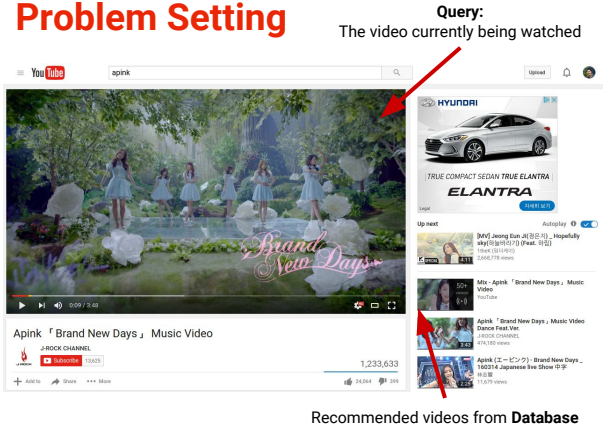
Content-based Related Video Recommendations

Joonseok Lee, Nisarg Kothari, Paul Natsev, Sami Abu-el-Haija, Jiang Wang (Google Research)

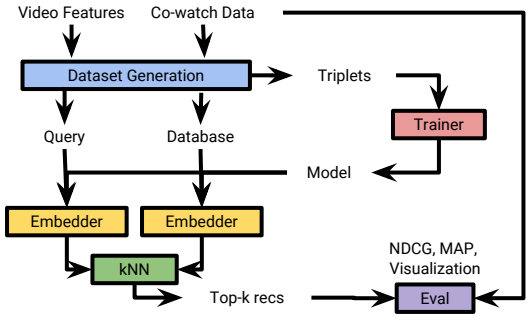
Motivation

- Collaborative filtering (CF) based recommendation performs well only if the target user and item have established co-watch history.
- For **cold-start** cases, CF may not be applied; no recommendations can be generated.
- **Content signals** can be useful for cold-start, assuming that we can learn user preference using content signals for recommendation.
- For videos, **visual signals** learned from deep neural network are very informative (in addition to text meta-data).

Problem Setting

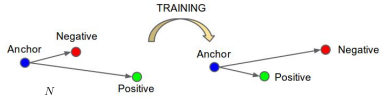


Method



- Training data are organized as **triplets**:
  { *anchor*, *positive*, *negative* }

- **Triplet loss[1]** training: make *anchor* closer to *positive* than *negative*.



$$L = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

- Video-level visual representation is the average of the visual embedding vectors (e.g, from an Inception model) from sampled frames.
- Starting from video-level visual representation , we train a **DNN** embedder to minimize triplet loss.

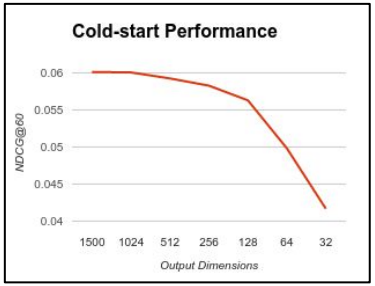
YouTube-8M Dataset [2]

- A recent, **large-scale labeled video dataset** that consists of 8 million YouTube videos.
- Comes with precomputed state-of-the-art vision **features** from billions of frames, and associated labels of 4,800 visual entities.

https://research.google.com/youtube8m/

Offline Experiments

Output dim	NDCG
1500	6.01%
1024	6.00%
512	**5.92%**
256	5.82%
128	5.63%
64	4.98%
32	3.94%



References

[1] F. Schroff, D. Kalenichenko, J. Philbin. **FaceNet: A Unified Embedding for Face Recognition and Clustering**, CVPR 2015.
[2] S. Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan. **YouTube-8M: A Large-Scale Video Classification Benchmark**, ArXiv Report arXiv:1609.08675, 2016.