Video Summarization by Learning Deep Side Semantic Embedding

Yitian Yuan, Tao Mei, *Senior Member, IEEE, Peng Cui,* and Wenwu Zhu, *Fellow, IEEE*

***Abstract*—With the rapid growth of video content, video summarization, which focuses on automatically selecting important and informative parts from videos, is becoming increasingly crucial. However, the problem is challenging due to its subjectiveness. Previous research, which predominantly relies on manually designed criteria or resourcefully expensive human annotations, often fails to achieve satisfying results. We observe that the side information associated with a video (e.g., surrounding text such as titles, queries, descriptions, comments, and so on) represents a kind of human-curated semantics of video content. This side information, although valuable for video summarization, is overlooked in existing approaches. In this paper, we present a novel Deep Side Semantic Embedding (DSSE) model to generate video summaries by leveraging the freely available side information. The DSSE constructs a latent subspace by correlating the hidden layers of the two uni-modal autoencoders, which embed the video frames and side information, respectively. Specifically, by interactively minimizing the semantic relevance loss and the feature reconstruction loss of the two uni-modal autoencoders, the comparable common information between video frames and side information can be more completely learned. Therefore, their semantic relevance can be more effectively measured. Finally, semantically meaningful segments are selected from videos by minimizing their distances to the side information in the constructed latent subspace. We conduct experiments on two datasets (Thumb1K and TVSum50) and demonstrate the superior performance of DSSE to several state-of-the-art approaches to video summarization.**

***Index Terms*—Video summarization, Deep learning, Side Semantics, Embedding.**

I. Introduction

TREMENDOUS popularity of video websites like YouTube, Yahoo Video, and social networks like Facebook, Google+ have stimulated massive growth of video contents over the Internet. In order to manage the growing number of videos on the web and also to extract effective information from them, more attentions have been paid to video summarization, a mechanism which aims to produce a short summary of a video, so as to give users a synthetic and useful visual abstract of video content. In general, there are two different forms of video summarization: static video summarization and dynamic video summarization. Static summary involves a set of key frames from video and there is no restriction with time and sequence issue. Dynamic summary contains a small portion of the video shots concatenated by chronological order and more like a shorter version of the original video. In addition, video thumbnail, which is the first thing a user sees when browsing or searching for videos, can be thought as a special kind of video summary at the highest level of abstraction, with only a single frame included.

(a) Video and its key frames selected by human annotator

Side semantic information

Title: How to lock your bike. The RIGHT way!
Query: Videos about lock up bikes
Description: Here, our host, Amanda, teaches YOU how to properly lock your bike. With the help of a U-Lock or a Chain lock, this can be done fairly easily.
Comment: Instead of using a cable with your D lock, just use a second lock, the thief will have to cut two locks instead of only one.

(b) Side semantic information around video

Fig. 1. An example of video from Youtube and its side semantic information. (a) is the video and five key frames selected by human annotator, (b) shows the side semantic information like video title, user query, video description, user comment of video in (a).

Video summarization is a challenging problem because of its subjectiveness — users have their own preferences over the summaries. Nevertheless, Gong et al. showed that there exists a high inter-annotator agreement of the summaries of the same video given by different evaluators [1]. It is therefore possible to select the important and informative parts from videos that can basically satisfying the majority of preferences. To solve this problem, unsupervised approaches [2]–[21] often picked frames or shots from videos with some manually designed criteria such as visual attention, representativeness and importance. However, handcrafted criteria often fail to suit diverse videos on the web. In contrast to unsupervised ones, supervised approaches [1], [22]–[27] taught the system to directly learn from human-created summaries how to select subsets, so as to meet evaluation metrics derived from human-perceived quality. Although effective sometimes, they relied on heavily human annotations which are hard to obtain.

In reality, humans are very good at summarizing information and experiences in words. After people watch a video, it is common for them to summarize the video content and share with each other by words, and what they say can be

Y. Yuan, P. Cui and W. Zhu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: yuanyt16@mails.tsinghua.edu.cn; cuip@tsinghua.edu.cn; wwzhu@tsinghua.edu.cn).

T. Mei is with Microsoft Research, Beijing 100080, China (e-mail: tmei@microsoft.com).

Corresponding author: Tao Mei.

Copyright © 2017 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

seen as the summary of video in textual form. Different people may express different contents, but the main topic and semantic meaning are still the same. Frames or shots that are most relevant to the common expression from people are really "summary worthy" since they reflect the semantic information from video and represent what people concern. It is not realistic to get textual summaries of videos directly from people, but as shown in Fig. 1, there are some side information (e.g., surrounding text such as titles, queries, descriptions, comments, and so on) associated with videos. The side information, regarded as the indirect feedback from people, represents a kind of human-curated semantics of video content. Although valuable for video summarization, the side information is often neglected by previous approaches.

In this paper, we investigate the problem of video summarization with side semantic information. Some recent works have attempted to use titles and queries associated with videos to infer the importance of video frames. Song et al. leveraged video titles to search images from the web. They hoped to get useful visual information from these web images to find representative parts from videos [28]. Liu et al. directly mapped the visual feature to a fixed textual space through a linear transformation [29], therefore video thumbnails can be selected by measuring their distances to the side information in textual space. When generating video summaries with the guidance of side formation, the critical problem is to select semantically meaningful video parts that are tightly correlated with side information. Therefore, how to effectively measure the semantic relevance between video frames and side information is essential. However, previous works [28], [29] did not give enough consideration to this point, leading to the following problems.

Firstly, there exists both common information and modality specific information in videos and their surrounding texts, whereas Liu et al. [29] ignored that the modality specific information is harmful to semantic relevance measurement. As shown in Fig. 1, even though the video and its descriptions share common information such as "bike" and "lock", there are still some characteristics that cannot be correlated. For example, "The RIGHT way!" is textual-specific information that is difficult to capture in the video. While wall and street are visual-specific information that cannot be depicted in the text. Intuitively, it's the common information, rather than modality specific information that helps us to match relevant items from two different sources. Therefore, constructing a new latent subspace where only the common information can be preserved is a better choice for semantic relevance measurement.

Secondly, the loss of common information in latent subspace will also cause performance degradation when matching video frames and side information, which is overlooked in [28]–[31]. For example, also as shown in Fig. 1, there are "bike, lock and street" depicted in a video frame, and there are "bike, lock and people" described in the video title. When embedding the video frame and title to latent subspace without any other constrain, there might be only partial finite field "bike" (or "lock") covered. This incomplete common information can lead to incomprehensive measurement of similarity. Hence, preserving more complete common information (both "bike" and "lock") in the latent subspace will make a more robust semantic relevance measurement.

To tackle the two problems mentioned above, we propose a Deep Side Semantic Embedding (DSSE) model which serves as a bridge between the diverse side semantic information and visual content. In our DSSE, two uni-modal autoencoders are used to encode the visual features of video frames and textual features of side information, respectively. By correlating the hidden layers of the two uni-modal autoencoders, we construct a latent subspace through interactively minimizing two novel loss terms, the semantic relevance loss and the feature reconstruction loss. The semantic relevance loss based on the hidden representations enables common information to be learned in the latent subspace. At the same time, the feature reconstruction loss of the two autoencoders will force common information to be preserved as much as possible. In addition, the feature reconstruction loss can also maintain the internal-similarity in visual and textual domains, which will indirectly benefit the propagation of semantic relevance between the two domains.

Moreover, we further employ a largely available click-through based video and image datasets to train a more effective DSSE model. Users predominantly tend to click on videos that are relevant to their queries when browsing videos in search engines, and thumbnails are the only visual contents that could be seen before they click on the video. The stronger the correlation between the video thumbnail and the user query, the higher the click rate. So the semantic relevance between video thumbnails and queries can be naturally indicated by the click number. In this paper, we use the $\{video\ thumbnail, query, click\ number\}$ triads generated from click-through based datasets to help to train our DSSE model. It is worth noting that we just choose query here, as it is one kind of side information. Any other side information (title, description, comment etc.) available can also be used.

By jointly integrating the semantic relevance loss and the feature reconstruction loss, and also with the help of large scale click-through training data, our DSSE model constructs a latent subspace where the semantic relevance between the video frames and side information can be more effectively measured. Finally, we generate a summary by minimizing the distances between the selected video frames and side semantic information in the latent subspace. We conduct two sets of experiments: video thumbnails selection and dynamic video summarization on two datasets Thumb1K [29] and TVSum50 [28] separately. Experimental results show that our DSSE outperforms several state-of-the-art methods in video summarization task.

II. Related Work

Conventional unsupervised video summarization methods generate summaries by leveraging handcrafted criteria based on low-level visual or motion cues [2]–[21]. The primary criteria include coverage or representativeness [7]–[11], visual quality [12], visual attention [2], [13], influence [14], [15],

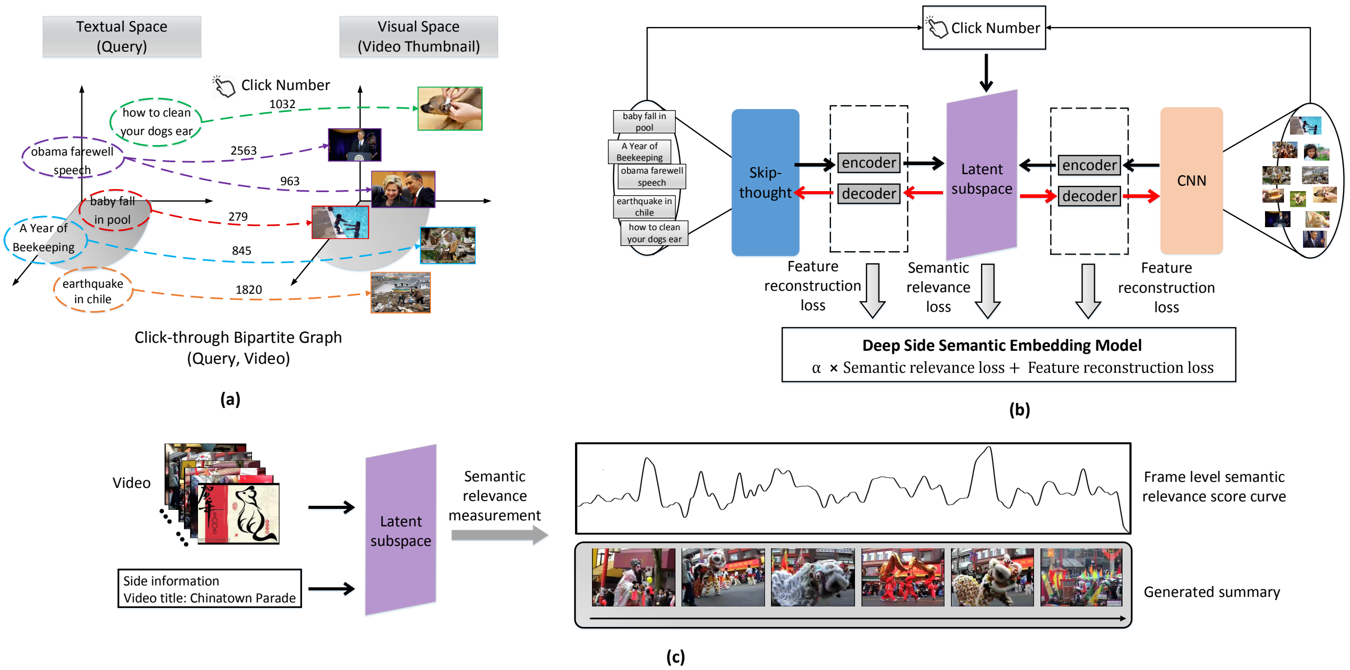

Textual Space (Query)
Visual Space (Video Thumbnail)
Click Number
how to clean your dogs ear
1032
obama farewell speech
2563
963
baby fall in pool
279
A Year of Beekeeping
845
earthquake in chile
1820
Click-through Bipartite Graph (Query, Video)
(a)
Click Number
baby fall in pool
A Year of Beekeeping
obama farewell speech
earthquake in chile
how to clean your dogs ear
Skip-thought
encoder
decoder
Latent subspace
encoder
decoder
CNN
Feature reconstruction loss
Semantic relevance loss
Feature reconstruction loss
Deep Side Semantic Embedding Model
α × Semantic relevance loss + Feature reconstruction loss
(b)
Video
Side information
Video title: Chinatown Parade
Latent subspace
Semantic relevance measurement
Frame level semantic relevance score curve
Generated summary
(c)

Fig. 2. Framework of video summarization by learning deep side semantic embedding. (a) Click-through bipartite graph between queries and videos. (b) Deep side semantic embedding model: A latent subspace is constructed by simultaneously minimizing the semantic relevance loss and feature reconstruction loss of the matched query and video thumbnail pairs. (c) Video frames and side information are embedded to the learned latent semantic subspace, then the distance in the latent subspace is directly taken as the relevance between side information and video frames. Frame level semantic relevance scores are gathered and video summary is generated by maximizing the overall semantic relevance score of the chosen parts.

[21], tracking of important object [16]–[19], [32] and so on. However, these handcrafted criteria are usually designed to deal with specific types of videos like egocentric videos or user videos, so that it is difficult to apply them to various kinds of online videos.

More recently, supervised methods which directly leverage human-edited summary examples to learn how to summarize videos have attracted much attention [1], [22]–[27]. Gong et al. proposed a supervised video summarization model, sequential determinantal point process (seqDPP), and trained seqDPP by the "oracle" summaries that agree the most among different users [1]. Based on seqDPP, Zhang et al. considered the long-short range dependencies in the sequential video frames and proposed a LSTM-based model for video summarization [25]. From another aspect, Gygli et al. generated video summaries by learning submodular functions from the user summaries [24]. One of the most important points for supervised methods is enough annotated data. People must watch the whole video and then decide if frames or shots should be included into the summary, so the annotation procedure can be very time consuming. Due to the resourcefully expensive annotation data, video summarization dataset often contains almost thousands of videos, which is far from enough to train a satisfying model so that cannot be scaled up.

Semi-supervised methods exploit some weakly supervised priors like video categories [33], domain knowledges [34], web images [35], [36] to facilitate the summarization process. While promising, these priors do not reflect the concrete contents of videos and often constrain to a limited number of object domains. Some other methods think that text associated with videos are good sources for inferring the semantic importance of video frames [28], [29], [37]. Song et al. learned canonical visual concepts shared between video frames and web images searched by video titles, and then measured the frame-level importance using the learned canonical concepts [28]. However, indirectly using video titles to grab web images will bring additional overhead for video summarization, and therefore causes a non-scalable system. Ideally, for a trained model, we hope to get its summarization result directly when meeting a new video instead of starting a learning procedure again. Liu et al. directly mapped the visual feature to a fixed textual space through a linear transformation [29], therefore the similarity between candidate video thumbnails and video query can be measured by their dot product in textual space. Candidate which is of high visual quality and is similar to the query, will be used as the final video thumbnail [29]. However, directly measuring the similarity between two different modalities in textual space will inevitably meet some textual modality interference mentioned above, which should be reduced in the semantic relevance learning procedure.

Choosing semantic meaningful parts from videos with the guidance of textual side information is quite related to several multi-modal retrieval methods [38]–[45], which aim at finding a multi-modal embedding space between image and tags/sentences so that the information in different domains can be represented in a unified subspace. Inspired by their works, we apply the subspace learning methodology in our video summarization task.

III. Deep Side Semantic Embedding Model

The basic idea of our deep side semantic embedding model is to construct a latent subspace with the ability of directly comparing side information and video frames. In this latent subspace, we hope the comparable common information between videos and side information can be more completely learned and the semantic relevance of them can be effectively measured. Therefore, we design two components in our DSSE objective function, i.e, learning semantic relevance and learning feature reconstruction. After we obtain the latent subspace, the relevance between a video frame and side information can be measured by their distance. Finally, the video frame with the highest relevance score can be seen as the thumbnail, and a dynamic video summary is generated by maximizing the total relevance score within a summary budget. The approach overview is shown in Fig. 2.

In the following, we will first construct the basic learning architecture of DSSE, and then introduce how to employ the large scale click-through data to strength our DSSE model. Finally, we will present the video summarization generation procedure. Some notations in this section are summarized in Table I.

TABLE I
NOTATIONS

Symbol	Definition
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	click-through video bipartite graph
$Q = \{q_1, q_2, \cdots, q_i \cdots\}$	set of queries in the bipartite graph
$V = \{v_1, v_2, \cdots, v_i \cdots\}$	set of videos in the bipartite graph
c_i	click number of v_i in response to q_i
$\mathbf{q}_i$	textual features of query q_i
$\mathbf{v}_i$	visual features of video v_i's thumbnail
X	a test video
T	the side information of X
$\mathbf{t}$	textual feature of T
n	the number of frames in X
$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$	the feature matrix of X
d_v	dimensionality of visual feature
d_t	dimensionality of textual feature
d_h	dimensionality of hidden layers

A. The Basic DSSE Learning Architecture

Learning semantic relevance: Given a video and its side information like query and title, the task of our method is to find a subset of frames in the video that are most relevant to its side information. Although the relevant video frames and side information are tightly correlated by the semantic meaning, the similarity in between, could not be directly computed since the representations of them are absolutely heterogenous (visual and textual). Some works directly mapped the visual feature of video frames to textual space by a linear transformation, and then measured their similarity in the textual space. However, the textual specific information which is not comparable to visual features is harmful to correlation learning. One solution, pursued in this paper, is to rely on the subspace learning, which assumes that a low-dimensional common subspace exists for the representations of video frames and side information. In this subspace, only the comparable common information between two different sources can be learned and shared, and the modality specific information is reduced.

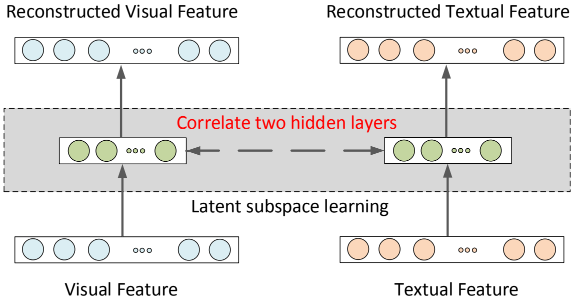

Reconstructed Visual Feature
Reconstructed Textual Feature
Correlate two hidden layers
Latent subspace learning
Visual Feature
Textual Feature

Fig. 3. The basic learning architecture of DSSE. We construct a latent subspace by correlating the hidden layers of the two uni-modal autoencoders.

Measuring the cross modal relevance between video frames and side information will result in the open problem of Semantic Gap, which is highly nonlinear in natural. Therefore we define the mappings from the visual space and textual space to the latent subspace as follows:

$$f(I_f; \mathbf{W}_f, \mathbf{b}_f) = \sigma(\mathbf{W}_f I_f + \mathbf{b}_f) \tag{1}$$

$$g(I_g; \mathbf{W}_g, \mathbf{b}_g) = \sigma(\mathbf{W}_g I_g + \mathbf{b}_g) \tag{2}$$

$I_f \in \mathbb{R}^{d_v}$ is the visual feature of video frame and $I_g \in \mathbb{R}^{d_t}$ is the textual feature of the side information. $\mathbf{W}_f \in \mathbb{R}^{d_h \times d_v}$, $\mathbf{W}_g \in \mathbb{R}^{d_h \times d_t}$ are the transformation matrices and $\mathbf{b}_f \in \mathbb{R}^{d_h}$, $\mathbf{b}_g \in \mathbb{R}^{d_h}$ are the bias vectors. d_v and d_t are the dimensions of visual features and textual features, respectively. d_h is the latent subspace dimension. In Eq. 1 and Eq. 2, we practically choose the sigmoid function $\sigma(x) = \frac{1}{1+exp(-x)}$ as the non-linear activation function.

To learn the transformation matrices and bias vectors above, we demand the matched video frame and side information to be close to each other in the latent subspace. Minimizing the distance of matched pair will force the comparable common information to be learned in the latent subspace, because the uncomparable modality specific information is almost impractical to align and will be dropped in this learning procedure. Based on the common information, the **semantic relevance loss** can be defined as the $L2$ distance between the video frame and side information in the latent subspace:

$$L_{rel}(I_f, I_g; \mathbf{W}, \mathbf{b}) = \|f(I_f; \mathbf{W}_f, \mathbf{b}_f) - g(I_g; \mathbf{W}_g, \mathbf{b}_g)\|_2^2 \tag{3}$$

We group the transformation matrices as $\mathbf{W}$ and the bias vectors as $\mathbf{b}$ here.

Learning feature reconstruction: Based on the above consideration that common information should be preserved in the latent subspace, another problem is that how much of it can be preserved? If both “bike” and “lock” are contained in the side information and video frame, but only one of them is captured in the latent subspace, then the common information we observed is insufficient. This will lead to an incomprehensive subspace and an unreliable similarity measure, resulting in the performance loss for video summarization.

To solve this problem, it's worth noting that the common information always comes from the original features, the loss of information in the latent subspace will also cause the original features cannot be well reconstructed from the latent subspace. From this point of view, minimizing the feature reconstruction loss will help us to preserve more common information in the latent subspace, and this can be naturally solved by introducing autoencoder [46] into our model.

Autoencoder is a kind of neural networks and aims to transform inputs into outputs with the least possible amount of distortion. The hidden layer of autoencoder can preserve some important characteristics of input and can be regarded as a more robust representation of input feature. Therefore, we rebuild our latent subspace on the hidden layer of autoencoder.

As shown in Fig. 3, our DSSE architecture is composed of two subnetworks, each with a uni-modal autoencoder. One autoencoder in DSSE encodes the video frame inputs, the other encodes the side information correlated to the videos. The two autoencoders share the same architecture but with different parameters. We constrain the hidden layers of the two autoencoders with the same unit number, and correlate the two layers with the semantic relevance loss term defined in Eq. 3. The transformation matrices and bias vector can be thought as the parameters of the two autoencoders at the first layer, the representations in latent subspace $f(I_f; \mathbf{W}_f, \mathbf{b}_f)$ and $g(I_g; \mathbf{W}_g, \mathbf{b}_g)$ can be thought as the hidden representations of the two autoencoders. Thus, the latent subspace is rebuilt on the hidden layers of the two correlated autoencoders.

As for the basic autoencoder, the **feature reconstruction loss** of the original input features is as follow:

$$L_{rec}(I_f, I_g; \Theta) = \left\| \widetilde{I_f} - I_f \right\|_2^2 + \left\| \widetilde{I_g} - I_g \right\|_2^2 \tag{4}$$

Here $\widetilde{I_f}$ and $\widetilde{I_g}$ are the reconstructed feature of I_f and I_g, respectively. To simplify the annotations, Θ represents all the parameters of the two correlated autoencoders.

If we only preserve partial parts of the common information in the latent subspace, the feature reconstruction loss is not optimal, because we will not reach the least possible amount of distortion without the remaining important common features captured in the hidden representations. So minimizing the feature reconstruction loss when constructing the latent subspace will help to preserve more valuable common information between two different sources. Additionally, the feature reconstruction loss can also maintain the internal-similarity within the textual and visual domains, which means that similar video frames (or side information) will have similar representations in the latent subspace, thus it indirectly benefits the propagation of semantic relevance between the two domains.

Overall loss: Combining the semantic relevance loss defined in Eq. 3 and the feature reconstruction loss defined in Eq. 4, the overall objective function of our DSSE is as follow:

$$\min_{\Theta} \alpha L_{rel}(I_f, I_g; \Theta) + L_{rec}(I_f, I_g; \Theta) \tag{5}$$

α is the parameter used to trade off between the two loss terms.

B. The click-through based DSSE learning

Based on the basic DSSE learning architecture, we further consider how to leverage the freely available click-through image and video datasets to learn a more effective DSSE model.

As shown in Fig. 2(a), a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ between the user queries and videos is constructed based on the search logs from a commercial search engine. $\mathcal{V} = Q \bigcup V$ is the set of vertices, which consists of a query set Q and a video set V. The query set Q can be thought as one kind of the side information of videos here. The number associated with an edge represents the click number of a video given a query. For most video search engines, users can only see the returned video thumbnails before clicking on a certain video, and they intend to choose the video whose thumbnail is more related to the query. Therefore, query and its relevant video thumbnail are closely bound up by their semantic meaning. Specifically, each edge and two vertices associated with it in the bipartite graph can generate a triad $\{q_i, v_i, c_i\}$, where c_i is the click number of video v_i in response to query q_i. Obviously, the larger the click number c_i, the higher the semantic relevance between q_i and v_i.

If a metric can be learned to measure the semantic relevance between different queries and video thumbnails, we can also measure the semantic relevance between video frames and their textual side information by this metric naturally. Therefore, we could use the click-through based data to strengthen the learning process of our DSSE model.

Specifically, we obtain a set of triplets $\mathcal{T}$ from our click-through bipartite graph, where each tuple $\langle q_i, v_i^+, v_i^- \rangle$ consists of a query q_i, a video thumbnail v_i^+ with higher click number c_i^+ and a lower clicked video thumbnail v_i^- with click number c_i^-. Also, we involve some thumbnails not clicked by query q_i as v_i^- in the triplets, enforcing the projections of video thumbnails with different semantics become far away in the learnt subspace. Therefore, the click-through based semantic relevance loss in DSSE is defined as:

$$\begin{aligned} &L_{rel}^*(\mathbf{q}_i, \mathbf{v}_i^+, \mathbf{v}_i^-; \Theta) = \\ &\qquad max(0, \gamma + c_i^+ L_{rel}(\mathbf{v}_i^+, \mathbf{q}_i; \Theta) - c_i^- L_{rel}(\mathbf{v}_i^-, \mathbf{q}_i; \Theta)) \end{aligned} \tag{6}$$

For v_i^- which is not clicked by query q_i, we set c_i^- as 1. The L_{rel}^* adopts the hinge rank loss form, it encourages the distance between a positive pair (q_i, v_i^+) to be smaller than the distance between a negative pair (q_i, v_i^-), and γ is the margin term. Compared with the typical hinge loss function, we further multiply the distance between video thumbnails and queries by their click numbers in L_{rel}^*, in order to strengthen the latent subspace learning. With this modification, the video thumbnails with higher click numbers will be closer to the query in the latent subspace, and therefore the model can better discriminate the irrelevant and relevant video thumbnails given a query.

The feature reconstruction loss based on the click data is defined as :

$$L_{rec}^*(\mathbf{q}_i, \mathbf{v}_i^+, \mathbf{v}_i^-; \Theta) = L_{rec}(\mathbf{v}_i^+, \mathbf{q}_i; \Theta) + L_{rec}(\mathbf{v}_i^-, \mathbf{q}_i; \Theta) \tag{7}$$

For training our DSSE model, we linearly combine L^*_{rel} and L^*_{rec} with the tradeoff parameter α, hence we get the following optimization problem:

$$\min_{\Theta} \sum_i \alpha L^*_{rel}(\mathbf{q}_i, \mathbf{v}_i^+, \mathbf{v}_i^-; \Theta) + L^*_{rec}(\mathbf{q}_i, \mathbf{v}_i^+, \mathbf{v}_i^-; \Theta) \quad (8)$$

Since all the terms in Eq. 8 are convex and smooth, we directly use gradient descent to optimize the overall objective function, which is convenient in some open source software library for machine learning like tensorflow and theano.

Because the training samples in the click-through video dataset are limited, we also use the large scale click-through image dataset to pretrain our model. The click-through image dataset is very similar to the video dataset, with only the visual data changed from video thumbnails to images. More details will be discussed in the Experiments section.

C. Video Summary Generation

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d_v \times n}$ denote the matrix of n video frames in a test video X, with each column $\mathbf{x}_i \in \mathbb{R}^{d_v}$ representing the visual feature of a frame. $\mathbf{t} \in \mathbb{R}^{d_t}$ is the textual feature of the side information T associated with video X. The side information T here can be various kinds of surrounding text of videos like query, title, description and so on.

After we get the optimized two uni-modal autoencoders of our DSSE model, we can map the frames in video X and the side information T into the learned latent subspace. Specifically, we measure the semantic relevance score between the ith video frame and the side information by computing the distance between $\mathbf{x}_i \in \mathbb{R}^{d_v}$ and $\mathbf{t} \in \mathbb{R}^{d_t}$ in the latent subspace. Smaller distance means higher relevance, that is:

$$R(\mathbf{x}_i) = 1 - \mathcal{N}(L_{rel}(\mathbf{x}_i, \mathbf{t}; \Theta)). \quad (9)$$

For simplicity, we normalize the distance between 0∼1 using min-max normalization $\mathcal{N}$ and then we subtract the normalized distance from 1 as the semantic relevance score.

Video Thumbnail Selection: Video thumbnail can be seen as the most concise static video summary since it should describe the video content in a single image. So we conduct some experiments for video thumbnail selection in this paper. When we have obtained the video frame level semantic relevance scores related to the side information, we can rank the frames by their score numbers and the frame with the highest semantic relevance score can be seen as the video thumbnail.

Dynamic Video Summarization: To generate a dynamic video summary of length l, we first employ a video segmentation algorithm to get video shots, and then compute the shot-level semantic relevance scores by taking an average of the frame level semantic relevance scores within each shot. Formally, we want to solve the following optimization problem:

$$\begin{aligned} &\max_z \sum_{i=1}^m z_i R(s_i) \\ &s.t. \sum_{i=1}^m z_i |s_i| \le l. \end{aligned} \quad (10)$$

Here m is the number of shots, $z_i \in \{0, 1\}$ and $z_i = 1$ indicates that shot s_i is selected. $R(s_i)$ is the shot-level semantic relevance score of the ith shot. This maximization is a standard knapsack problem, where $R(s_i)$ is the value of an item and the length $|s_i|$ is its weight. The problem can be solved globally optimal with dynamic programming. A dynamic video summary which maximizes the overall relevance score is then created by concatenating shots with $z_i = 1$ in chronological order. Following [28], l is set as 15% of the video length.

IV. Experiments

A. Datasets

We train our model on two click-through based datasets and evaluate the performance of our model on two video summarization datasets.

Clickture [47]: We leverage two different but similar click-through datasets to train our DSSE model.

- Click-through video dataset: Click-through video dataset is collected from Bing, which consists of 0.5 million $\{query, video\ thumbnail, click\ number\}$ triples, where query is a textual word or phrase, click number is an integer no less than one indicating the total clicked number.
- Click-through image dataset: Click-through image dataset is also collected from one year click-through data of Bing. The dataset comprises of two parts, i.e. the training and development (Dev) sets, the training set consists of 23.1 million $\{query, image, click\ number\}$ triples, and there are 79926 $\langle query, image \rangle$ pairs in Dev set. The relevance of each image to query in Dev set was manually annotated on a three point ordinal scale: Excellent, Good, and Bad.

The scale of click-through video dataset is limited and not enough to train a reliable model. Therefore, we leverage the large scale click-through image dataset to pretrain our model and then fine tune on click-through video dataset.

Thumb1K [29]: Thumb1K consists of 1037 query-video pairs collected from Bing. The dataset provides almost 20 key frames as candidate thumbnails for each video, and these candidate thumbnails are extracted by a representative attributes based method [48]. All the candidate thumbnails are labeled by five different scores: Very Good (VG), Good (G), Fair (F), Bad (B), and Very Bad (VB). We apply video thumbnail selection task on this dataset. Queries associated with videos provided in this dataset can be seen as the side information.

TVSum50 [28]: TVSum50 contains 50 videos downloaded from YouTube in 10 categories defined in the TRECVid Multimedia Event Detection (MED). The dataset provides video title and an important score of 1 (not important) to 5 (very important) to each of uniform-length (2s) shots for the whole video. Frame level important scores are labeled the same as their relevant shots and there are 20 different important scores labeled by 20 different people for each video. We apply dynamic video summarization task on this dataset. Titles of videos provided in this dataset can be seen as the side information.

B. Experimental Settings

Textual and Visual Features: For both video frames and side information, we need deep neural models which are well-suited for learning semantically-meaningful representations so that the high level semantic relevance can be measured in the latent subspace instead of the correlation between low level features. Since the effectiveness of the skip-thought sentence representations [49] in image-sentence retrieval and sentence classification tasks, we employ the skip-thought vectors to represent the side information of videos. The skip-thought model is trained on 11038 books from BookCorpus dataset [50] which includes about 74 million sentences, we can use the trained model as an off-the-shelf sentence embedding method as authors have concluded in the conclusion of the paper. Specifically, the 4800-D combine-skip vectors which combine both unidirectional and bidirectional sentence representations are chosed in this paper. Inspired by the success of deep convolutional neural networks (CNN), we employ AlexNet [51] to generate image representations in this work, the feature descriptor of each image or frame is obtained by extracting the output of the fc7 layer of the AlexNet model and we init the CNN with the parameters learned on ILSVRC-2012 [52]. The textual and visual features we extracted are normalized to (0,1) domain based on the domain restriction of the activation function Sigmoid.

Shot Segmentation: To generate dynamic video summary on TVSum50 dataset, we first temporally segment a video into disjoint intervals using KTS [33], a kernel-based change point detection algorithm which is widely used in video summarization tasks [25], [33].

Implementation details: Our DSSE model is implemented based on tensorflow [53]. When training our DSSE model by batch gradient descent, we set the batch size as 128, the learning rate lr is set as 0.0001 at first and we apply exponential decay to it. The training process will terminate if the average training loss difference between two consecutive epochs is less than the threshold, and the threshold value is based on the initial loss value. The margin term γ in L_{rel}^{*} is set as 0.5. The tradeoff parameter α is set as 100 and the latent subspace dimension is set as 256 by grid search. For more detailed analysis of them, see the next section. We spend about 13 hours to train our DSSE model on an Ubuntu 16.04 server with Intel Xeon CPU E5-2650, 128 GB Memory and NVidia Tesla M40 GPU. We use GPU only for extracting deep visual features.

Baseline methods: Although our method can apply to both video thumbnail selection and dynamic video summarization, Thumb1K [29] only provides almost 20 visual representative and comprehensive candidate thumbnails without the original videos, some video summarization methods are inapplicable to this dataset since video summarization is built on the whole video content. So we compare different methods on two different tasks below.

For video thumbnail selection, we compare:

- Random Selection: The method randomly selects one image from candidate thumbnails as final video thumbnail.
- Video Representative Attributes based Method (ATTR) [48]: The method considers the visual attributes of images and selects the most visual representative video frame as thumbnail.
- VSEM-VIDEO [29]: A deep visual-semantic embedding model trained on click-through video dataset for query dependent video thumbnail selection.
- MTL-VSEM [29]: A multi-task visual-semantic embedding model trained on click-through image and video dataset for query dependent video thumbnail selection.

For dynamic video summarization, we compare:

- Random Sampling: The method generates a summary by randomly selecting shots from videos such that the summary length is within the length budget l.
- Dictionary Selection based Video Summarization (DSVS) [7]: The method formulates video summarization as a dictionary selection problem using sparsity consistency, where a dictionary of key frames is selected such that the original video can be best reconstructed from this representative dictionary. Each frame in a video is assigned a representative score. Shot-level representative scores are calculated by averaging frame-level representative scores within each shot. We select shots with the highest representative scores that fit in the length budget.
- Co-archetypal Analysis (CA) [28]: The method develops a co-archetypal analysis technique that learns canonical visual concepts shared between video and web images retrieved by video titles. A summary is generated by maximizing the relevance and the representativeness of selected shots to canonical visual concepts, with length budget l.
- MTL-VSEM [29]: A multi-task visual-semantic embedding model mentioned above. Similarity between a video frame and the associated user query is measured by their inner product in the common space. Shot-level similarity scores are calculated by averaging frame-level similarity scores within each shot. A summary is generated by maximizing the overall similarity between the selected shots and the associated user query, and with length budget l.

Evaluation Metrics: We evaluate video thumbnail selection by two criteria: HIT@1 which computes the hit ratio for the first selected thumbnail and Mean Average Precision (MAP) which computes the mean precision for all the candidate thumbnails. The MAP is computed by

$$MAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}Precision\left(R_{jk}\right). \tag{11}$$

Where query set is defined as Q, for the jth query-video pair, there are m_j positive thumbnails, $Precision\left(R_{jk}\right)$ is the average precision at the position of returned kth positive thumbnails. Since video thumbnails are labeled by five different scores, we can calculate the HIT@1 and MAP in two different situations: set thumbnails with VG score as positive samples and set thumbnails with VG or G score as positive samples.

Following by [25], [28], we measure the quality of generated dynamic video summary by multiple human annotations. Specifically, given a proposed summary S and a user summary B_i provided by the ith annotator, we compute the precision p_i and recall r_i, according to the temporal overlap between the two. Then the pairwise F1 measure is computed as

$$F1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2p_i r_i}{p_i + r_i}. \tag{12}$$

Where N is the number of user summaries per video, and N is set as 20 in TVSum50 dataset. We evaluate video summarization results by the average F1 score of all the videos.

TABLE II
THE MAP OF DIFFERENT METHODS FOR VIDEO THUMBNAIL SELECTION

Method	Image	Video	MAP(VG)	MAP(VG&G)
Random	×	×	0.3870	0.6407
ATTR [48]	×	×	0.4852	0.7078
VSEM-VIDEO [29]	×	√	0.4729	0.7066
MTL-VSEM [29]	√	√	0.5228	0.7523
DSSE-VIDEO	×	√	0.5612	0.7557
DSSE-ALL	√	√	**0.5922***	**0.7821***
DSSE-ALL$_{glove}$	√	√	0.5763	0.7642
DSSE-ALL$_{nr}$	√	√	0.5744	0.7742
DSSE-ALL$_{click1}$	√	√	0.5541	0.7617

Notes: MAP(VG) means the MAP value when positive score equals VG; MAP(VG&G) means the MAP value when positive score equals VG and G. √(×) represents whether or not the click-through image or video dataset used in the training procedure of a method. *: Our method (DSSE) statistically significantly outperforms all other baselines (p $<$ 0.001) in pairwise t-test.

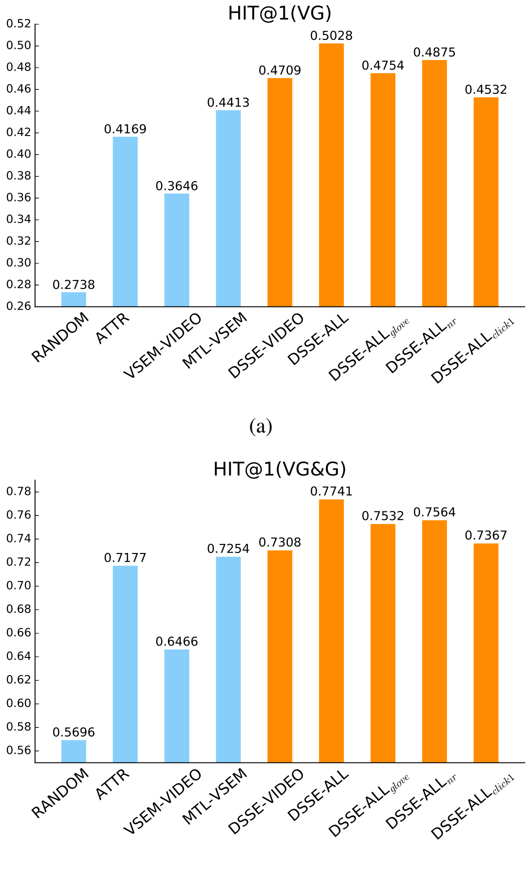
C. Evaluation of Video Thumbnail Selection

We first evaluate the performance of our DSSE method in video thumbnail selection task, Table II summaries the MAP scores, the HIT@1 results can be seen in Fig. 4.

There are some variations of our DSSE model that should be explained first. DSSE-VIDEO means we only employ the click-through video dataset to train our model, compared to it, DSSE-ALL means we pretrain our model on click-through image dataset and then fine tune on video dataset. The "√" and "×" in Table II also interpret this difference. Since our baseline method VSEM-VIDEO and MTL-VSEM applied "glove" word features to represent the video titles and queries, to be fair, we also use "glove" features in DSSE-ALL$_{glove}$ to represent words, and then average the word features as the sentence representation. In order to measure the usefulness of our feature reconstruction loss, we remove it from the overall loss in DSSE-ALL$_{nr}$ model, with only semantic relevance loss preserved. To justify the influence of the click numbers, we set them equal to 1 for all the positive query-thumbnail (and query-image) pairs in DSSE-ALL$_{click1}$.

From the results, we can find that when we pretrain on click-through image dataset and then fine tune on video dataset, our DSSE-ALL statistically significantly outperforms other methods whether on selecting one thumbnail or ranking several candidate thumbnails. The performance improvement between DSSE-VIDEO and DSSE-ALL proves that pretraining our model on click-through image dataset is beneficial. Specifically, our method achieves higher accuracy than ATTR method. It shows that compared to visual representative attributes, the semantic relevance between video thumbnails and queries measured by our DSSE model can better reflect people's concern when they watch videos, and so as to provide them a more satisfactory video thumbnail. With almost the same setting and the same training data, our DSSE-ALL$_{glove}$ outperforms MTL-VSEM at MAP(VG) by 10.2%, and HIT@1(VG) by 7.7%. It shows that constructing a new latent subspace for similarity measurement is a better choice because the harmful modality specific interference can be reduced. Compare DSSE-ALL with DSSE-ALL$_{glove}$, we find that a better sentence representation can further improve the performance of our method. Another observation is that there is a performance decrease from DSSE-ALL to DSSE-ALL$_{nr}$ in terms of all the evaluation metrics, and it verifies the effectiveness of the feature reconstruction loss. Considering the data completeness will help to preserve more useful common information in the latent subspace, and therefore benefits the semantic relevance measurement. Compared DSSE-ALL with DSSE-ALL$_{click1}$, we find that setting all the click numbers equal to 1 can cause a performance degradation, therefore the click number constrain has a great influence to our DSSE model. Besides considering the ranking relationship in different video thumbnails by hinge loss in L^{*}_{rel}, the click numbers further quantify the semantic relevance between queries and thumbnails, and can help the latent subspace learning procedure. For a trained DSSE model, it only takes 18ms on average to select a video thumbnail for a query-video pair.

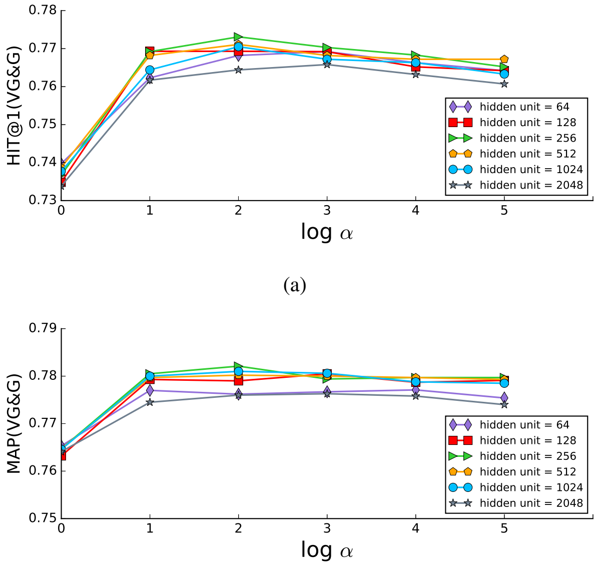
We further conduct experiments to measure the impacts of the tradeoff parameter α and the dimension of latent subspace (hidden layer unit number). The MAP and HIT@1 curves with different α and latent subspace dimensions are shown in Fig. 5. As for the tradeoff parameter α, both too small and too large values show poorer results, this is consistent with the impact of α in our DSSE model. Too small value of α overemphasizes the reconstruction loss of visual feature and ignores the semantic relevance between video thumbnails and side information, too large value reduces the influence of the feature reconstruction term. However, the curve is very smooth in a long range of α $\{10, 100, 1000, 10000\}$, thus the performance of our model is not very sensitive to the change of the tradeoff parameter. As for the dimension of latent subspace, the selection of its value do not have a great influence to our DSSE model. A smaller 64-D subspace and a larger 2048-D subspace get slightly poorer results. We infer that the semantic relevance between video frames and side information cannot be fully learned in a smaller subspace, and more training data are needed for a larger subspace. Considering both performance and run time efficiency, a moderate subspace dimension 256 or 512 is better for our DSSE model.



(a)

(b)

Fig. 4. HIT@1 score of different methods for video thumbnail selection.(a) positive score equals VG; (b) positive score equals VG and G.



(a)

(b)

Fig. 5. The HIT@1 and MAP performance curves with different tradeoff parameters α and different subspace dimensions (hidden layer unit numbers). (a) HIT1@1 performance curves when positive score equals VG and G; (b) MAP performance curves when positive score equals VG and G.

TABLE III
THE F1 SCORE OF DIFFERENT METHODS FOR VIDEO SUMMARIZATION

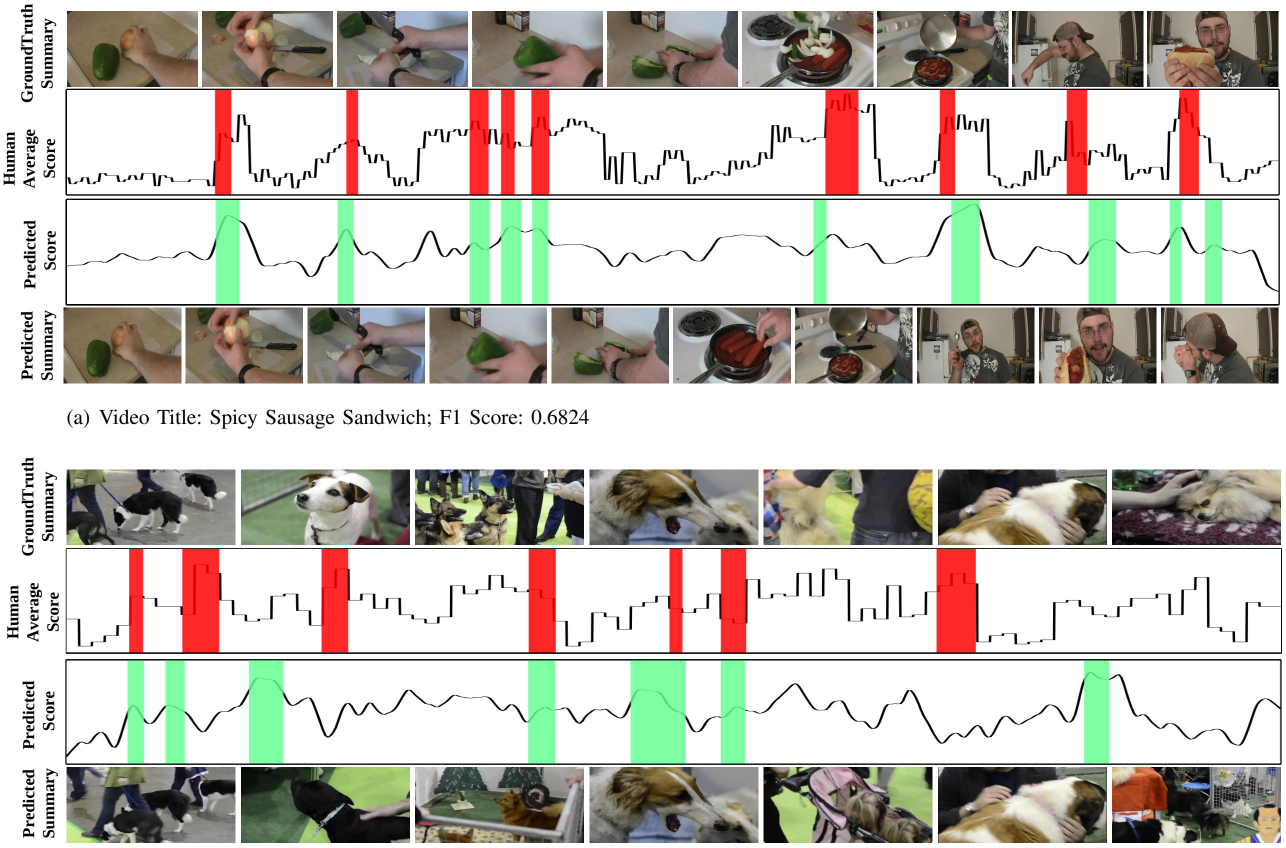
Category	Random	$CA^{\ddagger}$ [28]	DSVS [7]	MTL-VSEM [29]	DSSE
VT	0.31	0.52	0.54	**0.65**	0.63
VU	0.32	0.55	0.51	0.56	**0.60**
GA	0.35	0.41	0.53	**0.59**	0.58
MS	0.32	0.58	0.50	0.55	**0.64**
PK	0.34	0.44	0.40	0.48	**0.54**
PR	0.37	**0.53**	0.46	0.50	0.48
FM	0.32	0.51	0.47	**0.54**	0.52
BK	0.33	0.47	0.45	0.44	**0.51**
BT	0.28	0.49	0.56	0.56	**0.63**
DS	0.34	0.48	0.47	0.50	**0.55**
AVG	0.33	0.50	0.49	0.54	**0.57***

Notes: *: Our method (DSSE) statistically significantly outperforms all other baselines ($p < 0.001$) in paired t-test. ‡: CA used auxiliary grabbed web images for learning and we provide their published results here.

D. Evaluation of Dynamic Video Summarization

Table III shows the pairwise F1 scores of different methods on TVSum50 dataset. There are 10 categories in TVSum50 dataset: changing Vehicle Tire (VT), getting Vehicle Unstuck (VU), Grooming an Animal (GA), Making Sandwich (MS), Parkour (PK), Parade (PR), Flash Mob Gathering (FM), Beekeeping (BK), Attempting Bike Tricks (BT), Dog Show (DS). The performance on one category is represented in a row and the last row shows the average F1 score. Our method significantly outperforms all other methods and is particularly well on some categories like VU, MS, BT. It demonstrates that when watching these videos, users pay more attention to finding some specific content, so that the video titles can give a valuable guide to grab semantic meaningful frames or shots. However, the DSVS method which only considers the representativeness of video frames and ignores their semantic meaning is not effective to generate a good result. The CA method needs auxiliary web images related to the video titles because they need to learn co-archetypes between those web images and video frames, but they do not publish the auxiliary images for learning so that we report their published results there. To be fair, we use their evaluation code to evaluate our method. Although both of our DSSE and CA need auxiliary data for learning a video summarization model, the biggest difference is that they need to grab images from web and start a learning procedure again when meeting a new video, however, our model can directly generate the video summary in 4.8s for a 4 minutes' video.

Fig. 6 gives some qualitative results of our method. We can see that in the first example, the predicted summary aligns well with the human selection, it demonstrates that our method can find video content that most people concern with the help of video title. For the second example, we get a lower F1 score. In this case, the video always contains scenes related to dogs from the beginning to the end. Hence it is hard to choose from so many relevant shots. Even so, the generated summary is still a good depict of the video content. Some other video



(a) Video Title: Spicy Sausage Sandwich; F1 Score: 0.6824

(b) Video Title: The Dog Show; F1 Score: 0.5618

Fig. 6. Example summaries of videos from TVSum50. For each video we show the average human labeled important score curve in the second row and the predicted semantic relevance score curve based on our DSSE method in the third row. A peak in the human score curve indicates that this part is more likely to be selected by people, while a peak in the semantic relevance score curve indicates a high prediction for this part. We mark the selected segments by red and green bar on these two curves separately. Groundtruth summary based on the average human score and predicted summary based on the semantic relevance score are shown in the first and the last row respectively.

summarization results of our method are available online.

V. Conclusion and Future Work

In this work, we proposed a deep side semantic embedding model for video summarization, which aims to find semantic meaningful frames or shots of videos with the help of side semantic information. For this purpose, we construct a latent subspace by correlating the hidden layers of the two uni-modal autencoders, so that the comparable common information between video frames and side information can be learned more completely, and their semantic relevance can be measured more effectively. The large scale click-through based data also supply a massive resources to help to train a more robust model. Extensive experiments have verified the effectiveness of our method. The results demonstrate that when there are some specific content in videos that people are more purposeful to watch, side information is really a good guidance for video summarization because it can help to locate the crucial parts in videos that people concern.

Moving forward, we plan to improve our method by considering the temporal relationship between video frames, the motion features and other specific properties of videos when building the video summarization model. Other kinds of side information like video tags, captions, comments and so on will also be investigated in the future.

Acknowledgment

This paper was supported by National Natural Science Foundation of China Major Project No. U1611461 and National Program on Key Basic Research Project No. 2015CB352300.

References

[1] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, “Diverse sequential subset selection for supervised video summarization,” in *Advances in Neural Information Processing Systems*, 2014, pp. 2069–2077.

[2] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, “A generic framework of user attention model and its application in video summarization,” *IEEE transactions on multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[3] J. You, G. Liu, L. Sun, and H. Li, “A multiple visual models based perceptive analysis framework for multilevel video summarization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 273–285, 2007.

[4] A. Aner and J. R. Kender, “Video summaries through mosaic-based shot and scene clustering,” in *European Conference on Computer Vision*, 2002, pp. 388–402.

https://www.youtube.com/watch?v=Ldn8kcJ1Y-U

[5] N. Vasconcelos and A. Lippman, "A spatiotemporal motion model for video summarization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1998, pp. 361–366.

[6] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 435–441.

[7] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 66–75, 2012.

[8] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2513–2520.

[9] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan, "From keyframes to key objects: Video summarization by representative object proposal selection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1039–1048.

[10] S. E. F. De Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[11] T. Liu and J. R. Kender, "Optimization algorithms for the selection of key frame sequences of variable length," in *European Conference on Computer Vision*, 2002, pp. 403–417.

[12] T. Mei, X.-S. Hua, C.-Z. Zhu, H.-Q. Zhou, and S. Li, "Home video visual quality assessment with spatiotemporal factors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 6, pp. 699–706, 2007.

[13] N. Ejaz, I. Mehmood, and S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 34–44, 2013.

[14] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2714–2721.

[15] B. Xiong, G. Kim, and L. Sigal, "Storyline representation of egocentric videos with an applications to story-based search," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4525–4533.

[16] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1346–1353.

[17] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 982–990.

[18] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *International Journal of Computer Vision*, vol. 114, no. 1, pp. 38–55, 2015.

[19] D. Liu, G. Hua, and T. Chen, "A hierarchical visual model for video object summarization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 12, pp. 2178–2190, 2010.

[20] J. Tompkin, K. I. Kim, J. Kautz, and C. Theobalt, "Videoscapes: exploring sparse, unstructured video collections," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, p. 68, 2012.

[21] M. Tapaswi, M. Bauml, and R. Stiefelhagen, "Storygraphs: visualizing character interactions as a timeline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 827–834.

[22] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1059–1067.

[23] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *European conference on computer vision*, 2014, pp. 505–520.

[24] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3090–3098.

[25] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European Conference on Computer Vision*, 2016, pp. 766–782.

[26] A. Sharghi, B. Gong, and M. Shah, "Query-focused extractive video summarization," in *European Conference on Computer Vision*, 2016, pp. 3–19.

[27] H. Jiang, Y. Lu, and J. Xue, "Automatic soccer video event detection based on a deep neural network combined cnn and rnn," in *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on*. IEEE, 2016, pp. 490–494.

[28] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5179–5187.

[29] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3707–3715.

[30] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui, "Click-through-based cross-view learning for image search," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 717–726.

[31] T. Yao, T. Mei, and C.-W. Ngo, "Learning query and image similarities with ranking canonical correlation analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 28–36.

[32] Z. J. Zha, T. Mei, Z. Wang, and X. S. Hua, "Building a comprehensive ontology to refine video concept detection," in *ACM Sigmm International Workshop on Multimedia Information Retrieval*, 2007, pp. 227–236.

[33] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*, 2014, pp. 540–555.

[34] M. Sun, A. Farhadi, and S. Seitz, "Ranking domain-specific highlights by analyzing edited videos," in *European conference on computer vision*, 2014, pp. 787–802.

[35] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using web-image priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2698–2705.

[36] G. Kim, L. Sigal, and E. P. Xing, "Joint summarization of large-scale collections of web images and videos for storyline reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4225–4232.

[37] I. Naim, Y. C. Song, Q. Liu, H. Kautz, J. Luo, and D. Gildea, "Unsupervised alignment of natural language instructions with video segments," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 1558–1564.

[38] A. Karpathy, A. Joulin, and F. F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in neural information processing systems*, 2014, pp. 1889–1897.

[39] S. J. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *International journal of computer vision*, vol. 100, no. 2, pp. 134–153, 2012.

[40] S. Li, S. Purushotham, C. Chen, Y. Ren, and C.-C. J. Kuo, "Measuring and predicting tag importance for image retrieval," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[41] J. Lei Ba, K. Swersky, S. Fidler *et al.*, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4247–4255.

[42] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *Computer Science*, 2014.

[43] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 7–16.

[44] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, 2013, pp. 2121–2129.

[45] T. Mei, B. Yang, X. S. Hua, and S. Li, "Contextual video recommendation by multimodal relevance and user feedback," *Acm Transactions on Information Systems*, vol. 29, no. 2, pp. 1–24, 2011.

[46] A. Ng, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[47] X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li, "Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 243–252.

[48] H.-W. Kang and X.-S. Hua, "To learn representativeness of video frames," in *Proceedings of the 13th ACM international conference on Multimedia*, 2005, pp. 423–426.

[49] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in neural information processing systems*, 2015, pp. 3294–3302.

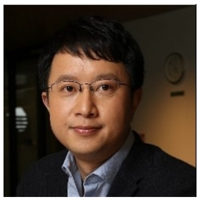
[50] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

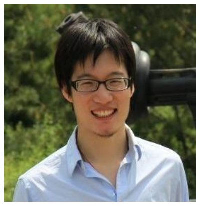
[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[53] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

Yitian Yuan is currently a Master student in the Department of Computer Science and Technology of Tsinghua University. She received her B.E. degree from the Department of Computer Science and Technology of Beijing Jiaotong University in 2016. Her main research interests include multimedia analysis, computer vision and deep learning.



Tao Mei is a Senior Researcher and Research Manager with Microsoft Research Asia. His current research interests include multimedia analysis and computer vision. He has authored or co-authored over 150 papers with 11 best paper awards. He holds 40 filed U.S. patents (with 18 granted) and has shipped a dozen inventions and technologies to Microsoft products and services. He is an Editorial Board Member of IEEE Trans. on Multimedia, ACM Trans. on Multimedia Computing, Communications, and Applications, IEEE MultiMedia Magazine, and Pattern Recognition. He is the General Co-chair of IEEE ICME 2019, the Program Co-chair of ACM Multimedia 2018, IEEE ICME 2015, and IEEE MMSP 2015. Tao is a Fellow of IAPR and a Distinguished Scientist of ACM. Tao received B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.



Peng Cui is an Associate Professor in Tsinghua University. He got his PhD degree from Tsinghua University in 2010. He is keen to promote the convergence of social media data mining and multimedia computing technologies. His research interests include network representation learning, human behavioral modeling, and social-sensed multimedia computing. He has published more than 60 papers in prestigious conferences and journals in data mining and multimedia. He is the Associate Editors of IEEE TKDE, ACM TOMM, Elsevier Journal on Neurocomputing, and Guest Editors of IEEE Intelligent Systems, Information Retrieval Journal, Machine Vision and Applications, etc.

Wenwu Zhu is currently a Professor and the Vice Chair of the Department of Computer Science, Tsinghua University, Beijing, China. Prior to his current position, he was a Senior Researcher and a Research Manager with Microsoft Research Asia, Beijing, China. He was the Chief Scientist and the Director with Intel Research China, Beijing, China, from 2004 to 2008. He was at Bell Labs, Murray Hill, NJ, USA, as a member of technical staff from 1996 to 1999. He received the Ph.D. degree from the New York University, New York, NY, USA, in 1996. His current research interests include the areas of multimedia computing, communications, and networking, as well as big data.

Wenwu is an IEEE Fellow, AAAS Fellow, an SPIE Fellow, and an ACM Distinguished Scientist. He has been serving as the Editor-in-Chief for the IEEE TRANSACTIONS ON MULTIMEDIA (T-MM) since January 1, 2017. He served on the Steering Committee for T-MM in 2016 and the IEEE TRANSACTIONS ON MOBILE COMPUTING (T-MC) from 2007 to 2010. He has served on various Editorial Boards, such as the Guest Editor for the PROCEEDINGS OF THE IEEE, the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (T-CSVT), as well as an Associate Editor for T-MM, the ACM Transactions on Multimedia, Communications, and Applications, T-CSVT, T-MC, and the IEEE TRANSACTIONS ON BIG DATA. He was the recipient of 5 Best Paper Awards including T-CSVT in 2001 and ACM Multimedia 2012.