Future-Supervised Retrieval of Unseen Queries for Live Video

Spencer Cappallo
University of Amsterdam
cappallo@uva.nl

Cees G. M. Snoek
University of Amsterdam
cgmsnoek@uva.nl

ABSTRACT

Live streaming video presents new challenges for retrieval and content understanding. Its live nature means that video representations should be relevant to current content, and not necessarily to past content. We investigate retrieval of previously unseen queries for live video content. Drawing from existing whole-video techniques, we focus on adapting image-trained semantic models to the video domain. We introduce the use of future frame representations as a supervision signal for learning temporally aware semantic representations on unlabeled video data. Additionally, we introduce an approach for broadening a query's representation within a pre-constructed semantic space, with the aim of increasing overlap between embedded visual semantics and the query semantics. We demonstrate the efficacy of these contributions for unseen query retrieval on live videos. We further explore their applicability to tasks such as no example, whole-video action classification and no-example live video action prediction, and demonstrate state of the art results.

ACM Reference format:
Spencer Cappallo and Cees G. M. Snoek. 2017. Future-Supervised Retrieval of Unseen Queries for Live Video. In *Proceedings of MM'17, October 23–27, 2017, Mountain View, CA, USA.*, , 9 pages.
DOI: https://doi.org/10.1145/3123266.3123437

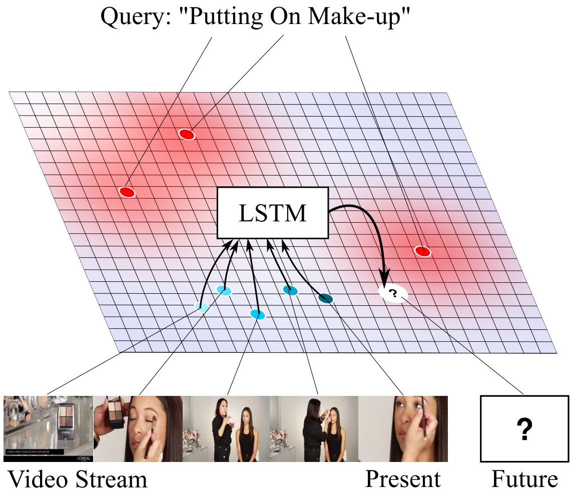

Figure 1: High quality image-trained concepts are useful, but lack temporal understanding. We enrich per-frame semantics with temporal awareness by using future representations for supervision. The model is trained within a semantic space on unlabeled video data, and can be used for unseen query retrieval across live video streams.

1 INTRODUCTION

Live, streaming video is increasingly prevalent and brings new twists on the task of video retrieval. The live nature of video streams alters the focus of understanding: with whole videos, retrieval systems seek to understand what the entire video is about [16, 20, 37], but in a streaming video scenario, current content reigns supreme. Content from the distant past may be irrelevant to the present content of the stream, and traditional pooling approaches e.g. [17, 23, 39] are ill-suited to such tasks. We address the problem of live video retrieval.

Just as the content of video streams may vary widely, so too can the possible queries. It is impossible to know beforehand what will be featured within a stream, and what users may wish to find. A hitherto unheard of natural disaster, such as "*tornado in Holland*" is unlikely to have a pre-trained classifier, but streams containing such content will absolutely be of interest should it happen. Similarly, semantic search within CCTV is a natural form of interaction for non-expert users. The attributes of interest can not be predetermined, and acquiring annotated data for classifier training is infeasible in a live scenario with hundreds of cameras. For these reasons, 'zero-example' approaches, e.g. [12, 22, 36], are almost a necessity for streaming video retrieval purposes.

Cappallo *et al.* introduce the task of live video stream retrieval in [4]. They target very long streams, and their approach consists of a hand-tuned temporal pooling over concept detections. In contrast, our approach learns semantic changes over time, and acts directly within the the semantic embedding space. We draw on the large body of existing work for no example retrieval of whole videos [12, 16, 36]. Many zero example video retrieval systems leverage annotation-rich image datasets [9] to build visual semantic understanding. In a whole-video setting, these representations can be pooled over the entirety of a video to generate video-level semantic descriptors [5, 10, 13, 15, 41]. In a stream, it's necessary for our representation to be temporally local. As our first contribution, we propose the use of future representations as a source of supervision for a recurrent neural network which acts directly within the semantic space. By using only future *representations*, our model can be trained off-line on abundant unlabeled video data, and subsequently applied to live video. An illustration of our proposed model

MM'17, October 23–27, 2017, Mountain View, CA, USA.

ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: https://doi.org/10.1145/3123266.3123437

is given in Figure 1. Past and present frames are fed into an LSTM trained to predict the semantic representation of future frames.

There is often a mismatch between the semantic space of a query and the visual semantic representation. This can be partially addressed through the use of a joint embedding [35], or can be directly learned in a supervised setting. In a no-example retrieval scenario where the system should be responsive to the broadest possible set of input queries, learning this mismatch directly becomes intractable. As a second contribution, we learn a broader mapping of the query within the semantic space, to increase the likelihood of alignment with the visual semantics.

We evaluate these two contributions against competitive baselines and explore their properties. As a third contribution we also demonstrate the abilities of the proposed approach on three tasks: continuous retrieval of video streams, no example whole video classification, and no example live action prediction.

2 RELATED WORK

2.1 Retrieval of Videos without Examples

The bulk of work on zero-example video retrieval has used some set of known video labels for knowledge transfer to unseen queries. Many submissions to the Zero Example TRECVID Multimedia Event Detection benchmark [24] have used pre-trained video concepts trained on large sets of video data [5, 7, 10, 14–16, 19, 36, 41]. For example, Wu *et al.* [36] learn weakly-supervised video concepts from web videos and their accompanying text descriptions. These concepts are combined with other concept banks in a mutual embedding space where distance to a query can be used to score video relevancy. Our work uses a semantic space to relate test queries and our visual representation, but our visual representation is built by adding temporal awareness to image concepts and requires no video labels of any kind.

One semantic space that has achieved strong success as a basis for retrieval tasks is the word2vec embedding [21]. A word2vec embedding is trained by attempting to predict the neighboring words to a given input word from some text corpus. The result is a space in which similar words lay close together. Norouzi *et al.* [22] applied a word2vec semantic space to relate pre-trained image categories to unseen class labels. Later papers adapted this approach to other domains, such as videos [12, 38]. The sidebar describes the basic approach used by [12, 22] to relate visual concepts to an unseen query. [4, 12, 38] base their methods on this approach. Our work builds on this by learning temporal qualities of visual representations within this semantic space, and proposing an alternative to broaden the mapping of a query within the semantic space.

2.2 Looking at the future

Some recent works have looked at the problem of trying to predict aspects of the future in video, e.g. [25, 33, 34, 40]. These works attempt to extrapolate future motion from a single frame, and focus on the visual representation without consideration of semantics. In contrast, our work uses the full range of past and present frames to attempt to predict a single future semantic representation. Closer to our own work, Vondrick *et al.* [32] looked at the problem of predicting future representations of video frames extracted with CNNs. Like the others, Vondrick *et al.* use only a single frame as input and are focusing on future prediction itself. The goal of our model is not future prediction, but use of the future as a source of supervision for learning our representation of the present. Vondrick *et al.* use a non-semantic representation, and learn to apply their predictions to target labels through training, while our model exists within a semantic space and can perform retrieval on unseen queries.

Word2vec [21] for No-Example Retrieval

Given some target query phrase, q, composed of N terms and a set of visual concept classifiers, $p(c|x)$ for classes $c \in C$, [22] and subsequent works relate these two inputs as follows. The query q is represented by the mean word2vec vector of its terms:

$$\omega(q) = \frac{1}{N} \sum_{w_i \in q} \omega(w_i) \tag{1}$$

where $\omega(\cdot)$ yields a word2vec representation for a constituent term $w_i \in q$. Likewise, an image is placed within the word2vec space using the average word2vec representation of high-scoring visual concepts, weighted by the their decision scores:

$$\omega(x) = \frac{1}{\sum_{c \in T} p(c|x)} \sum_{c \in T} p(c|x)\omega(c) \tag{2}$$

where T is some set of high-scoring terms to use, to avoid dilution from a long-tail of low-confidence class predictions. Once embedded, the score for an image and text term pair, (x, q), is calculated by the cosine similarity between their respective word2vec representations:

$$score(x, q) = sim(\omega(x), \omega(q)) \tag{3}$$

where $sim(\cdot, \cdot)$ returns the cosine similarity.

2.3 Retrieval of Live Video

There has been some recent work on live video retrieval [4] and the related task of live video action prediction [8, 29]. Prior work on live video action detection has been limited to supervised methods trained on labeled examples. Soomro *et al.* [29] perform both action classification and spatial localization on partially seen videos. They do not investigate temporal localization, and report on datasets with short, single-action clips. De Geest *et al.* [8] look at the problem of supervised live video action detection in television episodes, where many different actions may happen in a single video. They evaluate how well their algorithms work at predicting temporal action relevancy in a video. We apply our model to the problem of live video action prediction, but do so in a no-example setting.

Live video retrieval considers multiple concurrent video streams, unlike live action detection which is focused on a single video. Cappallo *et al.* [4] work on the task of live video stream retrieval for unseen queries. Their approach involves temporal pooling of an image-trained conceptbank, which is ultimately placed in a semantic embedding space. Our proposed approach instead works within the semantic embedding space itself. Their temporal pooling is discovered through validation on the data set and relies on labeled examples for the parameter setting. Our approach learns a temporal understanding directly from unlabeled video examples, and we demonstrate that our learned model possesses some general applicability across video data sets.

3 METHOD

A video stream retrieval algorithm can have no knowledge of the future, and the past becomes increasingly irrelevant. Unlike a whole video retrieval task, the semantics of a video stream at time 0 are of limited use for retrieval of the content at time t. Users will not be viewing earlier content, relevant or not, and its only purpose is to inform our interpretation of the present.

3.1 Future Supervision

Given a video stream s, with some semantic representation x_t at time t, we seek an improved representation $\hat{x}_t$ that exists in the same space as the x_t but which is informed by $x_0, ..., x_{t-1}$. This improved representation should capture knowledge of temporal semantics that are absent in the frame-level representation x_t. To accomplish this, we exploit abundantly available unlabeled video data to learn a model of how semantics change over time.

The future is unavailable in our targeted live video test setting, as input is restricted to present and past video. However, the future is a free and plentiful form of supervision during off-line training. We learn our representation off-line on unlabeled videos for which future frames are available. Instead of targeting some class label, our proposed model seeks to predict a future representation of the video, $x_{t+\Delta t}$, with some temporal gap Δt. For brevity, we adopt the convention that $x_{0..t} \equiv \{x_0, x_1, ..., x_{t-1}, x_t\}$, the set of frame-level semantic representations from time 0 to time t. The goal is to learn

$$\hat{x}_t = p(x_{t+\Delta t} | x_{0..t}) \tag{4}$$

To learn $\hat{x}_t$, a recurrent neural network with LSTM units is used. LSTMs work well when processing short sequences of information [11]. LSTMs have had success in off-line video tasks [1], but tend to forget distantly past inputs. This has led to temporal pooling of LSTM outputs [42]. This weakness becomes a strength for live video tasks, where we seek a representation which portrays only current and recent content.

The proposed method operates within the semantic embedding space. The model is largely independent from choices made when constructing the visual embedding, and remains adaptable to alternative embedding schemes. In this paper, the model operates within a pre-defined word2vec embedding, which has been constructed to maximize cosine similarity between similar vocabulary terms. The network seeks to minimize the cosine similarity loss:

$$L = 1 - \frac{x_{t+\Delta t}\hat{x}_t}{||x_{t+\Delta t}||||\hat{x}_t||} \tag{5}$$

It is important to stress that the goal of this approach is not predicting the future. The future is simply an available, reliable source of supervision for enriching image-trained semantics with some temporal awareness. Through operating solely within a semantic space, our model remains responsive to novel queries, rather than learning to be responsive to a particular set of training queries. There is a possibility of bias in the learned representation due to bias in the semantics of the training videos themselves, which we explore in the experiments.

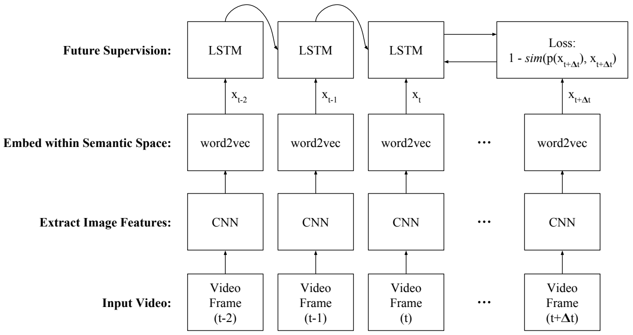

Figure 2: Overview of training pipeline for future supervision. Video frames are placed within a word2vec space, and our model enriches this representation with temporal awareness through trying to predict future representations. At test time, the learned representation can be used for live video retrieval of unseen queries.

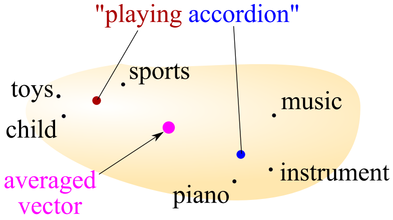

Figure 3: The mean of constituent terms might not capture the most pertinent aspects of the query. By also learning the space around individual terms, our model can capture semantics that would be overlooked.

This paper targets an off-line training/live video testing paradigm, but the model can be adapted to on-the-fly training. Instead of predicting a future frame based on current and past frames, the model can predict the current representation based on past frames, $p(x_t | x_{0..t-\Delta t})$. This adaptation could learn temporal semantics specific to a single long video stream, for tasks such as anomaly detection within surveillance footage.

3.2 Query Ambiguation

Retrieval of unseen queries generally relies on external, structured knowledge. This knowledge often comes from textual corpora where semantic relationships between terms can be identified based on co-occurrence or proximity within documents. Unfortunately, semantic gaps between visual semantics, linguistic semantics, and the semantics of a chosen model can limit performance. Indeed, even within a mutual embedding space, the manifestation of different modalities within the space may be misaligned. To alleviate this effect, we propose a method to broaden, or "ambiguate", the mapping of a query within the semantic space.

Assume some vector space Ω, which seeks to model semantics such that *leash* is closer to *dog* than to *lemonade*, and a related function $\omega(w)$ which yields the representation of word w within Ω. To place a query within this space, prior work uses the averaged representation of all words in the query, as described in section 2. Similarly, the embedding of video within Ω may be generated by a weighted combination of high-scoring concept detections. This assumes the averaged semantics of the query terms is well-aligned with the semantics of embedded video frames. We propose a method to learn a wider, more ambiguous query representation. Figure 3 illustrates the intuition behind our approach. Given a query *playing accordion*, the mean of the two terms may differ from the actual region of interest (musical instruments and performances). By covering a volume of the semantic space, rather than a point, the overlap between the query and the embedding of a relevant video may be increased.

Given a list of one or more words composing the target query, $\{w_0, ...w_l\} \in q$, we construct simulated data points within the semantic space. The simulated data consists of positively annotated examples in the space around and between class-relevant words, and negative annotated examples built with non-relevant terms. A positive data point, x_i is constructed using a set m^+ consisting of $k \leq l$ terms in or related to q:

$$x_i = \beta r_c + \frac{1}{k} \sum_{j \in m^+} \omega(j) \tag{6}$$

where r_c is a random vector perturbation within the semantic space and β is a scaling term for the perturbation. Negative examples are constructed in the same manner, using $m^- \notin q$. In a multi-class setting, other test labels can be used as negative examples in a one-vs-rest training regime. A classifier is trained on the simulated data.

3.3 Retrieval Among Video Streams

Video streams are usually present as many concurrent streams, as in the case of multiple surveillance cameras or large online streaming platforms such as Twitch. To perform retrieval of one stream from many concurrent ones, we select the highest scoring stream for the given query.

$$\text{stream} = \arg\max_{s \in S} p(q|\hat{x}_t^s) \tag{7}$$

where S is the set of all concurrent video streams, and $\hat{x}_t^s$ is therefore the future-supervised representation of stream s at time t.

3.4 Adapting to Whole-Video Tasks

The representation $\hat{x}_t$ is constructed to give a confident representation of the temporally local semantics, for live video tasks. It is possible to adapt the model for whole-video tasks by pooling its predictions over the entire video.

$$p(q|\hat{x}_{video}) = \max_t sim(q, \hat{x}_t) \tag{8}$$

where $sim(\cdot, \cdot)$ gives the cosine similarity between two vectors. The use of a max term ensures that short-term, high-confidence predictions can be exploited. These short-term predictions could be lost in a whole-video average. The motivation for query ambiguation also holds in the whole-video case.

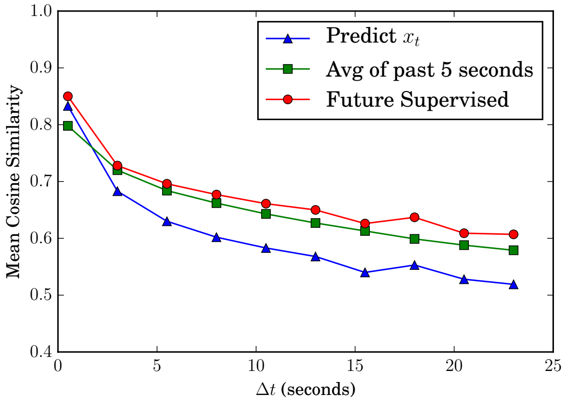

Figure 4: Performance of predicting future semantic representation. The future-supervised approach outperforms using the instantaneous representation as well as mean pooling over the previous 5 seconds.

4 EXPERIMENTS

In this section, we report on several experiments investigating the proposed approach and its efficacy, both on live video retrieval and other no-example video tasks.

4.1 Implementation Details

Image Features Frames of input videos are extracted twice per second. A convolutional neural network, based on the GoogLeNet architecture [30], creates concept scores. We use a network from [20] which generates confidence scores for 12,988 ImageNet concepts [9]. The weighted combination of the 15 highest scoring concepts are used to embed the frame within a word2vec embedding space. The threshold value of 15 was chosen following the analysis in [12] as well as preliminary experimentation.

Semantic Space All experiments use a 500-dimensional skip-gram word2vec embedding space [21]. The word2vec representation was trained on the title, description, and tags of the YFCC100M data set [31] as described in [3].

Future Supervision The future supervised model for all experiments uses a single-layer LSTM. This structure was determined in preliminary experimentation to perform well. Training of the model is performed with Keras [6] using its Theano backend.

An overview of the training pipeline can be seen in Figure 2.

4.2 Exp 1: Future Supervision

4.2.1 Performance with Δ*t.* We first investigate the effect of future supervision on our representation. Though our goal is to embed some temporal awareness into our understanding of the present, it is insightful to judge the predicted representation by its similarity to the target representation, as this is the objective the model is trained with.

Dataset We report on ActivityNet 1.2 [2]. ActivityNet consists of 4,819 training videos and 2,383 validation videos from 100 activity classes. Class annotations have temporal extent, and some videos contain multiple classes. As ActivityNet is an active challenge, the

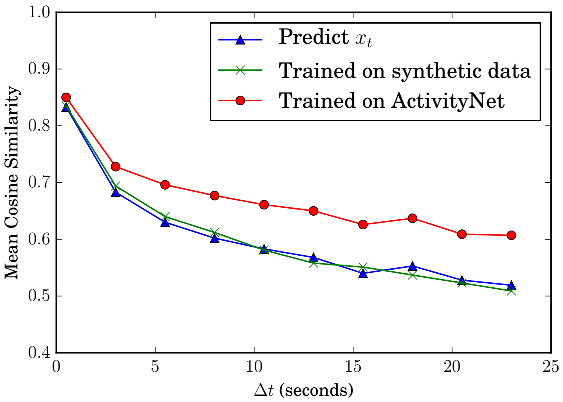

Figure 5: Performance of the model trained on synthetic data constructed to emulate only short term noisy fluctuation of semantics. Performance is similar to the baseline, suggesting that the model is learning temporal properties unique to actual video data.

labels for the official test set are not released, therefore all results for ActivityNet are reported on the validation set.

Evaluation Performance is reported in terms of Mean Cosine Similarity, which is a measure of how closely the vector directions align between the ground truth semantic representation and the predicted semantic representation.

Results In Figure 4, the performance of our trained model on predicting $x_{t+\Delta t}$ is compared against two baselines for increasing durations of Δt. The first baseline simply uses x_t as the prediction for future frames, while the second performs mean pooling over the previous 5 seconds.

At small Δt values, simply repeating the current observation works nearly as well as our model, as the difference between x_t and $x_{t+\Delta t}$ is small. As Δt increases, the two diverge. Naturally, the performance of all methods decreases, as the future grows increasingly unpredictable the further from the present you travel. The mean pooled prediction performs competitively, which suggests that much of what our model learns is how to most effectively pool relevant, recent information.

4.2.2 Training on Synthetic Stream Data. The proposed model aims to be robust to short-term changes in the semantic representation, but it is also hoped that it is learning something more nuanced about the evolution of semantics over time. To test if the model is only learning to be robust to noise, it can be trained on synthetic data representing static representations with increased noise.

Dataset To generate the synthetic data, a similar procedure to query ambiguation is followed. An individual frame of the synthetic video data is given by

$$x_t = \alpha r_g + \frac{1}{k} \sum_{i \in m} i \tag{9}$$

where r_g is a random noise vector in a normal distribution, α is a weighting parameter, and m is k randomly chosen members of some set of vectors M. M in this case is the set of 13k visual concepts, and k varies between 1 and 20. Subsequent frames use the same subset of M to maintain content cohesiveness, with a 2% chance per timestep to choose a new subset of M.

This model corresponds to a steady visual appearance with short-term noise or variation. By training on this model, the LSTM learns explicitly to focus only on semantics which are constant over time, while ignoring distracting fluctuations. If the sole benefit of the proposed model arises from this noise robustness, training on such synthetic data should yield comparable performance to training on real video data.

Evaluation As before, we report Mean Cosine Similarity.

Results Figure 5 shows the performance for a future-supervised model trained off of synthetic data. The model trained on synthetic data shows occasional improvement over the baseline, but fails to achieve similar results to models trained on actual video data. This suggests that a more nuanced model of how we *expect* video stream semantics to behave could be useful for bootstrapping our representation. The results show that our models trained on actual video data learn more than mere resistance to temporal fluctuation.

4.2.3 Generality of Future-Supervised Models. We investigate the extent to which our future-supervised representations are reflective of the training data set.

Datasets For testing the generality of the learned representation, we train and test our model on two video datasets in addition to ActivityNet: the TRECVID MED 2014 TestVal set [24], and the EVVE dataset [26]. The MED14 dataset consists of 27k web videos, which were used in the 2014 edition of the video event retrieval benchmark TRECVID. As such, some of the videos have been selected due to containing video events such as “rock climbing” or “wedding proposal”. The EVVE dataset consists of 3k web videos, selected based on YouTube queries for 13 specific events, such as “strokkur geyser” or “barcelona riots 2012”. These datasets were chosen for their real world video of topics different from those in ActivityNet, and for having a larger (MED) and smaller (EVVE) number of videos than ActivityNet.

Future supervised models are trained with varying Δt sizes on the ActivityNet, MED14, and EVVE data sets. Test sets are removed that consist of a random 5% of the videos for each respective data set. The models are trained on the remaining 95% of the data. Subsequently, the models are applied to each of the held out test sets.

Evaluation Results are again reported in Mean Cosine Similarity between the predicted future representation and the actual representation.

Results Figure 6 gives the results of the three models as a function of Δt on each of the test sets. The models trained on ActivityNet and MED'14 perform strongest on their respective test sets, but we see that they perform reasonably well on each other's data sets. For sufficiently large training sets of videos, much of what is learned appears to generalize to other videos. The EVVE-trained model is notable in its poor performance on the other test sets. As a smaller data set with less variety, it appears the model does not accumulate the knowledge necessary to generalize well.

4.3 Exp 2: Video Retrieval over Time

The quest to learn temporal semantics is useless if the representation does not help the target task of unseen query retrieval. In this experiment, we test the performance of the proposed method for

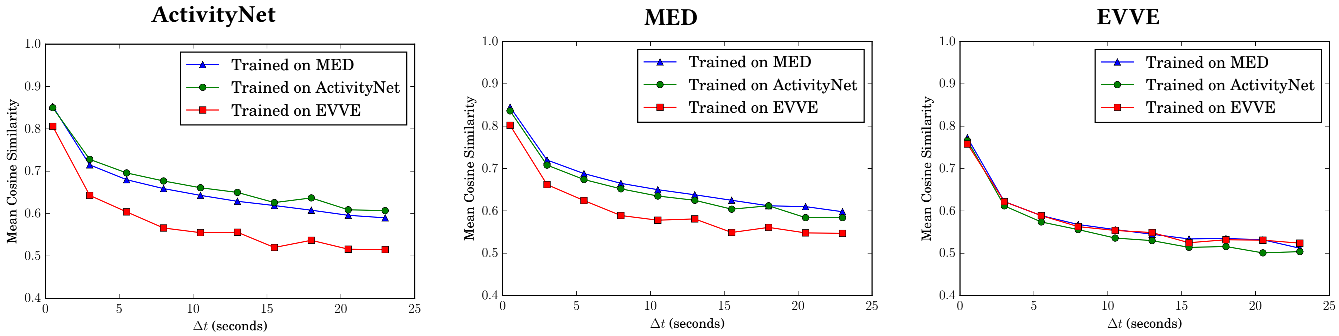

Figure 6: Cross-domain performance of future-supervised models. ActivityNet and MED trained models generalize well, while the EVVE-trained model generalizes poorly. This is likely due to the relative size and variety of the videos within the data sets.

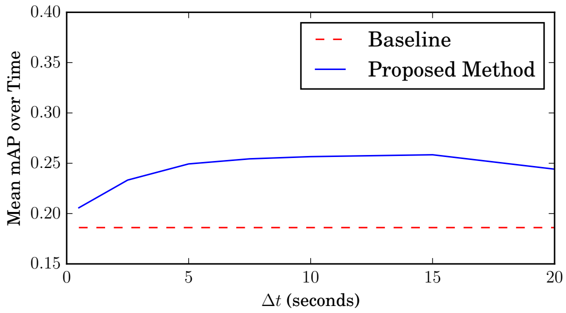

Figure 7: Retrieval performance of learned representation as a function of Δt, compared against a baseline of using only the current frame. The future supervision improves retrieval performance, with a plateau for Δt values between 5 and 15 seconds, beyond which performance begins to drop.

stream retrieval, as well as the performance of query ambiguation to improve retrieval.

Dataset The future supervised model is trained on the train set of ActivityNet [2], and results are reported on the validation set. As our model does not need training labels, we use the full 100 classes as unseen test queries.

Evaluation Live video stream retrieval requires retrieval of relevant streams among many concurrent streams. For this reason, we evaluate our test set as though it is many concurrent videos which all started simultaneously. Values are reported in Mean mAP over Time [4]. Mean mAP over Time is calculated by finding the mean average precision across all test labels at every timestep t, and subsequently averaging these mAP values across all the timesteps. It provides a measure of the retrieval performance of a system at any given time during the duration of the streams.

Results Figure 7 shows the retrieval performance of a future-supervised representation with varying values of Δt, and without query ambiguation. Unsurprisingly, small values of Δt yield little performance improvement over the baseline, as they correspond to training our model to emulate the baseline. As Δt increases, an improvement is observed, which diminishes as Δt extends into the increasingly unknowable future. The results demonstrate that future supervision is a valuable supervision channel for improving local semantic representations.

Table 1: Retrieval performance of proposed method on ActivityNet.

Method	Mean mAP over Time
Baseline	0.186
Future Supervision	0.258
Future Super. + Query Ambiguation	**0.302**

We also test the effectiveness of the query ambiguation proposed in Section 3.2. In Table 1, the mean retrieval performance across all classes is reported for future supervision with and without query ambiguation. Query ambiguation yields a considerable improvement over using only the mean word2vec representation. In Figure 8, the results are shown across all 100 test labels for the best performing Δt. The query ambiguation almost always outperforms the model using the mean vector, and yields significant improvement for some of the previously lowest performing queries. We notice that labels like "doing karate" are particularly improved, likely due to "doing" only dragging the mean away from the area of interest (the region around "karate"). Similarly, "walking the dog" is greatly improved by the ambiguation, likely because "dog" is much more discriminative than "walking" for identifying the activity.

4.4 Exp 3: Live Action Prediction

We apply our model for stream retrieval to the challenging and new task of no-example live action prediction. Live action prediction seeks to identify actions as they are happening in live video.

Dataset De Geest *et al.* [8] introduced the TVSeries dataset for the task of live action prediction. This dataset consists of the first few episodes of several popular TV shows, annotated with basic actions such as "pointing" or "picking up something". The episodes are split into training, validation, and testing datasets. We train our future supervised representation on the training set, and report results using the parameter settings for query ambiguation that yielded best results on ActivityNet.

Evaluation To compare with De Geest *et al.* , we report our results in mAP where ranking is performed along the temporal

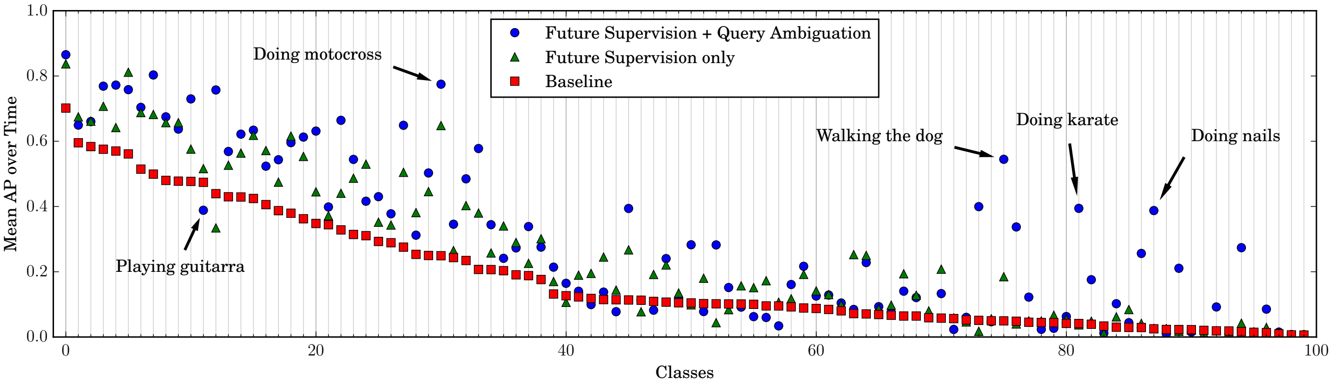

Figure 8: Per-class performance of proposed model on ActivityNet. Results are presented in to highlight relative performance of proposed method. Notably, some classes for which the baseline does poorly are significantly improved.

Table 2: Performance of no-example approach on live action prediction on the TVSeries dataset. Also reported are the results of [8]. Despite seeing only unlabeled videos, our approach surpasses a per-frame supervised method.

Method	mAP (within video)
Random	0.9%
Ours	2.2%
[8] CNN (Supervised)	1.9%
[8] LSTM (Supervised)	2.7%
[8] FV (Supervised)	5.2%

axis of a single video. This contrasts to the stream retrieval setting, where the algorithm ranks concurrent streams based on relevancy. Instead, this measures how well the algorithm can rank the temporal axis of the video in terms of relevancy.

Results In Table 2, the performance of our approach is compared to the results of De Geest *et al.* . De Geest *et al.* uses a supervised setting, where training labels are given, and reports results for three different approaches: one trained on the per-frame output of a CNN with a VGG-16 [28] architecture (CNN), one which trains an LSTM on top of the CNN features (LSTM), and one which uses Fisher vectors on top of video descriptors (FV).

Our model, despite having seen no labeled videos, manages to outperform a supervised model that has no temporal knowledge. The vague nature of the class labels in the TVSeries dataset is irrelevant with supervision, but is especially challenging for our no-example setting. Additional description of the actions could be an inexpensive way to improve performance. In Figure 9, we present an example where such additional information has been included. By extending "eat" to "eat food with utensil", we see a stronger response to an eating portion of the video.

4.5 Exp 4: Continuous Stream Retrieval

The motivating scenario for this task is a viewer who wants to watch streaming video content relevant to some query, let's say *wildlife*, over an extended duration. The ideal retrieval system not only returns relevant streams, but also does not needlessly switch among multiple relevant streams over time, as this makes a poor viewing experience.

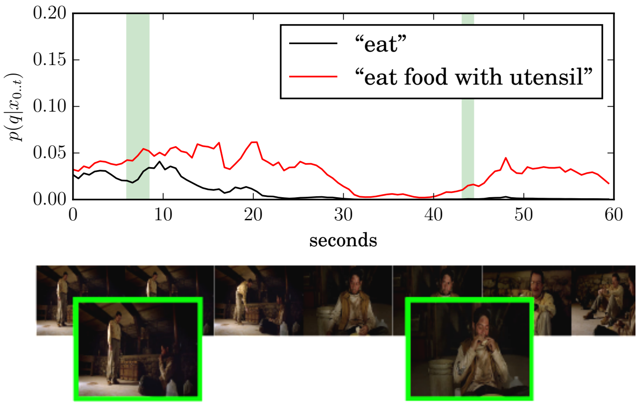

Figure 9: Possible value of added description, which fits with our setting. A longer description yields a stronger response.

Dataset AN-L is a dataset proposed in [4] which consists of ActivityNet videos concatenated into 30 minute long videos. This creates long videos in which the content changes drastically, and for which the annotations have temporal extent. The AN-L data set is composed of the 100 classes of ActivityNet, which have been divided into 40 training and validation classes, and 60 test classes.

Evaluation The Continuous Retrieval task is evaluated by the average top-1 accuracy over time for a given query, with a penalty for changing the prediction [4].

$$\text{ZP} = \frac{g_+ + r_+}{\sum_t y_t} \tag{10}$$

where g_+ is a count of every time step where the retrieved stream switches between $t-1$ and t correctly (*ie*, the new stream is relevant and the old one is no longer relevant), while r_+ is a count of every time step where the retrieved stream is both relevant and consistent with the previous time step. This metric therefore counts

Table 3: Performance of proposed method on Continuous Retrieval task on the AN-L dataset. By learning our representation, we can capture temporal dynamics which are unavailable to the pooling approach used in [4].

Method	ZP (%)
Cappallo et al. [4]	28.3
Future Supervision	31.9
Future Supervision & Query Ambiguation	**36.5**

transitions between simultaneously relevant streams as negative results, and favors both accuracy and temporal consistency in retrieval. To compare with [4], we report results in the ZP metric on their proposed AN-L dataset split, and using the same sampling frequency (2 frames per second).

Results Our proposed method's results are reported in Table 3. The future-supervised model improves over [4]'s hand-tuned temporal pooling, suggesting that the future supervised model is able to learn temporal cues that are missed by straightforward pooling. Query ambiguation offers further improvement.

4.6 Exp 5: Whole Video Classification

In this experiment, we investigate how well our model can adapt to a whole-video, no-example classification setting, using the setting laid out by Jain *et al.* [12]. To isolate the effect of our method, we also perform the method of Jain *et al.* , but using our features.

Dataset UCF Sports is a small action classification dataset consisting of 150 videos of 10 sports actions, such as "golf swing" or "diving" [27]. We report results on the test split described in [18] and used by [12].

Evaluation We report the average classification accuracy.

Results In Table 4, we report our results for whole-video classification on UCF Sports. We see that a large portion of our improvement over [12] is the result of the GoogLeNet-based CNN we use for extracting our features, instead of the AlexNet-based CNN used by [12]. We observe further improvement by the incorporation of future supervision and query ambiguation. As Jain *et al.* do not use the training set of UCF Sports, we also report results using the future supervised model that is trained on ActivityNet (and therefore has not seen any training videos from UCF Sports). In both cases, the proposed model improves performance.

In Figure 10, the performance of the proposed model is shown as a function of percentage of video seen. We report the temporally local class predictions as well as the performance over time of our whole-video modification. We present two baselines: use of the current representation x_t for class prediction, as well as using the mean representation $\frac{1}{t}\sum_{i=0}^{t} x_i$. We see that the baseline of the current representation performs erratically over time, while the averaged version is naturally steadier.

Our proposed method performs similarly to the baseline when very little of the video has been seen, but ultimately improves over the baselines. The live, temporally local predictions actually outperform the whole-video modification at times, but the two are ultimately equivalent by the end of the videos. It is likely that, due to the short length of UCF Sports videos, short-term high-confidence predictions are likely to still be captured in the final LSTM representation of the video. It is expected that for longer videos with more variation in relevancy to the class over time, the whole-video modification would improve over the learned representation at the final timestep.

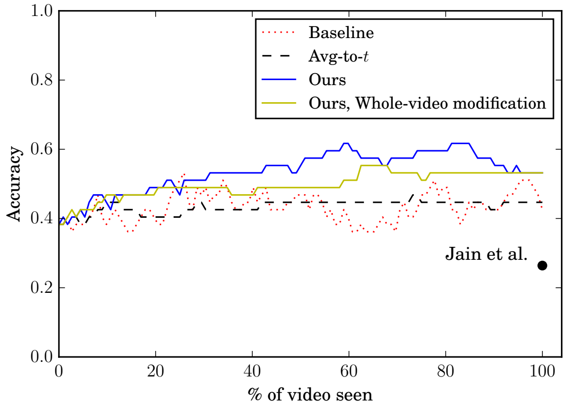

Figure 10: Classification performance on UCF Sports as a function of percentage of video seen. We present a baseline using only the current frame, as well as the average of frames from time 0 to t. Jain et al. [12] require the whole video. After only 20% of the video, the proposed model is already nearing its ultimate performance.

Table 4: Whole-video, zero-example classification on UCF Sports.

Method	Accuracy (%)
Jain et al. [12]	26.4
Jain et al. (our features)	44.6
Ours (Trained on ActivityNet)	51.1
Ours (Trained on UCF Sports)	**53.2**

5 CONCLUSION

In this paper, we introduced a new method for enriching image-based semantics with temporal awareness by exploiting future representations as a source for supervision. Future representations are a reliable source of temporal supervision because they only capture the truth of how content changes over time. Furthermore, it serves a cheap and abundant source of supervision because they can be generated for unlabeled video, which is effectively unlimited on modern video-sharing platforms. We explored the performance of this learned representation on the task of no-example live video stream retrieval, together with a query ambiguation approach for broader coverage of the semantic space. The applicability of the proposed model was demonstrated on a continuous retrieval task, live action prediction, and whole-video classification.

ACKNOWLEDGMENTS

This research is supported by the STW Story project.

REFERENCES

[1] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. 2010. Action classification in soccer videos with long short-term memory recurrent neural networks. In *ICANN*.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

[3] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. 2015. Image2emoji: Zero-shot emoji prediction for visual media. In *MM*.

[4] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. 2016. Video Stream Retrieval of Unseen Queries using Semantic Memory. In *BMVC*.

[5] Jiawei Chen, Yin Cui, Guangnan Ye, Dong Liu, and Shih-Fu Chang. 2014. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*.

[6] François Chollet. 2015. Keras. https://github.com/fchollet/keras. (2015).

[7] Yin Cui, Dong Liu, Jiawei Chen, and Shih-Fu Chang. 2014. Building A Large Concept Bank for Representing Events in Video. *arXiv:1403.7591* (2014).

[8] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. 2016. Online action detection. In *ECCV*.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[10] Amirhossein Habibian, Thomas Mensink, and Cees G. M. Snoek. 2014. Composite concept discovery for zero-shot video event detection. In *ICMR*.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] Mihir Jain, Jan C van Gemert, Thomas Mensink, and Cees GM Snoek. 2015. Objects2action: Classifying and localizing actions without any video example. In *ICCV*.

[13] Mihir Jain, Jan C. van Gemert, and Cees G. M. Snoek. 2015. What do 15,000 object categories tell us about classifying and localizing actions?. In *CVPR*.

[14] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. 2015. Self-paced curriculum learning. In *AAAI*.

[15] Lu Jiang, Teruko Mitamura, Shoou-I Yu, and Alexander G Hauptmann. 2014. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*.

[16] Lu Jiang, Shoou-I Yu, Deyu Meng, Yi Yang, Teruko Mitamura, and Alexander G Hauptmann. 2015. Fast and Accurate Content-based Semantic Search in 100M Internet Videos. In *MM*.

[17] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. 2010. Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching. In *NIST TRECVID Workshop*.

[18] Tian Lan, Yang Wang, and Greg Mori. 2011. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*.

[19] Masoud Mazloom, Amirhossein Habibian, Dong Liu, Cees G. M. Snoek, and Shih-Fu Chang. 2015. Encoding Concept Prototypes for Video Event Detection and Summarization. In *ICMR*.

[20] Pascal Mettes, Dennis C Koelma, and Cees GM Snoek. 2016. The imagenet shuffle: Reorganized pre-training for video event detection. In *ICMR*.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

[22] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. 2014. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *ICLR*.

[23] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. 2013. Action and event recognition with Fisher vectors on a compact feature set. In *ICCV*.

[24] Paul Over, Jon Fiscus, Greg Sanders, David Joy, Martial Michel, George Awad, Alan Smeaton, Wessel Kraaij, and Georges Quénot. 2014. Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*.

[25] Silvia L Pintea, Jan C van Gemert, and Arnold WM Smeulders. 2014. Déja Vu: Motion Prediction in Static Images. In *ECCV*.

[26] Jérôme Revaud, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. 2013. Event retrieval in large video collections with circulant temporal encoding. In *CVPR*.

[27] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. 2008. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*.

[28] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[29] K. Soomro, H. Idrees, and M. Shah. 2016. Predicting the where and what of actors and actions through online action localization. In *CVPR*.

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*.

[31] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[32] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Anticipating visual representations from unlabeled video. In *CVPR*.

[33] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. 2016. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*.

[34] Jacob Walker, Abhinav Gupta, and Martial Hebert. 2015. Dense optical flow prediction from a static image. In *ICCV*.

[35] Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*.

[36] Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Prem Natarajan. 2014. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*.

[37] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, and X. Xue. 2016. Multi-Stream Multi-Class Fusion of Deep Networks for Video Classification. In *MM*.

[38] Xun Xu, Timothy Hospedales, and Shaogang Gong. 2015. Semantic embedding space for zero-shot action recognition. In *ICIP*.

[39] Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2015. A discriminative CNN video representation for event detection. In *CVPR*.

[40] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. 2016. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*.

[41] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. 2015. EventNet: A Large Scale Structured Concept Library for Complex Event Detection in Video. In *MM*.

[42] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *CVPR*.