Leveraging Multi-modal Prior Knowledge for Large-scale Concept Learning in Noisy Web Data

Junwei Liang
Carnegie Mellon University
junweil@cs.cmu.edu

Lu Jiang
Carnegie Mellon University
lujiang@cs.cmu.edu

Deyu Meng
Xi'an Jiaotong University
dymeng@mail.xjtu.edu.cn

Alexander Hauptmann
Carnegie Mellon University
alex@cs.cmu.edu

ABSTRACT

Learning video concept detectors automatically from the big but noisy web data with no additional manual annotations is a novel but challenging area in the multimedia and the machine learning community. A considerable amount of videos on the web is associated with rich but noisy contextual information, such as the title and other multi-modal information, which provides weak annotations or labels about the video content. To tackle the problem of large-scale noisy learning, We propose a novel method called Multi-modal WEbly-Labeled Learning (WELL-MM), which is established on the state-of-the-art machine learning algorithm inspired by the learning process of human. WELL-MM introduces a novel multi-modal approach to incorporate meaningful prior knowledge called curriculum from the noisy web videos. We empirically study the curriculum constructed from the multi-modal features of the Internet videos and images. The comprehensive experimental results on FCVID and YFCC100M demonstrate that WELL-MM outperforms state-of-the-art studies by a statically significant margin on learning concepts from noisy web video data. In addition, the results also verify that WELL-MM is robust to the level of noisiness in the video data. Notably, WELL-MM trained on sufficient noisy web labels is able to achieve a better accuracy to supervised learning methods trained on the clean manually labeled data.

KEYWORDS

Video Understanding; Prior Knowledge; Web Label; Big Data; Webly-supervised Learning; Noisy Data; Concept Detection

ACM Reference format:
Junwei Liang, Lu Jiang, Deyu Meng, and Alexander Hauptmann. 2017. Leveraging Multi-modal Prior Knowledge for Large-scale Concept Learning in Noisy Web Data. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania, ,* 9 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079003

ICMR '17, June 6–9, 2017, Bucharest, Romania
 ACM ISBN 978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3079003

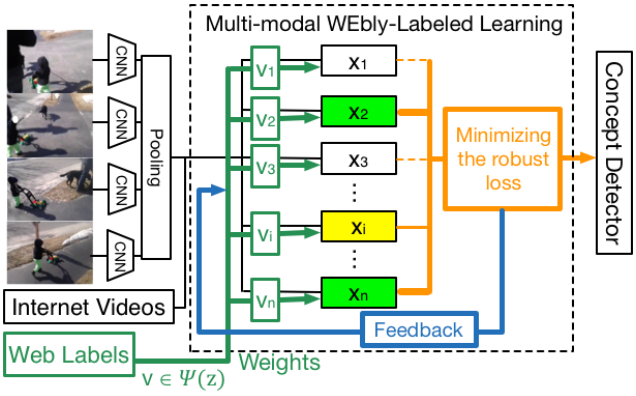

Figure 1: Overview of Multi-modal WEbly-Labeled Learning (WELL-MM). The algorithm jointly models the prior knowledge extracted from web labels and the current learned model at each iteration to overcome the noise labels. $\{\mathbf{x}_i\}_{i=1}^n$ are input samples and their current weights are determined by $\{\mathbf{v}_i\}_{i=1}^n$. Colored samples are the samples with nonzero weights at the current iteration. The blue line indicates the feedback from the previous objective function value.

1 INTRODUCTION

Millions of videos are being uploaded to the Internet every day. These videos capture different aspects of multimedia content about our daily lives. Automatically categorizing videos into concepts, such as people, actions, objects, etc., is an important topic. Recently many studies have been proposed to tackle the problem of concept learning [1, 6, 10, 14, 21, 23, 29, 30].

Many datasets acquire the clean concept labels via annotators. These datasets include ImageNet [10], TRECVID MED [36] and FCVID [22]. Collecting such datasets requires significant human effort, which is particularly expensive for video data. As a result, the labeled video collection is usually much smaller than the image collection. For example, FCVID [22], only contains about 0.09 million labels on 239 concept classes, much less than the 14 million labels on over 20,000 classes in the image collection ImageNet [10]. On the other hand, state-of-the-art concept models utilize deep neural networks [23, 46], which need more data to train. However, training only on manually labeled clean data seem insufficient for large-scale concept learning.

Images or videos on the web often contain rich contextual information, such as their titles or descriptions. We can infer the label by the metadata. Figure 2 shows an example of a video with inferred concept label "walking a dog". In this paper, we call the samples with inferred labels *weakly labeled* or *webly labeled*. The webly-labeled data are easy to collect and thus usually orders-of-magnitude larger than manually-labeled data. However, the web labels are very noisy and have both low accuracy and low recall.

Concept learning over weakly labeled data becomes popular as it allows for large-scale learning on big data. However, these methods have only focused on utilizing a single text modality to model the noisy labels [4, 8, 13, 14, 27]. For example, in Figure 2, the textual metadata is useful but also contain lots of noises. In fact, the video is of multiple modalities and our intuition is that the inference obtained from multiple modalities is more reliable than that from a single text modality. For example, we can more confident to say this video is about "walk a dog" if we spot the text in the title, hear the words "good boy" in the speech, and see a dog in some key frames. To this end, we can leverage the prior knowledge in automatically extracted multi-modal features from the video content such as pre-trained still image detectors, automatic speech recognition and optical character recognition. In some cases when videos have little textual metadata, multi-modal knowledge become the only useful clues in concept learning.

Recent studies on weakly labeled concept learning show promising results. However, since existing approaches only focuses on a single modality, two important questions have yet: 1) what are the important multi-modal prior knowledge, except textual metadata, for modeling noisy labels? 2) how to integrate the multiple modalities into concept learning in a theoretically sound manner?

In this paper, to utilize multi-modal prior knowledge for concept learning, we propose a learning framework called **Multi-modal WEbly-Labeled Learning (WELL-MM)**. The learning framework is motivated by human learning, in which the learner starts from learning easier aspects of a concept, and then gradually take more complex examples into the learning process[3, 20, 25]. Specifically. WELL-MM learns a concept detector iteratively from first using a few samples with more confident labels, then gradually incorporate more samples with noisier labels. Figure 1 shows the overview of the proposed framework. The algorithm integrates multi-modal prior knowledge, which is derived from the multi-modal video and image features, into the dynamic learning procedure. The idea of curriculum and self-paced learning paradigm has been proved to be efficient to deal with noise and outliers [8, 18, 25]. Our proposed method is the first to generalize the learning paradigm to leverage multi-modal prior knowledge into concept learning. Experimental results show that multi-modal prior knowledge is important in concept learning over noisy data. The proposed WELL-MM outperforms other weakly labeled learning methods on three real-world large-scale datasets, and obtains the state-of-the-art results with recent deep learning models.

The contribution of this paper is threefold. First, we propose a novel solution to address the problem of weakly labeled data learning through a general framework that considers multi-modal prior knowledge. We show that the proposed WELL-MM not only outperforms state-of-the-art learning methods on noisy labels, but also, notably, achieves comparable results with models trained using manual annotation on one of the video dataset. Second, we provide valuable insights by empirically investigating different multi-modal prior knowledge for modeling noisy labels. Experiments validate that by incorporating multi-modal information, our method is robust against certain levels of noisiness. Finally, the efficacy and the scalability have been demonstrated on three public large-scale benchmarks, which include datasets on both Internet videos and images. The promising results suggest that detectors trained on sufficient weakly labeled videos may outperform detectors trained on existing manually labeled datasets.

Figure 2: Multi-modal prior knowledge from web video.

2 RELATED WORK

Curriculum and Self-paced Learning: Recently a learning paradigm called *curriculum learning* (CL) was proposed by Bengio et al., in which a model is learned by gradually incorporating from easy to complex samples in training so as to increase the entropy of training samples [3]. A curriculum determines a sequence of training samples and is often derived by predetermined heuristics in particular problems. For example, Chen et al. designed a curriculum where images with clean backgrounds are learned before the images with noisy backgrounds [8], i.e. their method first builds a feature representation by a Convolutional Neural Network (CNN) on images with clean background and then they fine tune the models on images with noisy background. In [41], the authors approached grammar induction, where the curriculum is derived in terms of the length of a sentence. Because the number of possible solutions grows exponentially with the length of the sentence, and short sentences are easier and thus should be learned earlier.

The heuristic knowledge in a problem often proves to be useful. However, the curriculum design may lead to inconsistency between the fixed curriculum and the dynamically learned models. That is, the curriculum is predetermined a prior and cannot be adjusted accordingly, taking into account the feedback about the learner. To alleviate the issue of CL, Kumar et al. designed a learning paradigm, called *self-paced learning* (SPL) [25]. SPL embeds curriculum design as a regularizer into the learning objective. Compared with CL,

SPL exhibits two advantages: first, it jointly optimizes the learning objective with the curriculum, and thus the curriculum and the learned model are consistent under the same optimization problem; second, the learning is controlled by a regularizer which is independent of the loss function in specific problems. This theory has been successfully applied to various applications, such as matrix factorization [48], action/event detection [19], domain adaption [44], tracking [43] and segmentation [26], reranking [18], etc.

Learning Detectors in Web Data: Many recent studies have been proposed to utilize a large amount of noisy data from the Internet. For example, [35] proposed a Never-Ending Language Learning (NELL) paradigm and built adaptive learners that make use of the web data by learning different types of knowledge and beliefs continuously. In the image domain, existing methods try to tackle the problem of constructing qualified training sets based on the search results of text or image search engines [9, 11, 30, 47]. For example, NEIL [9] followed the idea of NELL and learned from web images to form a large collection of concept detectors iteratively via a semi-supervised fashion. By combining the classifiers and the inter-concept relationships it learned, NEIL can be used for scene classification and object detection task. [11] introduced a webly-supervised visual concept learning method that automatically learns a large amount of models for a wide range of variations within visual concepts. They discovered concept variances through the vocabulary of online books, and then downloaded images based on text-search from the web to train object detection and localization models. [30] presented a weakly-supervised method called Baby Learning for object detection from a few training images and videos. They first embed the prior knowledge into a pre-trained CNN. When given very few samples for a new concept, a simple detector is constructed to discover much more training instances from the online weakly labeled videos. As more training samples are selected, the concept detector keeps refining until a mature detector is formed. [47] proposed a noise estimation method for training convolutional neural network with large-scale e-commerce images. Another recent work in image domain [8] proposed a webly supervised learning of Convolutional Neural Network. They utilized easy images from search engine like Google to bootstrap a first-stage network and then used noisier images from photo-sharing websites like Flickr to train an enhanced model.

In video domain, only a few studies [12, 16, 46] have been proposed for noisy data learning since training robust video concept detectors is more challenging than the problem in the image domain. [12] tackled visual event detection problem by using SVM based domain adaptation method in web video data. [16] described a fast automatic video retrieval method using web images. Given a targeted concept, compact representations of web images obtained from search engines like Google, Flickr are calculated and matched to compact features of videos. Such method can be utilized without any pre-defined concepts. [46] discussed a method that exploits the YouTube API to train large-scale video concept detectors on YouTube. The method utilized a calibration process and hard negative mining to train a second order mixture of experts model in order to discover correlations within the labels.

Most of the existing methods are heuristic approaches as it is unclear what objective is being optimized on the noisy data. Moreover, results obtained from the web search results is just one approach to acquire prior knowledge or curriculum. To the best of our knowledge, there have been no systematical studies on exploiting the multi-modal prior knowledge in video concept learning on noisy data. Since search engine algorithm is changing rapidly, it is unclear that how noisy the web labels are and how the level of noisiness in the data will affect performance. In this paper, we proposed a theoretically justified method with a clear framework for curriculum constructing and model learning. We also empirically demonstrate its superior performance over representative existing methods and systemically verify that WELL-MM is robust against the level of noisiness of the video data.

3 MULTI-MODAL WEBLY-LABELED LEARNING (WELL-MM)

3.1 Problem Description

In this paper, following [46], we consider a concept detector as a classifier and our goal is to train concept detectors from webly-labeled video data without any manually annotated labels. Given a collection of training samples with noisy labels, we do not make any assumption over the underlying noise distribution. Formally, we represent the training set as $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{z}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^n$, where $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ are the d-dimensional features of the training set, and $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ represent each sample's corresponding noisy web labels. We assume that the noisy labels are given. The noisy web labels are often automatically inferred using the sample's textual metadata provided by its uploader, or from other modalities such as pre-trained convolutional neural network over still images[7], Automatic Speech Recognition [37], or Optical character recognition [40]. For example, for instance, a video might have a noisy label "cat" as its title and speech both contain the word cat. The $\tilde{\mathbf{y}}_i \subset \mathcal{Y}$ is the inferred concept label set for the i^{th} observed sample based on its noisy web label, and $\mathcal{Y}$ denotes the full set of target concepts. In our experiment, to simplify the problem, we employ one-versus-all strategy for multi-class classification, and discuss our method in the context of binary classification over the noisy web labels.

3.2 Model

3.2.1 Objective Function. In this section, we propose a model called Multi-modal WEbly-Labeled Learning (WELL-MM) to leverage multi-modal prior knowledge for weakly labeled data. Formally, given the training set $\mathcal{D}$ mentioned previously, Let $L(\tilde{y}_i, g(\mathbf{x}_i, \mathbf{w}))$,denote the loss function which calculates the cost between the inferred label $\tilde{y}_i$ and the estimated label given by the decision function $g(\mathbf{x}_i, \mathbf{w})$. Here $\mathbf{w}$ represents the model parameters. Our objective function is to jointly learn the model parameter $\mathbf{w}$ and the latent weight variable $\mathbf{v} = [v_1, \cdots, v_n]^T$ by:

$$\min_{\mathbf{w}, \mathbf{v} \in [0,1]^n} \mathbb{E}(\mathbf{w}, \mathbf{v}; \lambda, \Psi) = \sum_{i=1}^{n} v_i L(\tilde{y}_i, g(\mathbf{x}_i, \mathbf{w})) + f(\mathbf{v}; \lambda), \quad \text{subject to } \mathbf{v} \in \Psi \tag{1}$$

where the latent weight variable $\mathbf{v} = [v_1, \cdots, v_n]^T$ represents the inferred labels' confidence, and thus reflects the learning sequence of samples. In order to learn concept detectors in noisy data, we utilize the self-paced regularizer f [20] to control the learning process, where f is expect to assign greater weights to samples with

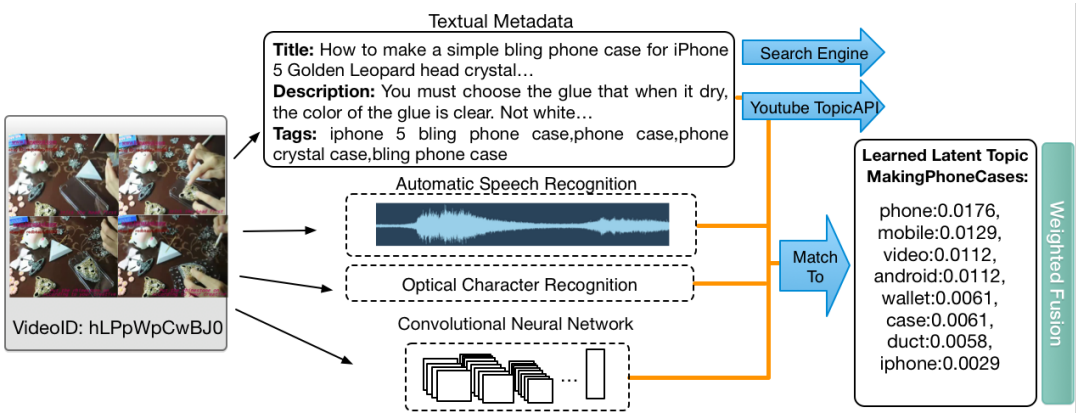

Figure 3: Curriculum Extraction Example. We automatically extract information using meaningful prior knowledge from several modalities and fuse them to get curriculum for WELL-MM. Our method makes use of text, speech, visual cues while common methods like search engine only extract textual information.

confident labels. For simplicity, we consider the linear regularizer Eq. (3) proposed in [20]:

$$f(\mathbf{v};\lambda) = \frac{1}{2}\lambda\sum_{i=1}^{n}(v_i^2 - 2v_i), \tag{2}$$

$\lambda \in (0,1)$ is a hyper-parameter that controls the pace of model training, which resembles the "age" of the model. We set λ to be small at the beginning and only samples of with small loss will be considered in training. As λ grows, more samples with larger loss will be gradually included. As stated in related studies [28, 33], the self-paced in Eq. (3) corresponds to a robust loss function. The robust loss in our problem tends to depress samples with noisy labels or outliers and thus may be instrumental in avoiding bad local minima.

In order to utilize the rich contextual information in the noisy data, we embed the multi-modal prior knowledge derived from the web labels $\mathbf{z}$ into a convex curriculum region Ψ for the latent weight variables. The shape of the region weakly implies the learning sequence, where favored samples have larger expected values. Generally, Ψ can be represented by $\Psi = \{\mathbf{v} \mid c(\mathbf{v},\mathbf{a}) \leq b\}$, where $\mathbf{a} = [a_1, \cdots, a_n]$ is the parameters of the region. In this paper, we use a linear constraint to form the curriculum region [20]:

$$\Psi = \{\mathbf{v} \mid \sum_{i=1}^{n} a_i v_i \leq b\} \tag{3}$$

The curriculum region is introduced to leverage the prior knowledge about the noisy labels and, as demonstrated in our experiments, is a crucial factor in weakly labeled data learning. We use multi-modal information to derive the probabilities of samples being positive of a class and if the probabilities are below a threshold (in our experiments it is set at zero) the samples will be consider as negatives. We assign value to a_i in correlated to samples 's probabilities being in the class, and b is set to 1. We use curriculum as a warm start in training, and set μ to zero after the first iteration. Since we empirically observed that curriculum constraints mostly benefit the first few iterations. We will discuss how to derive the multi-modal curriculum in details in the following section.

Eq. (1) is difficult to minimize over big data due to the constraints. In this paper, we propose to relax the constraints by introducing a Lagrange multiplier μ. The objective function then becomes:

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\mathbf{a},b,\mu) =$$
$$\sum_{i=1}^{n} v_i L(\tilde{y}_i, g(\mathbf{x}_i,\mathbf{w})) + \frac{1}{2}\lambda\sum_{i=1}^{n}(v_i^2 - 2v_i) + \mu(\sum_{i=1}^{n} a_i v_i - b), \tag{4}$$
$$\text{subject to } \mu \geq 0$$

The proposed Eq. (4) has two benefits over Eq. (1). First it enables the large-scale training on noisy data. This is important because as our experiments show that training on noisy data can outperform training on manually labeled data only when the noisy data are orders-of-magnitude larger. Second, it may tolerate the noise introduced in the curriculum region.

3.2.2 Multi-modal Curriculum. In this section we discuss the details on how to construct the curriculum region Ψ. Ψ is a feasible region that embeds the multi-modal prior knowledge extracted from the webly-labeled data as shown in Figure. 3. It geometrically corresponds to a convex feasible space for the latent weight variable. Given a set of training samples $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ with corresponding noisy labels $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n$, we want to extract the learning curriculum based on how related the training samples are to the target classes, which is modeled by the probability of the samples being the inferred class label (since we don't have the actual label in webly learning). The training samples with a greater value of probability mean that they are more confident to belong to the true class and should be learned earlier. Similar to Information Retrieval theory [31], here we use random variable $\mathbf{z}$ to represent the noisy web labels, $\mathbf{y}$ to represent the label classes, and the curriculum for a sample is then determined by:

$$\mathbf{P}(\mathbf{z} \mid \mathbf{y}) = \mathbf{P}(\mathbf{y} \mid \mathbf{z})\mathbf{P}(\mathbf{z})/\mathbf{P}(\mathbf{y}) \tag{5}$$

Since $\mathbf{P}(\mathbf{y})$ is the same for all samples, it can be regarded as a constant. The prior probability of a web video $\mathbf{P}(\mathbf{z})$ can be implemented with the duration, the view count and comments about the video. In

this paper we treat the prior as uniform so it can be ignored as well. Therefore, we calculate the curriculum simply based on $\mathbf{P}(\mathbf{y} \mid \mathbf{z})$, the probability of the sample being class $\tilde{y_i}$ given the noisy label. Since we want to incorporate the multi-modal prior information, we calculate the curriculum from:

$$\mathbf{P}(\mathbf{y} \mid \mathbf{z}) \propto \sum_m \theta_m \mathbf{P}(\mathbf{y} \mid \mathbf{z_m}) \tag{6}$$

We use random variable z_m to represent the m-th modality of the noisy labels for a sample and θ_m is the predetermined weight for modality m. In this paper, other than the textual metadata, we also utilize other modalities such as Automatic Speech Recognition (ASR) [37], Optical Character Recognition (OCR) [40] and basic image detector pre-trained on still images [38] (in this paper we use VGG net [39], extract keyframe-level image classification results and average them to get video-level results). Therefore the total number of the modalities is 4. We compare common ways to extract curriculum from web data for concept learning to the proposed novel method that utilizes state-of-the-art topic modeling techniques in natural language processing.

In the following methods (Word Hard Matching and Latent Topic with Word Embedding), we first extract bag-of-words features from different modalities for each video and then match them using specific matching methods to the concept words to get the probabilities in Eq. (6) as shown in Figure 3.

Word Hard Matching We build curriculum directly using exact word matching or stemmed word matching between the textual metadata of the noisy videos to the targeted concept names. This is the same method as stated in Webly Labeled Learning [28]. Noted that this method only utilizes one modality.

YouTubeTopicAPI The YouTube topic API is utilized to search for videos that are related to the concept words. The topic API uses textual information of the uploaded videos to obtain related topics of the videos from Freebase. This is the method used in [46].

SearchEngine The curriculum is built using the search result from a text-based search engine [32]. It is similar to related web-search based methods.

Ours We build the curriculum based on the latent topic we learned from the noisy label. We incorporate Latent Dirichlet Allocation (LDA) [5] to determine how each noisy labeled video is related to each target concept. The intuition is that each web video consists of mixtures of topics (concepts), and each topic is characterized by a distribution of words. We impose asymmetric priors over the word distribution so that each learned topic will be seeded with particular words in our target concept. For example, a topic will be seeded with words "make, phone, cases" for the target concept "MakingPhoneCases". we use the online variational inference algorithm from [17]. An example of the learned latent topic word distribution is shown on the right in Figure 3. We then match noisy labels from each modality $\mathbf{z_{im}}$ to the latent topic word distribution using word embedding soft matching [34]. The word embedding is pre-trained using Google News data.

Figure 3 shows an example of the noisy web video data and how the curriculum is extracted with different methods. Our method can utilize information from different modalities while common methods like search engine only consider textual information. We compare the performance of different ways of curriculum design by training detectors directly in Section 4.

3.3 Algorithm

As proven in recent studies [25, 33], Eq. (1) is a biconvex optimization problem. We utilize the alternative convex search algorithm (ACS) [2] to optimize Eq. (1) following [20, 25]. Algorithm 1 takes the input of the training set, an instantiated self-paced regularizer and the curriculum constraint function; it outputs an optimal model parameter $\mathbf{w}$. it derives the curriculum region from multi-modal noisy labels $\mathbf{Z} \in \mathbb{R}^{m \times n}$ and forms the curriculum constraint function. Then, it initializes the latent weight variables in the feasible region. In the while loop, the algorithm alternates between two steps until it finally converges: In step 4 given the most recent $\mathbf{v}^*$, the algorithm learns the optimal model parameters; In step 5, we fix the $\mathbf{w}^*$ and the algorithm learns the optimal weights $\mathbf{v}^*$ for each sample. Starting in the beginning, the model grows from learning with easy (less noisy) samples with a small model "age". The model "age" is gradually increased so that the model can incorporate more noisy samples in the training and become more robust over time. Step 4 can be implemented by existing off-the-shelf supervised learning methods such as the Support Vector Machine or back propagation. Gradient-based methods can be used to solve the convex optimization problem in Step 5. According to [15], the alternative search in Algorithm 1 converges as the objective function is monotonically decreasing and is bounded from below.

Algorithm 1: Multi-modal WEbly-Labeled Learning

input : Input dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{Z}, \tilde{\mathbf{Y}}\}$, self-paced function f and a curriculum constraint function c
output: Model parameter $\mathbf{w}$

1 Derive curriculum region from $\mathbf{Z} \in \mathbb{R}^{m \times n}$ into $\mathbf{a}, b$;
2 Initialize $\mathbf{v}^*$, λ in the curriculum region;
3 **while** *not converged* **do**
4 Update $\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathbb{E}(\mathbf{w}, \mathbf{v}^*; \lambda, \mathbf{a}, b)$;
5 Update $\mathbf{v}^* = \arg\min_{\mathbf{v}} \mathbb{E}(\mathbf{w}^*, \mathbf{v}; \lambda, \mathbf{a}, b)$;
6 **if** λ *is small* **then** increase λ by the step size;
7 **end**
8 **return** $\mathbf{w}^*$

4 EXPERIMENTS

In this section, we evaluate our method WELL-MM for learning video detectors on noisy labeled data. We first conduct our method on noisy learning in image domain. The efficacy of our methods are mainly verified on two major public benchmarks: FCVID and YFCC100M, where FCVID is by far one of the biggest manually annotated video dataset [22], and the YFCC100M dataset is the largest multimedia benchmark [45].

4.1 Experimental Setup

Datasets, Features and Evaluation Metrics Previous studies on noisy learning in image domain have been focusing on noise estimation [42, 47]. We compare our method with them on the synthesized noisy dataset CIFAR-10 generated using code from [47]. We

report accuracy on each setting along with the results reported in papers [42, 47] experimented on the same dataset.

Fudan-columbia Video Dataset (FCVID) contains 91,223 YouTube videos (4,232 hours) from 239 categories. It covers a wide range of concepts like activities, objects, scenes, sports, DIY, etc. Detailed descriptions of the benchmark can be found in [22]. Each video is manually labeled to one or more categories. In our experiments, we do not use the manual labels in training, but instead we automatically generate the web labels according to the concept name appearance in the video metadata. The manual labels are used only in testing to evaluate our and the baseline methods. Following [22], the standard train/test split is used. The second set is YFCC100M [45] which contains about 800,000 videos on Yahoo! Flickr with metadata such as the title, tags, the uploader, etc. There are no manual labels on this set and we automatically generate the curriculum from the metadata in a similar way. Since there are no annotations, we train the concept detectors on the most 101 frequent latent topics found in the metadata. There are totally 47,397 webly labeled videos on the 101 concepts for training.

On FCVID, as the manual labels are available, the performance is evaluated in terms of the precision of the top 5 and 10 ranked videos (P@5 and P@10) and mean Average Precision (mAP) of 239 concepts. On YFCC100M, since there are no manual labels, for evaluation, we apply the detectors to a third public video collection called TRECVID MED which includes 32,000 Internet videos [36]. We apply the detectors trained on YFCC100M to the TRECVID videos and manually annotate the top 10 detected videos returned by each method for 101 concepts.

Implementation Details We build our method on top of a pre-trained convolutional neural network as the low-level features (VGG network [39], except in the image experiment we use AlexNet [24] as in [47]). We extract the key-frame level features and create a video feature by the average pooling. The same features are used across different methods on each dataset. The concept detectors are trained based on a hinge loss cost function by SVM. Algorithm 1 is used to train the concept models iteratively and the λ stops increasing after 100 iterations. At each iteration, we apply a dropout of 0.5 when sampling negative samples. We automatically generate curriculum labels based on the video metadata, ASR, OCR and VGG net 1,000 classification results using latent topic modeling with word embedding matching as shown in Section 3.

Baselines in video domain experiment The proposed method is compared against the following five baseline methods which cover both the classical and the recent representative learning algorithms on webly-labeled data. *BatchTrain* trains a single SVM model using all samples in the multi-modal curriculum built with our method as described in section 3.2.2. *Self-Paced Learning (SPL)* is a classical method where the curriculum is generated by the learner itself [25]. *BabyLearning* is a recent method that simulates baby learning by starting with few training samples and fine-tuning using more weakly labeled videos crawled from the search engine [30]. *GoogleHNM* is a hard negative mining method proposed by Google [46]. It utilizes hard negative mining to train a second order mixture of experts model according to the video's YouTube topics. *FastImage* [16] is a video retrieval method that utilizes web images from search engine to match to the video with re-ranking. *WELL-MM* is the proposed method. The hyper-parameters of all methods including the baseline methods are tuned on the same validation set. On FCVID, the set is a standard development set with manual labels randomly selected from 10% of the training set (No training was done using ground truth labels) whereas on YFCC100M it is also a 10% proportion of noisy training set.

Table 1: Comparison of different curriculum using the BatchTrain learning method.

Method	P@5	P@10	mAP
WordHardMatching	0.782	0.763	0.469
YouTubeTopicAPI	0.587	0.563	0.315
SearchEngine	0.723	0.713	0.413
WordEmbedding	0.790	0.774	0.462
LatentTopic	0.731	0.716	0.409
WELL-MM	**0.838**	**0.820**	**0.486**

4.2 Experiments on FCVID

Curriculum Comparison As discussed in Section 3.2.2, we compare different ways to build curriculum for noisy label learning. Here we also compare their effectiveness by training concept detectors directly using the curriculum labels. The batch train model is used for all generated curriculum labels. In Table 1 we show the batch trained models' precision at 5, 10 and mean average precision on the test set of FCVID. For WELL-MM, we extract curriculum from different modalities as shown in Section 3.2.2, and combine them using linear weights. The weights are hyper-parameters that are tuned on the validation set, and the optimal weights for textual metadata, ASR, image classification and OCR results are 1.0, 0.5, 0.5 and 0.05, respectively. This attempt to combining curriculum from different modalities serves as a pilot study. However, experiments show that such simple linear weighting is already effective with WELL-MM. Further research in this direction is left for future work. We also compare WELL-MM with using only latent topic modeling and word embedding soft matching. Results show that the curriculum generated by combining latent topic modeling and word embedding using multi-modal prior knowledge is the most accurate, which indicates our claim of exploiting multi-modal information is beneficial.

Baseline Comparison Table 2 compares the precision and mAP of different methods where the best results are highlighted. As we see, the proposed WELL-MM significantly outperforms all baseline methods, with statistically significant difference at p-level of 0.05. Comparing SPL with BatchTrain, it shows that the self-paced learning model over-fits to the noise without prior knowledge and performs worse than the simple BatchTrain model. Comparing WELL-MM with SPL and BatchTrain, the effect of incorporating multi-modal curriculum makes a significant difference in terms of performance, which suggests the importance of prior knowledge and preventing over-fitting in webly learning. The promising experimental results substantiate the efficacy of the proposed method.

Robustness to Noise Comparison In this comparison we manually control the noise level of the curriculum in order to systematically verify how our methods would perform with respect to the noise level within the web data. To this end, we randomly select video samples with ground truth labels for each concept, so that

Figure 4: Illustration of representative videos selected by WELL-MM at different iterations

Table 2: Baseline comparison on FCVID

Method	P@5	P@10	mAP
BatchTrain	0.838	0.820	0.486
FastImage [16]	-	-	0.284
SPL [26]	0.793	0.754	0.414
GoogleHNM [46]	0.781	0.757	0.472
BabyLearning [30]	0.834	0.817	0.496
WELL-MM	**0.918**	**0.906**	**0.615**

the noise level of the curriculum labels are set at 20%, 40%, 60%, 80% and we fix the recall of all the labels. We then train WELL-MM using such curriculum and test them on the FCVID testing set. We also compare WELL-MM to three other methods with the same curriculum, among them *GoogleHNM* is a recent method to train video concept detector with large-scale data. We exclude *BabyLearning*, which relies on the returned results by the search engine, since in this experiment the curriculum is fixed. As shown in Table 3, as the noise level of the curriculum grows, WELL-MM maintains its performance while other methods drop significantly. Specifically, when the noise level of curriculum increased from 60% to 80%, other methods' mAP drops 46.5% on average while WELL-MM's mAP only drops 19.1% relatively. It shows that WELL-MM is robust against different level of noise, which shows great potential in larger scale webly-labeled learning as the dataset gets bigger, the noisier it may become.

Table 3: WELL-MM performance with curriculum consisting of multiple artificial noise levels.

Method \ Noise Level	20%	40%	60%	80%
BatchTrain	0.592	0.538	0.463	0.232
SPL	0.586	0.515	0.396	0.184
GoogleHNM	0.602	0.552	0.477	0.304
WELL-MM	**0.673**	**0.646**	**0.613**	**0.496**

Noisy Dataset Size Comparison To investigate the potential of concept learning on webly-labeled video data, we apply the methods on different sizes of subsets of the data. Specifically, we randomly split the FCVID training set into several subsets of 200, 500, 1,000, and 2,000 hours of videos, and train the models on each subset without using manual annotations. The models are then tested on the same test set. Table 4 lists the average results of each type of subsets. As we see, the accuracy of WELL-MM on webly-labeled data increases along with the growth of the size of noisy data while other webly learning methods' performance tend to be saturated.

Comparing to the methods trained using ground truth, In Table 4, WELL-MM trained using the whole dataset (2000 hours) outperforms Static CNN (trained using manual labels) using around 1400 hours of data and rDNN-F (trained using manual labels with three features) trained using around 450h of data. And since the incremental performance increase of WELL-MM is close to linear, we conclude that with sufficient webly-labeled videos (which are not hard to obtain) WELL-MM will be able to outperform the rDNN-F trained using 2000h of data, which is currently the largest manual labeled dataset.

Table 4: MAP comparison of models trained using web labels and ground-truth labels on different subsets of FCVID. The methods marked by * are trained using human annotated labels.

Method \ Dataset Size	200h	500h	1000h	2000h
BatchTrain	0.364	0.422	0.452	0.486
SPL [26]	0.327	0.379	0.403	0.414
GoogleHNM [46]	0.361	0.421	0.451	0.472
BabyLearning [30]	0.390	0.447	0.481	0.496
WELL-MM	0.487	0.554	0.595	0.615
Static CNN[22]*	0.485	0.561	0.604	0.638
rDNN-F[22]*	0.550	0.620	0.650	0.754

4.3 Experiments on CIFAR-10

Following [47], we generate synthesized noisy training data with a noise level of 30%, 40% and 50% on CIFAR-10 dataset. The models

Table 5: Experimental results on CIFAR-10

Methods \ Noise Level	30%	40%	50%
Noisy-CNN [42]	0.697	0.667	0.634
Massive-Learning [47]	0.698	0.668	0.630
WELL-MM	**0.709**	**0.700**	**0.682**

are trained on noisy data and tested on clean data. Classification Accuracy is reported. Our method doesn't assume any kind of noise distribution, while Noisy-CNN [42] assumes the noise distribution depends on classes and Massive-Learning [47] assumes it also depends on the image content. We show the experimental results in Table 5. The results show that WELL-MM outperforms the other methods at all noise levels. More interestingly, as the noise level rises from 30% to 50%, the performance of Massive-Learning [47] drops about 9.8%, while WELL-MM only drops 3.8%. It shows that WELL-MM can also effectively learn robust concept detectors in image domain.

4.4 Experiments on YFCC100M

In the experiments on YFCC100M, we train 101 concept detectors on YFCC100M and test them on the TRECVID MED dataset which includes 32,000 Internet videos. Since there are no manual labels, to evaluate the performance, we manually annotate the top 10 videos in the test set and report their precisions in Table 6. The MED evaluation is done by four annotators and the final results are averaged from all annotations. The Fleiss' Kappa value for these four annotators is 0.64. A similar pattern can be observed where the comparisons substantiate the rationality of the proposed webly learning framework. Besides, the promising results on the largest multimedia set YFCC100M verify the scalability of the proposed method.

Table 6: Baseline comparison on YFCC100M

Method	P@3	P@5	P@10
BatchTrain	0.535	0.513	0.487
SPL [26]	0.485	0.463	0.454
GoogleHNM [46]	0.541	0.525	0.500
BabyLearning [30]	0.548	0.519	0.466
WELL-MM	**0.667**	**0.663**	**0.649**

4.5 Qualitative Analysis

In this section we show training examples of WELL-MM. In Figure 4, we demonstrate the positive samples that WELL select at different stage of training the concept "baseball" and "birthday". For the concept "baseball", at the early stage (1/93, 25/93), WELL-MM selects easier and clearer samples such as the ones with camera directly pointing at the playground, while at the later stage (75/93, 93/93) WELL-MM starts to train with harder samples with different lighting conditions and untypical samples for the concept. For the concept "birthday", as we see, at later stage of the training, complex samples for birthday event like a video with two girl singing birthday song (75/84) and a video of celebrating birthday during hiking (84/84) are included in the training, while at the early stage, only typical "birthday" videos with birthday cake and candles are included.

5 CONCLUSIONS

In this paper, we proposed a novel method called WELL-MM for webly labeled video data learning. WELL-MM extracts multi-modal informative knowledge from noisy weakly labeled video data from the web through a general framework and achieves the best performance only using webly-labeled data on two major video datasets. The comprehensive experimental results demonstrate that WELL-MM outperforms state-of-the-art studies by a statically significant margin on learning concepts from noisy web video data. In addition, the results also verify that WELL-MM is robust to the level of noisiness in the video data. The result suggests that with more webly-labeled data, which is not hard to obtain, WELL-MM can potentially outperform models trained on any existing manually-labeled data.

ACKNOWLEDGEMENTS

This work was based in part on work supported by the National Science Foundation (NSF) under grant number IIS-1251187. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).
[2] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan M Shetty. 2013. *Nonlinear programming: theory and algorithms*. John Wiley & Sons.
[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*.
[4] Alessandro Bergamo and Lorenzo Torresani. 2010. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*.
[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.
[6] Shih-Fu Chang, Dan Ellis, Wei Jiang, Keansub Lee, Akira Yanagawa, Alexander C Loui, and Jiebo Luo. 2007. Large-scale multimodal semantic concept detection for consumer video. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*.
[7] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*.
[8] Xinlei Chen and Abhinav Gupta. 2015. Webly supervised learning of convolutional networks. In *ICCV*.
[9] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*.
[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.
[11] Santosh K Divvala, Alireza Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*.
[12] Lixin Duan, Dong Xu, IW-H Tsang, and Jiebo Luo. 2012. Visual event recognition in videos by learning from web data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34, 9 (2012), 1667–1680.
[13] Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. 2005. Learning object categories from Google's image search. In *ICCV*.
[14] Pierre Garrigues, Sachin Farfade, Hamid Izadinia, Kofi Boakye, and Yannis Kalantidis. 2016. Tag Prediction at Flickr: a View from the Darkroom. *arXiv preprint arXiv:1612.01922* (2016).
[15] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research* 66, 3 (2007), 373–407.

[16] Xintong Han, Bharat Singh, Vlad I Morariu, and Larry S Davis. 2015. Fast Automatic Video Retrieval using Web Images. *arXiv preprint arXiv:1512.03384* (2015).

[17] Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*. 856–864.

[18] Lu Jiang, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. 2014. Easy Samples First: Self-paced Reranking for Zero-Example Multimedia Search. In *MM*.

[19] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann. 2014. Self-Paced Learning with Diversity. In *NIPS*.

[20] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. 2015. Self-paced curriculum learning. In *AAAI*.

[21] Lu Jiang, Shoou-I Yu, Deyu Meng, Yi Yang, Teruko Mitamura, and Alexander G Hauptmann. 2015. Fast and accurate content-based semantic search in 100m internet videos. In *Proceedings of the 23rd ACM international conference on Multimedia*.

[22] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. 2015. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *arXiv preprint arXiv:1502.07209* (2015).

[23] Andrej Karpathy, George Toderici, Sachin Shetty, Tommy Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[25] M Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *NIPS*.

[26] M Pawan Kumar, Haithem Turki, Dan Preston, and Daphne Koller. 2011. Learning specific-class segmentation from diverse data. In *ICCV*.

[27] Li-Jia Li and Li Fei-Fei. 2010. Optimol: automatic online picture collection via incremental model learning. *International journal of computer vision* 88, 2 (2010), 147–168.

[28] Junwei Liang, Lu Jiang, Deyu Meng, and Alexander Hauptmann. 2016. Learning to Detect Concepts from Webly-Labeled Video Data. In *IJCAI*.

[29] Junwei Liang, Qin Jin, Xixi He, Gang Yang, Jieping Xu, and Xirong Li. 2014. Semantic Concept Annotation of Consumer Videos at Frame-Level Using Audio. In *Pacific Rim Conference on Multimedia*. Springer, 113–122.

[30] Xiaodan Liang, Si Liu, Yunchao Wei, Luoqi Liu, Liang Lin, and Shuicheng Yan. 2015. Towards computational baby learning: A weakly-supervised approach for object detection. In *ICCV*.

[31] Christopher D Manning, Prabhakar Raghavan, and others. *Introduction to information retrieval*. Vol. 1.

[32] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. 2010. *Lucene in Action: Covers Apache Lucene 3.0.* Manning Publications Co.

[33] Deyu Meng and Qian Zhao. 2015. What Objective Does Self-paced Learning Indeed Optimize? *arXiv preprint arXiv:1511.06049* (2015).

[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[35] T Mitchell, W Cohen, E Hruschka, P Talukdar, J Betteridge, A Carlson, B Dalvi, M Gardner, B Kisiel, J Krishnamurthy, and others. 2015. Never-Ending Learning. In *AAAI*.

[36] Paul Over, Jon Fiscus, Greg Sanders, David Joy, Martial Michel, George Awad, Alan Smeaton, Wessel Kraaij, and Georges Quénot. 2014. Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID*. 52.

[37] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, and others. 2011. The Kaldi speech recognition toolkit. (2011).

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.

[39] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[40] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *icdar*. IEEE, 629–633.

[41] Valentin I Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2009. Baby Steps: How fiLess is Morefi in unsupervised dependency parsing. *NIPS GRLL* (2009).

[42] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2014. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080* (2014).

[43] James Steven Supancic and Deva Ramanan. 2013. Self-paced learning for long-term tracking. In *CVPR*.

[44] Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. 2012. Shifting weights: Adapting object detectors from image to video. In *NIPS*.

[45] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The New Data and New Challenges in Multimedia Research. *arXiv preprint arXiv:1503.01817* (2015).

[46] Balakrishnan Varadarajan, George Toderici, Sudheendra Vijayanarasimhan, and Apostol Natsev. 2015. Efficient Large Scale Video Classification. *arXiv preprint arXiv:1505.06250* (2015).

[47] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[48] Qian Zhao, Deyu Meng, Lu Jiang, Qi Xie, Zongben Xu, and Alexander G Hauptmann. 2015. Self-Paced Learning for Matrix Factorization.. In *AAAI*.