Appendix of "Trusted Multi-view Learning with Label Noise"

Anonymous submission

In this appendix, we provide a more detailed description of the noise correlation matrix and types of label-noise, and show more implementation details of the proposed Trusted Multi-view Noise Refining (TMNR) method, hyper-parameter settings and performance comparisons under different situations.

1 Label-Noise Learning

1.1 Types of Label-Noise

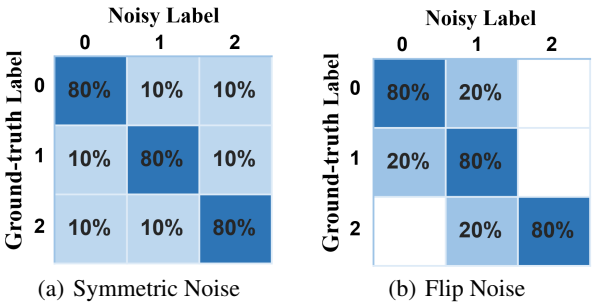
In Class-Conditional Noise (CCN), the process of label corruption is independent of the features of the instance itself. Its ground-truth labels are corrupted by a dataset-specific noise correlation matrix $\boldsymbol{T} = [t_{ij}]_{i,j=1}^{K} \in [0,1]^{K\times K}$, where $t_{ij} = P(\tilde{y} = j|y = i)$, $i, j \in \{1, ..., K\}$ and K is the number of all categories. In this hypothesis, when the noise rate is σ, the noise is said to be symmetric noise (as in Figure 1(a)) if $\forall_{i=j}\ t_{ij} = (1-\sigma)$ and $\forall_{i\neq j}\ t_{ij} = (\sigma/(K-1))$. As opposed to this, asymmetric noise means that for $\forall_i$, $\sum_{j=1,j\neq i}^{K} t_{ij} = \sigma$ and $\exists_{j\neq i}\ t_{ij} \neq [\sigma/(K-1)]$, then it means that when labels are incorrectly labelled, it is more likely that they will be corrupted to a specific category rather than randomly labelled into other categories with average probability. When a category of labels will only transform into a specific category of labels, as in $\forall_{i=j}\ t_{ij} = (1-\sigma)$ and $\exists_{j\neq i} t_{ij} = \sigma$, then it is called flip noise, shown in Figure 1(b).

In the real world, considering that samples with ambiguous features or poor collection quality are more likely to be mislabeled. Therefore, in Instance-Dependent Noise (IDN) modeling, it is assumed that the process of labeling corruption depends on its own features and category labels, and the sample $\boldsymbol{x}_n$ specific noise correlation matrix is defined as $t_{ij} = P(\tilde{y} = j|y = i, \boldsymbol{x}_n)$.

2 Supplementary Experimental Content

2.1 Noise Generation

To destroy the labels of the clean dataset used in the main text, we emulate the previous method [Cheng *et al.*, 2020] of destroying labels to generate datasets containing instance-dependent label noise. First we train a trusted classifier $g(\cdot)$ on a subset of the clean dataset using the same model as the Evidence Extraction Network part and make predictions for each sample that will be used for training. Then the noisy labels are obtained by taking into account the amount of evidence $\boldsymbol{e}$ obtained for each class and the uncertainty u of the opinions. Specifically, we measure the magnitude of the probability of whether a sample is corrupted in terms of uncertainty, i.e., elements with greater uncertainty are more likely to be corrupted. When a sample is selected for destruction, we pick $\tilde{y} = arg\ max(e_k)$ and $k \neq y$, $k \in \{1, ..., K\}$. This means that we set the category that the classifier believes to be most similar to the category to which the instance belongs as the noise label to satisfy our assumption of instance-related label noise. We added noise with a scale of 10% to 50% to all the datasets to evaluate the proposed method.



(a) Symmetric Noise (b) Flip Noise

Figure 1: Different types of CCN correlation matrices when the noise ratio $\sigma = 20\%$ and $K = 3$.

Datasets	Size	Classes	Dimensionality
UCI	2000	10	6/47/24
PIE	680	68	484/256/279
BBC	685	5	4659/4633/4665/4684
Caltech101	8677	101	48/40/254/1984/512/928
Leaves100	1600	100	64/64/64

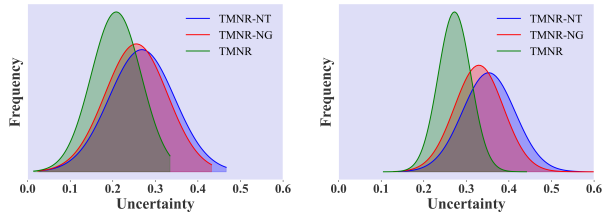
Table 1: Summary of feature dimensions for each view for all datasets.

2.2 Network Construction

The UCI Dataset, PIE Dataset, BBC Dataset, Caltech101 Dataset and Leaves100 Dataset are composed of pre-extracted vectorized features. Then we extract view-specific evidence using a fully connected layer network with ReLU

Datasets	Method			Instance-Dependent Noise					
	$\mathcal{L}_{acc}$	$\mathcal{M}$	$\mathcal{L}_{con}$	0%	10%	20%	30%	40%	50%
PIE	✔	-	-	89.85±2.69	84.41±3.13	82.65±3.24	77.65±2.11	66.47±3.50	60.44±5.82
	✔	✔	-	89.26±2.05	85.00±2.61	82.79±3.04	77.65±1.51	67.79±3.63	60.00±6.19
	✔	-	✔	89.71±2.23	84.41±3.06	82.79±3.28	77.06±1.43	67.94±3.44	60.88±4.25
	✔	✔	✔	**90.74±1.51**	**85.51±1.95**	**82.94±2.92**	**78.38±1.10**	**68.68±3.71**	**61.47±4.82**
Caltech101	✔	-	-	90.38±0.87	90.75±0.55	86.86±0.95	87.57±1.79	79.41±1.11	70.67±2.68
	✔	✔	-	**90.79±0.79**	90.59±1.08	87.82±1.29	87.62±1.13	79.62±1.33	71.21±2.35
	✔	-	✔	90.21±1.00	90.75±0.79	87.62±0.73	88.03±0.86	80.46±1.94	73.64±2.87
	✔	✔	✔	90.63±0.83	**91.05±0.83**	**88.08±0.98**	**88.08±1.15**	**80.63±1.74**	**73.93±3.00**
Leaves100	✔	-	-	63.75±2.70	61.19±1.50	61.56±1.98	55.50±3.30	55.31±6.24	47.56±6.62
	✔	✔	-	**69.25±3.62**	66.00±2.94	63.19±2.20	58.06±5.90	56.25±5.94	52.56±5.64
	✔	-	✔	68.56±2.76	70.50±3.66	64.62±2.25	61.69±4.23	58.06±2.48	58.81±4.81
	✔	✔	✔	69.19±2.15	**70.81±3.17**	**68.25±2.23**	**65.00±4.53**	**61.81±4.71**	**59.50±3.17**

Table 2: Ablation study on three datasets. $\mathcal{L}_{acc}$ denotes overall classification loss, $\mathcal{M}$ indicates uncertainty bootstrap regularization and $\mathcal{L}_{con}$ denotes loss of consistency of the inter-view correlation matrix. "✔" indicates that the corresponding component in the TMNR is applied, and "-" indicates that it is not used. Best results are highlighted by **bold**.



(a) BBC IDN-10%

(b) BBC IDN-40%

Figure 2: Evaluating the performance of TMNR with its different degradation methods in terms of uncertainty assessment on BBC datasets containing 10% and 40% noise.

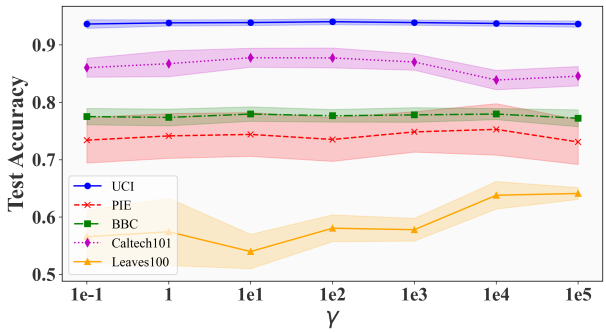


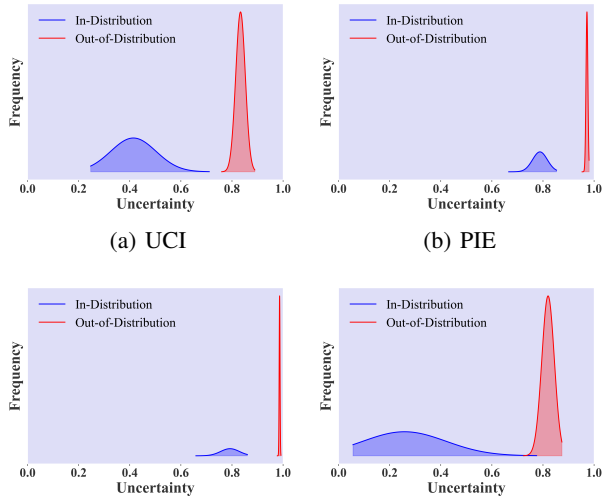
Figure 3: Classification accuracy when adjusting γ on all datasets with 30% noise rate.

activation functions. To be fair, for each dataset we add only one fully connected layer for each view with inputs as the feature dimensions of the view and outputs as the number of categories K. The feature dimensions of each view for all datasets are summarised in Table 1. For all datasets, the hyper-parameter β is set to $1e^{-2}$, and the non-diagonal elements of the correlation matrix are compared with their 5-nearest neighbors in the computation of the uncertainty bootstrap loss $\mathcal{M}$. The selection process of hyper-parameters γ for different datasets will be shown later.

2.3 Ablation Study

To demonstrate the effectiveness of our proposed uncertainty bootstrap regularisation $\mathcal{M}$ and loss of consistency of inter-view correlation matrices $\mathcal{L}_{con}$ in optimising noisy correlation matrices. We performed an ablation study: without adding any additional assumptions to the correlation matrix and using only one of the two modules. We verify their effectiveness by evaluating their performance under different levels of labelled noise. As shown in Table 2, each constraint almost always improves the performance of the experimental results under different noise levels, illustrating that the correlation patterns between clean and noisy evidence distributions can be learnt more effectively through uncertainty bootstrapping and inter-perspective consistency constraints. And the constraints on the correlation matrix work better when the noise percentage is high.

Meanwhile, in order to verify the effectiveness of the proposed noise correlation matrix in eliminating the effect of label noise, we obtain TMNR-NG and TMNR-NT by removing the additional constraints and correlation matrices sequentially. The uncertainty distributions obtained from the prediction of the test set using the TMNR with the two degradation methods trained on the BBC dataset containing 10% and 40% label noise are shown in Figure 2. The results show that although the uncertainty of the model inevitably increases with the increase of the noise ratio in the training data, the incorporation of the correlation matrix effectively reduces the uncertainty of the test sample decisions, especially when applying the constraints we designed. Lower uncertainty implies sharper Dirichlet distributions corresponding to the aggregation of opinions from multiple perspectives, indicating that sufficient evidence has been observed to make trusted decisions.



(a) UCI (b) PIE

(c) Leaves100 (d) Caltech101

Figure 4: Identification of in/out-of-distribution samples.

2.4 Hyper-parametric Analysis

We analyze it here the sensitivity of the hyper-parameter γ adjusting for the inter-view correlation matrix consistency loss $\mathcal{L}_{con}$ on all datasets containing 30% noise. The results is shown in Figure 3. We observe that although different datasets have different sensitivities to the parameter γ, appropriate parameter values can still improve the overall performance of the model. We set different γ values for each dataset based on the results of this experiment. The UCI dataset, PIE dataset, BBC dataset and the Leaves100 dataset are set to $1e^4$, and the Caltech101 dataset is set to $1e^3$.

2.5 Identification of Out-of-Distribution Data

To verify the effectiveness of our proposed TMNR as a trusted model in data noise identification, we add Gaussian noise with fixed standard deviation ($\sigma = 10$) to 50% of the test samples in the four datasets so that they constitute out-of-distribution (OOD) samples, and the remaining data serve as in-distribution samples. Their uncertainties were predicted using the model obtained on the training data without labeling noise and the results are shown in Figure 4. We can observe that the intra-distribution samples all obtained lower uncertainty than the OOD samples on all datasets. Meanwhile, this UCI dataset with higher prediction accuracy exhibits lower uncertainty overall, while the Leaves100 dataset with lower prediction accuracy has higher uncertainty. These results prove the reasonableness and effectiveness of TMNR's measure of data uncertainty to ensure that the decisions output by our model are trustworthy.

References

[Cheng *et al.*, 2020] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2020.