Learning with Label Noise for Image Retrieval by Selecting Interactions Supplementary Material

Sarah Ibrahimi† Arnaud Sors‡ Rafael Sampaio de Rezende‡ Stéphane Clinchant‡
† University of Amsterdam ‡ NAVER LABS Europe

A. Sensitivity of the τ hyperparameter

For T-SINT, τ is tuned, starting with an estimation of its value. We analyze how sensitive T-SINT is to the selection of the τ for different noise values for CUB by analyzing MAP@R results. This is shown in Figure F1.

B. Negative interactions

We stated that chances that an observed negative interaction is actually a positive interaction are very small whenever the number of different instances in a batch is much smaller than the number of total classes in the dataset. For T-SINT, we use a batch size of 80 samples, with 4 samples per class, resulting in 20 classes per batch.

We analyse what is the minimum number of classes for which this assumption holds and we performed experiments on subsets of the CARS train set, consisting of 20, 40, 60, and 80 classes. To each of these subsets, we add the noise ratios as for the full CARS dataset with 98 classes. Fewer classes leads to fewer training images which will result in a drop in performance. Therefore we do not compare the performances between those subsets, but the ratios of performance with noise to performance without noise. This results in Table T1, which shows the Precision@1 scores. From this table, we can see that only for the subset of 20 classes, the relative performance is much lower for increasing noise ratios, which indicates that the method does not work very well in this case. This might be due to using false negative interactions, but might also be caused by the small number of clean samples that are present in the subset of 20 classes for high noise rates.

From this table, we can conclude, that our method seems to work well for datasets with at least 40 classes, which is common for most real-world image retrieval datasets. For datasets with a smaller number of classes, one could think of adding an additional hyperparameter equivalent to τ to serve as a threshold for negative interactions.

C. Hyperparameters

For all methods, we tune the learning rate and batch size based on the datasets with 20% and 50% uniform noise.

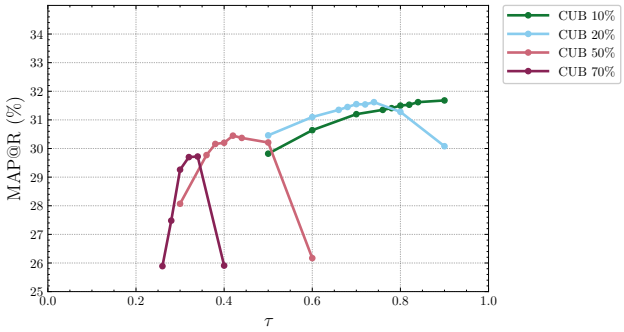

Figure F1: For low noise rates, T-SINT is not very sensitive to the τ value. For higher noise rates it is more sensitive, but even for 50% noise it has a range of possible τ values where the performance is quite stable.

Noise level	10%	20%	50%	70%
CARS_{full}	0.992	0.994	0.989	0.981
CARS_{80}	0.999	1.00	0.994	0.982
CARS_{60}	0.996	1.00	0.992	0.970
CARS_{40}	1.00	1.00	0.980	0.945
CARS_{20}	0.977	0.973	0.937	0.903

Table T1: Ratio P@1(p%)/P@1(0%) between no-noise performance and noise-corrected performance as the number of classes used during training is decreased

For the contrastive margin loss and PRISM, we find the best learning rate to be 1e-6 and a batch size of 40 for all datasets. For the SuperLoss, we find the best learning rate to be 1e-5 with a batch size of 128, except for CARS-98N, where we set this learning rate to 1e-6. For T-SINT, we use a learning rate of 1e-6 and a batch size of 80, except for SOP, where we use a learning rate of 3e-6. This is for using the ViT-B/32 backbone. For reproducing the experiments from [22], which consists of the Contrastive Margin Loss and PRISM with BN-inception and ResNet-50 backbones, we use the settings as provided in the paper [22].

We tune the hyperparameters related to noise as follows.

Dataset	Noise	λ	τ_+/τ_-
CUB	Uniform 0%	0.01	GlobalAvg
CUB	Uniform 10%	0.01	GlobalAvg
CUB	Uniform 20%	0.01	GlobalAvg
CUB	Uniform 50%	0.1	GlobalAvg
CUB	Uniform 70%	-	-
CARS	Uniform 0%	1	GlobalAvg
CARS	Uniform 10%	0.1	GlobalAvg
CARS	Uniform 20%	0.1	GlobalAvg
CARS	Uniform 50%	-	-
CARS	Uniform 70%	-	-
SOP	Uniform 0%	1	GlobalAvg
SOP	Uniform 10%	1	GlobalAvg
SOP	Uniform 20%	1	GlobalAvg
SOP	Uniform 50%	0.01	GlobalAvg
SOP	Uniform 70%	0.01	GlobalAvg
CARS-98N	Realistic	0.01	ExpAvg
Oxford	Realistic	0.25	ExpAvg
Landmarks	Realistic	0.05	GlobalAvg

Table T2: SuperLoss hyperparameters λ, τ_+ and τ

Dataset	Noise	τ
CUB	Uniform 0%	0.97
CUB	Uniform 10%	0.81
CUB	Uniform 20%	0.70
CUB	Uniform 50%	0.44
CUB	Uniform 70%	0.31
CARS	Uniform 0%	1.00
CARS	Uniform 10%	0.84
CARS	Uniform 20%	0.74
CARS	Uniform 50%	0.44
CARS	Uniform 70%	0.34
SOP	Uniform 0%	1.00
SOP	Uniform 10%	1.00
SOP	Uniform 20%	0.90
SOP	Uniform 50%	0.64
SOP	Uniform 70%	0.48
CARS-98N	Realistic	0.40
Oxford	Realistic	0.62
Landmarks	Realistic	0.66

Table T3: T-SINT hyperparameter τ

For PRISM on uniform noise, we set the estimated noise rate R according to the noise rates we use for uniform noise, e.g. R=0.7 for 70% uniform noise. For CARS-98N, we use R=0.5 from [22]. For Oxford and Landmarks we tune the noise rate and find the best values to be R=0.5 for Oxford and R=0.6 for Landmarks.

For the SuperLoss, we tune λ, τ_+ and τ_-. For λ we tried the values 0.001, 0.01, 0.05, 0.1, 0.25, 1.0 as these values were recommended in [5]. For the thresholds, [5] recommends three options: a global average, an exponential running average with a fixed smoothing parameter or a fixed value given by prior knowledge on the task. We experimented with the global average and the exponential running average. The best hyperparameters can be found in Table T2. Note that for Landmarks, we took the best hyperparameters according to [5].

For T-SINT, we use the estimation of τ from Equation 5 and tune it from there. An overview of these values for each dataset is presented in Table T3.

D. Results

In Tables T4, T5 and T6, the results for uniform noise are given that are presented in Figure 4. As a reference, we also present the results on all datasets for the Contrastive Margin Loss and PRISM when using the original backbone from [22]. Since [22] does not report MAP@R scores, we rerun all these experiments with the hyperparameters provided in the original paper.

Noise Rate	0%	10%	20%	50%	70%
Contrastive Margin $\text{Loss}_{\text{BN-inception}}$ [9]	57.41/20.87	56.65/19.11	56.33/18.64	40.02/7.97	34.45/6.32
Contrastive Margin $\text{Loss}_{\text{ViT-B/32}}$ [9]	71.17/29.43	67.88/25.99	62.49/20.62	- / -	- / -
$\text{PRISM}_{\text{BN-inception}}$ [22]	57.48/18.86	58.32/20.17	57.33/19.18	54.29/17.25	46.78/12.80
$\text{PRISM}_{\text{ViT-B/32}}$ [22]	72.06/31.11	72.43/31.27	71.93/31.10	70.78/29.66	64.45/23.85
$\text{SuperLoss}_{\text{ViT-B/32}}$ [5]	70.32/29.05	69.85/28.53	69.21/27.67	58.85/18.71	- / -
$\text{T-SINT}_{\text{ViT-B/32}}$ (Ours)	72.05/31.60	71.73/31.50	71.49/31.50	71.08/30.37	70.51/29.74

Table T4: Precision@1 / MAP@R (%) on CUB dataset with synthetic uniform label noise.

Noise Rate	0%	10%	20%	50%	70%
Contrastive Margin $\text{Loss}_{\text{BN-inception}}$ [9]	75.37/21.16	74.85/18.75	67.27/13.22	36.60/3.17	32.54/2.62
Contrastive Margin $\text{Loss}_{\text{ViT-B/32}}$ [9]	88.85/41.64	88.78/41.04	87.79/37.88	- / -	- / -
$\text{PRISM}_{\text{BN-inception}}$ [22]	80.02/22.95	78.02/21.37	76.93/19.76	70.15/15.77	52.75/7.50
$\text{PRISM}_{\text{ViT-B/32}}$ [22]	89.08/41.62	89.08/41.08	88.97/40.81	87.44/38.20	80.63/28.48
$\text{SuperLoss}_{\text{ViT-B/32}}$ [5]	87.89/39.03	87.18/36.94	86.69/34.55	- / -	- / -
$\text{T-SINT}_{\text{ViT-B/32}}$ (Ours)	89.67/42.71	88.97/42.21	89.10/41.90	88.69/40.41	87.94/36.33

Table T5: Precision@1 / MAP@R (%) on CARS dataset with synthetic uniform label noise.

Noise Rate	0%	10%	20%	50%	70%
Contrastive Margin $\text{Loss}_{\text{ResNet-50}}$ [9]	64.14/35.64	65.50/36.73	64.70/ 35.51	57.87/28.90	52.93/25.02
Contrastive Margin $\text{Loss}_{\text{ViT-B/32}}$ [9]	77.00/50.67	78.74/52.95	77.99/51.27	73.65/44.30	67.35/37.08
$\text{PRISM}_{\text{ResNet-50}}$ [22]	76.54/48.71	74.93/46.27	73.65/44.60	60.37/30.68	52.73/24.87
$\text{PRISM}_{\text{ViT-B/32}}$ [22]	77.87/50.28	77.84/49.30	74.75/45.67	68.18/37.65	61.32/31.37
$\text{SuperLoss}_{\text{ViT-B/32}}$ [5]	80.29/56.39	82.09/58.91	82.06/57.97	82.02/57.94	77.14/49.05
$\text{T-SINT}_{\text{ViT-B/32}}$ (Ours)	78.73/53.80	80.91/56.78	81.09/56.78	81.26/56.59	79.52/53.70

Table T6: Precision@1 / MAP@R (%) on SOP dataset with synthetic uniform label noise.