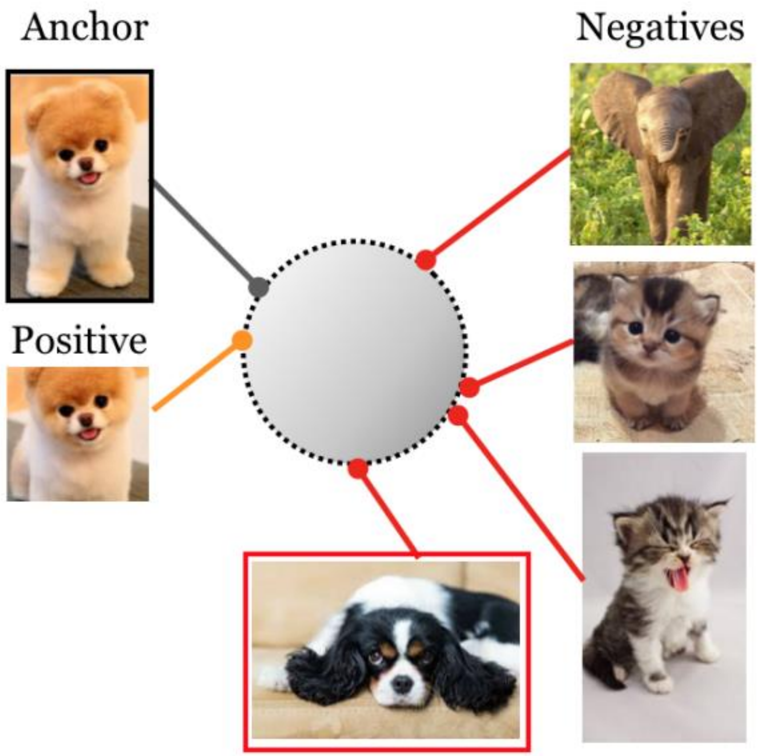
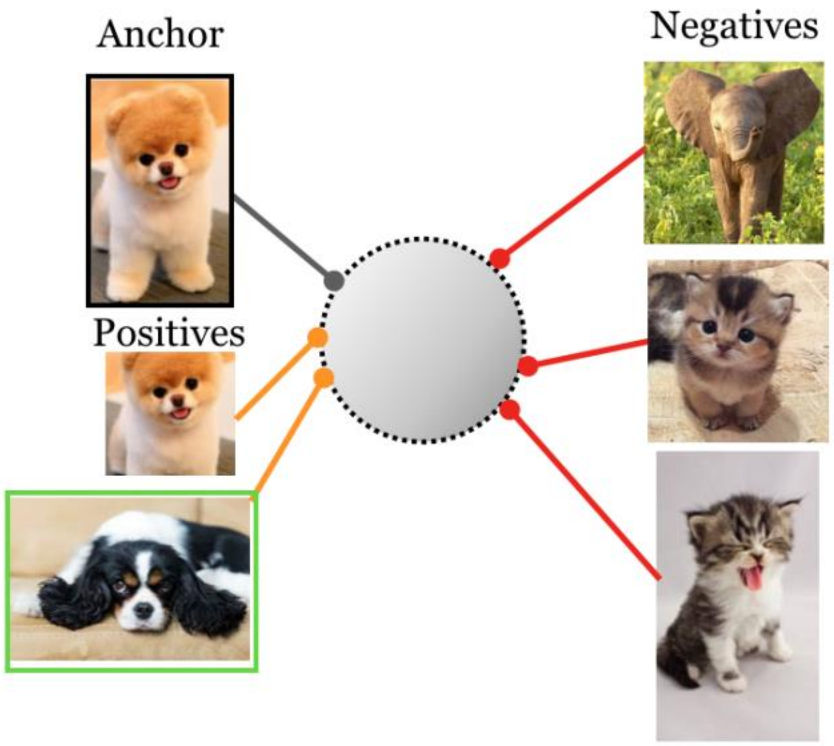
Multi-Objective Interpolation Training for Robustness to Label Noise

Diego Ortego, Eric Arazo, Paul Albert, Noel E. O'Connor and Kevin McGuinness
Insight Centre for Data Analytics, Dublin City University (DCU)
{diego.ortego, eric.arazo}@insight-centre.org

introduction



Self Supervised Contrastive



Supervised Contrastive

introduction

Robustness to label noise is usually pursued by identifying noisy samples to:

1. reduce their contribution in the loss
2. correct their label
3. abstain their classification
4. regularizing label noise information in DNN weights
5. small sets of correctly labeled data

directly learning image representations rather than a class mapping

Method

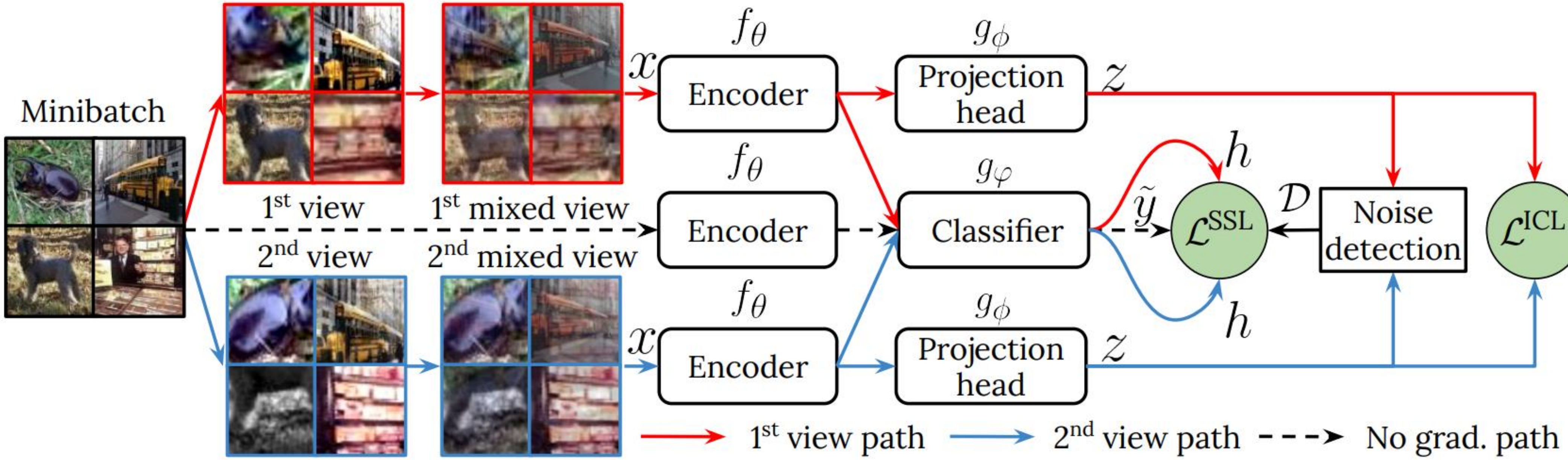


Figure 1. Multi-Objective Interpolation Training (MOIT) for improved robustness to label noise. We interpolate samples and impose the same interpolation in the supervised contrastive learning loss $\mathcal{L}^{ICL}$ and the semi-supervised classification loss $\mathcal{L}^{SSL}$ that we jointly use during training. Label noise detection is performed at every epoch to enable semi-supervised learning and its result is used after training to fine-tune the encoder and classifier to further boost performance.

Method

per-sample loss

对比损失

$$\mathcal{L}_i\left(z_i, y_i\right)=\frac{1}{2N_{y_i}-1}\sum_{j=1}^{2N}\mathbb{1}_{i\neq j}\,\mathbb{1}_{y_i=y_j}P_{i,j}$$

$$P_{i,j}=-\log\frac{\exp\left(z_i\cdot z_j/\tau\right)}{\sum_{r=1}^{2N}\mathbb{1}_{r\neq i}\exp\left(z_i\cdot z_r/\tau\right)}$$

$$x_i=\lambda x_a+(1-\lambda)\,x_b,\quad \lambda\in[0,1]\sim\mathrm{Beta}\,(\alpha,\alpha)$$

x_i denotes the training sample that combines two minibatch samples x_a and x_b

$$\mathcal{L}_i^{MIX}=\lambda\mathcal{L}_i\left(z_i, y_a\right)+(1-\lambda)\,\mathcal{L}_i\left(z_i, y_b\right)$$

Method

unless a large minibatch is used during training, few positive and negative samples are selected, which negatively affects the training process

Memory bank

$\mathcal{L}_i^{MEM}$ contrasts the 2N samples with the M memory samples, thus extending the number of positive and negative samples

$$\mathcal{L}^{ICL} = \mathcal{L}^{MIX} + \mathcal{L}^{MEM}$$

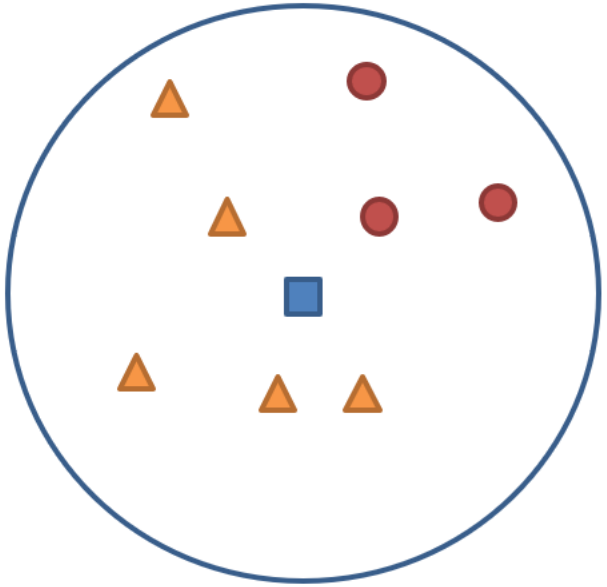
$\mathcal{L}_i^{MEM}$ Similar to $\mathcal{L}_i^{MIX}$

Method

Label noise detection

estimating a class probability distribution from the representation zi by doing a k-nearest neighbor (k-NN) search:

$$p\left(c \mid x_i\right)=\frac{1}{K} \sum_{\substack{k=1 \\ x_k \in \mathcal{N}_i}}^{K} \mathbb{1}_{y_k \neq c},$$



However, the labels y might be noisy, thus biasing the estimation of p.
We, therefore, estimate a corrected distribution $\hat{p}$ using:

$$\hat{p}\left(c \mid x_i\right)=\frac{1}{K} \sum_{\substack{k=1 \\ x_k \in \mathcal{N}_i}}^{K} \mathbb{1}_{\hat{y}_k \neq c},$$

$$\hat{y}=\arg\max_c\, p\left(c \mid x\right)$$

$$d_i=-y_i^T \log \left(\hat{p}\right),$$

$$\mathcal{D}_c=\left\{\left(x_i, y_i\right): d_i \leq \gamma_c\right\},$$

Method

Semi-supervised learning

We learn the classifier by performing semi-supervised learning where samples in $\mathcal{D}$ are considered as labeled and the remaining samples as unlabeled.

$$\mathcal{L}_i^{SSL} = -\lambda \tilde{y}_a^T \log(h_i) - (1-\lambda)\, \tilde{y}_b^T \log(h_i)$$

where the pseudo-label $\tilde{y}_a$ ($\tilde{y}_b$) for x_a (x_b) is estimated as

$$\tilde{y}_a = \begin{cases} y_a, & x_a \in \mathcal{D}_c \\ \bar{h}_a, & x_a \notin \mathcal{D}_c \end{cases}, \tag{11}$$

where $\bar{h}_a$ is the softmax prediction for image x_a without data augmentation. The final Multi-Objective Interpolation Training (MOIT) optimizes the loss:

$$\mathcal{L}^{MOIT} = \mathcal{L}^{ICL} + \mathcal{L}^{SSL}. \tag{12}$$

Classification refinement

$$\mathcal{L}_i^{MOIT+} = -\lambda \left[\left(\delta y_a + (1-\delta)\, \tilde{y}_a\right)^T \log(h_i) \right] - (1-\lambda) \left[\left(\delta y_b + (1-\delta)\, \tilde{y}_b\right)^T \log(h_i) \right],$$

experiment

Table 2. Weighted k-NN evaluation in CIFAR-100.

	Symmetric			Asymmetric	
	0%	40%	80%	10%	40%
SCL	72.66	58.32	41.00	71.11	68.00
ICL	75.30	66.38	53.60	74.34	72.04
MOIT	**75.76**	**67.42**	**55.58**	**74.86**	**72.60**

Table 3. Classification accuracy for different noise detection strategies and K values for 40% asymmetric noise in CIFAR-100.

	K	5	10	25	50	100	150	200	250	300	350
k-NN (p)	Acc.	59.42	61.74	64.84	66.10	67.18	67.42	67.46	67.68	67.14	66.94
k-NN ($\hat{p}$)	Acc.	62.28	65.30	68.58	70.56	71.16	71.22	71.24	**71.42**	70.98	70.80

experiment

Table 6. Performance in CIFAR-10 with symmetric and asymmetric noise. (*) Denotes that we have run the algorithm.

	Symmetric				Asymmetric			Avg.
	0%	20%	40%	80%	10%	30%	40%	
CE	93.85	78.93	55.06	33.09	88.81	81.69	76.04	72.50
Mix [53]	**95.96**	84.76	66.07	20.38	93.30	83.26	77.74	74.50
DB [1]	79.18	93.82	92.26	15.53	89.58	92.20	91.20	79.11
DMI [49]	93.88	88.33	83.24	43.67	91.11	91.16	83.99	82.20
PCIL [50]	93.89	92.72	91.32	55.99	93.14	92.85	91.57	87.35
DRPL [34]	94.08	94.00	92.27	61.07	**95.50**	92.98	92.84	88.96
DMix* [28]	94.27	**95.12**	**94.11**	35.36	93.77	92.47	90.04	85.02
ELR* [30]	95.49	94.49	92.56	38.23	95.25	**94.66**	92.88	86.22
MOIT	95.17	92.88	90.55	70.53	93.50	93.19	92.27	89.73
MOIT+	95.65	94.08	91.95	**75.83**	94.23	94.31	**93.27**	**91.33**

Table 7. Performance in CIFAR-100 with symmetric and asymmetric noise. (*) Denotes that we have run the algorithm.

	Symmetric				Asymmetric			Avg.
	0%	20%	40%	80%	10%	30%	40%	
CE	74.34	58.75	42.92	8.29	68.10	53.28	44.46	50.02
Mix [53]	77.90	66.40	52.20	13.21	72.40	57.63	48.07	55.40
DB [1]	64.79	69.11	62.78	45.67	67.09	58.59	47.44	59.35
DMI [49]	74.44	58.82	53.22	20.30	68.15	54.15	46.20	53.61
PCIL [50]	77.75	74.93	68.49	25.41	76.05	59.29	48.26	61.45
DRPL [34]	71.84	71.16	72.37	**52.95**	72.03	69.30	65.69	67.91
DMix* [28]	67.41	71.39	70.83	49.52	69.53	68.28	50.99	63.99
ELR* [30]	**78.01**	**75.90**	**72.89**	36.83	77.08	74.61	71.25	69.51
MOIT	75.83	72.78	67.36	45.63	75.49	73.34	71.55	68.85
MOIT+	77.07	75.89	70.88	51.36	**77.43**	**75.13**	**74.04**	**71.69**

Ablation study

Table 5. Ablation study for MOIT and MOIT+ in CIFAR-100. A: Asymmetric, S: Symmetric, SSL: semi-supervised learning, M: memory, B: Balanced clean set, r-t C: Re-training classifier, s-DA: strong data augmentation.

	S-40%	A-40%
(MOIT) w/o SSL	62.82	53.73
(MOIT) w/o M	66.10	68.88
(MOIT) w/o B	66.28	69.58
MOIT	**66.58**	**71.42**
(MOIT+) w/o r-t C	69.54	73.32
(MOIT+) w/ s-DA	67.98	71.90
MOIT+	**70.68**	**73.58**

experiment

Table 8. Performance evaluation on controlled web noise in mini-ImageNet. We run all methods.

		0%	20%	40%	80%
Mix [53]	Best	61.18	57.76	52.88	38.32
	Last	58.96	54.60	50.40	37.32
DMix [28]	Best	57.80	55.86	55.44	41.12
	Last	55.84	50.30	50.94	35.42
ELR [30]	Best	63.12	61.48	57.32	41.68
	Last	57.38	58.10	50.62	41.68
MOIT	Best	67.18	64.82	61.76	46.40
	Last	64.72	63.14	60.78	45.88
MOIT+	Best	**68.28**	**64.98**	**62.36**	**47.80**
	Last	67.82	63.10	61.16	46.78

thanks