Cross-modal Search Method of Technology Video based on Adversarial Learning and Feature Fusion

Xiangbin Liu, Junping Du*, Meiyu Liang, Ang Li
Beijing Key Laboratory of Intelligent Communication Software and Multimedia, School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876

Abstract: Technology videos contain rich multi-modal information. In cross-modal information search, the data features of different modalities cannot be compared directly, so the semantic gap between different modalities is a key problem that needs to be solved. To address the above problems, this paper proposes a novel Feature Fusion based Adversarial Cross-modal Retrieval method (FFACR) to achieve text-to-video matching, ranking and searching. The proposed method uses the framework of adversarial learning to construct a video multimodal feature fusion network and a feature mapping network as generator, a modality discrimination network as discriminator. Generator and discriminator are trained alternately based on adversarial learning, so that the data obtained by the feature mapping network is semantically consistent with the original data and the modal features are eliminated. Experimental results demonstrate that the proposed method performs better in text-to-video search on the self-built datasets of technology videos.

Keywords: Cross-modal search; Adversarial learning; Feature fusion; Technology video; Search ranking

1 Introduction

With the rapid development of the new generation of 5G Internet, video data information is exploding. More and more videos about science and technology information have an increasingly important role [1]. Technology videos mainly include academic conferences, lectures and other professional academic content, which is a valuable and rich information access channel for scientific researchers. Compared with the popular short videos on the Internet, technology videos are longer, more professional, and richer in content. For scientific researchers, when they want to obtain information related to a certain field, the most common way is to use text modal search for the relevant description text of technology videos, which has high specialization requirements for the description and labeled information of technology videos, and the unlabeled videos cannot be retrieved, which is not conducive to the dissemination of technology videos and academic information. The search of single-modal data can no longer meet the existing demand for technology video, and the demand for cross-modal information search of technology videos is increasing day by day.

In recent years, deep learning has been widely used on text and video data, and it provides support for cross-modal search by extracting data features accurately and efficiently. However, the data feature distributions between different modalities are not interoperable [2], and adversarial learning is needed to establish the association between text and video containing the same semantic content through the semantic space. Adversarial learning is very effective for generating a new data distribution and has been widely used for text, image, and speech generation [3].

In this paper, we propose FFACR for searching technology videos according to semantic similarity. The method mainly uses adversarial learning strategy to train three neural network models, namely, video multi-modal feature fusion network, feature mapping network and modal discrimination network. The network is involved in the training process and generates a better multimodal feature representation of the video. The modality discriminator network is used to distinguish the original modality of the data in the same semantic space.

The proposed FFACR method is intended to solve the cross-modal search problem of technology videos. The main contributions of this paper are as follows.

(1) We propose a novel cross-modal search method of technology video based on adversarial learning and feature fusion to achieve text-to-video searching. Through the joint learning of video multi-modal feature fusion network, feature mapping network and modal discrimination network, the text and video modal data are mapped into a unified semantic space.

(2) We propose a feature fusion-based approach to obtain a feature representation of the multi-modal of one video. Compared with the single-modal information, multi-modal video features can better represent the semantic information of the video and improve the accuracy in the subspace.

(3) We use natural semantic segmentation of video speech information to slice and dice the training samples. The speech text of videos has good natural segmentation information. Using automatic speech recognition method, we generate speech text and slice the long technology video into several short video samples for training.

(4) We use datasets of technology videos collected from the Internet, which can represent the types of videos that researchers are generally exposed to using. We use text search videos as the search task, and the results show that the FFACR algorithm can improve the cross-modal search accuracy of technology videos compared to other algorithms.

*Corresponding author: Junping Du (junpingdu@126.com).

978-1-6654-7735-2/22/$31.00 ©2022 IEEE
2 Related Work

2.1 Cross-modal search methods

Cross-modal search aims to solve the discrepancy problem between different modalities [4][5]. Feng et al constructed a model by correlating the hidden representations of two unimodal auto encoders, which is suitable for data with small amount of data and simple features. Hardoon et al used correlation analysis of kernel methods to learn the semantic representation of data pairs. Peng et al proposed cross-media multiple deep network (CMDN), which learns information and cross-modal correlations within the same modality and different modalities in two successive steps. The errors brought by training will produce cumulative, inaccurate cross-modal semantic descriptions [6][7][8]. Wang et al proposed a multimodal mapping method that can measure the similarity between different modal data and preserve the inter-modal and intra-modal similarity relationships. Wang et al proposed a coupled linear regression framework to achieve cross-modal search and solve the regularized linear regression problem. Lin[9] et al proposed a linear cross-modal factor analysis for simple cross-modal associations. Yao et al proposed ranking canonical correlation analysis to maximize the query text and image similarity to find the common subspace. Li[10] et al proposed deep canonical correlation analysis as nonlinear correlation analysis. The above methods are based on canonical correlation analysis method, which only considers the semantic correlation between pairs of text images, and does not consider the possible correlation between different data in the same modality [11][12][13]. Zhuang et al introduced dictionary learning to sparse coding for cross-modal search. Ngiam et al used multimodal learning task to learn multimodal features and used deep networks to learn cross-modal features. Yan et al used deep typical correlation to analyze linear model mappings, which could not explore cross-modal data linkage well. Song et al achieved fast search by XOR operation, which saves memory space, but increases the computational complexity of training [14][15][16].

2.2 Adversarial learning methods

Adversarial learning is used to obtain a generative model close to the original sample distribution by cyclically training the generator and discriminator [17]. Wang et al used adversarial learning to solve the cross-modal search problem by cyclically training the feature mapping and modal discriminator networks. This method is more accurate for cross-modal semantic portrayal[18], but the introduction of the triplet constraint increases the amount of data training and the training speed is slow [19][20][21]. Xu et al proposed deep adversarial metric learning, which introduces a deep network model for richer feature extraction, but only introduces adversarial learning to increase the cross-modal search regularization, making inter-modal semantics missing. Liang[22] et al proposed unsupervised cross-modal adversarial learning, which can utilize unlabeled data, but cross-modal semantics and homomodal semantics learning is more difficult, there is no uniform semantics as training targets, and it is not suitable for complex data. Li[23] et al proposed self-supervised adversarial hashing networks, using Hash coding to solve the cross-modal problem. Jiang et al proposed multi-feature fusion network to solve multi-feature cross-modal problem. But the fused features come from the same modality, which can not deal with the features of the different modalities of the video [24][25][26].

3 The Proposed Method

The FFACR method proposed in this paper includes a video multimodal feature fusion network, a feature mapping network, and a modal discrimination network.

The feature mapping network is optimized by minimizing the semantic deviation of the same modal data and the modal deviation of the same semantic data. The modal discrimination network is optimized by minimizing the error of the original modal decision of the data after mapping. The feature mapping network and the modal discriminant network are trained by adversarial learning. The structure is shown in Figure 1.

![Figure 1 FFACR network structure](image)

3.1 Natural semantic video segmentation

The natural semantic video segmentation method refers to the use of automatic speech recognition (ASR) information of technology videos with timestamp information to slice a video segment with a time period of a speaker’s sentence. Since most of the information of the technology video comes from speech, segmenting the long video according to the semantics of speech can better capture the semantic differences of the video segments.

For each video clip, we intercept the first and last frames of the clip as the image features of this sample, and the optical character recognition (OCR) of the current frame is spliced with ASR text as the text features of this sample. All videos are pre-processed above to form several image-text pairs, and the belongings are used as category labels during training, together with the video description text to form the sample set.

3.2 Feature fusion network

For the video vector \( v_i \), it needs to be obtained from
the image (video frame) feature vector $i_j$ and the text feature vector $t_i$ by a multimodal feature fusion network. Denote $v_i = f(v_i, t_i; \theta)$ as feature fusion process, where $f(v_i, t_i; \theta)$ is the fusion mapping function of image and text features, and $\theta$ is the parameters of the fusion network. For the fusion network structure, we will try three different structures to explore the difference of feature representation performance.

Through training, feature fusion networks that can effectively fuse and represent video multimodal information can be obtained to provide good video semantic representation information for subsequent cross-modal search tasks.

3.3 Feature mapping network

In order to unify the semantics of text and video across modalities, a feature mapping network is introduced to map the data of both modalities into the common semantic space $S$ through the feature mapping network. The feature mapping of text is $S_T = f_T(T; \theta_T)$, and the feature mapping of video is $S_V = f_V(V; \theta_V)$. $\theta_T$ and $\theta_V$ are the parameters representing the text feature network and the video feature network respectively. $S_T$ and $S_V$ denote the new features of text and video feature mapping in $S$ respectively, $S_T, S_V \in \mathbb{R}^{m\times n}$. The FFACR method proposed in this paper obtains suitable $S_T$ and $S_V$ in the common semantic space $S$ so that they maintain the semantic relationship before mapping, and the performance of the mapping is determined by the deviation of the semantic feature distribution of the input and output contents, and the feature mapping network is trained by the difference. Under the premise of maintaining the semantic invariance, it makes the data with similar semantics closer in $S$ and the different semantic data with the same modality farther in $S$.

The feature mapping network is divided into the text feature mapping network and the video feature mapping network, which are responsible for mapping the original data features into the same semantic space. In order to ensure that the mapped data maintain the original semantic features, a semantic prediction network is added after the feature mapping network, and the output of the classifier softmax is used as the result to predict the semantic distribution of the data mapped to the same semantic space. Let the parameters of the network be $\theta_{imd}$, and the $c$ imth dimensional values of the $i$ imth data semantic distribution in the text and video modalities are $p_{im}(t_i)$ and $p_{im}(v_i)$ respectively. And use the cross-entropy to calculate the deviation value of the semantics $L_{imd}$ in the subspace $S$, whose expression is

$$L_{imd}(\theta_{imd}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{c}(y_{ic} \log p_{im}(t_i) + \log p_{im}(v_i))$$

where $L_{imd}$ calculates the difference between the semantic distribution of each newly mapped data and the original data in the new semantic space, including the sum of the differences between the text and video components, in order to ensure that the data with similar semantics in the same modality are still close to each other in the new space $S$, and the data with farther semantics are still farther away in the new space after the transformation of $f_T(T; \theta_T)$ and $f_V(V; \theta_V)$.

In order to ensure that the data under different modalities are close to each other after feature mapping, and the data with different semantics of different modalities are far away from each other, the semantic similarity matrix $Sim_{i} \in \mathbb{R}^{m\times n}$ is constructed by using the semantic distribution of the original data based on the calculation of the semantic distribution of the input content $l_{1, n}$. The semantic distributions of any two data are $l_{a}$ and $l_{b}$, respectively, and their similarity is defined as

$$sim(l_{a}, l_{b}) = \frac{\sum_{i=1}^{d} l_{ai} \times l_{bi}}{\|l_{a}\| \cdot \|l_{b}\|} = \frac{\sum_{i=1}^{d} l_{ai} \times l_{bi}}{\sqrt{\sum_{i=1}^{d} (l_{ai})^2} \times \sqrt{\sum_{i=1}^{d} (l_{bi})^2}}$$

In this paper, $L_2$ parametric is chosen to describe the difference between two similarity matrices, and the difference value is defined as the modal deviation value $L_{imi}(\theta_{imi})$.

The overall loss function of the feature mapping network is defined as $L_{emb}$, by a linear weighted summation of the semantic deviations $L_{imd}$ and the modal deviations $L_{imi}$, where $\alpha$ and $\beta$ denote the contribution of the two deviation values to the loss function, respectively, and the mapping loss is calculated as shown below.

$$L_{emb}(\theta_{im}, \theta_{im}, \theta_{imd}) = \alpha \cdot L_{imd} + \beta \cdot L_{imi}$$

where $\alpha$ and $\beta$ are used as hyperparameters of the network to determine the optimal values by subsequent experiments.

3.4 Modal discriminant network

The modal Discriminant network mainly distinguishes the original modality of the data mapped to the common semantic space. Let the label of the data after mapping
through text be 0 and the label of the data after mapping through video be 1. The modal discrimination network determines the original modality of the data as accurately as possible. A neural network is used for the computation and the loss function of this network is defined as the deviation value of the modal prediction. The loss function $L_{adv}$ is calculated as shown below.

$$
L_{adv}(\theta_p) = -\frac{1}{n} \sum_{i=1}^{n} \log(1 - D(y_i; \theta_p)) + \log(D(x_i; \theta_p)) \\
= -\frac{1}{n} \sum_{i=1}^{n} \log(1 - D(f_i(t_i; \theta_p); \theta_p)) + \log(D(f_i(v_i; \theta_p); \theta_p))
$$

(4)

where $\theta_p$ is the modal discriminant network parameter, and $D(x, \theta_p)$ represents the probability that the network determines $X$ is a video.

The training process of adversarial learning consists of the co-training of the feature mapping network and the modal discrimination network. The goal of the feature mapping network is to maintain the semantic elimination of modality, and the goal of the modality discriminator is to distinguish the modality of different data in the common semantic space.

4 Experimental Results and Analysis

4.1 Datasets

In this paper, two technology video datasets are constructed independently, which are shown in Table I. The subject categories covered by datasets include computer, chemistry, biology, etc. The data sources include Wanfang Data Knowledge Service Platform, CCF China Software Conference video, etc. Each sample contains the images of the first and last frame of the video clip, OCR and ASR information during the clip time period; video description information as text modality, and video number as the real category information.

| Dataset Name                  | Number of Samples | Number of Labels |
|-------------------------------|-------------------|------------------|
| Wanfang Technology Video      | 3752              | 64               |
| Technology Lecture Video      | 6124              | 123              |

4.2 Comparative experiments of different methods

The MAP index is used to measure the performance of different cross-modal search methods, the MAP values are calculated from the top 5, top 10 and top 30 search results for the text-to-video task on the technology video dataset. The results of the comparison methods on the technology video dataset are shown in Table II and III.

As shown in Table II and III, the MAP metrics of the proposed FFACR method are better than the other methods. It can be observed that the performance of using video multimodal fusion features is better than that of video single-modal features, proving the effectiveness of it. The results are consistent with the previous assumption that the main source of technological video information is textual modality.

![Figure 2](image_url)

(a) PR curve on Wanfang technology video dataset (b) PR curve on technology lecture video dataset

We also use PR (precision recall) curves in Figure 2, to measure the effect of cross-modal search. As shown in Figures 2, the FFACR method outperforms the comparison algorithm on both datasets and is higher than FFACR_t and FFACR_f. The gap between FFACR_t and FFACR_f is more obvious on the technology lecture video dataset in Figure 2(b).

5 Conclusion

In this paper, we propose a novel feature fusion based adversarial cross-modal retrieval method (FFACR) to achieve text-to-video searching. FFACR method using adversarial learning to train a feature fusion network, a feature mapping network and a modal discrimination
network alternately, preserving semantic distributions and eliminating modal differences. Experimental results on technology video datasets show the effectiveness of the proposed method.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62192784, No.62172056, No.61877006).

Reference

[1] Li A, Du J, Kou F, Xue Z, Xu X, Xu M, Jiang Y. Scientific and Technological Information Oriented Semantics-adversarial and Media-adversarial Cross-media Retrieval. arXiv preprint arXiv:2203.08615, 2022.
[2] Li Y, Jiang D, Lian R, Wu X, Tan C, Xu Y, Su Z. Heterogeneous Latent Topic Discovery for Semantic Text Mining. IEEE Transactions on Knowledge and Data Engineering, 2021.
[3] Li Y, Liu Y, Yang B, Wu T. Application of interpretable machine learning models for the intelligent decision. Neurocomputing, 333(14):273-283, 2019.
[4] Meng D, Jia Y, Du J, and Yu F. Tracking Algorithms for Multiagent Systems, In IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(10): 1660-1676.
[5] Wei X, Du J, Liang M, and Ye L. Boosting deep attribute learning via support vector regression for fast moving crowd counting. Pattern Recognition Letters, 2019, 119: 12-23.
[6] Wang, Kaiyee, Qiuyue Yin, Wei Wang, Shu Wu, and Liang Wang. "A comprehensive survey on cross-modal retrieval." arXiv preprint arXiv:1607.06215 (2016).
[7] Li W, Jia Y, Du J. Distributed consensus extended Kalman filter: a variance-constrained approach. IET Control Theory & Applications, 11(3): 382-389, 2017.
[8] Xiao S, Shao Y, Li Y, et al. LECF: recommendation via learnable edge collaborative filtering. Science China Information Sciences, 65(1):1-15, 2022.
[9] Lin P, Jia Y, Du J, Yu F. Average consensus for networks of continuous-time agents with delayed information and jointly-connected topologies. 2009 American Control Conference, 2009: 3884-3889.
[10] Li W, Jia Y, and Du J. Tobit Kalman filter with time-correlated multiplicative measurement noise. IET Control Theory & Applications, 2016, 11(1): 122-128.
[11] Peng Yuxin, Huang X, and Zhao Y. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. IEEE Transactions on circuits and systems for video technology 28, no. 9 (2017): 2372-2385.
[12] Kou F, Du J, Yang C, Shi Y, Cui W, Liang M, and Geng Y. Hashtag recommendation based on multi-features of microblogs. Journal of Computer Science and Technology, 2018, 33(4): 711-726.
[13] Shao Y, Huang S, Li Y, et al. Memory-aware framework for fast and scalable second-order random walk over billion-edge natural graphs. The VLDB Journal, 30(5), 769-797, 2021.
[14] Peng Y, Zhu W, Zhao Y, Xu C, Huang Q, Lu H, Zheng Q, Huang T, and Gao W. "Cross-media analysis and reasoning: advances and directions." Frontiers of Information Technology & Electronic Engineering 18, no. 1 (2017): 44-57.
[15] Li Q, Du J, Song F, Wang C, Liu H, Lu C. Region-based multi-focus image fusion using the local spatial frequency. 2013 25th Chinese control and decision conference (CCDC), 2013: 3792-3796.
[16] Li W, Sun J, Jia Y, Du J, and Fu X. Variance-constrained state estimation for nonlinear complex networks with uncertain coupling strength. Digital Signal Processing, 2017, 67: 107-115.
[17] Meng D, Jia Y, and Du J. Robust iterative learning protocols for finite-time consensus of multi-agent systems with interval uncertain topologies. International Journal of Systems Science, 2015, 46(5): 857-871.
[18] Guan Z, Li Y, Xue Z, Liu Y, Gao H, Shao Y. Federated Graph Neural Network for Cross-graph Node Classification. In 2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems, 418-422, 2021.
[19] Deng C, Chen Z, Liu X, Gao X, and Tao D. Triplet-based deep hashing network for cross-modal retrieval. IEEE Transactions on Image Processing 27, no. 8 (2018): 3893-3903.
[20] Li M, Jia Y, and Du J. LPV control with decoupling performance of 4WS vehicles under velocity-varying motion. IEEE Transactions on Control Systems Technology 2014, 22(5): 1708-1724.
[21] Xu L, Du J, Li Q. Image fusion based on nonsubsampled contourlet transform and saliency-motivated pulse coupled neural networks. Mathematical Problems in Engineering, 2013.
[22] Liang M, Du J, Yang C, et al. Cross-Media Semantic Correlation Learning Based on Deep Hash Network and Semantic Expansion for Social Network Cross-Media Search. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(9): 3634-3648.
[23] Li Y, Yuan Y, Wang Y, et al. Distributed Multimodal Path Queries. IEEE Transactions on Knowledge and Data Engineering, 34(7):3196-321, 2022.
[24] Rehman, Sadaqat Ur, Muhammad Waqas, Shankshan Tu, Anis Koubaa, Obaid ur Rehman, Jawad Ahmad, Muhammad Hanif, and Zhu Han. Deep Learning Techniques for Future Intelligent Cross-Media Retrieval. CISTER-Research Centre in Realtime and Embedded Computing Systems, 2020.
[25] Li Y, Zeng I, Niu Z, Shi J, Wang Z, and Guan Z. Predicting vehicle fuel consumption based on multi-view deep neural network, Neurocomputing, 502:140-147, 2022.
[26] Cao T, Xu C, Du J, Li Y, Xiao H, Gong C, Zhong L, Niyato D. Reliable and efficient multimedia service optimization for edge computing-based 5G networks: game theoretic approaches. IEEE Transactions on Network and Service Management, 17(3):1610-1625, 2020.