Truncate-Split-Contrast

A Framework for Learning from Mislabeled Videos

Zixiao WANG

January 5, 2023

Department of Computer Science and Engineering
The Chinese University of Hong Kong

Introduction

Learning with Noisy Label(LNL)
Samples in a dataset can be mislabeled due to various reasons.

- Over-fitting mislabeled samples harms model generalization ability on unseen data[Zha+17][1]
- The existing LNL methods focus on image tasks.

[1] C. Zhang *et al.*, "Understanding deep learning requires rethinking generalization," *ICLR*, 2017.

Introduction

Learning with Noisy Label on Images

Depending on whether noisy instances are detected in training, the existing LNL methods can be roughly divided into two types.

- Noise robust models
  - robust loss functions, e.g., MAE, SCE[Wan+19][2]
- Noise detection methods
  - loss-based methods, e.g., M-Correction[Ara+19][3]
  - feature-based methods, e.g., CleanNet[LHZY18][4]

---

[2] Y. Wang *et al.*, "Symmetric cross entropy for robust learning with noisy labels," in *ICCV*, Oct. 2019.

[3] E. Arazo *et al.*, "Unsupervised label noise modeling and loss correction," in *ICML*, 2019.

[4] K.-H. Lee *et al.*, "Cleannet: Transfer learning for scalable image classifier training with label noise," in *CVPR*, 2018.

Introduction

Noise Utilization Methods

After detecting the potential noisy instances, models can utilize these noisy samples by,

- simply excluding them
- re-using them by estimating the pseudo labels of them in a semi-supervised fashion.

Introduction

Learning with Noisy Label on Videos

A straightforward migration from images to videos is not a sound choice.

- **computational cost**

  The pseudo label could be ambiguous and unreliable without sophisticated post-processing and data enrichment, which is time-consuming.

- **temporal semantics**

  This characteristic would be intuitively beneficial to noise detection.

- **high feature dimension**

  Feature-based methods may fail due to the curse of dimensionality.

Truncate-Split-Contrast

Pipeline

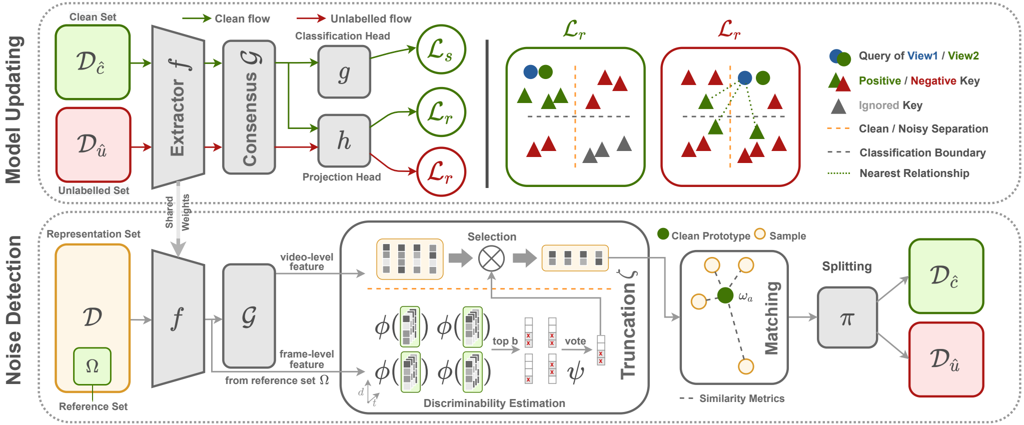

Figure 1: The pipeline of Truncate-Split-Contrast.

Formulation

Dataset $\mathcal{D}$, clean set $\mathcal{D}_c = \epsilon(\mathcal{D})$, and noisy set $\mathcal{D}_u = \mathcal{D} - \mathcal{D}_c$.

Classification head $g(\cdot)$

The loss is defined by,

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_r = -\sum_{\mathbf{x}_i \in \epsilon(\mathcal{D})} \sum_{k=1}^{K} \mathbf{y}_i(k) \cdot \log g(\mathbf{x}_i, k) + \lambda \mathcal{L}_r, \tag{1}$$

Truncate

In feature space, given the query $\boldsymbol{x}$ with label $\boldsymbol{a}$, and the clean-prototype $\boldsymbol{x}_a$ of category $\boldsymbol{a}$,

$$Similarity(\boldsymbol{x}, \boldsymbol{x}_a) = \boldsymbol{x} \cdot \boldsymbol{x}_a \tag{2}$$

We argue that to detect clean/noisy instances, utilizing all channels of a feature learned from classification supervision is not a must.

Truncate

A top $\boldsymbol{b}$ channel selection operation $\zeta_{\boldsymbol{b}}(\cdot)$

A category-level score function $\psi(\cdot)$

The reference set of category $\boldsymbol{a}$, Ω^{a}

The truncation function is defined as,

$$\boldsymbol{w} = \zeta_{b}\left(\boldsymbol{x},\ \psi\left(\Omega^{a}\right)\right). \tag{3}$$

Given the query $\boldsymbol{w}$ with label $\boldsymbol{a}$, and the clean-prototype $\boldsymbol{w}_{a}$,

$$\eta(\boldsymbol{w}) \equiv \mathit{Similarity}(\boldsymbol{w}, \boldsymbol{w}_{a}) = \boldsymbol{w} \cdot \boldsymbol{w}_{a} \tag{4}$$

Truncate

Oracle Selection

Ideally, when the wrongly annotated instances are known beforehand. The channel discriminative ability can be measured by the within-/ between-class variance.

$$\psi_o(\Omega) = \frac{(\boldsymbol{\mu}_c - \boldsymbol{\mu}_u)^2}{\boldsymbol{\sigma}_c^2 + \boldsymbol{\sigma}_u^2}, \quad \boldsymbol{\mu}_c = \frac{1}{|\Omega_c|} \sum_{\boldsymbol{x}_i \in \Omega_c} \boldsymbol{x}_i, \quad \boldsymbol{\sigma}_c^2 = \frac{1}{|\Omega_c|} \sum_{\boldsymbol{x}_i \in \Omega_c} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_c\right)^2, \tag{5}$$

Truncate

Approximate Selection

A video, in essence, consists of both the scene and motion semantics [WH18][5]. We simply utilize the temporal *average* and *variance* to roughly search the semantics-intense channels.

$$\phi_{ave}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T) = \sum_t \boldsymbol{v}_t / T, \quad \phi_{var}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T) = \sum_t (\boldsymbol{v}_t - \phi_{ave})^2 / T. \tag{6}$$

[5] Y. Wang and M. Hoai, "Pulling actions out of context: Explicit separation for effective combination," in *CVPR*, 2018, pp. 7044–7053.

Truncate

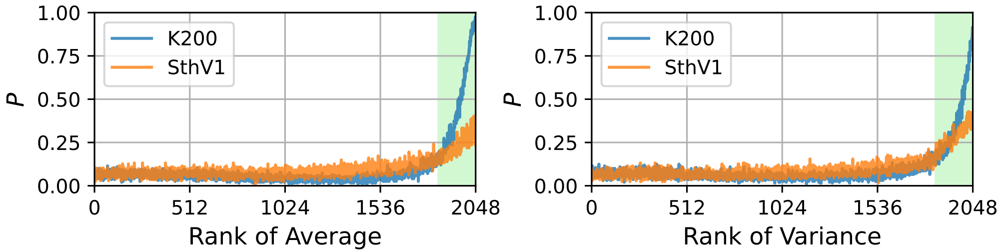

Figure 2: The statistical relation between oracle score and amplitudes/variance on K200 and SthV1 under symmetric-40% noise setting at fifth epoch. Each recorded point bears the coordinates (r, p). r is the ranking of the corresponding statistics. The higher the r, the larger the amplitudes/variance. p is the probability of the relevant channels picked by the top b *oracle selection*. The top b amplitudes/variance area is filled with green. $(b = 200)$

Split

A two-component Gaussian Mixture Model $\pi(\cdot)$ is utilized to fit the distribution of $\eta(\boldsymbol{w})$. The probability of $\boldsymbol{w}$ being noise of the $\boldsymbol{a}$-th category is then defined as $p(\text{noise}|\boldsymbol{w}; a) = \pi_a(\boldsymbol{w})$. Hence, the estimated clean set is obtained by thresholding $\pi_a(\boldsymbol{w})$,

$$\mathcal{D}_{\hat{c}} = \epsilon(\mathcal{D}) = \{\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{D},\ \pi_a(\boldsymbol{w}) < 0.5\}. \quad (7)$$

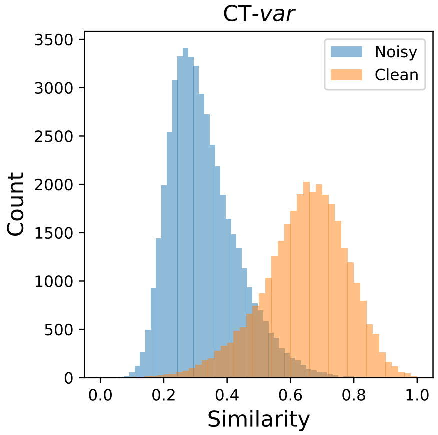

Figure 3: Similarity distribution of CT-*var*.

Contrast

Noise Contrastive Learning

The estimated clean and unlabelled splits, namely $\mathcal{D}_{\hat{c}}$ and $\mathcal{D}_{\hat{u}}$

The motivation of Noise Contrastive Learning (NCL) is to utilize the estimated noisy samples in model updating fully and further enlarge the margins among samples from different categories.

$$\mathcal{L}_r = -\alpha \sum_{\mathbf{x}_q \in \mathcal{D}} \beta_q \sum_{\mathbf{x}_p \in \mathcal{P}_q} \log \frac{\exp\left(\frac{\mathbf{z}_q \cdot \mathbf{z}_p}{\tau}\right)}{\sum_{\mathbf{x}_j \in \mathcal{P}_q \cup \mathcal{N}_q} \exp\left(\frac{\mathbf{z}_q \cdot \mathbf{z}_j}{\tau}\right)}, \tag{8}$$

Contrast

Given a clip pair $(\boldsymbol{G}, \widetilde{\boldsymbol{G}})$ from a video $\boldsymbol{V}$, the representations of the two clips are defined as $\Gamma = \{\boldsymbol{x}^{G}, \boldsymbol{x}^{\widetilde{G}}\}$. For a sampled clip $\boldsymbol{G}$ from *clean* cluster of category $\boldsymbol{a}$, the sets of positive and negative keys $\mathcal{P}_{\hat{c}}$, $\mathcal{N}_{\hat{c}}$ can be defined as,

$$\begin{aligned}
\mathcal{P}_{\hat{c}} =& \{\Gamma_j \mid \Gamma_j \subset \mathcal{D}_{\hat{c}}, \boldsymbol{i} \neq \boldsymbol{j}, \boldsymbol{a}_i = \boldsymbol{a}_j\} \cup \{\boldsymbol{x}_i^{\widetilde{G}}\}, \\
\mathcal{N}_{\hat{c}} =& \{\Gamma_j, \Gamma_l \mid \Gamma_j \subset \mathcal{D}_{\hat{u}}, \Gamma_l \subset \mathcal{D}_{\hat{c}}, \boldsymbol{a}_i = \boldsymbol{a}_j, \boldsymbol{a}_i \neq \boldsymbol{a}_l\},
\end{aligned} \tag{9}$$

When the sampled clip $\boldsymbol{G}_i$ is from *unlabelled* cluster, the sets of keys $\mathcal{P}_{\hat{u}}$ and $\mathcal{N}_{\hat{u}}$ is defined as,

$$\mathcal{P}_{\hat{u}} = \{\boldsymbol{x}_j \mid \boldsymbol{x}_j \in \mathrm{NN}(\boldsymbol{x}_i^{G})\} \cup \{\boldsymbol{x}_i^{\widetilde{G}}\}, \quad \mathcal{N}_{\hat{u}} = \mathcal{D} - \mathcal{P}_{\hat{u}} \cup \{\boldsymbol{x}_i^{G}\}. \tag{10}$$

Experiment

Noise Type	Symmetric				Asymmetric			Average
Noise Ratio	20%	40%	60%	80%	10%	20%	40%	
GCE[40]	53.1	49.6	42.1	23.4	54.0	52.0	41.4	45.5
SCE[32]	64.5	57.8	48.1	27.9	67.2	62.0	46.9	53.5
TopoFilter[34]	61.4	55.5	37.7	14.9	65.7	63.6	55.5	50.6
Co-teaching[9]	61.0	60.9	56.9	32.5	60.6	60.2	46.9	54.1
M-correction[1]	66.7	62.3	54.8	40.1	65.5	62.1	52.9	57.8
CT-*all*	68.4	64.2	58.3	43.3	69.1	67.4	51.9	60.4
CT-*var*	69.2	**67.1**	**61.1**	**48.4**	**70.0**	68.0	55.9	62.8
CT-*ave*	**69.4**	66.9	61.0	48.1	69.8	**68.6**	**58.4**	**63.2**
CT-*oracle*	70.4	67.7	61.5	49.6	70.6	70.0	58.9	64.1
Clean Only	70.5	68.3	64.9	58.8	70.9	70.3	68.6	67.5

Figure 4: Results on K200 (Part 1)

Experiment

DivideMix[16]	69.4	65.9	60.7	46.5	68.2	67.5	53.1	61.6
CT-*var* + PL-H	68.1	65.6	58.3	37.9	66.5	65.8	54.6	59.6
CT-*var* + PL-S	68.2	66.0	60.8	45.4	67.1	66.8	55.8	61.4
CT-*var* + PL-K	67.7	65.0	60.1	47.4	66.4	65.7	55.0	61.0
CT-*var* + CL	69.4	66.6	60.4	12.2	68.3	67.6	54.7	57.0
CT-*var* + SCL	70.3	67.6	61.4	46.9	70.2	68.1	57.0	63.1
CT-*var* + NCL	**70.9**	**68.6**	**63.4**	**49.9**	**70.5**	**69.5**	**59.2**	**64.6**

Figure 5: Results on K200 (Part 2)

Experiment

Table 2. Testing Accuracies (%) on Kinetics and Something V1 Dataset.

Dataset	Kinetics						Something V1					
Noise Type	Symmetric			Asymmetric			Symmetric			Asymmetric		
Noise Ratio	40%	60%	80%	10%	20%	40%	40%	60%	80%	10%	20%	40%
TopoFilter[34]	49.1	40.4	21.5	56.1	55.1	47.8	21.4	4.3	1.2	35.8	34.5	26.9
Co-teaching[9]	53.9	51.4	31.3	51.4	51.4	41.7	23.6	14.5	4.5	24.0	24.6	18.5
M-correction[1]	57.3	51.2	40.3	54.8	53.9	47.1	24.1	15.1	4.5	36.6	35.2	22.0
CT-*ave*	**60.3**	**56.6**	**46.3**	**62.9**	**61.3**	48.0	**36.3**	26.2	4.6	41.5	37.8	28.3
CT-*var*	60.1	55.8	45.7	62.8	61.0	**48.4**	36.2	**26.7**	**4.8**	**41.6**	**38.7**	**28.7**
Clean Only	61.6	58.8	54.3	70.3	68.0	61.7	40.7	36.0	24.8	44.0	43.5	40.6
CT*-*var*	61.1	56.7	48.6	63.6	62.5	57.8	36.8	27.0	5.8	41.7	40.2	32.8
CT-*var*+CL	60.0	55.3	7.6	61.7	59.5	45.2	31.8	1.6	1.1	37.8	36.2	27.1
CT-*var*+SCL	60.5	56.5	46.5	63.0	60.9	48.8	37.6	28.4	2.9	41.5	40.2	29.4
CT-*var*+NCL	**61.2**	**57.2**	**46.9**	**63.3**	**61.5**	**49.1**	**38.3**	**30.1**	**5.1**	**41.8**	**40.8**	**30.0**

Experiment

Table 1: Testing accuracies (%) on K200 with 60% Symmetric Noise and Different Hyper-parameter b.

b	100	200	400	1600	2048
CT-*var*	60.6	**61.1**	60.8	59.0	58.3
CT-*var*+NCL	62.9	**63.4**	63.3	62.0	61.6

Experiment

Noise Type	Symmetric				Asymmetric	
Noise Ratio	20%	40%	60%	80%	20%	40%
Bootstrap[27]	51.4	41.1	29.7	10.2	53.4	38.7
D2L[22]	54.0	29.7	-	-	43.6	16.9
L_{DMI}[35]	-	-	-	-	-	-
Data Param[28]	56.3	46.1	32.8	11.9	56.2	39.0
M-correction[1]	53.0	43.0	36.6	12.8	53.2	37.9
INCV[3]	58.6	55.4	43.7	23.7	56.8	44.4
AUM[26]	65.5	61.3	53.0	31.7	59.7	40.2
DivideMix†[16]	65.4	62.2	**62.4**	35.1	65.7	51.7
DivideMix[16]	67.6	**65.8**	**63.7**	**43.4**	**67.1**	**54.2**
CT-*img*	67.7	61.6	55.1	31.7	61.0	47.8
CT-*img* + PL-H	67.4	62.1	53.6	27.2	65.0	51.1
CT-*img* + PL-S	**68.2**	63.0	54.3	31.7	66.1	52.7
CT-*img* + PL-K	67.7	62.7	53.6	31.7	66.6	53.5
CT-*img* + NCL	**68.5**	**64.9**	57.5	**35.4**	**67.4**	**58.5**

Figure 6: Results on CIFAR100

Reference I

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *ICLR*, 2017.

[2] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *ICCV*, Oct. 2019.

[3] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. Mcguinness, "Unsupervised label noise modeling and loss correction," in *ICML*, 2019.

[4] K.-H. Lee, X. He, L. Zhang, and L. Yang, "Cleannet: Transfer learning for scalable image classifier training with label noise," in *CVPR*, 2018.

[5] Y. Wang and M. Hoai, "Pulling actions out of context: Explicit separation for effective combination," in *CVPR*, 2018, pp. 7044–7053.