TRECVID 2019
Video to Text Description

Asad A. Butt
NIST; Johns Hopkins University

George Awad
NIST; Georgetown University

Yvette Graham
Dublin City University

NIST
National Institute of Standards and Technology

Disclaimer
The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

Goals and Motivations

- ✓Measure how well an automatic system can describe a video in natural language.
- ✓Measure how well an automatic system can match high-level textual descriptions to low-level computer vision features.
- ✓Transfer successful image captioning technology to the video domain.

Real world Applications

- ✓Video summarization
- ✓Supporting search and browsing
- ✓Accessibility - video description to the blind
- ✓Video event prediction

SUBTASKS

- Systems are asked to submit results for two subtasks:
  1. Description Generation (Core):

     Automatically generate a text description for each video.
  2. Matching & Ranking (Optional):

     Return for each video a ranked list of the most likely text description from each of the five sets.

Video Dataset

The VTT data for 2019 consisted of two video sources:

- **Twitter Vine**:
  - Crawled 50k+ Twitter Vine video URLs.
  - Approximate video duration is 6 seconds.
  - Selected 1044 Vine videos for this year's task.
  - Used since inception of VTT task.

- **Flickr:**
  - Flickr video was collected under the Creative Commons License.
  - A set of 91 videos was collected, which was divided into 74,958 segments.
  - Approximate video duration is 10 seconds.
  - Selected 1010 segments.

Dataset Cleaning

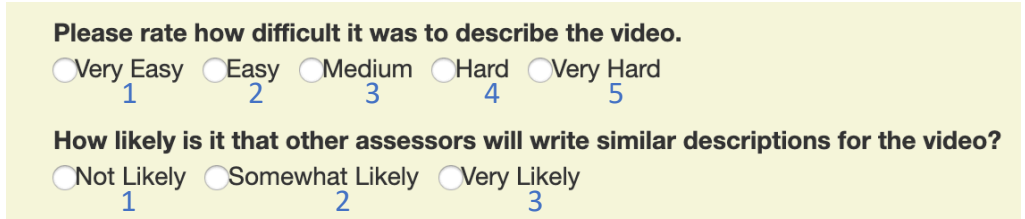
- Before selecting the dataset, we clustered videos based on visual similarity.
  - Resulted in the removal of duplicate videos, as well as those which were very visually similar (e.g. soccer games), resulting in a more diverse set of videos.
- Then, we manually went through large collection of videos.
  - Used list of commonly appearing topics to filter videos.
  - Removed videos with multiple, unrelated segments that are hard to describe.
  - Removed any animated (or otherwise unsuitable) videos.

Annotation Process

- A total of 10 assessors annotated the videos.
- Each video was annotated by 5 assessors.
  - Annotation guidelines by NIST:
    - For each video, annotators were asked to combine 4 facets *if applicable*:
      - Who is the video showing (objects, persons, animals, …etc) ?
      - What are the objects and beings doing (actions, states, events, …etc)?
      - Where (locale, site, place, geographic, ...etc) ?
      - When (time of day, season, ...etc) ?

Annotation – Observations

- Questions asked:



  - Q1 Avg Score: 2.03 (scale of 5)
  - Q2 Avg Score: 2.51 (scale of 3)

- Correlation between difficulty scores: -0.72

- Average Sentence Length for each assessor:

Assessor #	Avg. Length
1	17.72
2	19.55
3	18.76
4	22.07
5	20.42
6	12.83
7	16.07
8	21.73
9	16.49
10	21.16

2019 Participants (10 teams finished)

	Matching & Ranking (11 Runs)	Description Generation (30 Runs)
IMFD_IMPRESEE	✓	✓
KSLAB	✓	✓
RUCMM	✓	✓
RUC_AIM3	✓	✓
EURECOM_MeMAD		✓
FDU		✓
INSIGHT_DCU		✓
KU_ISPL		✓
PICSOM		✓
UTS_ISA		✓

Run Types

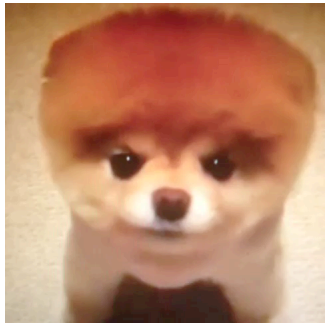
- Each run was classified by the following run type:
  - 'I': Only image captioning datasets were used for training.
  - 'V': Only video captioning datasets were used for training.
  - 'B': Both image and video captioning datasets were used for training.

Run Types

- All runs in Matching and Ranking are of type ‘V’.
- For Description Generation the distribution is:
  - Run type ‘I’: 1 run
  - Run type ‘B’: 3 runs
  - Run type ‘V’: 26 runs

Subtask 1: Description Generation

Given a video



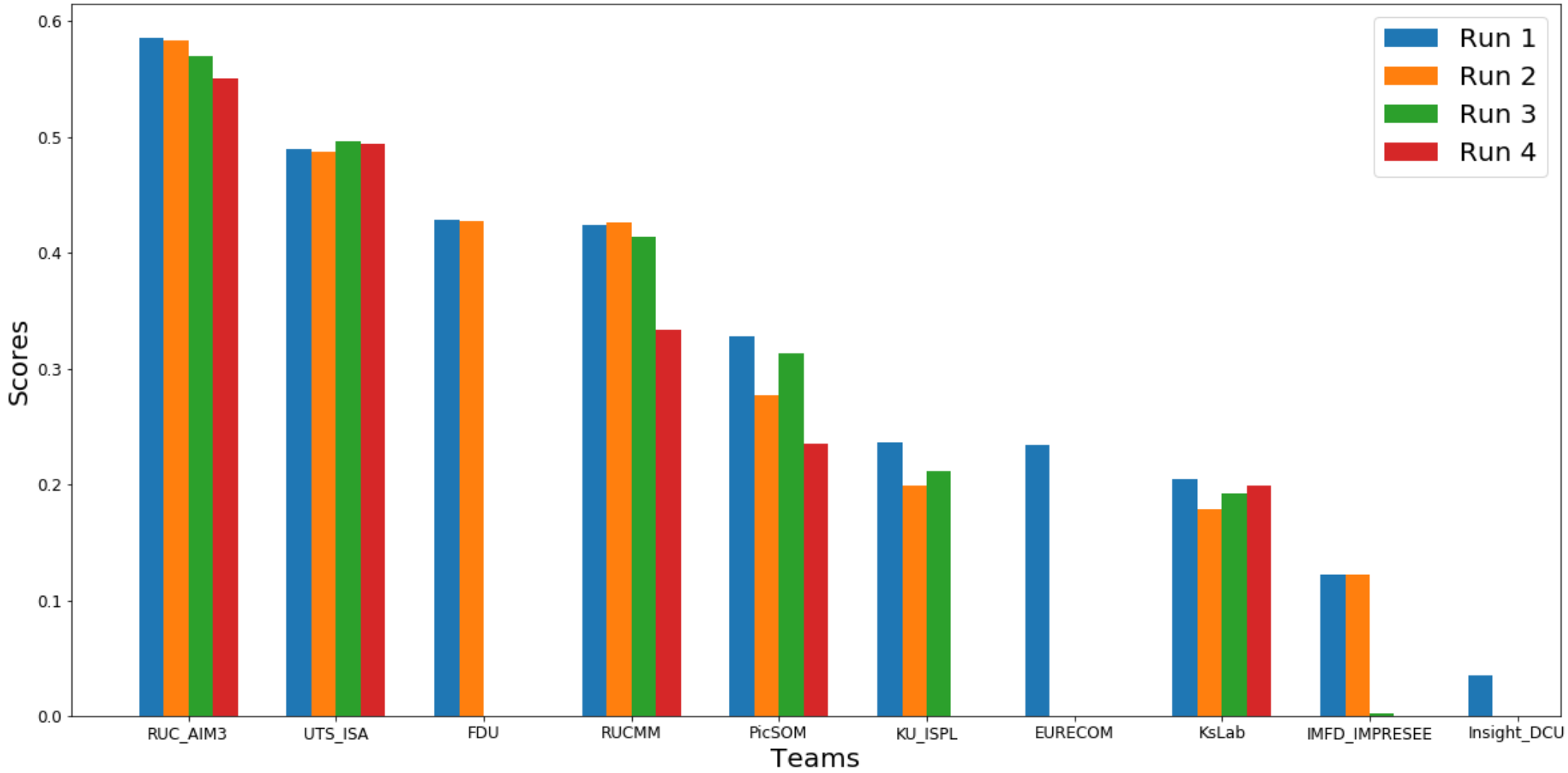
Generate a textual description

Who ? What ? Where ? When ?

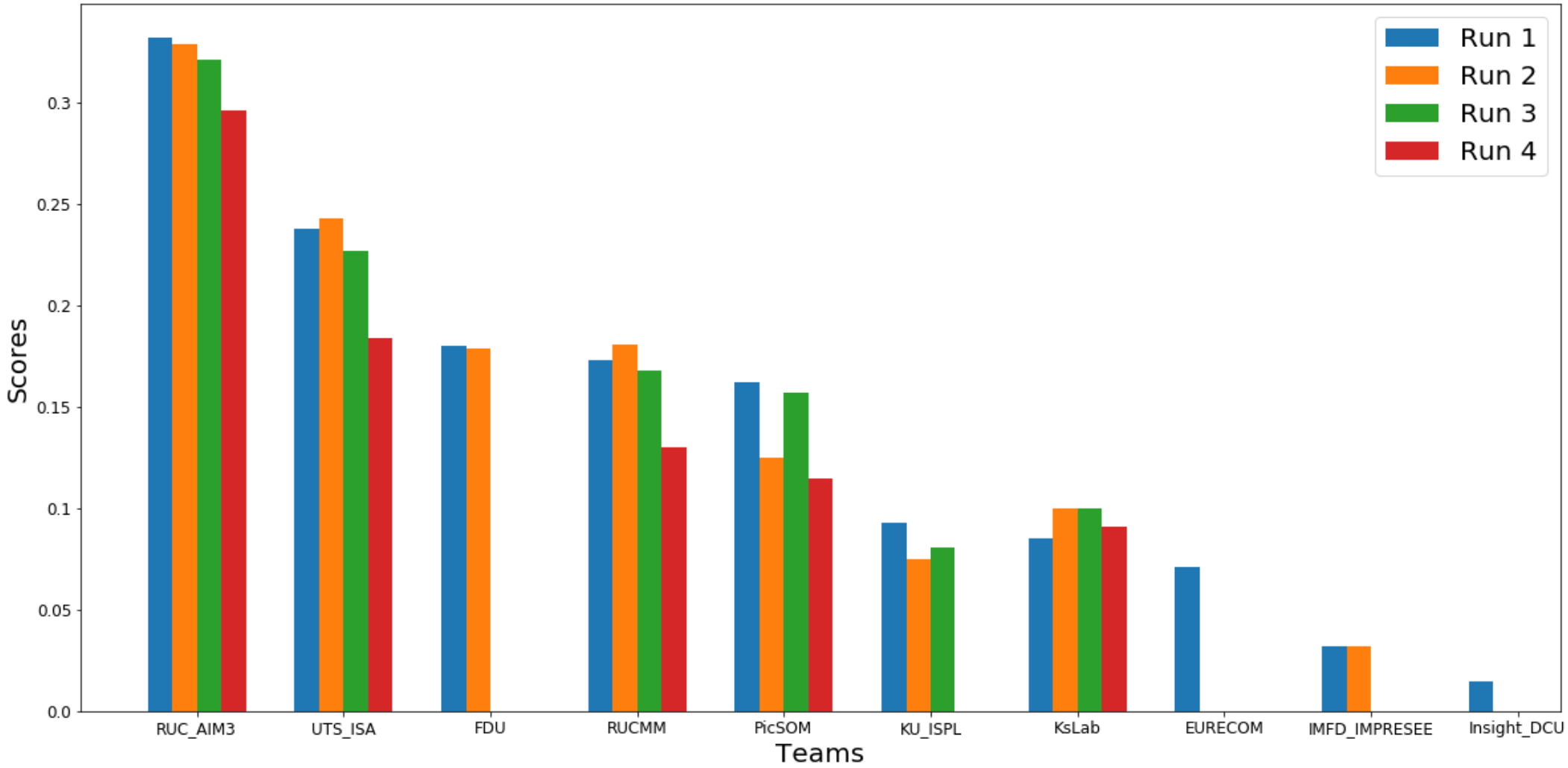
"a dog is licking its nose"

- Up to 4 runs in the *Description Generation* subtask.
- Metrics used for evaluation:
  - CIDEr (Consensus-based Image Description Evaluation)
  - METEOR (Metric for Evaluation of Translation with Explicit Ordering)
  - BLEU (BiLingual Evaluation Understudy)
  - STS (Semantic Textual Similarity)
  - DA (Direct Assessment), which is a crowdsourced rating of captions using Amazon Mechanical Turk (AMT)

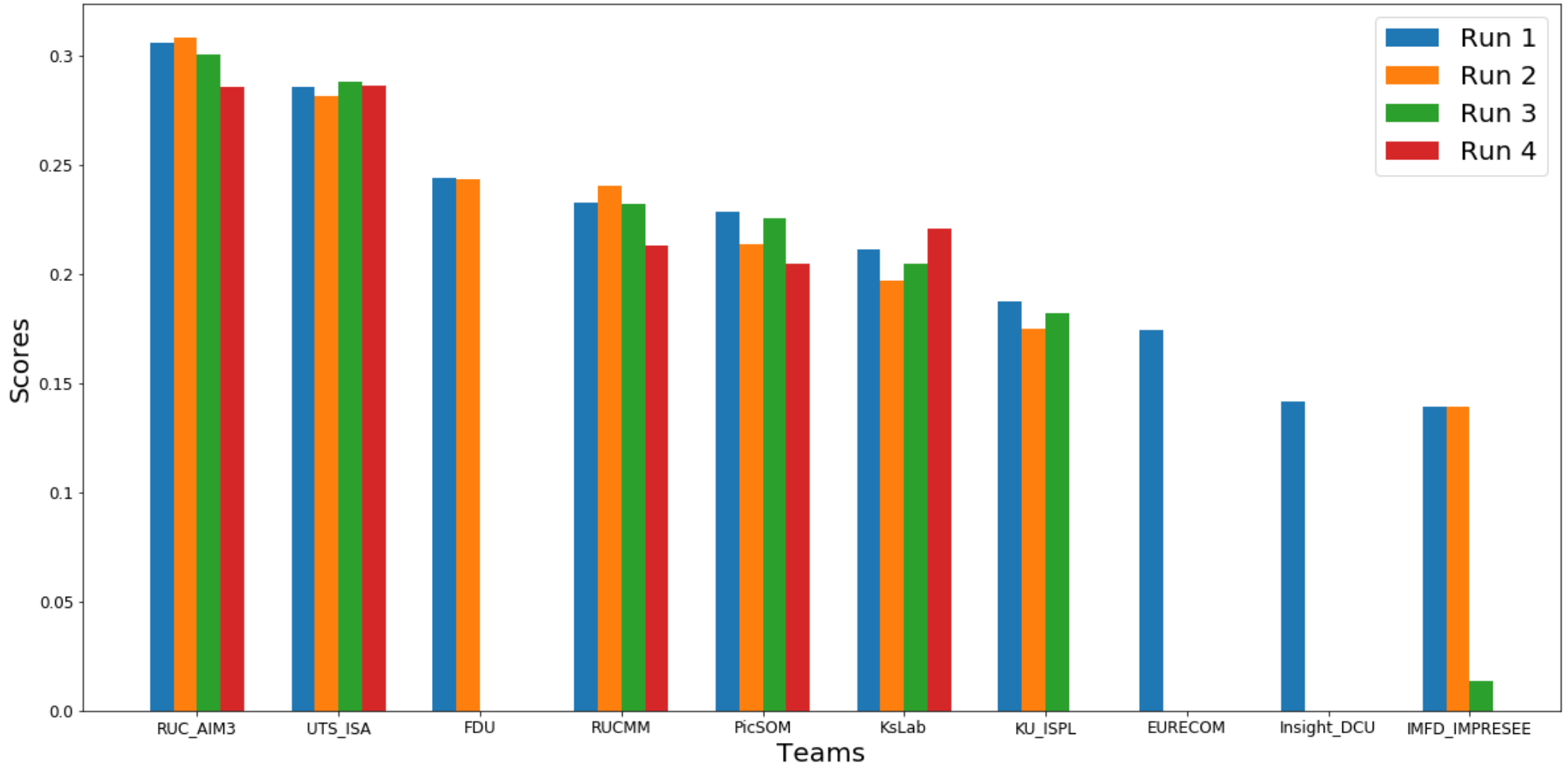
CIDER Results



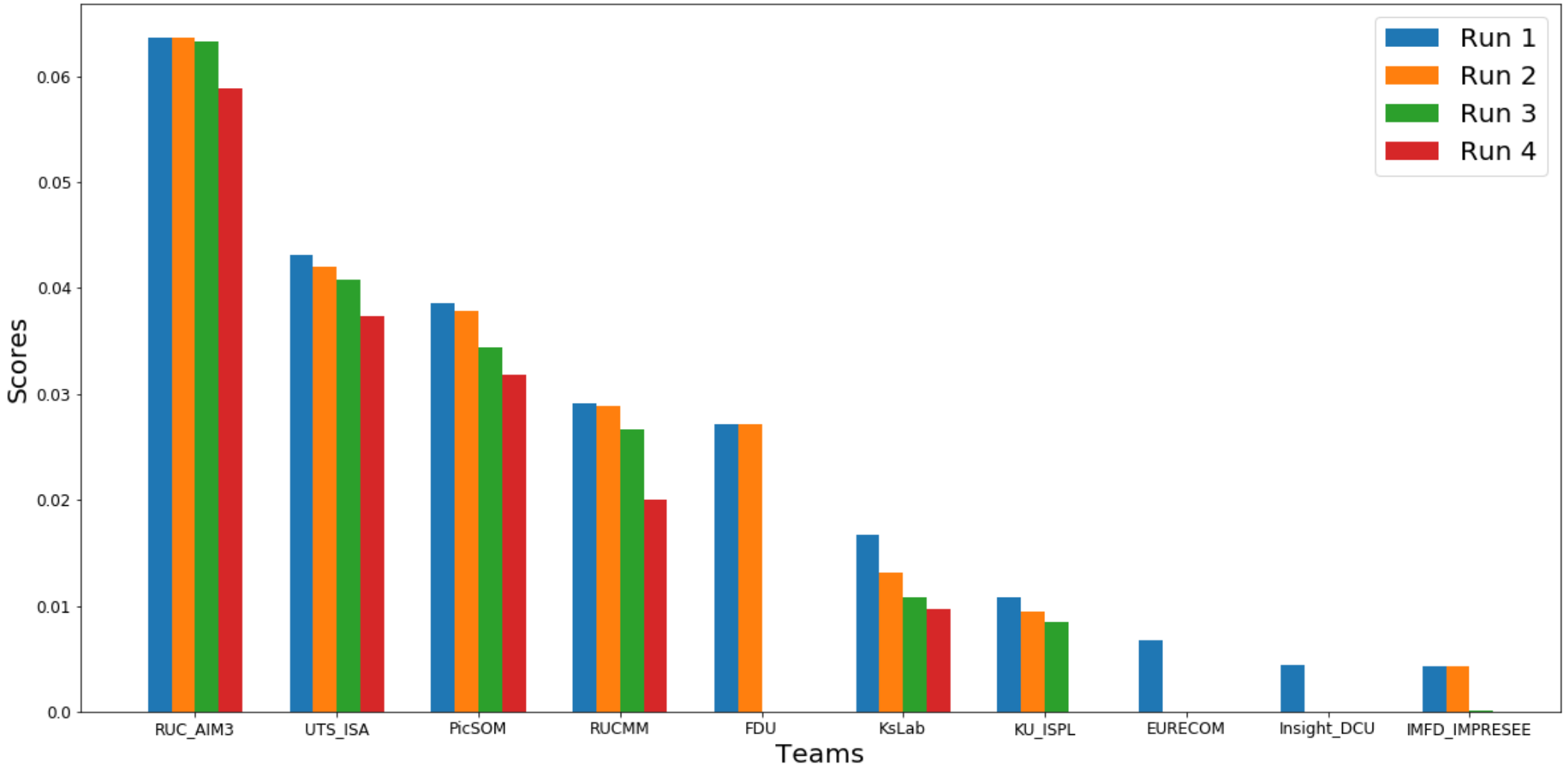
CIDER-D Results



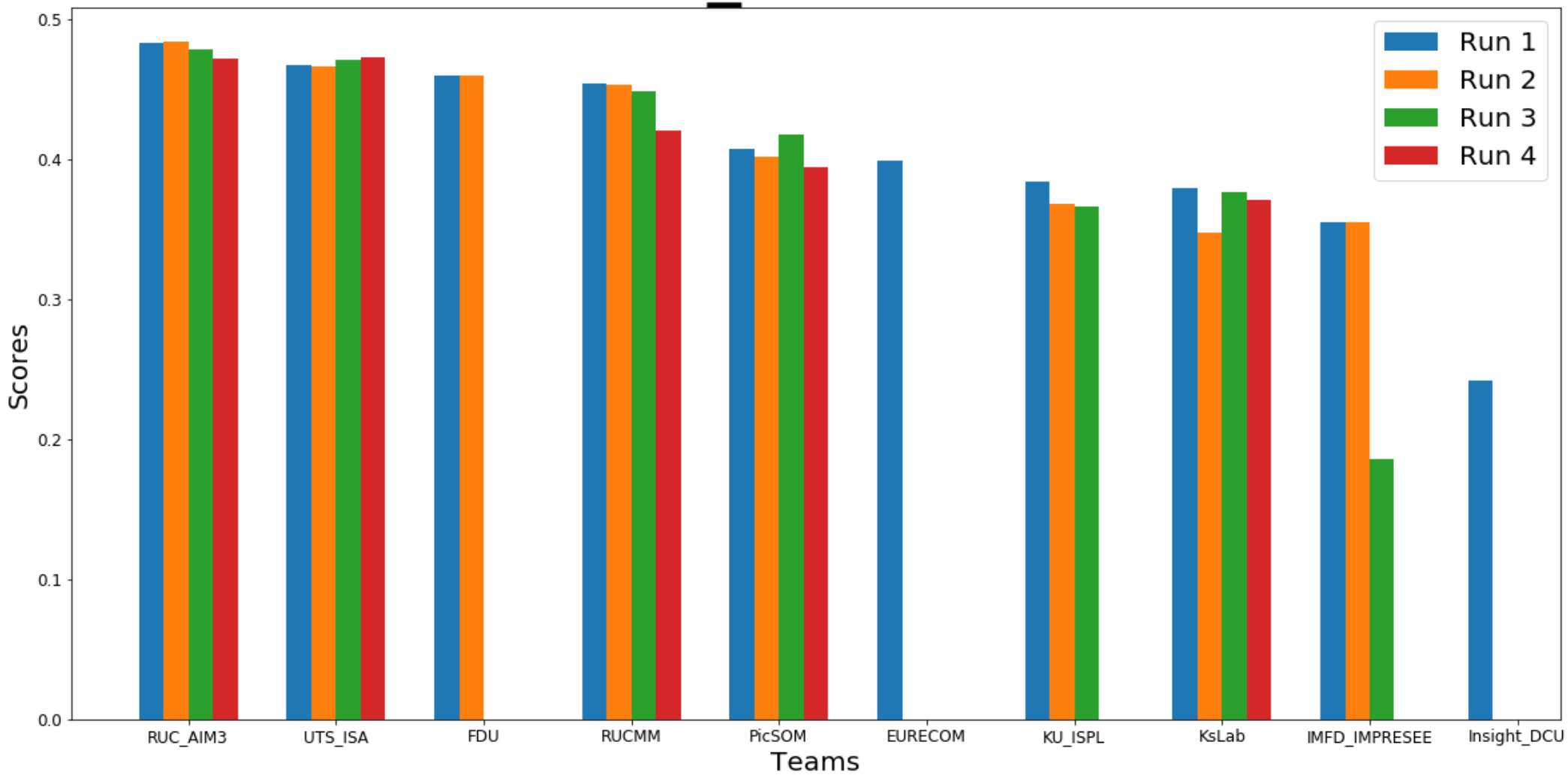
METEOR Results



BLEU Results

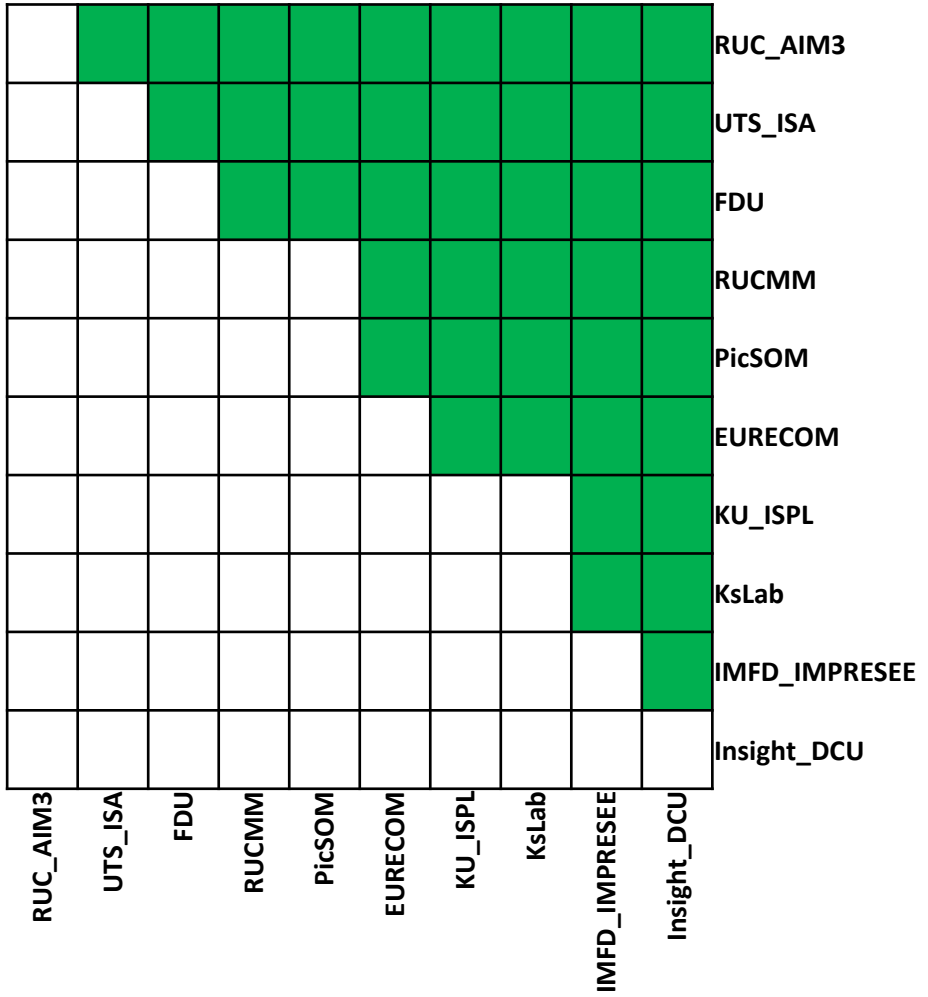


STS_1 Results



Significance Test – CIDEr

- Green squares indicate a significant “win” for the row over column using the CIDEr metric.
- Significance calculated at p<0.001
- RUC_AIM3 outperforms all other systems.



Metric Correlation

	CIDER	CIDER-D	METEOR	BLEU	STS_1	STS_2	STS_3	STS_4	STS_5
CIDER	1.000	0.964	0.923	0.902	0.929	0.900	0.910	0.887	0.900
CIDER-D	0.964	1.000	0.903	0.958	0.848	0.815	0.828	0.800	0.816
METEOR	0.923	0.903	1.000	0.850	0.928	0.916	0.921	0.891	0.904
BLEU	0.902	0.958	0.850	1.000	0.775	0.742	0.752	0.724	0.741
STS_1	0.929	0.848	0.928	0.775	1.000	0.997	0.998	0.990	0.994
STS_2	0.900	0.815	0.916	0.742	0.997	1.000	0.999	0.995	0.997
STS_3	0.910	0.828	0.921	0.752	0.998	0.999	1.000	0.995	0.997
STS_4	0.887	0.800	0.891	0.724	0.990	0.995	0.995	1.000	0.998
STS_5	0.900	0.816	0.904	0.741	0.994	0.997	0.997	0.998	1.000

Comparison with 2018

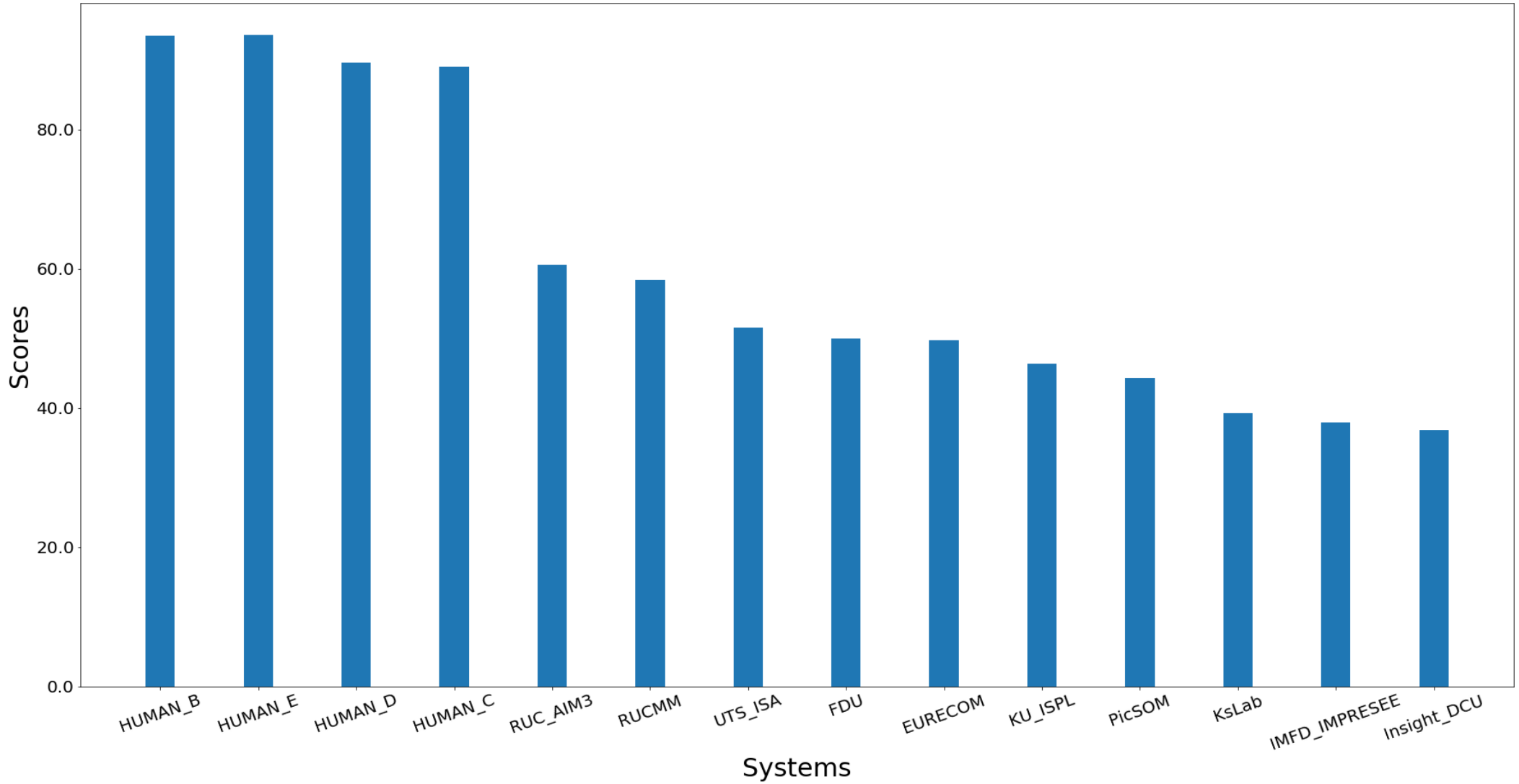
- Scores have increased across all metrics from last year.
- The table shows the maximum score for each metric from 2018 and 2019.

Metric	2018	2019
CIDEr	0.416	0.585
CIDEr-D	0.154	0.332
METEOR	0.231	0.306
BLEU	0.024	0.064
STS	0.433	0.484

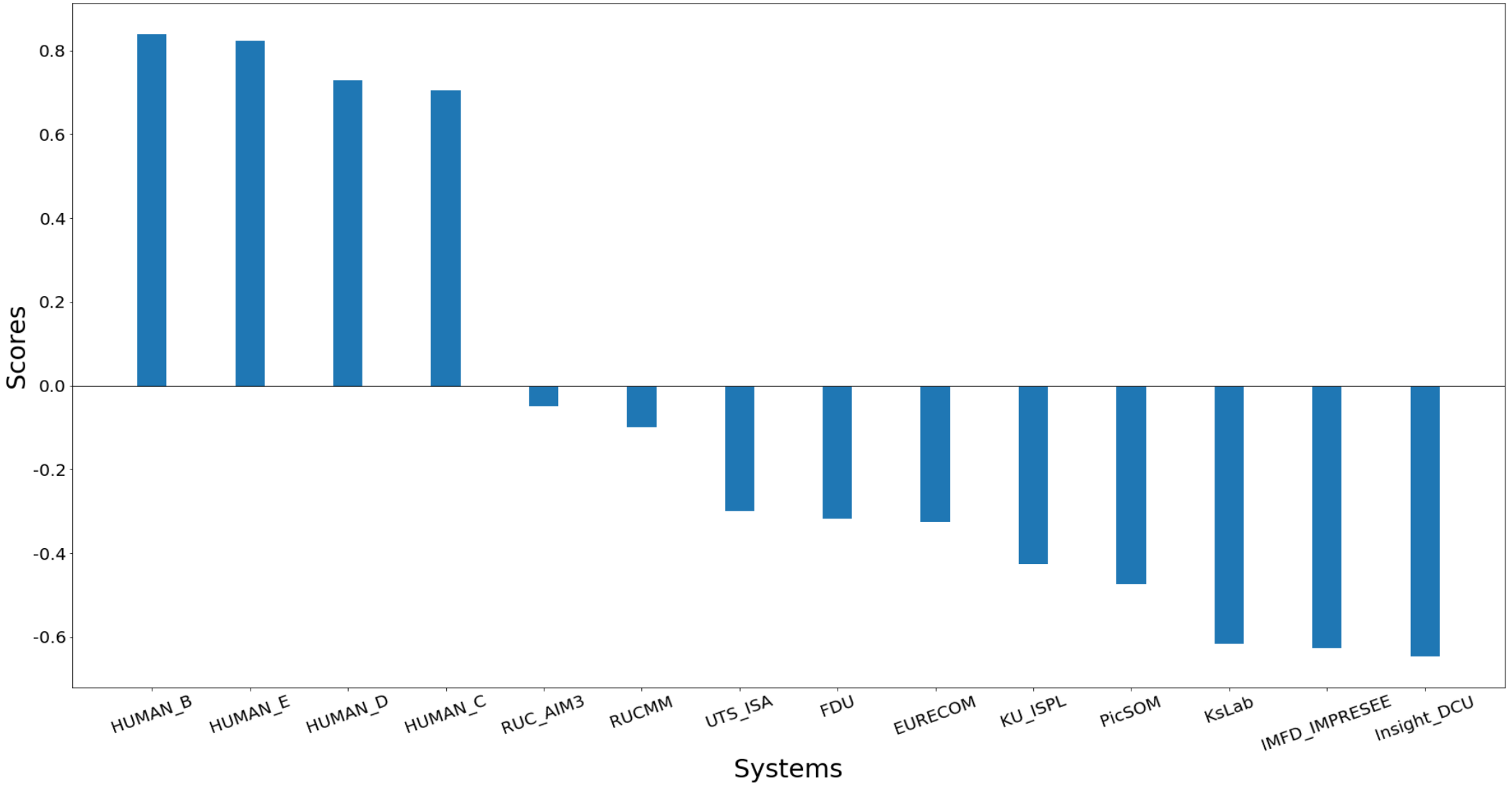
Direct Assessment (DA)

- DA uses crowdsourcing to evaluate how well a caption describes a video.
- Human evaluators rate captions on a scale of 0 to 100.
- DA conducted on only primary runs for each team.
- Measures …
  - **RAW**: Average DA score [0..100] for each system (non-standardized) – micro-averaged per caption then overall average
  - **Z**: Average DA score per system after standardization per individual AMT worker's mean and std. dev. score.

DA Results - Raw

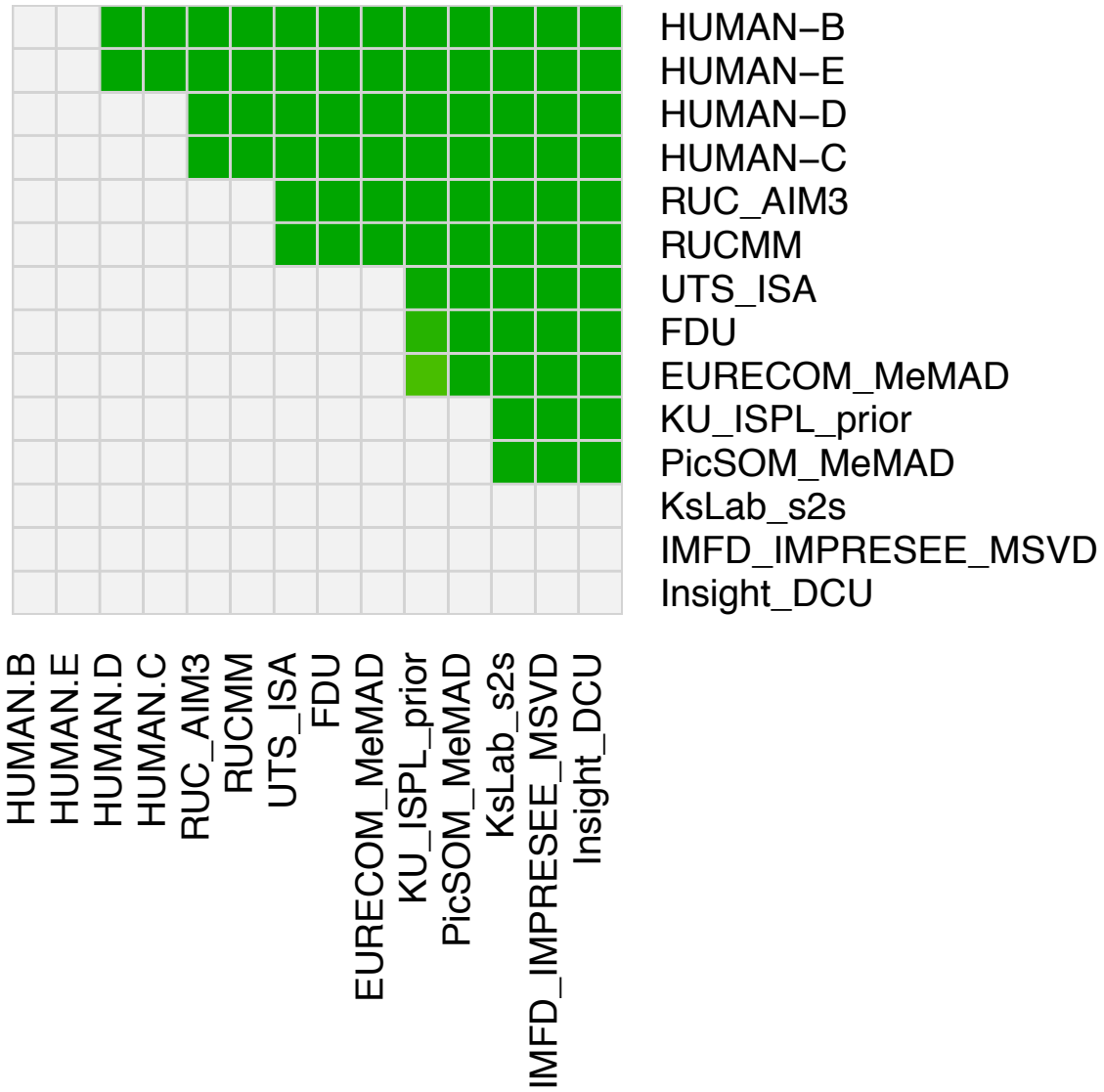


DA Results - Z


Scores
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
HUMAN_B
HUMAN_E
HUMAN_D
HUMAN_C
RUC_AIM3
RUCMM
UTS_ISA
FDU
EURECOM
KU_ISPL
PicSOM
KsLab
IMFD_IMPRESEE
Insight_DCU
Systems

What DA Results Tell Us ..

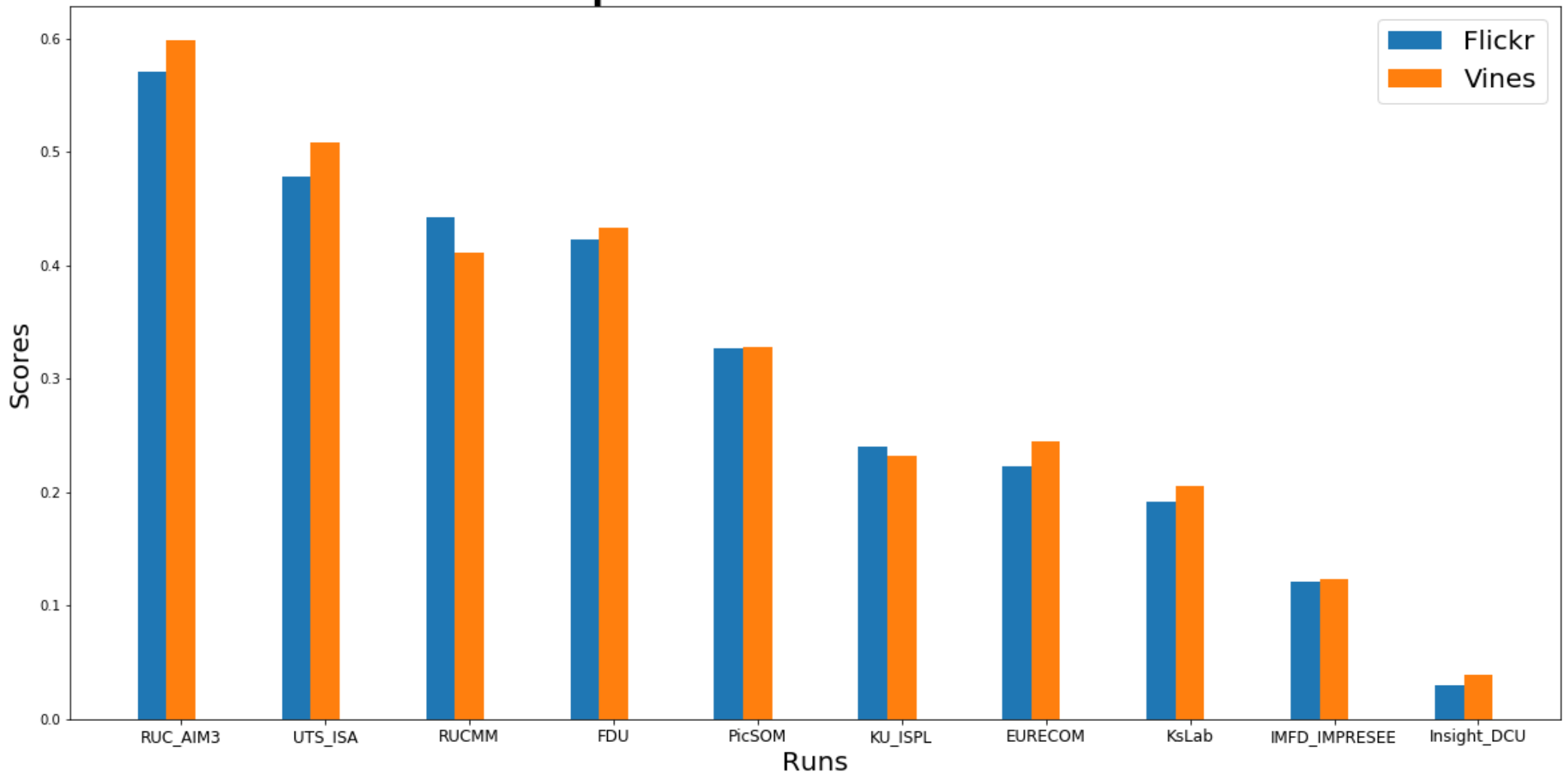
- Green squares indicate a significant “win” for the row over the column.
- No system yet reaches human performance.
- Humans B and E statistically perform better than Humans C and D. This may not be significant since each ‘Human’ system contains multiple assessors.
- Amongst systems, RUC-AIM3 and RUCMM outperform the rest, with significant wins.



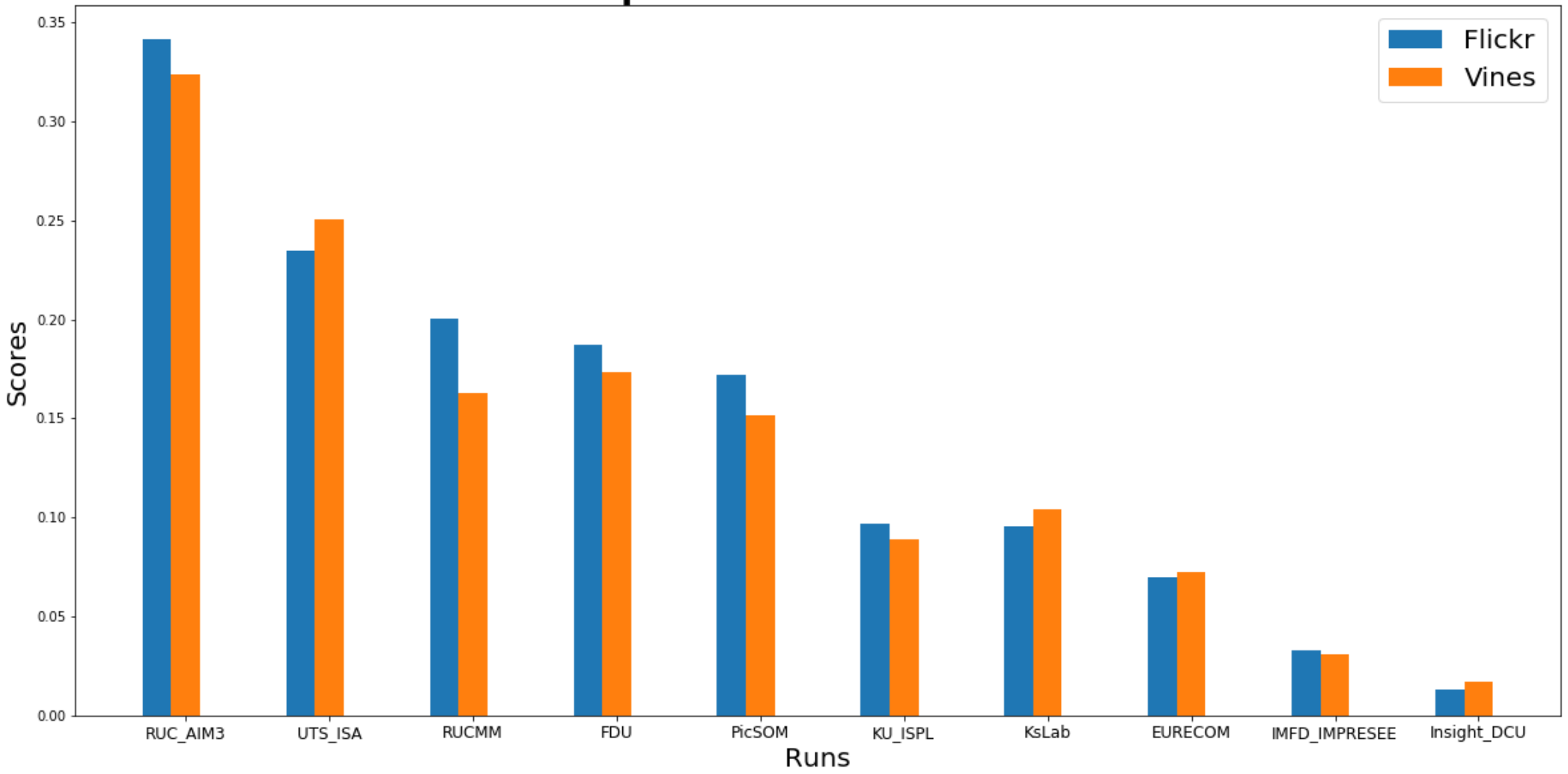
Correlation Between Metrics (Primary Runs)

	CIDER	CIDER-D	METEOR	BLEU	STS	DA_Z
CIDER	1.000	0.972	0.963	0.902	0.937	0.874
CIDER-D	0.972	1.000	0.967	0.969	0.852	0.832
METEOR	0.963	0.967	1.000	0.936	0.863	0.763
BLEU	0.902	0.969	0.936	1.000	0.750	0.711
STS	0.937	0.852	0.863	0.750	1.000	0.812
DA_Z	0.874	0.832	0.763	0.711	0.812	1.000

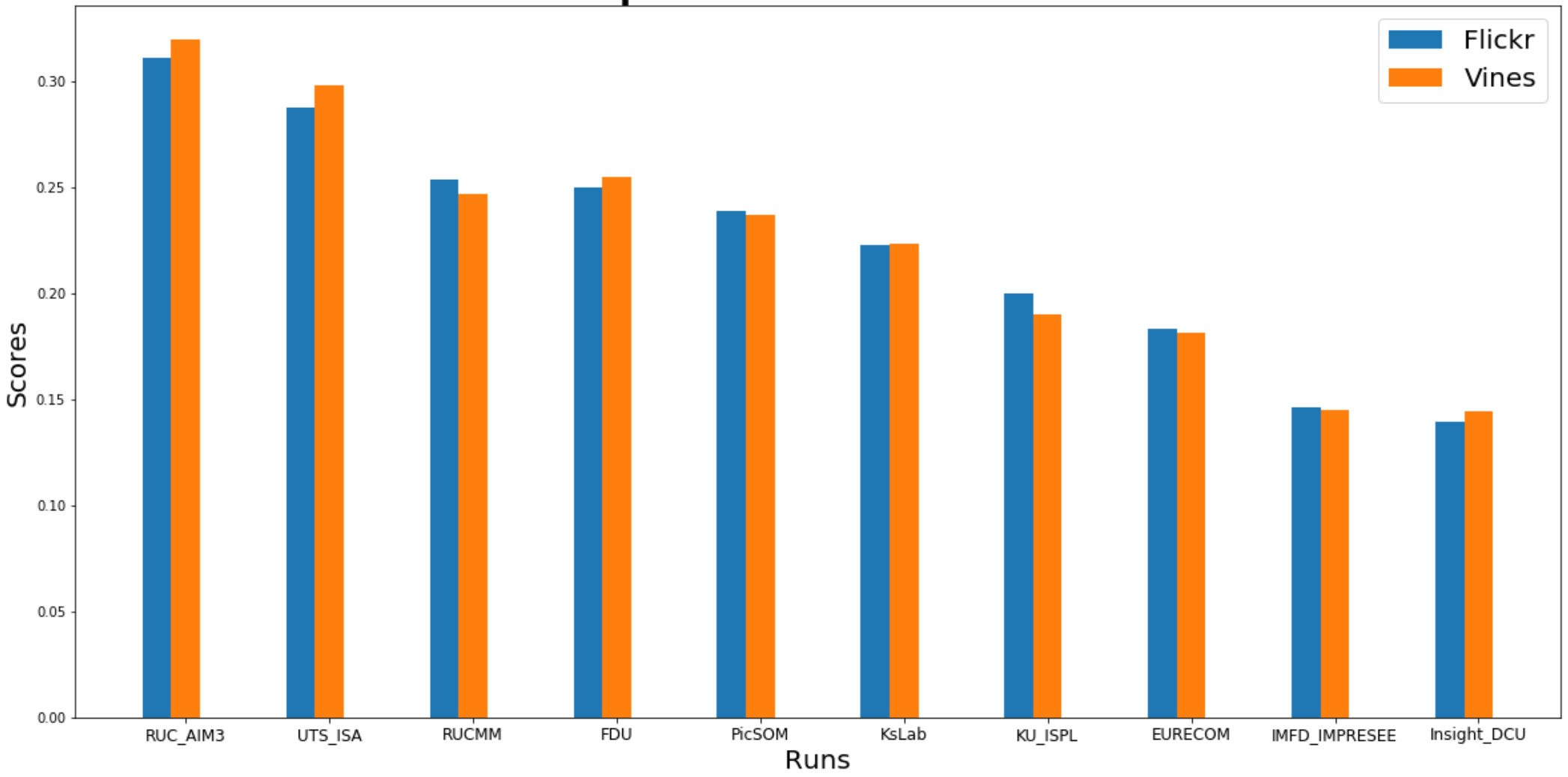
CIDEr Comparison - Flickr vs. Vines



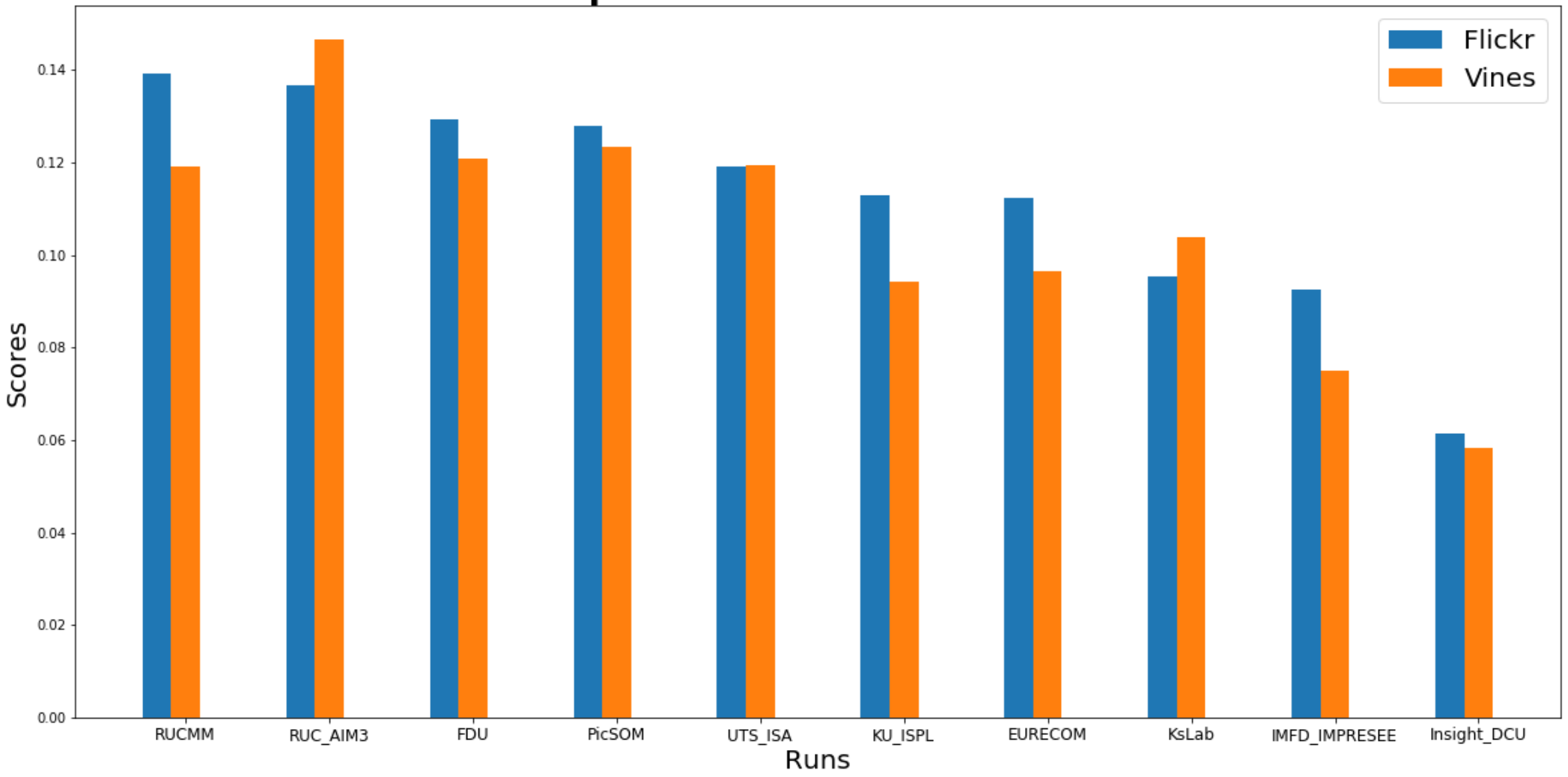
CIDErD Comparison - Flickr vs. Vines



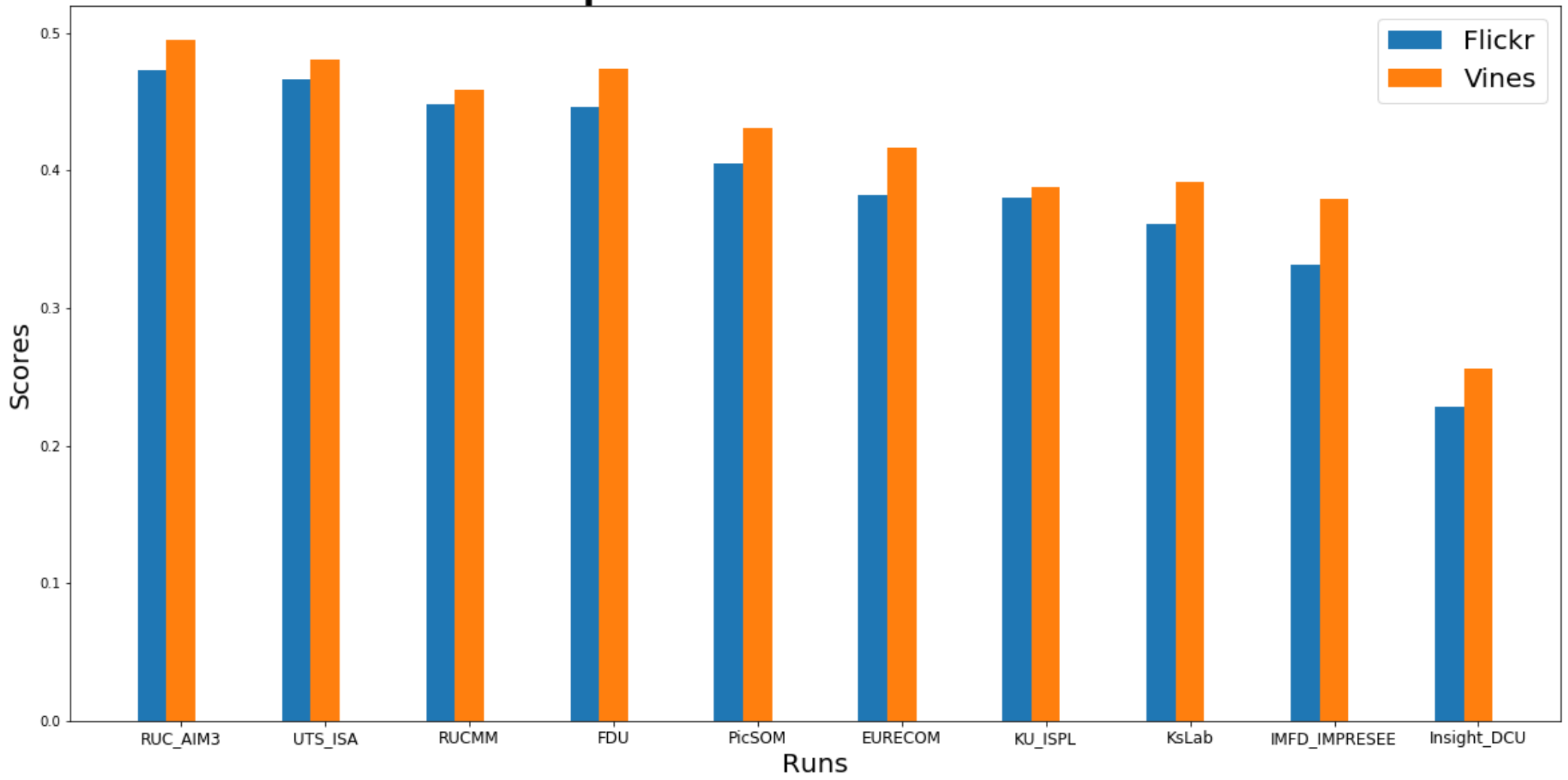
Meteor Comparison - Flickr vs. Vines



BLEU Comparison - Flickr vs. Vines



STS1 Comparison - Flickr vs. Vines



Flickr vs Vines

- Table 1 shows the average sentence lengths for different runs over the Flickr and Vines datasets.
- The GT average sentence lengths are as follows:

Flickr	Vines
17.48	18.85

- There is no significant difference to show that the sentence length played any role in score differences.
- It is difficult to reach a conclusion regarding the difficulty/ease of one dataset over the other.

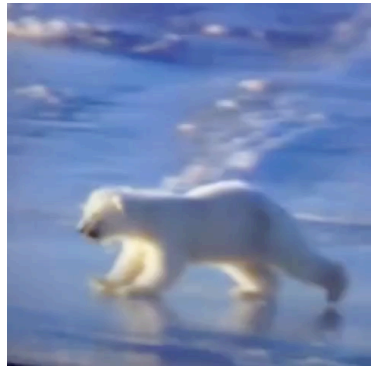
Team	Flickr	Vines
IMFD_IMPRESEE	5.49	5.41
EURECOM	6.16	6.21
RUCMM	7.63	7.93
KU_ISPL	7.72	7.64
PicSOM	8.58	9.09
FDU	9.06	9.44
KsLab	9.50	9.95
Insight_DCU	11.59	12.23
RUC_AIM3	12.62	11.63
UTS_ISA	15.16	15.32

Table 1

Top 3 Results – Description Generation

Assessor Captions:
1. White male teenager in a black jacket playing a guitar and singing into a microphone in a room
2. Young man sits in front of mike, strums guitar, and sings.
3. A man plays guitar in front of a white wall inside.
4. a young man in a room plays guitar and sings into a microphone
5. A young man plays a guitar and sings a song while looking at the camera.

#1439



#1080

#826

Bottom 3 Results – Description Generation

Assessor Captions:
1. Two knitted finger puppets rub against each other in front of white cloth with pink and yellow squares
2. two finger's dolls are hugging.
3. Two finger puppet cats, on beige and white and on black and yellow, embrace in front of a polka dot background.
4. two finger puppets hugging each other
5. Two finger puppets embrace in front of a background that is white with colored blocks printed on it.

#688

#1330

#913

Example of System Captions

1. a man is singing and playing guitar
2. a man is playing a guitar and singing
3. a man is playing a guitar
4. a man is playing a guitar and playing the guitar in front of a microphone
5. a man is sitting in a chair and playing a guitar and singing
6. a young man singing into a microphone in a room in front of a guitar
7. a man is sitting at a desk and talking
8. a man is talking about a video

Observations – Description Generation

- This subtask captures the essence of the VTT task as systems try to describe videos in natural language.
- It was made mandatory for VTT participants for the first time.
- A number of metrics were used to evaluate results.
- For the first time, multiple video sources were used.
  - No obvious advantage/disadvantage for the sources. Probably because care is taken to get a diverse set of real world videos.

Subtask 2: Matching & Ranking

Person reading newspaper outdoors at daytime

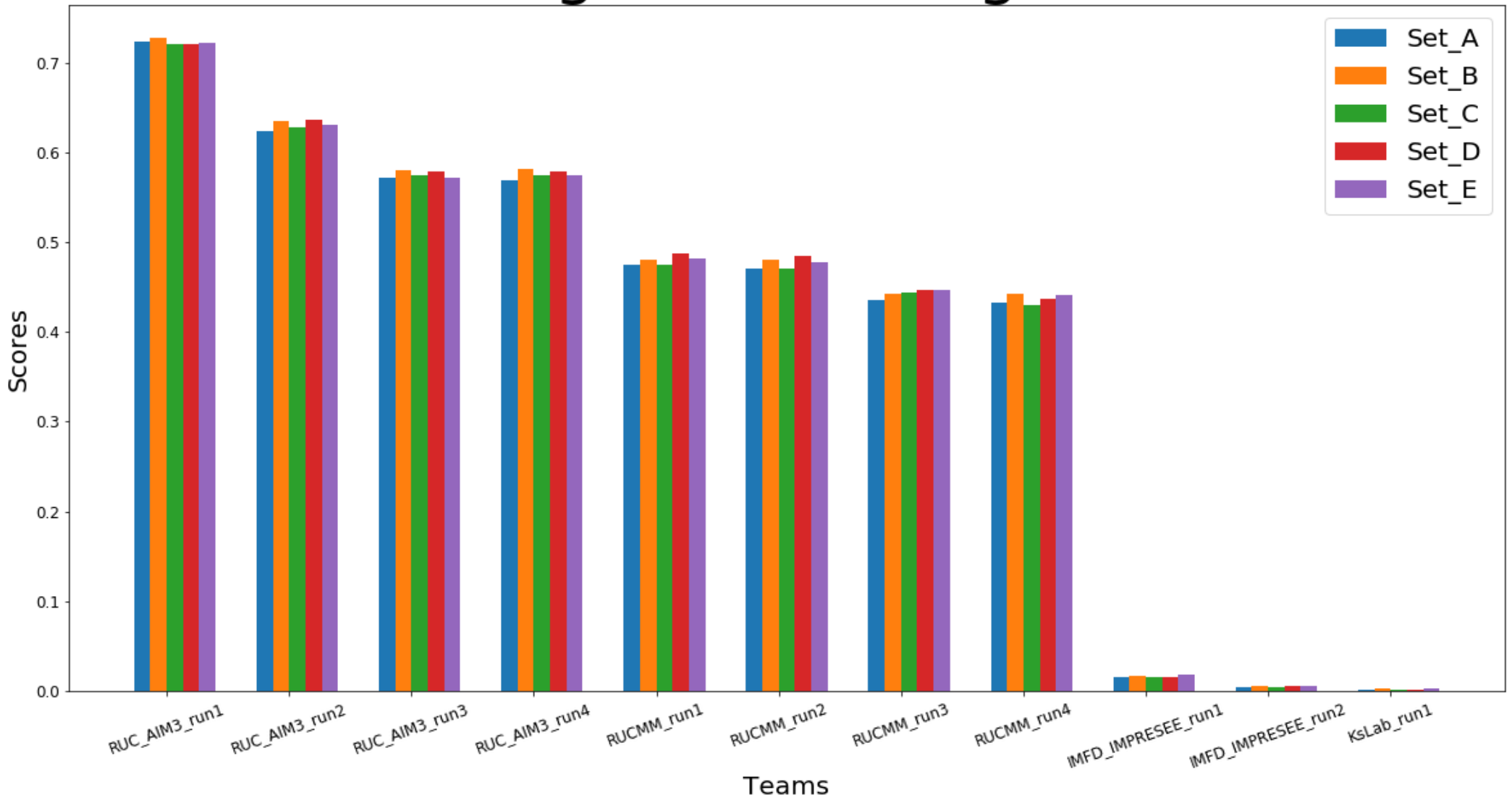
Person playing golf outdoors in the field

Three men running in the street at daytime

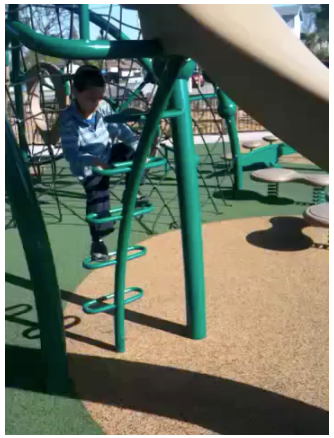
Two men looking at laptop in an office

- Up to 4 runs per site were allowed in the *Matching & Ranking* subtask.
- Mean inverted rank used for evaluation.
- Five sets of descriptions used.

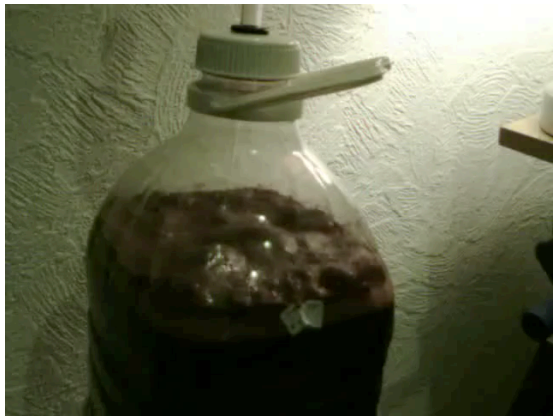
Matching and Ranking Results



Top 3 Results



#13

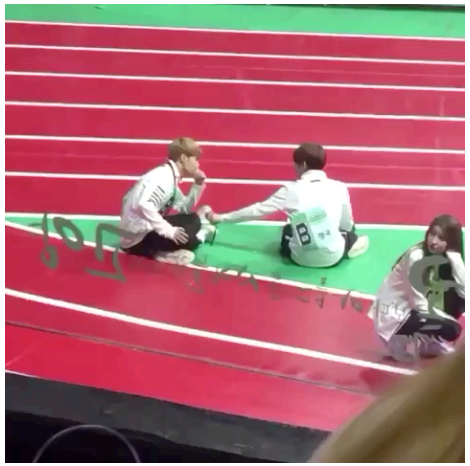


#455

#32

Bottom 3 Results

#1704



#1822

#205

Observations – Matching and Ranking

- 4 teams participated in this optional task.
- The overall mean inverted rank score increased from previous year. Table shows maximum scores for 2018 and 2019.

	2018	2019
Mean Inverted Rank	0.516	0.727

(Very) High Level Overview of Approaches

RUC_AIM3

- Matching & Ranking:
  - Dual encoding module used.
  - Given sequence of input features, 3 branches to encode global, temporal, and local information.
  - Encoded features are then concatenated and mapped into joint embedding space.

RUC_AIM3

- Description Generation:
  - Video Semantic Encoding: Video features extracted in temporal and semantic attention.
  - Description Generation with temporal and semantic attention
  - Reinforcement Learning Optimization: Fine tune captioning model through RL with fluency and visual relevance rewards.
    - Pre-trained language model for fluency.
    - For visual relevance, matching and ranking model used such that embedding vectors should be close in joint space.
  - Ensemble: Various caption modules used. Then relevance used to rerank captions.
- Datasets Used:
  - TGIF, MSR-VTT, VATEX, VTT 2016-17

UTS_ISA

- Framework contains three parts:
  - Extraction of high level visual and action features.
    - Visual features: ResnetXt-WSL, EfficientNet
    - Action + Temporal features: Kinect-i3d features
  - LSTM based encoder-decoder framework to handle fusion and learning. Recurrent neural network used.
  - An expandable ensemble module used. A controllable beam search strategy generates sentences of different lengths.
- Datasets Used:
  - MSVD, MSR-VTT, VTT 2016-18

RUCMM

- Matching & Ranking:
  - Dual encoding used. BERT encoder included to improve dual encoding.
  - Best result by combining models.
- Description Generation:
  - Based on classical encoder-decoder framework.
  - Video-side multi-level encoding branch of dual encoding framework utilized instead of common mean pooling.
- Datasets Used:
  - MSR-VTT, MSVD, TGIF, VTT-16

DCU

- Commonly used BLSTM Network. C3D as input followed by soft attention, which is fed again to a final LSTM.
- A beam search method is used to find the sentences with the highest probability.
- Glove embedding for output words.
- Datasets Used:
  - TGIF, VTT

IMFD-IMPRESSEE

- Matching & Ranking:
  - Deep learning model based on W2VV++ (developed for AVS).
  - Extended by using Dense Trajectories as visual embedding to encode temporal information of the video.
  - K-means clustering to encode Dense Trajectories.
  - Sentence and video embedding into a common vector space.
  - Run without batch normalization performed better than with.

IMFD-IMPRESSEE

- Description Generation:
  - Semantic Compositional Network (SCN) to understand effectively individual semantic concepts for videos.
  - Then a recurrent encoder based on a bidirectional LSTM used.

- Datasets Used:
  - MSR-VTT

FDU

- For visual representation, used Inception-Resnet-V2 CNN pretrained on the ImageNet dataset.
- Concept detection to remove gap between feature representation and text domain.
- LSTM to generate sentences.
- Datasets Used:
  - TGIF, VTT 2017

KSLab, Nagaoka University of Technology

- The goal is to decrease processing time.
- System processes 5 consecutive frames from the beginning and end of the video.
- Each frame converted to 2048 feature vector through Inception V3 Network. Encoder-decoder network is constructed by two LSTM networks.
- No connection observed between video length and score.
- Datasets Used:
  - TGIF, VTT 2016-17

PicSOM and EURECOM

- Combined notebook paper. Tried to answer multiple research questions.
- PicSOM
  - Comparison of cross-entropy and self-critical training loss functions.
    - Self-critical uses CIDER-D scores as reward in reinforcement learning.
    - As expected, self-critical training works better.
  - Use of both still image data and video features improves performance. For still images, video features were non-informative.

PicSOM and EURECOM

- EURECOM
  - Experimented with the use of Curriculum Learning in video captioning.
  - The idea is to present data in an ascending order of difficulty during training.
  - Captions are translated into a list of indices – bigger index for less frequent words.
  - Score of sample is the maximum index of its caption.
  - Video features extracted with an I3D neural network.
  - The process does not seem to be beneficial.
  - Datasets Used:
    - MS-COCO, MSR-VTT, TGIF, MSVD, VTT 2018

Conclusion

- Good number of participation. Task will be renewed.
- This year we used two video sources – Flickr and Vines.
- Each video had 5 annotations.
- Lots of available training sets.
- Multiple research questions on way to solve the VTT task.
- Metric scores for Description Generation and Matching & Ranking have increased over last year.
- A new dataset is in the works – Details to come.

Discussion

- Is there value in the matching and ranking sub-task? Should it be continued as an optional sub-task? Are any teams interested in only this particular sub-task?
- Is the inclusion of run types valuable?
- We may add other popular metrics, such as SPICE. Any suggestions for adding/removing metrics?
- What did individual teams learn?
- Do the participating teams have any suggestions to improve the task?