南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Active Learning for Noisy Data Streams Using Weak and Strong Labelers

Taraneh Younesian
Delft University of Technology
t.younesian@tudelft.nl

Dick Epema
Delft University of Technology
d.h.j.epema@tudelft.nl

Lydia Y. Chen
Delft University of Technology
y.chen-10@tudelft.nl

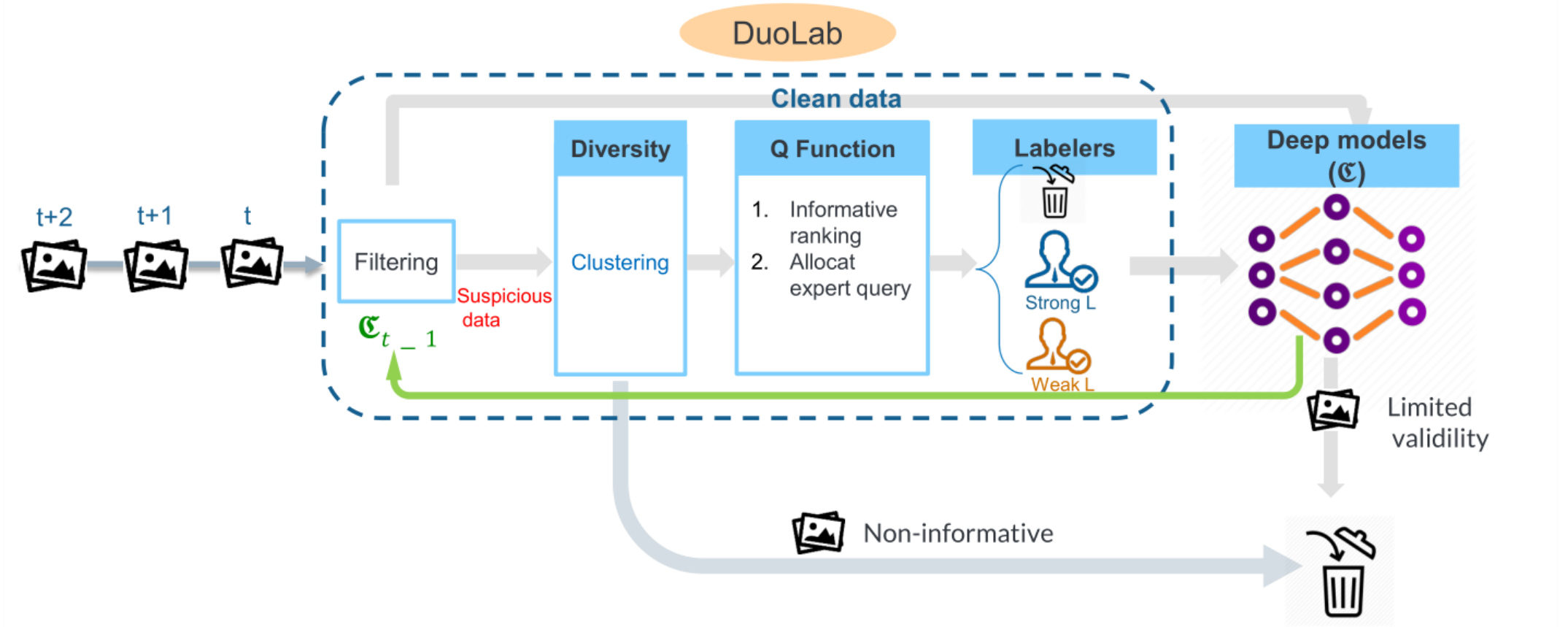


Figure 1: On-line learning scenario with noisy labels: data streams, classifier, and strong/weak labelers

DuoLab

- The objective is to train a classifier in the presence of noisy labels considering these two labelers with a limited budget B per batch for labeling to achieve a certain level of accuracy.

- DuoLab consists of the four stages of ***filtering***, ***clustering***, ***ranking*** and ***labeler selection***.

Filtering: Identifying the Suspicious Data

- The first step of DuoLab is to identify the suspicious data samples that might have corrupted labels.

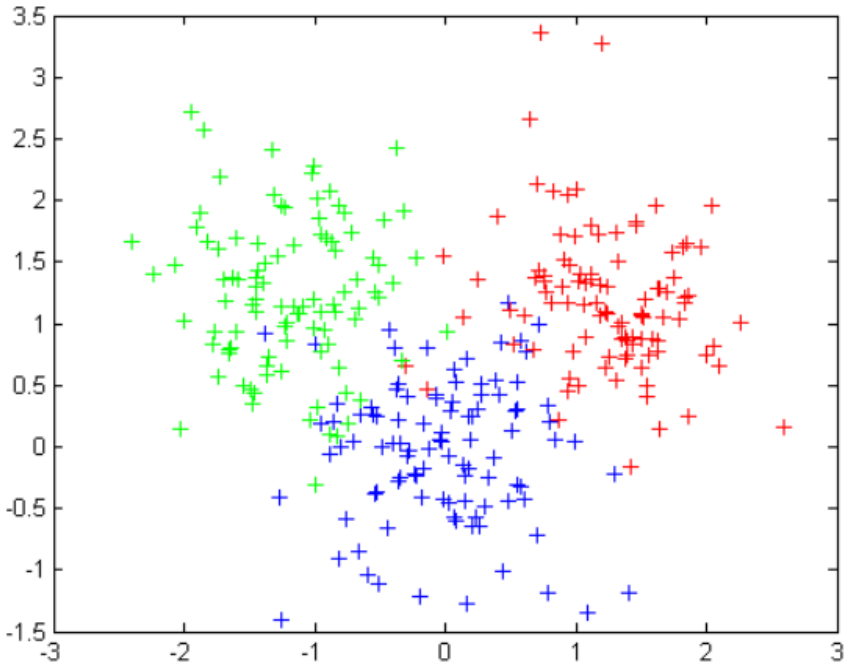
We consider the samples that have $\hat{y}_j^1 = \tilde{y}_j$ or $\hat{y}_j^2 = \tilde{y}_j$ to be clean

$$\text{clean set } C = \{(\mathbf{x}_j^c, y_j)\}$$

$$\text{suspicious set } U = \{(\mathbf{x}_j^u, \tilde{y}_j)\}$$

Clustering: Adding Diversity

- Only relying on the informative sample selection may **result in selecting similar samples** that would cause a waste of the labeling budget.
- Next, we select the k most informative samples of each cluster using active learning as explained in the next section, and discard the rest.



Ranking: Informativeness

Best-versus-second-best (BvSB)

$$I(\mathbf{x}_j) = p_{best}(\mathbf{x}_j) - p_{second-best}(\mathbf{x}_j)$$ ↓

The informativeness of $\mathbf{x}_j$ ↑

Labeler Selection: Cost Optimization

Cost Sensitive Labeler Selection Function

$$L_V(t) = \sum_{v=1}^{V} p(\mathbf{x}_v, y) \log(p(\mathbf{x}_v, \hat{y}^1))$$

$$Q(\mathbf{x}_j^u) = \frac{L_V(t)}{I(\mathbf{x}_j^u) \mathfrak{c} E_{\mathcal{S}}(t)}$$

$$E_{\mathcal{W}}(t) + \mathfrak{c} E\mathcal{S}(t) \leq B$$

Algorithm

Algorithm 1: Algorithm of DuoLab: filtering, clustering, labeler selection, and training

Input : Initial dataset D^I, Data batches D, weak labeler $\mathcal{W}$, strong labeler $\mathcal{S}$, cost of strong labeler $\mathfrak{c}$, maximum number of weak queries per sample $\bar{w}$, budget B, clustering parameter $\mathfrak{k}$

Output : Training set C for the classifier $\mathfrak{C}$

Train $\mathfrak{C}$ with D^I

foreach *arriving* $D = \{(\mathbf{x}_j, \tilde{y}_j)\}$ *and each* $(\mathbf{x}_j, \tilde{y}_j)$ **do**

  if $(\hat{y}_j^1 = \tilde{y}_j)$ ***or*** $(\hat{y}_j^2 = \tilde{y}_j)$ **then**

    $C = \{(\mathbf{x}_j^c, y_j), \mathbf{x}_j^c := \mathbf{x}_j\}$ `#clean set`

  else

    $U = \{(\mathbf{x}_j^u, \tilde{y}_j), \mathbf{x}_j^u := \mathbf{x}_j\}$ `#suspicious set`

Apply Kmeans clustering on U.

From each cluster select the most informative $\mathfrak{k}$ samples and add to $K = \{(\mathbf{x}_j^k, \tilde{y}_j)\}$. `#Discard the rest`

for $\mathbf{x}_i$ *in* K *where* i *is the informativeness index* **do**

  $w = 0$

  if $(Q > \bar{Q})$ ***and*** $(E_W(t) + \mathfrak{c}E_S(t) + \mathfrak{c} \leq B)$ **then**

    Query $\mathcal{S}$, update $E_{\mathcal{S}}$ and Q

    Add $\mathbf{x}_i$ to C

  else if $(E_{\mathcal{W}}(t) + \mathfrak{c}E_{\mathcal{S}}(t) + 1 \leq B)$ ***and*** $(w < \bar{w})$ **then**

    Query $\mathcal{W}$ based on Ω, update $E_{\mathcal{W}}$ and Q, $w = w + 1$

    if *(The answer is "Yes")* **then**

      Add $\mathbf{x}_i$ to C

    else

      Go to step 14

  else

    Discard $\mathbf{x}_i$

Experiments

Noise		30%					60%				
Method	**c**	**Acc(%)**	**no. S**	**no. W**	**TP(%)**	**FP(%)**	**Acc(%)**	**no. S**	**no. W**	**TP(%)**	**FP(%)**
DuoLab	2	**76.13**	22.0	57.0	61.43	3.81	**69.44**	38.4	63.0	33.57	7.84
DuoLab	10	**75.45**	4.7	52.5	60.90	3.80	67.42	9.8	51.8	32.95	7.71
DuoLab + Kmeans	2	75.99	15.3	67.1	61.26	3.66	68.61	21.5	78.7	33.38	7.84
DuoLab + Kmeans	10	75.33	4.1	56.1	60.70	3.78	67.53	21.5	78.7	33.38	7.84
Only $\mathcal{S}$	10	74.96	25.0	-	60.80	3.78	66.84	25.0	-	32.84	7.81
Only $\mathcal{W}$	-	75.11	-	67.6	60.87	3.84	**68.93**	-	89.5	33.53	7.75
Clean All Suspicious	-	77.26	343.2	-	61.72	3.91	75.71	-	575.1	35.03	7.46
No AL(only Filter)	-	73.60	-	-	60.77	3.78	63.34	-	-	32.26	8.11
Opt Filter	-	77.78	-	-	70.00	-	72.56	-	-	40.00	-
No Filter	10	62.21	7.3	96.7	-	-	36.91	17.7	24.2	-	-

Experiments

Noise		30%					60%				
Method	**c**	**Acc(%)**	**no. S**	**no. W**	**TP(%)**	**FP(%)**	**Acc(%)**	**no. S**	**no. W**	**TP(%)**	**FP(%)**
DuoLab	5	39.07	33.3	102.1	32.18	0.44	**34.45**	44.0	171.9	17.21	0.97
DuoLab	25	39.27	7.3	102.7	32.57	0.45	**34.14**	8.9	174.2	17.65	0.90
DuoLab + Kmeans	5	38.45	16.9	105.1	33.14	0.49	34.35	17.3	190.8	17.65	0.97
DuoLab + Kmeans	25	38.98	4.0	100.5	32.30	0.49	32.92	4.0	181.4	17.11	0.97
Only $\mathcal{S}$	25	**39.57**	90.0	-	33.14	0.47	33.13	90.0	-	17.12	0.95
Only $\mathcal{W}$	-	38.25	-	105.9	32.83	0.48	32.94	-	192.2	17.12	0.99
Clean All Suspicious	-	49.05	5085.4	-	40.91	0.45	49.53	6829.6	-	23.38	0.74
No AL(only Filter)	-	38.08	-	-	32.63	0.44	32.32	-	-	16.18	0.98
Opt Filter	-	42.28	-	-	70.00	-	37.33	-	-	40.00	-
No Filter	25	30.86	112.0	-	-	-	13.19	112.0	-	-	-

Experiments

Table 3: The accuracy of DuoLab and the noise resilient baselines for CIFAR-10 and CIFAR-100 with 30% noise.

	Method	Accuracy (%)	
		CIFAR-10	CIFAR-100
Baselines	D2L	52.77	11.55
	Forward	59.94	25.56
	Co-teaching	61.52	29.41
	Bootstrap soft	48.95	23.35
	Bootstrap hard	49.61	24.02
Our	DuoLab ($\mathfrak{c} = 2, 5$)	76.13	39.07
	DuoLab ($\mathfrak{c} = 10, 25$)	75.45	39.27

Active label cleaning: Improving dataset quality under resource constraints

Mélanie Bernhardt[1,*], Daniel C. Castro[1,*], Ryutaro Tanno[1], Anton Schwaighofer[1], Kerem C. Tezcan[1], Miguel Monteiro[1], Shruthi Bannur[1], Matthew Lungren[2], Aditya Nori[1], Ben Glocker[1], Javier Alvarez-Valle[1], and Ozan Oktay[1,†]

[1] *Health Intelligence, Microsoft Research Cambridge, Cambridge, CB1 2FB, UK*
[2] *Department of Radiology, Stanford University, Palo Alto, CA 94304, USA*

[*] *These authors contributed equally to this work.*
[†] *Corresponding author: ozan.oktay@microsoft.com*

Active label cleaning

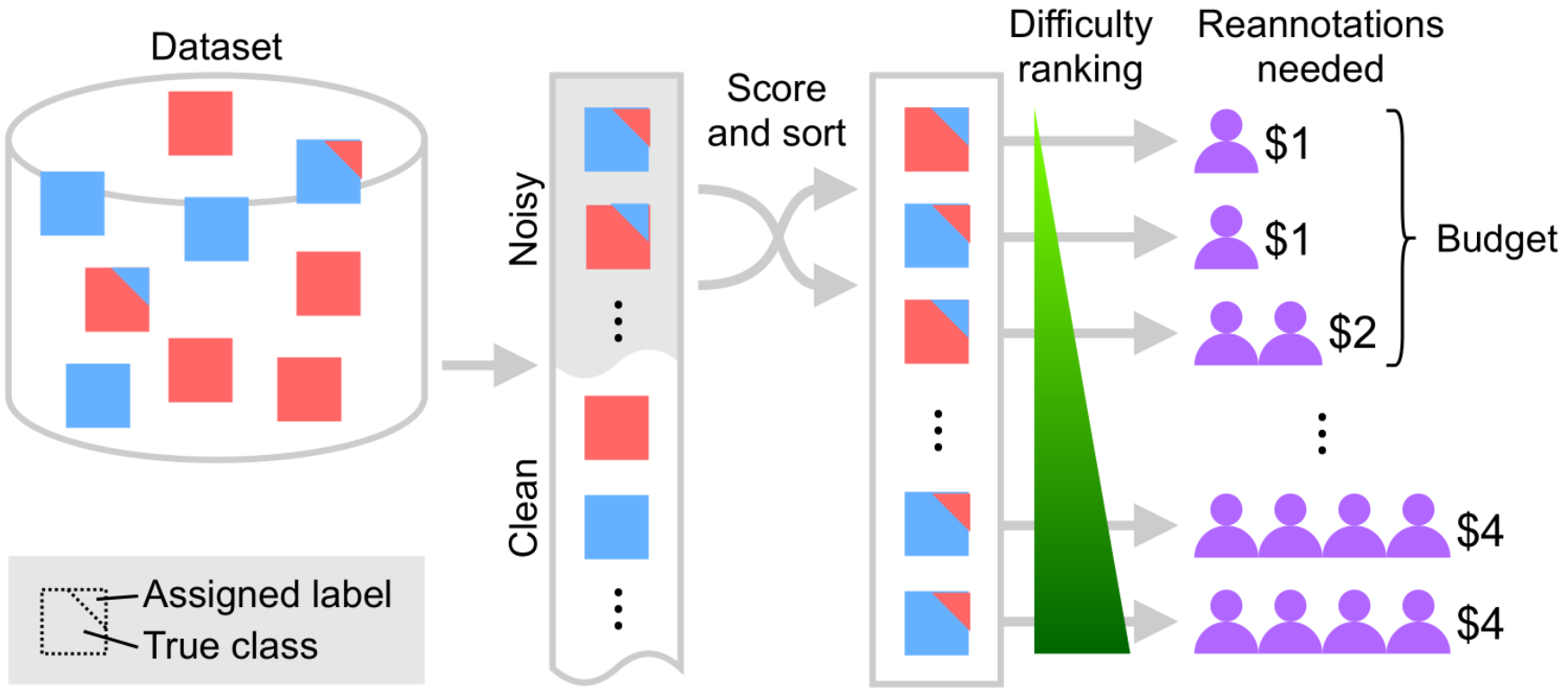


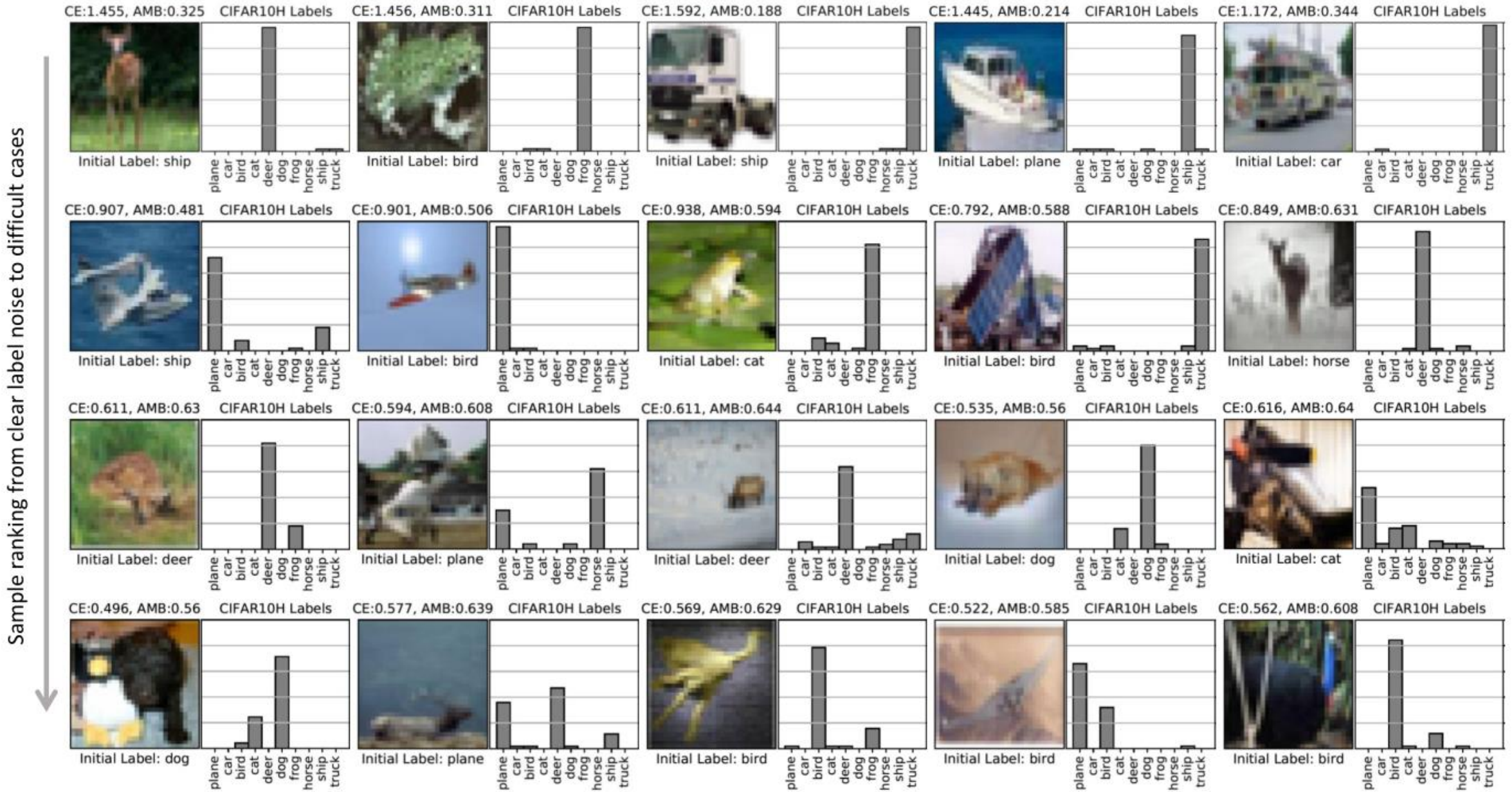
Fig. 1 Overview of the proposed active label cleaning. A dataset with noisy labels is sorted to prioritise clearly mislabelled samples, maximising the number of corrected samples given a fixed relabelling budget.

Active label cleaning

$$\max \underbrace{\frac{1}{N}\sum_{i=1}^{N} 1[\hat{y}_i = y_i]}_{\text{correctness of majority labels}} \quad \text{s.t.} \underbrace{\sum_{i=1}^{N} \|\hat{\ell}_i\|_1 \leq B}_{\text{budget constraint}}, \tag{1}$$

$$\Phi(\mathbf{x}, \hat{\ell}; \theta) = \underbrace{CE(\hat{\ell}, p_\theta)}_{\text{noisiness} \uparrow} - \underbrace{H(p_\theta)}_{\text{ambiguity} \downarrow}. \tag{2}$$

Active label cleaning



Conclusion

Table 1 Active label cleaning

Given: $Y = \{\ell_i\}_{i=1}^N$: True label distributions
Input: $\mathcal{D} = \{(\mathbf{x}_i, \hat{\ell}_i)\}_{i=1}^N$: Dataset with noisy labels
$B \in \mathbb{N}$: Relabelling budget
$b \in \mathbb{N}$: Update frequency

$\theta \leftarrow$ TRAINROBUSTMODEL($\mathcal{D}$)
$\mathcal{I}_{\text{avail}} \leftarrow \{1, \ldots, N\}, \quad \mathcal{I}_{\text{cleaned}} \leftarrow \emptyset$
count $\leftarrow$ 0
while count $< B$ **do** ▷ If budget remains
  $j \leftarrow \arg\max_{i \in \mathcal{I}_{\text{avail}}} \Phi(\mathbf{x}_i, \hat{\ell}_i; \theta)$ ▷ Rank (Eq. (2))
  repeat
    $\hat{\ell}_j \leftarrow \hat{\ell}_j +$ SAMPLE(ℓ_j) ▷ Acquire one-hot label
    count $\leftarrow$ count + 1
  until majority formed in $\hat{\ell}_j$
  $\mathcal{I}_{\text{avail}} \leftarrow \mathcal{I}_{\text{avail}} \setminus \{j\}, \quad \mathcal{I}_{\text{cleaned}} \leftarrow \mathcal{I}_{\text{cleaned}} \cup \{j\}$
  $\mathcal{D} \leftarrow \{(\mathbf{x}_i, \hat{\ell}_i) : i \in \mathcal{I}_{\text{avail}} \cup \mathcal{I}_{\text{cleaned}}\}$
  if count divisible by b **then**
    $\theta \leftarrow$ UPDATE($\theta, \mathcal{D}$) ▷ Fine-tune model
  end if
end while
return $\mathcal{D}$

THANKS