Renmin University of China at TRECVID 2020: Sentence Encoder Assembly for Ad-hoc Video Search

Xirong Li, Fangming Zhou, Aozhu Chen
AI & Media Computing Lab, School of Information, Renmin University of China
MOE Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China
https://ruc-aimc-lab.github.io

Abstract

In this paper we summarize our TRECVID 2020 [2] video retrieval experiments. We participated in the Ad-hoc Video Search (AVS) task with a fully deep learning based solution. Our solution is based on a newly developed model, which we term Sentence Encoder Assembly (SEA) [11]. The novelty of the SEA model is two-fold. First, different from the prior art that uses only a single common space, SEA supports text-video matching in multiple encoder-specific common spaces. Such a property prevents the matching from being dominated by a specific encoder that produces an encoding vector much longer than other encoders. Second, in order to explore complementarities among the individual common spaces, we propose to train SEA by multi-space multi-loss learning. We exploit MSR-VTT and TGIF as training data. For video representation, we use pre-trained ResNet-152 and ResNeXt-101 to extract frame-level features, and C3D to extract segment-level features. Video-level features are obtained by mean pooling. Using SEA alone obtains a mean infAP of 0.236 for the 2020 task. Having SEA pre-trained on the Google's Conceptual Captions dataset is helpful, obtaining a higher infAP of 0.251. We again find late average fusion of distinct models (consisting of SEA and W2VV++ trained in varied settings) beneficial, obtaining the best infAP of 0.269 among our four submissions, and ranked at the second place teamwise.

1 Our Approach

As in our participation in the last two years [8, 10], we continue our practice of a fully deep learning based approach to the Ad-hoc Video Search (AVS) task. Given a novel textual query s and an unlabeled video v, our approach computes their **c**ross-**m**odal **s**imilarity $cms(s, v)$ by a deep cross-modal representation learning network that is end-to-end trained on many paired visual instances and sentence descriptions. Besides the W2VV++ model [9] and Dual Encoding [7] used in our solution for the TRECVID 2019 AVS task, this year we experiment with Sentence Encoder Assembly (*SEA*, in short) [11], a novel model that provides a more flexible and more effective mechanism to exploit distinct sentence encoders for query representation learning.

1.1 The SEA Model

Different from the previous deep learning based models that either uses a single sentence encoder or uses multiple sentence encoders but with a single common space, the *SEA* model performs text-video matching in multiple encoder-specific common spaces.

As illustrated in Fig. 1, the query representation module of *SEA* utilizes k distinct sentence encoders, denoted as $\{e_{t,2}, e_{t,2}, \ldots, e_{t,k}\}$. Accordingly, there are k cross-modal matching subnetworks, each corresponding to a specific sentence encoder. Each subnetwork, indexed by i, consists of two fully connected (FC) layers, one on the text side to transform $e_{t,i}(s)$ into a $d_{c,i}$-dimensional vector, and the other FC on the video side that transforms the video feature vector $f(v)$ into another $d_{c,i}$-dimensional vector. Consequently, the sentence-video semantic relevance, denoted as $cms_i(s, v)$, is computed as the cosine similarity between the two embedding:

$$cms_i(s,v) := \text{cosine}(\underbrace{FC_{t,i}(e_{t,i}(s))}_{\text{text embedding}}, \underbrace{FC_{v,i}(f(v))}_{\text{video embedding}}), \quad (1)$$

where $FC_{t,i}$ and $FC_{v,i}$ indicate the two FC layers, each followed by a *tanh* function to increase their learning capacity.

By simply averaging the similarities computed in the individual common spaces, we have the overall cross-modal similarity as

$$cms(s,v) := \frac{1}{k}\sum_{i=1}^{k} cms_i(s,v). \quad (2)$$

Note that we do not go for more complicated alternatives, *e.g.* weighing the individual similarities by self-attention mechanisms. Rather, we opt for this simple combination strategy, not only for preventing the risk of over-fitting. Such a strategy also encourages the individual common spaces to be good enough to be combined, as they are set to be equally important.

1.1.1 Choice of Sentence Encoders

We experimented with the following five sentence encoders:

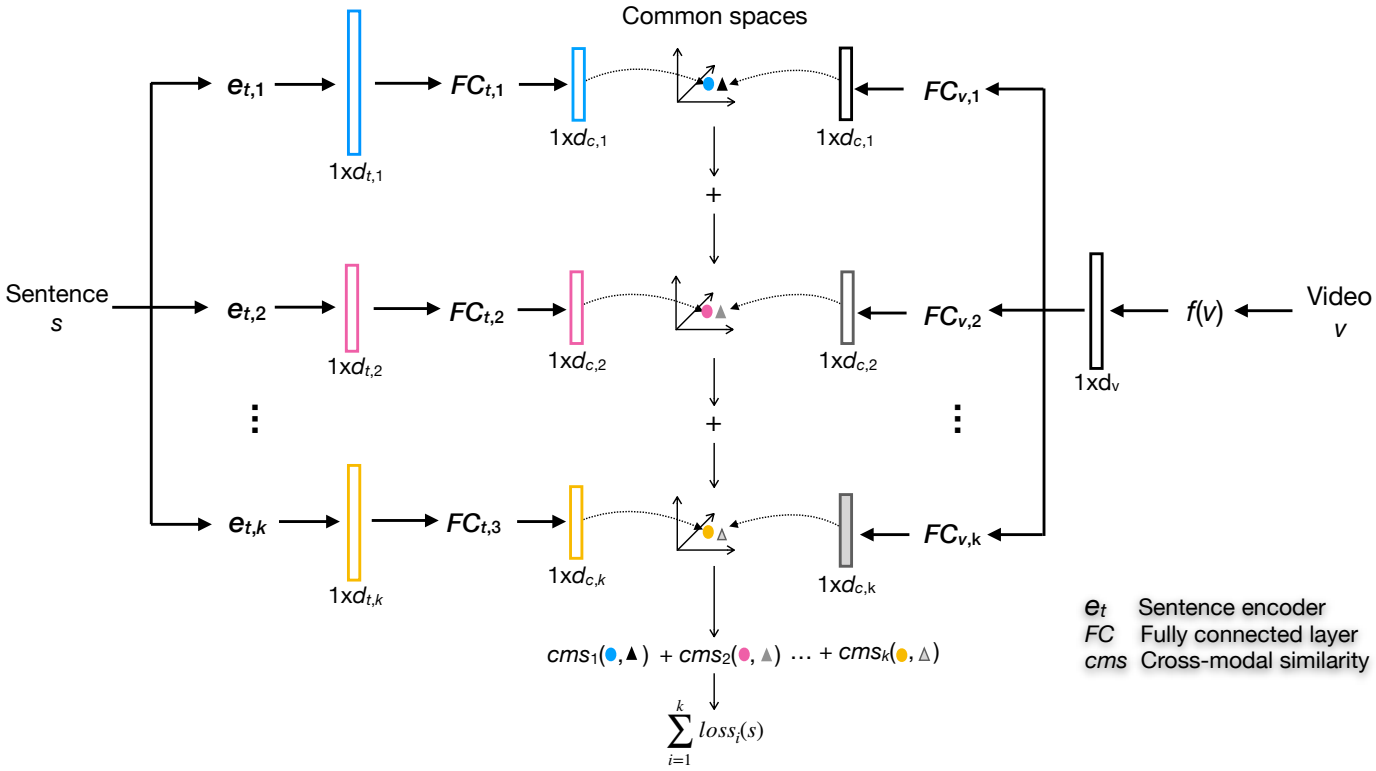

Figure 1: Conceptual diagram of the Sentence Encoder Assembly (SEA) model [11]. The key idea of SEA is to leverage multiple sentence encoders $\{e_{t,1}, e_{t,2}, \ldots, e_{t,k}\}$ and consequently build multiple encoder-specific common spaces for computing the cross-modal similarity between a given textual query s and an unlabeled video v.

- Bag-of-Words (BoW) [6]
- word2vec (w2v) [13]
- NetVlad [1]
- bi-GRU [4]
- BERT [5]

Varied combinations of the sentence encoders allow us to implement specific versions of the *SEA* model to handle queries of varying complexity.

1.1.2 Choice of Video Features

We use the same 4,096-dimensional ResNet+ResNeXt feature as in last year [10]. In addition, we extract a 2,048-dimensional C3D feature from video data, obtaining a combined ResNet+ResNeXt+C3D feature of size 6,144.

1.1.3 Choice of (Pre-)Training Data

Following our TV19 system [10], MSR-VTT [15] and TGIF [12] are merged as a common training set, while the development set of the TRECVID 2016 Video-to-Text Matching task [3] is used as an external validation set. However, different from [10] where all the models were trained from scratch, this year we experiment with the pre-training strategy. Some of the models, subject to the applicability of their video features, are pre-trained on the Google's Conceptual Captions (GCC) dataset [14], see Table 1.

2 Submissions

Based on the performance of the individual models and their combinations on the previous AVS tasks, we submitted the following four runs:

- *Run 4*: Model 4c, using *SEA* with BoW and w2v as its sentence encoders and ResNeXt-ResNet-C3D as video features.
- *Run 3*: Model 3b, using *SEA* with BoW and NetVlad as its sentence encoders, ResNeXt-ResNet as video features, and pre-trained on GCC.
- *Run 2*: Late average fusion of three base models, *i.e.* 3b, 3c and 4c.
- *Run 1* (primary run): Late average fusion of four base models, *i.e.* 3b, 3c, 4b and 4c.

The performance of our four runs and the base models on the TRECVID 2020 AVS task and all the previous AVS tasks is summarized in Table 2. It can be observed that pre-training is helpful. Adding the C3D feature is also helpful. Among the four runs, our primary run (*Run 1*) is the best.

Table 1: Varied models used in our experiments. A model postfixed with the letter *b* indicates that the model is pre-trained on the GCC dataset. As GCC is an image collection, it cannot be used to pre-train models that uses the C3D feature.

Model	Network	Sentence Encoders	Video Features	Pre-training
1	W2VV++	BoW	ResNet+ResNeXt	×
1b	W2VV++	BoW	ResNet+ResNeXt	√
1c	W2VV++	BoW	ResNet+ResNeXt+C3D	×
2	W2VV++	NetVlad	ResNet+ResNeXt	×
2b	W2VV++	NetVlad	ResNet+ResNeXt	√
2c	W2VV++	NetVlad	ResNet+ResNeXt+C3D	×
3	SEA	BoW, NetVlad	ResNet+ResNeXt	×
3b	SEA	BoW, NetVlad	ResNet+ResNeXt	√
3c	SEA	BoW, NetVlad	ResNet+ResNeXt+C3D	×
4	SEA	BoW, w2v	ResNet+ResNeXt	×
4b	SEA	BoW, w2v	ResNet+ResNeXt	√
4c	SEA	BoW, w2v	ResNet+ResNeXt+C3D	×
5	SEA	BoW, w2v, bi-GRU	ResNet+ResNeXt	×
5b	SEA	BoW, w2v, bi-GRU	ResNet+ResNeXt	√
5c	SEA	BoW, w2v, bi-GRU	ResNet+ResNeXt+C3D	×
6	SEA	BoW, w2v, bi-GRU, BERT	ResNet+ResNeXt	×
6b	SEA	BoW, w2v, bi-GRU, BERT	ResNet+ResNeXt	√
6c	SEA	BoW, w2v, bi-GRU, BERT	ResNet+ResNeXt+C3D	×

This is in line with our previous findings [8, 10] that late fusion always boosts the performance further.

An overview of the AVS task benchmark is shown in Fig. 2. Team-wise, our submissions are ranked at the second place among all the submissions.

Acknowledgments

The authors are grateful to the TRECVID coordinators for the benchmark organization effort. This research was supported by the National Natural Science Foundation of China (No. 61672523), Beijing Natural Science Foundation (No. 4202033), and the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 18XNLG19).

References

[1] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic. Netvlad: CNN architecture for weakly supervised place recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[2] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, A. F. Smeaton, Y. Graham, G. J. F. Jones, W. Kraaij, and G. Quénot. Trecvid 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of TRECVID 2020*. NIST, USA, 2020.

[3] G. Awad, J. Fiscus, D. Joy, M. Michel, A. Smeaton, W. Kraaij, G. Quénot, M. Eskevich, R. Aly, R. Ordelman, G. Jones, B. Huet, and M. Larson. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *TRECVID Workshop*, 2016.

[4] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[6] J. Dong, X. Li, and C. G. M. Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 2018.

[7] 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.

[8] X. Li, J. Dong, C. Xu, J. Cao, X. Wang, and G. Yang. Renmin University of China and Zhejiang Gongshang University at TRECVID 2018: Deep cross-modal embeddings for video-text retrieval. In *TRECVID Workshop*, 2018.

[9] X. Li, C. Xu, G. Yang, Z. Chen, and J. Dong. W2VV++: Fully deep learning for ad-hoc video search. In *ACM Multimedia*, 2019.

[10] X. Li, J. Ye, C. Xu, S. Yun, L. Zhang, X. Wang, R. Qian, and J. Dong. Renmin University of China and Zhejiang

Table 2: Performance of our four runs and the individual models on the TRECVID 2016–2020 AVS tasks.

	2016	*2017*	*2018*	*2019*	*2020*
Individual models:					
Model 1	0.161	0.220	0.122	0.161	0.201
Model 1b	0.189	0.229	0.136	0.156	0.241
Model 1c	0.151	0.246	0.115	0.160	0.230
Model 2	0.154	0.215	0.126	0.143	0.162
Model 2b	0.197	0.236	0.141	0.164	0.217
Model 2c	0.142	0.234	0.119	0.147	0.193
Model 3	0.158	0.236	0.134	0.162	0.202
Model 3b	0.189	0.238	0.143	0.170	0.251
Model 3c	0.161	0.265	0.143	0.176	0.236
Model 4	0.157	0.234	0.128	0.166	0.189
Model 4b	0.196	0.246	0.142	0.172	0.240
Model 4c	0.154	0.251	0.137	0.183	0.236
Model 5	0.164	0.228	0.125	0.167	0.186
Model 5b	0.191	0.240	0.128	0.165	0.227
Model 5c	0.164	0.250	0.121	0.176	0.227
Model 6	0.159	0.229	0.117	0.155	0.185
Model 6b	0.194	0.235	0.121	0.165	0.228
Model 6c	0.154	0.242	0.122	0.161	0.219
Our TV20 submissions:					
Run 4 (Model 4c)	0.154	0.251	0.137	0.183	0.236
Run 3 (Model 3b)	**0.196**	0.246	0.142	0.172	0.251
Run 2 (Late fusion of 3b, 3c, 4c)	0.182	0.270	0.152	**0.192**	0.263
Run 1 (Late fusion of 3b, 3c, 4b, 4c)	0.186	**0.272**	**0.153**	0.190	**0.269**

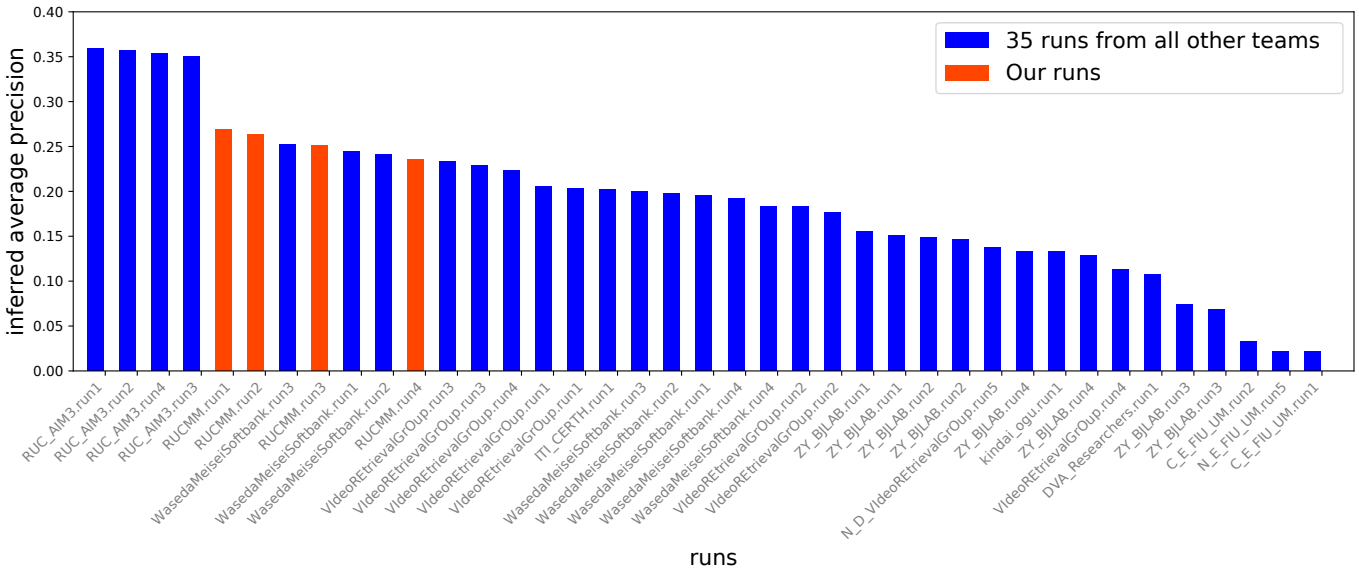

Figure 2: Overview of the TRECVID 2020 AVS benchmark evaluation.

Gongshang University at TRECVID 2019: Learn to search and describe videos. In *TRECVID Workshop*, 2019.

[11] X. Li, F. Zhou, C. Xu, J. Ji, and G. Yang. SEA: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 2021.

[12] 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.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[14] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

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