Evaluating Picture Description Speech for Dementia Detection using Image-text Alignment

Youxiang Zhu[1], Nana Lin[1], Xiaohui Liang[1], John A. Batsis[2], Robert M. Roth[3], Brian MacWhinney[4]

[1]University of Massachusetts Boston

[2]University of North Carolina

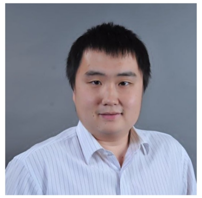
[3]Geisel School of Medicine at Dartmouth

[4]Carnegie Mellon University

This research is funded by the US National Institutes of Health National Institute on Aging, under grant No. R01AG067416.

Our team & Research

- NIH NIA R01 “SCH: INT: Exploiting Voice Assistant Systems for Early Detection of Cognitive Decline”, started on 2019.9.30-now
- Our research (past & ongoing): Speech-based dementia detection with:
  - Active speech tasks (including picture description)
  - Daily use/interaction with voice assistant (like Amazon Alexa)
  - Chat with ChatGPT



Xiaohui Liang

Associate Professor of Computer Science

University of Massachusetts Boston

John A. Batsis

Associate Professor, Division of Geriatric Medicine, School of Medicine, Department of Nutrition, The Gillings School of Global Public Health

University of North Carolina (UNC) at Chapel Hill

Robert M. Roth

Associate Professor of Psychiatry

Geisel School of Medicine at Dartmouth

Aleksandra C. Stark

Assistant Professor of Neurology

Geisel School of Medicine at Dartmouth

Brian MacWhinney

Professor of Psychology

Carnegie Mellon University

David Kotz

Professor of Computer Science

Dartmouth College

Youxiang Zhu

PhD Student

University of Massachusetts Boston

Nana Lin

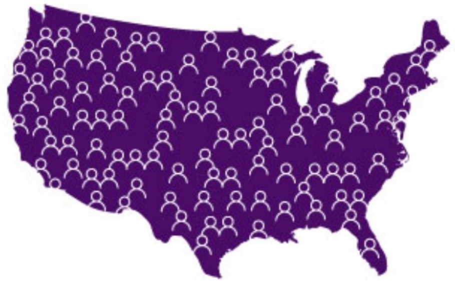
PhD Student

University of Massachusetts Boston

UMass Boston

Introduction - Dementia

- Dementia is a common disease for older adults
- Traditional diagnosis methods are costly and time consuming



More than **6 million** Americans are living with Alzheimer's. By 2050, this number is projected to rise to nearly 13 million.

1 in 3 seniors dies with Alzheimer's or another dementia. It kills more than breast cancer and prostate cancer combined.

In 2022, Alzheimer's and other dementias will cost the nation **$321 billion**. By 2050, these costs could reach nearly $1 trillion.

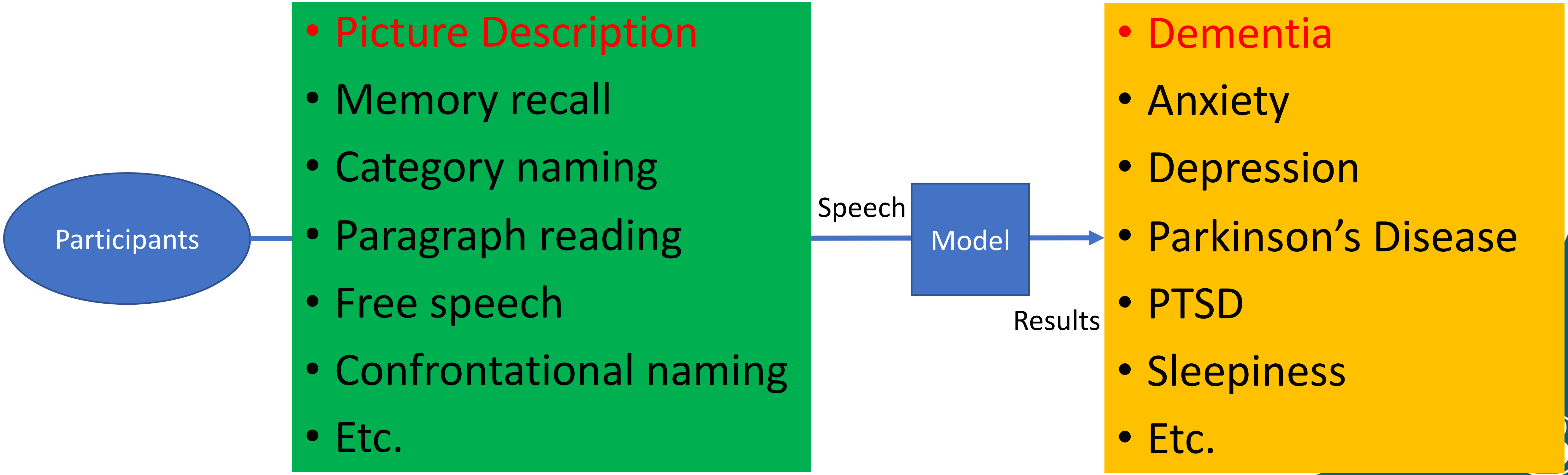
90% of physicians say it's important to diagnose MCI due to Alzheimer's, but **over half** say they are not fully comfortable diagnosing it.

https://www.alz.org/alzheimers-dementia/facts-figures

UMass Boston

Introduction – Speech-based health diagnosis

- Low cost
- Applicable for many diseases



Participants

- Picture Description
- Memory recall
- Category naming
- Paragraph reading
- Free speech
- Confrontational naming
- Etc.

Speech

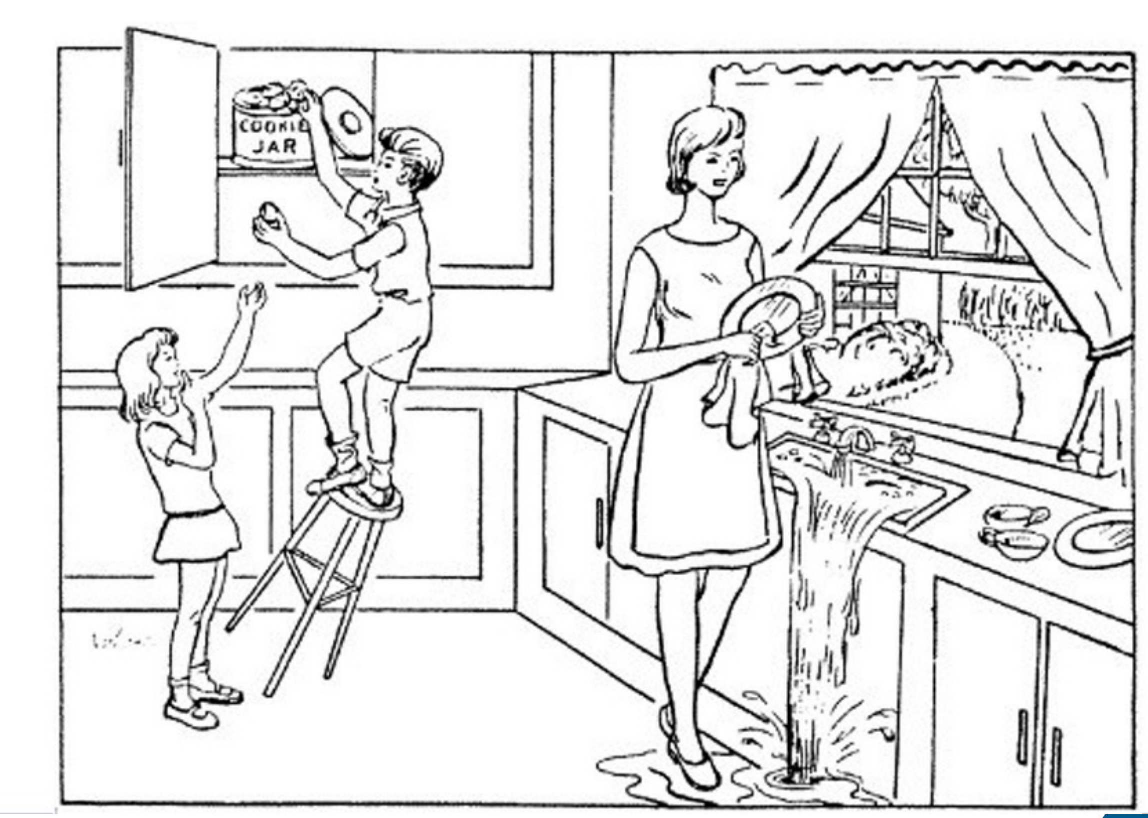
Model

Results

- Dementia
- Anxiety
- Depression
- Parkinson's Disease
- PTSD
- Sleepiness
- Etc.

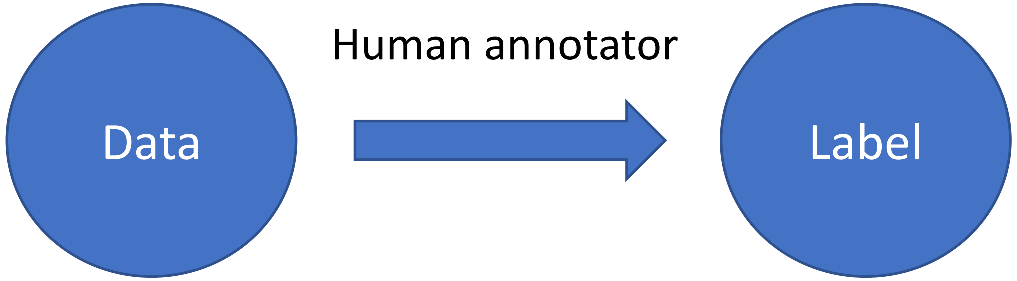
Cookie Theft Picture

- Participants are required to descript the picture via spontaneous speech
- Researchers aim to identify whether participants have dementia with such spontaneous speech



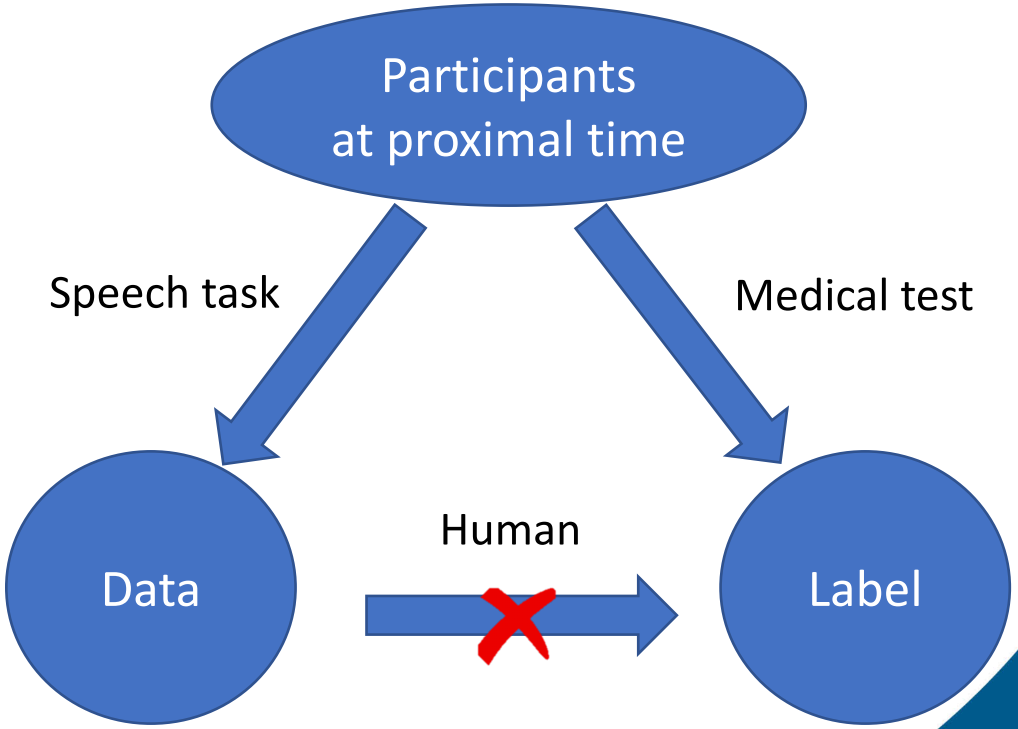
Dementia detection is challenging!

CV/NLP/SPEECH problems



We are building models to mimic human behavior.

Speech-based Health Diagnosis



We are expecting models to do something beyond human capacity!

ADReSS 2020 and ADReSSo 2021 Dataset

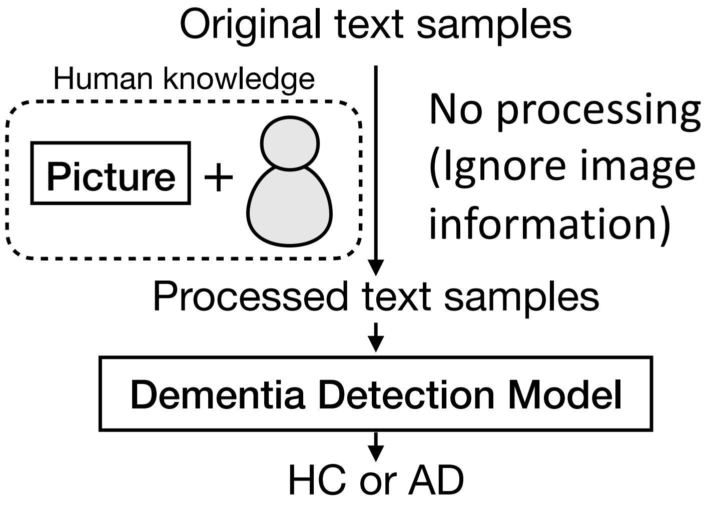
- All based on description of the Cookie Theft picture
- Balanced in AD (Alzheimer’s Dementia) / HC (healthy control), age, gender
- Standard train / test split
- Additional MMSE (an AD test) label

- ADReSS 2020
  - 108 training, 48 testing
  - Offer speech recording and manual transcription
- ADReSSo 2021
  - 166 training, 71 testing
  - Offer speech recording only
  - Additional cognitive decline (disease progression) inference task

Luz, S., Haider, F., Fuente, S.d.l., Fromm, D., MacWhinney, B. (2020) Alzheimer’s Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. Proc. Interspeech 2020
Luz, Saturnino, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. "Detecting cognitive decline using speech only: The adresso challenge." arXiv preprint arXiv:2104.09356 (2021).

Previous works on picture description task

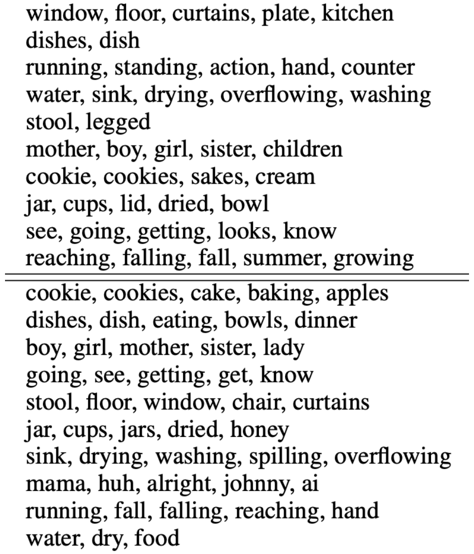
- Finding difference between AD and HC samples using speech and text only – ignore the image information
- Using human defined knowledge to interpret the image (e.g., information units)



Previous work using Picture information

- Drawbacks of previous work: Human defined words/sub-images, limited correlation between image and text information
- May be biased and take time consuming human efforts

A set of human defined words
Human labeled image-text correlation

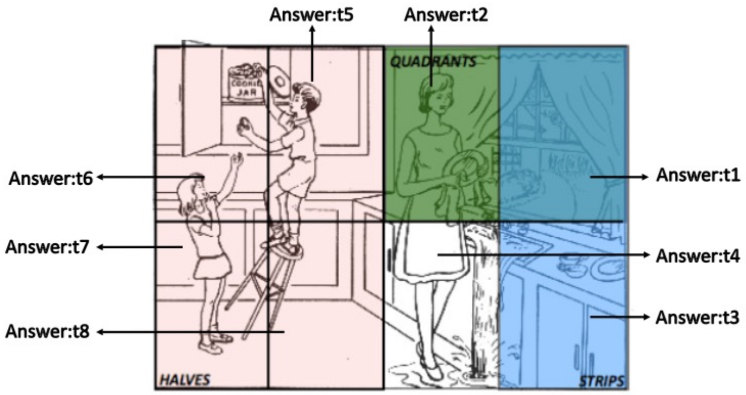


window, floor, curtains, plate, kitchen
dishes, dish
running, standing, action, hand, counter
water, sink, drying, overflowing, washing
stool, legged
mother, boy, girl, sister, children
cookie, cookies, sakes, cream
jar, cups, lid, dried, bowl
see, going, getting, looks, know
reaching, falling, fall, summer, growing

cookie, cookies, cake, baking, apples
dishes, dish, eating, bowls, dinner
boy, girl, mother, sister, lady
going, see, getting, get, know
stool, floor, window, chair, curtains
jar, cups, jars, dried, honey
sink, drying, washing, spilling, overflowing
mama, huh, alright, johnny, ai
running, fall, falling, reaching, hand
water, dry, food

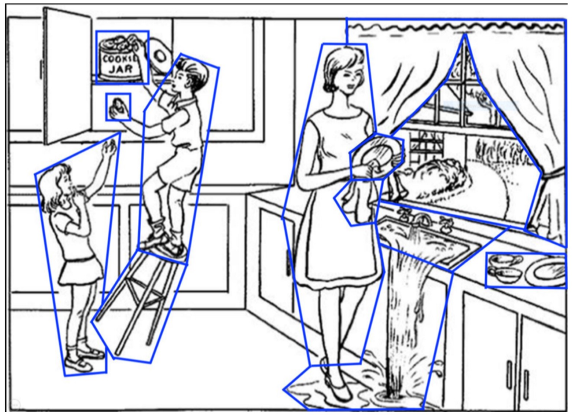
Information units
[Yancheva et al., ACL 2016]

8 Human defined areas
Human labeled image-text correlation



Dialogue acts
[Farzana et al., LREC 2020]

13 Human defined Areas
No image-text correlation



Eye tracking
[Barral et al., MLHC 2020]

Previous work vs. our work

- Previous works:
  - Finding difference between AD and HC samples using speech and text only – ignore the image information
  - Using human defined knowledge to interpret the image (e.g., information units)

- Our work: using pre-trained image-text alignment model (i.e., CLIP) to process the information from the image
- Reduce human bias and efforts

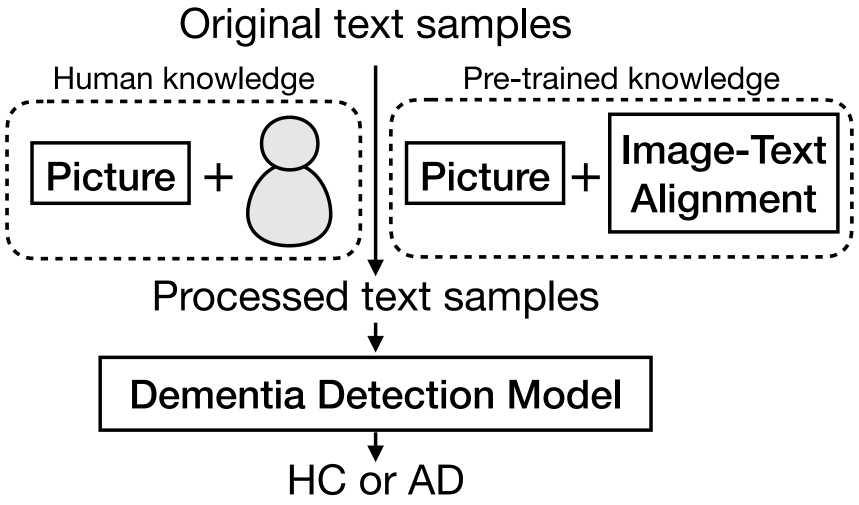
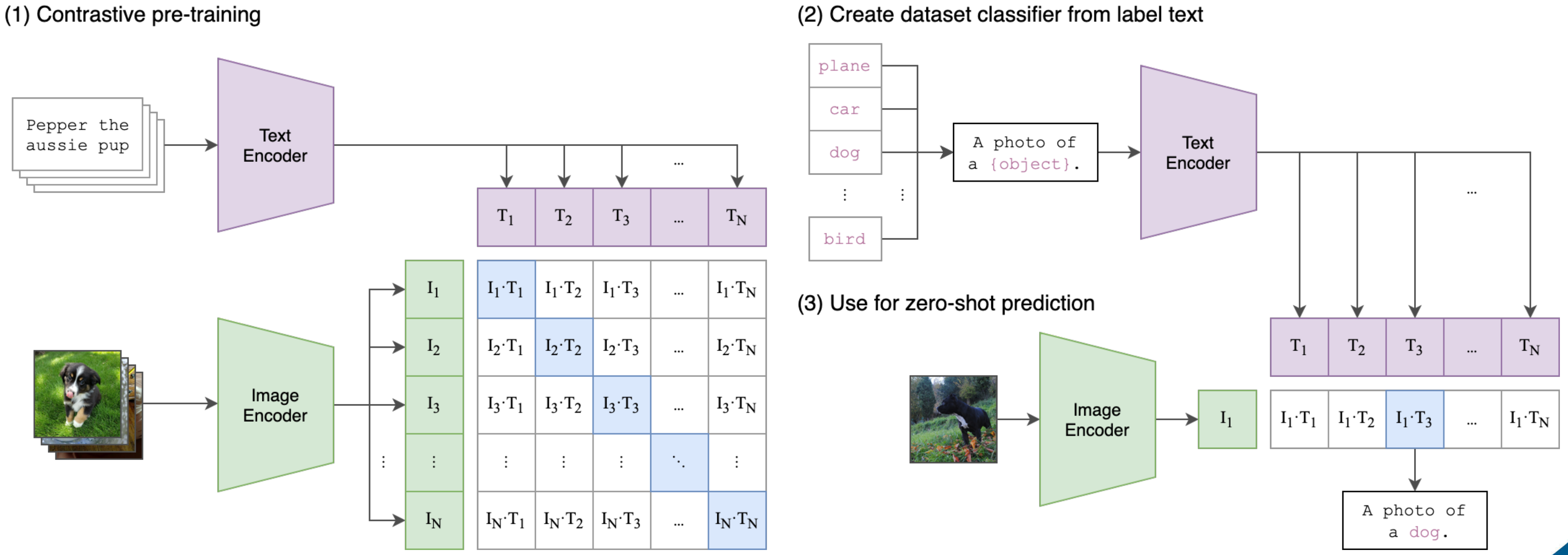


Image-text alignment – CLIP model

- Image-to-texts match: relevance scores of one image and multiple texts
- Text-to-image match: relevance scores of one text and multiple images

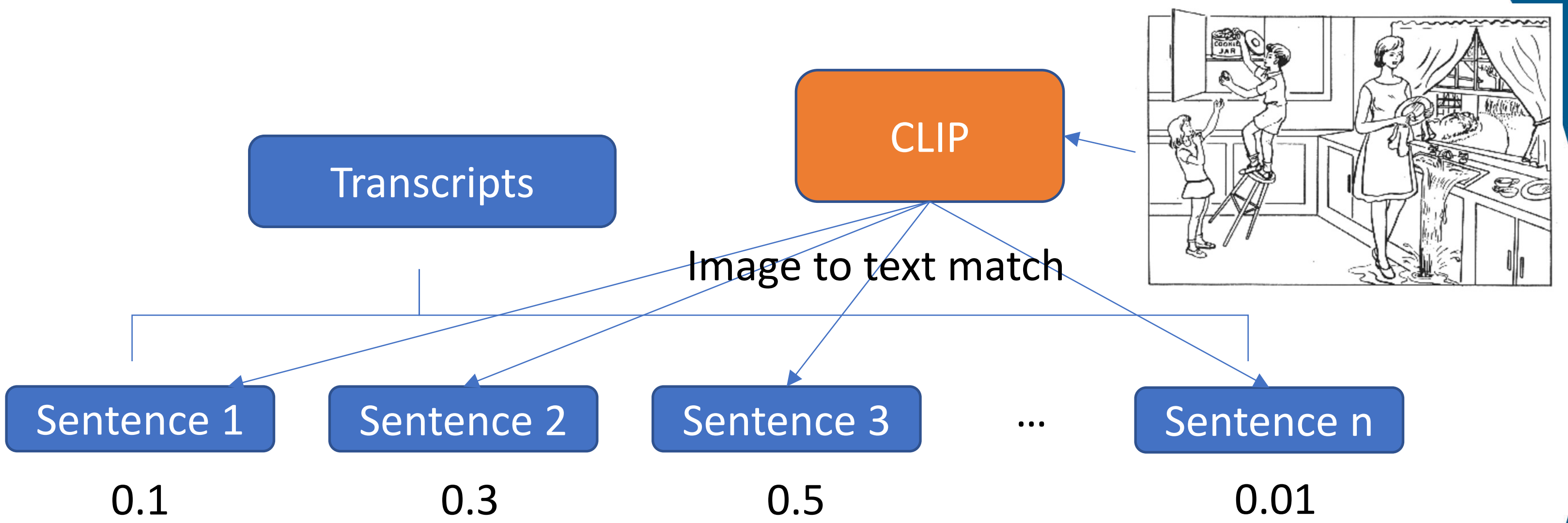

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
(2) Create dataset classifier from label text
plane
car
dog
⋮
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

Image-to-text match

- Relevance scores of one image and multiple texts
- Given a text transcript, split the text into sentences
- Use an image to generate matching probability for each sentence



Transcripts

CLIP

Image to text match

Sentence 1

Sentence 2

Sentence 3

…

Sentence n

0.1

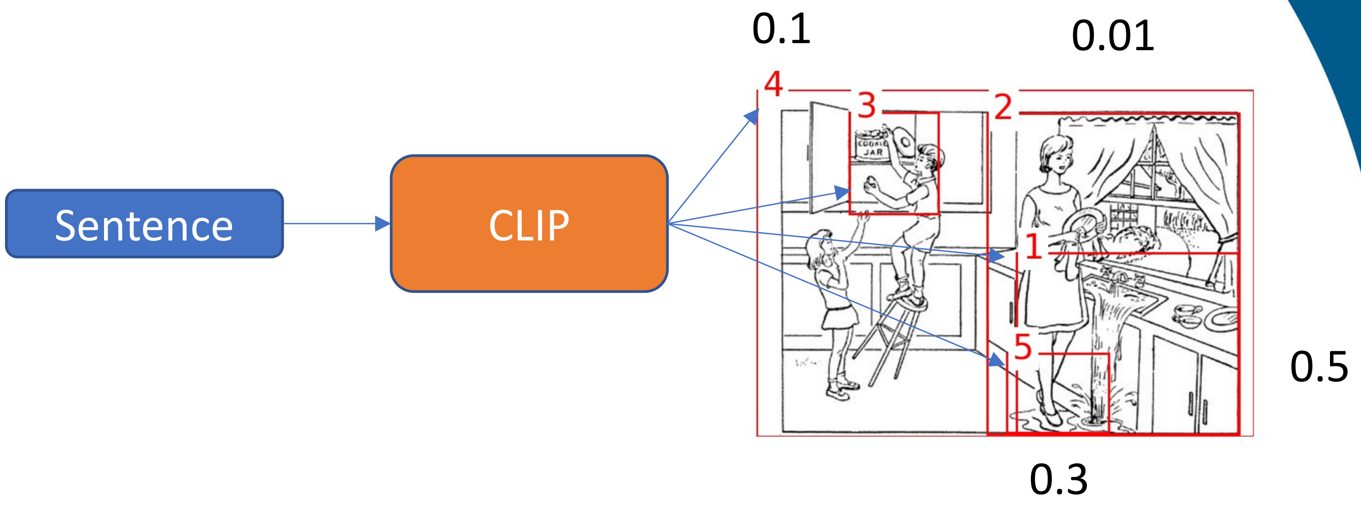
0.3

0.5

0.01

Text-to-image match

- Relevance scores of one text and multiple images
- Use selective search generate up to 400 bounding boxes (sub-images)
- Given a sentence, use CLIP to generate matching probability for each sub-image.



Sentence

CLIP

Preliminary result

- By image-to-texts match
  - HC produce lower number of word/sentence than AD
  - HC produce higher relevance text than AD
- By text-to-images match
  - Common focused areas: cookie jar and water on the floor
  - HC focuses on more areas than AD, i.e., the faucet area and the area outside of the window

	Relevance	sentence num/sample	word num/sample
HC	$c_{HC} = 19.66$	16.52	144.28
AD	$c_{AD} = 14.57$	17.70	158.35

Table 1: Preliminary results. Relevance scores are scaled by the total number of sentences in all samples.

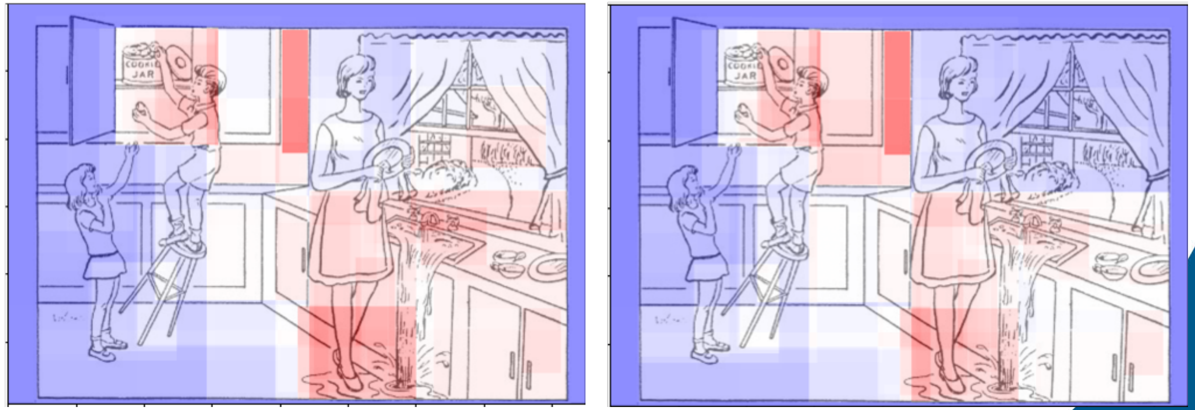


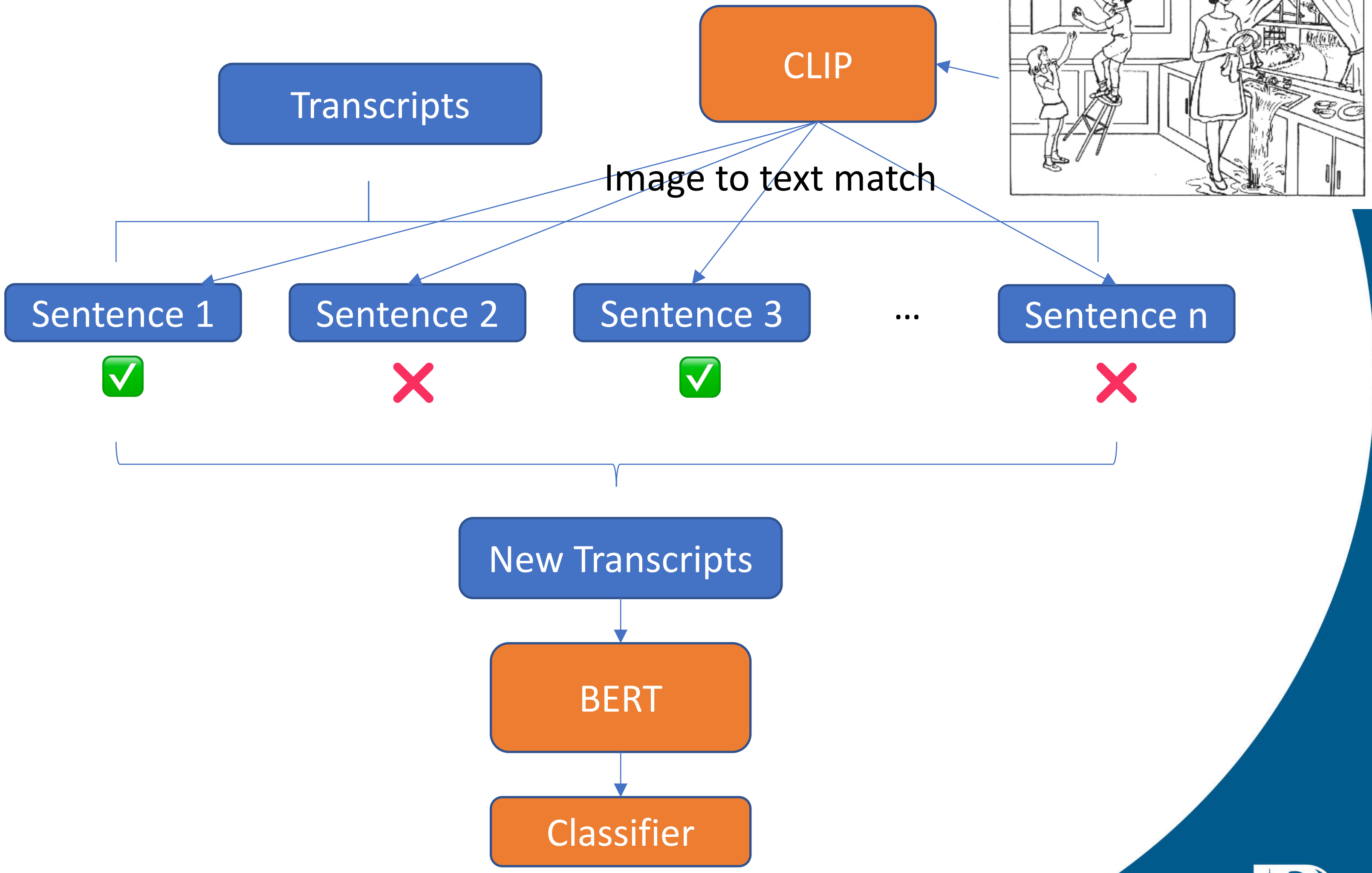
Figure 2: The focused area of HC (left) and AD (right). Red means highly focused and blue means lowly focused.

Image-text alignment - Methods

- Based on image-to-text match
    - Picture relevance model
    - Sub-image relevance model
- Based on text-to-image match
    - Focused area model

Picture relevance model

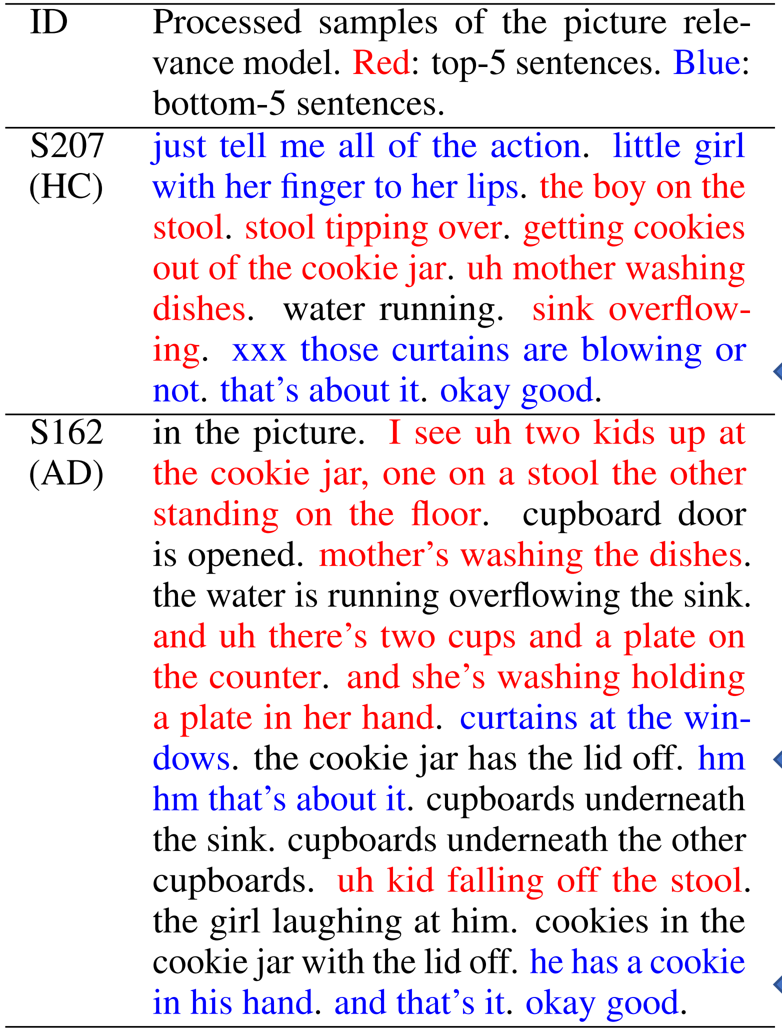
- Based on Image-to-text match
- The whole cookie theft picture as input
- Select the top-k and bottom-k sentences related to the picture
- We consider such selection emphasize dementia-related information from the text



Picture relevance model - Results

- Accuracy: 79.91% -> 80.63%
- (top-k , bottom-k) = (6, 9)

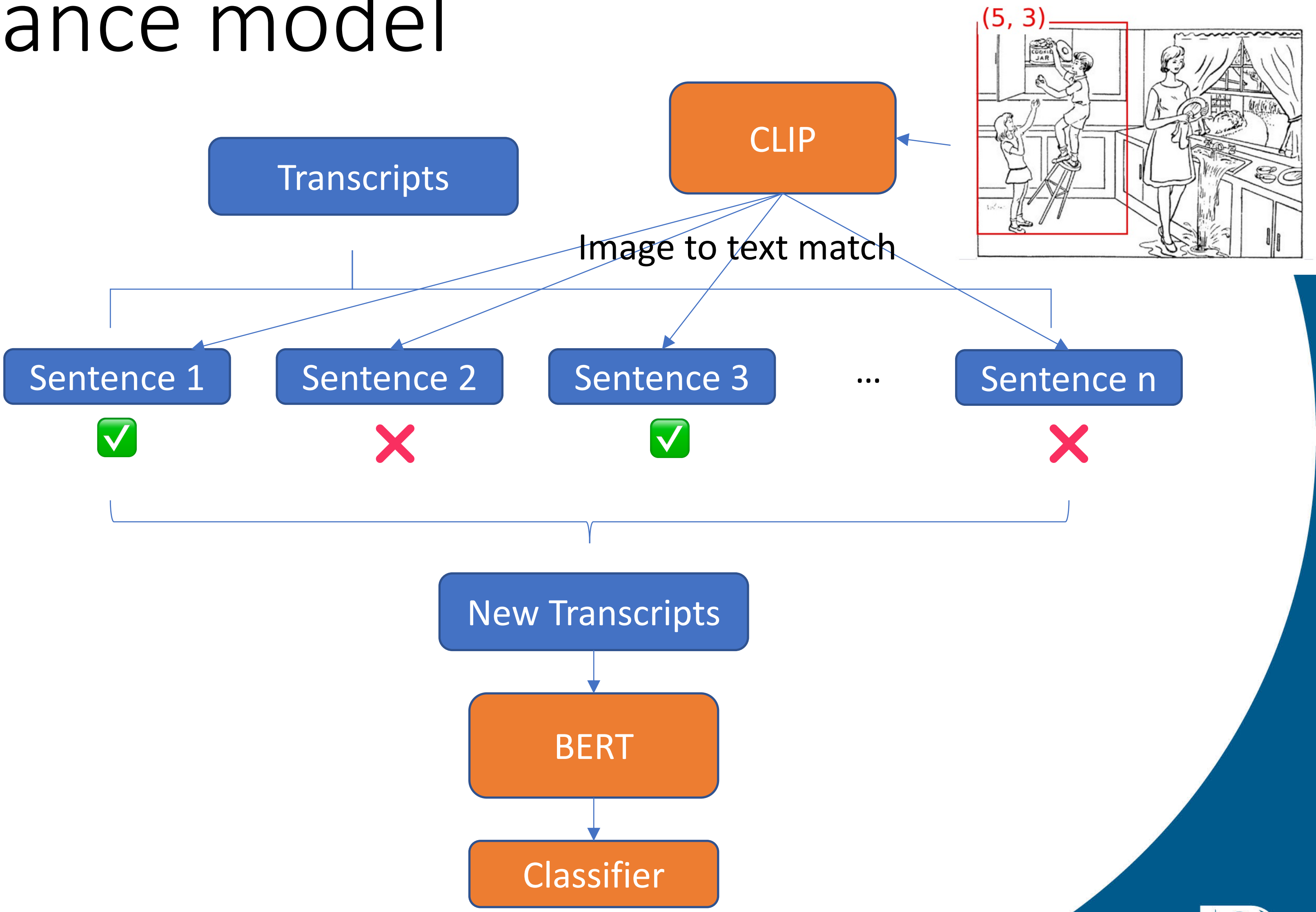
- There are irrelevant but necessary dialog acts such as acknowledgment, instruction, question and answering, stalling.
- Irrelevant sentences helps: AD participants speaks more irrelevant sentences



ID	Processed samples of the picture relevance model. Red: top-5 sentences. Blue: bottom-5 sentences.
S207 (HC)	just tell me all of the action. little girl with her finger to her lips. the boy on the stool. stool tipping over. getting cookies out of the cookie jar. uh mother washing dishes. water running. sink overflowing. xxx those curtains are blowing or not. that's about it. okay good.
S162 (AD)	in the picture. I see uh two kids up at the cookie jar, one on a stool the other standing on the floor. cupboard door is opened. mother's washing the dishes. the water is running overflowing the sink. and uh there's two cups and a plate on the counter. and she's washing holding a plate in her hand. curtains at the windows. the cookie jar has the lid off. hm hm that's about it. cupboards underneath the sink. cupboards underneath the other cupboards. uh kid falling off the stool. the girl laughing at him. cookies in the cookie jar with the lid off. he has a cookie in his hand. and that's it. okay good.

Sub-image relevance model

- The same as the picture relevance model except using a sub image as input
- Some of the contents in the picture may be more dementia-sensitive than the others
- Find out the most dementia-sensitive sub-image by maximize the embedding difference of AD and HC in the training set



(5, 3)

CLIP

Transcripts

Image to text match

Sentence 1 ✓

Sentence 2 ✗

Sentence 3 ✓

…

Sentence n ✗

New Transcripts

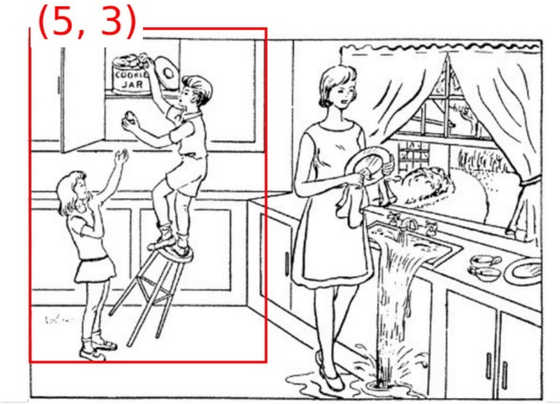
BERT

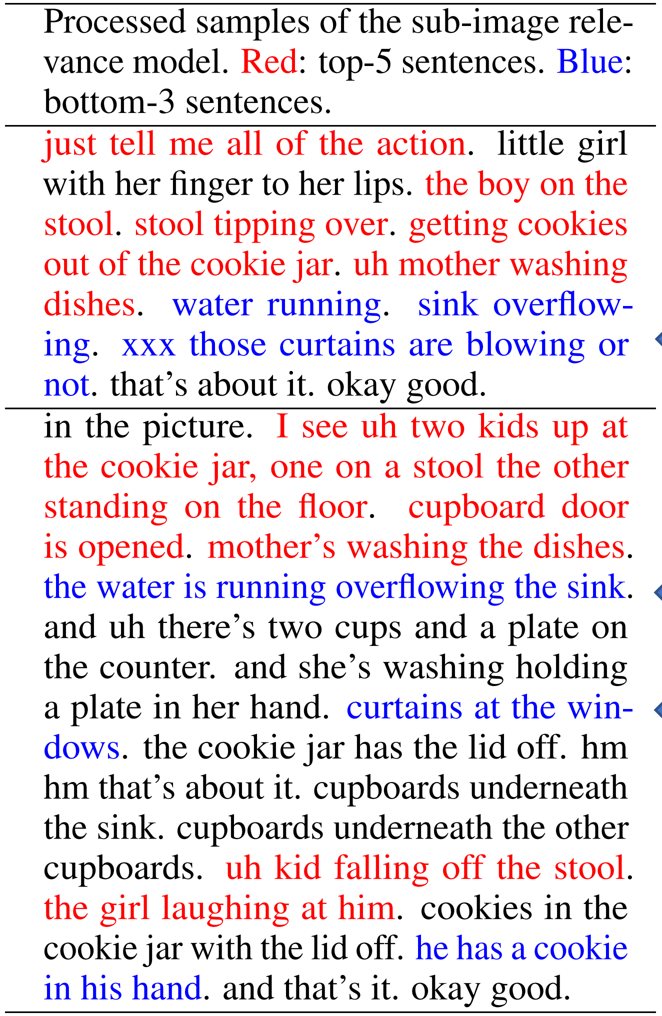
Classifier

Sub-image relevance model - Results

- Accuracy: 79.91% -> 83.44%
- (top-k , bottom-k) = (5, 3)

- The sentence describing the right part of image now consider as irrelevant.

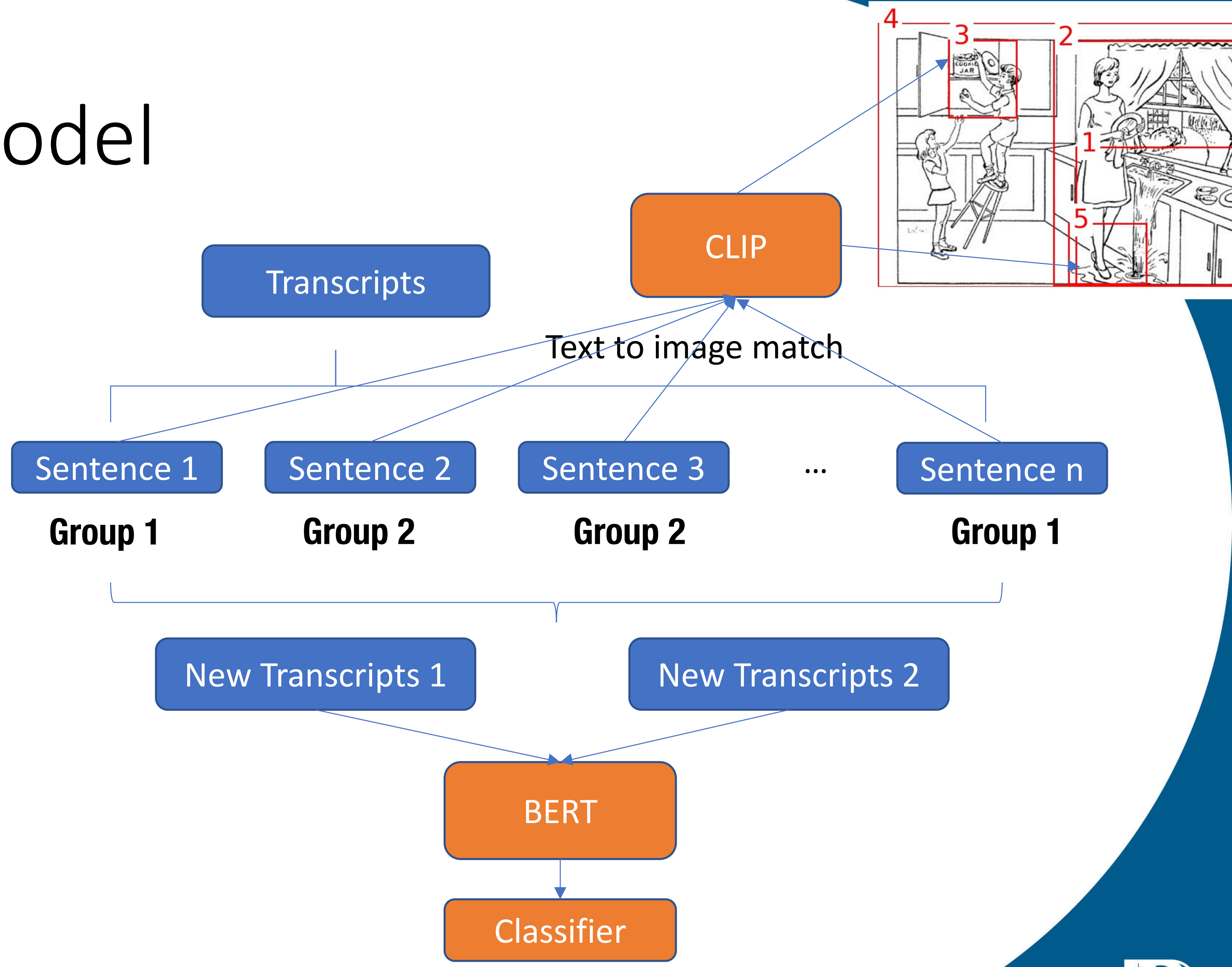




Processed samples of the sub-image relevance model. Red: top-5 sentences. Blue: bottom-3 sentences.
just tell me all of the action. little girl with her finger to her lips. the boy on the stool. stool tipping over. getting cookies out of the cookie jar. uh mother washing dishes. water running. sink overflowing. xxx those curtains are blowing or not. that's about it. okay good.
in the picture. I see uh two kids up at the cookie jar, one on a stool the other standing on the floor. cupboard door is opened. mother's washing the dishes. the water is running overflowing the sink. and uh there's two cups and a plate on the counter. and she's washing holding a plate in her hand. curtains at the windows. the cookie jar has the lid off. hm hm that's about it. cupboards underneath the sink. cupboards underneath the other cupboards. uh kid falling off the stool. the girl laughing at him. cookies in the cookie jar with the lid off. he has a cookie in his hand. and that's it. okay good.

Focused area model

- Based on text-to-image match
- Select top-n focused areas from the picture
- Group the sentences into corresponding focus areas
- Direct comparison within the same group



Transcripts

CLIP

Text to image match

Sentence 1

Sentence 2

Sentence 3

…

Sentence n

Group 1

Group 2

Group 2

Group 1

New Transcripts 1

New Transcripts 2

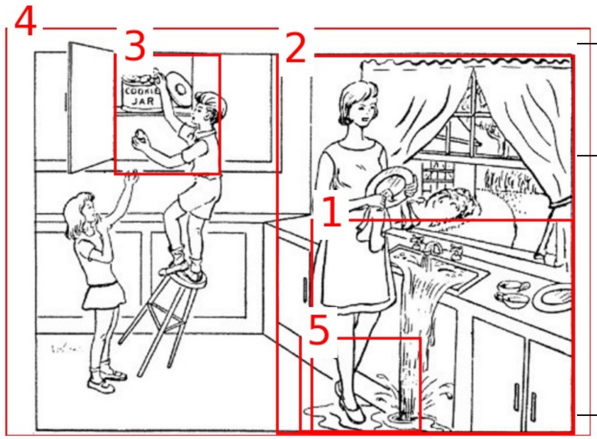
BERT

Classifier

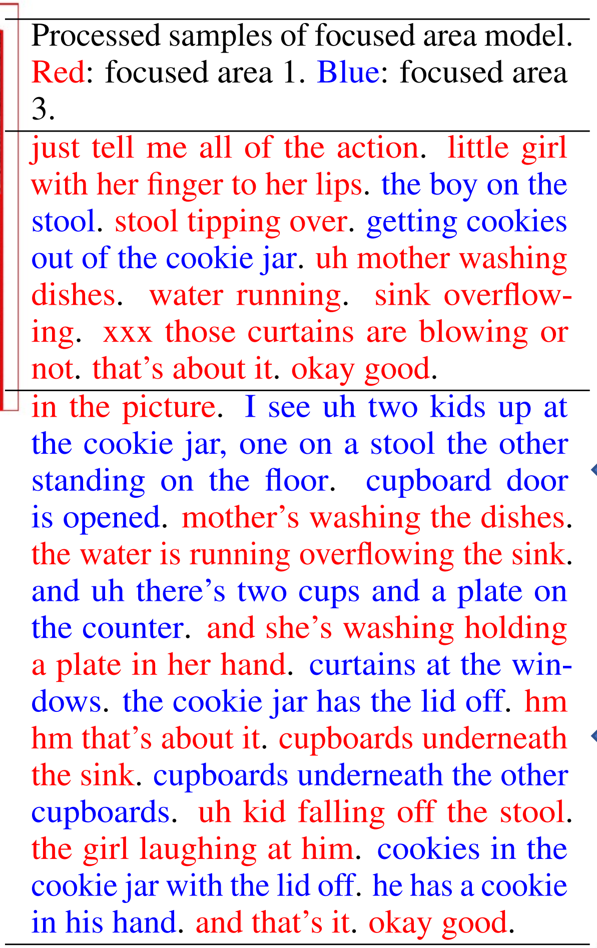
UMass Boston

Focused area model - Results

- Accuracy: 79.91% -> 82.49%
- Focus area: (1, 3)



- Most of sentences related to the areas are successful grouped
- AD participants may produce sentence hard to group (e.g., cupboards)



Processed samples of focused area model. Red: focused area 1. Blue: focused area 3.

just tell me all of the action. little girl with her finger to her lips. the boy on the stool. stool tipping over. getting cookies out of the cookie jar. uh mother washing dishes. water running. sink overflowing. xxx those curtains are blowing or not. that's about it. okay good.

in the picture. I see uh two kids up at the cookie jar, one on a stool the other standing on the floor. cupboard door is opened. mother's washing the dishes. the water is running overflowing the sink. and uh there's two cups and a plate on the counter. and she's washing holding a plate in her hand. curtains at the windows. the cookie jar has the lid off. hm hm that's about it. cupboards underneath the sink. cupboards underneath the other cupboards. uh kid falling off the stool. the girl laughing at him. cookies in the cookie jar with the lid off. he has a cookie in his hand. and that's it. okay good.

Conclusion & Future work

- We study the image text alignment for dementia detection and find out:
  - HC participants produce smaller number of word/sentence but with high picture relevance than AD
  - Common focused area exists, and HC have more focused areas than AD
- Based on the above findings, we propose models to process the text transcripts, and demonstrate the performance improvements than the baseline.
- The future work includes end-to-end training using the picture as input.

Thank you

UMass Boston