Video to Text, TRECVID 2018

Korea University (Intelligent Signal Processing Laboratory)
Youngsaeng Jin, Jeonggi Kwak, Younglo Lee, Jeongseop Yun, Hanseok Ko

Content

- Introduction
- Architecture
- Input descriptor
- Experiment
- Result
- Conclusion

Introduction

- Video to text
  - Describe a dynamic visual content with a natural language text

"Two guy are playing baseball."

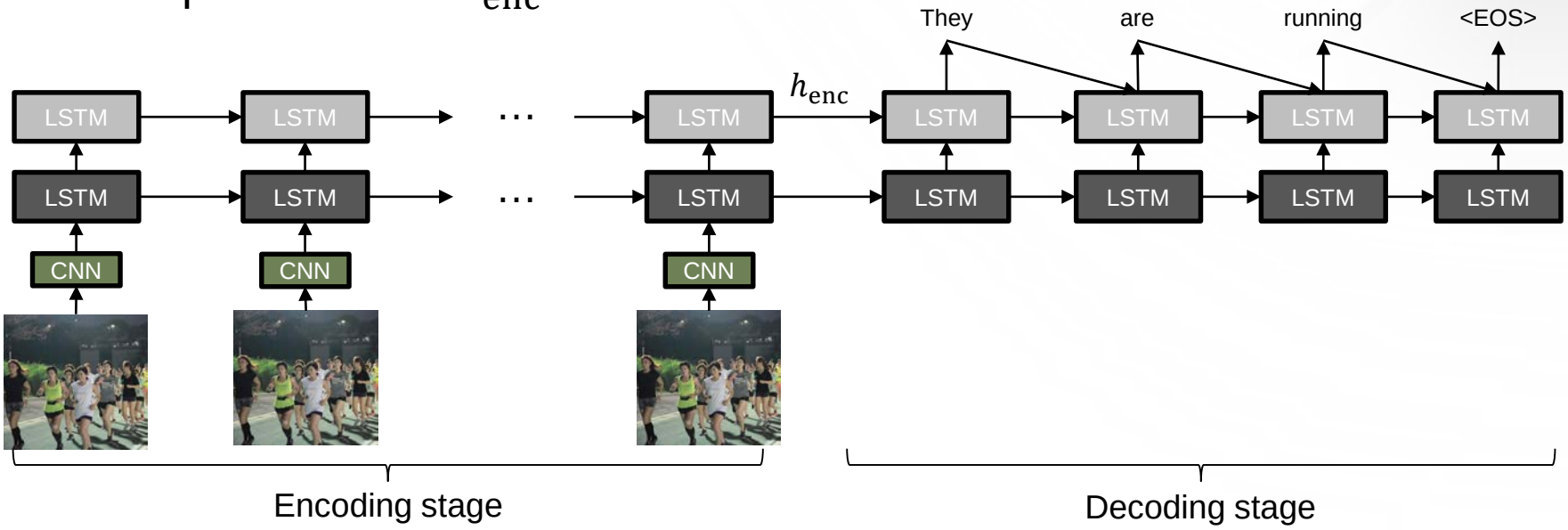
Introduction

- Issues in video to text
  - Scene is too complex for machine to describe
  - Hard to capture all video information
  - Variable length of input and output

"A man helps the other man get up from the ground"

Architecture

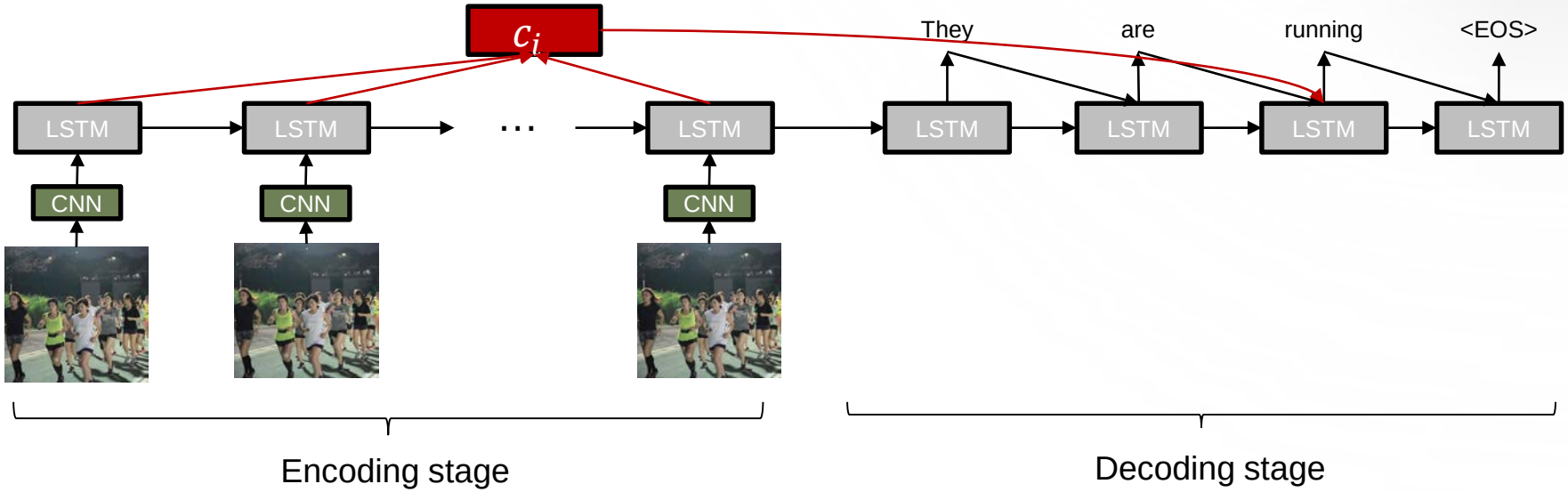
- Sequence to Sequence – Video to Text(S2VT) [1]
  A stack of two LSTM layers
  - Encoding stage : Frames → Visual representation
  - Decoding stage : Visual representation → Words
  - A limitation to represent all information of a video into a fixed-length representation h_{enc}



[1] Venugopalan, Subhashini, et al. "Sequence to sequence-video to text." *Proceedings of the IEEE international conference on computer vision*. 2015.

Architecture

- Attention mechanism [2]
  - Achieve great performance on similar sequence-to-sequence tasks like machine translation
  - Look over all the information included in the input frames
    - Pay more attention to important frames and generate the proper word through a context vector



[2] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Architecture

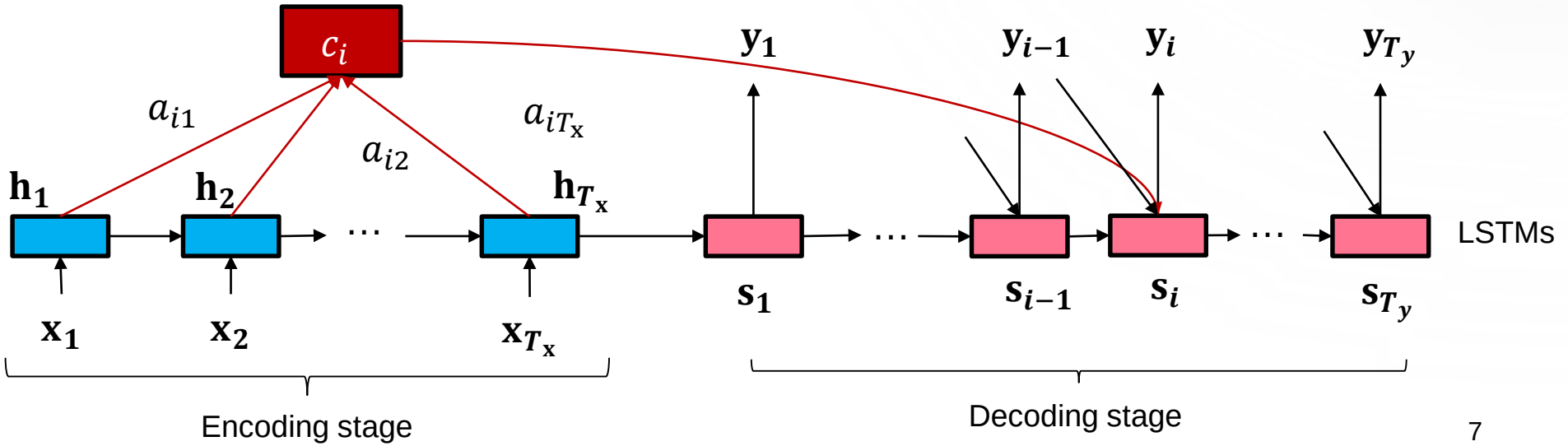
- Attention in encoder-decoder network
  - The decoder attends to difference parts of the encoder information
    - Learn how to generate a context vector c_i instead of a single fixed vector

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j$$

The weight $a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$

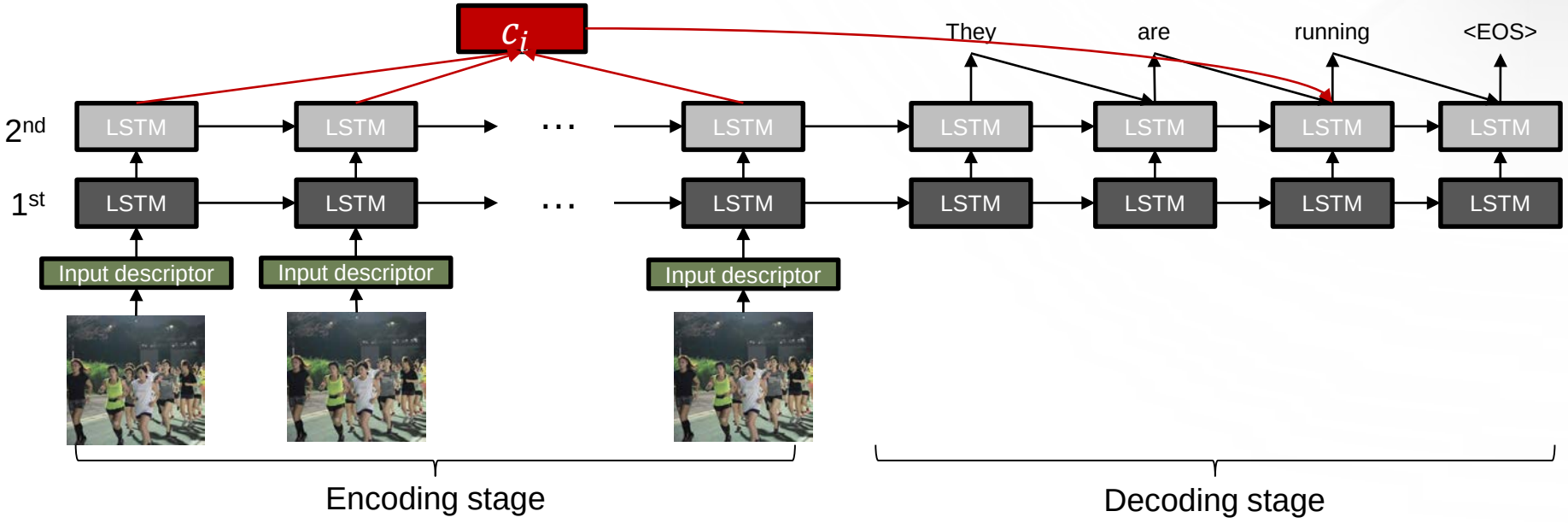
where $e_{ij} = \alpha(s_{i-1}, h_j)$

$\alpha(\cdot)$ is alignment model which is trained



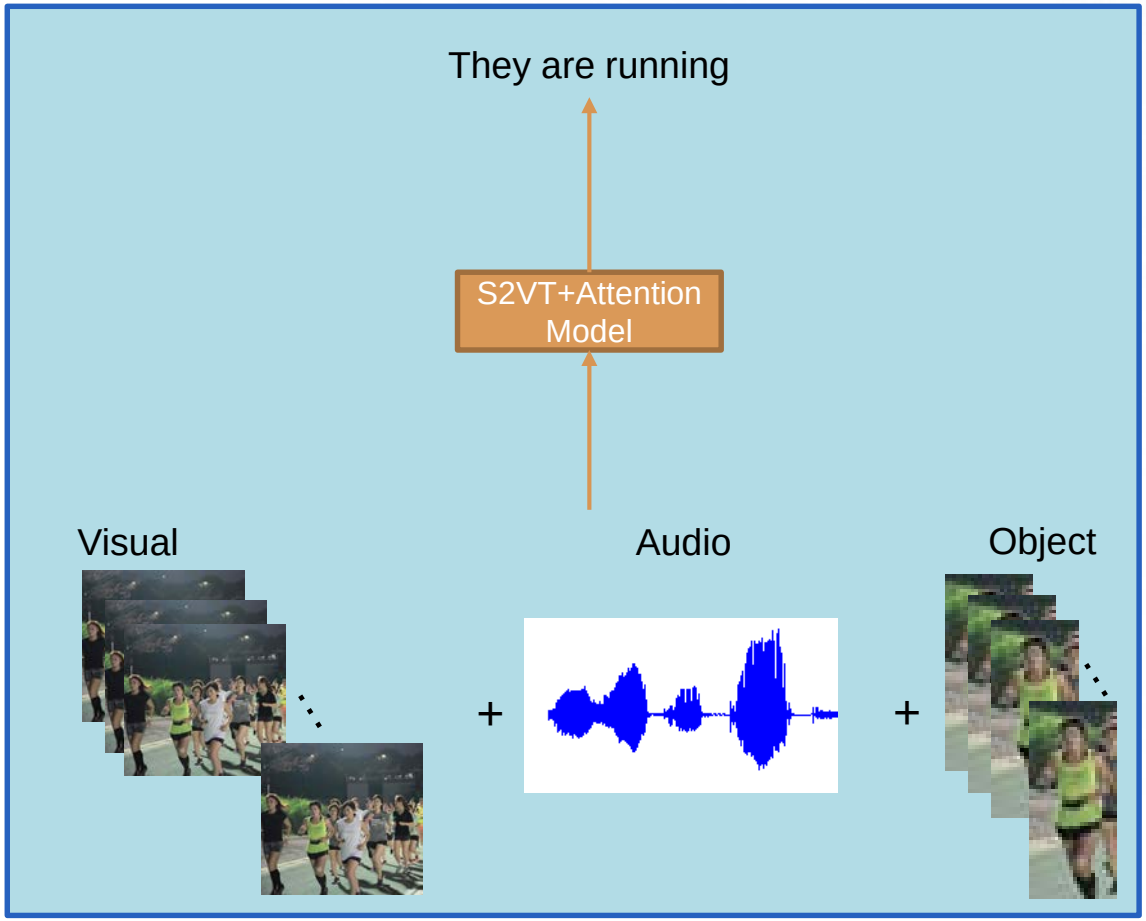
Architecture

- S2VT + Attention
  - A stack of two LSTMs with 1024 hidden units each.
  - The context vector is computed by hidden states of 2nd LSTM layer in encoder
  - The context vector is contributed to 2nd LSTM layer in decoder



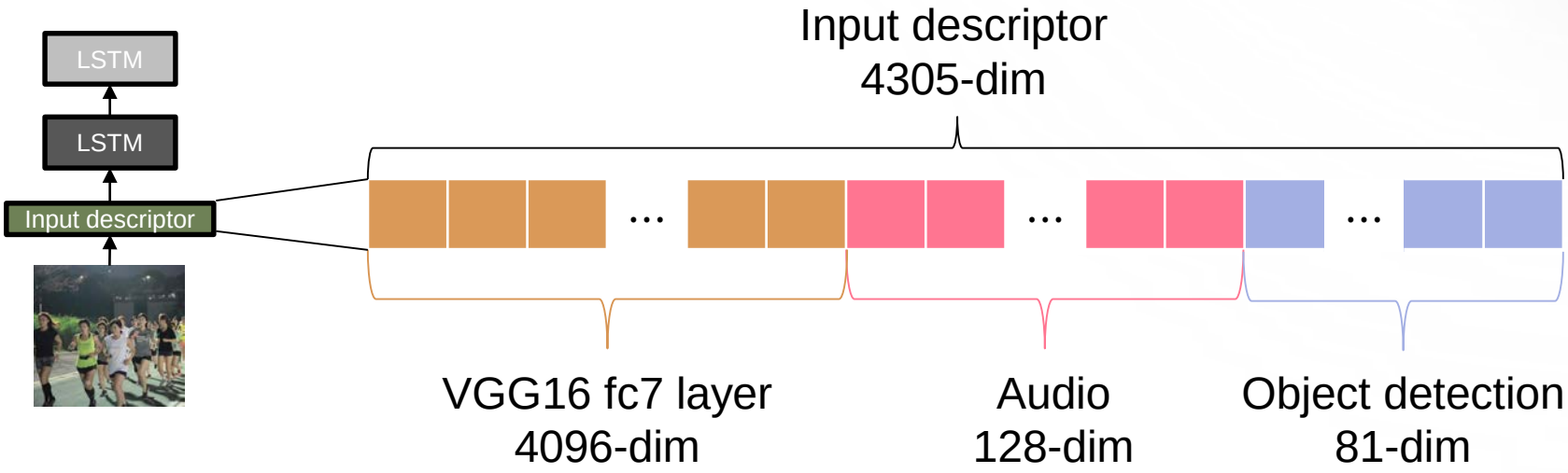
Input Descriptor

- Multimodal fusion strategy
  - Early fusion



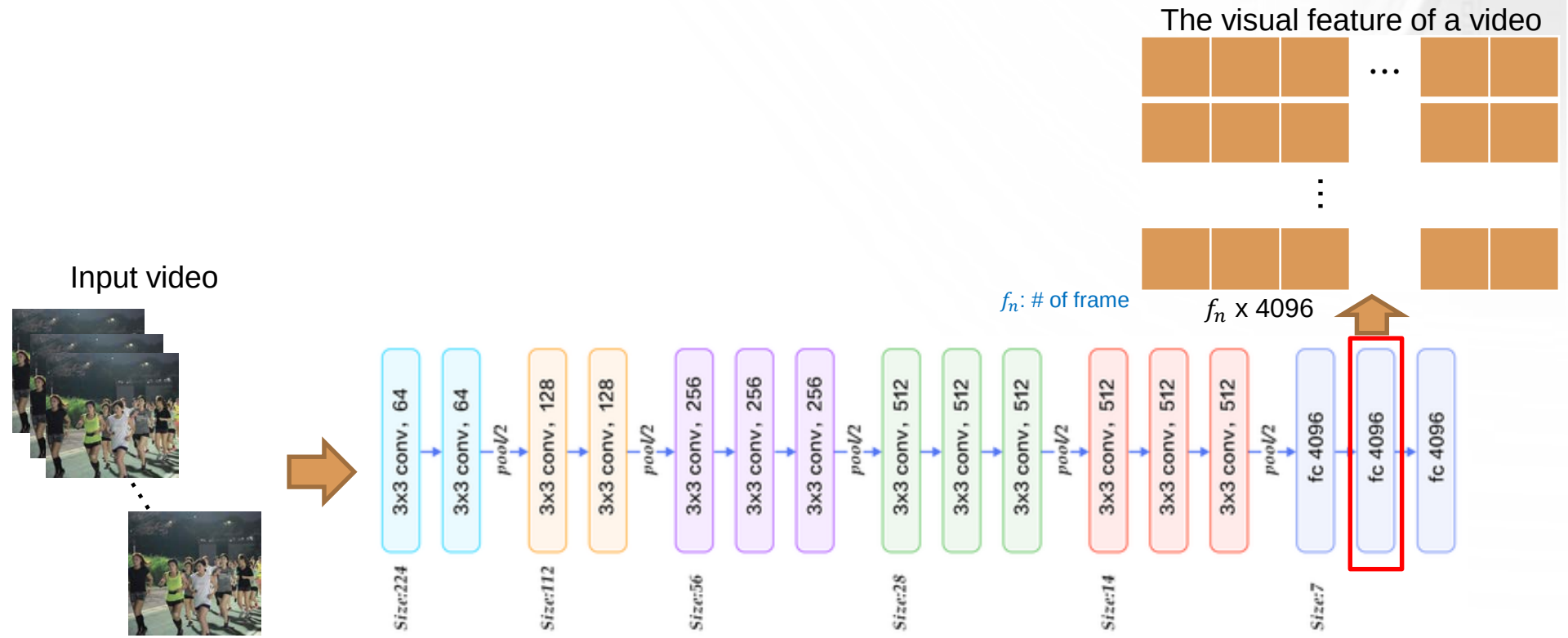
Input Descriptor

- Concatenation of visual, audio, detection information
  - Visual feature (4096 dim)
  - Audio feature (128 dim)
  - Object detection (81 dim)



Input Descriptor

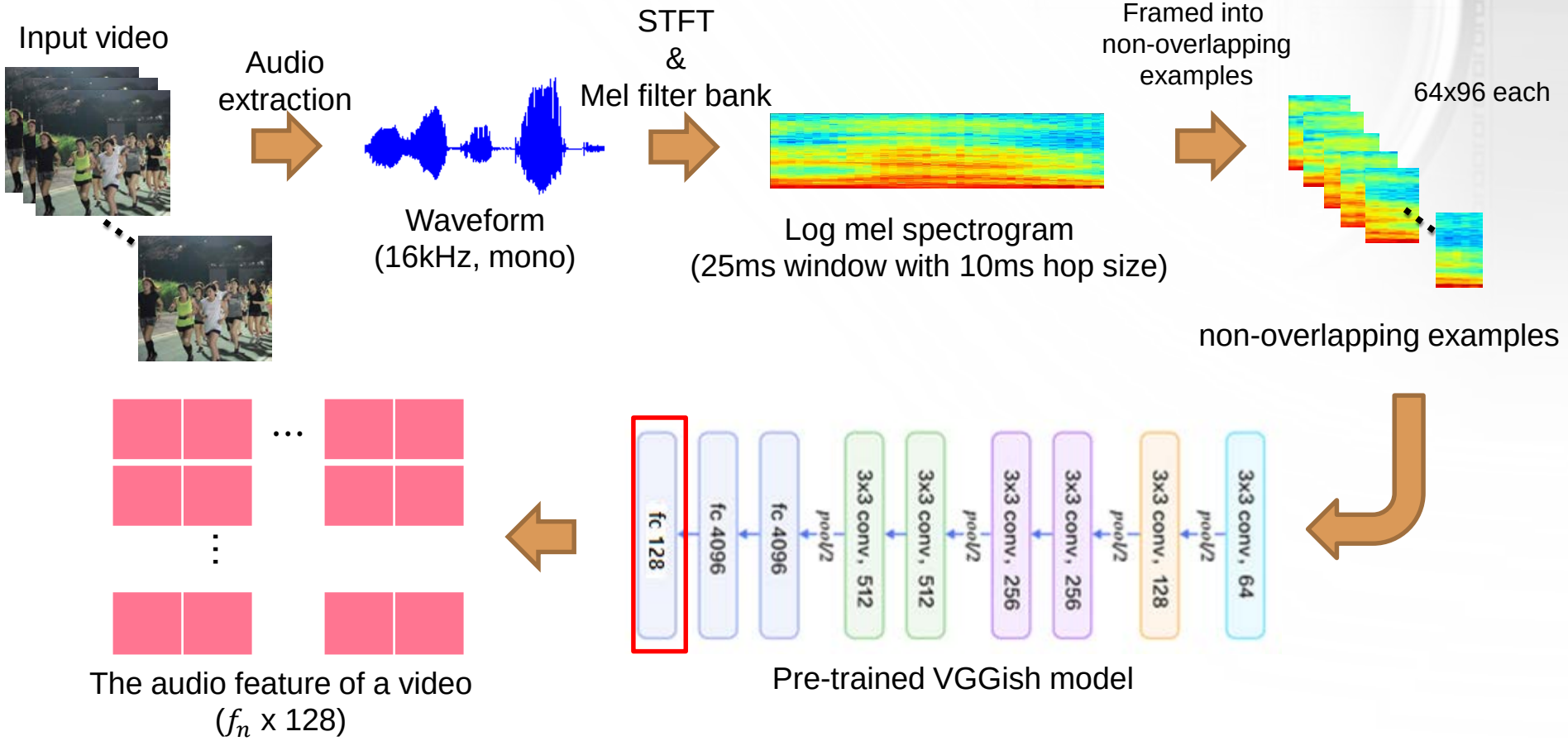
- Visual feature (4096 dim) – VGG-16 [3]
  - A pre-trained CNN is known as an effective visual feature extractor
    - High-dimensional visual features are extracted from fc7 layer of pre-trained VGG-16



[3] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

Input Descriptor

- Audio feature (128dim) – VGGish [4]
  - Extraction schema



[4] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

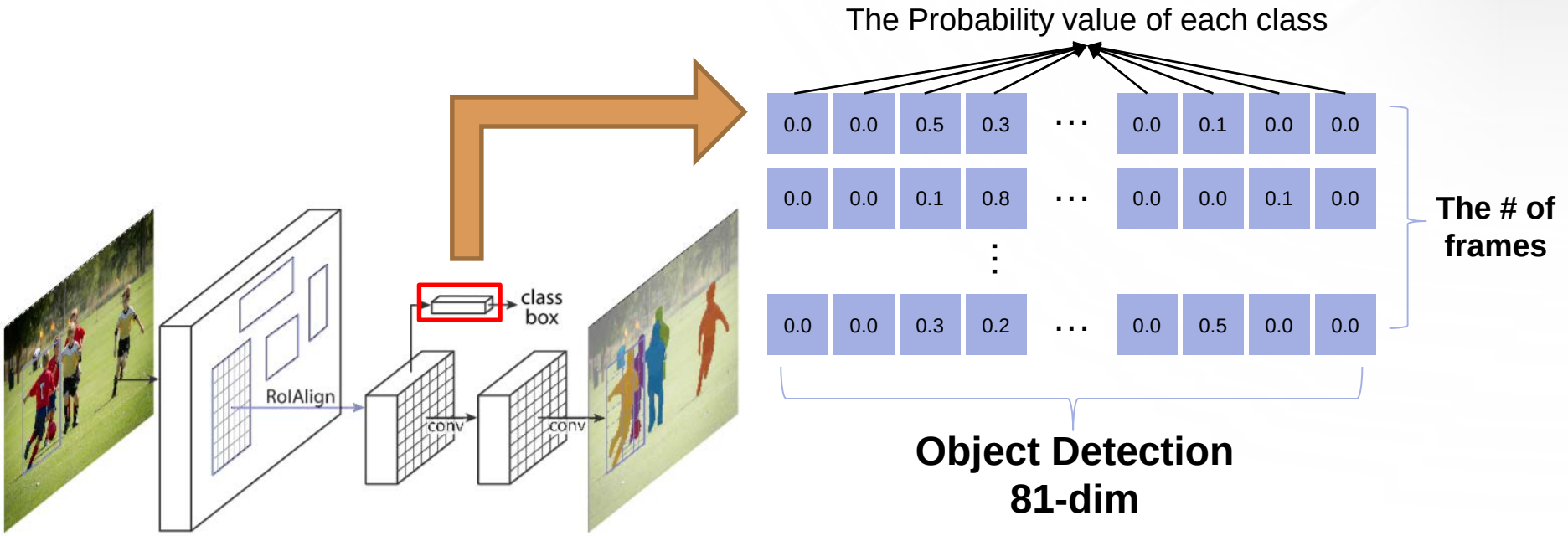
Input Descriptor

- Object detection (81 dim) - Mask R-CNN [5]
  - Maximum probability of each class among all region of interest(RoI)

	c_1	c_2		c_{80}	c_{81}
RoI 1	**0.5**	0.1	…	0.0	0.1
RoI 2	0.0	**0.8**	…	**0.1**	0.1

c_1	c_2		c_{80}	c_{81}
0.5	0.9	…	0.1	0.1

  - Classes : person, vehicle, dog etc



[5] He, Kaiming, et al. "Mask r-cnn." Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.

Technique

- Overfitting
  - Data augmentation
    - Additive random noise on input descriptor
  - Dropout
    - Apply on both LSTM (training phase)

Decision criterion

- Description generation
  - Choose a word with maximum probability at each timestep
  - Concatenate words in all timesteps
- Matching and ranking
  - Comparison : caption vs caption (Generation vs Reference)
  - Measurement metrics : METEOR, BLEU
  - Scoring : 1.2 * METEOR + BLEU

Experiment

- Database
  1. MSVD
  2. M-VAD
  3. MSR-VTT

Table 1. Statistics about 3 dataset in our task.

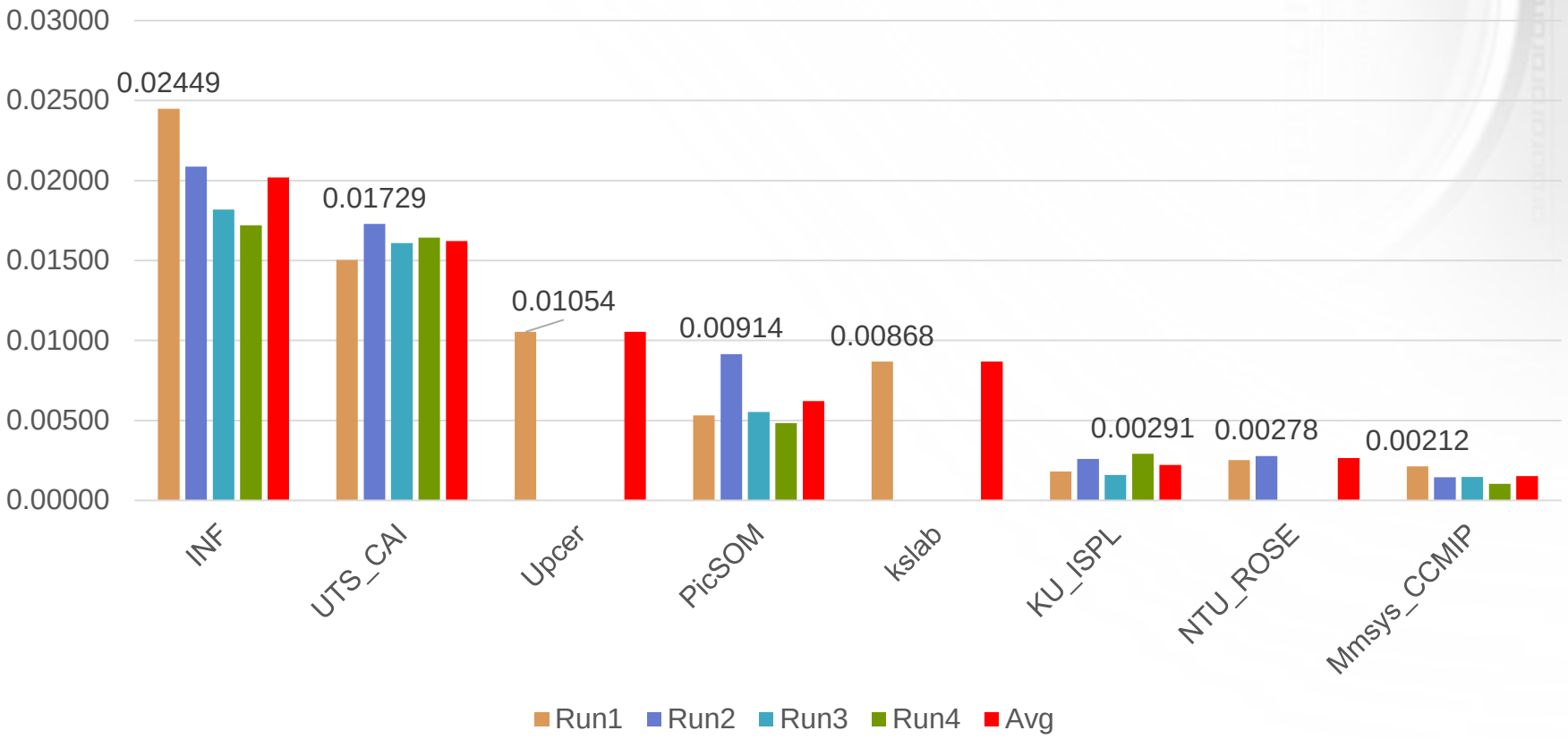
	MSVD	MVAD	MSR-VTT
# video	1,564	4,951	6,074
# description	67,139	4,951	121,021
# avg description	40	1	20
# vocab	12,316	10,984	22,451

Experiment

- Run types
  - Run 1: VGG (no attention)
  - **Run 2: VGG** (Primary)
  - Run 3: VGG + Sound
  - Run 4: VGG + Sound + Detection

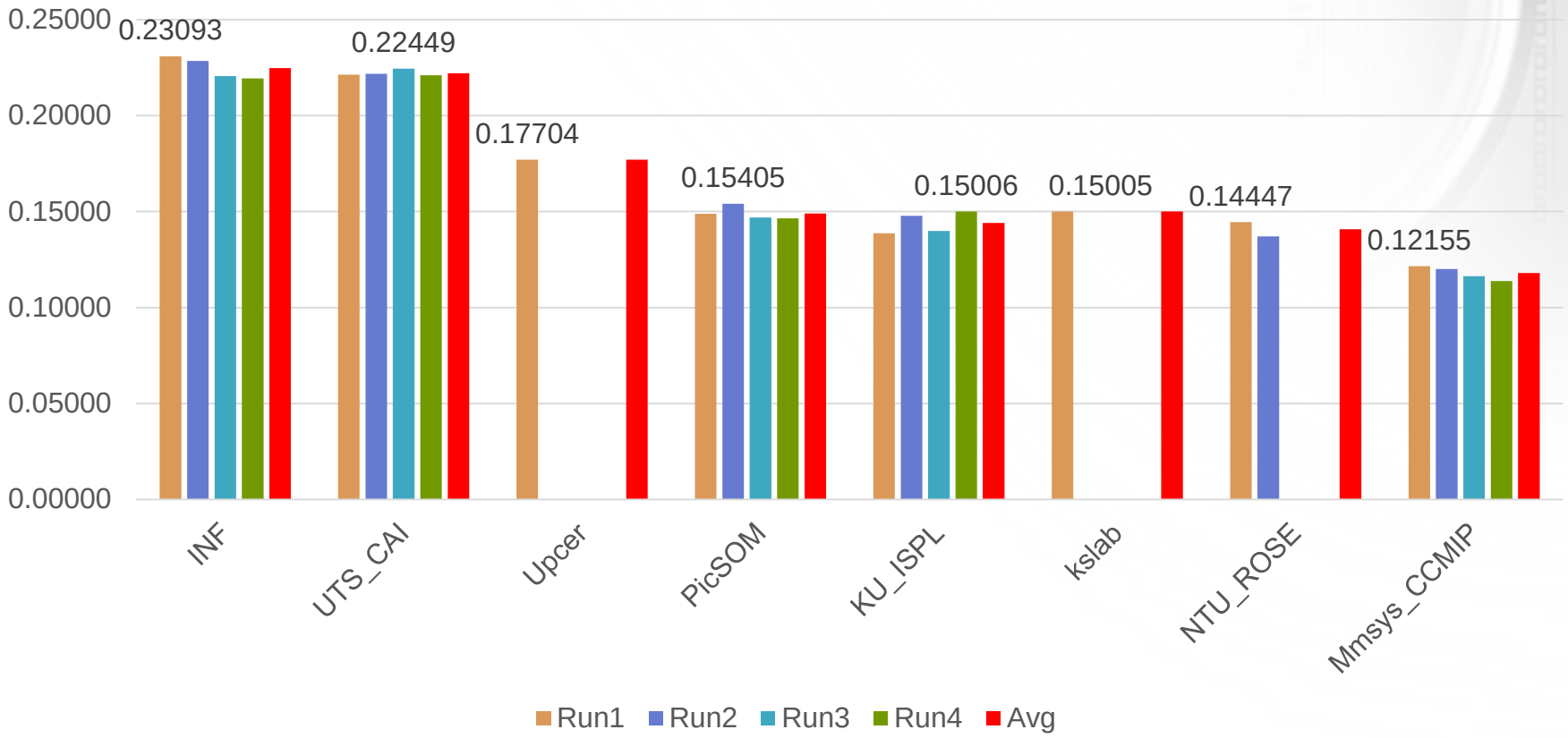
Results

- Generation – BLEU : 6 / 8



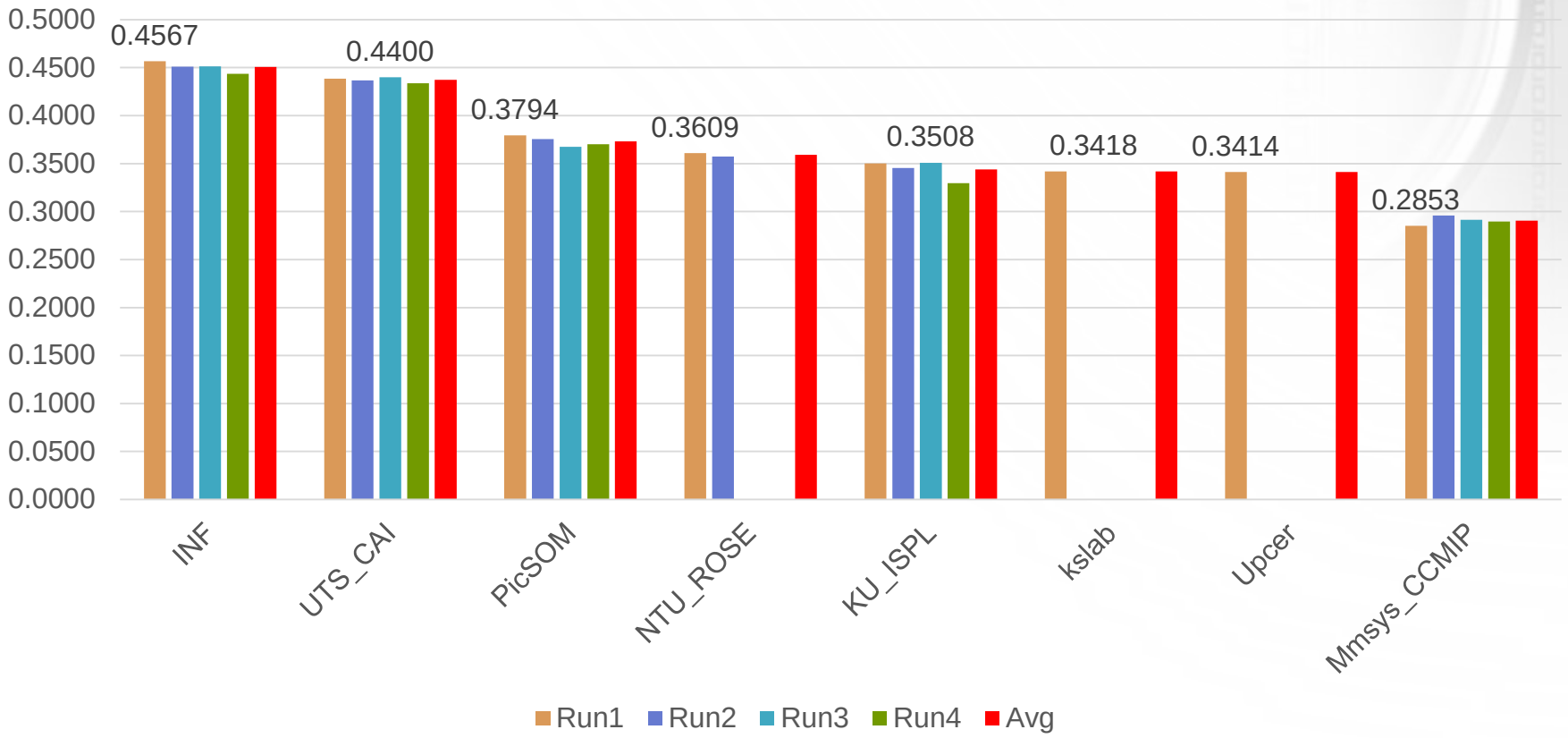
Results

- Generation – Meteor: 5 / 8



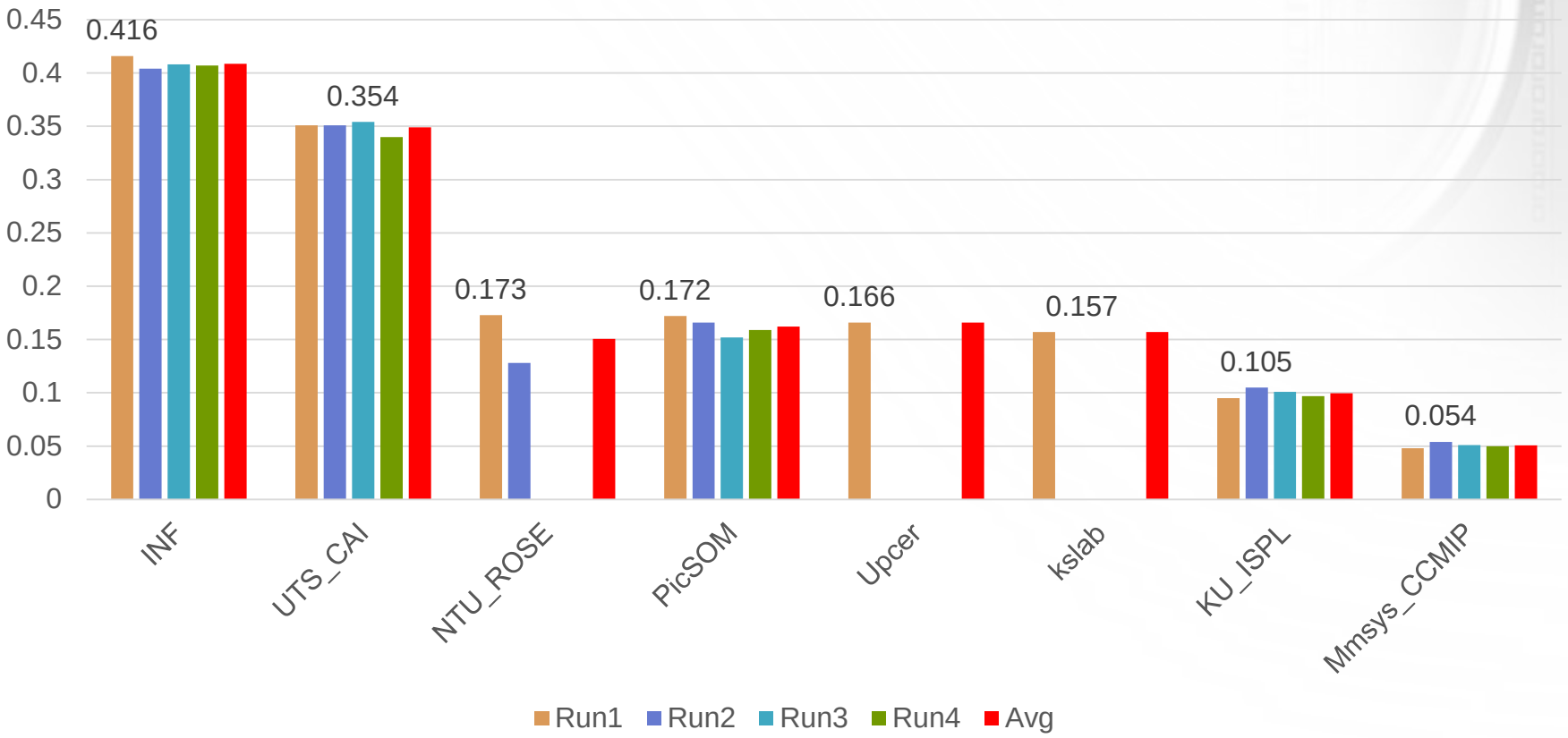
Results

- Generation – STS: 5 / 8



Results

- Generation – CIDEr: 7 / 8



Results

- Quantitative performance

Run	BLEU	METEOR	CIDEr	STS
1	0.00181	0.13876	0.095	0.3402
2	0.00259	0.14782	0.105	0.3454
3	0.00159	0.13998	0.101	0.3508
4	0.00291	0.15006	0.097	0.3298

Results

- Example 1

Run	Description
1	A baby is playing with
2	A man is playing with baby
3	A woman is holding a baby while two men are sitting on the couch playing with a baby
4	Someone his to the floor
GT	A small black girl dressed in a white top and black skirt making motions with her hands and fingers walking among many adults sitting down

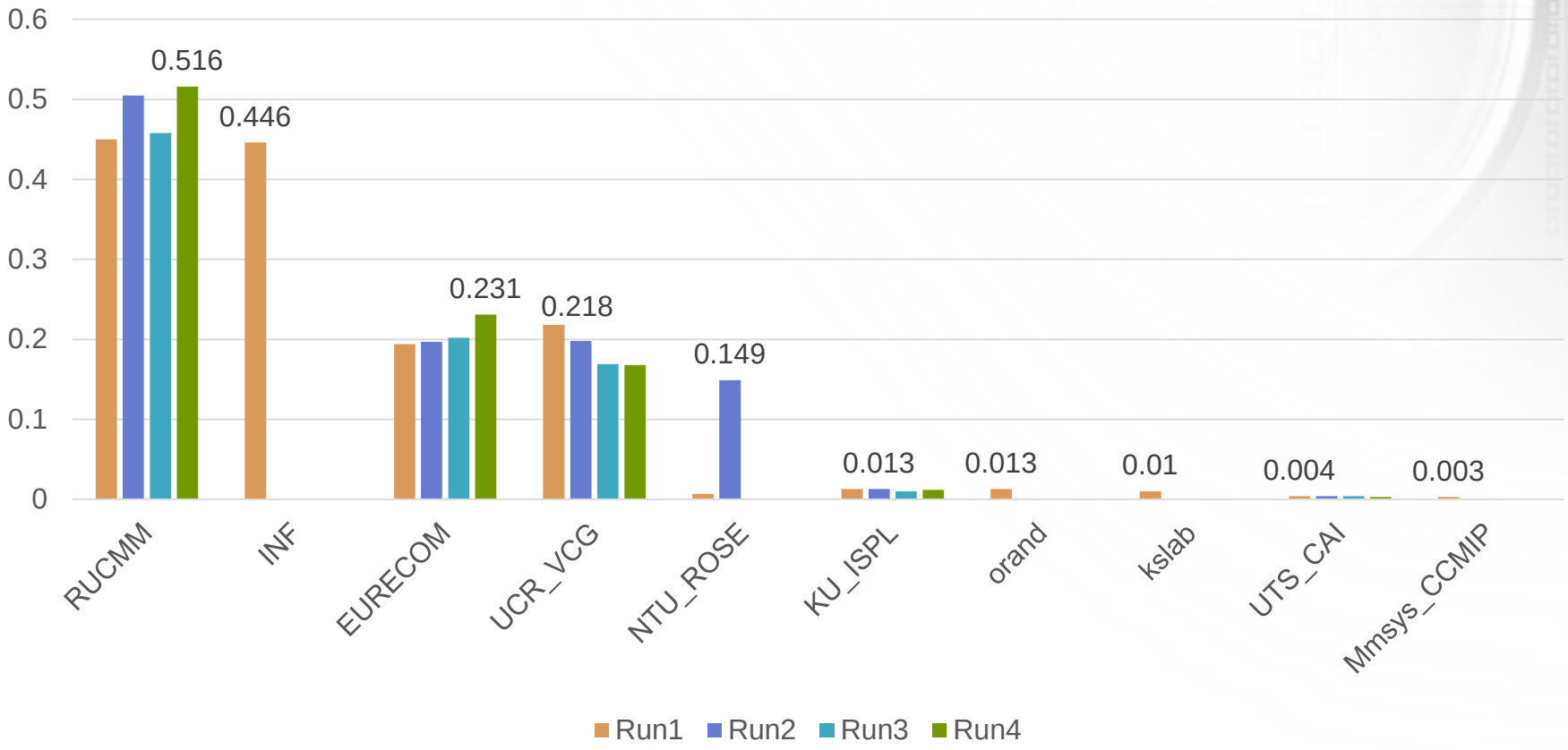
Results

- Example 2

Run	Description
1	A man is dancing
2	A man is sitting on a bench and playing with the dog
3	A man is standing by a horse
4	A girl is waiting in a pink bag
GT	A young man, dressed with an Abaya and Arabic outfit, is singing, to a donkey that is moving its head right and left, outdoors.

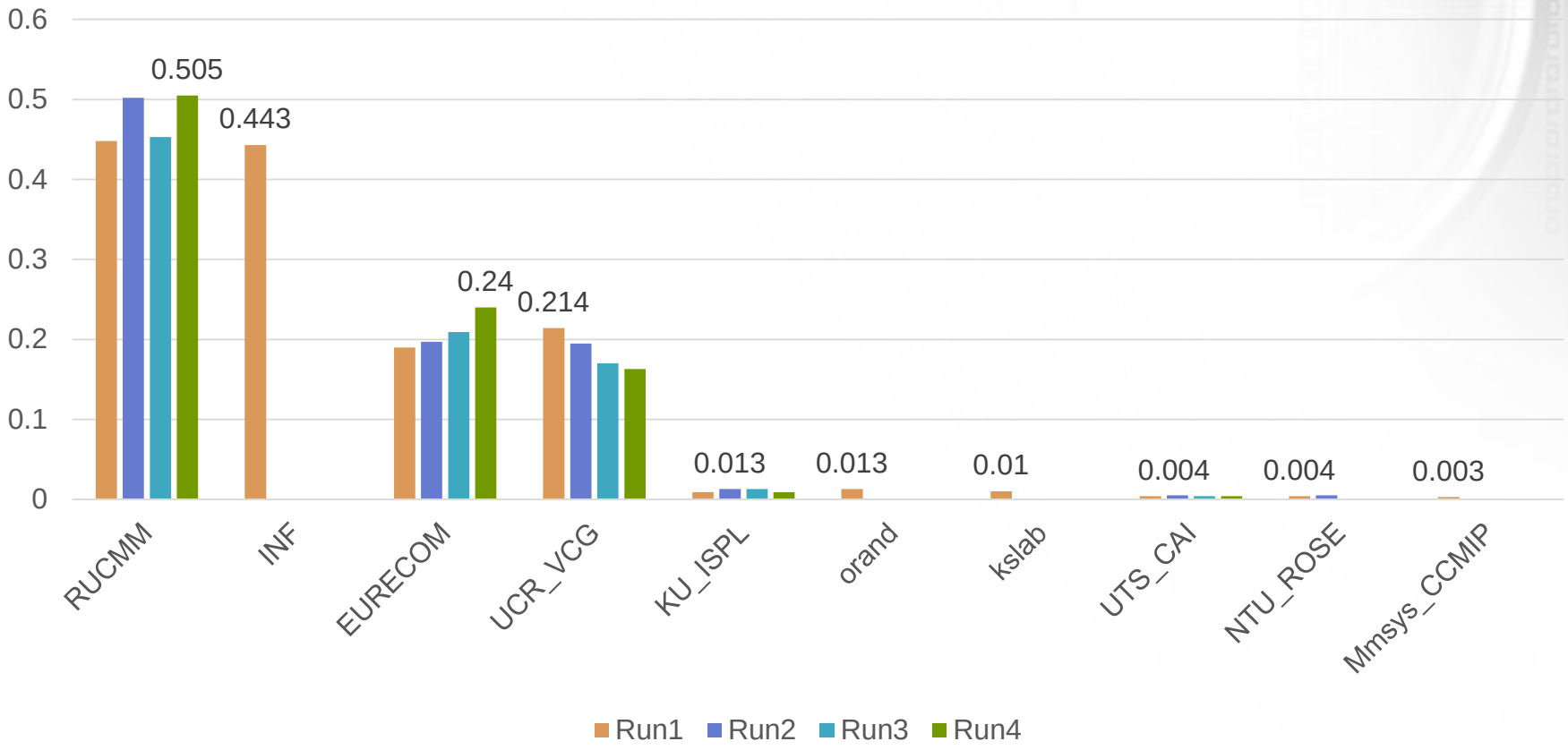
Results

- Matching & Ranking – Set A: 6 / 10



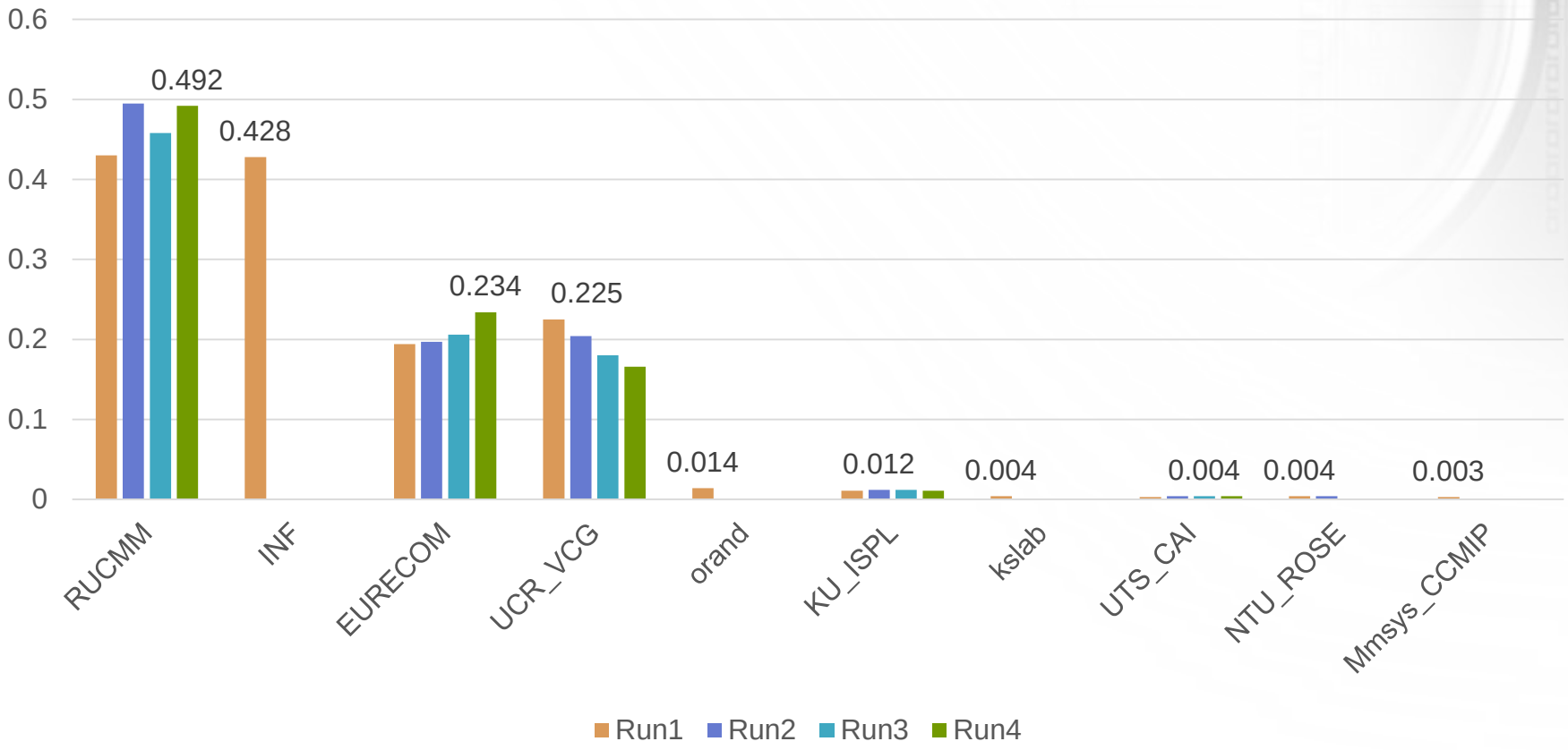
Results

- Matching & Ranking– Set B: 5 / 10



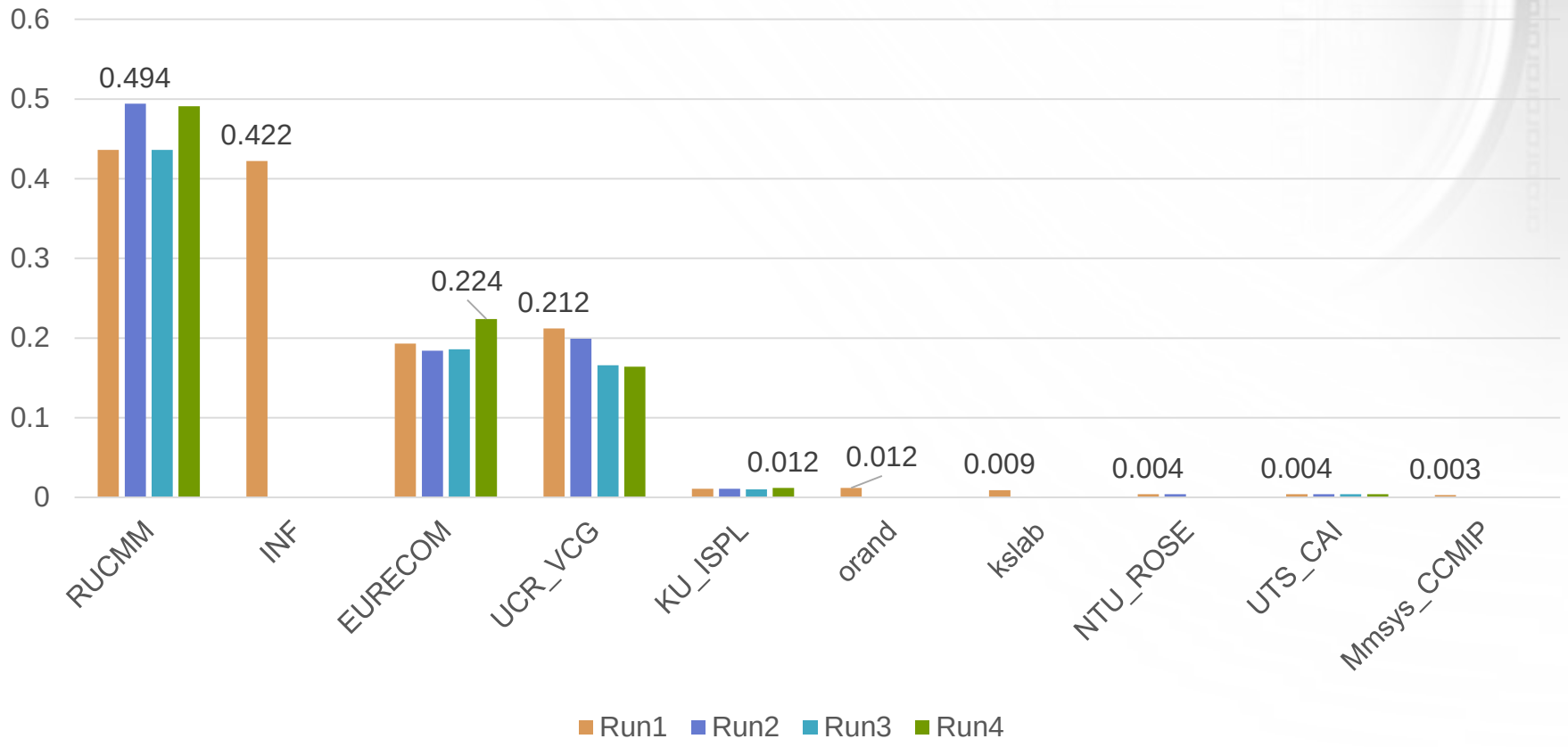
Results

- Matching & Ranking– Set C: 6 / 10



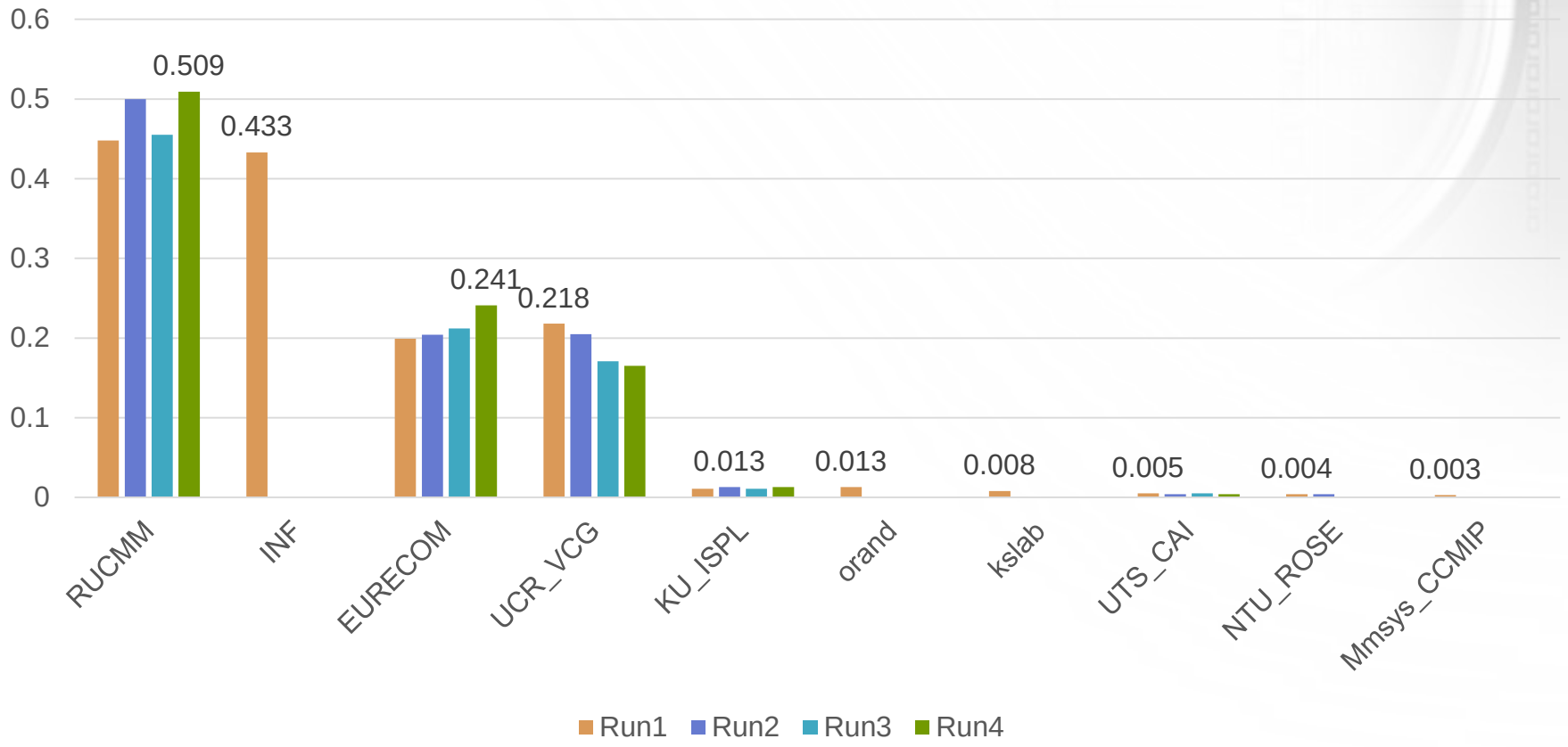
Results

- Matching & Ranking – Set D: 5/ 10



Results

- Matching & Ranking – Set E:5/ 10



Conclusion

- Attention-based sequence-to-sequence model was effective to accurate description generation of video clips.
- Various combinations of visual feature, acoustic feature and detection result are used for an input descriptor of the model.
  - The performance with additional features didn't show great improvement
- The performance is improved when attention mechanism is exploited.

Thank You!

Any questions?