Feature Re-Learning with Data Augmentation for Content-based Video Recommendation

Jianfeng Dong[1], Xirong Li[2], Chaoxi Xu[2], Gang Yang[2], Xun Wang[1]

1. Zhejiang Gongshang University
[2]AI & Media Computing Lab, Renmin University of China

Grand Challenge Session @ ACM Multimedia 2018

Videos are important

Video-sharing websites are very popular.

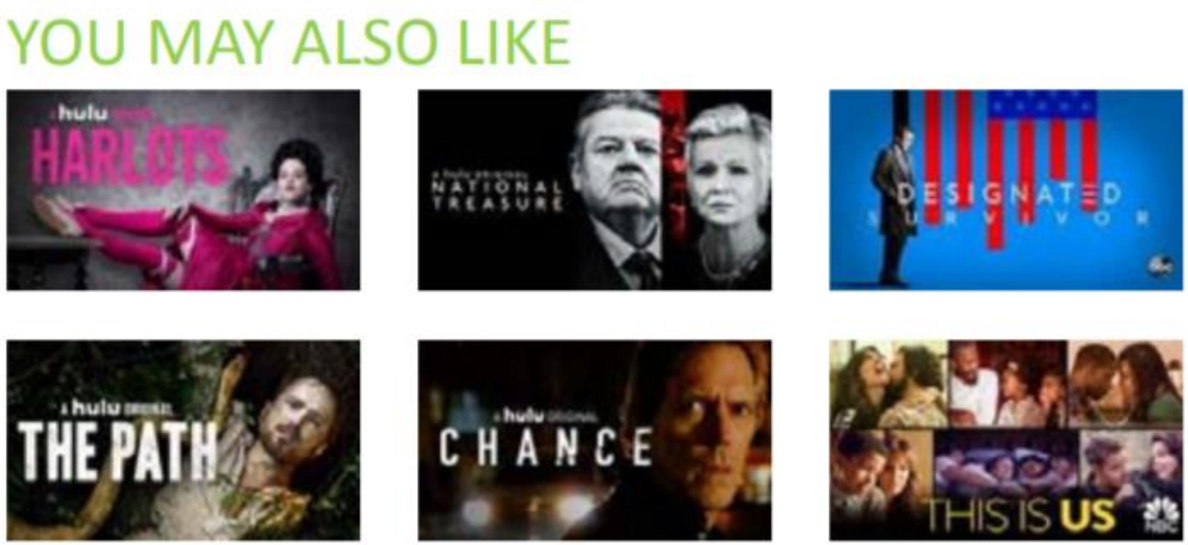
On YouTube:

- **300 hours** of video are uploaded every minute
- **5 billion** videos are watched per day
- **30 million** user visited YouTube per day
- **2.1 hours** consumed by visitors per day per person

Video recommendation

In a rich context

- User interaction: browsing, commenting and rating
- Meta-data: title, filename
- …



Cold-start video recommendation

- No contextual information
- Video content only

~~browsing~~

~~commenting~~

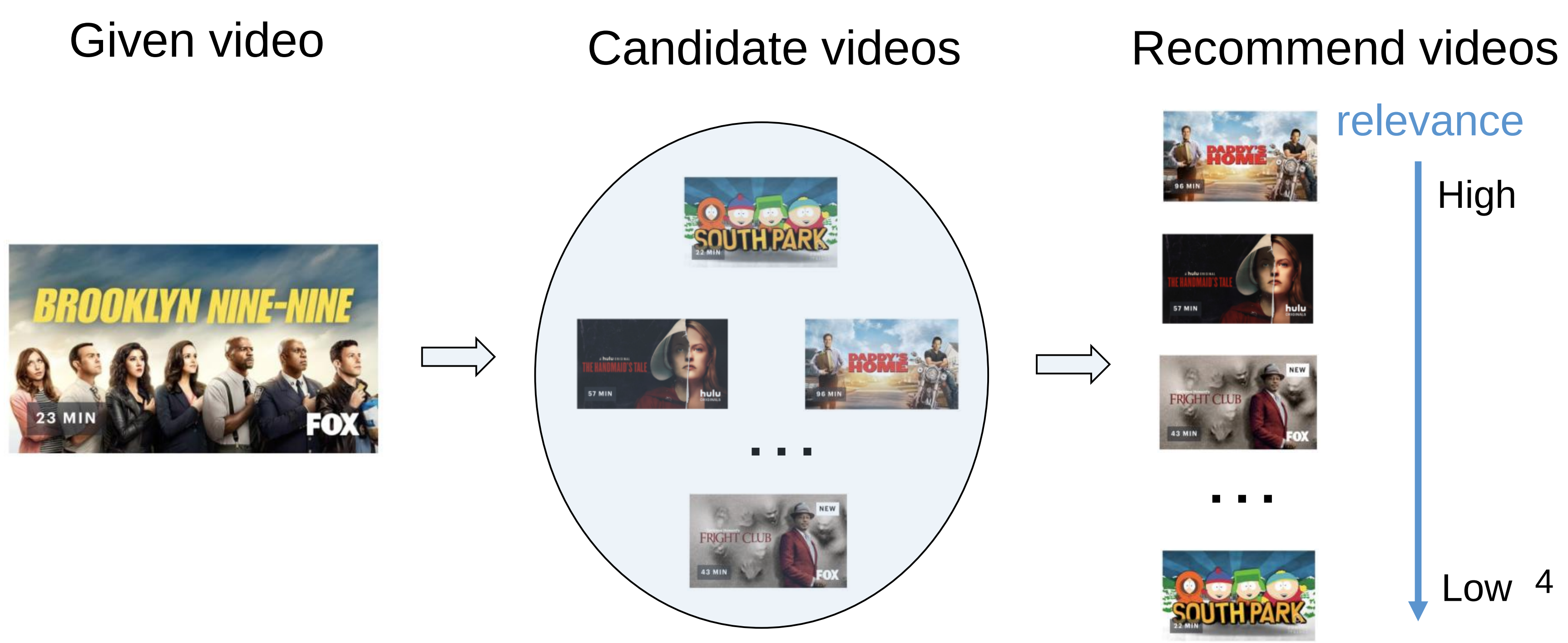
~~rating~~

…

Hulu task

Content-based Video Relevance Prediction Challenge

Given a video, participants are asked to rank a list of pre-specified videos in terms of their relevance.



Task setup

What we have

- Two tracks:
  - Movies Track
  - TV-shows Track
- Video relevance list
- Visual features
  - frame-level feature: Inception-v3
  - video-level feature: C3D

What we do not have

- Videos
- Frames
- Contextual information
  - user interaction
  - meta-data
  - …

Impossible to visually examine recommendation results

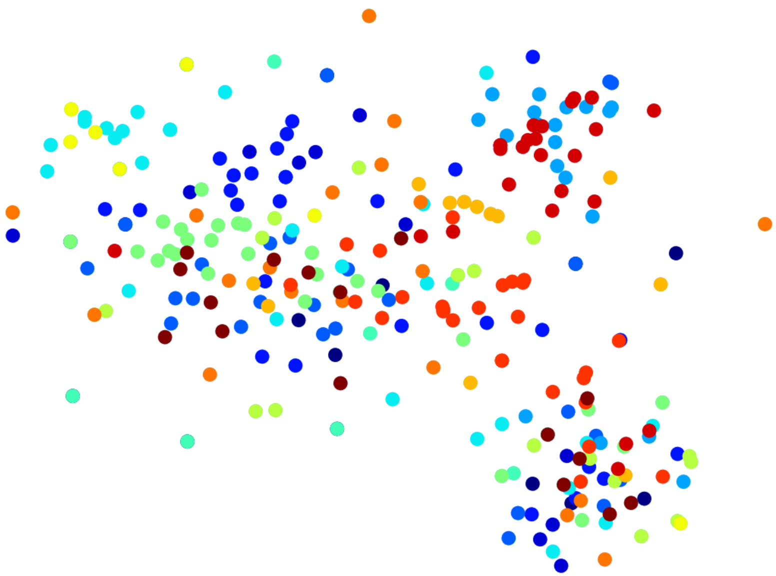
Challenge one

Limited training data.

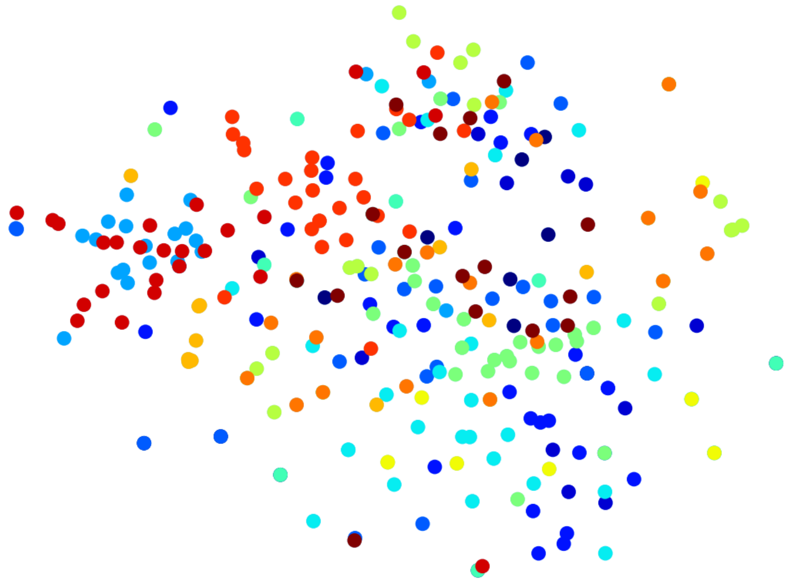
	train	validation	test
Movies Track	4500	1188	4500
TV-shows Track	3000	864	3000

Challenge two

Off-the-shelf CNN features are not optimal.



inception-v3



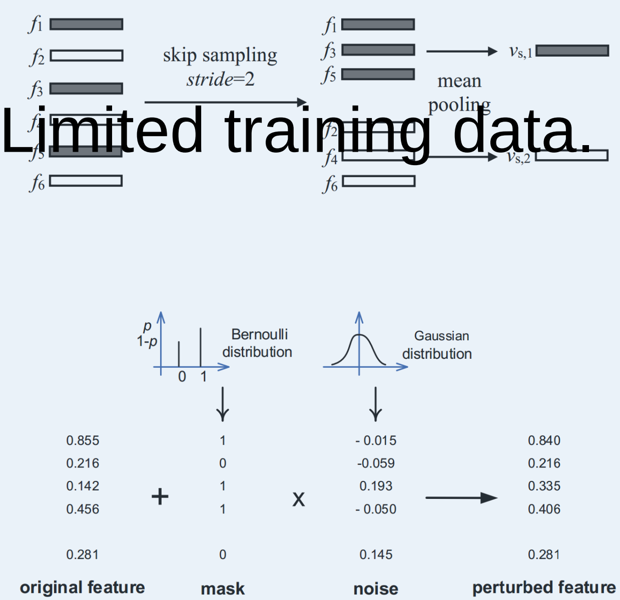
C3D

Our solution

Challenge one

Data augmentation

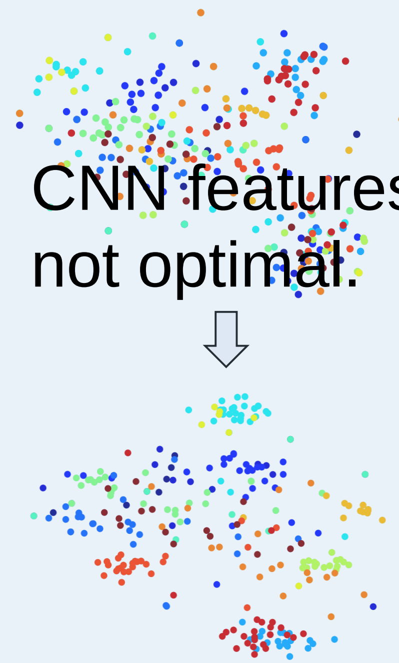
Limited training data.



Challenge two

Feature re-learning

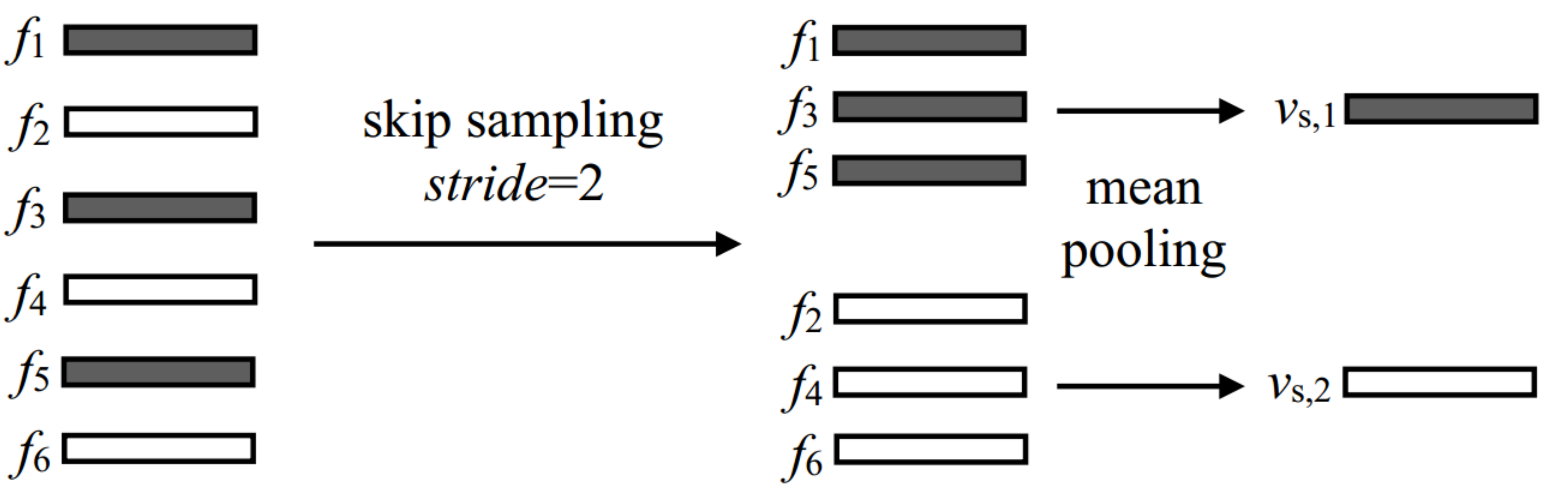
CNN features are not optimal.



Late fusion

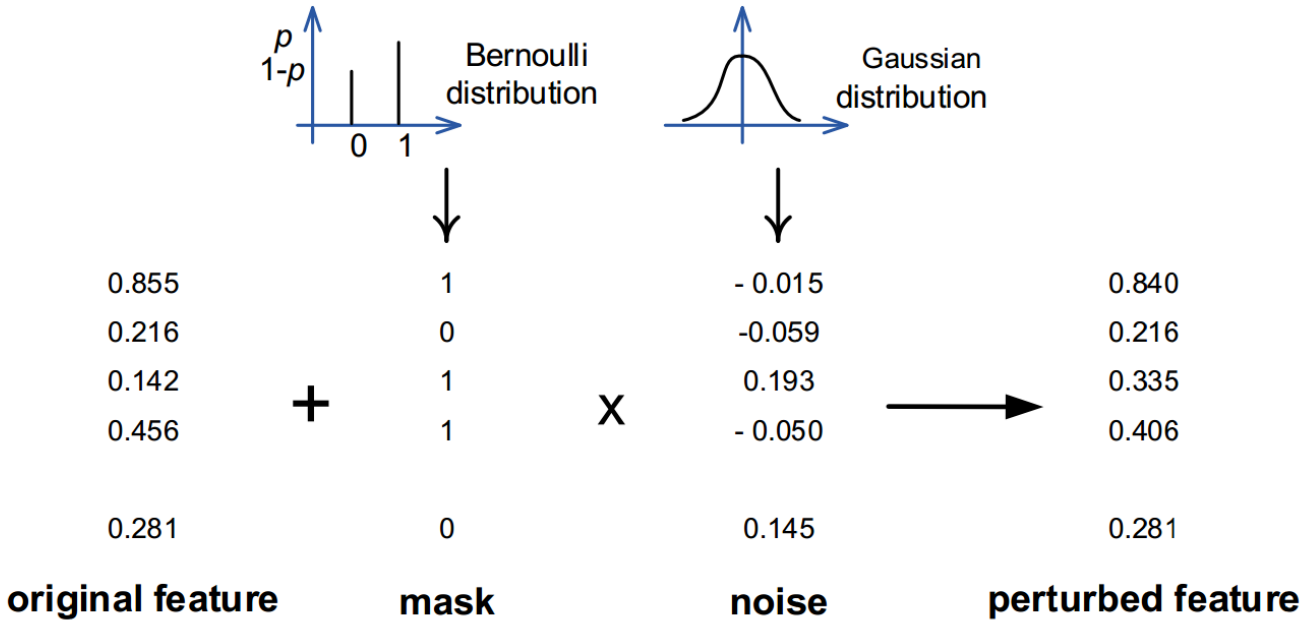
Augmentation for frame-level features

Inspired by the fact that humans could grasp the video topic after watching only several sampled video frames in order, we augment data by skip sampling.

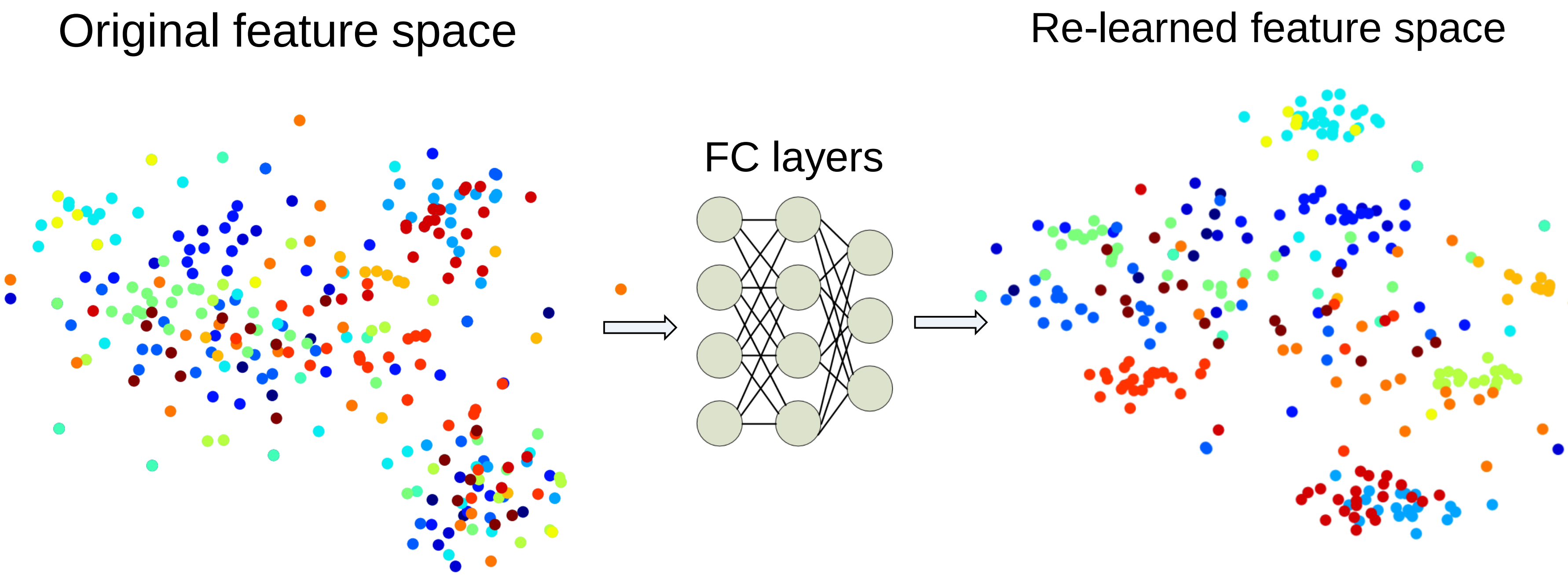


Augmentation for video-level features

As adding tiny perturbations to image pixels are imperceptible to humans, we introduce perturbation-based data augmentation.



Feature re-learning



Triplet ranking loss:

$$\mathcal{L}(v, v^+, v^-; W, b) = max(0, \alpha - cs_\phi(v, v^+) + cs_\phi(v, v^-))$$

Augmentation and re-learning

Both data augmentation and feature re-learning is effective.

Feature	Augmentation	Re-Learning	Movies	TV-shows
Inception-v3	×	×	0.099	0.124
	×	√	0.163	0.199
	√	√	**0.191**	**0.244**
C3D	×	×	0.112	0.145
	×	√	0.155	0.185
	√	√	**0.163**	**0.196**

Choice of loss functions

Triplet ranking loss consistently outperforms the other two loss functions on both two tracks.

Loss	Movies	TV-shows
Triplet ranking loss	**0.163**	**0.199**
Improved Triplet ranking loss [1]	0.125	0.181
Contrastive loss [2]	0.160	0.194

[1] F. Faghri, D. J Fleet, J. R. Kiros, and S. Fidler. 2018. VSE++: improved visual semantic embeddings. In BMVC.
[2] R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In CVPR

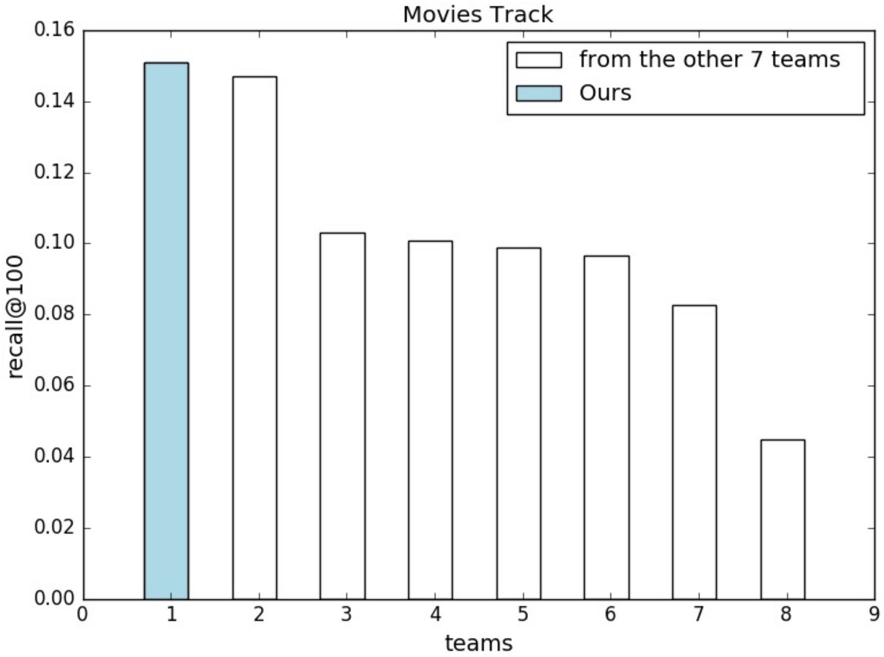
Late fusion

Late fusion is employed by averaging the relevance given by multiple models, which further boosts the performance.

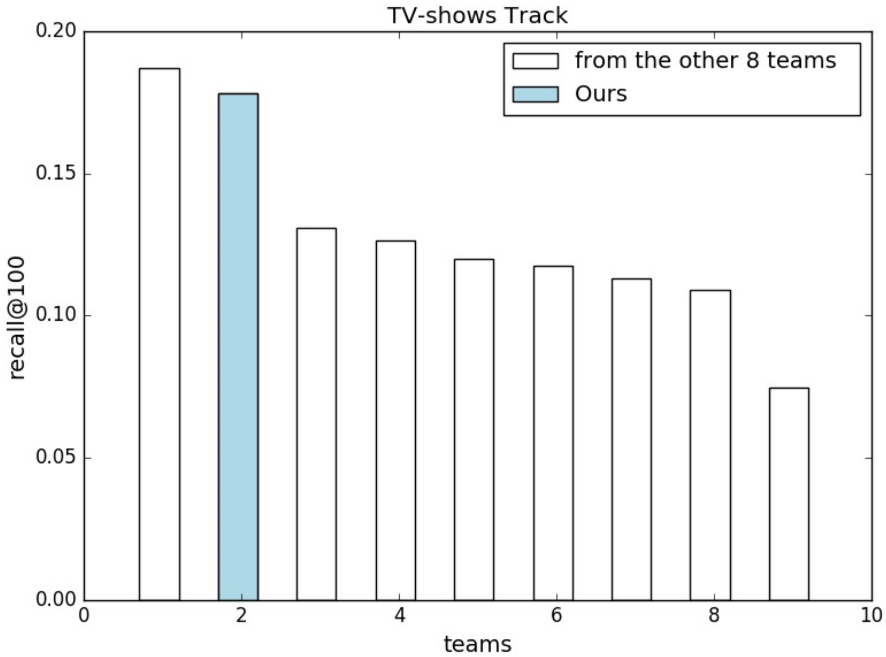
Late fusion	Movies	TV-shows
×	0.191	0.244
√	**0.211**	**0.276**

Official evaluation

Our runs are ranked first on Movies Track and second on TV-shows Track.



Movies Track



TV-shows Track

Take-home messages

Good practices

- data augmentation on features generating more training instances
- feature re-learning with the triplet ranking loss
- late fusion of multiple models

https://github.com/danieljf24/cbvr

Our runs

Track 1: TV-shows								
	hit@k				recall@k			
	k=5	k=10	k=20	k=30	k=50	k=100	k=200	k=300
Hulu	0.249	0.356	0.461	0.525	0.085	0.141	0.219	0.269
run 1	0.274	0.365	0.488	0.542	0.099	0.160 (↑ 13.5%)	0.248	0.302
run 2	0.287	0.381	0.492	0.550	0.104	0.167 (↑ 18.4%)	0.257	0.314
run 3	0.288	0.391	0.484	0.539	0.099	0.162 (↑ 14.9%)	0.249	0.305
run 4	**0.309**	**0.411**	0.506	0.567	0.109	0.173 (↑ 22.7%)	0.266	0.323
run 5	0.308	0.408	**0.522**	**0.589**	**0.112**	**0.178** (↑ 26.2%)	**0.273**	**0.331**

Track 2: Movies								
	hit@k				recall@k			
	k=5	k=10	k=20	k=30	k=50	k=100	k=200	k=300
Hulu	0.190	0.242	0.320	0.373	0.081	0.116	0.168	0.206
run 1	0.210	0.272	0.355	0.412	0.092	0.133 (↑ 14.7%)	0.192	0.237
run 2	0.211	0.278	0.368	0.427	0.095	0.139 (↑ 19.8%)	0.201	0.248
run 3	0.215	0.278	0.359	0.422	0.096	0.138 (↑ 19.0%)	0.198	0.244
run 4	**0.234**	0.298	**0.390**	**0.448**	0.104	0.148 (↑ 27.6%)	0.210	0.258
run 5	0.232	**0.302**	0.389	0.441	**0.105**	**0.151** (↑ 30.2%)	**0.215**	**0.263**