Coupled Confusion Correction: Learning from Crowds with Sparse Annotations

Hansong Zhang, Shikun Li, Dan Zeng, Chenggang Yan, Shiming Ge

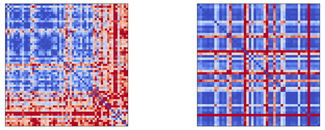
Summary

Background:

Precisely annotating large-scale datasets is very time-consuming. As a alternative, datasets collected by crowd-sourcing is much cheaper but contain noisy labels, which will eventually decrease the generalization performance of deep networks.

Motivation:

- In previous works, the modeling of annotator expertise rely solely on its individual annotations. Therefore, when the annotations are sparse (which is ubiquitous in crowd-sourcing datasets), the annotator-specific-confusion parameters (ASCPs) will be poorly learned, leading to decreased model performance.
- Moreover, when generating synthetic crowd-sourcing datasets, previous works neglected the unbalance of the number of annotations per worker, which is not consistent with the real-world ones.
- In crowd-sourcing datasets, there are some annotator groups that share similar expertise. Their ASCPs can be better modeled if they can be considered together.



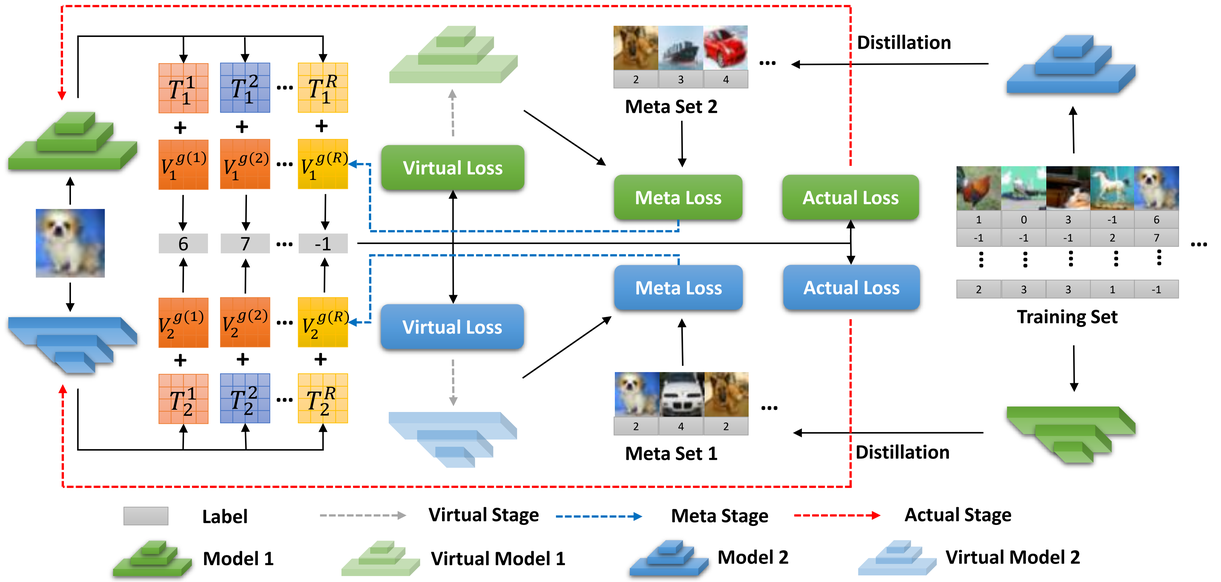
LabelMe CIFAR-10N

Contribution:

- We propose Coupled Confusion Correction (CCC) to mitigate the influence of annotation sparsity, where the confusion matrices of annotators can be updated under the supervision of not only its individual annotations but also the distilled annotations.
- We propose to imbalance the distribution of number of annotations via a Beta distribution, making the synthetic datasets more consistent with real-world ones.
- Extensive experiments are conducted on various datasets, where the results indicate the superiority of our method.

Methodology

Main idea: The core idea of CCC is (1) distill a small **meta set** based on small-loss criterion; (2) employ the distilled meta set to correct the learned confusion matrices of annotators by meta-learning; (3) cluster annotators with similar expertise using K-Means method, so that their confusion matrices can be corrected together.



Meta Set Distillation: For each class c, we fetch all the instances that are attached with the label c, then we select the ones with M/C smallest losses, where M is the size of the meta set. Further, to avoid the confirmation bias, we use two differently-initiated models to distill meta set for each other.

Correct Confusion Matrices via Meta-Learning: In our CCC, the corrected confusion matrices are obtained through the following bi-level optimization:

$$\{\boldsymbol{V}^{r\star}\}_{r=1}^{G} = \underset{\{\boldsymbol{V}^{r}\}_{r=1}^{G}}{\arg\min} \frac{1}{M}\sum_{j=1}^{M}\mathcal{L}_j^{meta}(\boldsymbol{\theta}^{\star}), \text{s.t.},$$

$$\boldsymbol{\theta}^{\star}, \{\boldsymbol{T}^{r\star}\}_{r=1}^{R} = \underset{\boldsymbol{\theta},\{\boldsymbol{T}^{r}\}_{r=1}^{R}}{\arg\min} \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}1[\mathcal{A}_{i,r}]\mathcal{L}_{i,r}^{tra}(\boldsymbol{T}_{cor}^{r},\boldsymbol{\theta})$$

where $\boldsymbol{T}_{cor}^{r} = \boldsymbol{T}^{r} + \boldsymbol{V}^{g(r)}$ denotes the corrected confusion matrix of r-th annotator with $g(r)$ representing the index of group which the r-th annotator belongs to. Let ℓ denotes the cross entropy loss function, the meta loss here is then defined as $\mathcal{L}_j^{meta}(\boldsymbol{\theta}^{\star}) = \ell(f(\boldsymbol{x}_j^{meta}|\boldsymbol{\theta}^{\star}), y_j^{meta})$, and the training loss is $\mathcal{L}_{i,r}^{tra}(\boldsymbol{T}_{cor}^{r},\boldsymbol{\theta}) = \ell((\boldsymbol{T}^{r}+\boldsymbol{V}^{g(r)})f(\boldsymbol{x}_i|\boldsymbol{\theta}), \tilde{y}_i^{r})$. $\mathcal{A}_{(N\times R)}$ is a matrix indicating the annotation presence, i.e., $\mathcal{A}_{i,j}$ is True if the i-th instance is labeled by j-th annotator, and False otherwise.

Experiments

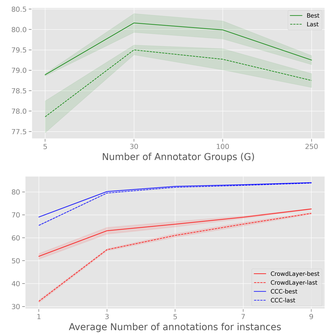
Experiments on the Synthetic Datasets with Independent Noise:

Dataset		CIFAR-10				Fashion-MNIST			
Method / Case		IND-I	IND-II	IND-III	IND-IV	IND-I	IND-II	IND-III	IND-IV
MajorVote	Best	55.08±0.38	47.85±2.20	54.27±1.91	66.94±1.01	84.89±2.24	79.94±2.30	76.72±1.00	89.03±0.75
	Last	41.14±0.37	32.90±0.71	42.00±0.39	60.18±1.95	61.00±0.65	51.31±0.56	56.88±0.29	78.02±0.52
CrowdLayer	Best	63.06±1.49	53.50±1.24	59.39±1.21	69.09±0.62	88.83±0.61	84.77±2.09	87.01±3.05	90.31±0.42
	Last	54.78±0.49	43.71±0.62	53.17±0.34	67.51±0.91	75.89±0.19	66.39±0.21	66.67±0.47	84.67±0.20
DoctorNet	Best	69.63±0.86	63.43±1.46	69.77±2.02	76.25±0.47	90.54±0.23	88.78±0.68	88.29±1.20	90.96±0.32
	Last	66.80±1.12	57.95±1.09	65.40±0.98	74.60±0.45	83.01±0.13	76.07±0.49	75.77±0.20	86.78±0.19
Max-MIG	Best	71.63±1.15	65.52±0.99	69.73±1.09	75.97±0.39	91.03±0.38	90.15±0.45	90.07±0.49	91.32±0.25
	Last	66.47±0.48	56.63±0.46	61.68±0.63	73.19±0.59	82.16±0.32	73.33±0.39	73.53±0.47	86.16±0.37
CoNAL	Best	67.28±1.24	59.27±1.53	65.93±0.15	78.16±0.27	89.95±0.30	87.62±0.88	89.13±0.44	91.72±0.14
	Last	56.18±0.30	47.02±0.65	57.16±0.49	71.96±0.53	75.27±0.17	65.80±0.32	71.16±0.29	86.67±0.25
UnionNet	Best	74.67±0.82	68.92±0.96	71.52±1.14	79.29±0.65	91.30±0.37	90.09±0.42	89.67±0.58	91.61±0.25
	Last	74.44±0.60	63.41±0.75	67.00±0.42	78.78±0.74	88.22±0.19	78.29±0.29	79.09±0.54	90.56±0.21
CCC (Ours)	Best	**80.16±0.23**	**75.33±0.43**	**78.28±0.25**	**83.06±0.26**	**92.59±0.01**	**91.93±0.18**	**92.50±0.13**	**93.23±0.06**
	M.I.	↑ 5.49	↑ 6.41	↑ 6.76	↑ 3.77	↑ 1.29	↑ 1.78	↑ 2.43	↑ 1.51
	Last	**79.50±0.12**	**73.52±0.39**	**77.20±0.26**	**82.56±0.07**	**92.51±0.06**	**91.58±0.39**	**92.17±0.04**	**93.14±0.06**
	M.I.	↑ 5.06	↑ 10.11	↑ 10.20	↑ 3.78	↑ 4.29	↑ 13.29	↑ 13.08	↑ 2.58

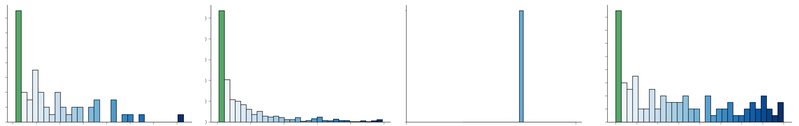
Experiments on Real-world Datasets:

Method / Dataset		LabelMe	CIFAR-10N	MUSIC
MajorVote	Best	80.72±0.45	81.08±0.34	67.67±0.98
	Last	77.19±1.18	80.89±0.25	62.00±1.05
CrowdLayer	Best	84.01±0.36	80.43±0.25	71.67±0.50
	Last	82.32±0.41	80.14±0.28	69.33±0.93
DoctorNet	Best	82.32±0.49	83.68±0.33	67.33±0.74
	Last	81.73±0.49	83.32±0.29	65.33±0.75
Max-MIG	Best	86.28±0.35	83.25±0.47	75.33±0.69
	Last	83.16±0.66	83.12±0.63	71.67±0.93
CoNAL	Best	87.46±0.53	82.51±0.15	74.00±0.89
	Last	**84.85±0.91**	80.92±0.28	68.67±1.88
UnionNet	Best	85.19±0.42	82.66±1.12	72.33±0.58
	Last	82.66±0.38	82.32±1.09	68.33±0.91
CCC (Ours)	Best	**87.65±1.10**	**86.36±0.05**	**76.22±0.42**
	M.I.	↑ 0.19	↑ 2.68	↑ 0.89
	Last	84.79±0.80	**86.12±0.12**	**72.89±0.68**
	M.I.	↓ 0.06	↑ 2.80	↑ 1.22

Ablation Studies:



Number of Annotations per Annotator:



LabelMe CIFAR-10N Previous Ours