Enhancing Video Summarization via Vision-Language Embedding

Bryan A. Plummer[‡*] Matthew Brown[†] Svetlana Lazebnik[‡]
[‡]University of Illinois at Urbana Champaign [†]Google Research
{bplumme2,slazebni}@illinois.edu mtbr@google.com

Abstract

This paper addresses video summarization, or the problem of distilling a raw video into a shorter form while still capturing the original story. We show that visual representations supervised by freeform language make a good fit for this application by extending a recent submodular summarization approach [9] with representativeness and interestingness objectives computed on features from a joint vision-language embedding space. We perform an evaluation on two diverse datasets, UT Egocentric [18] and TV Episodes [45], and show that our new objectives give improved summarization ability compared to standard visual features alone. Our experiments also show that the vision-language embedding need not be trained on domain-specific data, but can be learned from standard still image vision-language datasets and transferred to video. A further benefit of our model is the ability to guide a summary using freeform text input at test time, allowing user customization.

1. Introduction

People today are producing and uploading video content at ever increasing rates. To appeal to potential viewers, videos should be well edited, containing only significant highlights while still conveying the overall story. This is especially important for video from wearable cameras, which can consist of hours of monotonous raw footage. Automatic video summarization techniques [36] can facilitate more rapid video search [40, 42] and ease the burden of editing a long video by hand [35]. Consequently, many methods for computing video summaries have been proposed by researchers [1, 5, 6, 8, 15, 20, 25, 30, 42, 48, 47].

Summarizing video typically involves a tradeoff between including segments that are interesting in their own right and those that are representative for the story as a whole. Some events may be interesting in isolation, but if they are repeated too frequently the summary may become redundant or unrepresentative. Gygli *et al*. [9], whose work we build upon, proposed an optimization approach for balancing the criteria of interestingness and representativeness. Prior work has defined these criteria in abstract mathematical terms (e.g., using notions of sparsity, graph connectedness, or statistical significance) [2, 17, 49] or tried to learn them using implicit or explicit supervision [30, 35, 43, 47]. Generally, it is agreed that bringing in explicit semantic understanding, or the ability to associate video shots with high-level categories or concepts, is helpful for enabling meaningful summaries. A number of approaches have focused on learning limited vocabularies of concepts (often in a weakly supervised manner) from large databases of images and/or video collected from the web [1, 14, 15, 41]. When rich supervision in the form of freeform language (titles, on-screen text, or closed captioning) is available, it becomes possible to use more sophisticated joint vision-language models to capture a wider range of concepts and to extract a more meaningful video summary [33]. Joint modeling of visual content and text is becoming increasingly common for video summarization and retrieval, typically to help identify whether a given shot is relevant to the overall story of a video or a particular user query [23, 32, 34, 42].

Recently, we have seen a proliferation of powerful vision-language models based on state-of-the-art feedforward and recurrent neural networks. Such models have been used for cross-modal retrieval [16, 19, 26, 29, 27, 37, 39], image caption generation [12, 13, 27, 38, 44], and visual storytelling [11, 50]. Motivated by these successful applications, we experiment with a joint image-text embedding as a representation for video summarization. Such an embedding is given by functions trained to project image and text features, which may initially have different dimensionalities, into a common latent space in which proximity between samples reflects their semantic similarity. We use the two-branch neural network of Wang *et al*. [39] to learn a nonlinear embedding using paired images and text (or video and specially produced annotations). Then, at test time, we use the embedding to compute the similarity between two video segments without requiring any language inputs. As we can see from Figure 1, even an embedding trained on

[*]Major part of work done while an intern at Google

Figure 1. Example query video frames (left column) and their best-matching still images with their captions from the Flickr30k dataset [46] (right two columns). The similarity is computed by mapping the visual features describing both the video frames and the still images into a learned vision-language space, which provides a semantically consistent representation for video summarization.

a different domain, i.e., the Flickr30k dataset of still images and captions [46], can retrieve semantically consistent results for a query video frame (e.g. images of an outdoor market are returned for the second query, or a woman sitting at a table for the third).

An overview of our system is presented in Figure 2. We start with the approach of Gygli *et al*. [9], which creates a video summary based on a mixture of submodular objectives on top of vision-only features. We augment this method, which we will refer to as Submod in the following, with a set of vision-language objectives computed in the cross-modal embedding space. The effectiveness of this approach is experimentally demonstrated on the UT Egocentric [18] and TV Episodes [45] datasets, which have different statistics and visual content. Our experiments show that the embedding can be learned on traditional vision-language datasets like Flickr30k [46] while still providing a good representation for the target video datasets. We are able to leverage this improved representation to create more compelling video summaries and, using the same underlying model, allow a user to create custom summaries guided by text input.

2. Semantically-aware video summarization

A common way of summarizing video is by selecting a sequence of segments that best represent the content found in the input clip. Following the Submod method of Gygli *et al*. [9], we formulate this selection process as optimization of a linear combination of objectives that capture different traits desired in the output summary. We chose to build on the Submod framework due to its two attractive properties. First, it is generic and easily adaptable to different summarization tasks that may have different requirements. Second, by constraining the weights in the combination to be nonnegative and the objectives to be submodular, a near-optimal solution can be found efficiently [28].

Given a video V consisting of n segments, our goal is to select the best summary $Y \subset V$ (typically subject to a budget or cardinality constraint) based on a weighted combination of visual-only objectives $\phi_o(V, Y)$ and vision-language objectives $\phi_{o'}(V, Y)$:

$$\arg\max_{Y \subset V} \underbrace{\sum_{o} w_o \phi_o(V, Y)}_{\text{Visual-Only Objectives}} + \underbrace{\sum_{o'} w_{o'} \phi_{o'}(V, Y)}_{\text{Vision-Language Objectives}} . \quad (1)$$

The weights are learned from pairs of videos and output summaries as in [9]. The objectives are restricted to being submodular and the weights to being non-negative, which makes it possible to use a greedy algorithm to obtain approximate solutions Eq. (1) with guarantees on the approximation quality.

We start with the same visual-only objectives as in the original Submod method [9], which will be reviewed in Section 2.1. The contribution of our work is in proposing new vision-language objectives, which will be introduced in Sections 2.2 and 2.3.

2.1. Visual Objectives

Submod [9] splits the subshot selection task into a mixture of three objectives enforcing representativeness, uniformity, and interestingness, as explained below.

Representativeness. A good summary needs to include all the major events of a video. To measure how well the current summary represents the original video's content, visual features are extracted from each segment and a k-medoids loss function is employed. We can think of the summary as a set of codebook centers, and for each segment from the original video represented by some feature vector f_i, we can map it onto the closest codebook center f_s, and compute the total squared reconstruction error:

$$L(V, Y) = \sum_{i=1}^{n} \min_{s \in Y} ||f_i - f_s||_2^2 . \quad (2)$$

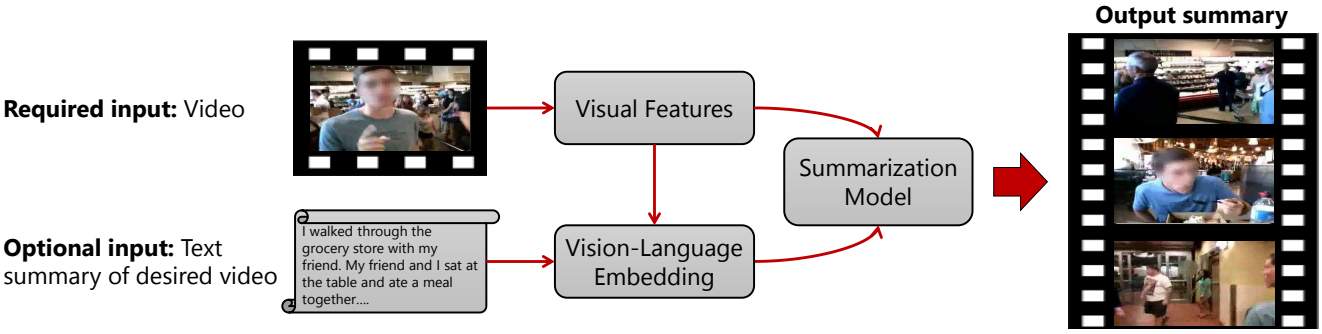

Figure 2. **Method Overview.** At test time, we assume we are given a video and, optionally, a written description of the desired summary. Our approach projects visual features into a learned vision-language embedding space where similarity reflects semantic closeness. By using this representation, we can produce more diverse and representative summaries than those created with visual features alone. The cross-modal embedding space further enables us to use text input to directly modify a summary.

This is reformulated into a submodular objective:

$$\phi_{rep}(V, Y) = L(V, \{p'\}) - L(V, Y \cup \{p'\}), \quad (3)$$

where p' represents a phantom exemplar [6], which ensures we don't take the minimum over an empty set.

As in [9], we represent a segment's visual content f_s by the average of the image features over all its frames. However, we replace the DeCAF features [4] used in [9] with more up-to-date Deep Residual Network features [10] (we use the 2048-dimensional activations before the last fully connected layer of the 152-layer ResNet trained on ImageNet [3]).

Uniformity. The second objective is designed to enforce temporal coherence, as excessively large temporal gaps between segments can interrupt the flow of the story, while segments that are too close to each other can be redundant. The uniformity objective $\phi_{unif}(V, Y)$ is completely analogous to Eq. (3), except that the feature representing each frame is simply its mean frame index (i.e., it is a scalar in this case).

Interestingness. Some segments might be preferred over others in a summary, even if they all represent the same event. For example, a segment where a child is smiling and waving at the camera might be preferred to one where they have their back to the camera. The notion of what is "interesting" is typically highly particular to the exact nature of the desired summary and/or application domain, although some generic definitions of "interestingness" have been proposed as well (e.g. [8, 18]). We use the same method as in [9] to produce a per-frame interestingness score for all the frames in a video segment. Since in principle it is possible for different segments to overlap, we sum over the interestingness scores $I(y)$ of all the unique frames y in the current summary Y:

$$\phi_{int}(V, Y) = \sum_{y \in \hat{Y}} I(y), \quad (4)$$

where $\hat{Y}$ denotes the union of all the frames in Y. In our experiments, we use this term on only one dataset, UT Egocentric [18], which has per-frame interestingness annotations that can be used for training a classifier for producing the scores $I(y)$. More details about this will be given in Section 3.

2.2. Vision-Language Objectives

We would like to project video features into a learned joint vision-language embedding space, in which we expect similarity to be more reflective of semantic closeness between different video segments. Due to its state-of-the-art performance on vision-language retrieval tasks, we chose to learn our embedding model using the two-branch network of Wang *et al*. [39]. One of the branches of this network takes in original visual features A and the other one takes in text features B. Each branch consists of two fully connected layers with ReLU nonlinearities between them, followed by L2 normalization. The network is trained with a margin-based triplet loss combining bi-directional ranking terms (for each image feature, matching text features should be closer than non-matching ones, and vice versa), and a neighborhood-preserving term (e.g. text features that correspond to the same image should be closer to each other than non-matching text features).

In this paper, we experiment with two different embeddings. The first one is trained using the dense text annotations that come with both our video datasets. However, due to the small size and vocabulary of these datasets, as well as the domain-specific nature of their descriptions, this embedding may not generalize well. Thus, we train a second embedding on the Flickr30k dataset [46], which contains 31,783 still images with five sentences each. By using Flickr30k, we can evaluate how well its representation can be transferred to video, which has quite different properties.

We train both embeddings using the code provided by the authors of [39]. On the visual side we use the same ResNet

features as in Section 2.1. On the text side, we use the same 6000-dimensional HGLMM features as in [16, 39]. The output dimensionality of the embedding space is 512.

After learning an embedding, we map our visual features to the shared semantic space and use them to compute two additional objectives we refer to as **semantic representativeness** and **semantic interestingness**. These share the forms of the visual-only versions, i.e., Eqs. (3) and (4), respectively. While one might assume that these semantic objectives should supersede their visual-only counterparts, our experiments will show that both are needed for best results. Just as semantic representativeness provides a notion of how semantically similar two video segments are, visual representativeness provides a notion of more low-level visual similarity. Ideally, a good summary will be both semantically and visually diverse so as to provide the maximum amount of information under the current budget.

2.3. Text-Guided Summarization

Including a vision-language embedding into our summarization model not only allows us to select segments that are more semantically representative and interesting, but also gives us a direct way to incorporate human input when creating a summary, as shown in Figure 2. A user can supply a freeform description of the desired summary, and the objective function can be augmented with a term that encourages the result to be consistent with this description. This is similar to the query-focused summarization framework of Sharghi et al. [32], but rather than consisting of keywords that can apply across many videos, our descriptions can be freeform sentences that are specific to the input video. We consider two scenarios corresponding to different assumptions about the form of the optional language input.

Constrained text guidance. In this version of text guidance, we assume that we are given a written description in which each sentence maps onto a single desired segment. That is, the first selected segment from the video should be consistent with the first sentence in the input description, the second segment should be consistent with the second sentence, and so on. We introduce an additional vision-language objective for Eq. (1) based on the sum of inter-modal scores between each summary segment and its corresponding sentence. More precisely, let g_s denote the feature representation of the segment s (i.e., the mean of per-frame feature vectors in the vision-language embedding space), t_s be the representation of the corresponding sentence from the description D, and $\text{sim}(g_s, t_s)$ be the cosine similarity between them. Then our new text guidance objective is given by

$$\phi_{text}(V, Y, D) = \sum_{s \in Y} \text{sim}(g_s, t_s)\,. \tag{5}$$

This is similar to what one would do for sentence-to-video retrieval, except the sentences are provided as a set and there are global costs for the summary as a whole (e.g., the uniformity and representativeness objectives). Since we assume that the sentences are given in correct temporal order, when a segment is chosen for a sentence it greatly restricts the available segments for the remaining sentences. Since our target videos have continuous shots with a lot of redundant segments, a standard retrieval approach would likely return a lot of very similar nearby segments in the top few results. The global summary-level costs are necessary to provide diversity.

Unconstrained text guidance. For videos that contain hours of footage, or in cases when a description of the desired summary cannot be written immediately after a video is shot, it may be difficult to remember the correct ordering of events or provide a temporally aligned description. In a related scenario, a user may want to summarize a video they did not shoot and maybe have not even seen – e.g., someone may want to summarize a soccer match and is particularly interested in corner kicks. For these reasons, we also implement an unconstrained version of text guidance, in which the input sentences and the associated video segments do not have to appear in the same order. This results in a bipartite matching problem between a set of candidate segments and the list of sentences which we solve using the Hungarian algorithm. After obtaining the assignments, we compute the text guidance objective using Eq. (5).

3. Experiments

3.1. Protocol and Implementation Details

Datasets. We evaluate our approach on two datasets for which detailed segment-level text annotations are available: the UT Egocentric (UTE) dataset [18] and the TV Episodes dataset [45]. The UTE dataset consists of four wearable camera videos capturing a person's daily activities. Each video is three to five hours long, for a total of over 17 hours. The TV Episodes dataset [45] consists of four videos of three different TV shows that are each 45 minutes long.

For both UTE and TV Episodes datasets, Yeung *et al*. [45] provided dense text annotations for each 5- and 10-second video segment, respectively. While the UTE dataset videos are first-person videos taken in an uncontrolled environment, the TV episodes are well edited, third-person videos. As a result of these variations, the text annotations also have some obvious differences in statistics (e.g. the UTE annotations typically begin with a self reference to the camera wearer in the first person, while the TV Episodes typically refer to people by their name in the episode).

Note that there exist other popular benchmarks for video summarization, including SumMe [8] and TVSUM [34] datasets. However, we did not include them in our evaluation as they do not have text annotations on which a vision-

language embedding model could be trained.

Training. For each video in the UTE and TV Episodes datasets, Yeung *et al*. [45] have supplied three human-composed reference text summaries. To train the weights for different objectives in the Submod method, these summaries need to be mapped to suitable subsets of segments in the videos by matching sentences from the summaries to the original per-segment video annotations. We follow the same greedy n-gram matching and ordered subshot selection procedures as previous work [9, 45] to obtain 15 training summaries for each video.

For each dataset, we use a four-fold cross-validation setup, training on each each subset of three videos and testing on the fourth one. This involves training the vision-language embedding (for models that do not use the Flickr30k-trained embedding), the interestingness function (only on the UTE dataset, as detailed in Section 3.2) and the weights in Eq. (1). For the latter step, the training data consists of 45 video-summary pairs.

Testing and evaluation. For both datasets, we set our budget (i.e., the maximum number of segments that can be selected) at 24, producing 2-minute summaries on the UTE dataset and 4-minute summaries on the TV Episodes dataset. Following [9, 32, 45], we evaluate video summarization in the text domain. At test time, given a video summary generated by our method, we create the corresponding text summary by concatenating the original text annotations of the segments that make up the summary. We use non-overlapping segments for each dataset so as to have a non-ambiguous mapping to the text annotations, though the Submod approach is still applicable to video segmentations that produce overlapping segments [9]. The automatically produced summary is compared against the three human-provided reference summaries using the recall-based ROUGE metric [22]. Note that this evaluation is content-based: multiple segments may score the same if they are associated with the same or a very similar text description regardless of their relative visual quality (e.g., a blurry segment may be considered as good as a sharper one). As in prior work [9, 32], we report the recall and f-measure on each dataset using the ROUGE-SU score, which demonstrated the strongest correlation with human judgment [45]. We use the same ROUGE parameters as in [9, 45], obtained through personal communication with the authors.

In our evaluation, we compare the following baselines and variants of our method:

1. *Sampling.* Baselines that sample segments in the testing video uniformly or randomly. We run these baselines five times each and report the mean results.

2. *Video MMR.* The approach of [21] as implemented by the authors of [45]. They provided us their output summaries on the UTE dataset only, and we evaluated them using our ROUGE settings.

3. *seqDPP.* The approach of [7] using their code. We replace their SIFT-based feature representation [24] with our ResNet features which we also use to compute the context-based representation required in this method. We concatenate these with features computed over a saliency map [31] as in [7].

4. *Submod-V.* The original Submod approach using the code of Gygli *et al*. [9] and their visual-only objectives.

5. *Submod-S.* Submod which replaces visual-only representativeness and interestingness with the semantic versions.

6. *Submod-V + Sem. Inter.* Combination of the semantic interestingness objective with the visual-only objectives.

7. *Submod-V + Sem. Rep.* Combination of the semantic representativeness objective with the visual-only objectives.

8. *Submod-V + Both.* Combination of the semantic interestingness and semantic representativeness objectives with the visual-only objectives.

Note that variants 6 and 8 above are only available on the UTE dataset since it is the only one that has an interestingness function.

3.2. UTE Dataset Results

For this dataset, Lee *et al*. [18] have provided importance annotations that can be used to train an interestingness classifier. Following [9], we learn to predict the interestingness of a video segment (as a binary label) using a support vector machine with a radial basis function kernel over our visual or vision-language features. As in [9], we compute features on the whole image rather than on regions as in [18]. For reference, the resulting classifier using the visual features has an average precision of 56.2 on the annotated frames.

We evaluate our approach on two-minute-long summaries in Table 1. Our new semantic features trained on UTE data provide a combined improvement in f-measure of nearly 5%, with a 4% improvement in recall as shown in the last line of Table 1(c). A majority of that gain comes from our semantic representativeness objective. Despite having very different text annotations on images with different statistics, the semantic features trained on the Flickr30k dataset perform nearly as well as the UTE-trained features.

Figure 4 visualizes the weights of the five objectives in our best-performing model. We can see that visual and semantic representativeness get the two highest weights,

Figure 3. Output summaries of UT Egocentric Video 2 produced with and without the semantic features, corresponding to models in the last lines of Table 1(c) and (a), respectively. Parts (a) and (b) show the first and second halves of the summary. For better readability, we add a color-coded timeline hand-annotated with high-level scenes (e.g., *Transit*, *Market*). **(a)** The first *Transit* scene is captured with the semantic features and missed otherwise. **(b)** While the two summaries represent each scene with an equal number of segments, we can see a difference in the precise segments that are selected: In the *Washing Dishes* scene, the summary based on semantic features selects segments more representative of dishwashing, rather than simply standing at the sink.

	Method	F-measure	Recall
(a)	Baselines		
	Random	26.51	25.23
	Uniform	28.13	25.76
	Video MMR [21]	22.73	20.80
	seqDPP [7]	28.87	26.83
	Submod-V [9]	29.35	27.43
(b)	Flickr30k Embedding		
	Submod-S	27.18	29.69
	Submod-V+Sem. Inter.	31.44	28.28
	Submod-V+Sem. Rep.	32.40	30.00
	Submod-V+Both	33.50	31.16
(c)	UTE Embedding		
	Submod-S	29.54	31.01
	Submod-V+Sem. Inter.	31.58	29.24
	Submod-V+Sem. Rep.	33.24	30.84
	Submod-V+Both	34.15	31.59

Table 1. **UT Egocentric summarization performance.** (a) contains our baselines including our reproduction of [7, 9] using their code with updated visual features. (b-c) demonstrates the effectiveness of our vision-language objectives on this task using embeddings trained on different datasets.

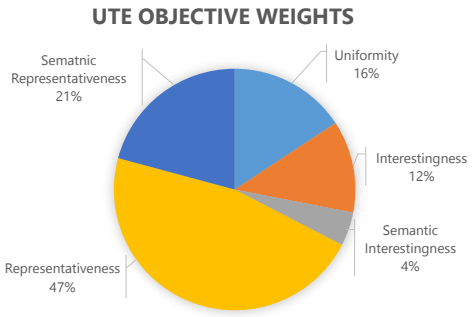

Figure 4. Learned weights for the five objectives of our best-performing model on the UT Egocentric dataset, averaged across the four training-test splits.

adding up to more than 60% of the total, followed by uniformity. The two interestingness objectives have the smallest (though still non-negligible) contribution, indicating that representativeness does most of the job of capturing story elements.

Qualitatively, the performance gains afforded by the semantic features appear to stem primarily from the addition of missing story elements. An example of this is shown in Figure 3(a), where the car drive to the market is completely

	Method	F-measure	Recall
(a)	Baselines		
	Random	32.83	28.88
	Uniform	33.90	29.15
	seqDPP [7]	35.39	32.12
	Submod-V [9]	38.18	33.47
(b)	Flickr30k Embedding		
	Submod-S	38.92	35.28
	Submod-V+Sem. Rep.	39.87	36.50
(c)	TV Episodes Embedding		
	Submod-S	37.29	32.75
	Submod-V+Sem. Rep.	40.90	37.02

Table 2. **TV Episodes summarization performance.** (a) Baselines, including our reproduction of [7, 9] using their code with updated visual features. (b,c) Different combinations of vision-language objectives using embeddings trained on Flickr30k and TV Episodes, respectively.

missing from the output summary without the semantic features. Another manifestation, although more subtle, can be seen in the *Washing Dishes* section of Figure 3(b). The segment being chosen with semantic features corresponds to an action that is common in washing dishes (rinsing a sponge), while without semantic features the user is just standing there.

3.3. TV Episodes Results

The TV Episodes dataset does not provide per-frame importance annotations with which to train a semantic interestingness classifier, so we do not use the interestingness objective of Eq. (4) here. The results in Table 2 show that augmenting the visual representativeness and uniformity objectives with the semantic representativeness objective once again provides an improvement. As can be seen from Table 2(c), semantic representativeness computed on top of the TV Episodes embedding increases the f-measure by 1.5%, and recall by just over 3%. As on the UTE dataset, the embedding trained on the dataset itself performs slightly better than the Flickr30K-trained one.

Overall, the absolute improvement in the ROUGE scores here is smaller than on UTE. In fact, of the four training-test splits, adding semantic representativeness improves results in two cases and actually makes them worse in the other two, though the absolute improvements end up being larger. We also see much higher variance in the per-objective weights learned by the Submod method on TV Episodes than on UTE. Part of the problem is the limited amount of training data. We also suspect an interestingness objective as used for the UTE dataset would help stabilize the summaries and make them more meaningful.

Figure 5 compares summaries produced with and without semantic representativeness on the fourth TV Episodes video. The result with the semantic objective more commonly agrees with the segments in the reference summary. On the left, the segment with the semantic features focuses on selecting a segment deemed more critical to the story of the original video (i.e. Joel being attacked vs. him walking around his house). For the center pair of segments, semantic representativeness selects the segment when a video of Joel's attack is shown at the police department, instead of a segment where the video is simply mentioned.

Dataset	Text Guidance	F-measure	Recall
UTE	Unconstrained	34.90	31.77
	Constrained	35.21	32.31
TV Eps.	Unconstrained	41.18	38.14
	Constrained	41.17	38.11

Table 3. Performance on text-constrained summarization, when the written description of the desired summary is given as an additional input at test time. We are using our full models with the vision-language embedding trained on the respective datasets (corresponding to the last lines of Tables 1 and 2).

3.4. Text-Guided Summarization Results

Table 3 shows the evaluation of text-guided summarization, where a reference text description is provided as an additional input at test time. These results are obtained with our full models with the vision-language embedding trained on the respective datasets. Comparing the results in Table 3 with the last lines of Tables 1 and 2, we see gains across both datasets. While one might think the constrained version, where the written description is provided in temporal order, would perform better, we only see this manifest on the UTE dataset. On the TV Episodes dataset, the two versions perform about the same. We believe this is not only due to the differences in length of the raw videos, but also the repetitive nature of the different scenes. Although the videos in the UTE dataset form a continuous stream and tend to change gradually, once a place is left is isn't often revisited. Looking at the different story elements in Figure 3(a) and Figure 3(b), only *Transit* and *Playing with LEGOs* is repeated. In contrast, the nature of the TV Episodes dataset means that the general visual elements corresponding to different sets may occur multiple times. The offices where the people work, the homes of suspects, or crime scenes (as these TV Episodes are of crime shows) are often repeated, making it challenging to identify the specific scene being described without considering the audio as well. The unconstrained model appears to be more robust to this kind of confusion.

4. Conclusion

In this paper we demonstrated that video summarization can be improved through the use of vision-language em-

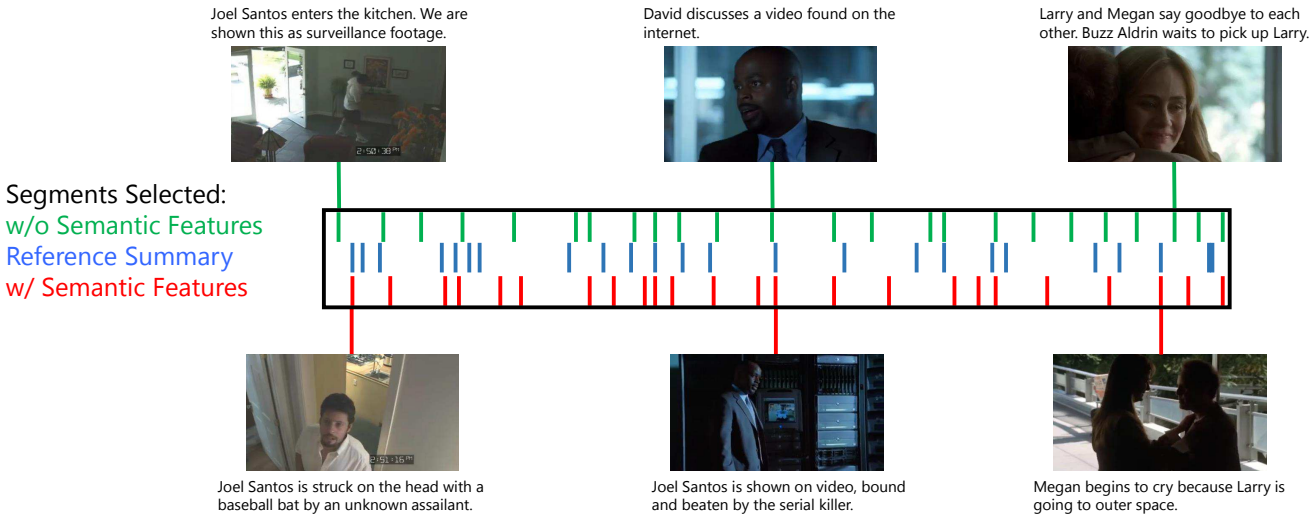

Figure 5. Comparison between the video summaries on Video 4 from the TV Episodes dataset produced with and without the semantic representativeness objective. For completeness, we also show the frames from the reference summary. The summary with semantic features more commonly selects segments found in the reference summary. The figure shows frames from three such occurrences along with the closest selected segment in the summary without semantic features.

beddings trained on image features paired with text annotations, whether from the same domain (i.e., videos of a similar type) or from a quite different one (still images with diverse content). The feature representation in the embedding space has the potential to better capture the story elements and enable users to directly guide summaries with freeform text input.

While our work shows the promise of video summarization datasets accompanied by rich text annotations, like the ones released by Yeung et al. as part of their VideoSET framework [45], it also shows their limitations. In particular, these datasets have only a few videos that can be highly variable. Thus, the amounts of training and test data are not necessarily sufficient to draw firm conclusions about the relative advantages of different summarization methods (in our case, we struggled with instability issues on the TV Episodes dataset). Compounding the problem are the inconsistencies in the kinds of annotations that are available for different datasets (in particular, annotations that can be used to train good interestingness objectives) and the evaluation methodologies that are proposed in the literature. While efforts like VideoSET are a good start, they need to be greatly expanded in scope.

Acknowledgements: We would like to thank Emily Fortuna and Aseem Agarwala for discussions and feedback on this work. This work is partially supported by the National Science Foundation under Grants CIF-1302438 and IIS-1563727, Xerox UAC, and the Sloan Foundation.

References

[1] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015. 1

[2] Y. Cong, J. Yuan, and J. Luo. Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE Transactions on Multimedia*, 14(1):66–75, 2012. 1

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013. 3

[5] M. Ellouze, N. Boujemaa, and A. M. Alimi. Im(s)2: Interactive movie summarization system. *J. Vis. Comun. Image Represent.*, 21(4):283–294, 2010. 1

[6] R. Gomes and A. Krause. Budgeted Nonparametric Learning from Data Streams. In *ICML*, 2010. 1, 3

[7] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In *NIPS*, 2014. 5, 6, 7

[8] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *ECCV*, 2014. 1, 3, 4

[9] M. Gygli, H. Grabner, and L. Van Gool. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 2015. 1, 2, 3, 5, 6, 7

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 3

[11] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, J. Devlin, A. Agrawal, R. Girshick, X. He, P. Kohli, D. Ba-

tra, L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling. In *NAACL*, 2016. 1

[12] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1

[13] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 1

[14] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013. 1

[15] G. Kim, L. Sigal, and E. P. Xing. Joint Summarization of Large-scale Collections of Web Images and Videos for Storyline Reconstruction. In *CVPR*, 2014. 1

[16] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015. 1, 4

[17] J. Kwon and K. M. Lee. A unified framework for event summarization and rare event detection. In *CVPR*, 2012. 1

[18] Y. J. Lee, J. Ghosh, , and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 1, 2, 3, 4, 5

[19] G. Lev, G. Sadeh, B. Klein, and L. Wolf. RNN fisher vectors for action recognition and image annotation. In *ECCV*, 2016. 1

[20] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu. Video summarization via transferrable structured learning. In *WWW*, 2011. 1

[21] Y. Li and B. Mérialdo. Multi-video summarization based on video-MMR. In *WIAMIS*, 2010. 5, 6

[22] C. Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 workshop. Volume 8.*, 2004. 5

[23] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*, 2015. 1

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 5

[25] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 1

[26] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015. 1

[27] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). In *ICLR*, 2015. 1

[28] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions - I. *Mathematical Programming*, 1978. 2

[29] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 2016. 1

[30] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014. 1

[31] E. Rahtu, J. Kannala, M. Salo, and J. Heikkil. Segmenting salient objects from images and videos. In *ECCV*, 2010. 5

[32] A. Sharghi, B. Gong, and M. Shah. Query-focused extractive video summarization. In *ECCV*, 2016. 1, 4, 5

[33] M. A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language. In *International Workshop on Content-Based Access of Image and Video Databases*, 1998. 1

[34] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. TVSum: Summarizing Web Videos using Titles. In *CVPR*, 2015. 1, 4

[35] M. Sun, A. Farhadi, and S. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *ECCV*, 2014. 1

[36] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1), Feb. 2007. 1

[37] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. *ICLR*, 2016. 1

[38] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1

[39] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 1, 3, 4

[40] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua. Event driven web video summarization by tag localization and key-shot identification. *IEEE Transactions on Multimedia*, 14(4):975–985, 2012. 1

[41] B. Xiong and K. Grauman. Detecting snap points in egocentric video with a web photo prior. In *ECCV*, 2014. 1

[42] B. Xiong, G. Kim, and L. Sigal. Storyline Representation of Egocentric Videos and Its Applications to Story-based Search. In *ICCV*, 2015. 1

[43] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *CVPR*, 2015. 1

[44] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1

[45] S. Yeung, A. Fathi, and L. Fei-Fei. Videoset: Video summary evaluation through text. *arXiv:1406.5824*, 2014. 1, 2, 4, 5, 8

[46] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 2, 3

[47] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: exemplar-based subset selection for video summarization. In *CVPR*, 2016. 1

[48] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *ECCV*, 2016. 1

[49] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *CVPR*, 2014. 1

[50] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *ICCV*, 2015. 1