1-2020

Neighbourhood structure preserving cross-modal embedding for video hyperlinking

Yanbin HAO

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Benoit HUET

Neighbourhood Structure Preserving Cross-Modal Embedding for Video Hyperlinking

Yanbin Hao, *Member, IEEE*, Chong-Wah Ngo, *Member, IEEE*, and Benoit Huet, *Member, IEEE*

***Abstract*—Video hyperlinking is a task aiming to enhance the accessibility of large archives, by establishing links between fragments of videos. The links model the aboutness between fragments for efficient traversal of video content. This paper addresses the problem of link construction from the perspective of cross-modal embedding. To this end, a generalized multi-modal auto-encoder is proposed. The encoder learns two embeddings from visual and speech modalities, respectively, whereas each of the embeddings performs self-modal and cross-modal translation of modalities. Furthermore, to preserve the neighbourhood structure of fragments, which is important for video hyperlinking, the auto-encoder is devised to model data distribution of fragments in a dataset. Experiments are conducted on Blip10000 dataset using the anchor fragments provided by TRECVid Video Hyperlinking (LNK) task over the years of 2016 and 2017. This paper shares the empirical insights on a number of issues in cross-modal learning, including the preservation of neighbourhood structure in embedding, model fine-tuning and issue of missing modality, for video hyperlinking.**

***Index Terms*—Video hyperlinking, cross-modal translation, structure-preserving learning.**

I. Introduction

THE ability to access and navigate within videos as efficiently as web pages by link traversal has been envisioned to improve user experience. Video hyperlinking [1], aiming to create links across video fragments, is one example of efforts towards enhancing accessibility of video archive. This research subject has recently been benchmarked in MediaEval [2] and TRECVid [1]–[4] during years 2014 to 2017. Figure 1 shows an example to motivate video hyperlinking task. This paper revisits video hyperlinking from the perspective of learning cross-modal features. We generalize the cross-modal neural networks popularly used for this problem [5]–[8]. More importantly, new insights are provided on the essential of learning structure-preserving cross-modal embedding for hyperlinking.

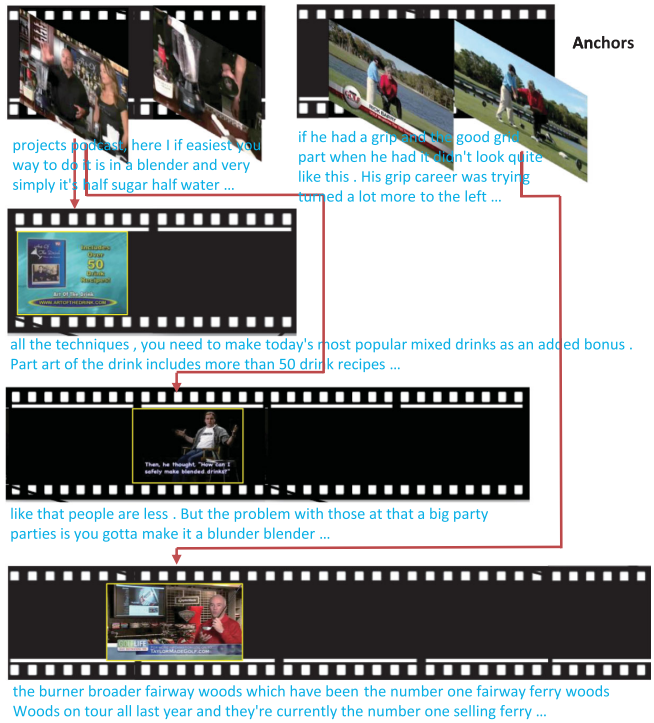


Fig. 1. Examples of video hyperlinking. The first anchor (left) showing the making of blended drinks is hyperlinked to two target fragments, where one is about blended drink recipes and the other is the trials and travails in preparing blended drinks for a Halloween party. The second anchor (right) about golf coaching is hyperlinked to a target fragment introducing fairway wood. The texts underneath the filmstrips are speech transcripts extracted from the audio stream of the video using ASR (Automatic Speech Recognition).

The two major issues of video hyperlinking are the selection of fragments as anchors (i.e., sources) and linking of anchors to target fragments (i.e., destinations) [9], [10]. This paper addresses the latter issue. Specifically, given an anchor as query, candidate targets for hyperlinking are ranked based on their "aboutness" in anchoring the multi-modal content of a query. The definition of "aboutness" is multi-perspective, such as zoom-into-details, contextually relevant, and provision of second opinion [11], [12]. This paper treats "aboutness" as anchoring of contextually relevant content from different modalities. For example, creating a hyperlink from an anchor showing pictures of London Parliament with verbal reference to the British Royal family to a target fragment with the Queen of England as main subject. In this case, despite that the delivered content by anchor and target

Manuscript received February 1, 2019; revised May 18, 2019; accepted June 5, 2019. Date of publication June 14, 2019; date of current version December 31, 2019. This work was supported in part by a grant from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China (CityU 11250716) and in part by a grant from PROCORE-France/Hong Kong Joint Research Scheme sponsored by the RGC of Hong Kong and the Consulate General of France in Hong Kong (F-CityU104/17). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Meng Wang. *(Corresponding author: Yanbin Hao.)*

Y. Hao and C.-W. Ngo are with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: haoyanbin@hotmail.com; cwngo@cs.cityu.edu.hk).

B. Huet is with the Data Science Department, EURECOM, Sophia-Antipolis 06904, France (e-mail: benoit.huet@eurecom.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2923121

can be visually and orally different, both fragments are linked for being related by a context that may trigger user interest. As such, video hyperlinking is fundamentally different from near-duplicate video retrieval [13]–[15] which focuses on detecting similar versions of fragments due to different video capturing viewpoints or audio-visual editing effects.

Video hyperlinking is generally treated as a two-stage process: query formulation and video search. The former is to mine topics of interest from an anchor as queries, and the latter is to search for targets of each topic [16]. Examples include [17] which performs hierarchical topic modeling of video archive and then represents each anchor as a vector of topics; and [18] which forms query by extracting multi-modal name entities from anchors for searching targets. Similar to conventional video retrieval algorithms, the search of targets relies mostly on early [5] or late fusion [19], [20] of multi-modal features. Along this direction, various approaches have been benchmarked in MediaEval [2] and TRECVid 2015 [1], and attained satisfactory performances. Since TRECVid 2016 [3], to further encourage the exploitation of multi-modal features, anchors are selected to contain a combination of verbal-visual information. Specifically, the selected anchors for benchmarking are accompanied either with verbal phrases like "seeing here" and "looks like" but the actual targets are not visually seen in videos, or vice versa where the objects and scenes crucial for a video are seen but not mentioned in speech track.

The strategic change in anchor selection has indeed pushed the investigation of cross-modal translation for video hyperlinking. Representative studies include variants of multi-modal deep auto-encoder networks (MmDAE) [7], [21], [22], which learns cross-modal embedding representation to translate from one modality feature to another. Video hyperlinking can thus be performed by transforming anchors to an embedding that enables cross-modal search of targets. There are two major ways of learning embedding, either learning a single embedding that projects different forms of modalities into a common space, or multiple embeddings each for a modality. Different from other cross-modal translation tasks, such as video/image captioning [23] or creation [24], metric learning [25], [26], retrieval [27]–[30] or content presentation [31], where visual and text cues are assumed fully correlated, the multi-modal features in general videos are more often complementary to each other rather than correlated. Hence, learning multiple embeddings is regarded as a more viable way of representation learning for video hyperlinking.

The main research issues studied in this paper are twofold: learning multiple embeddings to characterize the complementary nature of multiple modalities, and encapsulating local neighbourhood structure in the embeddings for robust hyperlinking. The first issue is addressed by unifying two variants of MmDAE [7], [8], [21] that learns embedding individually per modality while performing feature translation within and across modalities. The learning strategy can tolerate weakly correlated modalities, by ensuring that different modalities will not be blindly embedded into a common space. The second issue is inspired by the recent study in [32] where the preservation of local neighbourhood structure is taken into account for the learning of embedding. As studied in [10], the local data characteristics of an archive plays a critical role in hyperlinking, for example, to reduce the risk of false linking and to prevent linking to redundant targets. Therefore, a robust way of hyperlinking is by preserving local data structure while learning embeddings to bridge modality gap. In this way, the neighbourhood statistics among anchor-target pairs are less likely to be arbitrarily changed during the course of learning. Both issues are seldom investigated in the literature of video hyperlinking [8], [18], [20].

The novelty of this paper is on the proposal of a new neighbourhood preserving MmDAE, which jointly addresses the issues of feature learning and local structure preservation. The main contributions are summarized as follows:

- A generalized MmDAE (G-MmDAE), which unifies two previous versions of MmDAE [7], [8], is proposed.
- Variants of structure-preserving embedding based on G-MmDAE are explored for video hyperlinking. Along this direction, issues such as model training and missing modality are addressed.

The remaining of this paper is organized as follows. Section II reviews the existing cross-modal embedding neural networks being developed for video hyperlinking. Section III describes issues in video fragmentation and feature preparation. Section IV outlines formulas for generalization of cross-modal embedding, and extension of it to structure-preserving embedding for video hyperlinking. Section V presents results and empirical insights on Blip10000 dataset. Finally, Section VI concludes this paper.

II. Related Works

Auto-encoder (AE) is one of the most commonly used techniques for representation learning, aiming to minimize the reconstruction error between the input and its reconstructed estimation from the hidden layer representation. In this section, we review two variants of AEs which have been used successfully in video hyperlinking. These AEs are designed for multi-modal embedding and cross-modal translation. More concretely, a modality (e.g., keyframes) is embedded and then translated to a new modality (e.g., speech). The reconstruction loss between the decoded and original modalities is then leveraged as the yardstick to optimize the parameters of AEs. In video hyperlinking, only two modalities are generally considered: speech and visual tracks. With the fact that speech transcript can be noisy, direct translation from visual track to speech sequence is not likely to yield satisfactory performance. As a consequence, instead of adopting recurrent neural networks to decode a learnt representation into sentences, speech is decoded as a feature vector rather than a word sequence in these AEs.

The first AE, named as MmDAE-O [7], projects two different modalities into a common representation, as shown in Figure 2(a). The input layer for each modality is followed by one or multiple fully-connected layers, before being collapsed into a common representation layer. This layer is fully connected, and interacts two modalities by non-linear transformation to create a joint multi-modal embedding. The learnt embedding is then decoded to its original forms of modalities by passing through

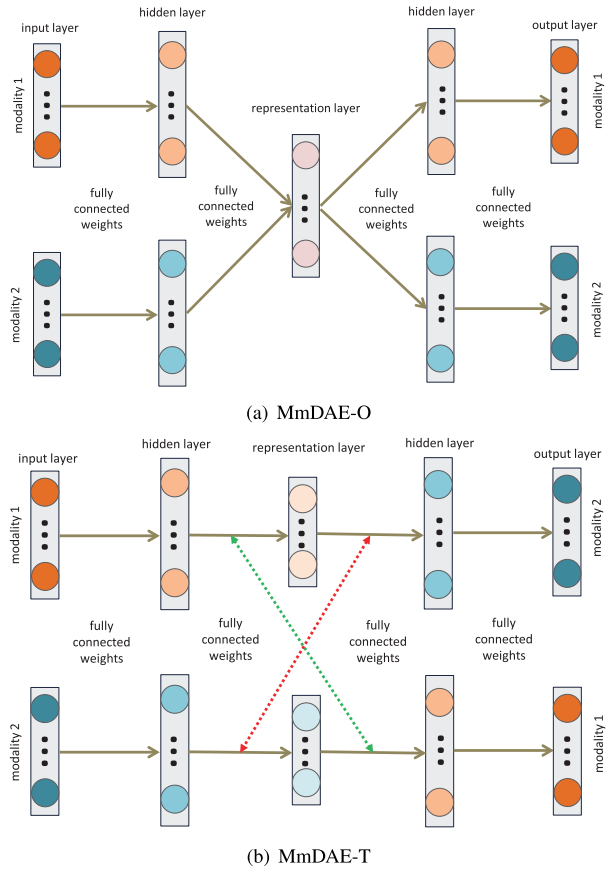


Fig. 2. Two variants of multi-modal auto-encoders used in the literature of video hyperlinking: (a) MmDAE-O learns one embedding, (b) MmDAE-T learns two embeddings for cross-modal translation.

two separate paths, each having equal number of fully-connected layers as the encoder. Reconstruction loss, which measures the distance between input and output modalities, is backpropagated end-to-end for network learning. The learnt embedding is expected to capture the latent representations of both modalities. With MmDAE, the embeddings of video fragments are extracted and indexed for video hyperlinking.

Instead of creating a common embedding, the second AE, named MmDAE-T [8], is composed of two subnetworks to generate two different embeddings, one for each modality. MmDAE-T, as shown in Figure 2(b), is formed by two encoder-decoder pairs for cross-modal translation. The two subnetworks share the same parameter weights for the layers adjacent to the representation layer (as marked in the green and red dotted lines). Rather than collapsing two modalities into a joint representation, MmDAE-T makes two different embeddings compatible through reconstructing the modalities of counterparts. In [5], the embeddings extracted from a video fragment are posteriorly concatenated as feature for hyperlinking. Note that both MmDAE-O and MmDAE-T can deal with the problem of missing modalities. MmDAE-O performs this by zeroing the input of missing modalities. MmDAE-T either duplicates embedding from another modality or treats the decoded modality as the input of missing modality [8].

In addition to MmDAE, other forms of embeddings that have been attempted for hyperlinking include generative adversarial networks (GAN) [33] and semantic representation network (SRN) [20]. In [33], the generator of GAN is applied to generate keyframe from speech. The speech embedding of the generator together with the convolution feature of a keyframe form a visual-speech pair, which is further predicted by discriminator as either true, false or synthetic pair. The embedding of the discriminator is ultimately treated as the representation of a video fragment. While being technically novel, the training of GAN is computationally more demanding than MmDAE. Furthermore, video context cannot be captured by GAN as the learning involves only speech-image pairs, rather than speech-video pairs as in MmDAE. SRN [20] treats hyperlinking as a classification task, by predicting probability of hyperlinking between two input video fragments. Compared with MmDAE [33], SRN suffers from the requirement to label anchor-target pairs for network learning. As demonstrated in this paper (Section V), MmDAE is feasible to be trained by using video-caption pairs which are more easily obtained than manually labelling anchor-target pairs.

While the idea of structure-preserving embedding is new for video hyperlinking, it has been recently studied in near-duplicate video search [15], [32], [34] and image retrieval [35]–[38] in the context of social media. In [36], for example, the learnt embeddings is trained to keep the local structure of the original visual and tag similarities. These local information are essential for applications such as hashing and retrieval. The structure-preserving embedding in this paper, although similar in spirit as [32], [39], [40], is formulated on the basis of encoder-decoder framework for self- and cross-modal embedding. As such, the formulation of the proposed deep network is fairly different from the structure-preserving formulation used in other context.

III. Content Preprocessing

This section starts by presenting the segmentation of videos into fragments (or clips), where each fragment is regarded as a target candidate (III-A). The extraction of multi-modal features from video fragments is also detailed (III-B).

A. Video Fragmentation

Video is treated as a sequence of visual and audio streams. Among various audio content, we consider only speech track, which is not only rich of textual information but also provides vivid cues for fragmentation through speech pause and speaker transition. The speech track is converted into transcripts, along with speech segment boundaries, by using LIMSI speech recognition system (ASR) [41]. In hyperlinking, as discussed in [42], a viewer expects that a target starts from the beginning rather than middle of a speech. Hence, this paper considers a video fragment as an uninterrupted speech sentence. Using speech boundaries, however, can result in excessive number of very short fragments, especially in conversation where utterances involves only a few spoken words. On the other hand, speech boundary is undefined for silent video segments, and thus the duration of

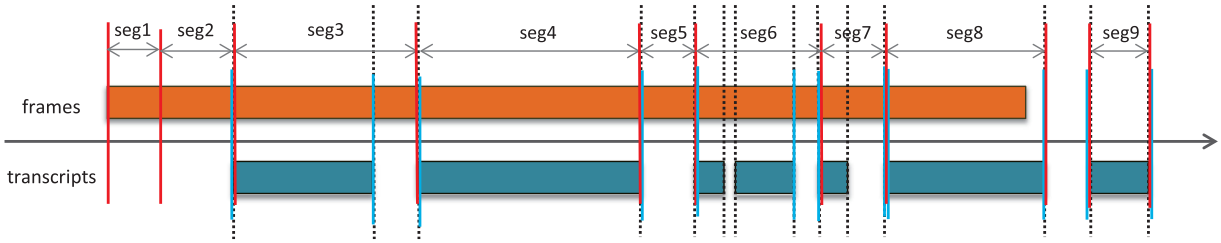


Fig. 3. An example illustrating the process of video fragmentation. A video is initially partitioned based on speech boundaries (black dash lines). Adjacent short fragments (e.g., seg6) are merged resulting in new boundaries (blue solid lines). Finally, fragments with only visual streams are either split or merged with fragments depending on length, ending up the final fragmentation (red solid lines).

a speechless fragment can be lengthy. To strike the balance, a heuristic speech-driven fragmentation algorithm is proposed as following.

The proposed algorithm starts by decomposing a video into fragments based on speech boundaries. The fragments with speech length less than a threshold τ will then be merged. Specifically, the algorithm scans a video from beginning to end, and progressively groups the adjacent fragments until the minimum target length τ reaches. Next, silent segments are split into fragments of fixed length τ. A fragment with length less than τ, as a result of splitting, is absorbed by its previous fragment along the time order. Figure 3 shows an example that illustrates the process of the proposed video fragmentation.

B. Fragment Descriptors

A fragment is represented by both visual and textual descriptors. Visual descriptor is characterized by a high dimensional vector, corresponding to either the features extracted from the convolutional layer of neural network [43] or the histogram of concept appearance [44]. Instead of employing content-based keyframe selection such as based on shot boundary detection [45], uniform sampling of keyframes at the rate of one frame per three seconds from a fragment is adopted for efficiency reason. As reported by [32], [34] in the context of near-duplicate video retrieval, the performance difference between uniform and content-based keyframe selection is insignificant. The features extracted from keyframes are averagely pooled to form a visual descriptor.

By treating the transcript as a document, textual descriptor is represented either as a bag-of-words vector using TF-IDF [46] or an encoded vector using word2vec trained based on Google news dataset [47]. For the latter, words in a transcript are first encoded by word2vec and then weighted averaged to form a descriptor. Specifically, let $\boldsymbol{v}_i = [v_{i1}, \ldots, v_{id_W}]^T$ denote the d_W-dimensional vector of the ith word in a vocabulary composed of U words. Furthermore, let $\boldsymbol{x} = [x_1, \ldots, x_U]^T$ be the bag-of-words vector of a transcript that captures the TF-IDF scores of every word, i.e., x_i. A textual descriptor, $\boldsymbol{z}$, is thus computed as following

$$\boldsymbol{z} = \frac{\sum_{i=1}^{U} x_i \boldsymbol{v}_i}{\sum_{l=1}^{U} x_l}. \quad (1)$$

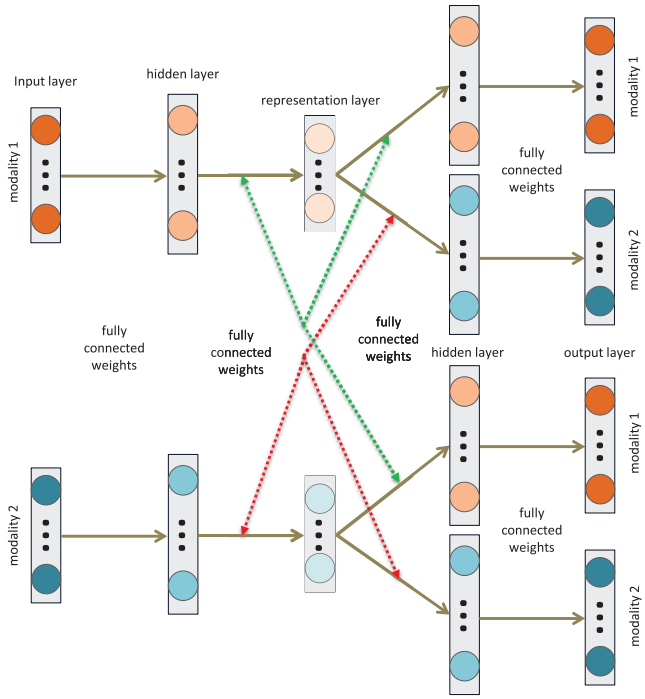


Fig. 4. Network architecture of G-MmDAE. The red and green arrows point to two groups of shared network parameters.

IV. Generalized MmDAE

This section generalizes MmDAE by integrating its variants (Figure 2) into a single network, named G-MmDAE. Similar to MmDAE-T, two embeddings (or representation layers) are learnt in G-MmDAE, one for each modality (i.e., visual and textual descriptors). Meanwhile, as MmDAE-O, each embedding of G-MmDAE is trained to reconstruct two different modalities. Referring to Figure 4, the network is formed by two subnetworks with four different encoder-decoder paths, performing either cross-modal translation or self-modal reconstruction. The network parameters before and after representation layers are shared (except biases) depending on the modality to be encoded or decoded. Specifically, the path to encode the representation of a modality shares the same parameter weights as the path to decode the representation of that modality. The weight sharing strategy enforces the mapping in the middle layers to be as close as possible to each other's inverses, which further enhances the

learnt representation to be more attainable from either one of the modalities [5]. In terms of learning effectiveness, this strategy enables simultaneous learning of the two sub-networks. Furthermore, the number of network parameters is reduced, which potentially avoids the problem of overfitting in learning.

A. Network Architecture

Denote an encoder branch as $\Phi_E(\boldsymbol{x}, \boldsymbol{\theta}_E)$ and a decoder branch as either $\Phi_{SD}(\boldsymbol{x}, \boldsymbol{\vartheta}_{SD})$ or $\Phi_{CD}(\boldsymbol{x}, \boldsymbol{\vartheta}_{CD})$. The subscript SD indicates decoding of the original modality (i.e., self-decoding), and CD indicates cross-modal decoding from visual to text or vice versa. The notation $\boldsymbol{x}$ refers to input, while $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ represent network parameters. Given a batch of visual-textual descriptor pairs, denoted as $\{(\boldsymbol{z}_t^{(1)}, \boldsymbol{z}_t^{(2)})\}_{t=1}^{B}$, as training samples, the parameters of G-MmDAE are optimized using the mean squared error (MSE), as follows

$$L_{MSE} = \alpha \left(L_S^{(1)} + L_S^{(2)} \right) + (1-\alpha) \left(L_C^{(1)} + L_C^{(2)} \right), \quad (2)$$

which performs weighted sum of losses accumulated from four different paths. For ease of reading, we use superscripts (1) and (2) to denote two different modalities, and subscripts S and C to indicate self- or cross-modal decoding respectively. The parameter $0 \leq \alpha \leq 1$ controls the weights of different losses. The loss at each path is quantified as

$$L_S^{(1)} = \frac{1}{B} \sum_{t=1}^{B} \left\| \Phi_{SD} \left(\Phi_E \left(\boldsymbol{z}_t^{(1)}, \boldsymbol{\theta}_E^{(1)} \right), \boldsymbol{\vartheta}_{SD}^{(1)} \right) - \boldsymbol{z}_t^{(1)} \right\|_2^2, \quad (3)$$

$$L_S^{(2)} = \frac{1}{B} \sum_{t=1}^{B} \left\| \Phi_{SD} \left(\Phi_E \left(\boldsymbol{z}_t^{(2)}, \boldsymbol{\theta}_E^{(2)} \right), \boldsymbol{\vartheta}_{SD}^{(2)} \right) - \boldsymbol{z}_t^{(2)} \right\|_2^2, \quad (4)$$

$$L_C^{(1)} = \frac{1}{B} \sum_{t=1}^{B} \left\| \Phi_{SD} \left(\Phi_E \left(\boldsymbol{z}_t^{(1)}, \boldsymbol{\theta}_E^{(1)} \right), \boldsymbol{\vartheta}_{CD}^{(2)} \right) - \boldsymbol{z}_t^{(2)} \right\|_2^2, \quad (5)$$

$$L_C^{(2)} = \frac{1}{B} \sum_{t=1}^{B} \left\| \Phi_{SD} \left(\Phi_E \left(\boldsymbol{z}_t^{(2)}, \boldsymbol{\theta}_E^{(2)} \right), \boldsymbol{\vartheta}_{CD}^{(1)} \right) - \boldsymbol{z}_t^{(1)} \right\|_2^2, \quad (6)$$

where there are two sets of mapping parameters $\{\boldsymbol{\theta}_E^{(g)}\}_{g=1}^{2}$ for encoding, two sets $\{\boldsymbol{\vartheta}_{SD}^{(g)}\}_{g=1}^{2}$ for self-modal decoding and two sets $\{\boldsymbol{\vartheta}_{CD}^{(g)}\}_{g=1}^{2}$ for cross-modal decoding. Here, $\|\cdot\|_2$ denotes the l_2-norm. Note that G-MmDAE degenerates to MmDAE-T by setting $\alpha = 0$, and to a simplified version of MmDAE-O without cross-modal decoding by setting $\alpha = 1$.

B. Training With Stochastic Structure Retaining

The network parameters of G-MmDAE are optimized by minimizing the reconstruction errors between paired inputs. As training samples are treated independently during optimization, the neighborhood structure of samples is not considered. As revealed in [10], local data structure characterizes the popularity and risk of hyperlinking, which provides cues for establishment of hyperlinks. Therefore, in principle, embedding should change data distribution under the guidance of the original neighborhood structure and not merely based on individual sample. With this motivation, we devise the training of G-MmDAE such that the learnt representation can preserve as much as possible the original data statistics and structure. This is achieved through stochastic structure retaining, where the key idea is to enforce the training samples in a mini-batch to exhibit similar neighbourhood structure in both the embedding and original spaces. In the following, we first present the construction of neighbourhood structure, followed by learning of structure-preserving cross-modal embedding.

Probabilistic encoding of neighbourhood structure. Given a video fragment i, the likelihood of picking another fragment j as its neighbour is represented by conditional probability. Denote d_{ij} as the distance between fragments i and j, the conditional probability is described as

$$\varphi(d_{ij}, \sigma_i) = \frac{\exp\left(-\frac{d_{ij}^2}{2\sigma_i^2}\right)}{\sum_{t \in I, t \neq i} \exp\left(-\frac{d_{it}^2}{2\sigma_i^2}\right)}, \quad (7)$$

where I is the set of training samples in a mini-batch. The Gaussian parameter σ_i is centered on fragment i, and characterizes how fast the similarity between fragments j and i will vanish when their distance increases. Any particular value of σ_i induces a probability distribution, $\{\varphi(d_{ij}, \sigma_i)\}_{j \in I}$, over I. The Shannon entropy of this distribution increases in proportion to the value of σ_i [48]. Given a batch of visual-textual descriptor pairs, denoted as $\{(\boldsymbol{z}_t^{(1)}, \boldsymbol{z}_t^{(2)})\}_{t=1}^{B}$, as training samples, we can compute two paired $B \times B$ matrices ($\mathbf{P}_1 = [p_{j|i}^{(1)}], \mathbf{P}_2 = [p_{j|i}^{(2)}]$) based on Eq. (7). The matrices represent the pairwise probabilities of visual and textual modalities respectively, where $p_{j|i}^{(g)} = \varphi(d_{ij}^{(g)}, \sigma_i^{(g)})$, $g \in \{1, 2\}$. The measure d_{ij}^{g} is computed using cosine distance metric as following

$$d_{ij}^{(g)} = 1 - \cos\left(\boldsymbol{z}_i^{(g)}, \boldsymbol{z}_j^{(g)}\right). \quad (8)$$

The parameter σ_i is estimated based on [49], which performs binary search of σ_i such that the Shannon entropy, i.e., $-\sum_{j \neq i} \varphi(d_{ij}, \sigma_i) \log_2 \varphi(d_{ij}, \sigma_i)$, approaches to $\log_2 K$. Here, K is an integer perplexity parameter assumed to be specified by user. As analysis in [32], [48], [50], K can be interpreted as a smooth measure of the effective number of neighbours. Performing binary search for each mini-batch, however, can be computational expensive. We approximate the value of σ_i by performing the binary search on several example sets of B (mini-batch size) fragments. In the implementation, each estimation of σ_i involves $B-1$ fragments uniformly sampled from a mini-batch. The process is repeated for five times, and the resulting values of σ_i are averaged as the final value. The estimation of σ_i using small samples still works satisfactorily, although structural properties (e.g., effective number of neighbours) can be more perfectly estimated with the whole training data at the expense of computational time. As studied in [49], the performance of stochastic neighbour embedding is fairly robust as long as the value of K is in a reasonable range (e.g., 5 to 50). To this end, two probability matrices are constructed similarly for the visual

and textual embeddings of training samples. Denote $(\boldsymbol{e}_t^{(1)}, \boldsymbol{e}_t^{(2)})$ as a pair of learnt embeddings, their $B \times B$ probability matrices, $(\mathbf{Q}_1 = [q_{j|i}^{(1)}], \mathbf{Q}_2 = [q_{j|i}^{(2)}])$, are computed via Eq. (7), where $q_{j|i}^{(g)} = \boldsymbol{\varphi}(1 - \cos(\boldsymbol{e}_i^{(g)}, \boldsymbol{e}_j^{(g)}), \frac{1}{\sqrt{2}})$, $g \in \{1, 2\}$. Here, we fix the Gaussian parameter σ to a constant of $\frac{1}{\sqrt{2}}$. As argued in [32], it is not necessary to scale the distances in both spaces, since similar effects can be achieved by scaling one and fixing the other.

Structure-preserving embedding. Following [15], [39], [50], Kullback-Leibler (KL) divergence is employed to measure the degree of matching between two distributions. Specifically, given two probability matrices, $\mathbf{P}$ and $\mathbf{Q}$, the KL divergence of sample distributions between the original and embedding spaces is

$$S_{KL}(\mathbf{Q}, \mathbf{P}) = \sum_{i \neq j} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}. \tag{9}$$

As KL divergence is not symmetric, different types of pairwise distance errors in the embedded space are not weighted equally. In particular, there is a large score when representing nearby datapoints with widely separated embedded points, but there is only a small score when representing widely separated datapoints with nearby embedded datapoints. In other words, this structure matching function is asymmetric and focuses on retaining the local structure of training data in the embedded space. Considering the architecture of G-MmDAE, we employ asymmetric KL divergence and the loss function is

$$\begin{aligned} L_{KL}(\boldsymbol{\theta}_E^{(1)}, \boldsymbol{\theta}_E^{(2)}) = {} & S_{KL}(\mathbf{Q}_1, \alpha\mathbf{P}_1 + (1-\alpha)\mathbf{P}_2) \\ & + S_{KL}(\mathbf{Q}_2, \alpha\mathbf{P}_2 + (1-\alpha)\mathbf{P}_1). \end{aligned} \tag{10}$$

The parameter $0 \leq \alpha \leq 1$ plays the same role as in Eq. (2) to weight the importance between the probability matrices constructed from visual and textual descriptors. Finally, this loss function is combined with Eq. (2) as a multi-objective loss

$$L_{MO} = \lambda L_{MSE} + L_{KL}, \tag{11}$$

where $\lambda \geq 0$ is a balancing parameter to control the relative importance of L_{MSE} to L_{KL}.

C. A Simplified G-MmDAE

G-MmDAE can also be learnt by directly minimizing the loss between the stochastic structure of embeddings and the original data samples without decoder. In other words, the balancing parameter is set to $\lambda = 0$ in Eq. (11). Due to the absence of decoder, nevertheless, the compatibility between visual and textual embeddings is not enforced. Therefore, correlation loss is introduced as following

$$L_{CR} = \frac{1}{B} \sum_{i}^{B} C_{cos}\left(\left(\boldsymbol{e}_i^{(1)}, \boldsymbol{e}_j^{(2)}\right), y_{ij}\right), \tag{12}$$

$$C_{cos}\left(\left(\boldsymbol{e}_i^{(1)}, \boldsymbol{e}_j^{(2)}\right), y_{ij}\right) = \begin{cases} 1 - \cos\left(\boldsymbol{e}_i^{(1)}, \boldsymbol{e}_j^{(2)}\right), & \text{if } y_{ij} = 1, \\ \max\left(0, \cos\left(\boldsymbol{e}_i^{(1)}, \boldsymbol{e}_j^{(2)}\right) - \gamma\right), & \text{if } y_{ij} = -1, \end{cases} \tag{13}$$

where $C_{cos}(\cdot)$ is the cosine similarity cost function and γ is the margin. The notation $y_{ij} = 1$ indicates that the ith and jth embeddings, $\boldsymbol{e}_i^{(1)}$ and $\boldsymbol{e}_j^{(2)}$, are a positive modality pair originated from the same fragment, and otherwise $y_{ij} = -1$ indicates a negative pair. Combining the two losses L_{KL} and L_{CR}, we can get the final objective as

$$L_{KC} = L_{KL} + \beta L_{CR}, \tag{14}$$

where $\beta \geq 0$ is a balancing parameter. Figure 5 depicts the architecture of this simplified G-MmDAE.

D. Convergence Analysis

The network training is mainly governed by the optimization of objective function, i.e., Eq. (2) for G-MmDAE, Eq. (11) for structure-preserving G-MmDAE and Eq. (14) for simplified G-MmDAE. As their underlying loss functions are smooth and differentiable, local optimum solution can be guaranteed by applying gradient descent algorithm [50]. As an example, the gradient formulation of S_{KL} in Eq. (9) with respect to the learned embedding $\boldsymbol{e}_i$ is given as follows

$$\frac{\partial S_{KL}}{\partial \boldsymbol{e}_i} = 2 \sum_j \left(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j}\right)\left(\boldsymbol{e}_i - \boldsymbol{e}_j\right). \tag{15}$$

With this formulation, back propagation is used by gradient descent algorithm to update network parameters. Similar procedure is applied to the loss functions of mean square error (MSE), i.e., Eqs. (3), (4), (5) and (6), and cosine similarity, i.e., Eq. (13).

The time complexity of network training depends on the numbers of layers, nodes in each layer, number of training examples and epochs. As an example, structure-preserving G-MmDAE, which has a total of 8 hidden layers and 14,336 nodes, takes about 20 minutes to converge using 20,000 training examples after 50 epochs. It is worth noting that the time complexity of MSE and cosine similarity is linear to the mini-batch size B. For KL divergence, i.e., Eq. (9), pairwise comparison between two groups of probabilities, $\mathbf{Q}$ and $\mathbf{P}$, is required. The time complexity is hence $O(n^2)$, where n is the number of examples in a mini-batch.

V. Experiments and Results

A. Dataset and Evaluation Metrics

The experiments are conducted on Blip10000 dataset collected from blit.tv [51]. The dataset contains 3,288 hours of videos with an average video length of 13 minutes. These videos cover a broad range of topics such as news, art and sport. There are 147 anchors provided by TRECVid LNK on this dataset. These anchors come along with a ground-truth which is composed of 9,602 anchor-target pairs, with 48.6% of positive pairs and 51.4% of negative pairs. These pairs were manually assessed, either as positive or negative pairs, by Mechanical Turk (MT) workers recruited by TRECVid LNK. On average, each anchor links to 65 targets. Note that some of videos are removed from Bip10000 by LNK, resulting in totally 11,482 videos for hyperlinking.

Following the evaluation mechanism used by TRECvid, the performance is measured by precision (P) up to a depth (i.e.,

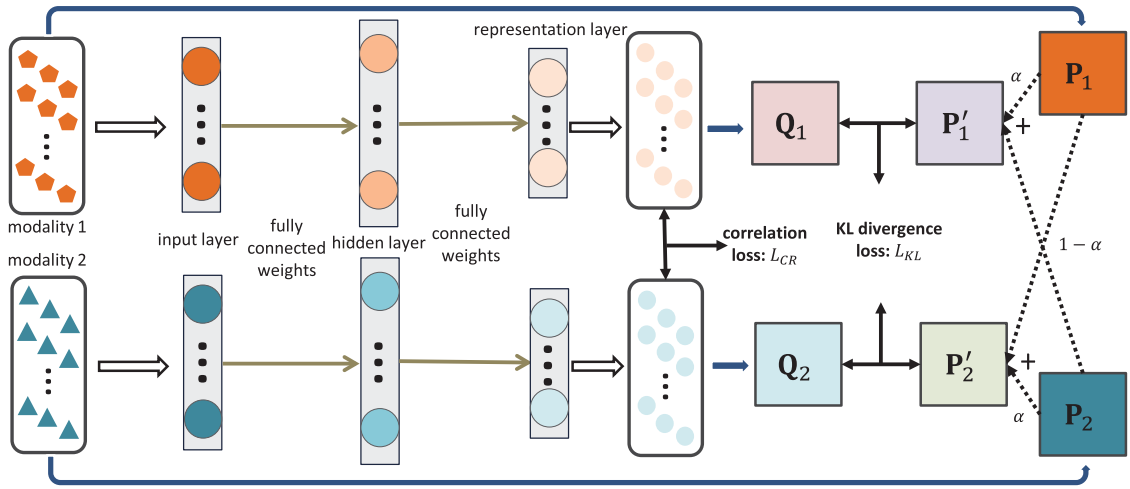


Fig. 5. Network architecture of the simplified G-MmDAE and its training mechanism. The distributions of training samples in the original and embedded spaces are captured by matrices $\mathbf{P}_i$ and $\mathbf{Q}_i$ respectively, where $i = \{1, 2\}$ corresponds to either text or visual modality. The learning aims to minimize the divergences between these two sets of distributions, while making the embedded features learnt from different modalities as similar as possible.

TABLE I
STATISTICS ON THE NUMBER OF FRAGMENTS FOR THREE DIFFERENT POOLS USED IN THE EXPERIMENTS. THE FRAGMENTS IN POOL-A ARE GENERATED BY THE PROPOSED ALGORITHM (SECTION III-A), AND POOL-B BY FIXED LENGTH SEGMENTATION. POOL-GT CONTAINS THE FRAGMENTS MANUALLY LABELED BY MT WORKERS

Stream	Pool-A	Pool-B	Pool-GT
Speech + Visual	116,170	70,430	9,415
Visual only[1]	20,728	9,156	172
Speech only	79	43	15
Total	136,977	79,629	9,602

P@5, P@10 and P@20), mean average precision (MAP) at the depth of 1,000, and mean average interpolated segment precision (MAiSP) [52]. Precision and MAP do not consider the level of overlap in frames between the ground-truth and detected fragments. MAiSP, in contrast, takes into account the degree of overlap and averages segment precisions at different recall levels.

B. Experiment Setup

Table I shows the number of fragments on Blip10000 dataset using the proposed approach (Section III-A) and fixed-length segmentation. We name the resulting pools of fragments as Pool-A (proposed approach) and Pool-B (fixed-length) respectively. For the proposed approach, the minimum fragment length is set to $\tau = 50$ seconds. The average length of fragments is 70 seconds. While for fixed-length segmentation, the fragment length is set as 120 seconds following [18]. Additionally, the set of fragments being labeled by MT workers are also collected to form Pool-GT. These fragments are the results pooled from different runs submitted by TRECVid participants, and hence the length of fragments varies. All the three pools will be experimented for hyperlinking. Note that in the experiments, we assume that the fragments which are not labeled by MT workers are all negative samples. Therefore, the performance attained in this paper should not be compared directly to the official results reported by TRECVid.

Two kinds of visual descriptors are extracted: concept histogram (CH) and pool5 (CP5) feature from convolutional neural networks (CNN). Five CNNs are trained separately on different datasets that cover 15,036 concepts mostly about objects and scenes: ImageNet (1,000 concepts) [44], ImageNet-Shuffle (12,988) [53], Places (205) [54], TRECVid MED research collection (497) [55] and SIN (346) [55]. CH visual descriptor is formed by concatenating the outputs of all five CNNs. CP5 descriptor, on the other hand, is set as the pool5 layer (2,048 dimensions) of the CNN trained using ImageNet. There are also two kinds of textual descriptors: bag-of-words (BoW) using TF-IDF, and word2vec representation using the improved Skip-gram word2vec model [47] trained on Google news dataset. The dimension of word2vec is set as $d_W = 300$.

C. Compared Methods

We group the compared methods into (I) baselines: CH, CP5, BoW, word2vec; (II) basic MmDAE variants: MmDAE-O, MmDAE-T and G-MmDAE; (III) the proposed MmDAE that preserves neighborhood structure: G-MmDAE* (Section IV-B) and G-MmDAE*-mini (Section IV-C); and (IV) fusion: CP5+word2vec by early fusion, CH+BoW +MmDAE-T, CH+BoW+G-MmDAE, CH+BoW+G-MmDAE* and CH+BoW+G-MmDAE*-mini by late fusion with average weights.

The two input modalities to the neural networks in groups II and III are CP5 and word2vec. All the networks are initially trained using MSR-VTT dataset [57], which is a

[1]As the used word embedding vectors are pre-trained based on the English corpus and all anchors are English video clips, in this work, we only use the visual contents of the non-English videos.

TABLE II
DETAILS OF MODEL TRAINING

Optim. Para.	**Pre-training with MSR-VTT**	**Fine-tuning with Blip10000**	
		Group II	Group III
Optimizer	torch.optim.SGD		torch.optim.Adam
Batch Size	1,000		
Epoch	300	100	
Learning Rate	0.01	0.001	0.0001
Momentum	0.9 (with Nesterov [57])		N/A
Dropout	0.2		

TABLE III
PERFORMANCE COMPARISONS ON POOL-GT. THE TWO BEST PERFORMANCES ARE BOLDED AND UNDERLINED RESPECTIVELY

Group	Methods	P@5	P@10	P@20	MAP	MAiSP
I	CH	0.256	0.218	0.178	0.181	0.088
	CP5	0.299	0.251	0.212	0.194	0.101
	BoW	0.393	0.327	0.284	0.205	0.135
	word2vec	0.343	0.297	0.234	0.183	0.133
II	MmDAE-O	0.348	0.294	0.237	0.209	0.122
	MmDAE-T	0.371	0.312	0.239	0.198	0.134
	G-MmDAE	0.373	0.312	0.246	0.209	0.132
III	G-MmDAE*	0.378	0.331	0.259	0.237	0.133
	G-MmDAE*-mini	0.386	0.331	0.260	0.235	0.140
IV	CP5+word2vec	0.315	0.280	0.225	0.206	0.108
	CH+BoW+MmDAE-T	0.445	0.407	0.331	0.294	0.161
	CH+BoW+G-MmDAE	0.445	0.407	0.331	<u>0.295</u>	0.161
	CH+BoW+G-MmDAE*	**0.450**	**0.415**	**0.334**	**0.299**	**0.165**
	CH+BoW+G-MmDAE*-mini	<u>0.448</u>	<u>0.409</u>	<u>0.332</u>	**0.299**	<u>0.162</u>

benchmark dataset used for video captioning. MSR-VTT contains 10,000 web video clips and 200,000 clip-sentence pairs. The pre-trained models are then fine-tuned by using fragments randomly picked from Blip10000 dataset. The total number of fragments used for fine-tuning is restricted to be about 20% of the total fragments. As Pool-GT has relatively small number of fragments than other pools, the neural networks experimented on Pool-GT are fine-tuned using fragments picked from Pool-A. Note that, for each compared method, we train the corresponding neural network for three times. The performances of the three networks are then averaged and reported in this paper.

The networks are implemented using PyTorch.[2] All the layers are fully connected and activated by Tanh function. For fair comparison, all the networks (except G-MmDAE*-mini) are set to have the same numbers of layers and neurons in each layer. The number of hidden neurons is 4,096 for visual descriptor CP5, and 2,048 for textual descriptor word2vec. The embedding size at representation layer is set as 1,024 for all the networks. Furthermore, the balancing factor α is set to an equal weight, i.e., $\alpha = 0.5$ (Eqs. (2) and (10)). The other balancing factor is set as $\lambda = 1.0$ (Eq. (11)) for G-MmDAE* such that the loss functions for modality encoding and structure preserving share equal weights. Following the work [58], the margin γ used in $C_{cos}(\cdot)$ (Eq. (13)) is fixed to 0.1 empirically. Moreover, to optimize the correlation loss ($\beta > 0$) employed in G-MmDAE*-mini, we pick a positive modality pair with 20% probability and a random negative modality pair with 80% probability from the training data. Finally, similar as λ, we set $\beta = 1.0$ to equally weight cross-modal consistency and structure preserving to report the performance.

[2]https://pytorch.org/

The performance of G-MmDAE* and G-MmDAE*-mini are not sensitive to the setting of the perplexity parameter K, we follow the suggestion in [32] and set $K = 20$. The detailed parameter setting for model training is listed in Table II.

D. Results

We report performances on three different Pools of dataset. Table III shows the result of hyperlinking for 147 anchors on Pool-GT. Single modality, either by visual or textual descriptor alone, already achieves greater than 25% in P@5. BoW appears to be a strong baseline outperforming all the deep neural networks. Among the compared methods in Group-II, using two embeddings (MmDAE-T and G-MmDAE) shows consistently better performances than network with only one layer (MmDAE-O). The result gives clues that having embeddings peculiar to each modality is a better strategy than collapsing both modalities into a common embedding. We speculate that this is because different modalities are complementary, and projecting them to a common representation layer may end up losing information specific to a modality. G-MmDAE, which learns two embeddings while enabling cross and self-modal decoding, exhibits the best performance among the three methods in Group-II in terms of precision and MAP. By further considering neighborhood structure, G-MmDAE* in Group-III shows higher performances in all the measures. The results verify our claim that the learnt embeddings should preserve as much as the original structure information for the task of hyperlinking. Interestingly, the simplified version, i.e., G-MmDAE*-mini, exhibits even better performance than G-MmDAE*. We believe that this is because visual-textual modalities in the training samples of Blip10000 are not as correlated as the samples of MSR-VTT.

TABLE IV
PERFORMANCE COMPARISONS ON POOL-A AND POOL-B. THE TWO BEST PERFORMANCES ARE BOLDED AND UNDERLINED RESPECTIVELY

Group	Method	Pool-A					Pool-B				
		P@5	P@10	P@20	MAP	MAiSP	P@5	P@10	P@20	MAP	MAiSP
I	CH	0.099	0.087	0.074	0.084	0.041	0.105	0.095	0.071	0.062	0.041
	CP5	0.166	0.131	0.100	0.113	0.052	0.172	0.133	0.102	0.092	0.051
	BoW	0.187	0.151	0.116	0.060	0.059	0.201	0.160	0.114	0.062	0.056
	word2vec	0.163	0.135	0.103	0.055	0.054	0.176	0.139	0.102	0.062	0.057
II	MmDAE-O	0.187	0.155	0.120	0.121	0.063	0.206	0.163	0.128	0.104	0.065
	MmDAE-T	0.201	0.165	0.119	0.092	0.064	0.220	0.165	0.126	0.094	0.071
	G-MmDAE	0.203	0.161	0.121	0.107	0.067	0.218	0.168	0.131	0.103	0.070
III	G-MmDAE*	0.223	0.166	0.125	0.123	0.067	0.229	0.173	0.136	0.111	0.069
	G-MmDAE*-mini	0.226	0.180	0.138	**0.128**	0.072	0.228	0.182	0.137	**0.112**	0.070
IV	CP5+word2vec	0.197	0.161	0.124	0.126	0.059	0.199	0.156	0.120	0.103	0.057
	CH+BoW+MmDAE-T	0.231	0.195	0.144	0.112	**0.079**	0.233	0.192	**0.148**	0.101	0.074
	CH+BoW+G-MmDAE	0.232	0.195	0.144	0.112	**0.079**	0.234	0.192	**0.148**	0.101	0.074
	CH+BoW+G-MmDAE*	0.246	0.199	0.148	0.123	0.078	0.251	**0.202**	0.147	0.107	0.076
	CH+BoW+G-MmDAE*-mini	**0.249**	**0.203**	**0.151**	0.123	**0.079**	**0.253**	**0.202**	**0.148**	0.107	**0.077**

TABLE V
SPEED EFFICIENCY OF DIFFERENT APPROACHES ON POOL-A. THE AVERAGE TIMES OF FEATURE EMBEDDING AND SEARCHING PER ANCHOR ARE REPORTED

Group	Method	Embedding (ms)	Searching (s)
I	CH	N/A	0.112
	CP5	N/A	0.195
	BoW	N/A	0.096
	word2vec	N/A	0.031
II	MmDAE-O	1.11	0.041
	MmDAE-T	1.62	0.077
	G-MmDAE	1.64	0.078
III	G-MmDAE*	1.69	0.094
	G-MmDAE*-mini	1.57	0.085

TABLE VI
IMPACT OF SELF-MODAL VERSUS CROSS-MODAL DECODING. NOTE: $\alpha = 0, 0.5, 1$ REFER TO CROSS, SELF+CROSS, AND SELF-MODAL DECODING RESPECTIVELY (SEE EQS. (2) AND (10)). THE EXPERIMENT IS CONDUCTED ON POOL-A

Method	α	P@5	P@10	P@20	MAP	MAiSP
G-MmDAE	0.0	0.201	0.165	0.119	0.092	0.064
	0.5	0.203	0.161	0.121	0.107	0.067
	1.0	0.199	0.161	0.121	0.110	0.067
G-MmDAE*	0.0	0.221	0.168	0.126	0.122	0.068
	0.5	0.223	0.166	0.125	0.123	0.067
	1.0	0.208	0.165	0.126	0.124	0.068

Further performing cross-modal decoding during fine-tuning stage may actually hurt the feature learning of G-MmDAE*. In other words, preserving neighborhood structure plays a more important role than cross-modal decoding in learning embedding. We believe that the network of G-MmDAE* can be further optimized with more training examples for performance boosting, as more data can generally enhance the construction of data structure. Similar as other reported results [5], [18], [20], fusion of different modalities or methods leads to improvement than that of single modality or method. Particularly, as the inputs of G-MmDAE variants are CP5 and word2vec, further late fusing the results with CH and BoW leads to significantly larger improvement. The results show that the learnt embeddings are complementary to low-level hand-crafted features.

Table V shows the running time required for an anchor from feature embedding to target hyperlinking. Excluding the time spent on extracting anchor descriptors, variants of MmDAE take less than 0.1 s to complete the search of targets. The time is compatible to baselines such as CH and CP5, and varies depending on feature dimension.

Effect of fragmentation. Table IV shows the performances of various methods on Pool-A and Pool-B. As these pools include significantly more fragments than Pool-GT, it is not surprising to notice performance degradation. Different from the performances observed on Pool-GT, all deep models consistently outperform BoW. Except this, the performance trend is similar. Structure-preserving embedding exhibits better performance, and further fusion with CH and BoW leads to the overall best performances in P@K, where K={5, 10, 20}. Fusion, however, either does not improve or degrade the MAP performance. In other words, although fusion boosts the ranking of some true positives to $K \leq 20$ position, overall it does not promote or recall more true positives within the depth of $K = 1,000$. Fusion basically help in pushing the ranking position if different methods equally rank a true positive high. Otherwise, the rank of a true positive is likely to degrade after taking average fusion. Comparing two different fragmentation methods, most methods show higher performance in P@5 on Pool-B and in MAP on Pool-A. which may due to the fact that fragments with longer length are more likely to overlap with ground-truth. The result can be interpreted as that more sophisticated fragmentation scheme as presented in Section III-A can help recalling more number of relevant targets, but pushing them to higher rank position for hyperlinking remains difficult. On the other, the result may also imply that aligning the start and end times of a fragment with speech is possibly not critical. Users can still forward or backward a video as long as the desired content is captured by the fragment.

Effect of fine-tuning. We also study the effect of fine-tuning on the performance. By using the model trained based on MSR-VTT dataset only, the methods in Group-II show similar performance as with the version with model fine-tuning. The result can verify that the embeddings learnt by these methods indeed properly capture the cross-modal signals. As a consequence, the network parameters are not overwhelmingly overridden when

TABLE VII
EFFECT OF REPLACING MISSING MODALITY WITH QUERY EXPANSION

	G-MmDAE($\alpha = 0.5$)					G-MmDAE*($\alpha = 0.5$)				
	P@5	P@10	P@20	MAP	MAiSP	P@5	P@10	P@20	MAP	MAiSP
Without query expansion	0.174	0.136	0.103	0.079	0.053	0.193	0.149	0.110	0.094	0.060
With query expansion	0.187	0.142	0.104	0.081	0.056	0.200	0.150	0.113	0.096	0.061

TABLE VIII
PERFORMANCE BREAKDOWN FOR THREE SETS OF ANCHORS. RESULTS ON POOL-A ARE REPORTED. THE TWO BEST PERFORMANCES ARE BOLDED AND UNDERLINED RESPECTIVELY

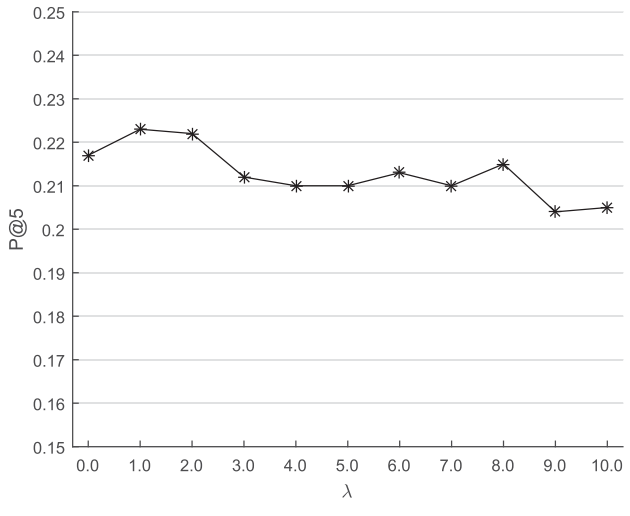
Group	Method	Dev'16(anchor IDs:1-28)			Test'16(anchor IDs:29-122)			Test'17(anchor IDs:123-147)		
		P@5	MAP	MAiSP	P@5	MAP	MAiSP	P@5	MAP	MAiSP
I	CH	0.021	0.157	0.025	0.111	0.067	0.047	0.144	0.062	0.037
	CP5	0.086	0.254	0.042	0.180	0.085	0.058	0.208	0.058	0.039
	BoW	0.129	0.094	0.110	0.122	0.042	0.044	0.488	0.088	0.055
	word2vec	0.143	0.065	0.086	0.113	0.065	0.046	0.368	0.063	0.046
II	MmDAE-O	0.116	**0.285**	0.075	0.169	0.080	0.061	0.333	0.086	0.056
	MmDAE-T	0.148	0.197	0.099	0.170	0.064	0.055	0.376	0.076	0.053
	G-MmDAE	**0.157**	0.240	0.101	0.167	0.073	0.060	0.384	0.085	0.057
III	G-MmDAE*	0.141	0.254	0.098	0.202	0.094	0.060	0.389	0.084	0.058
	G-MmDAE*-mini	0.136	0.247	0.100	**0.209**	0.104	0.068	0.384	0.083	0.057
IV	CP5+word2vec	0.093	0.268	0.051	0.195	0.094	0.063	0.320	0.085	0.053
	CH+BoW+MmDAE-T	**0.157**	0.193	**0.125**	0.178	0.087	**0.069**	**0.504**	0.109	**0.065**
	CH+BoW+G-MmDAE	**0.157**	0.193	**0.125**	0.179	0.087	0.068	**0.504**	0.110	**0.065**
	CH+BoW+G-MmDAE*	**0.157**	0.193	**0.125**	0.203	**0.105**	0.066	**0.504**	**0.111**	**0.065**
	CH+BoW+G-MmDAE*-mini	**0.157**	0.193	**0.125**	0.206	**0.105**	0.068	**0.504**	0.110	**0.065**

being fine-tuned on a dataset of very different content. On the other hand, as the methods in Group-III aim to preserve neighborhood statistics of the original dataset, model fine-tuning is a necessary step for learning data distribution. As experimented, fine-tuning boosts the performance of G-MmDAE* by 5%–10% (P@5, Pool-A and Pool-B) compared to pre-trained model. Nevertheless, as discussed in the previous paragraph, detaching decoder of G-MmDAE* from fine-tuning, i.e., G-MmDAE*-mini, can achieve higher performance boost. We speculate that the fine-tuning of G-MmDAE* decoder could be effective if applying methods such as [59] to filter the fragments with weak correlation between visual and textual modalities from training.

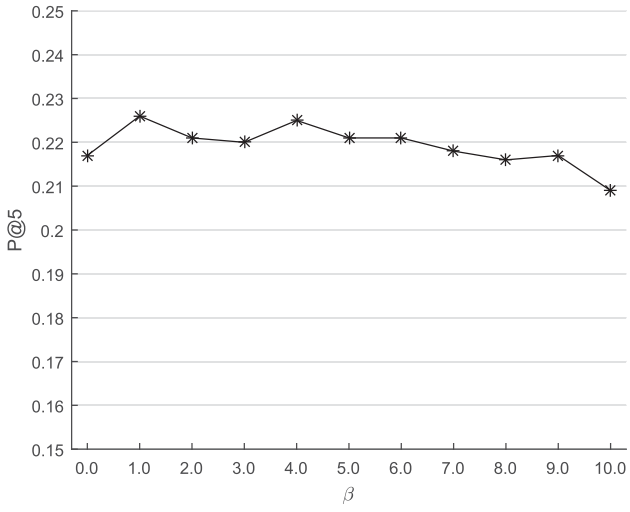
Self vs. cross-modal decoding. We further investigate the fundamental difference in performances between self and cross-modal decoding, by tuning the parameter α for G-MmDAE and G-MmDAE*. Table VI lists the results of comparison. Cross-modal translation attains higher precision (P@5 and P@10), but shows either competitive or lower performances than self-modality decoding in other measures. In other words, cross-modal embedding is helpful in pushing relevant targets, which is important for video hyperlinking. However, the embedding also introduces noise resulting in performance fluctuation. Decoding both self and cross modalities compromises their respective performances and attains the overall best P@5.

We also experiment the effect of G-MmDAE variants in dealing with missing modalities. For example, for a fragment without speech track, text embedding is obtained by treating the decoded modality of visual descriptor as input. Similarly to [5], we refer to this method as "query expansion". The experiment is conducted on Pool-A by randomly removing one modality of a fragment with a probability of 0.3. Through this process, we make sure that, in the resulting dataset, there are 15% of fragments with missing speech and 15% with missing visual modality. Table VII shows the experimental result. As shown, better performances are consistently observed in both versions of G-MmDAE when query expansion strategy is adopted.

Performance variations among anchors. The 147 anchors are created across different years and with different emphasis for hyperlinking. For example, the first 28 anchors are picked to convey speech-to-speech hyperlinking and included as the development set of TRECVid LNK 2016 (Dev'16). Hence, visual descriptor plays a minor role. On the other hand, the remaining anchors are selected to convey speech-to-visual information, and thus cross-modal features are expected to be more effective. Table VIII shows the performances across the three set of anchors on Pool-A. For the first 28 anchors in Dev'16, the result shows that performance using textual descriptor is significantly better than visual descriptor. Although these anchors emphasize more on speech-to-speech hyperlinking, combining both visual and speech manages to boost the performance. Performances on the remaining anchors in Test'16 and Test'17, nevertheless, show different trends. In Test'16, which has 94 anchors, the performance gap between visual and textual descriptors is relatively smaller. Hence, cross-modal embedding leads to larger improvement than single modality compared to the performance observed in Dev'16. In Test'17, in contrast, BoW appears to be a strong baseline significantly better than all the methods in groups I to III. We speculate that this is due to bias in ground-truth annotation because there are relatively few teams participating in TRECVid LNK 2017 benchmarking. In summary,



(a) G-MmDAE*



(b) G-MmDAE*-mini

Fig. 6. Performance changes in P@5 for (a) G-MmDAE* and (b) G-MmDAE*-mini when varying the values of λ and β respectively. The experiments are conducted on Pool-A and the value of α is fixed to 0.5.

despite different observations made on different subsets of anchors, late fusion of groups I and III yields either very competitive or best overall performances.

Parameter sensitivity. Here, we investigate the impacts of two hyper parameters, λ and β in Eq. (11) and Eq. (14) respectively, which balance different loss functions in G-MmDAE* and G-MmDAE*-mini. When their values are set to 0, only structure preservation is enforced while cross-modal embedding compatibility is not considered. On the other hand, when larger values are set, the importance of structure preservation w.r.t cross-modal translation decreases. Figure 6 shows the performance trends by varying the values of λ and β in the ranges of [0, 10], respectively. Either ignoring ($\lambda, \beta = 0$) or over-emphasizing ($\lambda, \beta > 1$) cross-modal decoding leads to suboptimal performance. Indeed, the proposed settings of $\lambda, \beta = 1$, which are intuitively to balance both factors, successfully leverage them to boost performance.

VI. Conclusion

We have presented a generalized version of multi-modal auto-encoder popularly used in the literature of video hyperlinking. The proposed encoder (G-MmDAE) learns two feature embeddings, while performing both self and cross-modal decoding. Empirical results on Blip10000 dataset show that G-MmDAE manages to compromise the performances between MmDAE-O and MmDAE-T, which either learns one embedding or performs only cross-modal decoding. Furthermore, the paper contributes by introducing structure-preserving embedding learning on top of G-MmDAE. A simplified version without explicit decoding of modalities is also proposed. The new proposals (G-MmDAE* and G-MmDAE*-mini) show performance improvement over G-MmDAE.

Comprehensive experimental verification has been conducted using 147 anchors provided by TRECVid on Blip10000 dataset. One key conclusion is the feasibility of learning G-MmDAE and its variants (without structure preserving) using video captioning datasets. Our results verify that the parameters learnt by these models properly capture cross-modal signals for video hyperlinking. G-MmDAE*, on the other hand, requires model fine-tuning for learning data distribution. As in reality visual and speech modalities are not necessarily strongly correlated as in video captioning datasets, learning through self and cross-modal decoding sometimes will hurt the performance. As a consequence, learning correlation between embedding features (i.e., G-MmDAE*-mini) appears to be a safer strategy than performing modality decoding (G-MmDAE*). Nevertheless, in the case when a dataset consists of fragments with missing modalities, G-MmDAE* is more applicable for effectiveness in "filling in" for missing modalities with query expansion. Finally, throughout the experiments, there is no obvious difference between using fixed length video fragmentation and the algorithm proposed in this paper (i.e., sentence level segmentation). Aligning fragments with start and end times of speech may not be critical although helpful in finding more relevant targets of an anchor.

References

[1] P. Over *et al.*, "TRECVID 2015: An overview of the goals, tasks, data, evaluation mechanisms, and metrics," in *Proc. TREC Video Retrieval Eval.*, Jan. 2015, pp. 1–52.
[2] M. Eskevich *et al.*, "The search and hyperlinking task at MediaEval 2014," in *Proc. MediaEval Workshop*, 2014, pp. 16–17.
[3] G. Awad *et al.*, "Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking," in *Proc. TREC Video Retrieval Eval.*, 2016, pp. 1–54.
[4] G. Awad *et al.*, "Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking," in *Proc. TREC Video Retrieval Eval.*, 2017, pp. 1–39.
[5] V. Vukotić, C. Raymond, and G. Gravier, "Multimodal and crossmodal representation learning from textual and visual features with bidirectional deep neural networks for video hyperlinking," in *Proc. ACM Workshop Vis. Lang. Integration Meets Multimedia Fusion*, 2016, pp. 37–44.
[6] M. Demirdelen, M. Budnik, G. Sargent, R. Bois, and G. Gravier, "Irisa at trecvid 2017: Beyond crossmodal and multimodal models for video hyperlinking," in *Proc. Working Notes TREC Video Retrieval Eval. Workshop*, 2017, pp. 1–8.
[7] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.
[8] V. Vukotić, C. Raymond, and G. Gravier, "Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 343–346.

[9] R. Ordelman, R. Aly, M. Eskevich, B. Huet, and G. J. Jones, "Convenient discovery of archived video using audiovisual hyperlinking," in *Proc. Third Edition Workshop Speech, Lang. Audio Multimedia*, 2015, pp. 23–26.

[10] Z.-Q. Cheng, H. Zhang, X. Wu, and C.-W. Ngo, "On the selection of anchors and targets for video hyperlinking," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2017, pp. 287–293.

[11] R. Aly, R. J. Ordelman, M. Eskevich, G. J. Jones, and S. Chen, "Linking inside a video collection: What and how to measure?" in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 457–460.

[12] S. Yang, L. Pang, C.-W. Ngo, and B. Huet, "Serendipity-driven celebrity video hyperlinking," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 413–416.

[13] C.-L. Chou, H.-T. Chen, and S.-Y. Lee, "Pattern-based near-duplicate video retrieval and localization on web-scale videos," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 382–395, Mar. 2015.

[14] H. Wang, T. Tian, M. Ma, and J. Wu, "Joint compression of near-duplicate videos," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 908–920, May 2017.

[15] Y. Hao *et al.*, "Unsupervised t-distributed video hashing and its deep hashing extension," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5531–5544, Nov. 2017.

[16] R. J. Ordelman, M. Eskevich, R. Aly, B. Huet, and G. Jones, "Defining and evaluating video hyperlinking for navigating multimedia archives," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 727–732.

[17] A.-R. Simon *et al.*, "Hierarchical topic models for language-based video hyperlinking," in *Proc. Third Edition Workshop Speech, Lang. Audio Multimedia*, 2015, pp. 31–34.

[18] B. Huet, E. Baralis, P. Garza, and M. Reza Kavoosifar, "Eurecom-Polito at TRECVID 2017: Hyperlinking task," in *Proc. TRECVID 21st Int. Workshop Video Retrieval Eval.*, Gaithersburg, MD, USA, Nov. 2017.

[19] B. Merialdo, P. Pidou, M. Eskevich, and B. Huet, "Eurecom at trecvid 2016. The adhoc video search and video hyperlinking tasks," in *Proc. TREC 20th Int.Workshop Video Retrieval Eval.*, Gaithersburg, MA, USA, Nov. 2016.

[20] P. A. Nguyen *et al.*, "Vireo @ trecvid 2017: Video-to-text, ad-hoc video search and video hyperlinking," in *Proc. TREC Video Retrieval Eval. Workshop*, 2017.

[21] M. Cha, Y. Gwon, and H. T. Kung, "Multimodal sparse representation learning and applications," 2015, arXiv: 1511.06238.

[22] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 7–16.

[23] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 171–184, 2002.

[24] X. Yang, T. Zhang, and C. Xu, "Text2video: An end-to-end learning framework for expressing text with videos," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2360–2370, Sep. 2018.

[25] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, Jun. 2017.

[26] Z. Li and J. Tang, "Weakly supervised deep metric learning for community-contributed image retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1989–1999, Nov. 2015.

[27] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.

[28] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 128–141, Jan. 2018.

[29] R. Hong *et al.*, "Coherent semantic-visual indexing for large-scale image retrieval in the cloud," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4128–4138, Sep. 2017.

[30] R. Hong, Z. Hu, R. Wang, M. Wang, and D. Tao, "Multi-view object retrieval via multi-scale topic models," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5814–5827, Dec. 2016.

[31] M. Wang, R. Hong, X.-T. Yuan, S. Yan, and T.-S. Chua, "Movie2comics: Towards a lively video content presentation," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 858–870, Jun. 2012.

[32] Y. Hao *et al.*, "Stochastic multiview hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 1–14, Jan. 2017.

[33] V. Vukotić, C. Raymond, and G. Gravier, "Generative adversarial networks for multimodal representation learning in video hyperlinking," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2017, pp. 416–419.

[34] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1997–2008, Dec. 2013.

[35] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, Nov. 2012.

[36] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2827–2840, Sep. 2015.

[37] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, 2018, doi: 10.1109/TPAMI.2018.2852750.

[38] C. Luo, B. Ni, S. Yan, and M. Wang, "Image classification by selective regularized subspace learning," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 40–50, Jan. 2016.

[39] T. Mu, J. Y. Goulermas, and S. Ananiadou, "Data visualization with structural control of global cohort and local data neighborhoods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1323–1337, Jun. 2018.

[40] Y. Hao *et al.*, "Cross-domain sentiment encoding through stochastic word embedding," *IEEE Trans. Knowl. Data Eng.*, to be published, 2019, doi: 10.1109/TKDE.2019.2913379.

[41] L. Lamel and J.-L. Gauvain, "Speech processing for audio indexing," in *Proc. Adv. Natural Lang. Process.*, 2008, pp. 4–15.

[42] M. Eskevich and G. J. F. Jones, "Time-based segmentation and use of jump-in points in DCU search runs at the search and hyperlinking task at mediaeval 2013," in *Proc. MediaEval Workshop*, 2013.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[44] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[45] M. Wang *et al.*, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 975–985, Aug. 2012.

[46] G. Salton, "Developments in automatic text retrieval," *Science*, vol. 253, no. 5023, pp. 974–980, 1991.

[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[48] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *J. Mach. Learn. Res.*, vol. 11, no. Feb., pp. 451–490, 2010.

[49] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 857–864.

[50] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[51] S. Schmiedeke *et al.*, "Blip10000: A social video dataset containing spug content for tagging and retrieval," in *Proc. 4th ACM Multimedia Syst. Conf.*, 2013, pp. 96–101.

[52] D. N. Racca and G. J. Jones, "Evaluating search and hyperlinking: An example of the design, test, refine cycle for metric development," in *Proc. MediaEval Workshop*, 2015.

[53] P. Mettes, D. C. Koelma, and C. G. Snoek, "The imagenet shuffle: Reorganized pre-training for video event detection," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 175–182.

[54] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[55] P. Over *et al.*, "Trecvid 2014—an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. TREC Video Retrieval Eval.*, Orlando, FL, USA, 2014, pp. 1–52.

[56] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.

[57] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5288–5296.

[58] A. Salvador *et al.*, "Learning cross-modal embeddings for cooking recipes and food images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3020–3028.

[59] C. A. Henning and R. Ewerth, "Estimating the information gap between textual and visual representations," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2017, pp. 14–22.

Yanbin Hao received the B.E. and Ph.D. degrees from the Hefei University of Technology, Hefei, China, in 2012 and 2017, respectively. During his Ph.D., he was also a visiting Ph.D. (2015–2017) student with the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, U.K. He is currently a Senior Research Associate with the Department of Computer Science, City University of Hong Kong, Hong Kong. His research interests mainly include machine learning and multimedia data analysis, such as large-scale multimedia indexing and retrieval, multimedia data embedding, and video hyperlinking.

Chong-Wah Ngo received the B.Sc. and M.Sc. degrees both in computer engineering from Nanyang Technological University of Singapore, Singapore, and the Ph.D. degree in computer science from Hong Kong University of Science and Technology, Hong Kong. He is a Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. Before joining City University of Hong Kong, he was a Postdoctoral Scholar with the Beckman Institute, University of Illinois, Urbana-Champaign. His main research interests include large-scale multimedia information retrieval, video computing, multimedia mining and visualization. He was an Associate Editor for the IEEE TRANSANCTION ON MULTIMEDIA and is currently steering committee member of TRECVid, ICMR (International Conference on Multimedia Retrieval) and ACM Multimedia Asia. He is program Co-Chair of ACM Multimedia 2019, and general Co-Chairs of ICIMCS 2018 and PCM 2018. He was named ACM Distinguished Scientist in 2016 for contributions to video search and semantic understanding.

Benoit Huet received the B.Sc. degree in computer science and engineering from the Ecole Superieure de Technologie Electrique (Groupe ESIEE, France), in 1992, the M.Sc. degree in artificial intelligence from the University of Westminster, London, U.K., in 1993 with distinction, where he then spent two years working as a research and teaching Assistant, and the D.Phil degree in computer science from the University of York, York, U.K. for his research on the topic of object recognition from large databases. He is an Assistant Professor with the Data Science Department, EURECOM, Biot, France. He has co-authored more than 150 papers in Books, Journals, and International Conferences. His current research interests include large scale multimedia content analysis, mining and indexing C video understanding and hyperlinking—multimodal fusion—socially-aware multimedia. He was awarded the HDR (Habilitation to Direct Research) from the University of Nice Sophia Antipolis, France, in October 2012 on the topic of Multimedia Content Understanding: Bringing Context to Content. He is an Associate Editor for the IEEE TRANSACTION ON MULTIMEDIA, MULTIMEDIA TOOLS AND APPLICATION (Springer) and Multimedia Systems (Springer) and has been Guest Editor for a number of special issues (*EURASIP Journal on Image and Video Processing*, IEEE MULTIMEDIA, etc.). He regularly serves on the technical program committee of the best conference of the field (ACM MM/ICMR, IEEE ICME/ICIP). He served as Technical Program Co-Chair of ACM Multimedia 2016 and ACM ICMR 2018 and the General Co-Chair for Multimedia Modeling 2019 in Thessaloniki, Greece and ACM Multimedia 2019 in Nice, France.