arXiv:2110.08355v1 [cs.CL] 15 Oct 2021

LEARNING WITH NOISY LABELS BY TARGETED RELABELING

Derek Chen
ASAPP Inc.
New York, NY
dchen@asapp.com

Zhou Yu
Dept. of Computer Science
Columbia University, NY
zy2461@columbia.edu

Samuel R. Bowman
New York University, NY
bowman@nyu.edu

ABSTRACT

Crowdsourcing platforms are often used to collect datasets for training deep neural networks, despite higher levels of inaccurate labeling compared to expert labeling. There are two common strategies to manage the impact of this noise, the first involves aggregating redundant annotations, but comes at the expense of labeling substantially fewer examples. Secondly, prior works have also considered using the entire annotation budget to label as many examples as possible and subsequently apply denoising algorithms to implicitly clean up the dataset. We propose an approach which instead reserves a fraction of annotations to *explicitly* relabel highly probable labeling errors. In particular, we allocate a large portion of the labeling budget to form an initial dataset used to train a model. This model is then used to identify specific examples that appear most likely to be incorrect, which we spend the remaining budget to relabel. Experiments across three model variations and four natural language processing tasks show our approach outperforms both label aggregation and advanced denoising methods designed to handle noisy labels when allocated the same annotation budget.

1 INTRODUCTION

Modern machine learning often depends on heavy data annotation efforts. To keep costs in check while maintaining speed and scalability, most people turn to non-specialist crowd-workers through platforms like Mechanical Turk. Although crowdsourcing lowers costs to a reasonable level, it also tends to produce substantially higher error rates compared with expert labeling. The classic approach for improving reliability is to perform redundant annotations which are later aggregated using a majority vote to form a single gold label (e.g. Snow et al. (2008); Sap et al. (2019); Potts et al. (2021)). This simple solution is easy to understand and easy to implement, but comes at the expense of severely reducing the number of examples with an associated label. Additionally, majority voting can cause biased results when the task is difficult since assumptions about mutual independence in label noise become inaccurate (Cao et al., 2019).

As a result, researchers have made great strides in designing automatic label cleaning methods, noise-insensitive training schemes and other denoising mechanisms (Sukhbaatar et al., 2015; Han et al., 2018; Tanaka et al., 2018). For example, some methods learn a noise transition matrix for reweighting the label (Dawid & Skene, 1979; Goldberger & Ben-Reuven, 2017), while others modify the loss (Ghosh et al., 2017; Patrini et al., 2017). Another set of options generate cleaned examples from noisy ones through semi-supervised pseudo-labeling (Jiang et al., 2018; Li et al., 2020). However, empirically getting many of these techniques to well in practice is often a struggle.

To avoid this complexity of repairing or reweighting the labels of existing annotations, we sidestep the problem by simply obtaining new annotations for a small selected subset of examples. In doing so, we are able to take advantage of high quantity similar to denoising methods, while still maintaining the high quality benefits of label aggregation. Concretely, we start by allocating a large portion of the labeling budget to obtain an initial training dataset. The examples in this dataset were annotated in a single pass, so we would expect a small percentage of them to be incorrectly labeled. However, enough of the labels should be correct to train a reasonably performing model. Next, we consider several methods that take advantage of the newly trained model to identify the incorrectly
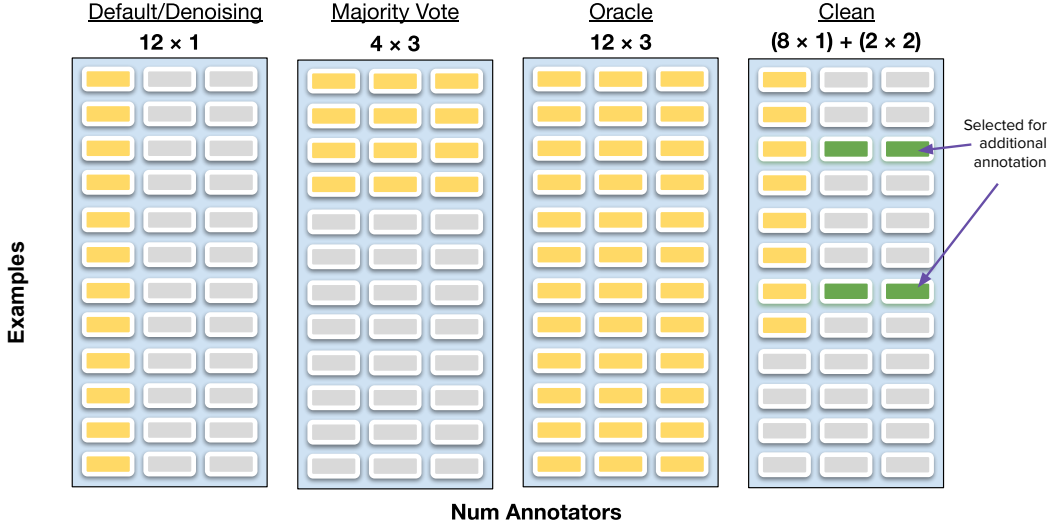
Figure 1: Data cleaning reserves a small portion of the annotation budget for targeted relabeling of examples that are identified as especially likely to be noisy. In contrast, the default and denoising methods spend all the entire budget upfront, yielding lower quality data.

labeled examples. We then spend the remaining budget to relabel those examples. Finally, we train a new model with the original data combined with the explicitly cleaned data.

A key ingredient of our method is a function for selecting which examples to annotate. We consider multiple approaches for identifying candidates for relabeling, including how the examples performed during training, which examples the model gets wrong and the hidden representations produced by the model. In all cases, relabeling the target examples relies on neither training any extra model components nor on tuning sensitive hyper-parameters. Instead, the boost in learning signal comes directly from human supervision. By using the existing annotation pipeline, the implementation becomes relatively trivial.

To test the generalizability of our method, we compare against multiple baselines on four tasks spanning multiple natural language formats. The control baseline and denoising baselines perform a single annotation per example. The majority vote baseline triples the annotations per example, but consequently is comprised of only one third the number of examples to meet the annotation budget. We lastly include an Oracle baseline that lifts the restriction on a fixed budget and instead uses all available annotations. We additionally test across a three variety of different model types, ranging from small neural models taking minutes to train up to large state-of-the-art transformer models. We find that under the same fixed annotation budget, cleaning methods either match or surpass all baselines we compared against.

In summary, our contributions include:

1. We propose an alternative direction for learning with noisy labels through explicit cleaning

2. We build four versions of our approach that vary in how they target examples to relabel

3. We compare against a number of baselines, many of which have never been implemented before in the language setting.

4. We conduct analysis to understand the strengths and weaknesses of our proposal

Overall, the Large Loss method, which selects examples for relabeling by the size of its training loss, performs the best out of all variations we considered. We plan to release all code and data.[1]

---

[1] https://github.com/asappresearch/clean

2 RELATED WORK

The standard method for learning in the presence of unreliable annotation is to perform redundant annotation, where each example is annotated multiple times and a simple majority vote determines the final label (Snow et al., 2004; Russakovsky et al., 2015; Bowman et al., 2015). While effective, this can be costly since it multiplicatively reduces the amount of data collected. To tackle this problem, previous researchers have developed several alternative methods.

Denoising Techniques Noisy training examples can be thought of as the result of perturbing the true, underlying labels by some source of noise. One class of methods assume that the noise is derived from label corruption where one *label class* is confused for another (Hendrycks et al., 2018). A noise transition matrix is often learned to keep track of which labels get confused with others, and then used during inference to reverse the effect (Sukhbaatar et al., 2015; Goldberger & Ben-Reuven, 2017). Other methods work under the assumption that labeling errors occur due to *annotator* biases (Raykar et al., 2009; Rodrigues & Pereira, 2018). Annotators may lack relevant expertise (Welinder et al., 2010; Guan et al., 2018) or engage in spamming behavior (Hovy et al., 2013; Khetan et al., 2018), leading to incorrect labels. Finally, some methods model the noise of each individual *example* since the assumption is that each example's label is generated independently of other examples. Traditional methods of this sort leverage expectation-maximization to iteratively reweight each example (Dawid & Skene, 1979; Whitehill et al., 2009; Mnih & Hinton, 2012), while more modern methods take advantage of deep neural networks to model the noise instead (Felt et al., 2016; Jindal et al., 2019).

Loss Modification Another way to deal with noisy annotations is to design an alternate training scheme (Patrini et al., 2017). Some methods add a regularization term to make the model more robust to noise (Tanno et al., 2019). One can also change the loss function by, for example, bounding the amount of loss contributed by individual training examples, thereby limiting the negative impact of mislabeled items (Ghosh et al., 2017; Zhang & Sabuncu, 2018). The learning procedure can also be modified such that the importance of training examples are dynamically reweighted to prevent overfitting to noise (Jiang et al., 2018).

Pseudo-Labeling Rather than change the model, pseudo-labeling methods correct the data themselves by devising new labels for existing data or by generating wholly new training examples. Semi-supervised methods create clean labels by bootstrapping off of existing labels (Reed et al., 2015; Tanaka et al., 2018). More sophisticated techniques perform data augmentation to formulate new examples (Arazo et al., 2019; Li et al., 2020). Still other approaches use two distinct networks to produce examples for each other to learn from (Han et al., 2018; Yu et al., 2019). Some methods are similar to ours in that they identify certain labels to clean (Lee et al., 2018). However, they use machine learning to implicitly fix the examples, whereas we rely on human experts to explicitly relabel the data. We compare against a pseudo-label method, but not Lee et al. (2018) in particular since their method requires training a large model to convergence multiple times for each run.[2]

Our work also falls under research studying how to maximize the benefit of labeled data given a fixed annotation budget. Khetan & Oh (2016) find that model-based EM can be quite powerful in modeling annotator noise, allowing singly-labeled data to outperform multiply-labeled data when annotation quality goes above a certain threshold. Bai et al. (2021) show that similar trade-offs exist when performing domain adaptation on a constrained budget. Finally, Zhang et al. (2021) observe that difficult examples benefit from additional annotations, so optimal spending of a finite budget actually varies the amount of labels given to each example. We differ from their work since our approach actively targets examples for relabeling based on its likelihood of noise, whereas they randomly select examples for multi-labeling without considering its characteristics.

3 PROPOSED METHODS

We study how to maximize the model performance given a finite data annotation budget. Concretely, we are given some model M for a given task, along with a budget as measured by B number of

---

[2]This would amount to roughly four months of continuous training given our experimental setup.

annotations, where each annotation allows us to apply a possibly noisy labeling function $f_r(x)$, where r is the number of redundant annotations applied to a single example. Annotating some set of unlabeled instances produces noisy examples $(X, f_r(X)) = (X, \tilde{Y})$. Our goal is to achieve the best score possible for some primary evaluation metric S on a given task by cleaning the noisy labels $\tilde{Y} \xrightarrow{clean} Y$. Afterwards, we train a model with the cleaned data and then test it on a separate test set. For all our experiments, we set $B = 12,000$ as the total annotation budget.

As a default setting, we start with a *Control* baseline which uses the entire budget to annotate 12K examples, once each $(n = 12,000; r = 1)$. To simulate a single annotation, we randomly sample a label from the set of labels offered for each example by the dataset. To obtain more accurate labels, people will often perform multiple annotations on each example and use *Majority Vote* to aggregate the annotations. Accordingly, as a second baseline we annotate 4k examples three times each $(n = 4,000; r = 3)$, matching the same total budget as before. In the event of a tie, we randomly select one of the candidate labels. Finally, we also include an *Oracle* baseline which uses the gold label for 12k examples $(n = 12,000; r = 3|5)$. The gold label is either given by the dataset or generated by majority vote, where the label might result from aggregating five annotations rather than just three annotations.

3.1 Noise Correction Baselines

We consider four advanced baselines that each approach the problem of learning with noisy labels in a slightly different manner. All of them perform a single annotation per example $(n = 12,000, r = 1)$ as seen in Figure 1. (1) Goldberger & Ben-Reuven (2017) propose applying a noise *Adaptation* layer which models the error probability of label classes. This noise transition matrix is learned as a non-linear layer on top of the baseline model M to denoise predictions. The layer is then discarded during final inference since gold labels are used during test time and are assumed to no longer be noisy. (2) The *Crowdlayer* also operates by modeling the error probability, but assumes the noise arises due to annotator error, so a noise transition matrix is created for each worker (Rodrigues & Pereira, 2018). Once again, this matrix is learned with gradient descent and removed for final inference. (3) The *Forward* correction method from Patrini et al. (2017) adopts a loss correction approach which modifies the training objective. Given $-\log p(\hat{y} = \tilde{y}|x)$ as the original loss, Forward modifies this to become $-log \sum_{j=1}^{c} T_{ji} p(\hat{y} = y|x)$ where c is the number of classes being predicted, and both i and j are used to index the number of classes. Matrix T represented as a neural network that is a learned jointly during pre-training. (4) Lastly, the *Bootstrap* method proposed by Reed et al. (2015) generates pseudo-labels by gradually interpolating the predicted label $\hat{y}$ with the given noisy label $\tilde{y}$. We apply their recommended *hard* bootstrap variant which uses the one-hot prediction for interpolation.

3.2 Cleaning through Targeted Relabeling

Rather than maximizing the number of examples annotated given our budget, we propose reserving a portion of the budget for reannotating the labels most likely to be incorrect. Specifically, we start by annotating a large number of examples one time each using the majority of the budget $(n_a = 10,000; r = 1)$. We then pretrain a model M_1 using this noisy data, and observe either the model's training dynamics or output predictions to target examples for relabeling. We then use the remaining budget to annotate those examples two more times $(n_b = 1,000; r = 2)$, allowing us to obtain a majority vote on those examples. The final training set is formed by combining the 1k multiply-annotated examples with the remaining 9k singly-annotated examples. We wrap up by initializing a new model M_2 with the weights from M_1 and fine-tune it with the cleaned data until convergence. We experiment with four approaches for discovering the most probable noisy labels.

Area Under the Margin AUM identifies problematic labels by tracking the margin between the likelihood assigned to the target label class and the likelihood of the next highest class as training progresses (Pleiss et al., 2020). Intuitively, if the gap between these two likelihoods is large, then the model is confident of its argmax prediction, presumably because the training label is correct. On the other hand, if the gap between them is small, or even negative, then the model is uncertain of its prediction, presumably because the label is noisy. AUM averages the margins over all training epochs and targets the examples with the smallest margins for relabeling.

Cartography Dataset Cartography is a technique for mapping the training dynamics of a dataset to diagnose its issues (Swayamdipta et al., 2020). We take the suggestion from Section 5 of their paper to detect mislabeled examples by tracking the mean model probability of the true label across epochs. Note that unlike AUM, Cartography tracks the final model outputs after the softmax, rather than the logits before the softmax. These can lead to different rankings since Cartography does not take the other probabilities in the distribution into account. With that said, the intuition is largely the same, such that Cartography also chooses consistently low-confidence (ie. low probability) examples for relabeling.

Large Loss Arpit et al. (2017) found that correctly labeled examples are easier for a model to learn, and thus incur a small loss during training, whereas incorrectly labeled examples produce a large loss. Inspired by this observation and other similar works (Jiang et al., 2018), the Large Loss method selects examples for cleaning by ranking the top n_b examples where the model achieves the largest loss during the optimal stopping point. The ideal stopping point is the moment after the model has learned to fit the clean data, but before it has started to memorize the noisy data (Zhang et al., 2017). We approximate this stopping point by performing early stopping during training when the progression of the development set fails to improve for three epochs in a row.

Prototype We lastly consider identifying noisy labels as those which are outliers compared to the other training data Lee et al. (2018). More specifically, we use a pretrained model to map all training examples into the same embedding space. Then, we select the vectors for each label class to form clusters where the centroid of each cluster is the "prototype" Snell et al. (2017). Finally, we define outliers as those far away from the centroid for their given class, as measured by Euclidean distance.

4 EXPERIMENTS

4.1 DATASETS AND TASKS

To test our proposal, we select datasets that span across four natural language processing tasks. We choose these datasets because they provide multiple labels per example, allowing us to simulate single- and multiple-annotation scenarios.

Offense The Social Bias Frames dataset collects instances of biases and implied stereotypes found in text (Sap et al., 2020). We extract just the label of whether a statement is offensive for binary classification.

NLI We adopt the MultiNLI dataset for natural langauge inference (Williams et al., 2018). The three possible label classes for each sentence pair are *entailment*, *contradiction*, and *neutral*.

Sentiment Our third task uses the first round of the DynaSent corpus for four-way sentiment analysis (Potts et al., 2021). The possible labels are *positive*, *negative*, *neutral*, and *mixed*.

QA Our final task is question answering with examples coming from the NewsQA dataset (Trischler et al., 2017). The input includes a premise taken from a news article, along with a query related to the topic. The target label consists of two indexes representing the start and end locations within the article that extract a span of text answering the query. Unlike the other tasks, the format for QA is span selection rather than classification. Due to this distinction, certain denoising methods that assume a fixed set of candidate labels are omitted from comparison.

4.2 TRAINING CONFIGURATION

In our experiments, we fine-tune parameters during initial training with only six runs, which is composed of three learning rates and two levels of dropout at 0.1 and 0.05. Occasionally, when varying dropout had no effect, we consider doubling the batch size instead from 16 to 32. We found an appropriate range of learning rates by initially conducting some sanity checks on a sub-sample of development data for each task and model combination. When a technique contained method-specific variables, we defaulted to the suggestions offered in their respective papers. We do not expect any of the methods to be particularly sensitive to specific hyperparameters.

Methods	FastT	DRoB	DeXL	Avg
Oracle	*78.0*	*81.8*	*86.4*	*82.0*
Control	76.8	81.3	86.2	81.4
Majority	76.8	80.4	84.8	80.7
Adaptation	77.5	81.4	85.5	81.4
Crowdlayer	77.0	81.2	85.4	81.2
Bootstrap	77.0	81.2	85.2	81.1
Forward	77.4	80.9	85.2	81.2
Large Loss	77.6	**81.6**	85.5	**81.5**
AUM	77.2	81.4	85.5	81.3
Cartography	77.4	81.0	85.2	81.2
Prototype	**77.7**	81.1	**85.7**	**81.5**

(a) Offensive Language Detection from SBF

Methods	FastT	DRoB	DeXL	Avg
Oracle	*40.9*	*49.5*	*88.5*	*59.6*
Control	40.0	48.4	87.2	58.5
Majority	38.6	46.6	86.2	57.1
Adaptation	40.5	**49.3**	87.6	**59.1**
Crowdlayer	40.1	48.8	87.0	58.7
Bootstrap	**40.6**	49.2	87.4	59.0
Forward	40.5	48.9	87.3	58.9
Large Loss	40.4	49.2	**87.9**	**59.1**
AUM	40.1	49.0	87.3	58.8
Cartography	40.2	48.8	87.2	58.7
Prototype	40.5	48.8	**87.9**	**59.1**

(b) Natural Language Inference from MNLI

Methods	FastT	DRoB	DeXL	Avg
Oracle	*57.3*	*57.2*	*72.7*	*62.4*
Control	54.8	57.6	72.3	61.6
Majority	53.6	55.9	70.9	60.1
Adaptation	54.3	57.0	71.9	61.1
Crowdlayer	54.7	57.1	72.1	61.3
Bootstrap	54.7	57.5	72.2	61.5
Forward	51.8	57.5	72.6	60.6
Large Loss	**57.0**	**57.6**	72.5	**62.4**
AUM	55.9	56.6	72.3	61.6
Cartography	56.6	56.6	71.5	61.6
Prototype	56.5	57.0	**72.8**	62.1

(c) Sentiment Analysis from DynaSent

Methods	FastT	DRoB	DeXL	Avg
Oracle	—	*8.24*	*56.1*	*32.1*
Control	—	7.23	53.7	30.5
Majority	—	6.02	45.4	25.7
Adaptation	—	—	—	—
Crowdlayer	—	—	—	—
Bootstrap	—	6.80	53.8	30.3
Forward	—	—	—	—
Large Loss	—	**7.31**	55.4	**31.4**
AUM	—	6.76	**56.0**	**31.4**
Cartography	—	6.47	54.0	30.2
Prototype	—	—	—	—

(d) Question Answering from NewsQA

Figure 2: Aggregated results for all method and model combinations. Results shown are the average over two seeds. Model names are abbreviated for space: FastT is FastText, DRoB is DistilRoBERTa, and DeXL is DeBERTa-XLarge. Avg is the average across models for that method. FastText doesn't produce context-dependent representations, and so is not usable on the QA task.

4.3 MODEL VARIATIONS

We select three models for comparison that represent strong options at their respective model sizes. We repeat the process of example identification and re-annotation separately for each model. We use all models as a pre-trained encoders to embed the text inputs of the different tasks we study.

DeBERTa-XLarge is our large model, which contains 750 million parameters and currently is the state-of-the-art on many natural language understanding tasks (He et al., 2021).

DistilRoBERTa represents a distilled version of RoBERTa-base (Liu et al., 2019). It contains 82 million parameters, compared to the 125 million parameters found in RoBERTa. Learning follows the distillation process set by DistillBERT where a student model is trained to match the soft target probabilities produced by the larger teacher model (Sanh et al., 2019). Fine-tuning DistilRoBERTa is approximately 60-70 times faster compared to fine-tuning DeBERTa-XLarge on the the same task.

For the final model, we avoid using Transformers altogether and instead use the FastText bag-of-words encoder Joulin et al. (2017). The FastText embeddings are left unchanged during training, so the only learned parameters are in the 2-layer MLP we use for producing the model's final output. The same output prediction setup is used for all models, with a 300-dimensional hidden state. Training the FastText models run roughly 100-120 faster compared to working with DeBERTa-XLarge.

5 RESULTS AND DISCUSSION

5.1 OVERALL RESULTS

Figure 2 show the results across all models types and tasks, with each row representing a different technique. All rows except the Oracle were trained using the same label budget of 12,000 annotations.[3] In some cases, a method may surpass the Oracle since we conducted limited hyperparameter tuning, but as expected, the Oracle model outperforms all other methods overall. Notably, the Control setting always beats the Majority setting. In fact, Majority is consistently the lowest-performing method on all models and tasks, showing that increased label quality is never quite enough to overcome the reduction in annotation quantity. While advanced denoising methods often achieve better accuracy than the control, a cleaning method always ends up as the best average performer. In fact, Large Loss is the best overall method with the highest average performance in all four tasks. Prototypical is a strong runner-up, and Adaptation is the best among the denoising methods.

Breakdown by Task   Table 2a contains the results for offense language detection (Sap et al., 2020), where we see that Large Loss, Prototypical, and Adaptation are the only methods to overtake the Control. These three are also the best overall performers on natural language inference as seen in Table 2b. The cleaning methods really shine on sentiment analysis and question answering where even the worst cleaning method often tops the best denoising method. We hypothesize this happens because the denoising methods work best in simple classification tasks, which we further explore in the next section. A handful of results are not reported in Table 2d since they refer to methods that are designed exclusively for classification tasks, and cannot be directly transferred to span selection.

Breakdown by Model   The larger models perform better than the smaller models, but curiously, patterns on method performance do not seem to transfer across model types. Focusing on specific models, we note that Prototypical seems to do particularly well with DeBERTA-XLarge. This is possibly a stronger model produces more accurate hidden state representations for determining outliers. When studying FastText models, the Clean methods seem to achieve higher accuracy on average. There do not seem to be any consistent trends on DistillRoBERTa.

5.2 FURTHER ANALYSIS

Figure 4 displays the correlations among the different targeted cleaning methods. For all methods, we first ranked the examples chosen for relabeling by their likelihood of annotation error. Taking Large Loss as an example, we ordered the 1,000 items selected for relabeling by their loss, with the largest loss items coming first. Then for all pairs of methods, we calculate the Spearman rank correlation coefficients. All four methods show weak correlation to each other with the highest score at 0.065 between Large Loss and Cartography.

We additionally attempt to measure the similarity of the targeted cleaning methods by using Jaccard similar, as shown in Figure 3. To calculate the similarity score, we start again with the top 1,000 examples selected for relabeling. For a given pair of methods, we first calculate the size of the intersection of these examples and then divide by the size of the union. The scores for all pairs are quite low except for the AUM and Large Loss pairing. Figures 4 and 3 imply the cleaning methods behave very differently from each other, with the possible exception of AUM and Large Loss.

We display qualitative examples for sentiment analysis and NLI in Tables 1 and 2, respectively. Following the same procedure for calculating correlation, we rank all examples chosen by their likelihood of error for all four cleaning methods. Then we select the five examples from each method that are deemed most likely to be noisy along with their labels to study.

Large Loss for sentiment analysis consistently discovers 'neutral' labels that were mis-labeled as 'mixed'. Prototype for sentiment analysis also does a good job uncovering label errors, finding 'positive' examples mislabeled as 'negative' and 'mixed' examples mislabeled as 'neutral'. Cartography seems to have picked up on more nuanced examples that may be harder to learn. For example,

---

[3]This number of annotations is often much less than the total available data for a task so our results are not directly comparable to prior state-of-the-art methods. For example, the full DynaSent training set includes 94,459 examples and Social Bias Frames contains 43,448 examples.

	Large	AUM	Cart	Proto
Large Loss	1.000	0.007	0.065	0.025
Area Margin	---	1.000	-0.002	0.034
Cartography	---	---	1.000	-0.011
Prototype	---	---	---	1.000

(a) Natural language inference correlation coefficients

	Large	AUM	Cart	Proto
Large Loss	1.000	0.007	0.065	0.025
Area Margin	---	1.000	-0.002	0.034
Cartography	---	---	1.000	-0.011
Prototype	---	---	---	1.000

(b) Sentiment analysis correlation coefficients

Figure 3: Spearman correlation for all pairs of targeted relabeling methods on the natural language inference and sentiment analysis tasks.

	Large	AUM	Cart	Proto
Large Loss	1.000	0.449	0.001	0.050
Area Margin	---	1.000	0.000	0.057
Cartography	---	---	1.000	0.049
Prototype	---	---	---	1.000

(a) Natural language inference Jaccard distance

	Large	AUM	Cart	Proto
Large Loss	1.000	0.524	0.001	0.043
Area Margin	---	1.000	0.000	0.052
Cartography	---	---	1.000	0.064
Prototype	---	---	---	1.000

(b) Sentiment analysis Jaccard distance

Figure 4: Jaccard distance for all pairs of targeted relabeling methods on the natural language inference and sentiment analysis tasks.

'The servers were pleasant.' sounds positive, but could also be seen as negative if the statement is interpreted to mean that the servers were previously pleasant, but no longer are any more. Both AUM and Cartography are less reliable in selecting labeling errors, which helps to explain their lower quantitative performance.

For NLI, Large Loss is once again quite accurate in picking up labeling errors. Prototype for NLI does a great job at finding examples labeled as 'entailment' which should be something else. The hypotheses for all the selected examples contain negative sentiment, which may be located far away from the entailment examples in the embedding space. Cartography exhibits a pattern of always choosing examples labeled as 'contradiction'. However, once again AUM and Cartography have selected examples there were correctly labeled and did not need additional annotation.

6 CONCLUSION

Noisy data is a common problem when scaling data annotation, especially for natural language processing (NLP) tasks. We propose a new direction of targeted relabeling and perform a study of learning with noisy data in practical settings. We apply our methods on multiple model sizes and NLP tasks of varying difficulty. Our comparisons show that saving a portion of a fixed labeling budget for re-annotation outperforms other baselines for learning with noisy data. Our strongest method, Large Loss, not only achieves high model accuracy, but is also one of the easiest techniques to implement, requiring no extra parameters to train or hyper-parameters to tune. We believe a best practice for machine learning under a fixed annotation budget should include targeting examples for relabeling. Future work includes exploring more sophisticated techniques for identifying examples to relabel and understanding how such cleaning models perform on additional NLP tasks such as machine translation or dialogue state tracking, which have distinct output formats.

Method	Input Text	Label
Large Loss	That's usually how it go goes.	MIXED
	I always order "to-go"	MIXED
	It's $15 bucks for a beer since I used a drink ticket	MIXED
	We usually frequent the settlers ridge location.	MIXED
	I went on June 4th around 10:30.	MIXED
AUM	So fine, no problem.	POSITIVE
	A sirloin hotdog wrapped in bacon.	NEUTRAL
	For many years, I have gone to the Pet Smart down the street.	NEUTRAL
	I was always so happy here when it was managed by Johnny.	NEUTRAL
	I ordered the pad Thai noodles, chicken chow mien and egg rolls.	POSITIVE
Cartography	The food and customer service was fantastic when you first opened	POSITIVE
	The servers were pleasant.	POSITIVE
	Our waiter was overly friendly and informational.	MIXED
	Family owned and operated these folks are killing it	POSITIVE
	I really thought the young folks behind the counter were outgoing and seemed to enjoy their jobs.	POSITIVE
Prototype	This should be a fun family place!	NEGATIVE
	Hotel was awesome.	NEGATIVE
	Great service for many years on our cars, but always at an additional price.	NEUTRAL
	Salad was great but a bit small.	NEUTRAL
	We had to specify the order *multiple* times, but eventually when the food came it was actually pretty good.	NEUTRAL

Table 1: Sentiment Analysis examples each method identified as being most likely to be label errors.

Method	Premise	Hypothesis	Label
Large Loss	Why shouldn't he be?	He doesn't actually want to be that way.	ENTAILMENT
	How do they feel about your being a Theater major?	They don't know you're a theater major, do they?	ENTAILMENT
	Defication of humankind as supreme.	Humankind is not supreme.	ENTAILMENT
	These are artists who are either emerging as leaders in their fields or who have already become well known.	These artists are becoming well known in their fields.	CONTRADICTION
	As he stepped across the threshold, Tommy brought the picture down with terrific force on his head.	Tommy stepped across a threshold and put a picture down on his head.	CONTRADICTION
AUM	And if, as ultimately happened, no settlement resulted, we could shrug our shoulders, say, 'Hey, we tried.'	Even if an agreement could not be reach we could say we tried.	ENTAILMENT
	Companies that were foreign had to accept Indian financial participation and management.	Foreign companies had to take Italian money	CONTRADICTION
	... he's been tireless about pursuing both celebrity and the cause of popular history ever since.	He never wanted any attention and kept to himself all the time.	CONTRADICTION
	Two more weeks with my cute TV satellite dish have increased my appreciation of it.	My appreciation of my satellite dish has increased.	ENTAILMENT
	Each working group met several times to develop recommendations for ... legal services delivery system	Each working met more than once to discuss changes to the legal services delivery system.	ENTAILMENT
Cartography	A detailed English explanation of the plot is always provided, and wireless recorded commentary units ...	You'll have to figure the plot out on your own.	CONTRADICTION
	I just loved Cinderella . I also saw my sisters as the wicked stepsisters sometimes, and I was Cinderella ...	I really disliked Cinderella and could never relate to her.	CONTRADICTION
	The judge gave vent to a faint murmur of disapprobation and the prisoner in the dock leant forward angrily.	The prisoner in the dock remained still and and expressionless	CONTRADICTION
	Jon was about to require a lot from her.	Jon needed nothing to do with her.	CONTRADICTION
	I know you'll enjoy being a part of the Herron School of Art and Gallery.	You will detest the Herron School of Art and Gallery and have nothing to do with it	CONTRADICTION
Prototype	Why shouldn't he be?	He doesn't actually want to be that way.	ENTAILMENT
	I like this area a whole lot and it's, it's growing so much and I just want to be near my family ...	I really despise living in this location and would prefer to be farther away from my relatives.	ENTAILMENT
	The air is warm.	The arid air permeates the surrounding land.	ENTAILMENT
	Inside the Oval: White House Tapes From FDR to Clinton	No tapes were recorded in the white house	ENTAILMENT
	He became even more concerned as its route changed moving into another sector's airspace.	He wasn't worried at all for the plane	ENTAILMENT

Table 2: Natural Language Inference examples that each method identified as being most likely to be label errors. Sentences were truncated in some cases for brevity.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 312–321. PMLR, 2019. URL `http://proceedings.mlr.press/v97/arazo19a.html`.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017. URL `http://proceedings.mlr.press/v70/arpit17a.html`.

Fan Bai, Alan Ritter, and Wei Xu. Pre-train or annotate? domain adaptation with a constrained budget. *ArXiv*, abs/2109.04711, 2021. URL `https://arxiv.org/abs/2109.04711`.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 632–642. The Association for Computational Linguistics, 2015. doi: 10.18653/v1/d15-1075. URL `https://doi.org/10.18653/v1/d15-1075`.

Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. Max-MIG: an information theoretic approach for joint learning from crowds. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=BJg9DoR9t7`.

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

Paul Felt, Eric K. Ringger, and Kevin D. Seppi. Semantic annotation aggregation with conditional crowdsourcing models and word embeddings. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad (eds.), *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pp. 1787–1796. ACL, 2016. URL `https://aclanthology.org/C16-1168/`.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In Satinder P. Singh and Shaul Markovitch (eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 1919–1925. AAAI Press, 2017. URL `http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14759`.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=H12GRgcxg`.

Melody Y. Guan, Varun Gulshan, Andrew M. Dai, and Geoffrey E. Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3109–3118. AAAI Press, 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16970`.

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

In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8536–8546, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html`.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=XPZIaotutsD`.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10477–10486, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/ad554d8c3b06d6b97ee76a2448bd7913-Abstract.html`.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. Learning whom to trust with MACE. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pp. 1120–1130. The Association for Computational Linguistics, 2013. URL `https://aclanthology.org/N13-1132/`.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2309–2318. PMLR, 2018. URL `http://proceedings.mlr.press/v80/jiang18c.html`.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew S. Nokleby. An effective label noise model for DNN text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 3246–3256. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1328. URL `https://doi.org/10.18653/v1/n19-1328`.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pp. 427–431. Association for Computational Linguistics, 2017. doi: 10.18653/v1/e17-2068. URL `https://doi.org/10.18653/v1/e17-2068`.

Ashish Khetan and Sewoong Oh. Achieving budget-optimality with adaptive schemes in crowdsourcing. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4844–4852, 2016. URL `https://proceedings.neurips.cc/paper/2016/hash/03e7ef47cee6fa4ae7567394b99912b7-Abstract.html`.

Ashish Khetan, Zachary C. Lipton, and Animashree Anandkumar. Learning from noisy singly-labeled data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=H1sUHgb0Z`.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5447–5456. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00571. URL `http://openaccess.thecvf.com/content_cvpr_2018/html/Lee_CleanNet_Transfer_Learning_CVPR_2018_paper.html`.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=HJgExaVtwr`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

Volodymyr Mnih and Geoffrey E. Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL `http://icml.cc/2012/papers/318.pdf`.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2233–2241. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.240. URL `https://doi.org/10.1109/CVPR.2017.240`.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/c6102b3727b2a7d8b1bb6981147081ef-Abstract.html`.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. Dynasent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 2388–2404. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.186. URL `https://doi.org/10.18653/v1/2021.acl-long.186`.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna K. Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman (eds.), *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pp. 889–896. ACM, 2009. doi: 10.1145/1553374.1553488. URL `https://doi.org/10.1145/1553374.1553488`.

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6596`.

Filipe Rodrigues and Francisco C. Pereira. Deep learning from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 1611–1618. AAAI Press, 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16102`.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL `https://doi.org/10.1007/s11263-015-0816-y`.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 3027–3035. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33013027. URL `https://doi.org/10.1609/aaai.v33i01.33013027`.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 5477–5490. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.486. URL `https://doi.org/10.18653/v1/2020.acl-main.486`.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4077–4087, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html`.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pp. 1297–1304, 2004. URL `https://proceedings.neurips.cc/paper/2004/hash/358aee4cc897452c00244351e4d91f69-Abstract.html`.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 254–263. ACL, 2008. URL `https://aclanthology.org/D08-1027/`.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional neural networks with noisy labels. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL `http://arxiv.org/abs/1406.2080`.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 9275–9293. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.746. URL `https://doi.org/10.18653/v1/2020.emnlp-main.746`.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5552–5560. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00582. URL `http://openaccess.thecvf.com/content_cvpr_2018/html/Tanaka_Joint_Optimization_Framework_CVPR_2018_paper.html`.

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 11244–11253. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01150. URL `http://openaccess.thecvf.com/content_CVPR_2019/html/Tanno_Learning_From_Noisy_Labels_by_Regularized_Estimation_of_Annotator_Confusion_CVPR_2019_paper.html`.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP at ACL 2017, Vancouver, Canada, August 3, 2017*, pp. 191–200. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-2623. URL https://doi.org/10.18653/v1/w17-2623.

Peter Welinder, Steve Branson, Serge J. Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pp. 2424–2432. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/0f9cafd014db7a619ddb4276af0d692c-Abstract.html.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pp. 2035–2043. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/hash/f899139df5e1059396431415e770c6dd-Abstract.html.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1101. URL https://doi.org/10.18653/v1/n18-1101.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7164–7173. PMLR, 2019. URL http://proceedings.mlr.press/v97/yu19b.html.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. Learning with different amounts of annotation: From zero to many labels. *ArXiv*, abs/2109.04408, 2021. URL https://arxiv.org/abs/2109.04408.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8792–8802, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f2925f97bc13ad2852a7a551802feea0-Abstract.html.