Renmin University of China and Zhejiang Gongshang University at TRECVID 2018: Deep Cross-Modal Embeddings for Video-Text Retrieval

Xirong Li[†], Jianfeng Dong[*], Chaoxi Xu[†], Jing Cao[†], Xun Wang[*], Gang Yang[†]

[†]AI & Media Computing Lab
Renmin University of China
Beijing, China

[*]School of Computer and Information Engineering
Zhejiang Gongshang University
Hangzhou, China

Abstract

In this paper we summarize our TRECVID 2018 [1] video retrieval experiments. We participated in two tasks: Ad-hoc Video Search (AVS) and Video-to-Text (VTT) Matching and Ranking. For the AVS task, we develop our solutions based on W2VV++, a super version of Word2VisualVec (W2VV) [4]. For the VTT task, our entry is built on the top of a recently proposed dual encoding network [5], which encodes an input, let it be a video or a natural language sentence, at multiple levels. The 2018 edition of the TRECVID benchmark has been a fruitful participation for our joint-team, resulting in the best overall result for both AVS and VTT tasks. Retrospective experiments show that our ad-hoc video search system, used as is, also outperforms the previous best results of the TRECVID 2016 and 2017 AVS tasks. We have released feature data at https://github.com/li-xirong/avs.

1 Ad-hoc Video Search

1.1 Approach

For the ad-hoc video search task, we develop a super vision of Word2VisualVec (W2VV) [4], which we term W2VV++. The original W2VV model is a deep neural network that projects a given sentence into a visual feature space by first vectorizing the sentence by a multi-scale encoding strategy. Then, the encoding result goes through a multilayer perceptron (MLP) to produce a feature vector $r(s)$. The network is trained such that the loss between a given video v and the sentence s, defined as the Mean Square Error (MSE) between $r(s)$ and the video feature $\phi(v)$, is minimized. For cross-modal matching the cosine similarity between $\phi(v)$ and $r(s)$ is used, denoted by $S_\theta(v,s)$, where θ indicates all the trainable parameters in the model. While W2VV is shown to be effective in the 2016 and 2017 TRECVID video-to-text matching tasks [11, 12], the MSE based loss limits its ability to exploit many negative samples during the training stage.

W2VV++ improves over W2VV by substituting an improved marginal ranking loss [7] for the MSE loss. Given a video-sentence pair (v,s), the new loss is defined as:

$$l(v,s;\theta) = \max_{v}(0, \alpha + S_\theta(v^-,s) - S_\theta(v,s)), \qquad (1)$$

where v^- is a hardest negative video sample of the sentence s. Following [7], we define the hardest negative example as the most similar sample to s in a mini-batch. As illustrated in Fig. 1, we investigate three variants of W2VV++, depending on *1)* the choice of video features and *2)* whether an extra fully connected layer is added for video *feature relearning* [6].

For video representation, we uniformly sample frames with an interval of 0.5 second. Deep visual features are extracted per frame by pre-trained CNN models. In particular, we adopt a ResNet-152 model used in [3] and a ResNeXt-101 model used in [12]. Accordingly, two 2,048-dim video-level features are obtained by mean pooling over the frames.

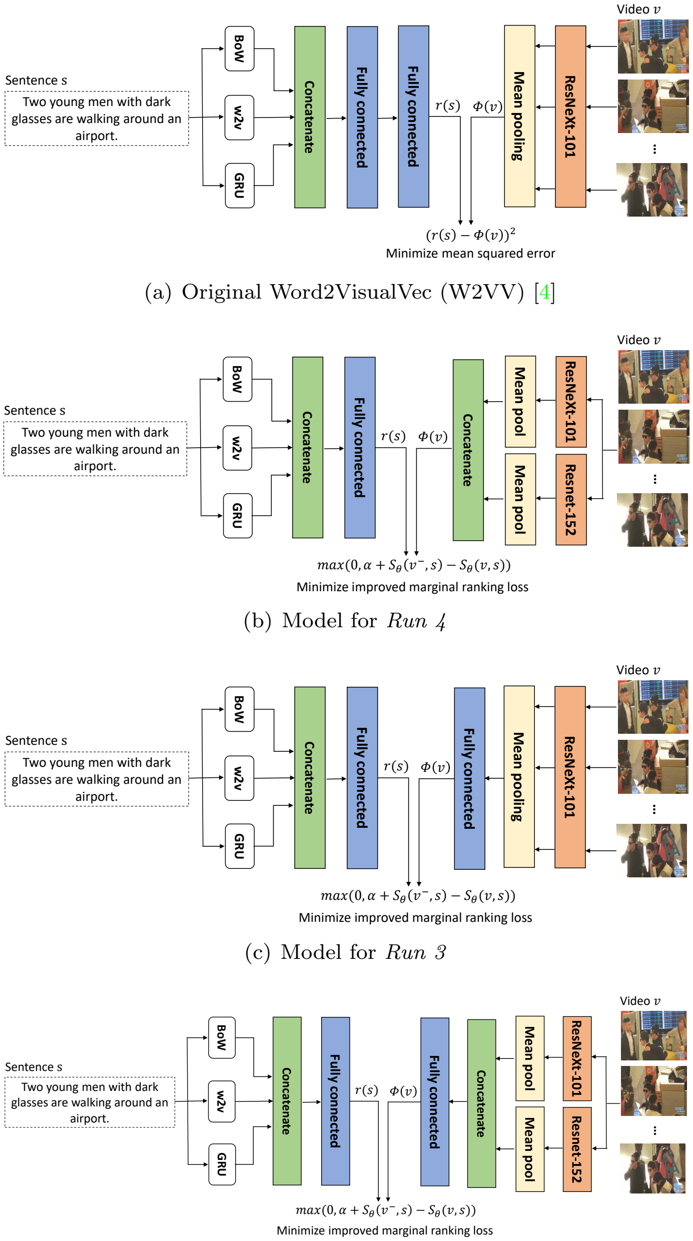
For sentence representation, we keep the sentence encoder of W2VV. That is, a given sentence is firstly vectorized in parallel by three vectorization strategies including Bag-of-Words (BoW), word2vec and a Gated Recurrent Units (GRU). The output of the three encoding blocks is concatenated and forwarded to a fully connected layer for common space learning.

We train W2VV++ on a joint collection of MSR-VTT [8] and TGIF [10], with hyper-parameters tuned on the training sets of the previous TRECVID VTT task.

1.2 Submissions

We submit the following runs:

- *Run 4* is W2VV++ that predict the ResNeXt-101 + ResNet-152 feature for a given sentence. The cross-modal similarity between the given sentence and any video from the IACC.3 collection is implemented as the cosine similarity between their corresponding feature vectors.
- *Run 3* differs *Run 4* in two ways. First, the former uses only the ResNeXt-101 feature. Second, it adds a fully connected layer on the top of the video feature.


Video v
Sentence s
Two young men with dark glasses are walking around an airport.
BoW
w2v
GRU
Concatenate
Fully connected
Fully connected
$r(s)$
$\Phi(v)$
Mean pooling
ResNeXt-101
$(r(s)-\Phi(v))^2$
Minimize mean squared error
Video v
Sentence s
Two young men with dark glasses are walking around an airport.
BoW
w2v
GRU
Concatenate
Fully connected
$r(s)$
$\Phi(v)$
Concatenate
Mean pool
Mean pool
ResNeXt-101
Resnet-152
$max(0, \alpha + S_\theta(v^-, s) - S_\theta(v, s))$
Minimize improved marginal ranking loss
Video v
Sentence s
Two young men with dark glasses are walking around an airport.
BoW
w2v
GRU
Concatenate
Fully connected
$r(s)$
$\Phi(v)$
Fully connected
Mean pooling
ResNeXt-101
$max(0, \alpha + S_\theta(v^-, s) - S_\theta(v, s))$
Minimize improved marginal ranking loss
Video v
Sentence s
Two young men with dark glasses are walking around an airport.
BoW
w2v
GRU
Concatenate
Fully connected
$r(s)$
$\Phi(v)$
Fully connected
Concatenate
Mean pool
Mean pool
ResNeXt-101
Resnet-152
$max(0, \alpha + S_\theta(v^-, s) - S_\theta(v, s))$
Minimize improved marginal ranking loss

(a) Original Word2VisualVec (W2VV) [4]

(b) Model for *Run 4*

(c) Model for *Run 3*

(d) Model for *Run 2*

Figure 1: Conceptual diagrams of W2VV++ models used in our runs. Our best run, *i.e. Run 1*, equally combine the W2VV++ models used in the other runs.

- *Run 2* is similar to *Run 3* but re-using the ResNeXt-101 + ResNet-152 feature.
- *Run 1* equally combines models from all the other runs and trained with different setups.

An overview of the AVS task benchmark is shown in Fig. 2. *Run 4* servers as our baseline, better than all submissions from the other teams. The result shows the effectiveness of W2VV++. *Run 2*, by adding an additional feature re-learning layer, outperforms *Run 4*. While the improvement appears to be marginal, their ensemble, as demonstrated by *Run 1*, gives a noticeable performance boost. The result indicates that these single models are complementary to each other.

Table 1: Retrospective experiments on the AVS tasks of the previous years. Our runs outperform the previous best runs.

	TRECVID edition		
	2016	*2017*	*2018*
Previous best run	0.054 [9]	0.206 [12]	–
Ours:			
Run 4	0.149	0.176	0.104
Run 3	0.140	0.171	0.103
Run 2	**0.151**	0.213	0.106
Run 1	0.149	**0.220**	**0.121**

A retrospective experiment on the AVS tasks of the previous years is reported in Table 1. Our models, used as is, outperform the previous best runs. The results again confirm the effectiveness of W2VV++. Moreover, considering that the video pool stays the same while the queries change each year, the retrospective experiment suggests that the 2018 topics are the most difficult, while the 2017 topics seem to be the easiest.

2 Video to Text Matching

For video-to-text matching, we participate in the Matching and Ranking subtask. Given a video, participants were asked to rank a list of pre-defined candidate sentences in terms of their relevance with respect to the given video. In the 2018 edition, the test video set consists of 1,904 videos collected from Twitter Vine. Five sentence sets are provided by the task organizers, denoted as setA, setB, setC, setD and setE. Each sentence set has 1,921 sentences.

2.1 Approach

Our entry is built on the top of a recently proposed dual encoding network [5]. The dual encoding network uses the same architecture to learn powerful representations for two sequential input of distinct modalities, *i.e.* video as a sequence of frames and sentence as a sequence of words, at multiple levels. In particular, the encoding network consists of three encoding blocks that are implemented by mean pooling, bidirectional GRU (biGRU) and biGRU-CNN respectively. The three blocks are stacked to explicitly model global, local and temporal patterns in both videos and sentences. The output of a specific encoding block is not only used as input of a follow-up encoding block, but also re-used via skip connection to contribute to the final output of the entire encoding network. It generates new, higher-level features progressively. These features, generated at distinct levels, are powerful and complementary to each other. So we

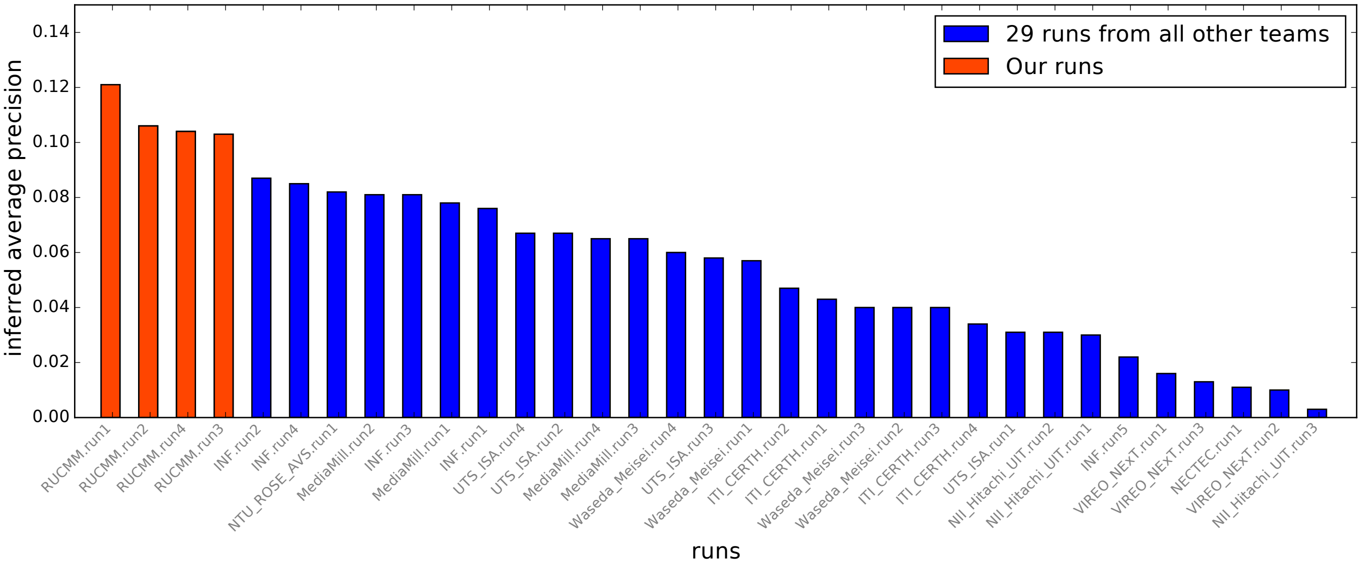


Figure 2: Overview of the TRECVID 2018 ad-hoc video search task benchmark, all runs ranked according to mean infAP. We were the best overall performer.

combine them to obtain robust video / sentence representations by a simple concatenation operation. After videos and sentences being encoded by the dual encoding network, we employ a state-of-the-art common space learning algorithm [7] to project the two modalities into a common space. Consequently, the relevance between a given video v and a sentence s is computed as the cosine similarity between the video feature $f(v)$ and the sentence feature $f(s)$ in the common space. For more technical details we refer readers of interest to [5].

The model is trained on a combined set of MSR-VTT [8], MSVD [2] and TGIF [10], with hyper-parameters tuned on the TRECVID 2016 VTT training set.

2.2 Submissions

We submit the following four runs:

- *Run 0* is the dual encoding model using the ResNeXt-101 feature.
- *Run 1* equally combines either models, among which four models are based on *Run 0* with their last FC layer varies. That is, a FC layer, a FC layer with a tanh activation, a FC layer with a BN layer, and a FC layer with a BN layer and a tanh activation. The other four models are trained in a similar manner, but using the ResNext-152 feature.
- *Run 2* equally combines eight W2VV++ models. Four models are separately trained using the ResNeXt-101 feature, with sentence vectorization varies. That is, BoW, GRU using the last output, GRU using the mean of all outputs, and multi-scale sentence vectorization, respectively. The other four models are trained in a similar manner, but using the ResNeXt-101 + ResNet-152 feature.

Table 2: Our runs in the TRECVID 2018 VTT matching and ranking task.

Ours	setA	setB	setC	setD	setE
Run 0	0.450	0.448	0.430	0.436	0.448
Run 1	0.505	0.502	**0.495**	**0.494**	0.500
Run 2	0.458	0.453	0.448	0.436	0.455
Run 3	**0.516**	**0.505**	0.492	0.491	**0.509**

- *Run 3* combines run 1, run 2 and eight VSE++ models [7]. Besides the original VSE++, we train multiple variants, including 1) VSE++ with two FC layers and 2) substituting BoW for GRU. Four models use the ResNeXt-101 feature, while the other four models use the ResNet-152 feature.

An overview of all submissions is shown in Fig. 3 and Fig. 4. Concrete numbers are summarized in Table 2. The leading position of our runs clearly demonstrates the effectiveness of the dual encoding network.

Acknowledgments

The authors are grateful to NIST and the TRECVID coordinators for the benchmark organization effort. The Renmin University of China acknowledges support by NSFC (No. 61672523, No. 61773385), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 18XNLG19). The Zhejiang Gongshang University acknowledges support by NSFC (No.

U1609215, No. 61672460), the Zhejiang Provincial Natural Science Foundation of China (No. Q19F020013).

References

[1] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, J. Magalhaes, D. Semedo, and S. Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018.

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011.

[3] J. Dong, S. Huang, D. Xu, and D. Tao. Dl-61-86 at TRECVID 2017: Video-to-text description. In *TRECVID Workshop*, 2017.

[4] J. Dong, X. Li, and C. G. M. Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 2018.

[5] J. Dong, X. Li, C. Xu, S. Ji, Y. He, G. Yang, and X. Wang. Dual encoding for zero-example video retrieval. In *CVPR*, 2019.

[6] J. Dong, X. Li, C. Xu, G. Yang, and X. Wang. Feature relearning with data augmentation for content-based video recommendation. In *ACMMM*, 2018.

[7] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

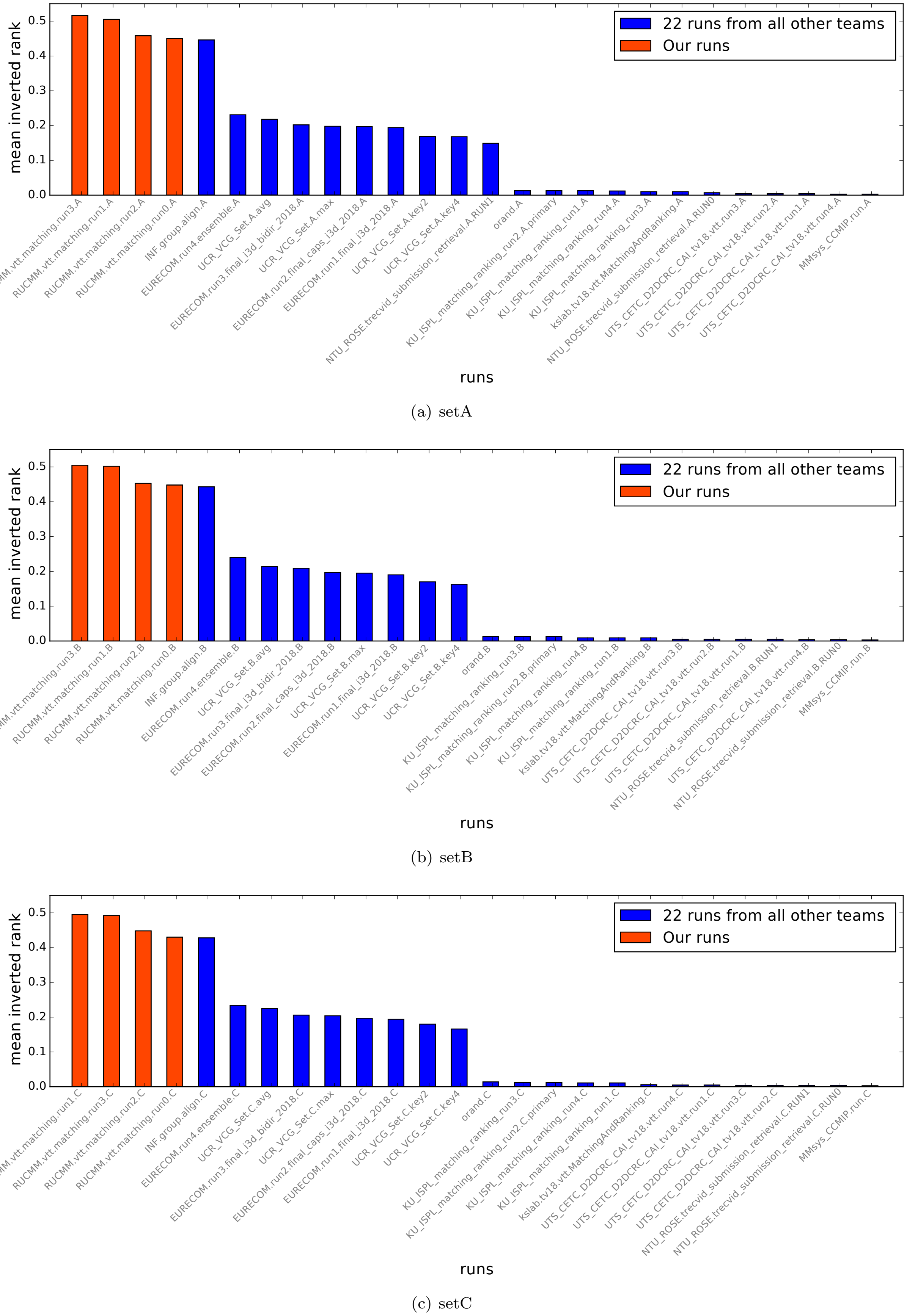
[8] T. Y. J. Xu, T. Mei and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[9] D.-D. Le, S. Phan, V.-T. Nguyen, B. Renoust, T. A. Nguyen, V.-N. Hoang, T. D. Ngo, M.-T. Tran, Y. Watanabe, M. Klinkigt, et al. NII-HITACHI-UIT at TRECVID 2016. In *TRECVID Workshop*, 2016.

[10] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. TGIF: A new dataset and benchmark on animated gif description. In *CVPR*, 2016.

[11] C. G. M. Snoek, J. Dong, X. Li, X. Wang, Q. Wei, W. Lan, E. Gavves, N. Hussein, D. C. Koelma, and A. W. M. Smeulders. University of Amsterdam and Renmin University at TRECVID 2016: Searching video, detecting events and describing video. In *TRECVID Workshop*, 2016.

[12] C. G. M. Snoek, X. Li, C. Xu, and D. C. Koelma. University of Amsterdam and Renmin university at TRECVID 2017: Searching video, detecting events and describing video. In *TRECVID Workshop*, 2017.

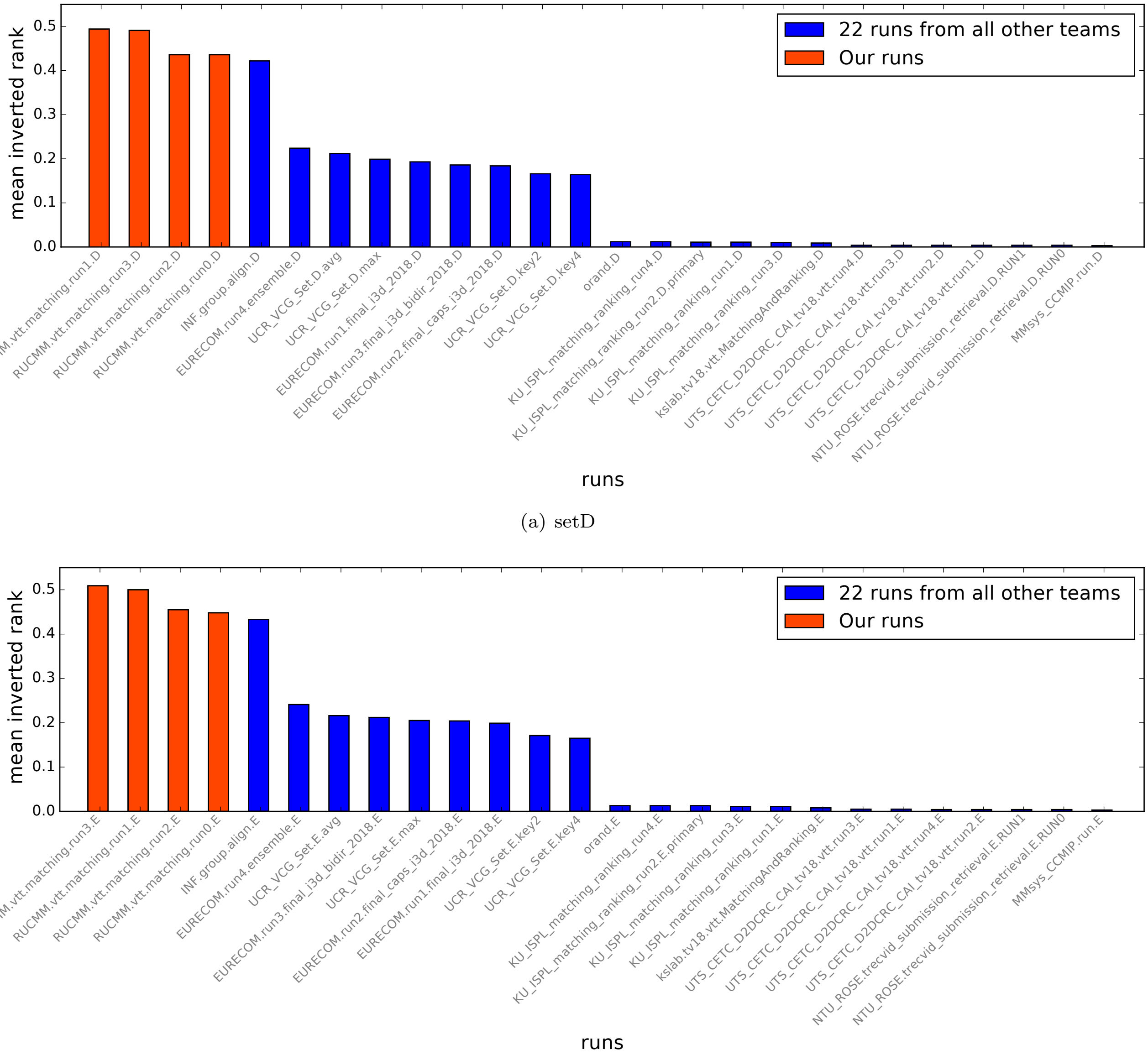


(a) setA

(b) setB

(c) setC

Figure 3: Overview of the TRECVID 2018 video-to-text matching and ranking task benchmark on setA, setB and setC, all runs ranked according to MIR. We were the best overall performer.



(a) setD

(b) setE

Figure 4: Overview of the TRECVID 2018 video-to-text matching and ranking task benchmark on setD and setE, all runs ranked according to MIR. We were the best overall performer.