Detecting Educational Content in Online Videos by Combining Multimodal Cues

Anirban Roy, Sujeong Kim, Claire Christensen, Madeline Cincebeaux
SRI International

NEURAL INFORMATION PROCESSING SYSTEMS

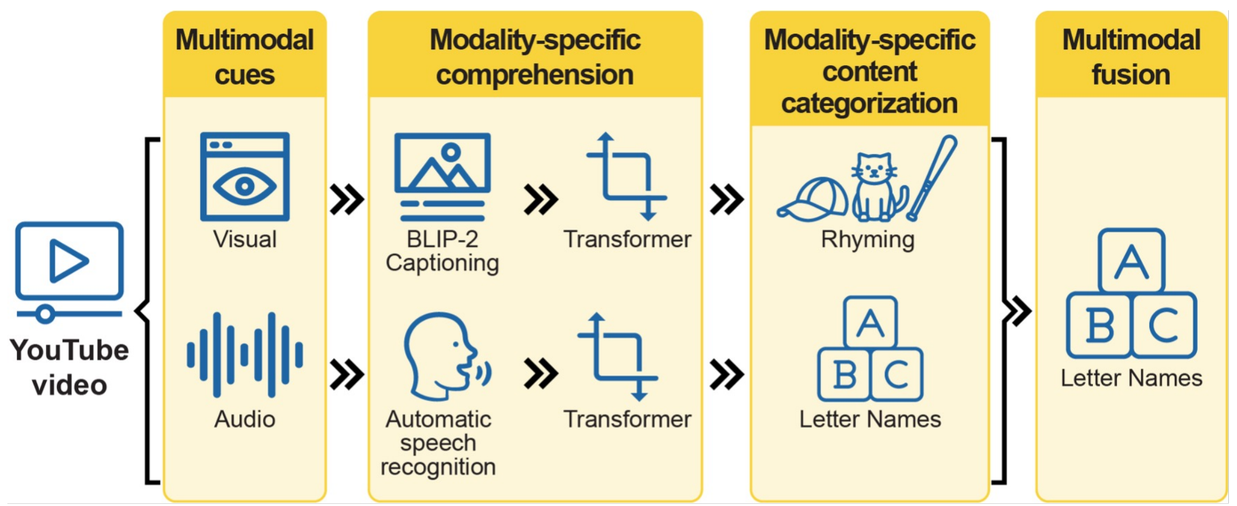
Motivation

- Recent surveys show 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 [Pew Research, 7/28/20; Rideout and Robb, 2019]
- Watching appropriate educational videos supports healthy child development and learning [Burkhardt and Lenhard, 2022; Hurwitz and Schmitt, 2020]

Contributions

- We propose a multimodal framework to combine visual, textual, and audio cues to detect educational content in online videos
- We leverage existing vision, text, and audio foundational models to extract and process the multimodal cues.
- We evaluate the proposes framework on various literacy and math codes following the common core standards.

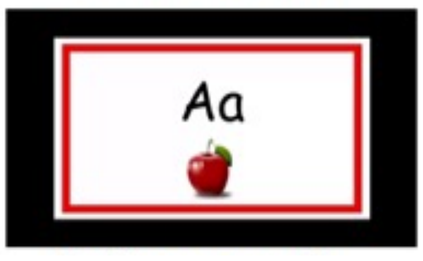
Multimodal framework for detecting educational contents



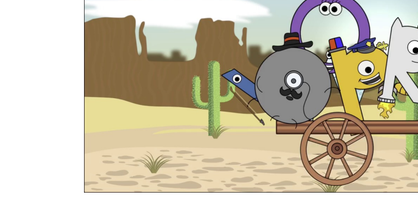
Proposed multimodal framework consists of two components.
Top: processing visual cues, bottom: processing audio cues.
Both the cues are combined to make final predictions.

Education content classes

Literacy

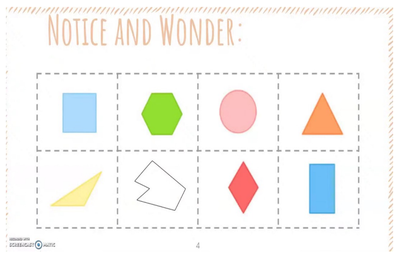


Letter sounds

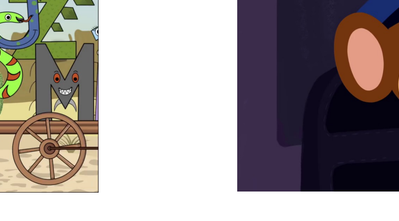


Sounds in words

Math codes

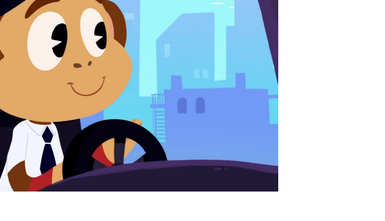


Analyze and compare shape



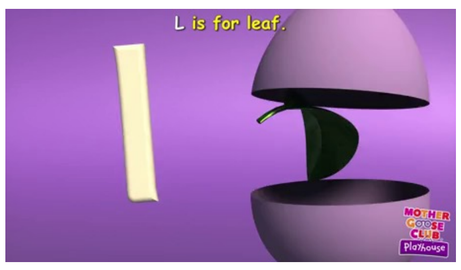
Building and drawing shapes

Background



Caption to capture visual cues

Caption: a cartoon image of a hat with the letter c on it



Caption: A leaf and a letter L on a purple background

Caption: a cartoon wagon with many different characters on it

Caption: a cartoon monkey in uniform driving a car

Processing audio cues

Video

Speech extraction

```
Mother Goose, Club
Playhouse  Phonics song.    A is
for apple. I'm whole the is for
ball ball.  Z is for cat.  Yeah, D
is for Dish Dish.   He is for egg,
and F is 45.  Five.  G is for gift.
Give H is for heart heart.  I is
for Igloo, it Igloo. J is for jar,
just just dark.  hey, is for car, I
might  L is for lie, low leave  M
is for Moon. Mama, mama, and is for
needle, the needle. So is for
allof. Olive, he is four parts
Part.   Q is for question.
Question. Question question ...
```

Speech to text conversion

Experiments and Results

Experiments are conducted on the APPROVE dataset [Gupta et al., 2023] containing literacy and math contents for pre-kindergarten and kindergarten levels

Results for the literacy codes

Literacy classes	Audio only	Video only	Multimodal
Follow words	69	67	86
Sight words	74	72	75
Letter sounds	66	52	64
Sounds in words	72	65	74
Letter names	82	78	83
Letter in words	53	49	61
Rhyming	98	85	98
Average	74	67	77

Accuracy (%) for literacy classes at the pre-kindergarten level.

Literacy classes	Audio only	Video only	Multimodal
Follow words	73	79	76
Sight words	57	63	64
Letter sounds	76	67	79
Sounds in words	76	62	73
Letter names	83	76	81
Letter in words	55	55	59
Rhyming	98	96	100
Average	74	71	76

Accuracy (%) for literacy classes at the kindergarten level.

Results for the math codes

Math classes	Audio only	Video only	Multimodal
Counting	86	67	84
Written numerals	80	70	81
Addition and Subtraction	98	73	98
Building and drawing shapes	89	71	83
Shape identification	89	62	87
Subitizing	73	60	77
Patterns	93	89	93
Cardinality	66	56	73
Analyzing and comparing shapes	89	51	89
Comparing groups	95	51	89
Measurable attributes	88	61	90
Sorting	87	60	90
Spatial language	79	67	78
Average	85	64	86

Accuracy (%) for math classes at the pre-kindergarten level.

Math classes	Audio only	Video only	Multimodal
Counting	83	69	83
Written numerals	79	72	81
Addition and Subtraction	98	84	98
Building and drawing shapes	77	56	87
Shape identification	88	71	89
Cardinality	66	54	70
Analyzing and comparing shapes	89	63	89
Comparing groups	97	72	97
Measurable attributes	91	57	91
Sorting	95	69	93
Spatial language	79	64	79
Average	86	66	87

Accuracy (%) for math classes at the kindergarten level.

Acknowledgment: NSF grant 2139219