Master Thesis Colloquium

Voice input based story generation

HOCHSCHULE
SRH HEIDELBERG
Intelligence in Learning

30.11.2020 by:

Rohit Keshav Bewoor (11011831)

Big Data and Business Analytics 2018-20 batch
SRH Hochschule Heidelberg

Agenda

- Problem Statement and Objectives
- Implementation
  - High Level Approach
  - Individual Stages
- Survey Design
- Results
- Contributions
- Conclusions
- Future Scope
- Virtual Demo

Problem Statement and Objectives

- **Problem Statement**:
  - Generate shorts stories with accompanying images by accepting voice input describing the story required.
  - Target audience for stories: young children (aged 5-8 years old)

- **Objectives**:
  - Independent evaluation of results
  - Avoid use of paid services if possible
  - Accept exactly 3 sentences as user input
  - Output story to have 1 to 3 images along with text
  - Graphical User Interface for ease of use

Implementation - High Level Approach

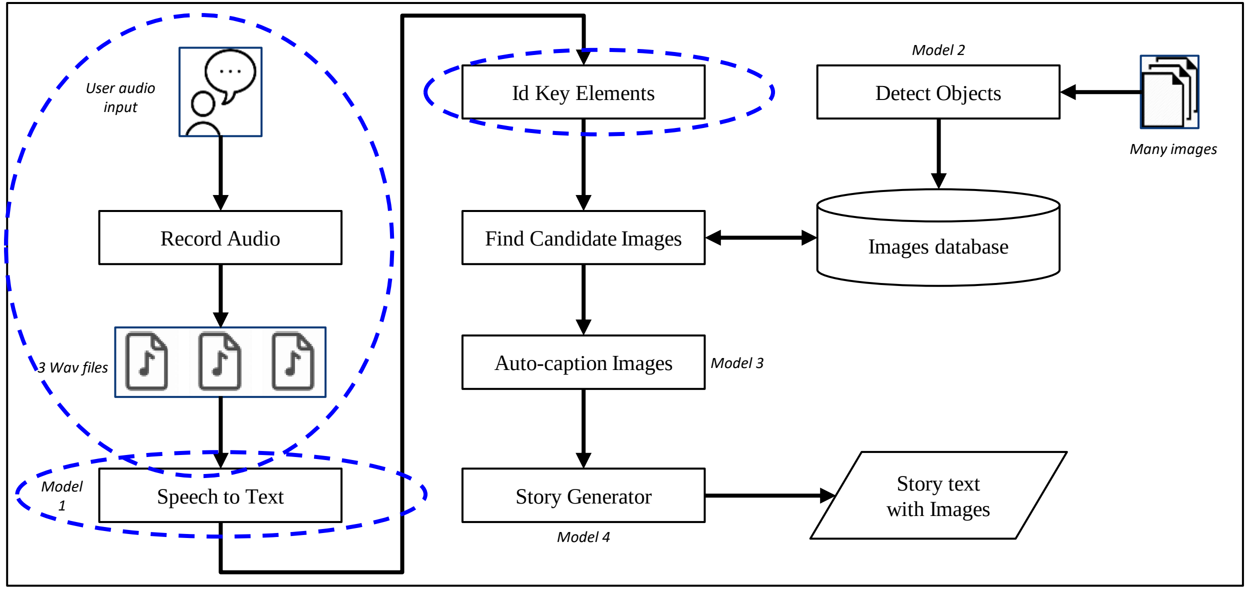
STAATLICH
ANERKANNTE
HOCHSCHULE

- Local service as Python-3 programs executing on laptop
- Models trained on free cloud computing platforms – Kaggle and Google Colab

- 3 wav files (recording optional)

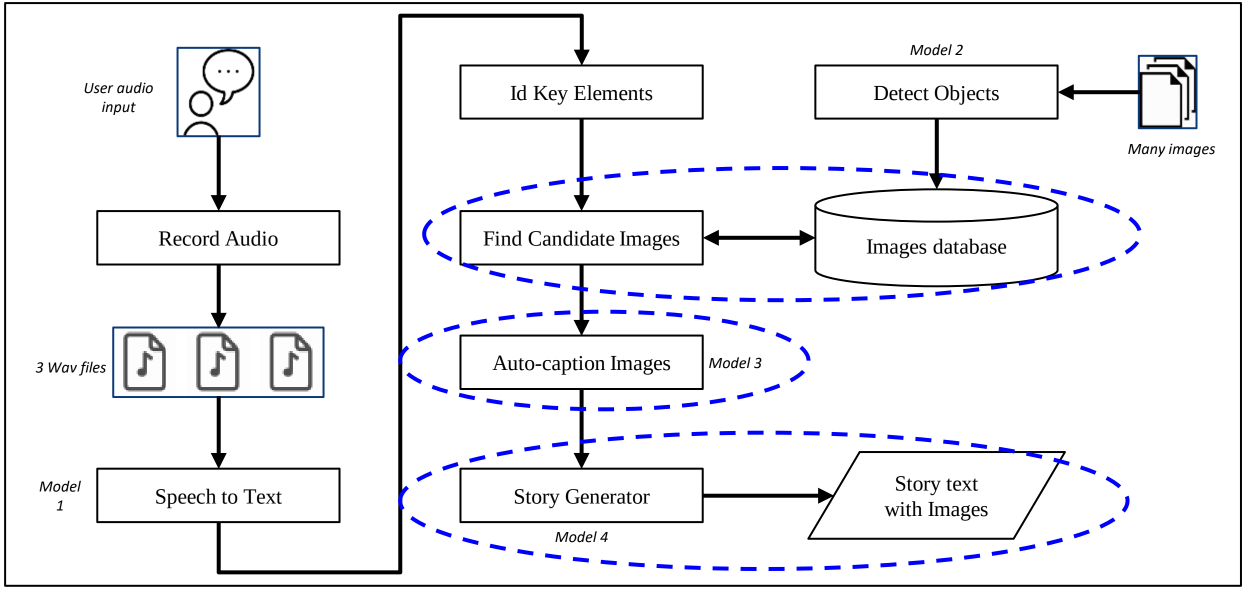
- Model 1: Perform Speech-to-Text
  - ➢ Output 3 sentences

- Identify Key Elements
  - ➢ Output Noun-type words
  - ➢ Only objects detectable



Implementation - High Level Approach

- Retrieve candidate images
  - Query database
  - Model 2: Object detection (optional)
  - GUI: selection (max 5/sentence)
- Model 3: Perform Image Captioning
  - GUI: selection and optional correction
  - Output captions
- Model 4: Story Generator
  - Output Story Text using captions



❖ ***Processed input voice in stages and output story text with images***

Neural network models used

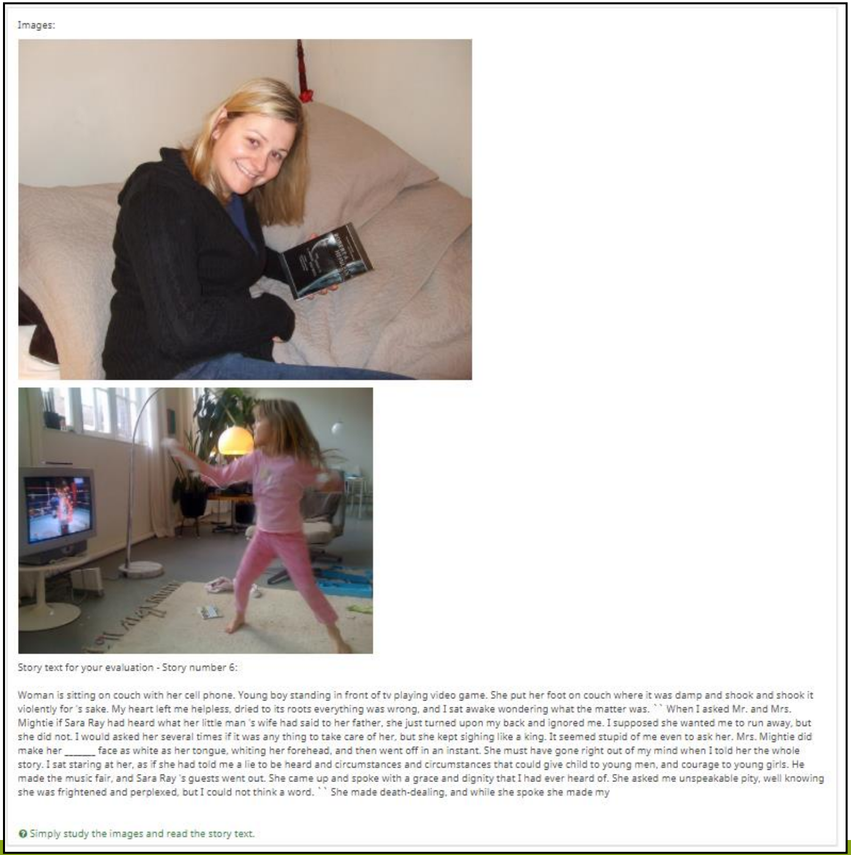
- **Speech-to-Text**
  - DeepSpeech version 0.7.3
  - Pre-trained model

- **Object Detector:**
  - YOLOv3 trained on Common Objects in Context (COCO) 2017 dataset
  - 80 labels can be detected

- **Image Captioning:**
  - Without-attention: “Show and Tell: A Neural Image Caption Generator”
  - With-attention: “Show, Attend Tell: Neural Image Caption Generator with Visual Attention”
  - Both models trained on approx. 100k images of COCO dataset

- **Story Generator:**
  - GPT-2 “medium” sized model – 355 million parameters
  - Fine-tuned on 11 files from Children's Book Test (CBT) dataset

Survey Design - Stories

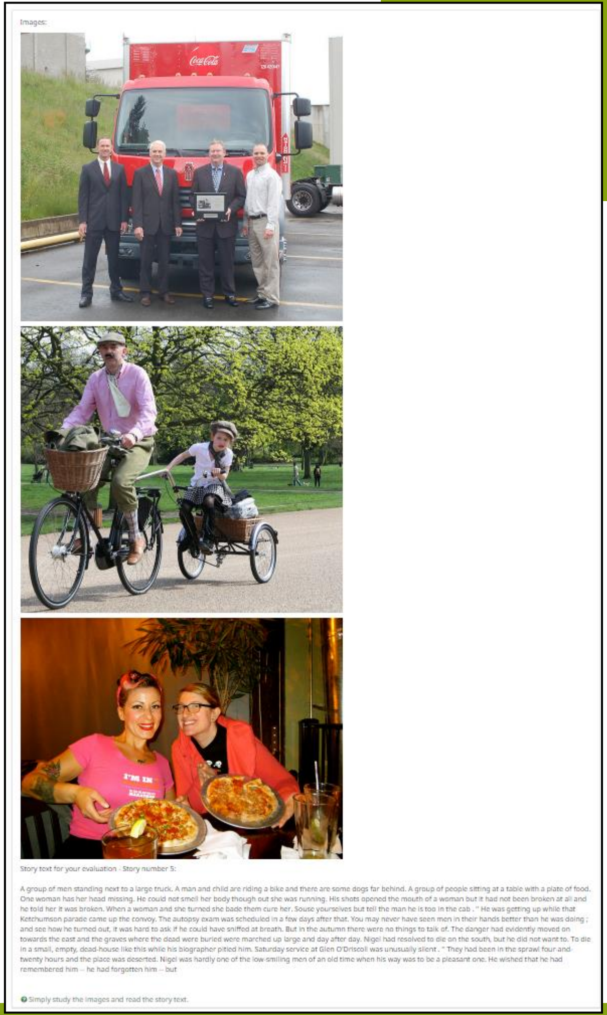


Images:

Story text for your evaluation - Story number 6:

Woman is sitting on couch with her cell phone. Young boy standing in front of tv playing video game. She put her foot on couch where it was damp and shook and shook it violently for 's sake. My heart left me helpless, dried to its roots everything was wrong, and I sat awake wondering what the matter was. `` When I asked Mr. and Mrs. Mightie if Sara Ray had heard what her little man 's wife had said to her father, she just turned upon my back and ignored me. I supposed she wanted me to run away, but she did not. I would asked her several times if it was any thing to take care of her, but she kept sighing like a king. It seemed stupid of me even to ask her. Mrs. Mightie did make her ______ face as white as her tongue, whiting her forehead, and then went off in an instant. She must have gone right out of my mind when I told her the whole story. I sat staring at her, as if she had told me a lie to be heard and circumstances and circumstances that could give child to young men, and courage to young girls. He made the music fair, and Sara Ray 's guests went out. She came up and spoke with a grace and dignity that I had ever heard of. She asked me unspeakable pity, well knowing she was frightened and perplexed, but I could not think a word. `` She made death-dealing, and while she spoke she made my

? Simply study the images and read the story text.



Images:

Story text for your evaluation - Story number 5:

A group of men standing next to a large truck. A man and child are riding a bike and there are some dogs far behind. A group of people sitting at a table with a plate of food. One woman has her head missing. He could not smell her body though out she was running. His shots opened the mouth of a woman but it had not been broken at all and he told her it was broken. When a woman and she turned she bade them cure her. Souse yourselves but tell the man he is too in the cab . " He was getting up while that Ketchumson parade came up the convoy. The autopsy exam was scheduled in a few days after that. You may never have seen men in their hands better than he was doing ; and see how he turned out, it was hard to ask if he could have sniffed at breath. But in the autumn there were no things to talk of. The danger had evidently moved on towards the east and the graves where the dead were buried were marched up large and day after day. Nigel had resolved to die on the south, but he did not want to. To die in a small, empty, dead-house like this while his biographer pitied him. Saturday service at Glen O'Driscoll was unusually silent . " They had been in the sprawl four-and-twenty hours and the place was deserted. Nigel was hardly one of the low-smiling men of an old time when his way was to be a pleasant one. He wished that he had remembered him -- he had forgotten him -- but

? Simply study the images and read the story text.

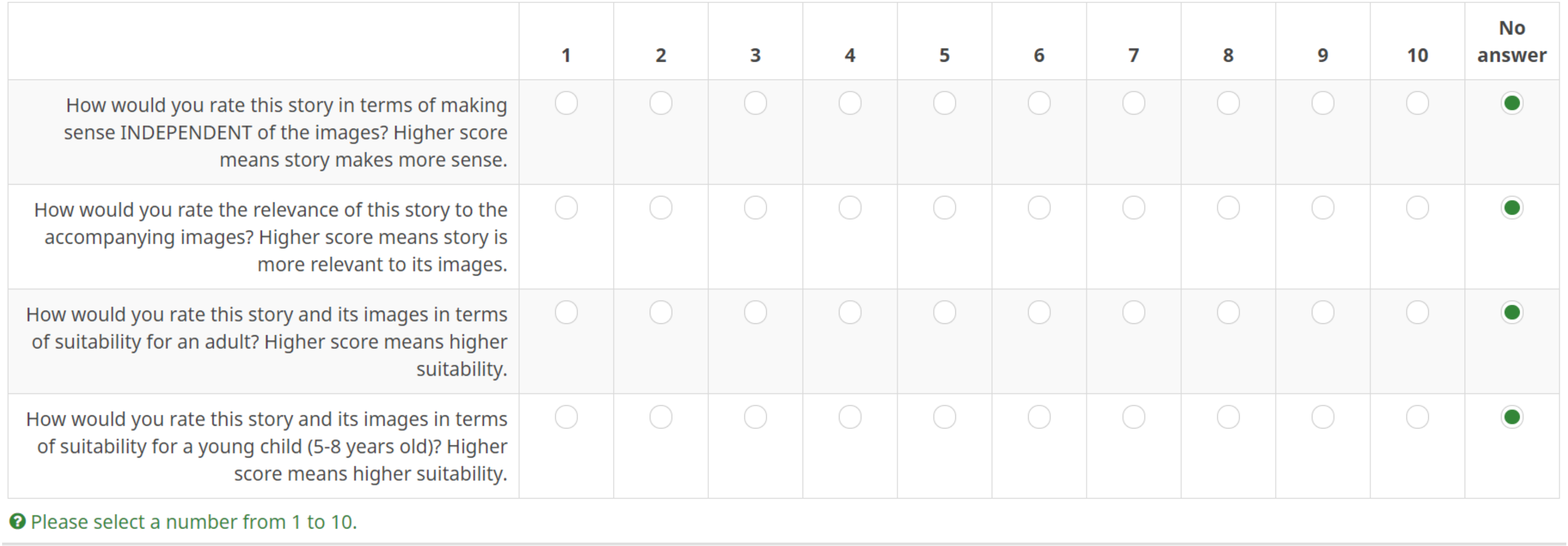
Survey Questions

Q1: Coherent independent of images?

Q2: Coherent with images?

Q3: Suitability for adults?

Q4: Suitability for children?



	1	2	3	4	5	6	7	8	9	10	No answer
How would you rate this story in terms of making sense INDEPENDENT of the images? Higher score means story makes more sense.	○	○	○	○	○	○	○	○	○	○	●
How would you rate the relevance of this story to the accompanying images? Higher score means story is more relevant to its images.	○	○	○	○	○	○	○	○	○	○	●
How would you rate this story and its images in terms of suitability for an adult? Higher score means higher suitability.	○	○	○	○	○	○	○	○	○	○	●
How would you rate this story and its images in terms of suitability for a young child (5-8 years old)? Higher score means higher suitability.	○	○	○	○	○	○	○	○	○	○	●

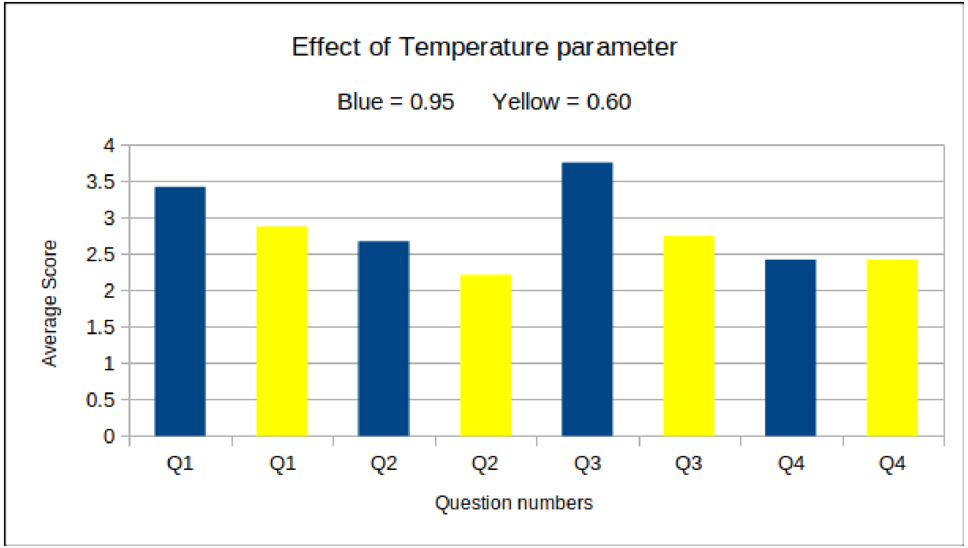
❓ Please select a number from 1 to 10.

Survey Results

STAATLICH
ANERKANNTE
HOCHSCHULE

Effect of Temperature: 0.95 vs 0.60

- Temperature = 0.95 consistently scored higher
- Only for “suitability for children”: slightly lower score

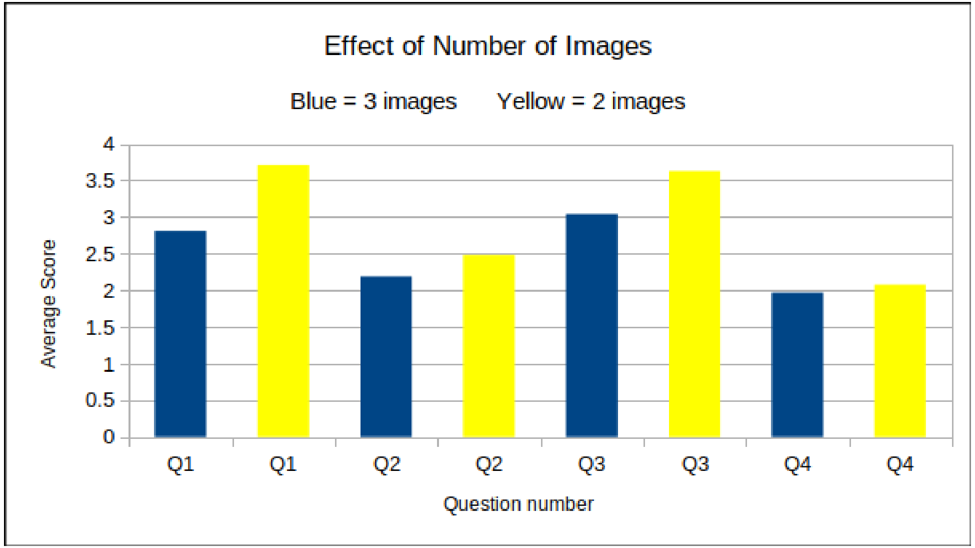


Q1: Coherent independent of images?
Q2: Coherent with images?
Q3: Suitability for adults?
Q4: Suitability for children?

Survey Results

Effect of Number of Images: 2 vs 3

- Stories with 2 images consistently scored higher



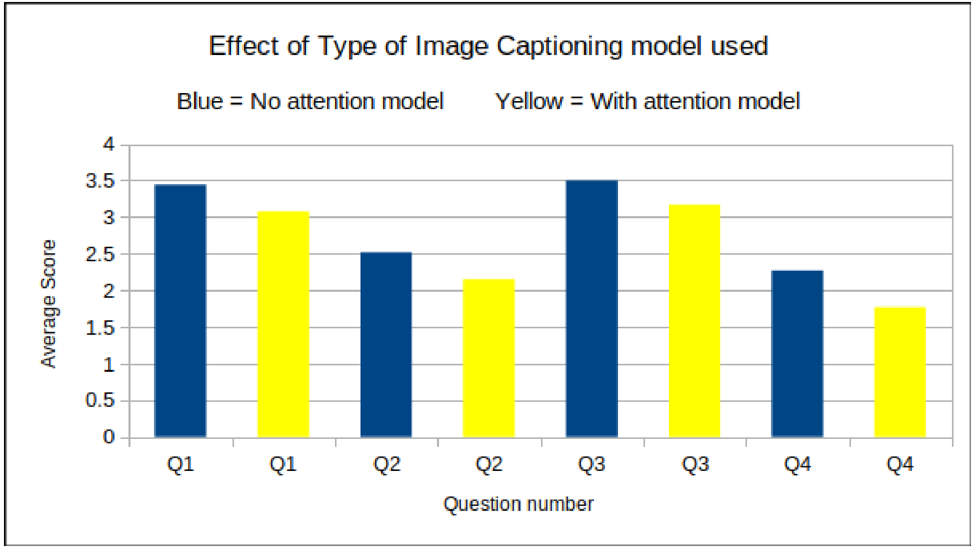
Q1: Coherent independent of images?
Q2: Coherent with images?
Q3: Suitability for adults?
Q4: Suitability for children?

Survey Results

STAATLICH ANERKANNTE HOCHSCHULE

Effect of Image Caption model type

- Without-attention model consistently scored higher



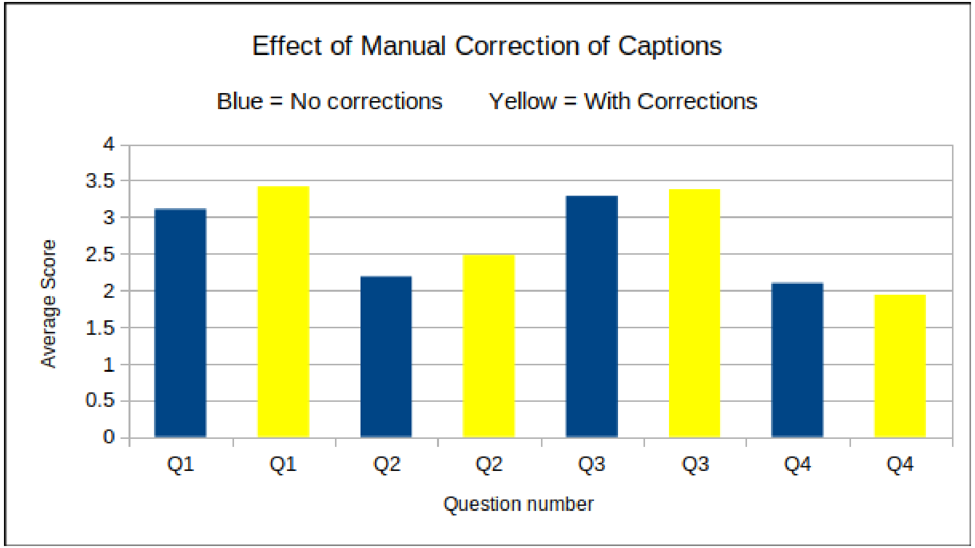
Q1: Coherent independent of images?
Q2: Coherent with images?
Q3: Suitability for adults?
Q4: Suitability for children?

Survey Results

STAATLICH
ANERKANNTE
HOCHSCHULE

Effect of manual correction of Captions

- Manual correction improved scores generally
- Only for “suitability for children”: slightly lower score



Q1: Coherent independent of images?
Q2: Coherent with images?
Q3: Suitability for adults?
Q4: Suitability for children?

Conclusions

- Based on survey results:
  - ➢ Stories were not suitable for young children (average score = 2.02)
  - ➢ Ratings for adults slightly higher (average score = 3.33), but still overall low scores
  - ➢ Stories with 2 images scored higher 2 > 3
  - ➢ Manually corrected captions had limited impact on scores
  - ➢ Without-attention model scored higher

Conclusions

- “Medium” size GPT-2 model unable to meet requirements
  - ➢ But, higher Temperature => better stories
- Image captioning models had reasonable BLEU scores (0.6 and above) – but insufficient for use-case
- Single sentence seed values produced more coherent story

Contributions

- Overall concept itself

- Pipeline architecture and choice of neural network models

- Graphical user interface incorporation
  - Decision to implement
  - User Interface design

- Designing and conducting of survey for objective results

Future Scope

- More training for with-attention model (data and epochs)

- Improve database design and information captured
  - Verbs (action being performed)
  - Prepositions (relative arrangement of the objects)

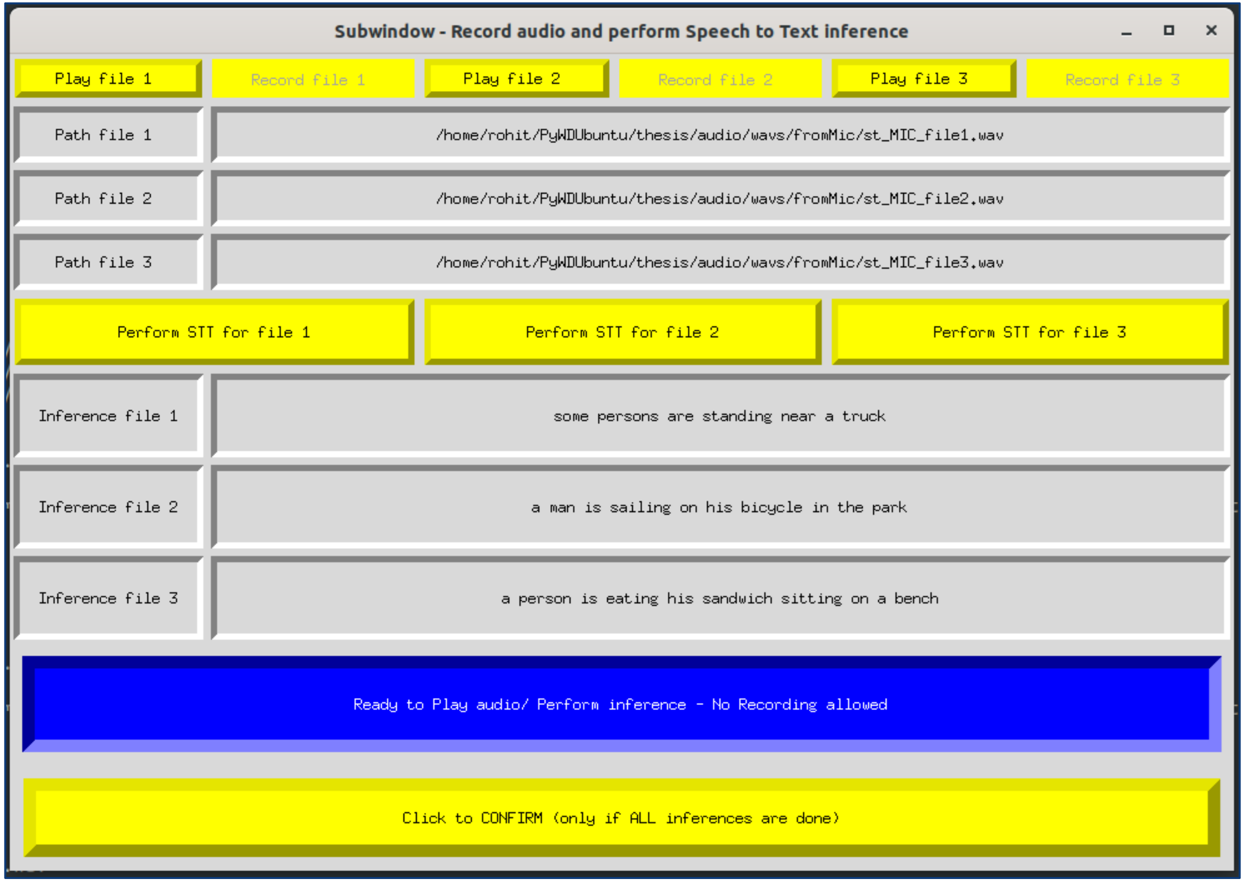
- Adapt code to show more than 20 images to allow more diversity during user selection

- More exhaustive survey
  - Number of respondents
  - Number of questions

Virtual Demo

- Screenshots of the user interaction screens
- Speech-to-Text stage to Image Captioning Results selection

Process Wav files and Perform STT



/home/rohit/PyWDUbuntu/thesis/audio/wavs/fromMic/st_MIC_file1.wav

/home/rohit/PyWDUbuntu/thesis/audio/wavs/fromMic/st_MIC_file2.wav

/home/rohit/PyWDUbuntu/thesis/audio/wavs/fromMic/st_MIC_file3.wav

some persons are standing near a truck

a man is sailing on his bicycle in the park

a person is eating his sandwich sitting on a bench

Ready to Play audio/ Perform inference - No Recording allowed

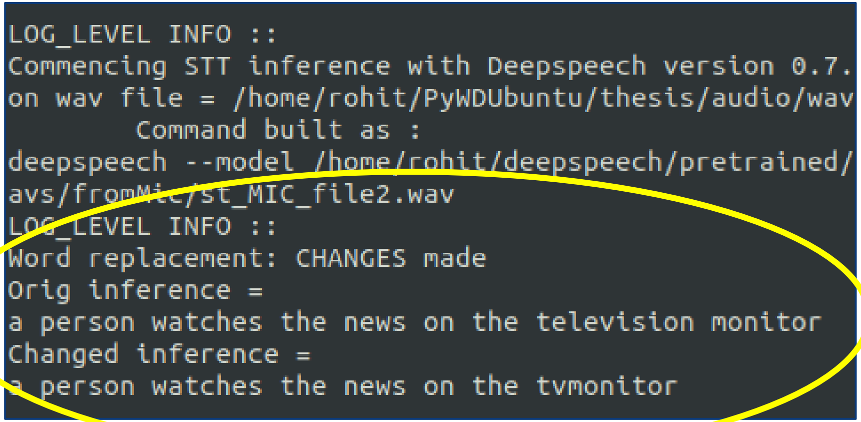
- User ran inference for all Wav files
- Happy with output
- Clicks Confirm button (bottom of window)

Replace inference word (special cases only)

- Replacement of special words to allow downstream processing to succeed
  - 80 predefined lables of COCO dataset

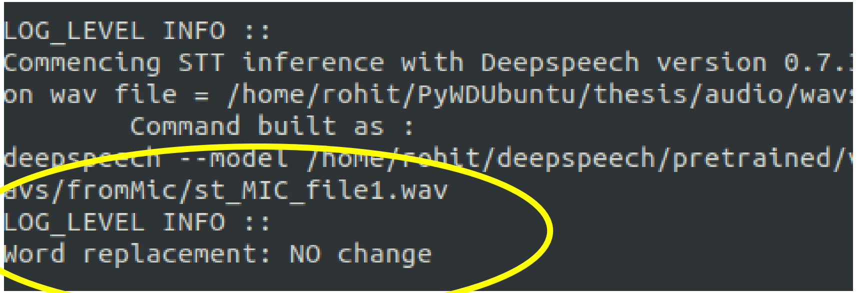
- Inference output of “*the **hand bag** has many items in it*” **will not match** “hand bag” with the **label “handbag”** and logic fails!

Changes made to inference output in this case



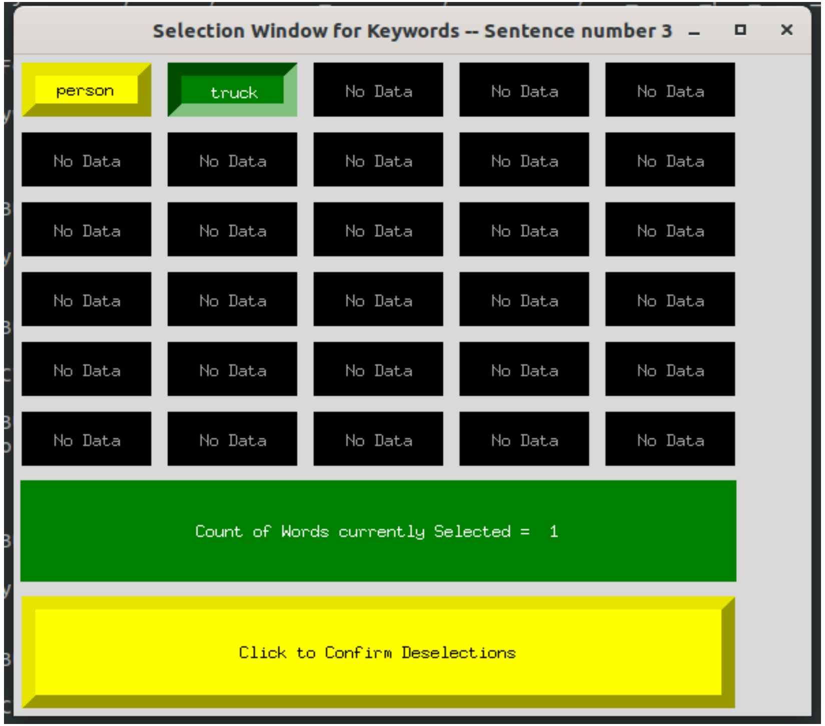
```
LOG_LEVEL INFO ::
Commencing STT inference with Deepspeech version 0.7.
on wav file = /home/rohit/PyWDUbuntu/thesis/audio/wav
        Command built as :
deepspeech --model /home/rohit/deepspeech/pretrained/
avs/fromMic/st_MIC_file2.wav
LOG_LEVEL INFO ::
Word replacement: CHANGES made
Orig inference =
a person watches the news on the television monitor
Changed inference =
a person watches the news on the tvmonitor
```

No change to inference output in thisi case



```
LOG_LEVEL INFO ::
Commencing STT inference with Deepspeech version 0.7.
on wav file = /home/rohit/PyWDUbuntu/thesis/audio/wavs
        Command built as :
deepspeech --model /home/rohit/deepspeech/pretrained/
avs/fromMic/st_MIC_file1.wav
LOG_LEVEL INFO ::
Word replacement: NO change
```

Keywords - Select / Deselect



Selection Window for Keywords -- Sentence number 3

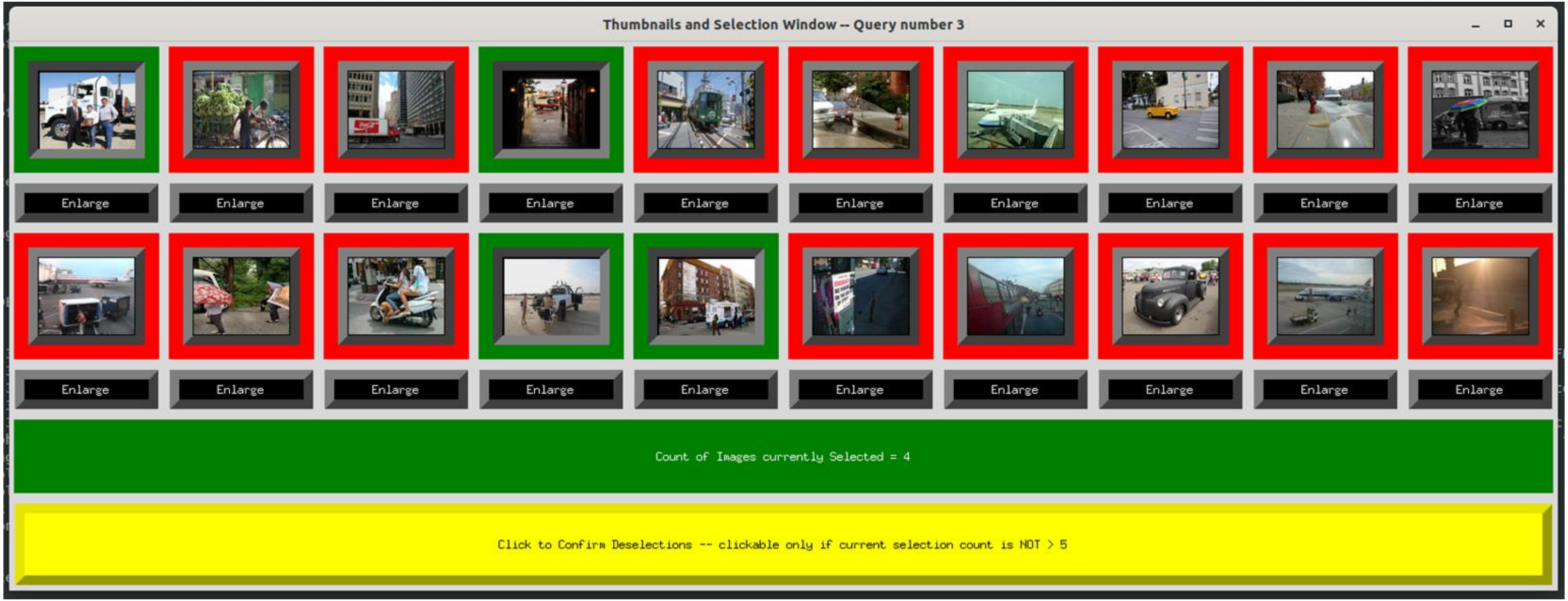
person	truck	No Data	No Data	No Data
No Data	No Data	No Data	No Data	No Data
No Data	No Data	No Data	No Data	No Data
No Data	No Data	No Data	No Data	No Data
No Data	No Data	No Data	No Data	No Data
No Data	No Data	No Data	No Data	No Data

Count of Words currently Selected = 1

Select Images from database

STAATLICH
ANERKANNTE
HOCHSCHULE

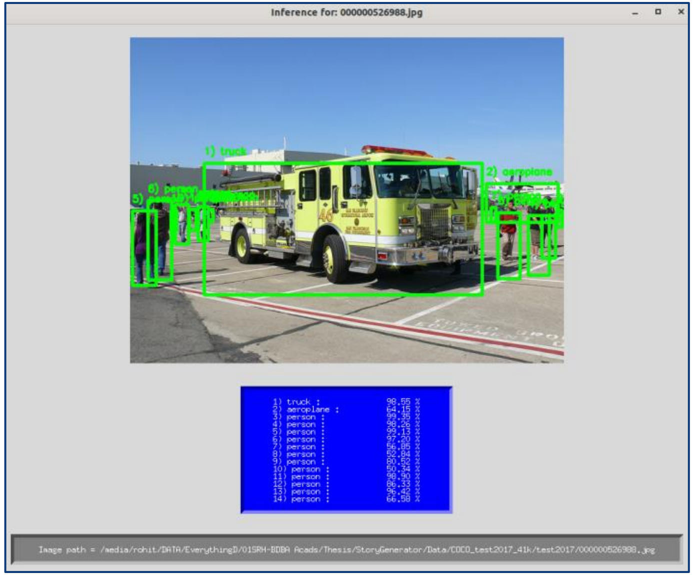
- Here user selected 4 images of the 20 originally returned by query



Count of Images currently Selected = 4

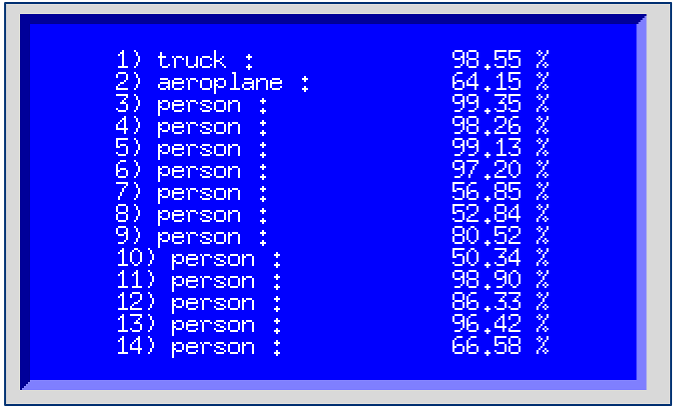
Click to Confirm Deselections -- clickable only if current selection count is NOT > 5

Optional: Object Detection



```
1) truck :                98.55 %
2) aeroplane :            64.15 %
3) person :               99.35 %
4) person :               98.26 %
5) person :               99.13 %
6) person :               97.20 %
7) person :               56.85 %
8) person :               52.84 %
9) person :               80.52 %
10) person :              50.34 %
11) person :              98.90 %
12) person :              86.33 %
13) person :              96.42 %
14) person :              66.58 %
```

```
Image path = /media/rohit/DATA/EverythingD/01SRH-BDBA Acads/Thesis/StoryGenerator/Data/COCO_test2017_41k/test2017/000000526988.jpg
```

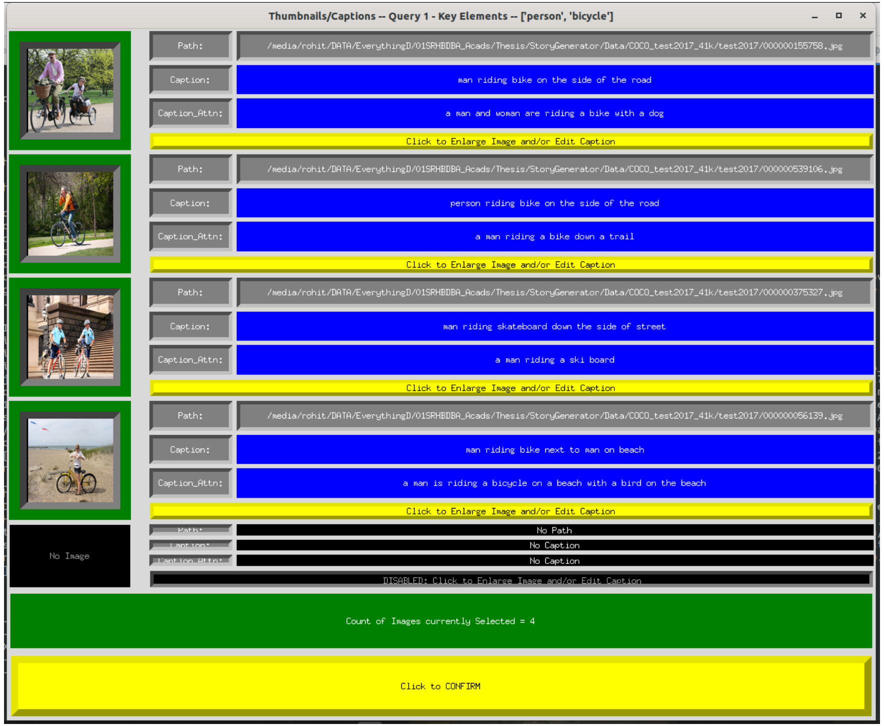

1) truck : 98.55 %
2) aeroplane : 64.15 %
3) person : 99.35 %
4) person : 98.26 %
5) person : 99.13 %
6) person : 97.20 %
7) person : 56.85 %
8) person : 52.84 %
9) person : 80.52 %
10) person : 50.34 %
11) person : 98.90 %
12) person : 86.33 %
13) person : 96.42 %
14) person : 66.58 %

Checks HAS relationship score > 0.90 for the objects specified in query.

Note: At least one object of “Truck” and “Person” have scores > 90%

Image Captioning

- Displaying images with original captions - *4 of 5 maximum possible images selected and displayed for this query*



Thumbnails/Captions -- Query 1 - Key Elements -- ['person', 'bicycle']

Path: /media/rohit/DATA/EverythingD/01SRHBDBA_Acads/Thesis/StoryGenerator/Data/COCO_test2017_41k/test2017/000000155758.jpg

Caption: man riding bike on the side of the road

Caption_Attn: a man and woman are riding a bike with a dog

Click to Enlarge Image and/or Edit Caption

Path: /media/rohit/DATA/EverythingD/01SRHBDBA_Acads/Thesis/StoryGenerator/Data/COCO_test2017_41k/test2017/000000539106.jpg

Caption: person riding bike on the side of the road

Caption_Attn: a man riding a bike down a trail

Click to Enlarge Image and/or Edit Caption

Path: /media/rohit/DATA/EverythingD/01SRHBDBA_Acads/Thesis/StoryGenerator/Data/COCO_test2017_41k/test2017/000000375327.jpg

Caption: man riding skateboard down the side of street

Caption_Attn: a man riding a ski board

Click to Enlarge Image and/or Edit Caption

Path: /media/rohit/DATA/EverythingD/01SRHBDBA_Acads/Thesis/StoryGenerator/Data/COCO_test2017_41k/test2017/000000056139.jpg

Caption: man riding bike next to man on beach

Caption_Attn: a man is riding a bicycle on a beach with a bird on the beach

Click to Enlarge Image and/or Edit Caption

No Image

Path: No Path

Caption: No Caption

Caption_Attn: No Caption

DISABLED: Click to Enlarge Image and/or Edit Caption

Count of Images currently Selected = 4

Click to CONFIRM

Image thumbnails on far left: Clickable to deselect

Image absolute path

Original Captions for Image (blue boxes)

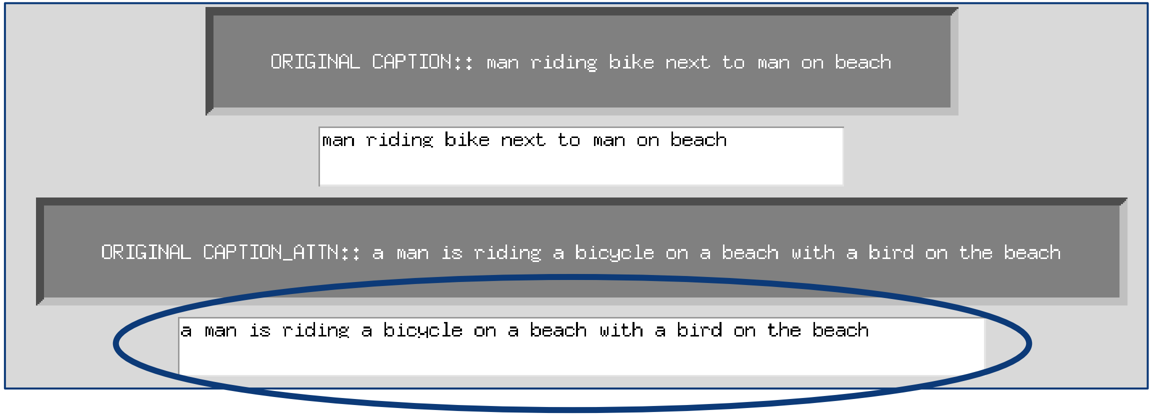
Clickable button to Enlarge Image + Edit Caption (yellow box)

Fifth image placeholder black as no Image selected

Count of currently selected Images (defaults to ALL)

Final Confirm button

Manual correction of captions (optional)



ORIGINAL CAPTION:: man riding bike next to man on beach

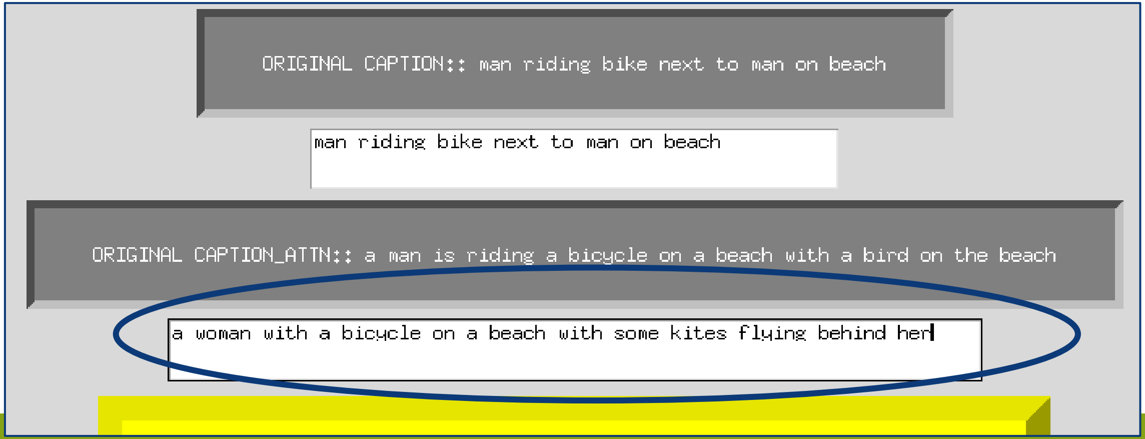
man riding bike next to man on beach

ORIGINAL CAPTION_ATTN:: a man is riding a bicycle on a beach with a bird on the beach

a man is riding a bicycle on a beach with a bird on the beach

User edited only the caption in the second white box (With-Attention caption)

Caption changes



ORIGINAL CAPTION:: man riding bike next to man on beach

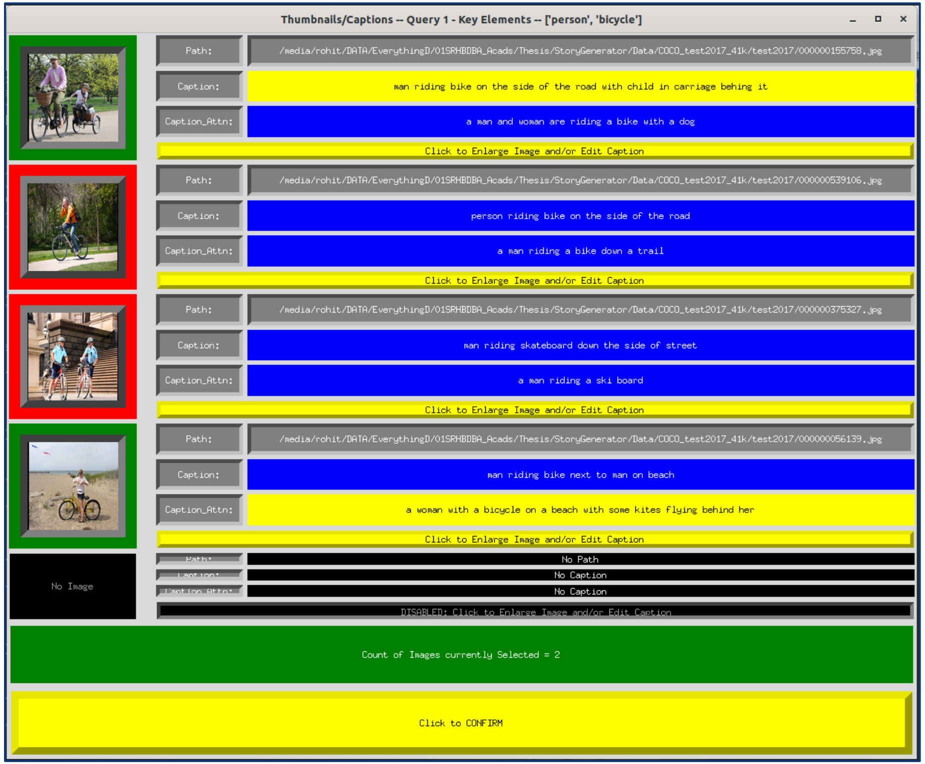
man riding bike next to man on beach

ORIGINAL CAPTION_ATTN:: a man is riding a bicycle on a beach with a bird on the beach

a woman with a bicycle on a beach with some kites flying behind her

Then clicks yellow button to CONFIRM

Ready for final Confirmation for this Query



Thumbnails/Captions -- Query 1 - Key Elements -- ['person', 'bicycle']

Path: /media/rohit/DATA/EverythingD/01SRHBDBA_Acads/Thesis/StoryGenerator/Data/COCO_test2017_41k/test2017/000000155758.jpg

Caption: man riding bike on the side of the road with child in carriage behing it

Caption_Attn: a man and woman are riding a bike with a dog

Path: /media/rohit/DATA/EverythingD/01SRHBDBA_Acads/Thesis/StoryGenerator/Data/COCO_test2017_41k/test2017/000000539106.jpg

Caption: person riding bike on the side of the road

Caption_Attn: a man riding a bike down a trail

Path: /media/rohit/DATA/EverythingD/01SRHBDBA_Acads/Thesis/StoryGenerator/Data/COCO_test2017_41k/test2017/000000375327.jpg

Caption: man riding skateboard down the side of street

Caption_Attn: a man riding a ski board

Path: /media/rohit/DATA/EverythingD/01SRHBDBA_Acads/Thesis/StoryGenerator/Data/COCO_test2017_41k/test2017/000000056139.jpg

Caption: man riding bike next to man on beach

Caption_Attn: a woman with a bicycle on a beach with some kites flying behind her

No Image

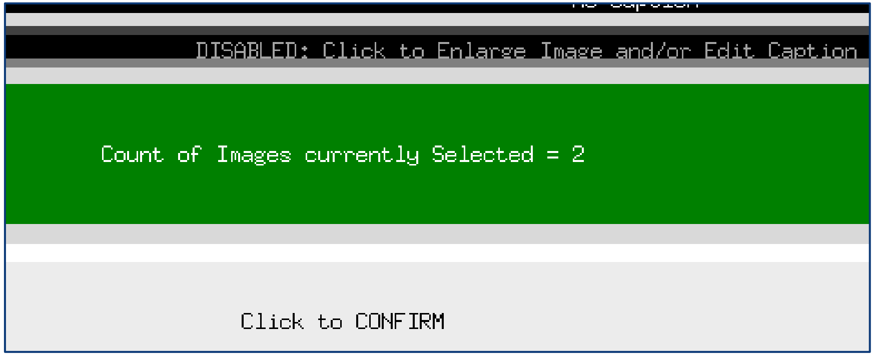
Path: No Path

Caption: No Caption

Caption_Attn: No Caption

Count of Images currently Selected = 2

- For this query:
  - > User edited one caption (first and last images)
    - > only two caption boxes are yellow
  - > User Deselected middle two images
    - > note red border around thumbnail



Count of Images currently Selected = 2

Thank you for your attention!