CML(ours)	**82.82**	**82.86**	**82.58**	**82.71**

Table 4: Experiment result for CML and baselines.

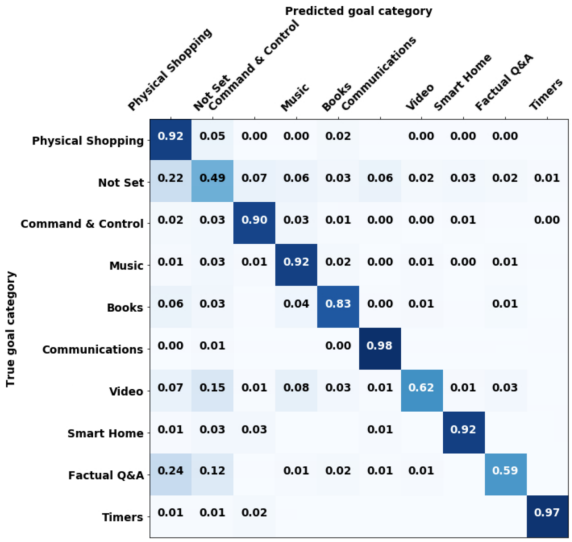


Figure 4: Confusion matrix(goal category prediction with CML)

5.4.3 Experiment results

From Table 4 we can see that our approach outperforms the other three baselines. We not only improve the quality of feature representation, but also improve model performance under noisy label scenario. We also render a confusion matrix for our method as illustrated in Figure 4. As we can see, the "Not Set" category does not perform good since annotators would choose "Not set" category when the category is ambiguous and they are not sure the correct answer. For another example, for "Timers" category, both recall and precision are very high as a result of less ambiguity compared to other categories.

6 Conclusion

In this paper, we propose a contrastive meta learning framework (CML) for estimating human label confidence scores and lowering data collection costs. We use contrastive learning and meta learning to jointly address the main challenges of label scarcity and poor feature representation. We design three sets of experiments with two application settings and three state-of-the-art baseline models to test the effectiveness of our proposed method. Our experiments on a commercial voice assistant GSR dataset show that our method can predict a reliable confidence score for annotations while also effectively lowering the cost of ground truth data collection. Moreover, our proposed method outperforms several SOTA approaches.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Bo Dong, Jinghui Guo, Zhuoyi Wang, Rong Wu, Yang Gao, and Latifur Khan. 2019. Regression prediction for geolocation aware through relative density ratio estimation. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1644–1649.

Bo Dong, Md Shihabul Islam, Swarup Chandra, Latifur Khan, and Bhavani M. Thuraisingham. 2018. GCI: A transfer learning approach for detecting cheats of computer game. In *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*, pages 1188–1197.

Bo Dong, Yifan Li, Yang Gao, Ahsanul Haque, Latifur Khan, and Mohammad M. Masud. 2017. Multistream regression with asynchronous concept drift detection. In *2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017*, pages 596–605.

Bo Dong, Cristian Lumezanu, Yuncong Chen, Dongjin Song, Takehiko Mizoguchi, Haifeng Chen, and Latifur Khan. 2020. At the speed of sound: Efficient audio scene classification. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 301–305, New York, NY, USA. Association for Computing Machinery.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1919–1925. AAAI Press.

Aritra Ghosh and Andrew Lan. 2021. Contrastive learning improves model robustness under label noise.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR.

Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2021. Self-guided contrastive learning for bert sentence representations.

Yi-Fan Li, Bo Dong, Latifur Khan, Bhavani Thuraisingham, Patrick T. Brandt, and Vito J. D'Orazio. 2021. Data-driven time series forecasting for social studies using spatio-temporal graph neural networks. In *Proceedings of the Conference on Information Technology for Social Good*, GoodIT '21, page 61–66, New York, NY, USA. Association for Computing Machinery.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. 2017. Learning from noisy labels with distillation.

Jie Liu, Wenqian Dong, Qingqing Zhou, and Dong Li. 2021. Fauce: fast and accurate deep ensembles with uncertainty for cardinality estimation. *Proceedings of the VLDB Endowment*, 14(11):1950–1963.

Jie Liu, Jiawen Liu, Wan Du, and Dong Li. 2019. Performance analysis and characterization of training deep learning models on mobile device. In *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 506–515. IEEE.

Jie Liu, Jiawen Liu, Zhen Xie, and Dong Li. Flame: Aself-adaptive auto-labeling system for heterogeneous mobile processors.

Eran Malach and Shai Shalev-Shwartz. 2018. Decoupling "when to update" from "how to update".

F. J. Massey. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

Mahdi Namazifar, John Malik, Li Erran Li, Gokhan Tur, and Dilek Hakkani Tür. 2021. Correcting automated and manual speech transcription errors using warped language models.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4334–4343. PMLR.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting.

David Q. Sun, Hadas Kotek, Christopher Klein, Mayank Gupta, William Li, and Jason D. Williams. 2020. Improving human-labeled data through dynamic automatic conflict resolution.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019a. Symmetric cross entropy for robust learning with noisy labels.

Zhuoyi Wang, Bo Dong, Yu Lin, Yigong Wang, Md Shihabul Islam, and Latifur Khan. 2019b. Co-representation learning framework for the open-set data classification. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 239–244.

Zhuoyi Wang, Yigong Wang, Bo Dong, Sahoo Pracheta, Kevin Hamlen, and Latifur Khan. 2020. Adaptive margin based deep adversarial metric learning. In *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 100–108.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization.

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels.

Guoqing Zheng, Ahmed H. Awadallah, and Susan Dumais. 2021. Meta label correction for noisy label learning. In *AAAI 2021*.