DL-61-86 at TRECVID 2017: Video-to-Text Description

Jianfeng Dong[1,2], Shaoli Huang[1], Duanqing Xu[2], Dacheng Tao[1]

1. UBTECH Sydney AI Centre, SIT, FEIT, University of Sydney
2. College of Computer Science and Technology, Zhejiang University

Matching and Ranking subtask

Query Videos:

Candidate sentences to be ranked:

a man speaks to audiences indoors

a person skates indoors

Athletics make a choreography in gym. ✓

a woman is holding a phone to her ear.

Cross-media Similarity

Video

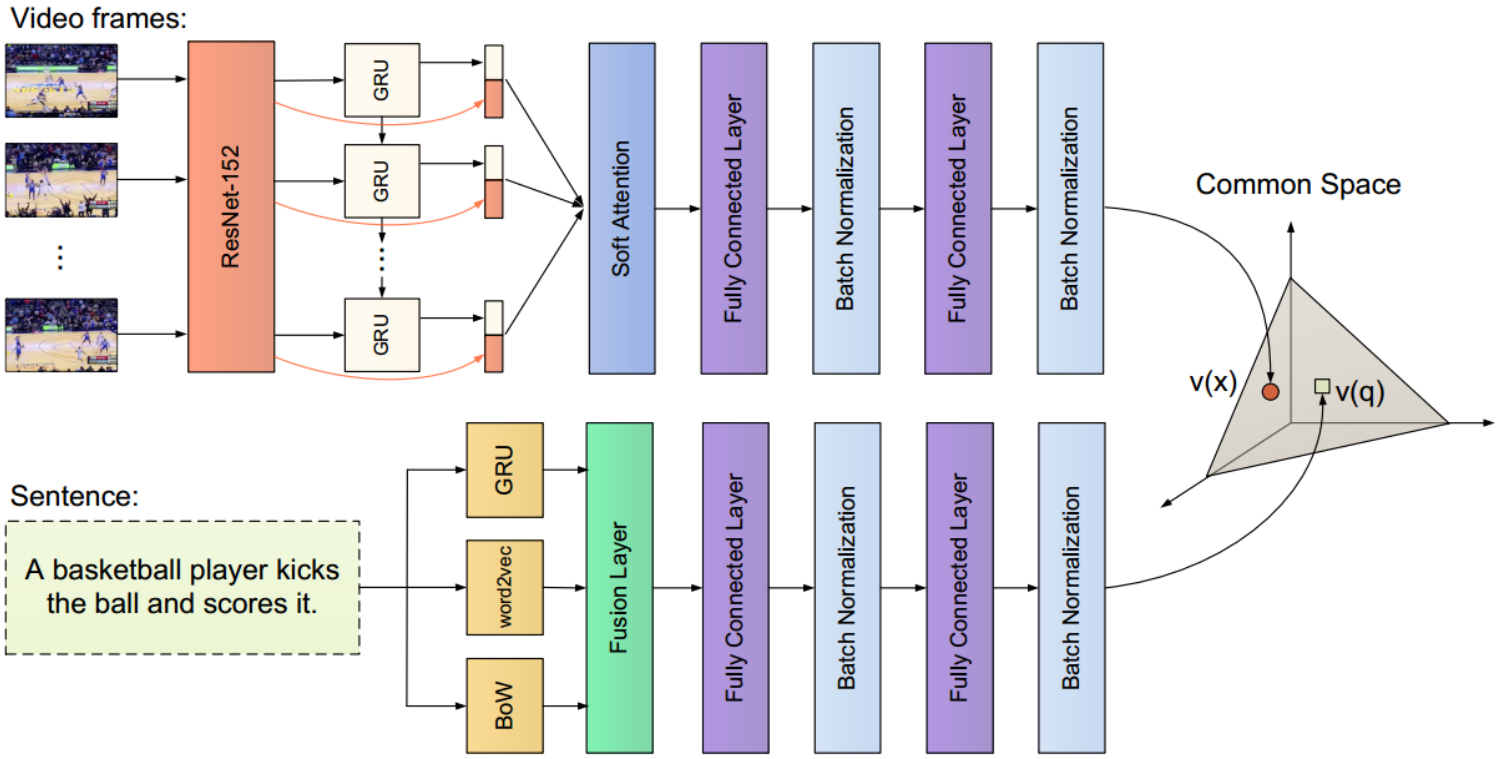
?

Sentence

Athletics make a choreography in gym.

Key question: how to compute cross-media similarity?

Our Model

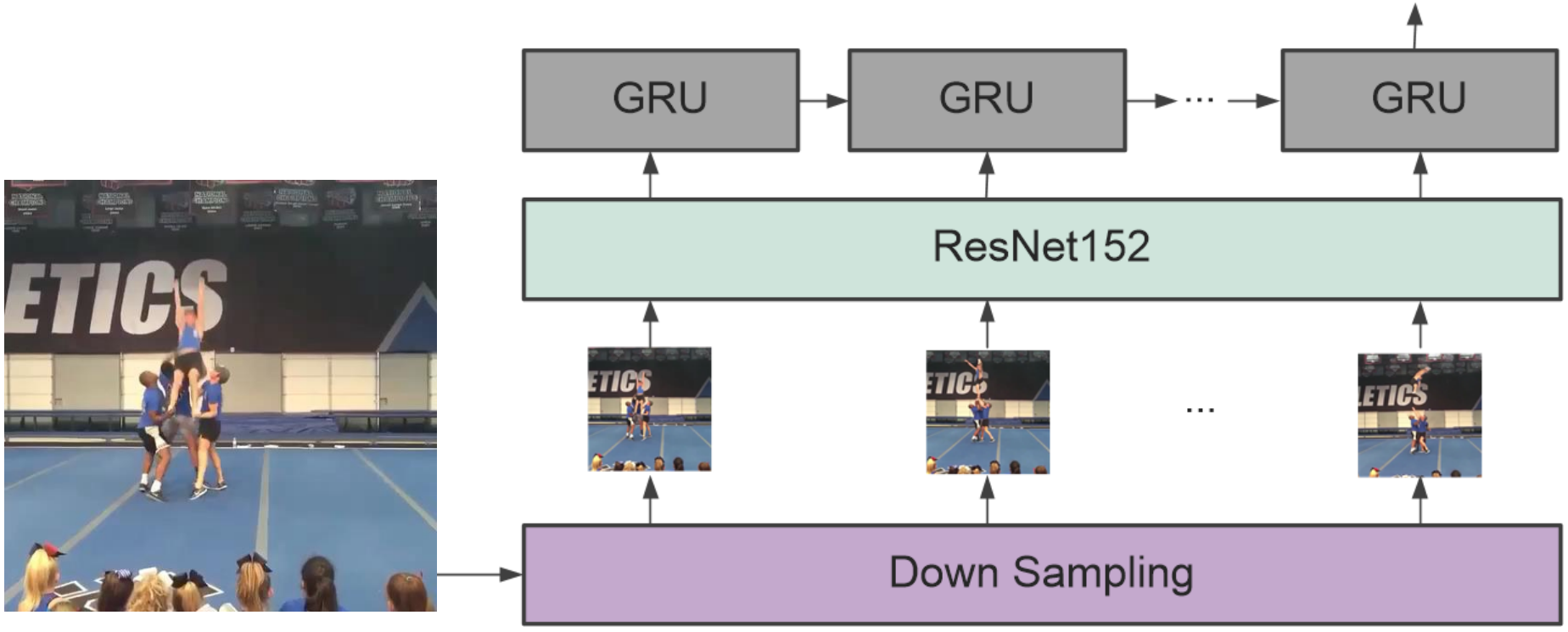


Key components:

- Spatial Enhanced Video Representation
- Multi-scale Sentence Representation

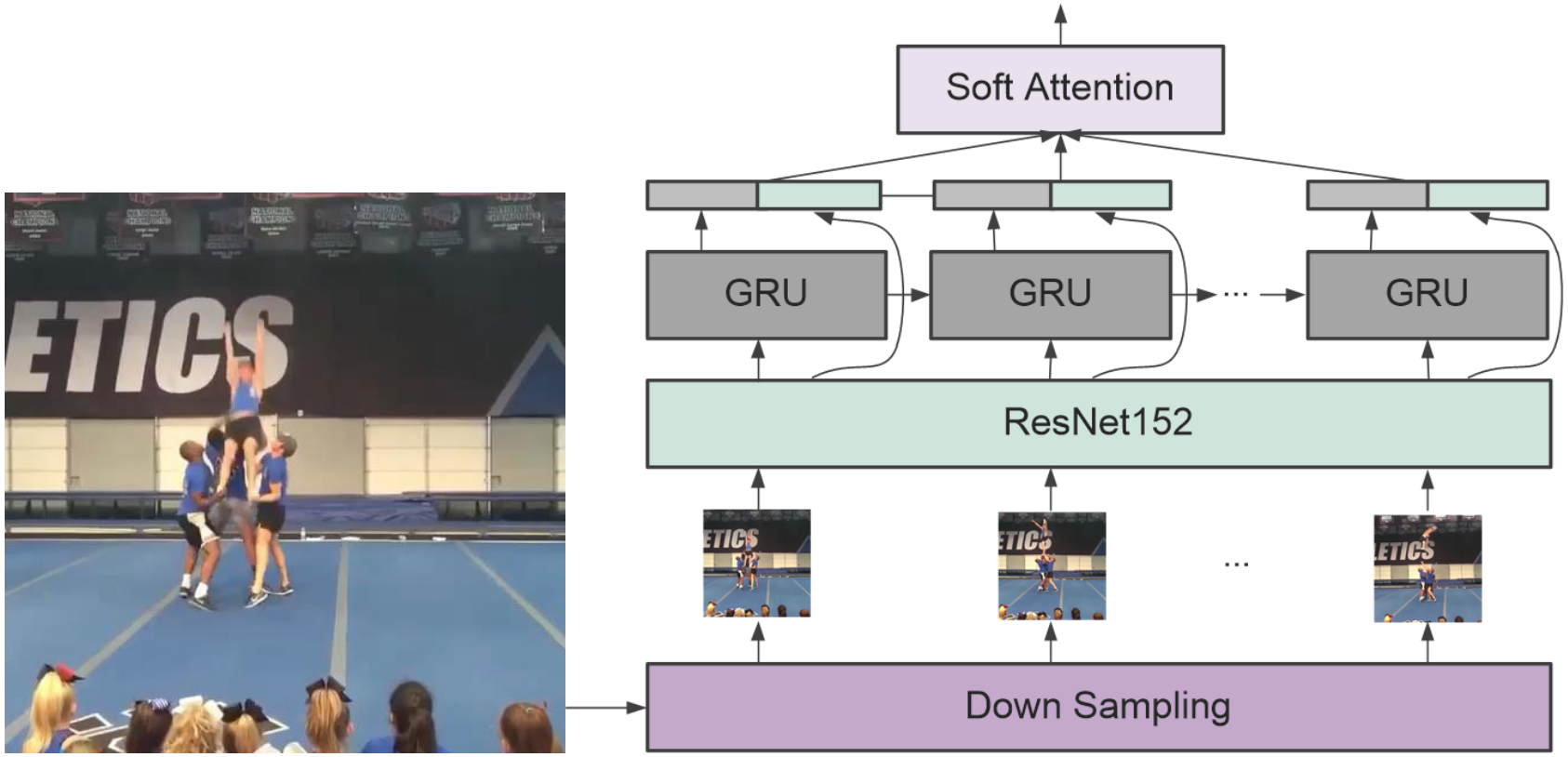
Common way of video representation

Use RNN to capture spatio-temporal information.



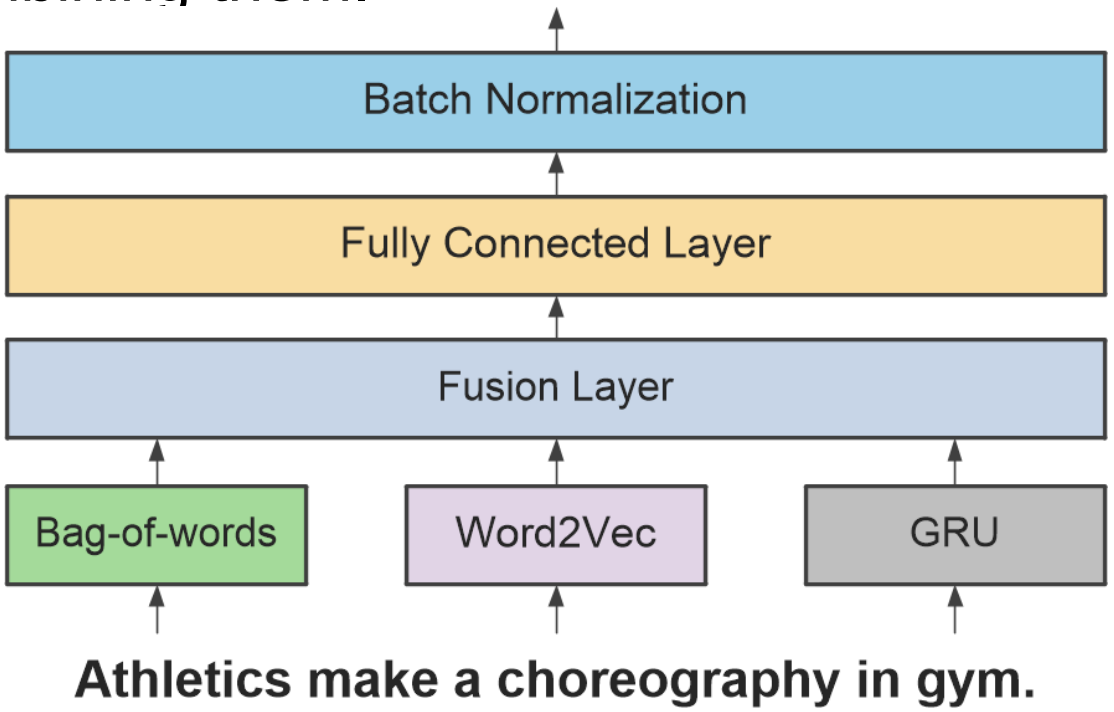
Spatial Enhanced Video Representation

Learn a GRU with skip-connections that allow bypassing of the spatial features.



Multi-scale Sentence Representation

It merges bag-of-words, word2vec and GRU sentence features and letting the model figure out the optimal way for combining them.



J. Dong, X. Li, and C. G. M. Snoek. Predicting Visual Features from Text for Image and Video Caption Retrieval. arXiv 2017.

Objective Function

Triplet Ranking Loss:

$$l(x, q; \theta) = \sum_{q'} [\alpha + s(x, q') - s(x, q)]_+$$

$$[x]_+ \equiv \max(x, 0)$$

Improved Triplet Ranking Loss: (using hardest example)

$$l(x, q; \theta) = \max_{q'} [\alpha + s(x, q') - s(x, q)]_+$$

Other wining components

1. Use more training data and fine-tune the model on the data provided by TRECVID
2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU
3. Use batch normalization after the FC layer
4. Fuse different models

Datasets

External datasets

Table 1. Overview of datasets used in our submission.

	Dataset	# Videos	# Sentences
Train	MSVD	1,970	80,863
	MSR-VTT	10,000	200,000
	TGIF	101,980	125,672
Validation	tv2016train	200	400
Fine-Tune	tv2016test	1,915	3,830

Datasets provides by TRECVID 2016

Other wining components

1. Use more training data and fine-tune the model on the data provided by TRECVID
2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU
3. Use batch normalization after the FC layer
4. Fuse different models

Pre-trained word2vec

1. Word2vec trained on the Google news documents

2. Word2vec trained on the tags of Flickr images

Word2vec with the dimensionality of 500 trained on 30 million Flickr tags.

URL:https://drive.google.com/open?id=0B1OT7LFjhrF_RWptMjY2TVBqLWc

J. Dong, X. Li, and C. G. M. Snoek. Predicting Visual Features from Text for Image and Video Caption Retrieval. arXiv 2017.

Other wining components

1. use more training data and fine-tune the model on the data provided by TRECVID
2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU
3. Use batch normalization after the FC layer
4. Fuse different models

Other wining components

1. use more training data and fine-tune the model on the data provided by TRECVID
2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU
3. Use batch normalization after the FC layer
4. Fuse different models

Fuse with Word2VisaulVec

Model fusion is simple but it is effective.

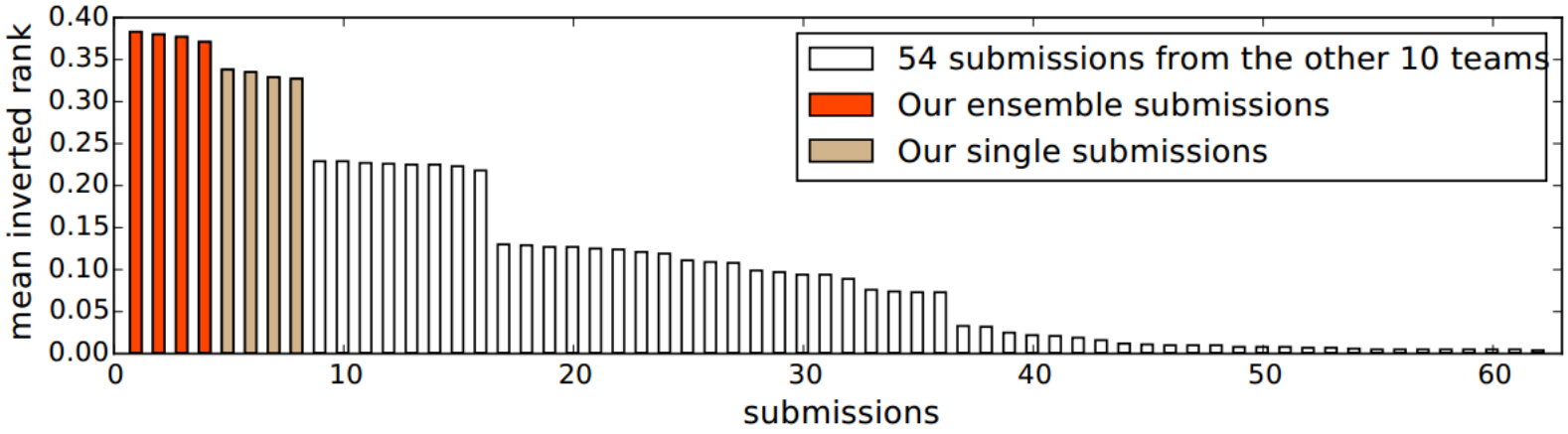
Improve Word2VisualVec:

1. Use multi-scale text representation to embed sentence
2. Use the improved triplet ranking loss

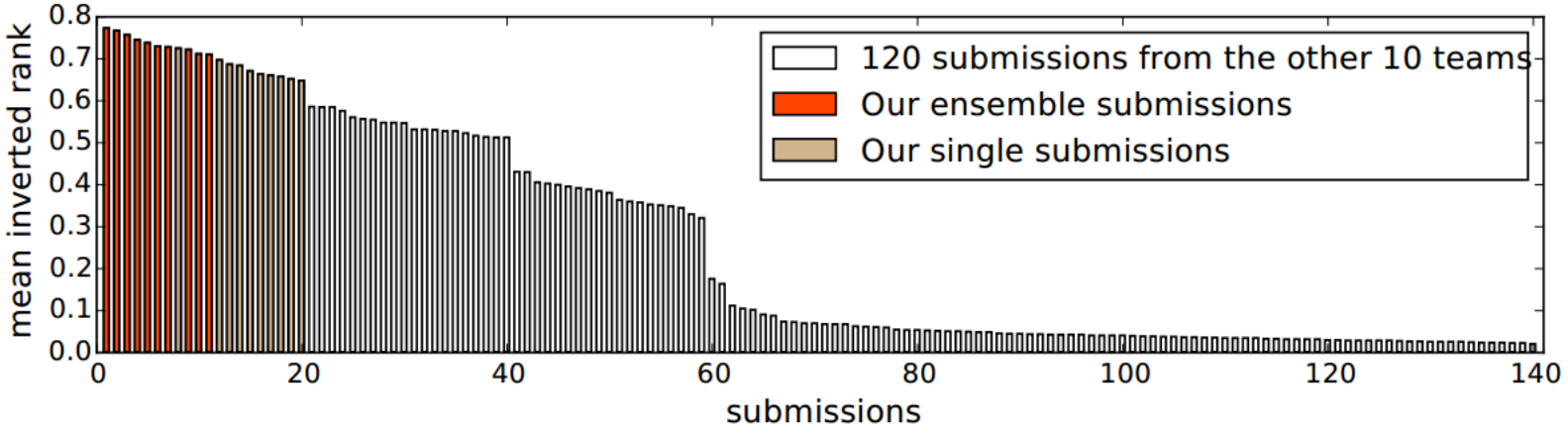
C. G. Snoek, J. Dong,et al. University of amsterdam and renmin university at trecvid 2016: Searching video, detecting events and describing video. In *TRECVID Workshop*, 2016.

Evaluation Results

Our submissions lead the evaluation with a great margin.



(a) Results on the test set 2



(d) Results on the test set 5

Take-home Messages

- Use Spatial Enhanced Video Representation to embed videos
- Use Multi-scale Sentence Vectorization to embed sentences
- Some other wining components
  1. use more training data and fine-tune the model on the data provided by TRECVID
  2. Use pre-trained word2vec to initialize word embedding before the LSTM/GRU
  3. Use batch normalization after the FC layer
  4. Fuse different models