Class Prototypes based Contrastive Learning for Classifying Multi-Label and Fine-Grained Educational Videos

Rohit Gupta [1] *   Anirban Roy [2]   Claire Christensen [2]   Sujeong Kim [2]   Sarah Gerard [2]

Madeline Cincebeaux [2]   Ajay Divakaran [2]   Todd Grindal [2]   Mubarak Shah [1]

[1] Center for Research in Computer Vision, University of Central Florida   [2] SRI International

rohitg@knights.ucf.edu, shah@crcv.ucf.edu   anirban.roy@sri.com

Abstract

The recent growth in the consumption of online media by children during early childhood necessitates data-driven tools enabling educators to filter out appropriate educational content for young learners. This paper presents an approach for detecting educational content in online videos. We focus on two widely used educational content classes: literacy and math. For each class, we choose prominent codes (sub-classes) based on the Common Core Standards. For example, literacy codes include 'letter names', 'letter sounds', and math codes include 'counting', 'sorting'. We pose this as a fine-grained multilabel classification problem as videos can contain multiple types of educational content and the content classes can get visually similar (e.g., 'letter names' vs 'letter sounds'). We propose a novel class prototypes based supervised contrastive learning approach that can handle fine-grained samples associated with multiple labels. We learn a class prototype for each class and a loss function is employed to minimize the distances between a class prototype and the samples from the class. Similarly, distances between a class prototype and the samples from other classes are maximized. As the alignment between visual and audio cues are crucial for effective comprehension, we consider a multimodal transformer network to capture the interaction between visual and audio cues in videos while learning the embedding for videos. For evaluation, we present a dataset, APPROVE, employing educational videos from YouTube labeled with fine-grained education classes by education researchers. APPROVE consists of 193 hours of expert-annotated videos with 19 classes. The proposed approach outperforms strong baselines on AP-PROVE and other benchmarks such as Youtube-8M, and COIN. The dataset is available at https://nusci.csl.sri.com/project/APPROVE.

---

1. Introduction

With the expansion of internet access and the ubiquitous availability of smart devices, children increasingly spend a significant amount of time watching online videos. A recent nationally representative survey reported that 89% of parents of children aged 11 or younger say their child watches videos on YouTube [4]. Moreover, it is estimated that young children in the age range of two to four years consume 2.5 hours and five to eight years consume 3.0 hours per day on average [45,46]. Childhood is typically a key period for education, especially for learning basic skills such as literacy and math [20, 25]. Unlike generic online videos, watching appropriate educational videos supports healthy child development and learning [7, 22, 23]. Thus, analyzing the content of these videos may help parents, teachers, and media developers increase young children's exposure to high-quality education videos, which has been shown to produce meaningful learning gains [22]. As the amount of online content produced grows exponentially, automated content understanding methods are essential to facilitate this.

In this work, given a video, our goal is to determine whether the video contains any educational content and characterize the content. Detecting educational content requires identifying multiple distinct types of content in a video while distinguishing between similar content types. The task is challenging as the education codes by Common Core Standards [3, 40] can be similar such as 'letter names' and 'letter sounds', where the former focuses on the name of the letter and the latter is based on the phonetic sound of the letter. Also, understanding education content requires analyzing both visual and audio cues simultaneously as both signals are to be present to ensure effective learning [3, 40]. This is in contrast to standard video classification benchmarks such as the sports or generic YouTube videos in UCF101 [54] Kinetics400 [53], YouTube-8M [1], where visual cues are often sufficient to detect the different classes. Finally, unlike standard well-known action videos,

education codes are more structured and not accessible to common users. Thus, it requires a carefully curated set of videos and expert annotations to create a dataset to enable a data-driven approach. In this work, we focus on two widely used educational content classes: literacy and math. For each class, we choose prominent codes (sub-classes) based on the Common Core Standards that outline age-appropriate learning standards [3,40]. For example, literacy codes include 'letter names', 'letter sounds', 'rhyming', and math codes include 'counting', 'addition subtraction', 'sorting', 'analyze shapes'.

We formulate the problem as a multilabel fine-grained video classification task as a video may contain multiple types of content that can be similar. We employ multi-modal cues since besides visual cues, audio cues provide important cues to distinguish between similar types of educational content. We propose a class prototypes based supervised contrastive learning approach to address the above-mentioned challenges. We learn a prototype embedding for each class. Then a loss function is employed to minimize the distance between a class prototype and the samples associated with the class label. Similarly, the distance between a class prototype and the samples without that class label is maximized. This is unlike the standard supervised contrastive learning setup where inter-class distance is maximized and intra-class distance is minimized by considering classwise positive and negative samples. This approach is shown to be effective for single-label setups [26]. However, it is not straightforward to extend this for the proposed multilabel setup as samples cannot be identified as positive or negative due to the multiple labels. We jointly learn the embedding of the class prototypes and the samples. The embeddings are learned by a multimodal transformer network (MTN) that captures the interaction between visual and audio cues in videos. We employ automatic speech recognition (ASR) to transcribe text from the audio. The MTN consists of video and text encoders that learn modality-specific embedding and a cross-attention mechanism is employed to capture the interaction between them. The MTN is end-to-end learned through the contrastive loss.

Due to the lack of suitable datasets for evaluating fine-grained classification of education videos, we propose a new dataset, called APPROVE, of curated YouTube videos annotated with educational content. We follow Common Core Standards [3, 40] to select education content suitable for the kindergarten level. We consider two high-level classes of educational content: literacy and math. For each of these content classes, we select a set of codes. For the literacy class, we select 7 codes and for the math class, we select 11 codes. Each video is associated with multiple labels corresponding to these codes. The videos are annotated by trained education researchers following standard validation protocol [41] to ensure correctness. APPROVE also con-

sists of carefully chosen background videos, i.e., without educational content, that are visually similar to the videos with educational content. APPROVE consists of 193 hours of expert-annotated videos with 19 classes (7 literacy codes, 11 math codes, and a background) where each video has 3 labels on average.

Our contributions can be summarized as follows:

- APPROVE, a fine-grained multi-label dataset of education videos, to promote exploration in this field.
- Class prototypes based contrastive learning framework along with a multi-modal fusion transformer suitable for the problem where videos have multiple fine-grained labels.
- Outperforming relevant baselines on three datasets: APPROVE, YouTube-8M [1] and COIN [55].

2. Related Works

Self-Supervised Contastive Learning (CL) has been an effective paradigm for visual representation learning. Methods such as SimCLR [8], MoCo (Momentum Contrastive learning) [18], Augmented Multiscale Deep InfoMax (AMDIM) [5], Contrastive Predictive Coding (CPC) [39], MoCov2 [10], MoCov3 [12] and SimCLRv2 [9] have achieved strong performance on image classification benchmarks. The shared property between these CL frameworks is that data augmentation is used to generate positive pairs for CL from a single instance, where other data instances are treated as negatives. Prototypical Contrastive Learning (PCL) [30] extends self-supervised contrastive learning with the idea of clustering data representations during training to generate unsupervised prototypes which represent *intra-class variation*. We utilize *class prototypes* instead in the supervised setting, to learn fine-grained distinctions *between classes*.

Supervised CL methods such as SupCon [26] utilize labels to enhance contrastive learning by forming positive and negative pairs using labels instead of data augmentation. Supervised Contrastive Learning has also been used for other tasks such as image segmentation [57] and classification in the presence of noisy labels [32]. Hierarchical CL [63] extends SupCon to the hierarchical classification case. However, SupCon cannot be extended to the multi-label case in a straightforward manner, as pairs of data samples with multiple labels cannot be clearly classified just into positives and negatives.

Weakly-Supervised Multi-Modal CL: Weakly aligned text-image/video datasets scraped from the web such as Conceptual Captions [50] and WebVid-10M [6] enable learning of multi-modal representations. CLIP [42] applies a cross-modal contrastive loss to train individual text and image encoders. Everything at Once [51] is able to additionally utilize the audio modality and incorporates a pairwise fusion encoder which encodes pairs of modalities, as a re-
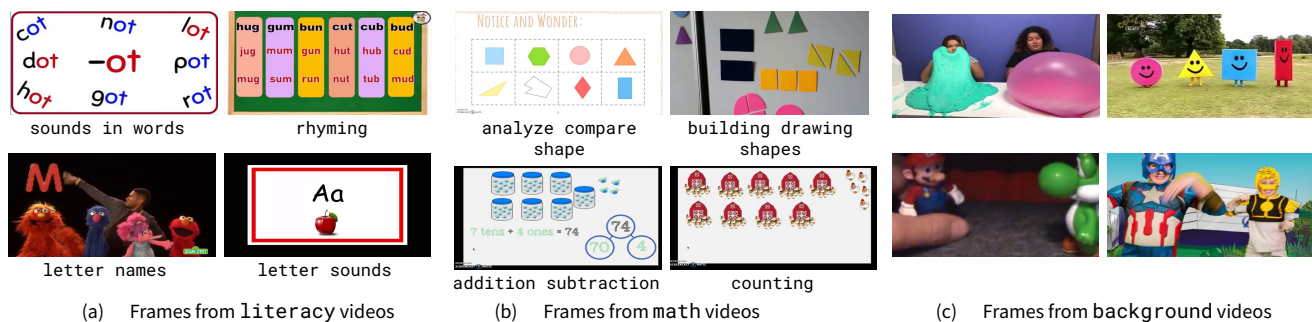
(a) Frames from `literacy` videos   (b) Frames from `math` videos   (c) Frames from `background` videos

Figure 1. Sample video frames from the APPROVE dataset. Videos belong to the **(a) literacy classes**, **(b) math classes**, and **(c) background**. Background videos do not contain educational content but share visual similarities with educational videos. The videos are labeled with fine-grained sub-classes, e.g., `letter names` vs `letter sounds`.

sult, 6 forward passes of the fusion model are required for 3 modalities. Frozen in Time [6] is able to utilize both image-text and video-text datasets through the use of a Space-Time Transformer Visual Encoder. Visual Conditioned GPT [37] uses a single cross-attention fusion layer to combine pre-trained CLIP text and visual features. Flamingo [2] adds cross-attention layers interleaved with language decoder layers to fuse visual information into text generation. MER-LOT [60,61] and Triple Contrastive Learning [58] combine contrastive learning and generative language modeling to learn aligned text-image representations.

Supervised Multi-Modal Learning: Supervised Multi-Modal Learning typically relies on crowd-captioned datasets such as Flickr30k [59] and MS-COCO Captions [11]. Some prior works such as OSCAR [33] and VinVL [62] have utilized pre-trained object detectors and multi-modal transformers to learn image captioning using supervised aligned datasets. BLIP [28] takes a hybrid approach where it bootstraps an image captioner using a labeled dataset and uses it to generate captions for web images. This generated corpus is then filtered and used for learning an aligned representation. ALign BEfore Fuse [29] highlights the importance of aligning text and image tokens before fusing them using a multi-modal transformer.

In this paper, we focus on the fine-grained classification of multilabel educational videos. Due to the lack of suitable datasets, we propose a new dataset, APPROVE, which is described next.

3. APPROVE Dataset

We propose a dataset, called APPROVE, of curated YouTube videos annotated with educational content. AP-PROVE consists of 193 hours of expert-annotated videos with 19 classes (7 literacy codes, 11 math, and background) and each video is associated with approximately 3 labels on average. We follow the Common Core Standards [3,40] to select education content suitable for kindergarten level. The Common Core Standards outline what students are expected to know and do at various age ranges and grades. This is a widely accepted standard followed by a range of

educators. We consider two high-level classes of educational content: literacy and math. For each of these content classes, we select a set of codes. For the literacy class, we select 7 codes including letter names, letter sounds, follow words, sight words, letters in words, sounds in words, and rhyming. For the math class, we select 11 codes including counting, individual number, comparing groups, addition subtraction, measurable attributes, sorting, spatial language, shape identification, building drawing, analyzing and comparing shapes. More details about the standard and the description of the codes are provided in the supplementary material. APPROVE also consists of carefully chosen background videos, i.e., without educational content, that are visually similar to the videos with educational content. We present frames corresponding to these classes in Fig. 1.
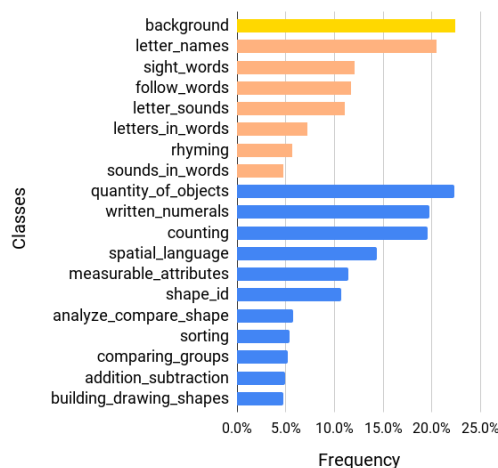

Figure 2. Frequency of the classes in APPROVE.
Math codes are in **Orange** and literacy codes in **Blue**.

To ensure the quality and correctness of the annotations, we consider educational researchers to annotate the videos and follow a standard validation protocol [41]. Each annotator is trained by an expert and annotations on a selected set are examined before engaging the annotator for the final annotation. Annotators start once they reach more than 90% agreement with the expert. Further, we estimate inter-
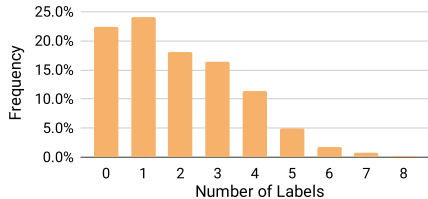
Figure 3. Distribution of the number of labels per video.

annotator consistency to filter out anomalies. Details about the validation process are provided in the supplementary material. It takes a month to train an education researcher to match expert-level coding accuracy. On average, it takes the trained annotators 1 min to annotate 1 min of video.

The videos are curated from YouTube and are annotated by the trained annotators to determine educational content in them. Each video can have multiple class labels that are quite similar making the task a multi-label and fine-grained classification problem. For example 'letter names' and 'letter sounds' where visual letters are shown in both but in 'letter sounds', the phonetic sound on the letter is emphasized (Fig. 1 (a)). Similarly, in both 'build and draw shapes' and 'analyzing and comparing shapes', multiple shapes can appear but the latter focuses on comparing multiple shapes by shape and size (Fig. 1 (b)). Class-wise stats are presented in Fig. 2. Note that the task is different from common video classification setups where either multi-label or fine-grained aspects are dealt separately. Single-label datasets such as HMDB51 [27], UCF101 [54], Kinetics700 [53] and multi-label ones such as Charades [52] are widely used benchmarks for this problem. YouTube-Birds and YouTube-Cars [64] are analogous datasets for object recognition from videos. Multi-Sports [34] and FineGym [49] label fine-grained action classes for sports. HVU [15] also adds scenes and attributes annotations along with action and objects. However, action, object and scene recognition are not enough for fine-grained video understanding. For instance, videos from a given education provider might share similar objects (person, chalkboard, etc.) and actions (writing on chalkboard) while covering different topics (counting, shape recognition etc.) in each video.

4. Proposed Approach

In this section, we first describe the proposed class prototypes based contrastive learning framework suitable for videos containing multiple educational codes. Then we present the approach to learning the class prototypes and finally describe the multimodal transformer network that learns features by fusing visual and text cues from videos.

4.1. Class prototypes based contrastive learning

In a contrastive learning framework, feature representations are typically learned by simultaneously minimizing the distance between positive samples and maximizing

| Dataset	Size (in hr)	Multi-Label	Fine Grained	Type	Annotators
Action Recognition					
HMDB	5	✗	✗	V	Authors
UCF	27	✗	✗	V	Authors
Kinetics	800	✗	✗	V+A	Crowd
Video Classification					
COIN	476	✗	✗	V+A	Crowd
YT-8M	-	✔	✗	F	Machine
APPROVE	193	✔	✔	V+T+A	**Experts**

Table 1. APPROVE dataset compared with selected prior datasets. V→Video Frames, A→Audio, T→Text, F→Features only.

the distance between negative samples (See Figure 4.(a)). The positive and negative samples are determined with respect to an anchor sample usually based on the class labels. For example, supervised contrastive learning (SupCon) [26] learns a representation to minimize the intra-class distances and maximize inter-class distances. We denote x_i and y_i as the ith sample and its label, respectively. Let's define z_i as the representation of the ith sample in a batch A, and $sim(z_i, z_j) = \frac{z_i \cdot z_j}{|z_i||z_j|}$ the cosine similarity, then the SupCon loss [26] is defined as:

$$\mathcal{L}_{\text{SupCon}} = \sum_{i \in A} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(sim(z_i, z_p)/\tau)}{\sum_{a \in A \setminus i} \exp(sim(z_i, z_a)/\tau)},$$ (1)

where $P(i)$ is the set of positive samples, i.e., with the same label as z_i, in the batch excluding i and $a \in A \setminus i$ is the index of all samples in the batch excluding the ith sample. τ is a scalar temperature parameter used for scaling similarity values. The positive pairs are grouped into the numerator and minimizing the loss minimizes their distance in the learned representation and vice versa for negative pairs. SupCon is known to be effective for classifying samples with a single label. However, it is not straightforward to extend this for the multilabel setup as beyond positive samples, where all labels are the same, and negative samples, where none of the labels is the same, there can be a third scenario where labels are partially overlapping. Though SupCon has been extended to hierarchical classification [63], it cannot be directly extended to the true multi-label case.

To address this issue, we learn class prototypes as the representative for each class and consider these as anchors while determining positive and negative samples. Specifically, for a specific class prototype, a representation is learned to minimize distances between the prototype and samples with this class label and maximize the distances between the prototype and samples without this class label. We compare the proposed approach with the standard single-label contrastive learning in figure 4. We iteratively update the class prototypes while learning the feature rep-
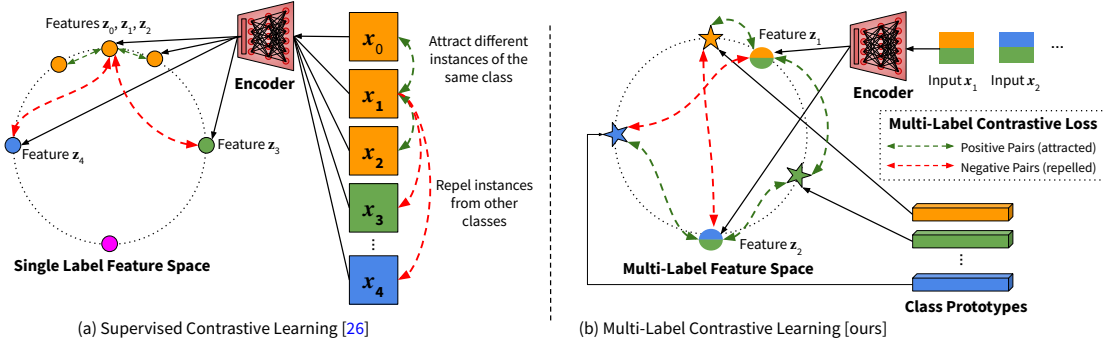
Figure 4. Contrastive learning operates on the feature space by bringing the representations of similar samples close and pushing distinct samples apart. Prior work in **(a) Supervised Contrastive Learning** [26] trains the network by treating instances from the same class as positive pairs and instances from different classes as negative pairs. This approach doesn't generalize to multi-label classification tasks, as some instance pairs have partially overlapping labels. We propose the use of class prototypes to enable **(b) Multi-Label Prototypes Contrastive Learning**. Each sample and the class prototypes corresponding to the labels associated with the sample are treated as positive pairs. Similarly, negative pairs are determined based on the missing class labels. Prototypes are represented by stars ($\star$) and inputs as circles ($\circ$) colored with all their relevant labels. We discuss strategies for initializing and learning the label prototypes in Sec. 4.2.

resentations. We define $C = \{c_1, \ldots, c_K\}$ as the set of classes where K is the number of classes. For a sample $\boldsymbol{x}$, let's define $P_{ml}(\boldsymbol{x}) = \{c_k^+\}, c_k^+ \in C$ as the set of multiple class labels associated with $\boldsymbol{x}$ (positive classes) and $c_k^- \in C \setminus P_{ml}(\boldsymbol{x})$ denotes the missing classes (negative classes). We define $CP = \{\boldsymbol{cp}_1, \ldots, \boldsymbol{cp}_K\}$ as the set of class prototypes. Considering $\boldsymbol{z}$ is the representation for the sample $\boldsymbol{x}$, the class prototypes based multilabel contrastive loss is defined as:

$$\mathcal{L}_{mlc}(\boldsymbol{x}) =$$
$$\frac{-1}{|P_{ml}(\boldsymbol{x})|} \sum_{c_k^+ \in P_{ml}(\boldsymbol{x})} \left[\log \frac{\exp(sim(\boldsymbol{z}, \boldsymbol{cp}_k)/\tau)}{\sum\limits_{c_j^- \in C \setminus P_{ml}(\boldsymbol{x})} \exp(sim(\boldsymbol{z}, \boldsymbol{cp}_j)/\tau)} \right].$$
$$(2)$$

Here, minimizing the loss of the positive class prototype and instance pairs in the numerator minimizes the distance between the representation z and the class prototypes corresponding to the sample, and vice versa for negative classes. We also utilize negative sampling to account for the class imbalance between positives and negatives.

4.2. Learning class prototypes

We aim to learn class-specific prototypes such that the multilabel samples can be thought of as the combinations of the class prototypes selected based on the associated labels. Lets assume Z_t is an $N \times d$ matrix of d-dimensional representations, i.e., $\boldsymbol{z}$s), of N samples and $L \in \{0, 1\}^{N \times K}$ is corresponding labels matrix with K classes. Let's denote CP_t is a matrix of size $K \times d$ of K class prototypes at a training iteration t. Then, $Z_t = L \times CP_t + \varepsilon$, where ε is the residual noise term. Assuming a Gaussian noise that is unbiased and uncorrelated with the labels L, the class prototypes can be approximated as $CP_t^* \approx (L^T L)^{-1} L^T Z_t$, where operation $(L^T L)$ results in a square matrix amenable to inversion. Note that for single labels, this implies averag-

ing the features of the instances belonging to a given class as the prototype for that class. In a multi-label setup, additionally, the co-occurrence between the labels is considered. Finally, the class prototypes are updated with learning iterations as:

$$CP_{t+1} = \beta \cdot CP_t + (1 - \beta) \cdot CP_t^*, \qquad (3)$$

where β is the decay parameter for the exponential moving averaging. The moving averaging avoids collapsing prototypes across the training iterations [31].

4.3. Multi-Modal Fusion Transformer

Given a sample video $\boldsymbol{x}$, we consider multimodal cues such as visual and audio cues to learn its representation $\boldsymbol{z}$. Considering multimodal cues is crucial for content recognition, specifically for education videos as effective comprehension requires attending to both visual demonstration and audio explaining the educational content [3, 40]. To capture audio cues, we consider the audio track of the video and extract speech by removing the background such as tunes or instruments [21]. Then we employ the automatic speech recognition (ASR) technique to transcribe text from the speech [43]. We notice that separating speech from the background is important for an accurate ASR transcription. Given the video frames and text transcription, we propose a multimodal transformer network (MTN) to fuse these cues using cross-modal attention.

Our MTN (Figure 5) has three components: image encoder, text encoder, and fusion encoder to learn visual ($\boldsymbol{z}^v$), text ($\boldsymbol{z}^t$), and fusion ($\boldsymbol{z}^f$) representations, respectively. The sample representation is comprised of these three representations $\boldsymbol{z} = \{\boldsymbol{z}^v, \boldsymbol{z}^t, \boldsymbol{z}^f\}$. The image encoder is implemented by a vision transformer (ViT) [16] that learns frame embeddings from the video frames along with a special CLS for each frame [14]. We pool the CLS tokens across
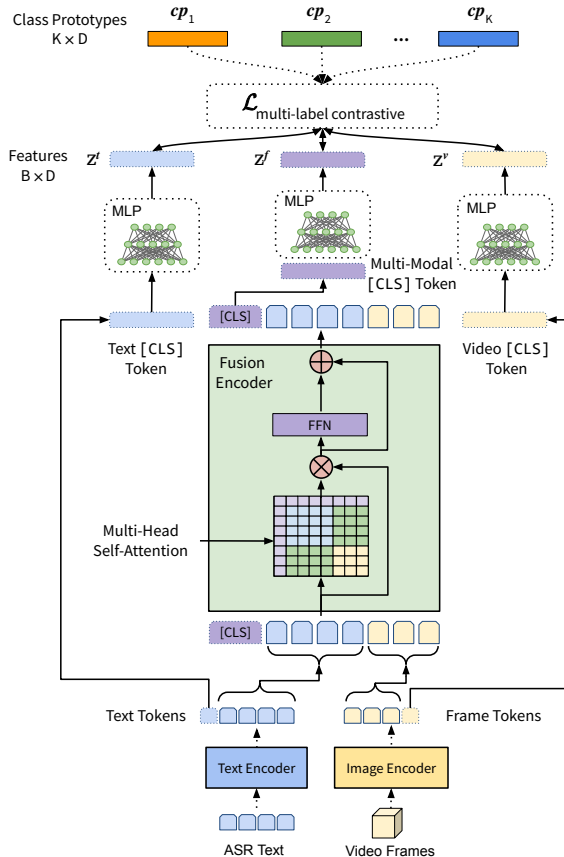
Figure 5. Multi-Modal Classification Network. A text encoder is used to encode ASR text from the video, while an Image Encoder is used to get tokens representing each frame of the video. Unimodal pre-training is carried out on the text & image encoders respectively. Multi-label contrastive loss is used along with shared prototypes to align the representations across both modalities. This is followed by joint end-to-end learning of the whole multimodal network including the fusion encoder which applies multihead self-attention within and across the modalities. The prototypes are further refined during the multi-modal training phase.

the frames and consider this as a compact video representation. Similarly, we consider the BERT-based text transformer [24] to learn the word embeddings along with the text CLS token for the entire transcription. We consider the text CLS as the text representation. Finally, the fusion encoder fuses the visual and text cues by leveraging cross-modal attention between frame and word embeddings.

4.4. Inference using prototypes

Existing approaches in contrastive learning [8, 26] usually discard the MLP layer after the contrastive training, and a linear classifier is trained on the frozen backbone network. We rely on the class prototypes to carry out inference by utilizing the cosine distance between the learned prototypes and test features. Given our prototype loss-based training, the estimated probability of a given class should be proportional to the normalized temperature scaled cosine distance

(Equation 4). In practice, we normalize the cosine distance such that -1 and 1 correspond to a confidence of 0 & 1 respectively. Thus the prediction is made following:

$$\hat{p}(k|\boldsymbol{x}) \propto \exp(sim(\boldsymbol{z}, \boldsymbol{cp}_k)/\tau), \qquad (4)$$

where $\boldsymbol{z}$ is the multimodal representation of the sample $\boldsymbol{x}$.

4.5. Implementation Details

Training strategy for the multimodal fusion: We follow a two-stage training process: during the initial unimodal training phase, we utilize *fixed* prototypes in each modality to align the representations. Then in the second stage, we train the unimodal encoders and the multi-modal fusion encoder end-to-end. The cross-modal alignment learned during the first stage improves the learning of the multi-modal representation. The multi-modal learning phase consists of alternating optimization steps of training the network using our contrastive loss per Sec. 4.1 and refining the class prototypes as described in Sec. 4.2.

Image encoder: For video frames, `Random Resized Crop` and `RandAugment` [13] augmentations are used from `torchvision`. We use ImageNet pretrained vision encoders ResNet50 [19], ViT-B/32 (224×224 resolution) and ViT-B/16 (384×384 resolution) [17].

Text encoder: We generate text from the videos using Whisper [43], an open source ASR model. For data augmentation, we generate four versions of the ASR text by back-translation using the `Helsinki-NLP/opus-mt-{en-de, en-nl, en-fr}` models through the `nlpaug` library [38]. Synonym replacement, text span removal and random word swapping augmentations are also used for the text data. We use DistilBERT-Base-uncased [48], and t5-small [44] from HuggingFace `transformers` library as the text encoder.

Optimizer: We employ AdamW optimizer [36] for training with a learning rate of 0.0005. Weight decay of 1e-6 is utilized only on the MLP head during contrastive training and the classifier during BCE/Focal/Asym. loss. Since pre-trained vision and text backbones are used, the backbone learning rate is set to 1/10th of the learning rate for the head. Exponential Moving Averaging every 10 steps with a decay of 0.999 was used for the model parameters.

5. Experiments

Datasets. In addition to the proposed APPROVE dataset, we evaluate our approach on a subset of Youtube-8M [1] and COIN [55] datasets. YT-8M consists of a diverse set of YouTube videos with video and audio modalities. We consider a subset of YT-8M dataset with 46K videos and 165 classes. COIN consists of instructional videos covering a wide variety of domains and spanning over 180 classes.

Baselines. We compare the efficacy of our multilabel classification framework against the following baselines:

Binary cross-entropy: In this baseline, loss for multiple labels is computed by combining the binary cross-entropy losses for individual classes.

Focal loss [35]: This considers a modified binary cross-entropy to assign a higher weight to hard samples by adjusting a focusing parameter γ. Negative samples can also be down-weighted by using a weight α. The focal loss for a positive label is given as $\mathcal{L}_{focal}(p) = -\alpha(1-p)^{\gamma}log(p)$. We set $\gamma = 2$ and $\alpha = 0.2$ in our experiments.

Asymmetric loss [47]: This builds upon the focal loss by utilizing different focusing parameters γ_+ and γ_- for positive and negative labels. It also ignores the negative samples with a prediction probability lower than a margin m. Asymmetric Loss for prediction p corresponding to label y is given as: $\mathcal{L}_{asym}(p, y) = -yL_+ - (1-y)L_-$, where $L_+ = (1-p)^{\gamma_+} \cdot \log(p)$ and $L_- = (\max(p-m, 0))^{\gamma_-} \cdot \log(1 - \max(p-m, 0))$. We follow the 5-step procedure recommended by the original authors to train this baseline. We experimentally set $\{\gamma_- = 2, \gamma_+ = 1, m = 0.1\}$ corresponding to the best performance on APPROVE.

Metrics. In order to develop a reliable education content detection framework, achieving high precision is crucial. Thus we consider Recall @ 80% Precision (R@80) as the primary metric. We also consider the standard area under the precision-recall curve (AUPR) that is not sensitive to a specific threshold for making the final prediction. We also consider a label ranking average precision (LRAP) [56] metric that is more suitable for the multilabel setup. This estimates whether the ground truth classes are predicted with higher scores than the rest:

$$\text{LRAP} = \frac{1}{n}\sum_{i=1}^{m}\frac{1}{|Y_i|}\sum_{\lambda \in Y_i}\frac{|\lambda' \in Y_i : rnk_i(\lambda') \le rnk_i(\lambda)|}{rnk_i(\lambda)},$$

where $rnk_i(\lambda)$ is the predicted rank of class λ for sample i. LRAP is a ranking metric and in independent of a threshold.

Results on APPROVE. We compare the proposed approach with the baselines in Tab. 2. Our approach outperforms the strongest baselines by 3.1% and 2.3% with respect to R@80 and AUPR, respectively. We also present results for separate models trained on Math and Literacy subsets of Approve respectively. Results on the Math subset are higher compared to the Literacy subset, which indicates the literacy classes are harder to distinguish mostly due to the high inter-class similarity. The top three hardest classes are `follow_words`, `letters_in_words`, and `sounds_in_words` and these are from the literacy set.

Results on YT-46K and COIN. Beyond the APPROVE dataset, we also test our approach on two public datasets: YT-46K and COIN. Here we provide the results for the multi-modal models and the results for the single-modality models are in Section D of the Supplementary Material.

Results on YT-46K are provided in Table 3. As YT-8M

Subset	Modality	Method	AUPR	LRAP	R@80
All	V	BCE	45.5	54.3	6.9
		Focal	45.9	56.6	15.0
		Ours	**46.7**	**57.9**	**19.6**
	T	BCE	79.8	85.1	63.3
		Focal	79.9	85.7	72.8
		Ours	**82.5**	**87.4**	**75.4**
	V+T	BCE	84.3	88.4	76.3
		Focal [35]	86.1	89.1	82.2
		Asym. [47]	86.0	89.2	82.4
		Ours	**88.4** +2.3	**90.7** +1.5	**85.5** +3.1
MTH	V+T	BCE	86.3	92.4	80.3
		Focal	87.2	92.1	82.4
		Ours	**88.4** +1.2	**93.2** +1.1	**83.2** +0.8
LIT	V+T	BCE	72.1	82.9	50.7
		Focal	72.7	83.5	50.9
		Ours	**73.6** +0.9	**84.7** +1.2	**54.7** +3.8

Table 2. Results on APPROVE dataset. All metrics in %. V→Video & T→Text. M→ Math & L→ Literacy Subsets.

Modality	Method	AUPR	LRAP	R@80
V+T	BCE	64.6	70.2	42.3
V+T	Focal [35]	69.7	72.7	44.6
V+T	**Ours**	**70.9** +1.2	**74.9** +2.2	**49.1** +4.5

Table 3. Results on YT-46K. V→Video Frames and T→Text.

Modality	Method	Top-1 Accuracy
V+T	CE	53.7
V+T	BCE	54.9
V+T	Focal [35]	56.1
V+T	SupCon [26]	54.7
V+T	**Ours**	**57.5** +1.4

Table 4. Results on COIN. V→Video Frames and T→Text.

was primarily collected with the intention of visual classification, the additional use of text data leads to a smaller improvement compared to APPROVE. Results on COIN are provided in Table 4. As each video from COIN is mapped to a single task. Thus, we consider the Top-1 accuracy as the metric. On COIN we compare our approach with SupCon [26] which is effective for single labels. Note that our approach outperforms SupCon and this justifies the effectiveness of the class prototypes based training in a generic contrastive learning framework.

5.1. Ablation Studies

We perform the following ablation studies to quantify the impact of our approach for learning class prototypes, the multimodal fusion module, and the choice of visual and text encoding frameworks.

Table 5. Ablation studies

(a) Class Prototypes

Variant	APPROVE	COIN
Random	84.1	56.6
Orthogonal	84.8	57.0
Learned	85.5	57.5
Hierarchical	86.0	57.8

(b) Fusion Encoder Size

Layers	APPROVE	COIN
1	84.9	57.1
2	85.5	57.5
4	85.3	57.6

(c) Vision Encoder

Vision Model	APPROVE	COIN
R50	84.8	55.2
ViT-B/32	85.5	57.5
ViT-B/16	83.8	57.8

(d) Text Encoder

Text Model	APPROVE	COIN
DistilBert-B	85.5	57.5
t5-S	87.3	57.9

Table 5. Ablation studies. R@80% Precision for APPROVE and Top-1 Accuracy for COIN. Default setup is underlined.

Table 6. Robustness analysis

(a) Noisy modalities.

% missing	APPROVE	COIN
0%	85.5	57.5
10% V	80.1	53.3
10% T	75.8	57.2
30% T	68.9	42.8

(b) Run-to-Run Variance

Method	APPROVE	COIN
BCE	76.3 ± 0.7	54.9 ± 0.6
Focal	82.2 ± 0.5	56.1 ± 0.3
Ours	85.5 ± 0.5	57.5 ± 0.8

(c) Initialization

Method	APPROVE	COIN
ImageNet & Wiki-en+TBC	85.5	57.5
CLIP	86.7	63.5

Table 6. Robustness analysis. Our method is robust to partially missing modality and has similar run-to-run variance as baselines.

Learning class prototypes: We compare the two strategies where after initializing the class prototypes, we 1) keep them fixed and learn only the multimodal embedding of the samples, and 2) class prototypes and sample embedding are learned iteratively. The initialization can be done either randomly, or with orthogonal constraints. We note that orthogonal initialization performs best in our experiments and iterative adjusting the class prototypes achieves better performance as shown in Table 5 (a). We also consider hierarchical prototypes, for APPROVE, using a 2-level hierarchy where the first level consists of 18 classes, and the second level is the 3 super-classes: math, literacy, and background. The 180 task categories of COIN are organized into 12 domains in the taxonomy provided with the dataset. This hierarchy imposes an additional constraint on learning the embeddings during training.

Fusion Encoder: This evaluates the effect of the number of layers in the fusion encoder on the final performance. As expected. the performance improves with more layers and saturates around 4 layers (Table 5 (b)).

Vision Encoder: This evaluates the effect of the image encoder on the final performance. We consider ResNet [19] and ViT variants [16]. We notice that the ViT-B/16-384 encoder works well for the larger COIN dataset, whereas the ViT-B/32-224 encoder works best for APPROVE (Table 5 (c)).

Text Encoder: This evaluates the effect of the text encoder on the final performance. We test DistilBERT and T5 backbones for the text encoder. BERT is trained to predict masked spans of text. T5's unsupervised objective is similar, however, it trains on predicting the entire sequence instead of just the masked spans. GPT2 takes an auto-regressive approach to language modeling (Table 5 (d)).

5.2. Robustness analysis

We evaluate the robustness of our approach as follows:

Noisy modality: YouTube videos may have noisy modalities where some of the video frames are missing or ASR transcription is noisy. We show that our approach is robust against the cases where a percentage of video frames or text words are missing as shown in Tab 6(a).

Run-to-Run variance: The low variance across runs (Tab 6(b)) indicates that our approach is not sensitive to random initialization of the class prototypes.

Initializing the encoders: We consider the ImageNet pretraining for the image encoder, while English Wikipedia + Toronto Book Corpus is used to pre-train the text encoder. We provide results where the backbones are initialized with CLIP [42], which provides a more aligned vision-text representation. As expected, the results are better with the CLIP initialization (Tab 6(c)). The improvements are more significant on COIN than APPROVE as CLIP models may not be exposed to educational videos.

6. Conclusion

We have proposed an approach for detecting educational content in online videos. The problem is formulated as a fine-grained multilabel video classification task and we have considered class-prototypes based contrastive learning to address this. We have employed a multimodal transformer network to fuse visual and audio cues. This is crucial for comprehending educational content as both visual and audio cues are to be aligned to ensure effective comprehension. Our approach is shown to be effective in distinguishing fine-grained educational content with high interclass similarity. We have introduced APPROVE - a dataset with 193 hours of expert-annotated educational videos. Beyond APPROVE, we have evaluated our approach on COIN and YouTube-8M datasets where our approach outperforms the competitive baselines.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1, 2, 6

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 3

[3] National Governors Association et al. Common core state standards. *Washington, DC*, 2010. 1, 2, 3, 5

[4] Brooke Auxier, Monica Anderson Andrew Perrin, and Erica Turner. Parenting children in the age of screens. Technical report, Pew Research Center, 2020. 1

[5] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[6] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 2, 3

[7] Johanna Burkhardt and Wolfgang Lenhard. A meta-analysis on the longitudinal, age-dependent effects of violent video games on aggression. *Media Psychology*, 25(3):499–512, 2022. 1

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 6

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020. 2

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 2

[11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3

[12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. 2

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 6

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[15] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision*, pages 593–610. Springer, 2020. 4

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 5, 8

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 8

[20] Lowry Hemphill and Terrence Tivnan. The importance of early vocabulary for literacy achievement in high-poverty schools. *Journal of Education for Students Placed at Risk*, 13(4):426–451, 2008. 1

[21] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. Deezer Research. 5

[22] Lisa B Hurwitz. Getting a read on ready to learn media: A meta-analytic review of effects on literacy. *Child Development*, 90(5):1754–1771, 2019. 1

[23] Lisa B Hurwitz and Kelly L Schmitt. Raising readers with ready to learn: A six-year follow-up to an early educational computer game intervention. *Computers in Human Behavior*, 104:106176, 2020. 1

[24] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 6

[25] Nancy C Jordan, David Kaplan, Chaitanya Ramineni, and Maria N Locuniak. Early math matters: kindergarten number competence and later mathematics outcomes. *Developmental psychology*, 45(3):850, 2009. 1

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2, 4, 5, 6, 7

[27] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 4

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[29] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 3

[30] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021. 2

[31] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C.H. Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021. 5

[32] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 316–325, June 2022. 2

[33] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 3

[34] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13536–13545, 2021. 4

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[37] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. A frustratingly simple approach for end-to-end image captioning, 2022. 3

[38] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019. 6

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[40] Andrew Porter, Jennifer McMaken, Jun Hwang, and Rui Yang. Common core standards: The new us intended curriculum. *Educational researcher*, 40(3):103–116, 2011. 1, 2, 3, 5

[41] J. S. Radesky, A. Schaller, S. L. Yeo, , H. M. Weeks, and M. B. Robb. Young kids and youtube: How ads, toys, and games dominate viewing. common sense media. Technical report, Common Sense Media, 2020. 2, 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 8

[43] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. Technical report, Tech. Rep., Technical report, OpenAI, 2022. 5, 6

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 6

[45] Victoria Rideout and Michael B. Robb. The common sense census: Media use by tweens and teens, Oct 2019. 1

[46] V Rideout and M. B. Robb. The common sense census: Media use by tweens and teens. common sense media. Technical report, Common Sense Media, 2019. 1

[47] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021. 7

[48] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 6

[49] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020. 4

[50] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[51] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multimodal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022. 2

[52] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 510–526, Cham, 2016. Springer International Publishing. 4

[53] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset, 2020. 1, 4

[54] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 4

[55] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 6

[56] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. *Mining Multi-label Data*, pages 667–685. Springer US, Boston, MA, 2010. 7

[57] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7303–7313, 2021. 2

[58] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 3

[59] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3

[60] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16375–16387, June 2022. 3

[61] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 3

[62] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021. 3

[63] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16660–16669, June 2022. 2, 4

[64] Chen Zhu, Xiao Tan, Feng Zhou, Xiao Liu, Kaiyu Yue, Errui Ding, and Yi Ma. Fine-grained video categorization with redundancy reduction attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 136–152, 2018. 4