Learning from Weakly-Labeled Web Videos via Exploring Sub-concepts

Kunpeng Li[1*], Zizhao Zhang[2], Guanhang Wu[2], Xuehan Xiong[2], Chen-Yu Lee[2], Zhichao Lu[2], Yun Fu[1], Tomas Pfister [2]

[1] Northeastern University [2] Google Cloud AI

Abstract

Learning visual knowledge from massive web videos has attracted growing research interest thanks to the large corpus of easily accessible video data on the Internet. However, for video action recognition, the action of interest might only exist in arbitrary clips of untrimmed web videos, resulting in high label noises in the temporal space. To address this issue, we introduce a new method for pre-training video action recognition models using queried web videos. Instead of trying to filter out potential noises, we propose to provide fine-grained supervision signals by defining the concept of Sub-Pseudo Label (SPL). Specifically, SPL spans out a new set of meaningful "middle ground" label space constructed by extrapolating the original weak labels during video querying and the prior knowledge distilled from a teacher model. Consequently, SPL provides enriched supervision for video models to learn better representations and improves data utilization efficiency of untrimmed videos. We validate the effectiveness of our method on four video action recognition datasets and a weakly-labeled image dataset to study the generalization ability. Experiments show that SPL outperforms several existing pre-training strategies and the learned representations lead to competitive results on several benchmarks.

Introduction

Remarkable successes (Feichtenhofer et al. 2019) have been achieved in video recognition in recent years thanks to the development of deep learning models. However, training deep neural networks requires a large amount of human-annotated data. It requires tremendous human labor and huge financial cost and therefore oftentimes sets out to be the bottleneck for real-world video recognition applications.

Web videos are usually acquired by online querying through label keywords. A keyword, which we refer as a weak label, is then assigned to each untrimmed video obtained. Although large-scale videos with weak labels are easier to be collected, training with un-verified weak labels poses another challenge in developing robust models. Recent studies (Ghadiyaram, Tran, and Mahajan 2019; Kuehne et al. 2019; Chang et al. 2019) have demonstrated that, in addition to the label noise (e.g. incorrect action labels on untrimmed videos), there is temporal noise due to the lack of accurate temporal action localization. This means an untrimmed web video may include other non-targeted content or only contain a small proportion of the target action. Reducing noise effects for large-scale weakly-supervised pre-training is critical but particularly challenging in practice. (Ghadiyaram, Tran, and Mahajan 2019) suggests querying short videos (e.g., within 1 minute) to obtain more accurate temporal localization of target actions. However, such data pre-processing methods prevent models from fully utilizing widely available web video data, especially longer videos with richer contents. (Duan et al. 2020) applies a teacher model to do filtering before pre-training, but it will remove a large proportion of data that is potentially useful.

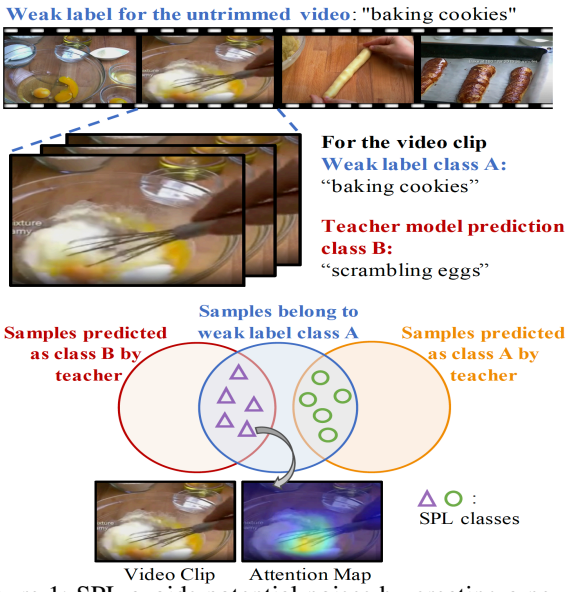

Figure 1: SPL avoids potential noises by creating a new set of "middle ground" pseudo labels (i.e. sub-concepts) via extrapolating two related action classes. Enriched supervision is provided for effective model pre-training.

In this work, we propose a novel learning method to conduct effective pre-training on untrimmed videos from the web. Instead of simply filtering the potential temporal noise, we propose to convert such "noisy" data to useful supervision by leveraging the proposed concept of Sub-Pseudo Label (SPL). As shown in Figure 1, SPL creates a new set

[*]Work done while the author was an intern at Google; now at Meta Reality Labs. Email: kunpengli@fb.com

of meaningful "middle ground" pseudo-labels to expand the original weak label space. Our motivation is based on the observation that, within the same untrimmed video, these "noisy" video clips have semantic relations with the target action (weak label class), but may also include essential visual components of other actions (such as the teacher model predicted class). Our method aims to use the extrapolated SPLs from weak labels and distilled labels to capture the enriched supervision signals, encouraging learning better representations during pre-training that can be used for downstream fine-tuning tasks.

Discovering meaningful SPLs has critical impact on learning high quality representations. To this end, we take advantage of the original weak labels as well as the prior knowledge distilled from a set of target labeled data from human annotations. Specifically, we first train a teacher model from the target labeled data and perform inference on every clip of web videos. From the teacher model predicted labels and the original weak labels of the web video, we design a mapping function to construct SPLs for these video clips. We then perform large-scale pre-training on web videos utilizing SPLs as the supervision space. In addition, we study different space reduction strategies for SPL to tackle high-dimensional label space when the number of classes is high. We construct weakly-labeled web video data based on two video datasets: Kinetics-200 (Xie et al. 2018) and SoccerNet (Giancola et al. 2018). Experimental results show that our method can consistently improve the performance of conventional supervised methods and several existing pre-training strategies on these two datasets. We also follow recent works (Miech et al. 2020; Stroud et al. 2020) to conduct experiments on HMDB-51 (Kuehne et al. 2011) and UCF-101 (Soomro, Zamir, and Shah 2012) to show that learned representations by SPL are generic and useful when SPL pre-training is conducted on web data from different domains. SPL achieves competitive results on these datasets compared with recent state-of-the-art pre-training methods.

In summary, our contributions can be concluded as follows: (a) We propose a novel concept of SPL for learning from weakly-labeled web videos. It efficiently avoids the potential noise in data by creating meaningful sub pseudo-labels, which provides enriched supervision and improves data utilization efficiency of untrimmed videos. (b) We investigate several space reduction strategies for SPL. utilizing weak labels as well as the knowledge distilled from the teacher model trained on the labeled dataset. (c) Comprehensive experiments on multiple video datasets show that our method can consistently improve the performance of baselines on both common and fine-grained action recognition datasets. (d) SPL maintains high originality and shows good generalization ability to the weakly-labeled image classification task. We believe it has potentials to impact more tasks where there exists uncertainty in labels.

Related Work

Learning from the web data. There are growing studies taking use of information from the web that aim to reduce the cost of data collection and annotations (Mahajan et al. 2018). For video classification, early works (Gan et al. 2016a) focus on utilizing web action images to boost action recognition models. However these image-based methods do not consider the rich temporal dynamics of videos. Interactions between web videos and images are further studied in (Gan et al. 2016b; Rupprecht et al. 2018), where a CNN network trained on web videos is refined using web images in a curriculum learning manner. (Ghadiyaram, Tran, and Mahajan 2019) recently shows that better pre-training models can be obtained by learning from very large scale noisy web videos with short length e.g., within 1 minute. (Duan et al. 2020) explores pre-training using multiple sources of web data. A teacher model is further applied to conduct filtering on the collected web data to reduce noise. Differently, SPL handles the temporal noise in untrimmed videos by exploring valid sub-concepts, so that enriched supervision can be provided for effective pre-training.

Knowledge distillation. Our work is also related to Knowledge Distillation (Hinton, Vinyals, and Dean 2015; Yan et al. 2020) where the goal is to transfer knowledge from one model (the teacher) to another (the student). Several recent distillation methods (Xie et al. 2020) focus on teachers and students with the same architecture which can be trained sequentially for image classification. (Müller, Kornblith, and Hinton 2020) proposes to improve the transfer by forcing the teacher to expand the network output logits during the supervised training. The student is then trained to match the subclass probabilities. While to some extend SPL and it share similar high-level motivation of exploring sub-concepts, the differences are significant. In addition to focusing on different problems, methods are also different: our method uses the extrapolated SPLs from weak labels and distilled labels, while (Müller, Kornblith, and Hinton 2020) learns to expend teacher network logits, which relies on a pre-defined hyper-parameter of subclass numbers.

Video recognition. Video analysis has long been tackled using hand-crafted features (Laptev 2005). Following the success of deep learning in the image domain (He, Zhang, and Sun 2016), State-of-the-art action recognition models either use two-stream (RGB and flow) networks (Simonyan and Zisserman 2014) or 3D ConvNets (Carreira and Zisserman 2017; Tran et al. 2015; Feichtenhofer et al. 2019). Our work aims to improve the performance of action recognition models by utilizing weakly-labeled web videos.

Method

Problem Formulation and Method Overview

For pre-training on a dataset D_p with N target actions, we aim to learn representations that can benefit the downstream task by afterwards fine-tuning on the target dataset D_t. This pre-training process of model M is usually achieved by minimizing the cross-entropy loss between the data samples x and their corresponding labels y, as follows:

$$L_{\text{CE}} = -\mathbb{E}_{(x,y)\sim D_p} \sum_{c=1}^{N} y_c \log M(x), \tag{1}$$

where $y_c \in \{0, 1\}$ indicates whether x belongs to class $c \in [0, N-1]$.

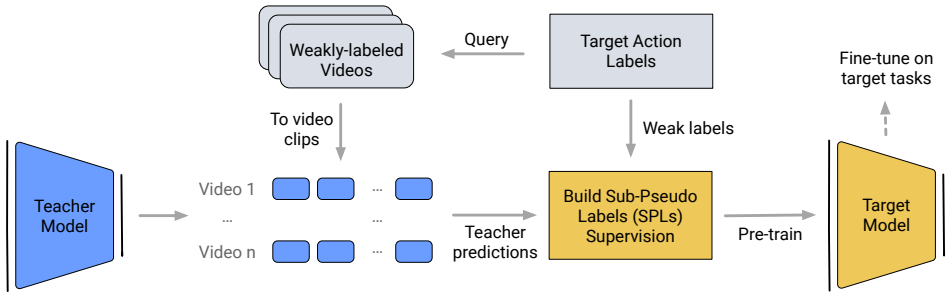

Figure 2: The overall pre-training framework for learning from web videos via exploring SPLs.

In the case of pre-training on a web video set as D_p, we sample clips from these untrimmed web videos to construct the training data. Since there are no ground-truth annotations, assigning a valid label y for each clip sample x is a key. A common practice (Ghadiyaram, Tran, and Mahajan 2019; Mahajan et al. 2018) is to treat the text query or hash tags that come together with the web videos as weak labels l. However this causes high label and temporal noises as target actions might exist in arbitrary clips of the entire video that occupy a very small portion. In addition to relying on the weak labels, we can also distill knowledge from a teacher model T trained on the target dataset D_t using Eq. 1, where D_p is replaced by D_t. A basic teacher-student training pipeline (Furlanello et al. 2018; Xie et al. 2020) can be applied by treating the teacher model prediction as the pseudo-label to train a student model on D_p. But there will be information lost as the original informative weak labels are totally ignored. Another strategy is to use agreement filtering to select reliable data whose weak labels match their teacher model predictions. However, in practice we find this strategy will discard a large amount of training data from D_p (over 60% in our experiments on the Kinetics-200 dataset), which limits the data scale for training deep neural networks.

Instead of treating the potential noise in D_p as useless data to filter out, we propose to migrate such noisy data to useful supervision by defining the concept of Sub-Pseudo Label (SPL). Specifically, SPL creates a new set of meaningful "middle ground" pseudo-labels, which are discovered by taking advantage of the original weak labels and the prior knowledge distilled from the teacher model. Figure 2 illustrates the overall framework to utilize SPLs.

SPL for Individual Training Sample

To determine the SPL class for each video clip in D_p, we first perform inference on each video clip in D_p using the teacher model T trained on D_t. A 2-dimensional confusion matrix $C \in \mathbb{R}^{N \times N}$ can be obtained to summarize the alignments between the teacher model inferences (columns) and the original weak annotations (rows).

Specifically, video clips at the diagonal location (w, w) of C can be roughly treated as samples belonging to class w, which is agreed by the original weak label as well as the teacher model T. For other samples at off-diagonal location (h, w) of C, we interpret them as follows: from the view of the weak labels, these clips come from videos retrieved using text query of the action class h. Therefore, they include context information that may not exactly represent the action class h but is semantically related to it. However, from the view of the teacher model T, visual features that belong to action class w can also be found in these clips based on knowledge learned from the target dataset D_t. Instead of allocating these samples to either action class h or w with the risk of leading to label noise, we convert such confusion to a useful supervision signal by assigning SPLs. For each data sample (x, y) in D_p, the sub-pseudo label $y \in [0, N^2 - 1]$ of the video clip x is obtained by

$$y = N \cdot l + T(x), \tag{2}$$

where l is the weak label of x and $T(x)$ is its teacher prediction, where $l, T(x) \in [0, N - 1]$.

Reduction of the Quadratic SPL Space

Given N categories of original labels, SPL results in a quadratic label space $O(N^2)$. When N is big, the softmax output layer becomes too large to train 3D video models efficiently. Moreover, when classes of a task are diverse, some classes could share very few or none concepts, resulting in less data included in such SPL classes. The distribution of SPL classes could be long-tailed (explained in Fig. 1 of the Appendix) as some semantically distant classes can be unlikely confused with each other. We explore several space reduction strategies for SPL to solve this problem.

Merge to Diagonal (SPL-M): Suppose we are targeting at SPLs with a total number of $K \times N$ classes, we kept N in-diagonal classes and then select the most frequent $(K - 1) \times N$ off-diagonal classes as new SPLs. For the samples of un-selected off-diagonal classes, we merge them into the diagonals of their corresponding rows. Since each row belongs to a class of weak labels, this strategy promotes original weak labels over teacher predictions.

Discard Tail Off-diagonal (SPL-D): The confusion matrix itself encodes information about label noises: the less frequent SPL classes have higher potentials to contain mislabeled data. Therefore, unlike SPL-M merging the unselected classes to corresponding diagonals, SPL-D discards these training samples, resulting in a smaller yet potentially less noisy training dataset.

Binary Merge (SPL-B): We explore using agreement and disagreement between weak labels and teacher predictions of video clips as a criterion to reduce the SPL space. In this case, the confusion matrix entries are reduced to $2 \times N$ classes, including N in-diagonal classes (agreement) and N off-diagonal classes (disagreement) of each row. In this way, each original class (in-diagonal classes) has exactly one new

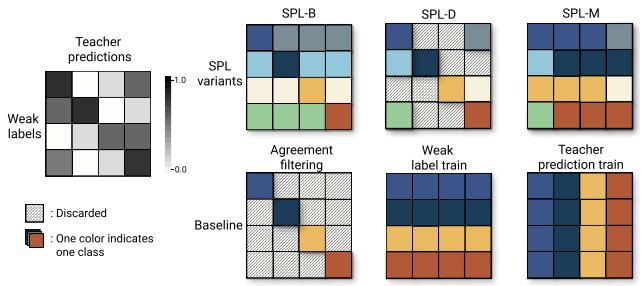

Figure 3: The left side is the confusion matrix. The top-right shows space reduction strategies for SPL, SPL-B: Using agreed and disagreed entries of each row as SPL classes. SPL-D: Keeping the top frequent entries. SPL-M: Merging less frequent off-diagonal entries to diagonals. The bottom-right shows baseline pseudo-label strategies.

sub-class (off-diagonal classes) to represent related contexts corresponding to this original class. As a result, SPL-B creates and maintains all sub-classes for the whole label space compared with SPL-D. It tends to have more even distribution of SPL classes than SPL-M. Such properties bring SPL-B potential strength when the original label space is large and classes are quite diverse.

SPL is simple and can be viewed as two-dimensional expansion of confusion matrix between week labels and pseudo labels. All these space reduction strategies can be viewed as pruning the confusion matrix, as illustrated in Figure 3. Figure 3 provides a holistic ("general") illustration that helps intuitively understand connections and differences between SPL and existing pseudo-label strategies as discussed in Section . Specifically, *Weak label train* is the weakly-supervised training studied by (Ghadiyaram, Tran, and Mahajan 2019). *Teacher prediction train* is a basic teacher-student self-training methods studied by (Furlanello et al. 2018; Xie et al. 2020). *Agreement filtering* only takes samples whose the weak label is matched with the teacher model prediction on it. That being said, existing strategies more or less explore partial knowledge inside the confusion matrix. However, SPL considers higher-level class affinities that haven't been considered yet. We will investigate these alternative strategies in experiments.

Experiments

We evaluate the proposed SPL algorithm on both common action recognition as well as fine-grained action recognition datasets. For the common action dataset, we mainly use Kinetics-200 (K200) (Xie et al. 2018) which is a subset of Kinetics-400 (Carreira and Zisserman 2017). In total, it contains 200 action categories with around 77K videos for training and 5K videos for validation. Studies (Xie et al. 2018; Choi et al. 2019) show that evaluations on K200 can be well generalized to the full Kinetics and other cases. Due to the huge computation resources required for extreme large-scale pre-training when taking full Kinetics as the target dataset, K200 results in an optimal choice for new algorithm explorations with an appropriate scale. Evaluation is conducted using Top-1 and Top-5 accuracy. To validate the learned representations by SPL are generic and useful, we also follow recent works (Miech et al. 2020; Patrick et al. 2020; Stroud et al. 2020) to conduct experiments on popular HMDB-51 (Kuehne et al. 2011) and UCF-101 (Soomro, Zamir, and Shah 2012) datasets. Specifically, we follow standard protocol to directly fine-tune the SPL pre-trained model on these two benchmarks to obtain results.

Pre-train Method	Top-1	Top-5
ImageNet Pre-train	80.6	94.7
Weak Label Train	82.8	95.6
Teacher Prediction Train	81.9	95.0
Agreement Filtering Train	82.9	95.4
Data Parameters	83.2	95.5
SPL-B (Ours)	**84.3**	**95.7**

Table 1: Comparisons with different pre-training strategies on WebK200-147K-V set with 6.7×10^5 clips. Fine-tuning results on Kinetics-200 dataset are shown.

We also conduct complete evaluations on fine-grained action dataset SoccerNet (Giancola et al. 2018), which is proposed for action event recognition in soccer broadcast videos and belongs to an important application domain. It includes three foreground action classes: *Goal, Yellow/Red Card, Substitution* and one *Background* class for the rest contexts in soccer games. We use 5547 video clips for training and 5547 clips for validation obtained from different full-match videos. For the evaluation matrix, we focus more on the performance of classifying foreground action classes, which are sparsely occurred in the broadcast videos. Therefore, mean average precision without background class is adopted. We also discuss cases for dataset whose class names cannot be used as reliable search queries in Appendix.

Weakly-Labeled Data Collection

To construct the pre-training dataset D_p for each target dataset (K200 and SoccerNet), we collect untrimmed web videos retrieved by a text-based search engine similar to (Caba Heilbron et al. 2015; Chen and Gupta 2015) and construct several dataset versions for following studies. Also see Appendix for more details.

WebK200. We treat the class names of Kinetics-200 dataset as the searching queries and use 4 languages for each query, including English, French, Spanish and Portuguese. We construct two web videos sets with different sizes: WebK200-147K-V with 147K videos and WebK200-285K-V with 285K videos (including more low-ranked videos returned by the search engine). Duplicate checking has been done on these two sets and videos with addresses appearing in the validation set of Kinetics-200 are removed. We also find only 1.27% of WebK200 videos are shared with untrimmed K400 videos. We sample a number of video clips with the length of 10 seconds from the retrieved videos. The number of clip samples for each class is roughly balanced.

WebS4. For the three foreground classes in SoccerNet: Goal, Yellow/Red Card, Substitution, we obtain the searching queries based on related terms from Wikipedia such as "free kick goal", "corner kick goal" resulting in 9 kinds of queries in total for these 3 foreground classes. For the Background class, we use "soccer full match" as the query. For

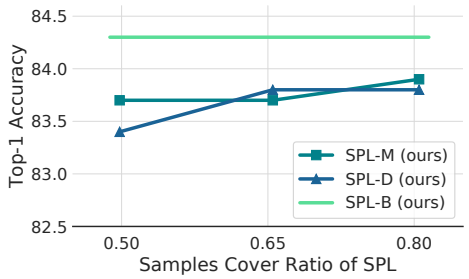

Figure 4: SPL-M and SPL-D with different samples cover ratio (SCR) defined in the Section for pre-training on WebK200-147K-V with 6.7×10^5 clips. Fine-tuning results on Kinetics-200 dataset are shown.

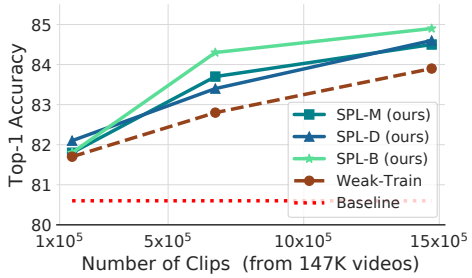

Figure 5: Effect of different numbers of clips given a fixed number of videos (WebK200-147K-V). Fine-tuning results on Kinetics-200 dataset are shown.

each searching query, we use 3 languages including English, French and Spanish. We sample 10 seconds video clips and keep the clip numbers for each class roughly balanced. Two web video sets are obtained with different number of total clips: WebS4-73K-C and WebS4-401K-C, where 73K and 401K represent the number of video clips in these two sets.

Implementation Details

We use 3D ResNet-50 (Wang et al. 2018) with self-gating (Xie et al. 2018) as the baseline model and more details are described in the Appendix. Following (Wang et al. 2018), the network is initialized with ResNet-50 pre-trained on ImageNet (Deng et al. 2009). At training stage, we use the batch size of 6 and take 16 RGB frames with temporal stride 4 as the input. The spatial size of each frame is 224×224 pixels obtained from the same randomly cropping operation as (Wang et al. 2018). For the pre-training on WebK200 sets, we set warm up training for 10 epochs with starting learning rate as 0.04 and then use learning rate 0.4 with cosine decay for 150 epochs. For the fine-tuning, we initialize the model from the last epoch of the pre-training and conduct end-to-end fine-tuning. We set warm up training for 10 epochs with starting learning rate as 0.04 and then use learning rate of 0.4 with cosine decay for 60 epochs. For SoccerNet dataset, we use the same pre-training setting with WebK200 to conduct pre-training on the WebS4 sets. For the fine-tuning, we use learning rate of 0.005 for 20 epochs. More settings of hyper-parameters are described in the appendix. They are obtained to get the best performance of baselines. Our method is implemented using TensorFlow (Abadi et al. 2015).

Pre-train Method	Number of videos	
	147K-V	285K-V
Synthetic Noise Ratio [1]	58.9 %	65.5%
Weak Label Train	83.9	84.0
SPL-M (Ours)	84.5	84.8
SPL-D (Ours)	84.6	84.9
SPL-B (Ours)	**84.9**	**85.3**

Table 2: Effect of different number of videos given a fixed number of clips (around 1.4×10^6). Results are reported on Kinetics-200 (Top-1 accuracy).

Results on Kinetics-200 Dataset

In this section, we verify the effectiveness of the proposed method on Kinetics-200 via studies of different perspectives and explorations. Fine-tuning results are reported.

Comparisons with other pre-training strategies. Section and Figure 3 categorize different pseudo-label strategies. Here we compare these strategies to ours. We report results based on their pre-training on our WebK200-147K-V set with 6.7×10^5 clips. From Table 1, we find they can all improve upon the baseline ImageNet pre-training. The performance gap between pre-training using Weak Label (Ghadiyaram, Tran, and Mahajan 2019) and Teacher Prediction (Xie et al. 2020) indicates there are more useful information included in weak labels. Although Agreement Filtering can do some noise reduction to the web videos, it discards around 60% of training samples resulting in a comparable performance with Weak Label. We also adopt Data Parameters (Saxena, Tuzel, and DeCoste 2019), one of the recent state-of-the-art methods for learning with noisy labels, to conduct pre-training on web videos. Our SPL-B (strategy with the best performance on Kinetics-200) outperforms these baselines and is able to take use of all noisy data.

Comparisons among different space reduction strategies for SPL. To compare these space reduction strategies, we conduct experiments on our WebK200-147K-V set with 6.7×10^5 clips for pre-training. We start with total number of SPL classes as $K \times N = 400$ so that the label space is consistent for the three variations. The label space of SPL-D and SPL-B is controlled by hyper-parameter K and their space is reduced by merging or discarding samples belong to infrequent SPLs. There is a question about how many frequent SPLs to keep. More classes introduce more fine-grained tail SPLs yet higher computation cost. We define samples cover ratio (SCR) = $\frac{\text{\# of samples in selected SPLs}}{\text{\# of total samples}}$. Specifically, 400 SPL classes give SCR= 49.81%. We evenly increase SCR by 15% to get 1600 and 4500 SPL classes with SCR of 65% and 80% respectively. From the result in Figure 4, we find that including more SPL classes can generally improve the performance of SPL-M and SPL-D. But the overall improvement gain is limited. SPL-B outperforms other space reduction alternatives. To verify our hypothesis about the advantages of SPL-B stated in the method section, we conduct further studies shown in Fig. 1 of the Appendix. SPL-B shows

[1] We use $\frac{\text{\# of off-diagonal elements}}{\text{\# of total elements}}$ in the confusion matrix as a synthetic measurement of the noise ratio.

Method	Video #	Top-1	Top-5
S3D	77K	78.4	-
R3D-50	77K	75.5	92.2
R3D-50-NL	77K	77.5	94.0
R3D-50-CGNL	77K	78.8	94.4
Omni (Slow-R50)	77K + 808K Web	82.9	95.8
SPL-B (R3D-50-G)	77K + 285K Web	**85.3**	**96.6**

Table 3: Results of other methods on Kinetics-200.

Method	Extra data	Video #	HMDB	UCF
Self-supervised:				
SpeedNet	K400	240K	48.8	81.1
AVTS	Audioset	2M	61.6	89.0
XDC	IG65M	65M	67.4	94.2
GDT	K400	240K	57.8	88.7
BYOL	K400	240K	**73.6**	95.5
Webly-supervised:				
MIL-NCE	HT100M	1.2M	61.0	91.3
WVT	WVT-70M	70M	65.3	90.3
Webly-supervised with Teacher Models:				
Omni's Teacher	K400	240K	69.4	94.7
Omni	Omni-8M	8.7M	70.7	96.0
Our Teacher	K200	77K	66.4	93.2
SPL-B (Ours)	WebK200	285K	70.4	**96.2**

Table 4: Fine-tuning results on HMDB-51 and UCF-101 benchmarks. Results validate that pre-training with SPL achieves much better performances than the teacher model, has good transferability while relies on much less web data.

clearly more even distribution of sample numbers to overcome long-tail distributions when the original label space is large. Note that SPL-B is designed for reducing quadratic SPL space. When the classes of a task is highly fine-grained or space is small, it is not necessary as we demonstrated in following datasets like SoccerNet and Clothing1M.

Effect of the number of training samples, more clips or more videos? It is a common practice to improve the performance of deep neural networks by increasing the number of training samples (Ghadiyaram, Tran, and Mahajan 2019). There are two practical options to do so given untrimmed web videos: (1) sampling more clips from a given number of web videos or (2) collecting more videos to sample clips. Both ways have potential drawbacks. The first one may result in duplication of visual contents. The second one may lead to lower quality of weak labels because this means we have to include more low-ranked videos returned by the search engine. In the following experiments, we aim to study this practical question and verify the effectiveness of SPL.

Effect of more clips. We sample different numbers of clips from WebK200-147K-V set (described in Section) and plot results of different pre-training strategies in Figure 5. The baseline result with red dot line represents the performance of using ImageNet pre-trained models. Results show that sampling more video clips from a given number of untrimmed videos can help improve the model performance. We also find that with a sufficient number of video clips available, SPL methods consistently outperform weak label pre-training by providing enriched supervision.

Pre-train Method	Pre-train Set	
	WebS4-73K-C	WebS4-401K-C
Baseline	73.7	73.7
Weak Label Train	74.8	75.3
Agreement Filter	75.1	75.4
Teacher Predication	74.1	75.2
SPL-B (Ours)	75.8	76.4
SPL (Ours)	**76.1**	**76.8**

Table 5: Results on the SoccerNet dataset. "Baseline" represents the ImageNet pre-training.

Effect of more videos. We sample a similar number of total video clips, around 1.4×10^6, from WebK200-147K-V (147K videos) and WebK200-285K-V (285K videos) to obtain two training sets. We conduct teacher model inference on these two sets to get a synthetic measurement of the noise ratio and find this ratio is larger in WebK200-285K-V set. The comparison in Table 2 indicates that, though synthetic noise ratio is higher with the increase of videos, enriched visual contents are beneficial to some extent. Besides, SPL-B obtains more performance gain than directly using weak labels, which shows its robustness to label noise.

Comparisons with other methods on Kinetics-200. In Table 3, we list results of other existing methods on this benchmark: S3D (Xie et al. 2018), R3D-NL (Wang et al. 2018), R3D-CGNL (Yue et al. 2018) and Omni (Duan et al. 2020). Our method is able to outperform the state-of-the-art methods by a clear margin. Specifically, Omni uses a teacher model to do data filtering and relies on a large amount of extra web data (808K web videos and 6704K web images) to conduct pre-training. Compared with it (Duan et al. 2020), SPL achieves better performance (2.4% Top-1 accuracy) while using less web data. As for backbone, SPL uses a backbone including 33.2 M parameters and it is comparable to Slow-R50 (Feichtenhofer et al. 2019) with 32.4M used in Omni. We also noticed these existing methods are trained using different input length and we conduct more studies with details in Appendix. We use 8-frame inputs to do the model inference with SPL models that were trained using 16-frames inputs. In this case, our baseline teacher has the same performance (78.6% top-1) with Omni's teacher. SPL (83.2% top-1) still outperforms Omni that relies on much more web data. As for improvements upon its own baseline "Weak Label Train", SPL obtains more gains (1.2% V.S. 0.5%) than Omni. These results show the strength of exploring enriched supervision by SPL.

Attention visualization of SPL classes. We visualize the visual concepts learned from SPLs via attention visualization. Specifically, we extend Grad-CAM (Selvaraju et al. 2017) to 3D CNNs to show the model's focus when making predictions. In Figure 6, we show some examples of SPL classes along with attention maps of the model trained using SPL. It is interesting to observe some meaningful "middle ground" concepts that can be learnt by SPL, such as mixing the eggs and flour, using the abseiling equipment.

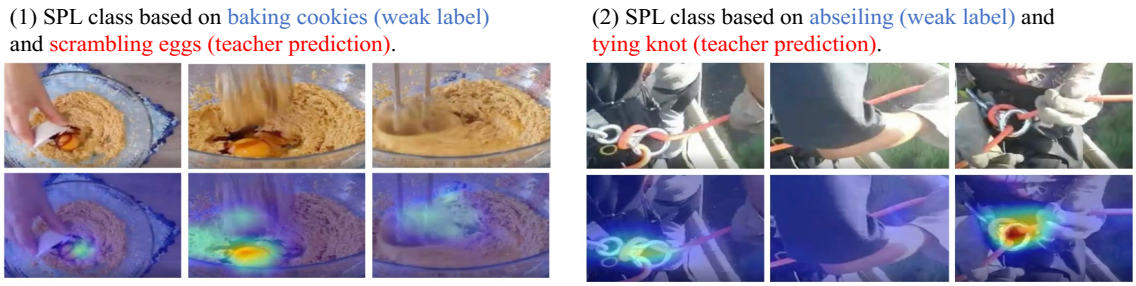

Figure 6: Examples of attention visualization for SPL classes. Original weak label (blue) and the teacher model prediction (red) are listed. Some meaningful "middle ground" concepts can be learnt by SPL, such as mixing up the eggs and flour (top) and using the abseiling equipment (bottom).

Transferability of Features Learned by SPL

To validate the learned features by SPL are generic and can be transferred to other video recognition tasks, we follow recent works (Duan et al. 2020; Miech et al. 2020; Stroud et al. 2020) to conduct experiments on HMDB-51 and UCF-101 datasets. Note that we directly fine-tune the SPL-B pre-trained model on these two benchmarks to obtain results - during this process, no new noisy web dataset is collected when treating HMDB-51 and UCF-101 datasets as the target dataset In Table 4, we list results of recent self-supervised and webly-supervised methods: SpeedNet (Benaim et al. 2020), BYOL (Feichtenhofer et al. 2021), AVTS (Korbar, Tran, and Torresani 2018), XDC (Alwassel et al. 2019), GDT (Patrick et al. 2020), MIL-NCE (Miech et al. 2020), WVT (Stroud et al. 2020) and Omni (Duan et al. 2020). We also report results of our K200 pre-train teacher model baseline. It is worth to note that conducting totally direct comparisons between recent webly-supervised video understanding frameworks is hard, due to the public unavailable web data used in each work. We have tried to set up fair comparisons as far as we can. Compared with Omni that also relies on a teacher model, SPL achieves competitive results using a weaker teacher (66.4 VS 69.3 top-1 accuracy on HMDB) and much less web data (285K V.S. 8.7M extra web videos).

Experiments on SoccerNet Dataset

We also conduct experiments on SoccerNet (Giancola et al. 2018), a fine-grained action recognition dataset. Different from Kinetics-200 actions, this dataset contains broadcast videos from soccer matches, so all classes contain sports actions sharing very similar visual (background) content. Therefore there exists high confusion between different classes. Moreover, we find actions in untrimmed web videos is transitory, leading to high temporal noise. We use two web video sets WebS4-73K-C and WebS4-401K-C with 73K clips and 401K clips respectively as described in Section . Since the label space is not large, we are able to test the full version of SPL that generates 4^2 SPL classes as well as SPL-B. In Table 5, we show the fine-tuning results on SoccerNet val set based on different types of pre-training.

Our SPL method consistently outperform other pre-training strategies, suggesting the advantages of SPL to learn short actions from untrimmed web videos. Besides, SPL-B is slightly worse than SPL, which indicates that reductions of SPL space may be not necessary when the original label space is not large and these original classes share a lot of class concepts. In such cases, long-tailed distribution is less severe and details of sub-pseudo class information in SPL can be better maintained without reduction.

Method	Top-1 (%)
None (Patrini et al. 2017)	79.43
Forward (Patrini et al. 2017)	80.38
CleanNet (Lee et al. 2018)	79.90
NoiseRank (Sharma et al. 2020)	79.57
SPL-B (Ours)	80.31
SPL (Ours)	**80.50**

Table 6: Results on the Clothing1M image classification dataset. All methods use the ResNet-50 backbone.

Generalizing SPL to Image Classification

We also test SPL on Clothing1M (Xiao et al. 2015), a large-scale image datasets with real-world label noises. Clothing1M contains $47,570$ training images with human-annotated labels and $\sim$1M images with noisy labels. There are 14 original fashion classes. This dataset is challenging and 1% improvement is regarded as important. SPL can directly be used for this task. Since the label space is not large here, we test the basic version of SPL that generates 14^2 SPLs classes as well as SPL-B. We follow common experimental setting (Lee et al. 2018) and starts ResNet-50 pre-training with random initialization. Then we fine-tune on the clean set. Table 6 compares against previous methods, suggesting generalizability of the proposed method. SPL either outperforms or achieves competitive results. Besides, SPL is still better than SPL-B when the label space is not large.

Conclusion

We propose a novel and particularly simple method of exploring SPLs from untrimmed weakly-labeled web videos. SPL does not increase training complexity and can be treated as an off-the-shelf technique to integrate with teacher-student like training frameworks in orthogonal. In addition, we believe it is a promising direction to discover meaningful visual concepts by bridging weak labels and the knowledge distilled from teacher models. SPL also demonstrates promising generalization to the image recognition domain, suggesting promising future directions like applying SPL to other tasks where there exists uncertainty in labels.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; and Zheng, X. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Alwassel, H.; Mahajan, D.; Torresani, L.; Ghanem, B.; and Tran, D. 2019. Self-supervised learning by cross-modal audio-video clustering. *arXiv:1911.12667*.

Benaim, S.; Ephrat, A.; Lang, O.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; Irani, M.; and Dekel, T. 2020. SpeedNet: Learning the Speediness in Videos. In *CVPR*.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Chang, C.-Y.; Huang, D.-A.; Sui, Y.; Fei-Fei, L.; and Niebles, J. C. 2019. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *CVPR*.

Chen, X.; and Gupta, A. 2015. Webly supervised learning of convolutional networks. In *ICCV*.

Choi, J.; Gao, C.; Messou, J. C.; and Huang, J.-B. 2019. Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. In *NeurIPS*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Duan, H.; Zhao, Y.; Xiong, Y.; Liu, W.; and Lin, D. 2020. Omni-sourced Webly-supervised Learning for Video Recognition. In *ECCV*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *ICCV*.

Feichtenhofer, C.; Fan, H.; Xiong, B.; Girshick, R.; and He, K. 2021. A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning. In *CVPR*.

Furlanello, T.; Lipton, Z. C.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born again neural networks. In *ICML*.

Gan, C.; Sun, C.; Duan, L.; and Gong, B. 2016a. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *ECCV*.

Gan, C.; Yao, T.; Yang, K.; Yang, Y.; and Mei, T. 2016b. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *CVPR*.

Ghadiyaram, D.; Tran, D.; and Mahajan, D. 2019. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*.

Giancola, S.; Amine, M.; Dghaily, T.; and Ghanem, B. 2018. Soccernet: A scalable dataset for action spotting in soccer videos. In *CVPR Workshops*.

He, K.; Zhang, X.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. In *NeurIPS Workshop*.

Korbar, B.; Tran, D.; and Torresani, L. 2018. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*.

Kuehne, H.; Iqbal, A.; Richard, A.; and Gall, J. 2019. Mining YouTube-A dataset for learning fine-grained action concepts from webly supervised video data. *arXiv:1906.01012*.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *ICCV*.

Laptev, I. 2005. On space-time interest points. *International Journal of Computer Vision*.

Lee, K.-H.; He, X.; Zhang, L.; and Yang, L. 2018. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*.

Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; and van der Maaten, L. 2018. Exploring the limits of weakly supervised pretraining. In *ECCV*.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*.

Müller, R.; Kornblith, S.; and Hinton, G. 2020. Subclass Distillation. *arXiv:2002.03936*.

Patrick, M.; Asano, Y. M.; Fong, R.; Henriques, J. F.; Zweig, G.; and Vedaldi, A. 2020. Multi-modal self-supervision from generalized data transformations. *arXiv:2003.04298*.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*.

Rupprecht, C.; Kapil, A.; Liu, N.; Ballan, L.; and Tombari, F. 2018. Learning without prejudice: Avoiding bias in webly-supervised action recognition. *CVIU*.

Saxena, S.; Tuzel, O.; and DeCoste, D. 2019. Data Parameters: A New Family of Parameters for Learning a Differentiable Curriculum. In *NeurIPS*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Sharma, K.; Donmez, P.; Luo, E.; Liu, Y.; and Yalniz, I. Z. 2020. NoiseRank: Unsupervised Label Noise Reduction with Dependence Models. *arXiv:2003.06729*.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*.

Stroud, J. C.; Ross, D. A.; Sun, C.; Deng, J.; Sukthankar, R.; and Schmid, C. 2020. Learning Video Representations from Textual Web Supervision. *arXiv:2007.14937*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *CVPR*.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *CVPR*.

Xie, Q.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2020. Self-training with Noisy Student improves ImageNet classification. In *CVPR*.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*.

Yan, X.; Misra, I.; Gupta, A.; Ghadiyaram, D.; and Mahajan, D. 2020. Clusterfit: Improving generalization of visual representations. In *CVPR*.

Yue, K.; Sun, M.; Yuan, Y.; Zhou, F.; Ding, E.; and Xu, F. 2018. Compact generalized non-local network. In *NeurIPS*.