Video Stream Retrieval of Unseen Queries using Semantic Memory

Spencer Cappallo
cappallo@uva.nl
Thomas Mensink
tmensink@uva.nl
Cees G. M. Snoek
cgmsnoek@uva.nl

Institute of Informatics
University of Amsterdam
Science Park 904
Amsterdam
The Netherlands

Search among live, user-broadcast videos is an under addressed and increasingly relevant challenge. Every day, more content is shared via services like Meerkat, Periscope, and Twitch. As streaming video becomes more prevalent, it is necessary to develop retrieval systems that can address the unique consequences of live video. In contrast to pre-recorded videos, live streams frequently are transmitted without any accompanying textual description. The nature of streaming video means that even if text is available, there is no guarantee that it will adequately describe the content of a live broadcast. For this reason, the video content itself must be used. The full range of possible future search queries is unknowable, which motivates the framing of stream retrieval as a no-example retrieval problem, where visual examples of a query are assumed to be unavailable beforehand.

We adapt existing approaches from the zero shot classification community, and rely on a word2vec semantic embedding to relate textual queries to pre-trained visual classifier confidence scores [2]. For a given query q, we score a stream with

$$\text{score}(q, x_t) = s(q)^\intercal \phi(x_t)$$

where x_t is the softmax scores of a deep neural network across some set of pre-trained classifiers C, $s(q)$ denotes the semantic similarity between q and C in the semantic embedding, and $\phi(x_t)$ encodes the classifier scores in a sparse manner.

Traditional video tasks assume the *whole* video is available for, a luxury that is not possible in a streaming setting. Also, especially in longer streams, content can change significantly and abruptly throughout the stream. A stream retrieval approach must provide up-to-date representations of the stream content. We explore three ways to emphasize only recent stream content. Two of these methods, Mean Memory Pooling and Max Memory Pooling, perform pooling over a fixed window from the past into the present. We introduce Memory Welling,

$$w(x_t) = \max\left(\frac{m-1}{m} w(x_{t-1}) + \frac{1}{m} x_t - \beta, 0\right)$$

where the current value of the well, w, is built on its previous state, diminished by a memory parameter m, and a constant leaking term β. Memory Wells emphasize recent, reliable content.

We test our approach and competitive baselines on the ActivityNet data set and a motivated subset of the FCVID data set. We synthesize two additional data sets of longer videos through concatanation of random videos. Two tasks are identified and targeted: Instantaneous Retrieval of relevant video streams at one moment, and Continuous Retrieval of streams relevant to a query over a long viewing session. Scoring metrics for both tasks are developed, and videos are evaluated in a simulated streaming setting. To test responsiveness to unseen queries, the test set queries are disjoint from the validation set.

In both target tasks, and on all data sets, either Memory Welling or an adaptation, Max Memory Welling, performs the strongest. We further validate our approach through comparison to state of the art on a traditional, non-streaming video task. Max Memory Welling demonstrates improvement over [1] on zero-shot event retrieval within the TRECVID MED 13 data set, using the setting described in [1].

[1] M. Jain, J. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, 2015.

[2] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *ICLR*, 2014.