JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

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

Rohit Gupta[1] · Anirban Roy[2] · Claire Christensen[2] · Sujeong Kim[2]
Sarah Gerard[2] · Madeline Cincebeaux[2] · Todd Grindal[2]
Ajay Divakaran[2] · Mubarak Shah[1]

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

[2] SRI International

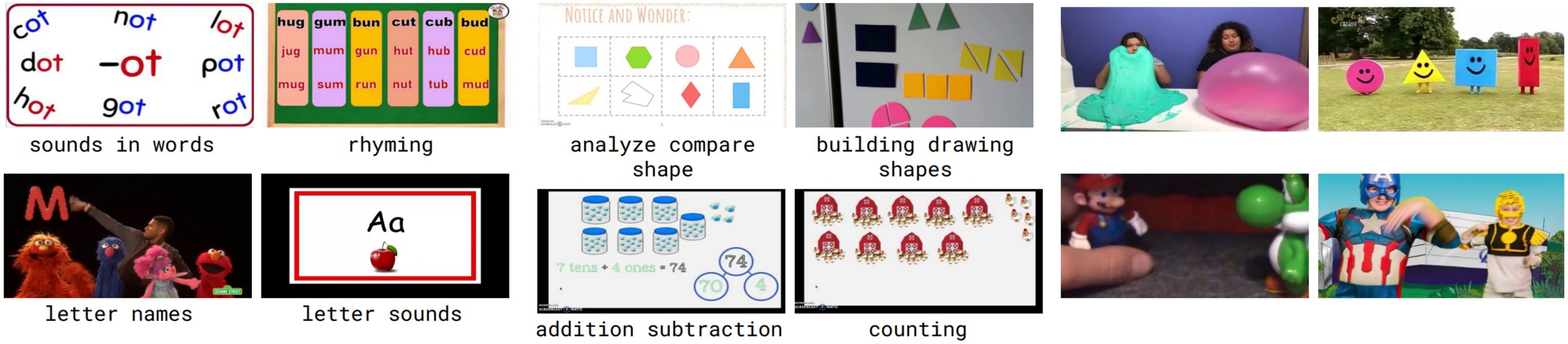
Motivation

- Young children, two to four years old, consume 2.5 hours of online video per day on average.
- Watching appropriate educational videos supports healthy child development and learning

APPROVE Dataset

- Curated educational YouTube videos expert-annotated into 19 classes (7 literacy codes, 11 math, and background)
- 193 hours

Fine-Grained Education Code Labels



(a) Frames from literacy videos (b) Frames from math videos (c) Frames from background videos

Challenges Addressed

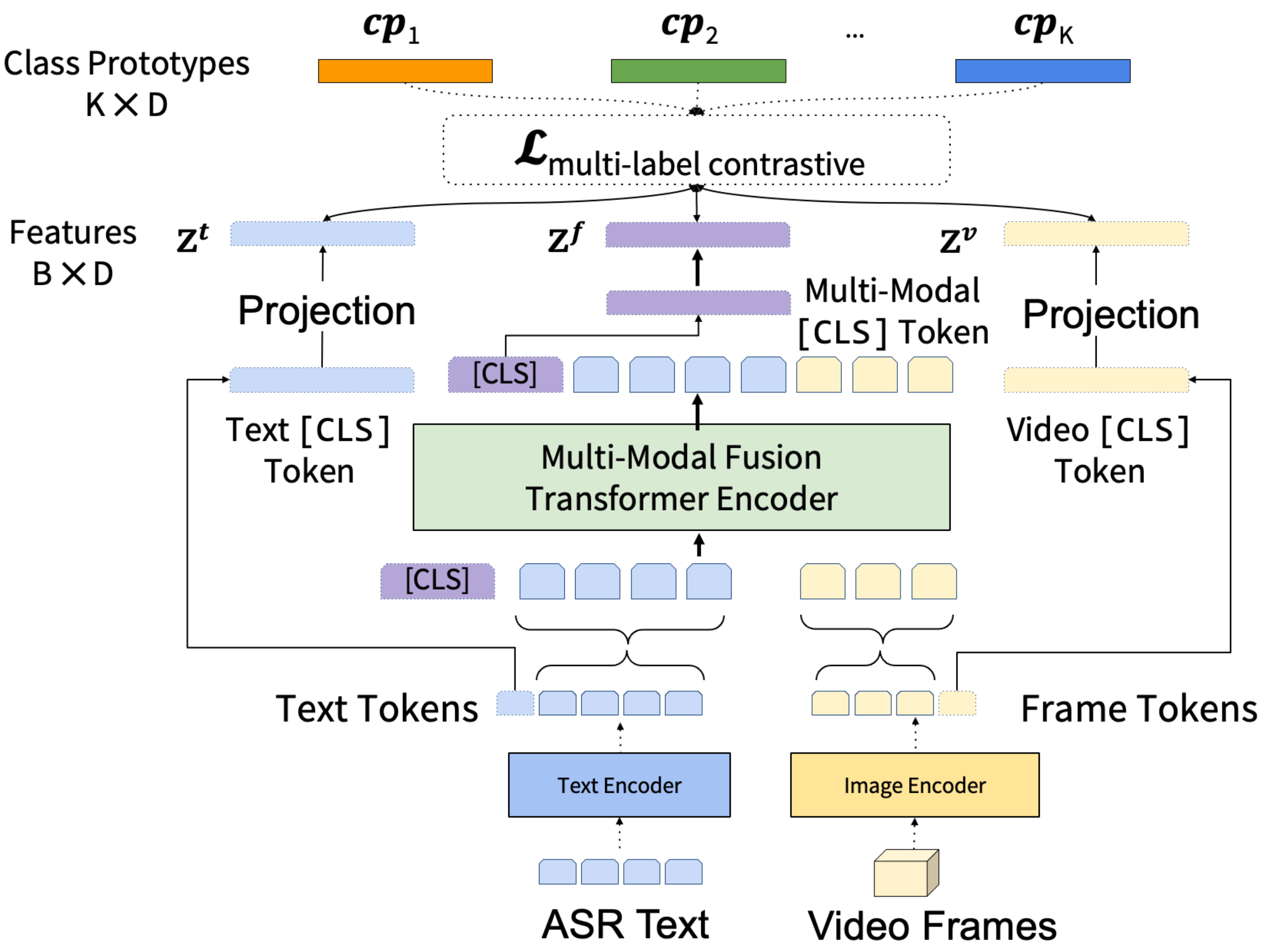
- Fine-grained classification requires multi-modal understanding
- Supervised Contrastive Learning is limited to single label case

Proposed Approach

Class Prototype Contrastive Learning solves two problems in one shot:

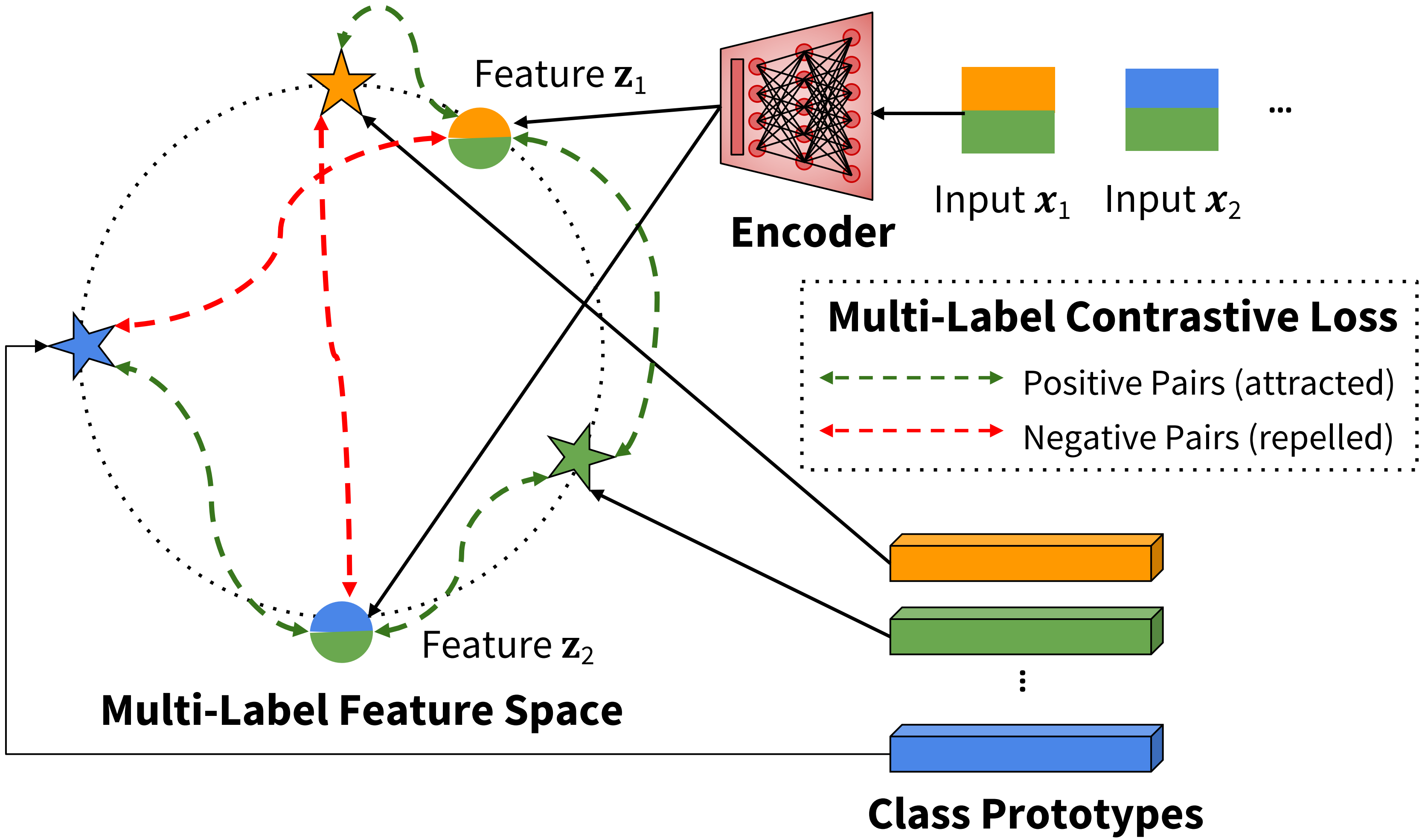
1. Use of shared prototypes across modalities allows for alignment
   - Features for video across modalities are pulled together
2. Generalizing Contrastive Learning to multi-label Setting can be achieved through use of class prototypes
   - Video features are attracted towards class prototypes of labels which are present in the video and repelled from labels that are not present

Overview



cp_1
cp_2
...
cp_K
Class Prototypes
K × D
$\mathcal{L}_{\text{multi-label contrastive}}$
Features
B × D
z^t
z^f
z^v
Projection
Multi-Modal
[CLS] Token
Projection
[CLS]
Text [CLS]
Token
Multi-Modal Fusion
Transformer Encoder
Video [CLS]
Token
[CLS]
Text Tokens
Frame Tokens
Text Encoder
Image Encoder
ASR Text
Video Frames

Multi-Label Contrastive Learning



Results: APPROVE

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 [36]	86.1	89.1	82.2
		Asym. [48]	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.

Results: YT-8M (1% subset)

Modality	Method	AUPR	LRAP	R@80
V+T	BCE	64.6	70.2	42.3
V+T	Focal [36]	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.

Results: COIN

Modality	Method	Top-1 Accuracy
V+T	CE	53.7
V+T	BCE	54.9
V+T	Focal [36]	56.1
V+T	SupCon [27]	54.7
V+T	**Ours**	**57.5** +1.4

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