Supplementary Material for CNVid-3.5M: Build, Filter, and Pre-train the Large-scale Public Chinese Video-text Dataset

Tian Gan[1]*, Qing Wang[2]*, Xingning Dong[1], Xiangyuan Ren[2], Liqiang Nie[3], Qingpei Guo[2]†
[1]Shandong University, [2]Ant Group, [3]Harbin Institute of Technology (Shenzhen)
gantian@sdu.edu.cn, wq176625@antgroup.com, dongxingning1998@gmail.com
xiangyuan.rxy@antgroup.com, nieliqiang@gmail.com, qingpei.gqp@antgroup.com

1. Introduction

This supplementary material contains the following four components: 1) In Sec.2, we discuss the potential influence of "good", "hard", and "noisy" samples on the pre-training. 2) In Sec.3, we present additional statistics of our CNVid-3.5M dataset. 3) In Sec.4, we detail the image-text model pre-trained on the Wukong dataset for weakly-paired data filtering. And 4) in Sec.5, we present the quantitative analyses of Positive Column Weighting (PCW) and Negative Line Sampling (NLS) in the proposed Hard Sample Curriculum Learning (HSCL) strategy.

2. "Good", "Hard", and "Noisy" Samples

In our work, we divide all video-text pairs into three categories, *i.e.*, "good" samples, "hard" samples, and "noisy" ones. This classification is based on the video-text consistency. As an example shown in Figure 1, for the same music video (MV), a "good" sample would describe the video content clearly and accurately. For a "hard" sample, there exists both useful information related to the given video (*e.g.*, "music video") and ineffective one that is redundant (*e.g.*, "click rate"). In comparison, a "noisy" sample describes things that nearly have no correlation with the video (*e.g.*, lyrics of this music video).

Based on the rough classification of "good", "hard", and "noisy" samples, we have three conjectures about their potential influence on the video-text pre-training as follows: 1) "Noisy" samples should be filtered out since they would degrade the pre-training performance. As the text in "noisy" samples is not in line with the associated video, it would confuse the pre-trained model and hinder the cross-modal alignment. 2) When the model is under convergence, "good" samples would benefit the pre-training more compared with "hard" ones. Since "good" samples are easy to learn, they would accelerate the optimization procedure. 3) When the model is close to convergence, emphasizing "hard" samples would further promote the pre-training performance. At this time, the model is capable of discriminating "good" ground-truth video-text pairs. Therefore, gradually turning to learn "hard" samples would increase the robustness of the pre-trained model.

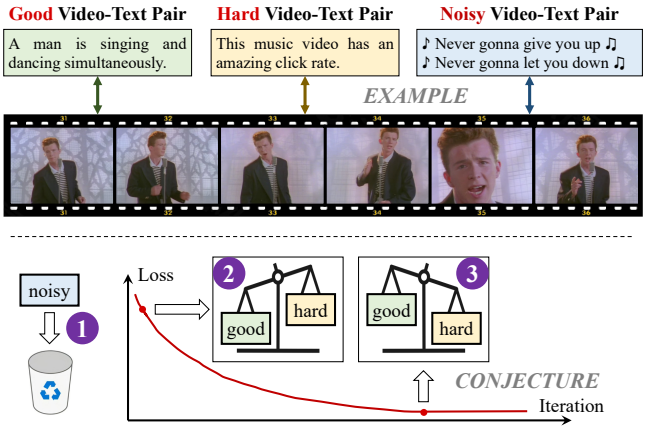

Figure 1. An example of "good", "hard", and "noisy" video-text samples (the top part of this figure). Based on which we come up with three conjectures (the bottom part of this figure).

The innovations of our methods are closely related to the above three conjectures. The proposed weakly-paired data filtering strategy is motivated by conjecture 1), which aims to remove "noisy" samples for better pre-training performance. Inspired by conjectures 2) and 3), we propose the Hard Sample Curriculum Learning strategy to gradually and smoothly emphasize those "hard" samples for better contrastive learning. Experiments in the main paper prove the effectiveness of our conjectures and proposed methods.

3. Additional Statistics of CNVid-3.5M

In this section, we present additional statistics of our CNVid-3.5M dataset, including the distributions of topics, keywords, video durations, and Part-of-Speech (POS) tags

*Equal contribution.
†Corresponding author.

No.	Pre-training Dataset	PCW	NLS	CL	VATEX	MSRVTT	DiDemo
					R@1/5/10 ↑ (MdR ↓)	R@1/5/10 ↑ (MdR ↓)	R@1/5/10 ↑ (MdR ↓)
Pre-training and Fine-tuning on CHINESE Video-text Dataset.							
F1	CNVid-3.5M	-	-	-	39.9 / 77.2 / 87.0 (2)	20.7 / 47.9 / 61.2 (6)	13.3 / 34.6 / 45.9 (13)
F2	CNVid-3.5M	-	√	√	40.7 / 75.9 / 86.5 (2)	22.8 / **48.9** / 60.5 (6)	13.3 / **35.0** / 47.2 (12)
F3	CNVid-3.5M	√	-	√	40.1 / 77.0 / 86.9 (2)	21.7 / 47.5 / 60.8 (6)	12.9 / 34.3 / 47.3 (12)
F4	CNVid-3.5M	√	√	-	40.7 / 76.8 / 86.9 (2)	21.1 / 47.3 / 60.1 (6)	13.4 / 34.3 / 46.4 (12)
F5	CNVid-3.5M	√	√	√	**41.5** / **78.2** / **87.2** (2)	**23.3** / 48.0 / **61.2** (6)	**13.6** / 34.4 / **47.3** (**12**)
Pre-training and Fine-tuning on ENGLISH Video-text Dataset.							
G1	COCO+VG+CC	-	-	-	45.5 / 81.5 / 91.0 (2)	28.6 / 55.5 / 66.1 (4)	25.3 / 50.3 / 62.4 (5)
G2	COCO+VG+CC	-	√	√	47.9 / 83.0 / 90.9 (2)	30.2 / 55.7 / **67.6** (4)	26.4 / 52.9 / **63.9** (4)
G3	COCO+VG+CC	√	-	√	47.6 / 82.8 / 90.9 (2)	29.6 / 55.9 / 65.9 (4)	24.9 / 51.2 / 63.2 (5)
G4	COCO+VG+CC	√	√	-	48.4 / **83.9** / 91.1 (2)	29.7 / 55.3 / 65.1 (4)	25.2 / 51.3 / 63.7 (5)
G5	COCO+VG+CC	√	√	√	**48.7** / 83.1 / **91.1** (2)	**31.6** / **56.5** / 66.7 (4)	**27.1** / **53.8** / 63.8 (**4**)

Table 1. Detailed ablation study of the proposed Hard Sample Curriculum Learning (HSCL) strategy. As HSCL contains three key components, we successively remove one of these three components to verify its effectiveness in promoting the video-text pre-training.

of the ASR text. We also discuss fairness and privacy for safety and ethics.

Topics: CNVid-3.5M contains a total of 5.7M topics, and each video has 1.63 topics on average. There are more than 58.8K categories of topics with an average of 9.73 samples. The maximum topic has 53,059 samples. To achieve an intuitive perception, we build the word cloud of Top-200 topics in Figure 2.

Keywords: CNVid-3.5M contains a total of 21.5M keywords, and each video has 6.15 keywords on average. There are more than 48.6K categories of keywords with an average of 44.34 samples. The maximum keyword has 189,246 samples. To achieve an intuitive perception, we build the word cloud of Top-200 keywords in Figure 3.

Video Durations: The maximum, minimum, and average video duration of CNVid-3.5M is 1,974s, 4s, and 36.34s, respectively. The total duration of all videos in CNVid-3.5M lasts for 1,475 days. The mode of duration is 17s with a number of 676,690 videos. We visualize the statistics of videos whose duration ranges from 5s to 60s in Figure 4.

Part-of-Speech Tags of the ASR Text: We employ an open-source NLP toolkit LAC * to obtain the Part-of-Speech (POS) tags of all the ASR Text in CNVid-3.5M. CNVid-3.5M contains a total of 17.0M verbs, 9.1M nouns, and 3.2M adjectives, respectively. There are 3,524, 3,567, and 1,107 categories of verbs, nouns, and adjectives with an average of 4.8K, 2.5K, and 2.9K samples, respectively. The maximum verb, noun, and adjective have 810K, 228K, and 148K samples, respectively. We visualize the statistics of Top-50 verbs, nouns, and adjectives in Figure 5, 6, and 7.

Fairness: We make sure that all data collected and shared is done in an ethical and responsible manner. Under the strict censorship on China's content platform, we ensure our dataset avoids sensitive words, improper video content, or any prejudicial information. Our dataset is **for research use only**, and **any commercial usage will be permitted.**

Privacy: Only the video id is released to the public. With our provided ID, researchers can build direct links to the source website, avoiding copyright issues. If a user deletes a video from those platforms, it becomes removed from our dataset as well. This way content creators can reserve their right to refuse the use of their videos.

4. Settings for Pre-training On Wukong

The image-text model employed for weakly-paired data filtering is pre-trained on the Wukong (100M image-text pairs) dataset. We adopt the two-encoder-fusion (*2-E-F*) paradigm as the basic pre-training architecture. For proxy tasks, We set the conventional MLM and VTM task. For hyper-parameters, the model is pre-trained by the Adam optimizer with a momentum of 0.9. The total pre-training stage lasts for 50,000 iterations with a batch size of 512. The initial learning rate is 5e-5. The whole pre-training takes about 5 days to complete on 32 NVIDIA V100 GPUs.

5. Detailed Ablation Study of HSCL

As mentioned in the main paper, the proposed Hard Sample Curriculum Learning (HSCL) strategy contains the following three key components: 1) **Positive Column Weighting** (PCW) to emphasize hard positive samples by re-weighting, 2) **Negative Line Sampling** (NLS) to better leverage hard negative examples by re-sampling, and 3) **Curriculum Learning** (CL) to smoothly and gradually lead the pre-trained model to perform the PCW and NLS calculation. We conduct several ablation study on both Chinese and English pre-training datasets to verify the effectiveness of these three components. As illustrated in Table 1, in general, the performance would decrease when removing any component of the proposed HSCL strategy, which verifies the effectiveness of PCW, NLS, and CL.

*Official Website: https://github.com/baidu/lac

Figure 2. The word cloud generated with Top-200 TOPICs in the CNVid-3.5M dataset.

Figure 3. The word cloud generated with Top-200 KEYWORDs in the CNVid-3.5M dataset.

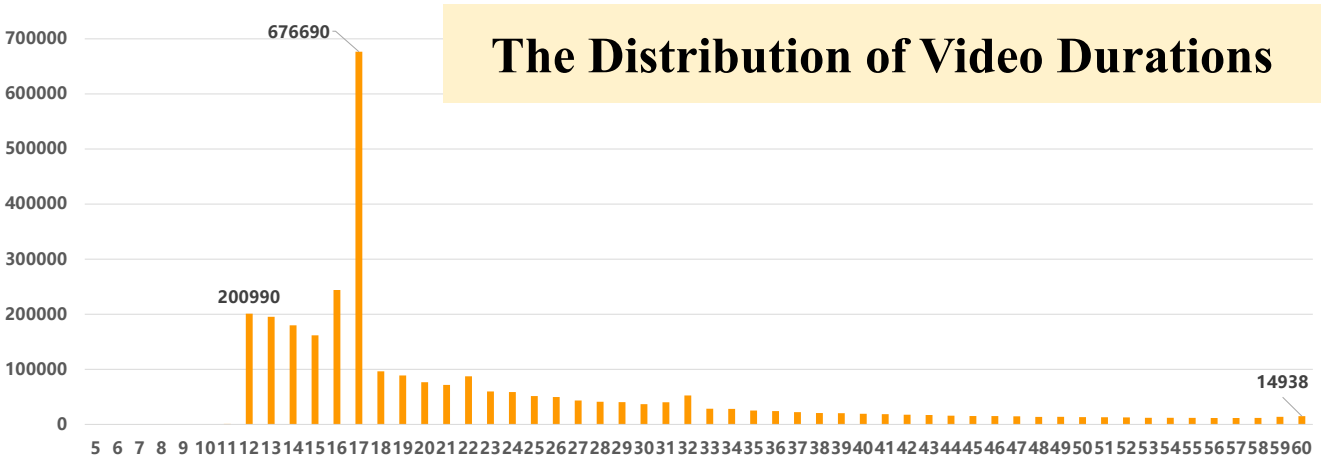

Figure 4. Statistics of videos whose duration ranges from 5(s) to 60(s).

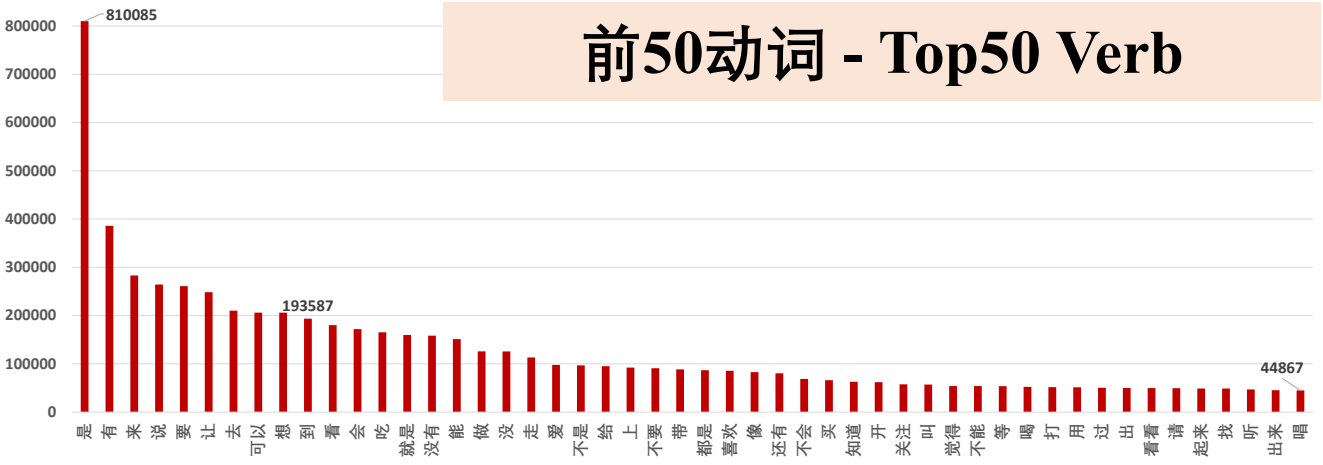

Figure 5. Statistics of Top-50 VERBs in the CNVid-3.5M dataset.

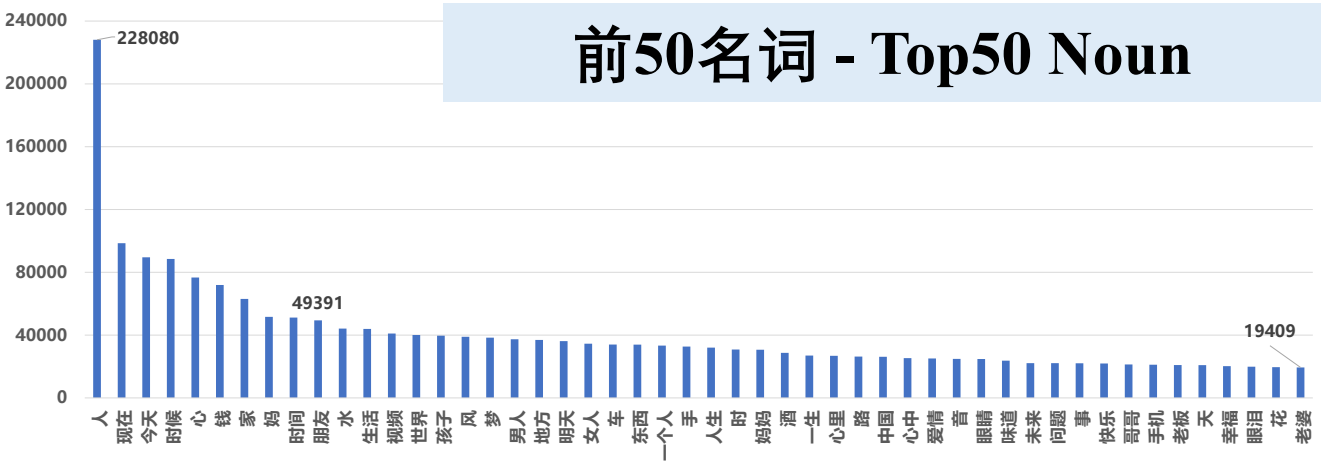

Figure 6. Statistics of Top-50 NOUNs in the CNVid-3.5M dataset.

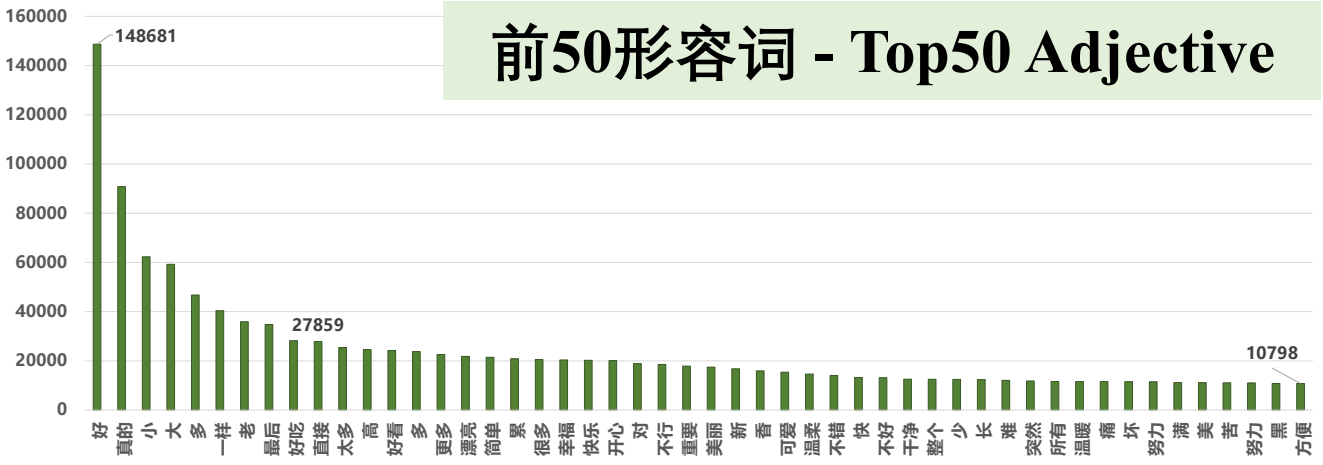

Figure 7. Statistics of Top-50 ADJECTIVEs in the CNVid-3.5M dataset.