Content-based Related Video Recommendations

Joonseok Lee
Google Research
joonseok@google.com

Nisarg Kothari
Google Research
ndk@google.com

Paul Natsev
Google Research
natsev@google.com

1 Introduction

This is a demo of related video recommendations, seeded from random YouTube videos, and based purely on video content signals. Traditional recommendation systems using collaborative filtering (CF) approaches suggest related videos for a given seed based on how many users have watched a particular candidate video right after watching the seed video. This does not take the video content into account but relies on aggregate user behavior. Traditional CF approaches work very well when the seed and the candidate videos are relatively popular – they must be watched in a sequence by many users in order for them to be identified as related by the CF system.

In this demo, we focus on the cold-start problem, where either the seed and/or the candidate video are freshly uploaded (or undiscovered) so the CF system cannot identify any related videos for them. Being able to recommend freshly uploaded videos as well as recommend good related videos for fresh video seeds are important for improving freshness and user engagement. We model this as a video content-based similarity learning problem, and learn deep video embeddings trained to predict ground-truth video relationships (identified by a CF co-watch-based system) but using only visual content. The system does not depend on availability on video metadata or any click information, and can generalize to both popular and tail content, as well as new video uploads. It embeds any new video into a 1024-dimensional representation based on its content and pairwise video similarity is computed simply as a dot product in the embedding space. We show that the learned video embeddings generalize beyond simple visual similarity and are able to capture complex semantic relationships.

2 Data and Method

We start by running a pre-trained Inception-style [2] image annotation model on video frames sampled at 1 frame-per-second and extract the last fully-connected layer as a frame-level feature representation. We average and PCA the frame-level features into a video-level input representation, and train an additional 3-layer deep network with fully connected layers to produce a fine-tuned and task adapted output video embedding. We train the second network on randomly chosen 3M YouTube videos and their co-watch history—that is, videos frequently co-watched together by multiple users are brought closer together in the embedding space, while videos with no such relationship are pushed apart. This is accomplished using triplet loss training with semi-hard negative example mining [1].

3 Experiments

We target two cold-start scenarios: 1) recommending recently uploaded videos for an established video seed (*cold-start candidates*), and 2) recommending established videos for recently uploaded seed videos (*cold-start seeds*). For each seed in the test set, we recommend 60 videos and compute the Normalized Discounted Cumulative Gain (NDCG) [1] scores. The baseline is the 1500-dimensional input video-level feature representation (average-pooled frame-level features from the Inception bottleneck layer, no fine-tuning). We compare embeddings with varying dimensions in Table 1, which shows that fine-tuning improves the NDCG metric (higher is better) even with 20 times more compact embeddings. The best performance is achieved with 512-dimensional embeddings, but even 128 dimensions achieve 8-20% relative NDCG gain over the baseline (which is 12 times larger).

[1] https://en.wikipedia.org/wiki/Discounted_cumulative_gain

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

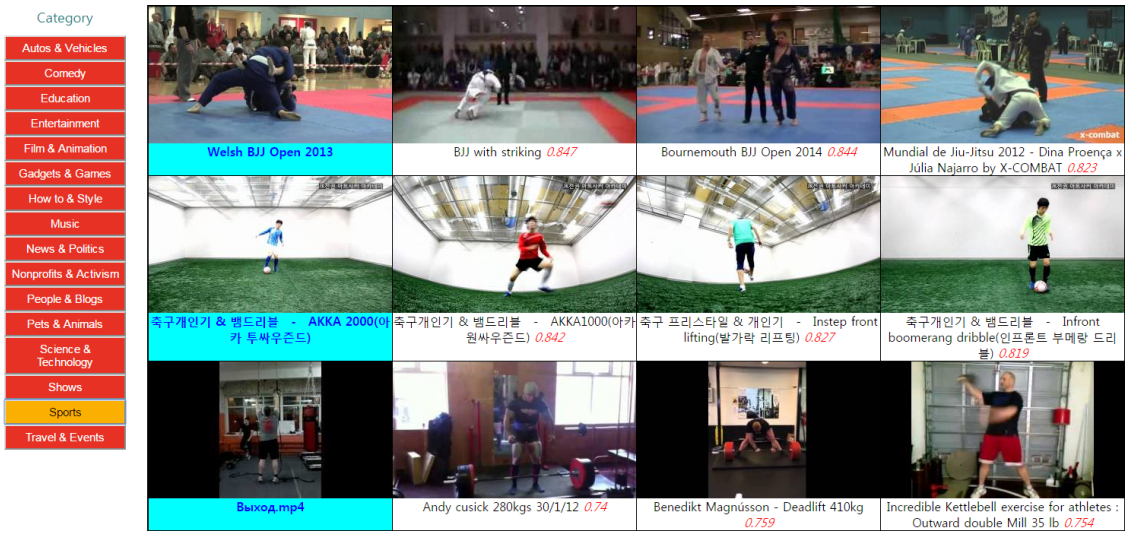

Figure 1: Demo screen with pre-computed videos

Output dimension	32	64	128	256	512	1024	Baseline (1500)
Cold-start candidates	2.19%	2.58%	2.87%	2.96%	3.00%	2.98%	2.40%
Cold-start seeds	4.20%	4.93%	5.42%	5.61%	5.69%	5.67%	5.02%

Table 1: NDCG@60 scores (the higher, the better) with various output embedding dimensions

4 Demonstration

We show some examples in Figure 1. We categorized about 300 videos into several high-level categories, such as music, sports, vehicles, which can be selected on the left menu. In each category, we show random seed videos and the top 10 recommended videos for each seed, with corresponding thumbnails and scores. Clicking on any of the thumbnails plays the corresponding YouTube video. In the demo, participants will be able to select categories of interest and browse corresponding seed videos and related video suggestions.

Figure 2 shows some examples of the top recommendation for 3 seed videos. In Figure 2a, the seed and recommended videos are both about cats but in different languages (English and Russian), and the system is able to infer a link between them based on visual and semantic similarity. The seed and top recommendation in Figure 2b are both compilations of K-pop music videos. Even though we do not use audio features and the thumbnails look very different visually, the system is still able to infer a link between these two related videos, including the compilation aspect of both. The last example (Figure 2c), the seed is a lecture about numerical analysis and the top recommendation is also a lecture (visually similar) but on a different topic. This shows the limitation of the current approach, where visually similar videos will be mapped close to each other (since their input representations will be close), even though they may be semantically different. Still, we believe that content-based video relatedness can be a useful signal to complement traditional CF recommendation systems and an effective way to deal with the cold-start recommendation problem.

(a) Cat videos in different languages (b) K-pop music video compilations (c) Lectures on a blackboard

Figure 2: Some examples of video seeds (left) and top recommendations (right) for them

References

[1] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.