VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

NII-UIT at TRECVID 2023: Deep video understanding

[1] *University of Information Technology (UIT), Ho Chi Minh City, Vietnam*
[2] *National Institute of Informatics (NII), Tokyo, Japan*

Overview

- Task Introduction
- Challenges
- Our approach
- Experimental Results
- Conclusion

DVU Task Introduction

Input

- Full length feature movie.
- List of scenes *(segmented manually based on semantic)*
- Ontology (vocabulary) *(for the entire dataset)*
- List of entities *(Person - location)*
- 5-7 images of each entities
- Transcript of movies (generated by OpenAI's whisper)

- A barrage of queries *(2 types movie level, 4 types scene level)*

Movie level queries

Fill in the graph query

Given a list of edges (relation) (relationship, event, action, etc…) for node (entity) X. Some (0-3) of those edges may have its target left BLANK.

Return a ranked list of of candidate for node X

Movie level queries

Multiple choice Q&A (151 of them)

- One natural language question
- 6 natural language choices per question
- More questions and higher difficulty than 2022

Scene level queries

Type 1: Find unique scene (optional in 2022) (40 of them)

Given a **full, inclusive** list of interaction between persons (unknown) in a scene, find that scene number.

Type 2: Find person (not required) 18 of them

Given scene number, list of interactions (both to and from) a person X, find X

Scene level queries

Type 3 Find Next interaction (20 of them)

Given a reference scene number, 1 interaction between 2 People X and Y in the scene, and a target scene number.

Find the immediate next interaction between X and Y in target scene number. (Target scene and reference number can be the same

Type 4: Find previous interaction (20 of them)

Type 1,2,3,4 have to be submit in one group.

Scene level query

Type 5: Match scene to text (50 of them)

Given 1 natural language sentence description of a scene and 10 scene numbers. Find the number of the scene best match the description.

Type 6: Scene sentiment classification (50 of them)

Given 1 scene number, a list of 6 sentiment words. Find the correct sentiment label for that scene.

Dataset

video_name	time_of_movie *(hh:mm:ss)*	# scene	time_of_scene *(s)*		
			min	max	avg
Memphis	**1:18:39**	47	17	294	97
Archipelago	**1:50:04**	57	21	389	113
Bonneville	1:32:39	41	19	269	124
heart_machine	1:23:37	**28**	22	**451**	158
Little_Rock	1:22:48	39	24	289	121

Challenges

Challenges

- Very difficult questions (especially this year test set)
  - Required very deep multimedia analysis of the movie to answer
  - Difficult even to human (could required a rewatch of the whole movie to answer a question)
- Full-length-feature-movie
  - Movies are long and diverse in genre
  - Required real-world knowledge to understand
  - Understanding scenes require whole movie level information

Our approach

Scene based entity recognition

- ➔ For ‘character’ entities:
  - ◆ *MTCNN + ArcFace*
- ➔ For ‘Location’ entities:
  - ◆ ResNet 50

Movie - query 1 - Find entity from relations

Only considered “share-scene” relation

Entity rarely have relations with itself

1. Get the list of candidate_entities, i.e *not mentioned in query*.
2. Sort the list of candidate by the number of scenes they share with any mentioned entities
3. Return the ranked list

Movie - query 2 - free form Q&A

Use google's pre-trained universal sentence encoder model.
https://tfhub.dev/google/universal-sentence-encoder-qa/3

Run 1:

- Use the whole movie's ASR text as context.

Run 2:

- Use only the scene where mentioned entities appears as context

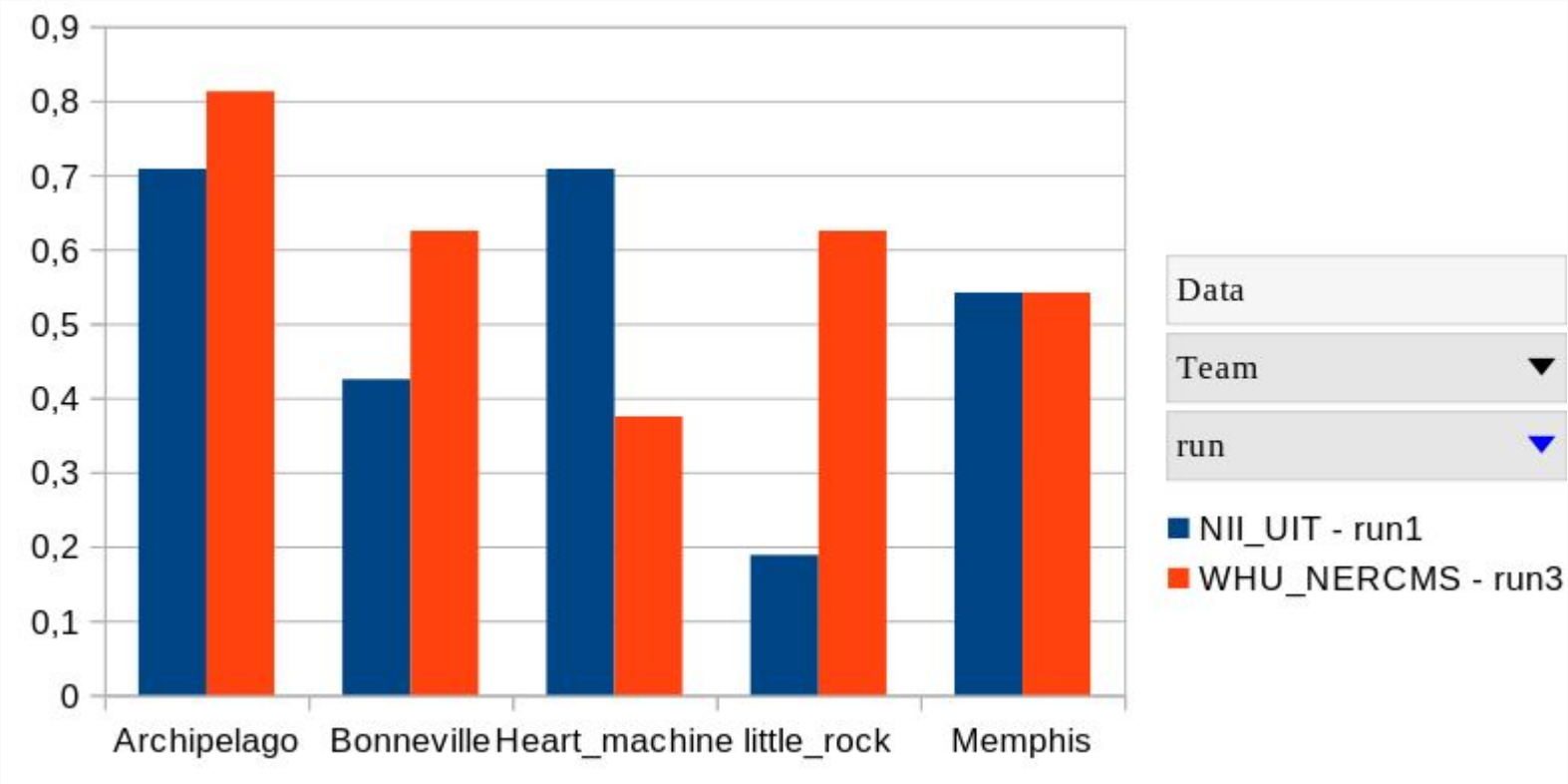
Scene level query type 4-5

Just use Google Universal sentence encoder to match the scenes' ASR with query

- Type 4: Match given description with each scene's ASR in the choices, return highest match
- Type 5: Match given scene's ASR with each sentiment label, return highest match.

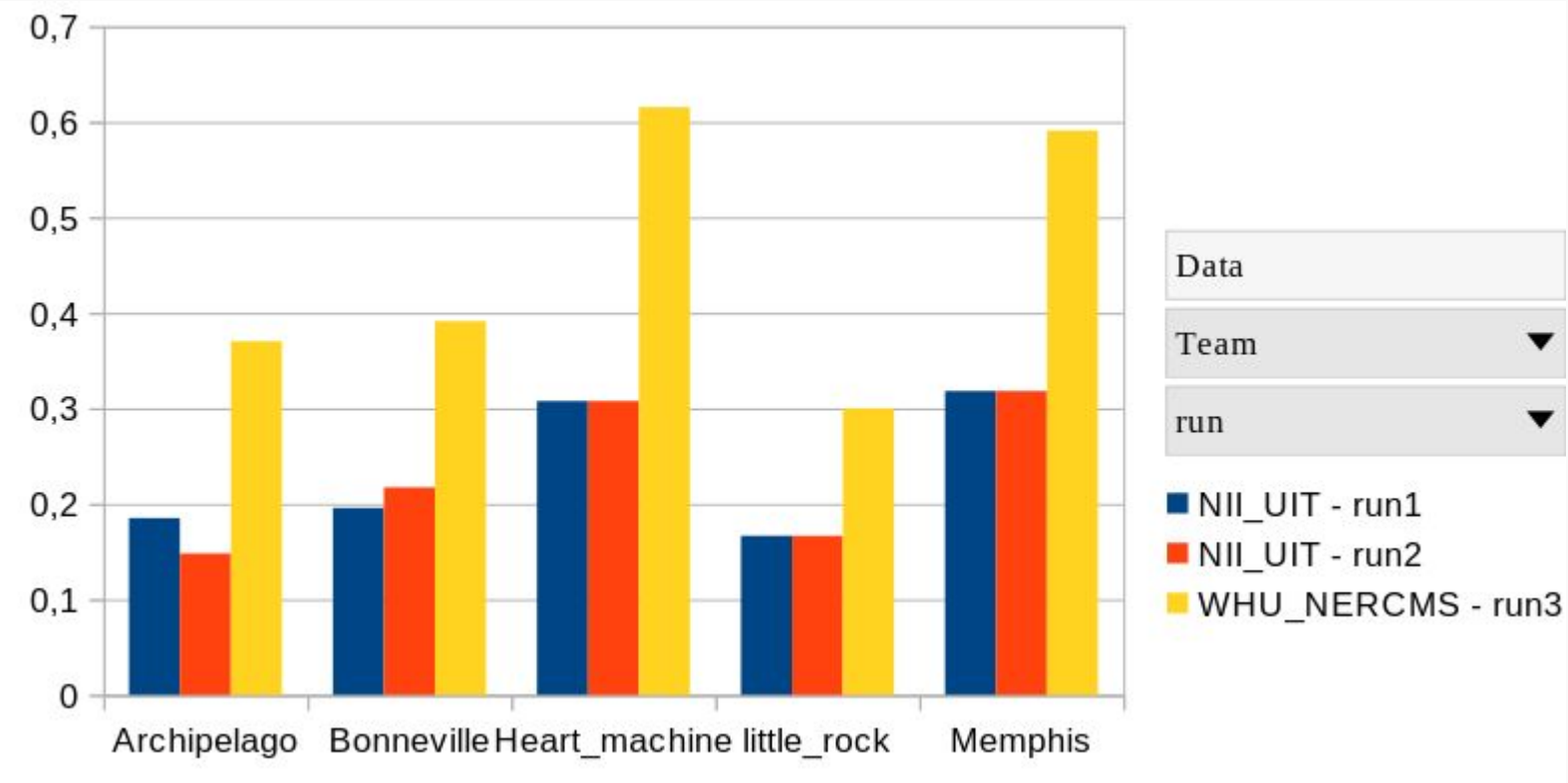
Submission result

Movie level - query 1

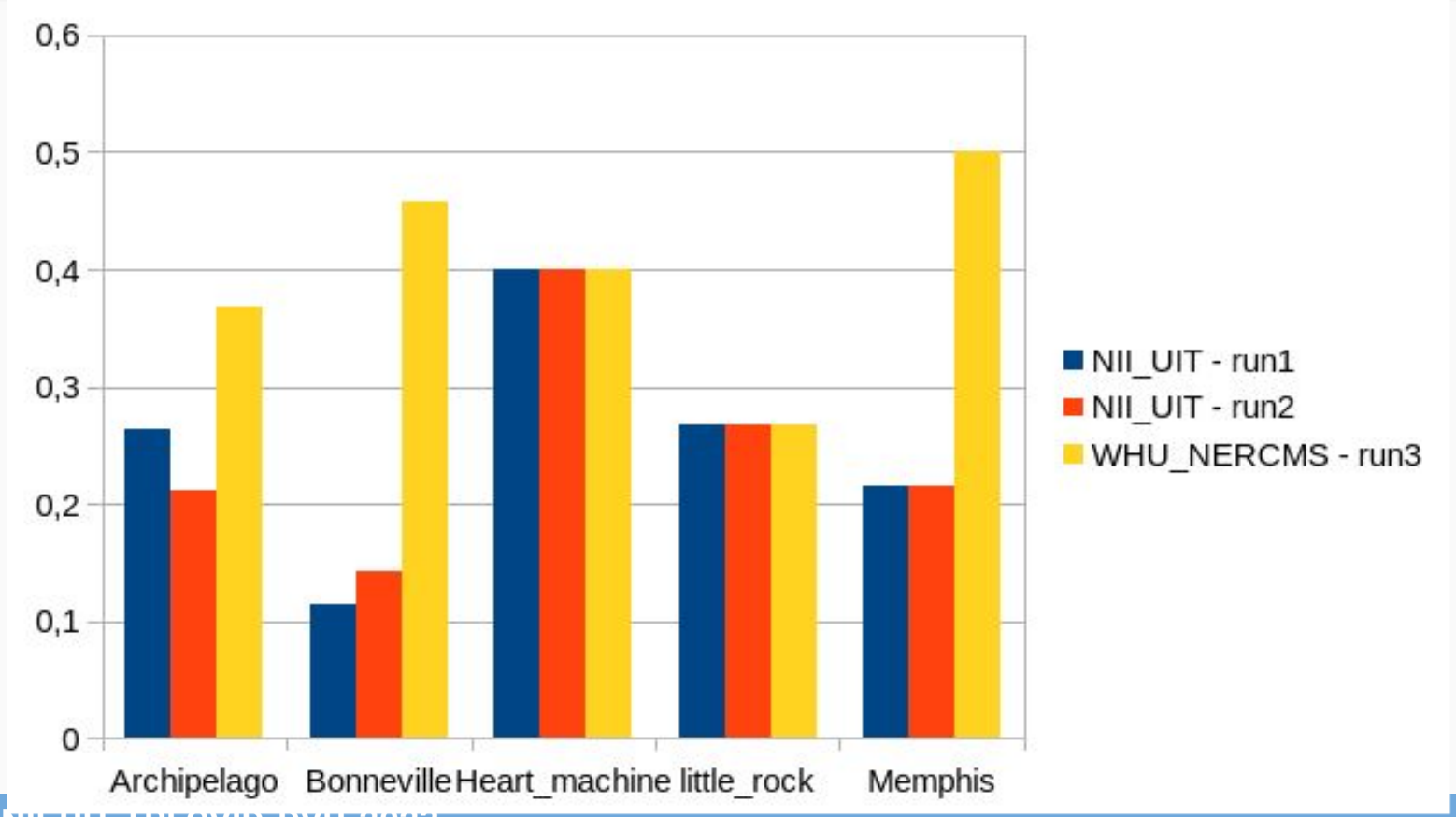


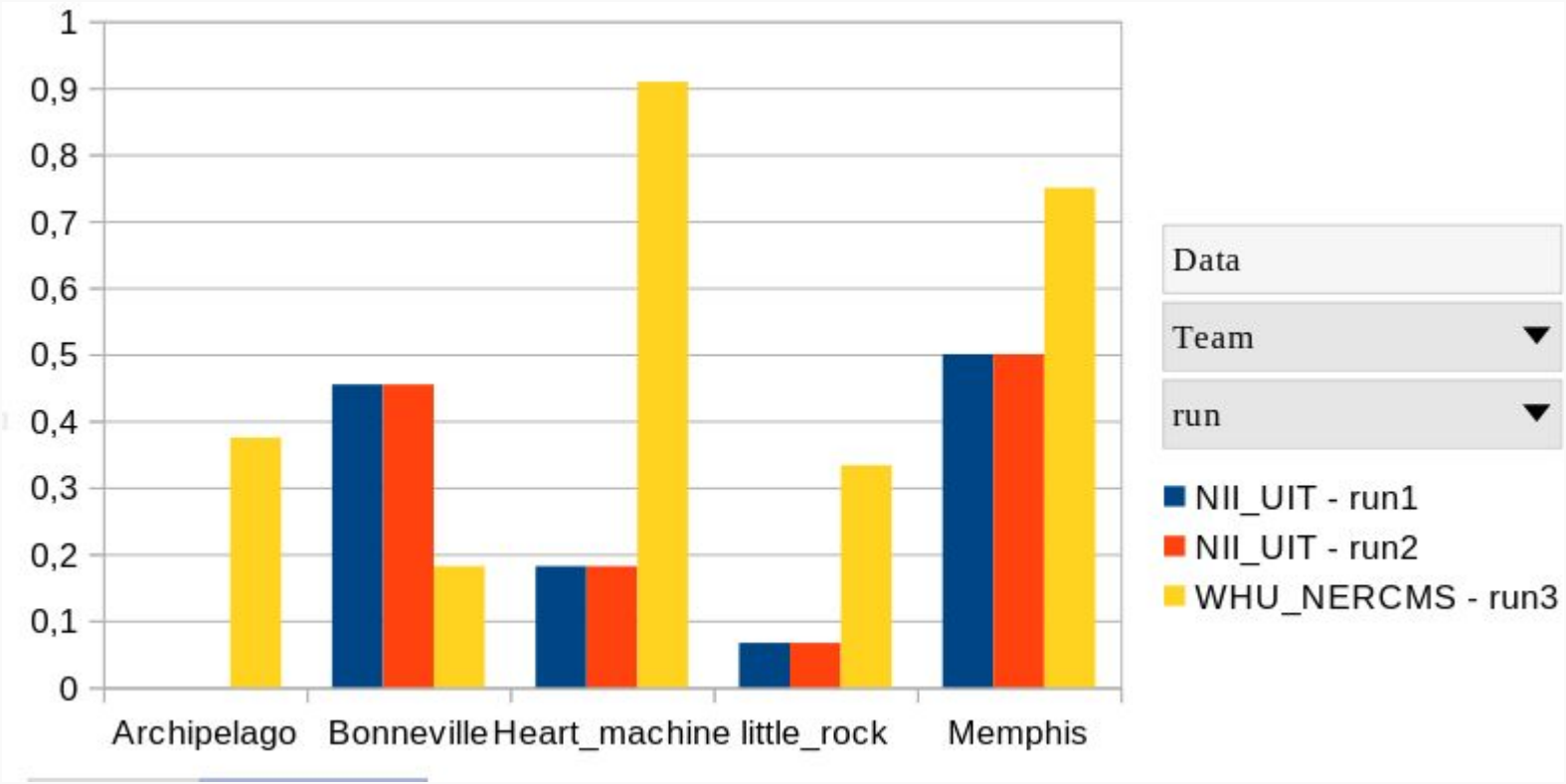
-

Movie - query 2 - natural language Q&A

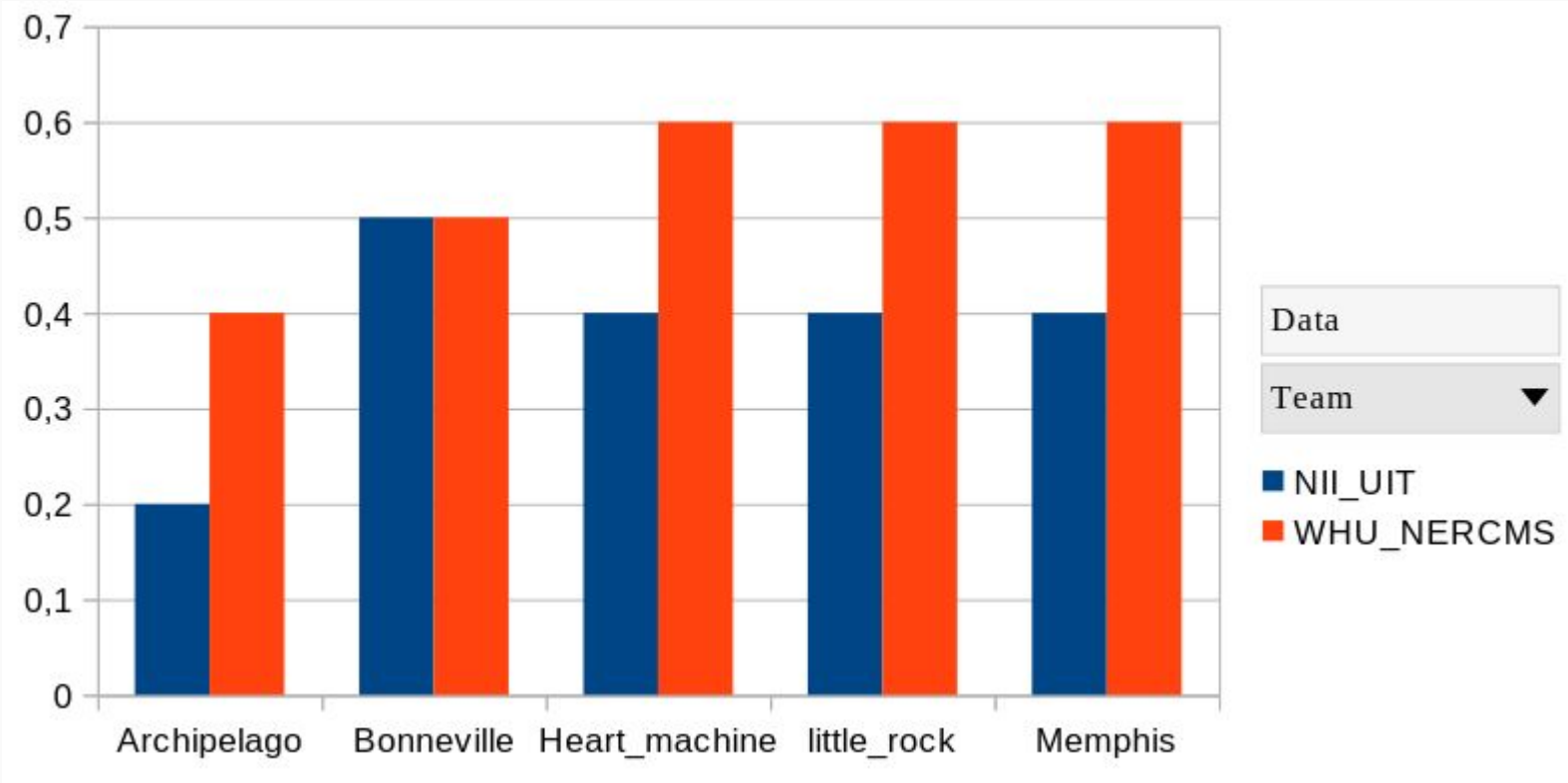


Human generated question

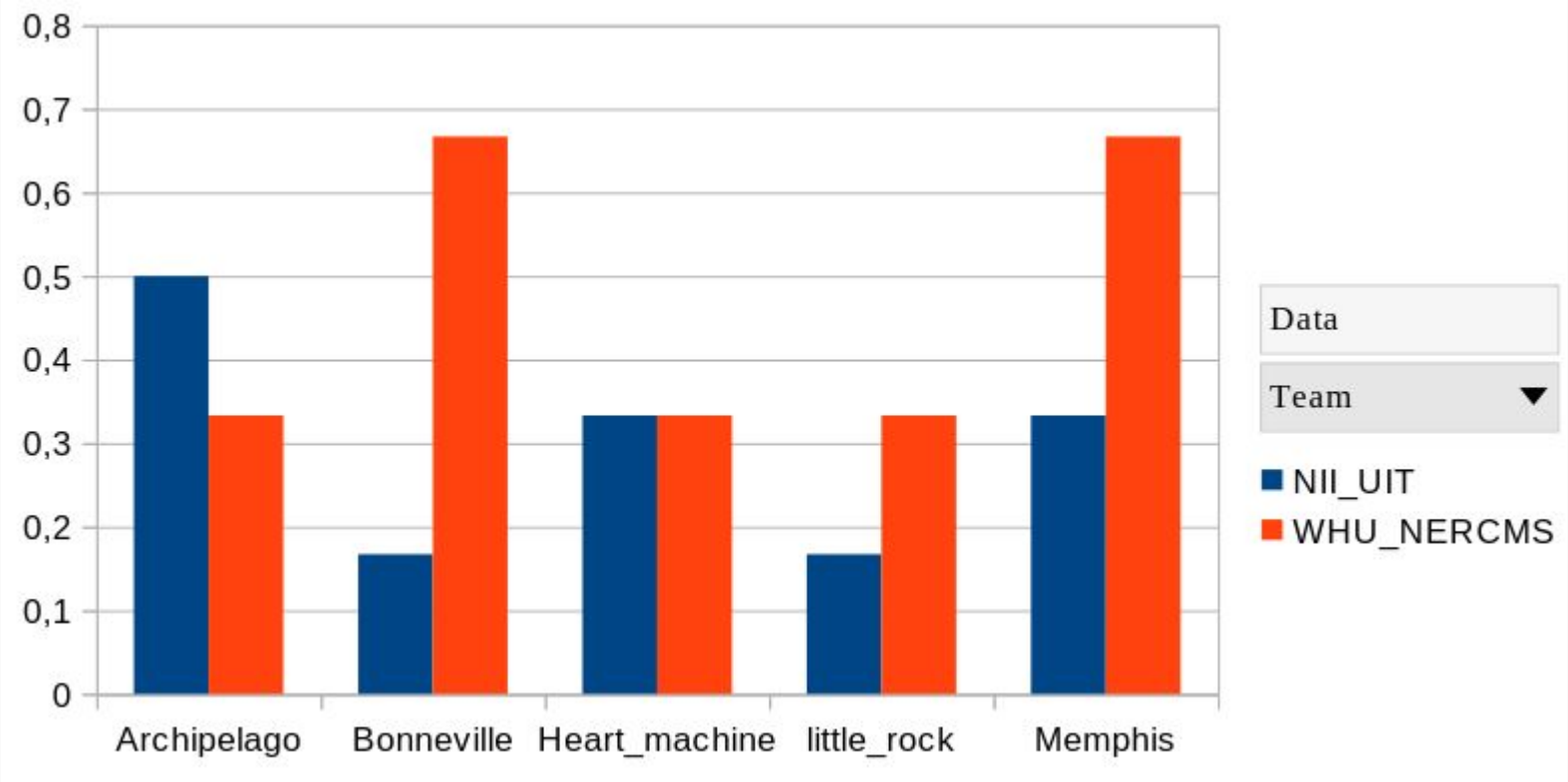



1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Archipelago
Bonneville
Heart_machine
little_rock
Memphis
Data
Team
run
NII_UIT - run1
NII_UIT - run2
WHU_NERCMS - run3

Scene query4-Match scene with description



Scene query - Scene sentiment labeling



Conclusion

Conclusion

- A very simple baseline was proposed for DVU task.
  - Leveraging visual entity recognition and LLM for ASR processing.
  - Stronger baseline than expected
- Very challenging task