Truncate-Split-Contrast: A Framework for Learning from Mislabeled Videos

Zixiao Wang, Tsinghua University
Junwu Weng, Tencent AI Lab
Chun Yuan, Tsinghua University
Jue Wang, Tencent AI Lab

Abstract

Learning with noisy label (LNL) is a classic problem that has been extensively studied for image tasks, but much less for video in the literature. A straightforward migration from images to videos without considering the properties of videos, such as computational cost and redundant information, is not a sound choice. In this paper, we propose two new strategies for video analysis with noisy labels: 1) A lightweight channel selection method dubbed as Channel Truncation for feature-based label noise detection. This method selects the most discriminative channels to split clean and noisy in- stances in each category; 2) A novel contrastive strategy dubbed as Noise Contrastive Learning, which constructs the relationship between clean and noisy instances to regularize model training. Experiments on three well-known bench- mark datasets for video classification show that our proposed truNcatE-split-contrAsT (NEAT) significantly outperforms the existing baselines. By reducing the dimension to 10% of it, our method achieves over 0.4 noise detection F1-score and 5% classification accuracy improvement on Mini-Kinetics dataset under severe noise (symmetric-80%). Thanks to Noise Contrastive Learning, the average classification accuracy improvement on Mini-Kinetics and Sth-Sth-V1 is over 1.6%.

The Pipeline of Our Method

The framework of truNcatE-split-contrAsT (NEAT) is shown in Fig. 1. Each video feature is first *truncated* for clean/noisy instance *splitting*. Then, the detected clean/noisy instances are utilized separately under the supervision of cross entropy and noise *contastive* loss for model updating. Our main contributions are summarized as follows:

- A light weight channel selection method for feature-based label noise detection is proposed. It discards the redundant channels to increase the effectiveness and efficiency of noisy/clean instance splitting.
- Noise Contrastive Loss is designed to construct the relationship among instances by referring the estimated clean/noisy splits, and utilizes this relationship to learn visual representations without involving wrong labels.
- To the best of our knowledge, this is the first efficient framework for LNL in video analysis. Extensive experiments show the effectiveness of our method on several video recognition datasets with noisy label settings.

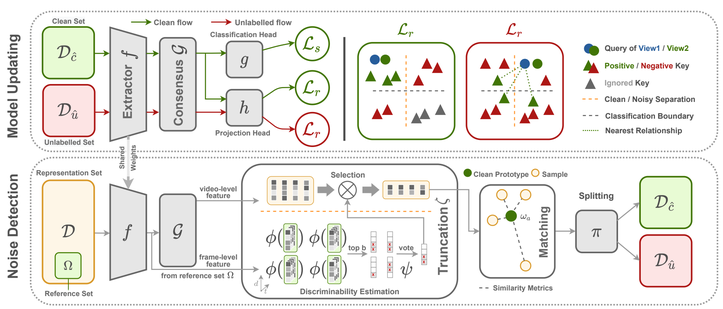

Figure 1: The pipeline of our framework NEAT on the noisy dataset in training.

Noise Detection Phase: Truncate-Split

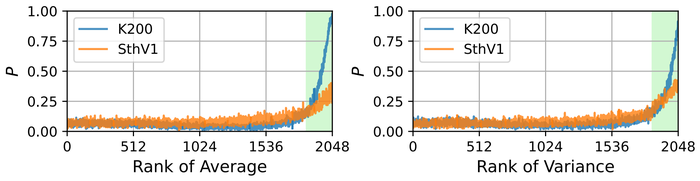

Figure 2: The statistical relation between oracle score and amplitudes/variance on K200 and SthV1.

Truncate. Generally, feature-based noise detection methods detect noisy labels by computing the similarity between the query x and the clean-prototype of this category. The higher the similarity, the more likely the query x to be clean. As full channels are learned to differentiate multiple categories, they are unnecessary for a much simpler task, i.e. differentiating clean/noisy instances in each category. We use a score function to select discriminative channels.

Ideally, when the true splits of correctly and wrongly annotated instances are known beforehand, we have Oracle Score Function. The channel discriminative ability can be measured by the within-/~between-class variance.

$$\psi_o(\Omega) = \frac{(\mu_c - \mu_u)^2}{\sigma_c^2 + \sigma_u^2}, \quad \mu_c = \frac{1}{|\Omega_c|}\sum_{\mathbf{x}_i \in \Omega_c} \mathbf{x}_i, \quad \sigma_c^2 = \frac{1}{|\Omega_c|}\sum_{\mathbf{x}_i \in \Omega_c} (\mathbf{x}_i - \mu_c)^2,$$

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

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

The relationship between oracle and approximation score can be found in Fig. 2.

Split. We observe that the similarity distribution of the clean and noisy instances gradually becomes a two-peak form during training. A two-component Gaussian Mixture Model is utilized to fit the distribution of similarity in truncated feature space. The estimated clean set is obtained by thresholding post-probability. we fix it as 0.5 in the experiments.

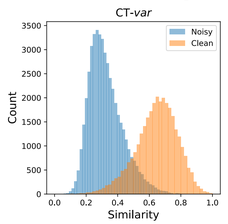

Figure 3: Similarity distribution.

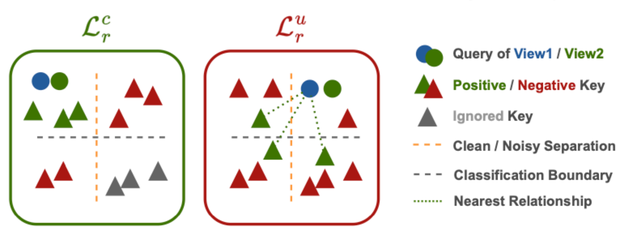

Figure 4: An illustration of Noise Contrastive Learning.

Model Updating Phase: Contrast

Contrast. 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.
We have different strategies for clean queries and noisy queries. If query x is from a clean video, the positive keys are from the clean samples with the same label as x, and the negative keys have two parts. The first part is those noisy samples with the same label with x, and the second part is those clean samples with a different label with x. those noisy samples with different labels with x are ignored. If query x is from a noisy video, the positive keys are searched by the nearest neighborhood algorithm and another view from the same video. All other samples are treated as negative keys. An illustration can be found in Fig. 4.

Experiments

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(2020)	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(2018)	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(2019)	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**

Table 1: Test Accuracies on K400 and Something-Something-V1.

b	100	200	400	1600	2048 (all)
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

Table 2: Ablation study on the number of kept channels. b=2048 means all channels are kept.

Visualization