MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://mmc.committees.comsoc.org/

MMTC Communications – Review

IEEE COMMUNICATIONS SOCIETY

Vol. 10, No. 5, October 2019

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the October 2019 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises four reviews that cover multiple facets of multimedia communication research including semantics and locality preserving correlation projections, resource allocation in MCC, secure edge caching, and visual and language representation. These reviews are briefly introduced below.

The first paper is published in IEEE ICME 2017 and edited by Dr. Yan Hua. It leverages correlation learning to capture the relationship between heterogeneous data available on the Internet.

The second paper is published in IEEE IEEE Transactions on Multimedia 2018 and edited by Dr. Jinbo Xiong. The paper studies the resource management issue in the emerging mobile cloud computing (MCC).

The third paper is published in IEEE Internet of Things Journal 2019 and edited by Dongfeng Fang. It focuses on securing edge caching in mobile cyber-physical system (MCPS) with blockchain techniques. It proposes a more efficient caching scheme for MCPS to take full advantage of the limited caching capacity of edge nodes.

The last paper is published in ACM Transactions on Multimedia Computing, Communications, and Applications 2019 and edited by Dr. Rui Wang. The paper studies video captioning and proposes three effective methods to generate more accurate and semantic sentences.

All the authors, nominators, reviewers, editors, and others who contribute to the release of this issue deserve appreciation with thanks.

IEEE ComSoc MMTC Communications – Review Directors

Qing Yang
University of North Texas, USA
Email: qing.yang@unt.edu

Roger Zimmermann
National University of Singapore, Singapore
Email: rogerz@comp.nus.edu.sg

Wei Wang
San Diego State University, USA
Email: wwang@mail.sdsu.edu

Zhou Su
Shanghai University, China
Email: zhousu@ieee.org

A Semantic and Local Correlation Method

A short review for "Semantics and Locality Preserving Correlation Projections"

Edited by Yan Hua

Y. Hua, J. Du, Y. Zhu and P. Shi, "Semantics and Locality Preserving Correlation Projections," 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, 2017, pp. 913-918.

Massive amounts of multimedia data, including image, video, text and audio, have been emerging on the Internet. Different kinds of data could describe relevant topics, for example, images and their surrounding texts on web pages, faces captured at different poses and sentences of multilingual language. Correlation learning methods aim to learn the relationship between these heterogeneous data and achieve tasks in computer vision and multimedia field, such as face recognition [1] and cross-media retrieval [2]. In particular, in cross-media retrieval scenarios where the data are from different media, a lot of correlation learning methods have been studied [3].

Correlation learning usually maps heterogeneous data into a common subspace, in which cross-media distance can be directly compared and retrieval is thus achieved by ranking. Correlation mapping model and the relationship between multimedia data utilized for model learning are two research keystones, such as linear mapping with one-to-one correspondence in classical canonical correlation analysis method [4] and deep learning with classification in semantic matching method [5].

In the paper, a multi-label semantics and locality preserving correlation projection method is proposed, which learns a semantic common subspace by view-specific projections from both intra-view and inter-view perspectives. Since multimedia data are described in heterogeneous feature space, view-specific projection matrixes are firstly utilized to transform the features into a comparable subspace. With the comparable representations, the distances of samples are forced to satisfy semantic relationship and local structure of original data by optimizing the parameters of the projections. The semantic relationship is provided by multi-labels of multimedia data, which is utilized for inter-view and intra-view modeling. The locality item aims to preserve local neighborhood structure of intra-view data by learning the projections.

Thus, the authors' main contributions in modeling aspect are that 1) the complicated semantic relationship is incorporated into projection learning when inter-view modeling, and 2) the

semantics and the local structure of data are both taken into account when intra-view modeling.

In the paper, the semantic relationship is preserved both in inter-view and intra-view data. For example, two pairs from image and text dataset shown in their paper, one pair (Image 1 and Text 1) is annotated with labels "glacier, mountain, nighttime, person, sign, sky, sun, sunset" and another pair (Image 2 and Text 2) with "mountain, sky, sunset". The multi-label information is not only shared between the inter-view samples (Image 1 vs. Text 1, Image 1 vs. Text 2, Image 2 vs. Text 1 and Image 2 vs. Text 2), but also between intra-view samples (Image 1 vs. Image 2 and Text 1 vs. Text 2). The more same semantic labels the samples have, the closer they are expected to be. In other words, semantically similar samples should be close in the learned subspace. Thus, semantic constraints between inter-view and intra-view data are constructed when learning the projections.

In addition, local neighborhood structure of intra-view data is preserved in the common subspace. It is an extension of single-view data analysis method Locality Preserving Projections (LPP) to multi-view scenario. More than a simple extension, semantics and locality are simultaneously preserved in the common subspace when intra-view modeling. The samples in local neighbor with more similar semantics are expected to be closer. A product of the original feature similarity and the semantic similarity is utilized to represent the similar relationship. The samples with larger similarity are learned to be closer in the common subspace. For intra-view modeling, it is proven effective to simultaneously consider the semantic and local constraints.

The optimization is a generalized eigenvalue problem by concatenating the view-specific projections and computing its corresponding Laplacian matrix. In the paper, experiments are conducted by retrieval tasks on image and text data. When validating the proposed method, the experiments are designed by varying the parameters of local neighbor. They also validate the effect of multi-label semantics on inter-view and intra-view modeling. Related methods utilizing one-to-one corresponding relationship and classification information are compared. On two retrieval tasks, i.e., querying with image to retrieve texts and querying with text to retrieve images, the proposed method realizes compelling good performances.

Image features are extracted with pre-trained deep learning model in the experiments. It is now a trend to construct end-to-end deep learning model in computer vision and multimedia retrieval field. Semantic and local relationship studied in the paper could be utilized and extended to design reasonable loss function in deep model. In summary, a correlation learning method is proposed, which learns view-specific projections by imposing restrictions to preserve the semantics and local structure of multi-view data. With the projections, multimedia data are

mapped as comparable representations and retrieval is achieved by distance ranking.

References:

[1] M. Kan, S. Shan, H. Zhang, S. Lao and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188–194, Jan. 2016.

[2] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," *2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, pp. 3441–3450, 2015.

[3] Y. Peng, X. Huang, Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[4] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," *Proceedings of the 18th ACM international conference on Multimedia*, New York, NY, USA, pp. 251–260, 2010.

[5] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu and S. Yan, "Cross-modal retrieval with CNN visual features: A new baseline", *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

Yan Hua, Ph.D., is an Assistant Professor in School of Information and Communication Engineering at Communication University of China. She received the B.S. degree in Communication Engineering from China Agricultural University in 2010, and the Ph.D. degree from Beijing University of Posts and Telecommunications in 2015. Her research interests include multimedia retrieval, image understanding and machine learning. She has published papers in prestigious journals such as IEEE Transactions on Multimedia and Neurocomputing, in prestigious conferences such as IEEE ICDM and IEEE ICME.

A Robust Joint Resource Reservation and Allocation Algorithm in MCC

A short review for "Joint Optimization of Radio and Virtual Machine Resources with Uncertain User Demands in Mobile Cloud Computing"

Edited by Jinbo Xiong

Y. Li., J. Liu, B. Cao, and C. Wang, " Joint Optimization of Radio and Virtual Machine Resources with Uncertain User Demands in Mobile Cloud Computing," IEEE Transactions on Multimedia, vol. 20, no. 9, Sept. 2018.

The wide application of mobile cloud computing (MCC) has greatly improved the data transmission efficiency of mobile network, and greatly improved the user service experience of voice or video traffic in mobile terminals. However, because of the dynamic nature of the MCC scenario, its resource management is a major challenge [1]. For cloud virtual machine resources (VMRs) and radio resources (RRs) required by mobile applications, cloud service providers (CSPs) currently provide on-demand provisioning and reservation plan. The resource reservation scheme can reduce the total provisioning cost by 30% compared with on-demand provisioning while its biggest challenge is that the resource demand of mobile users cannot be determined or predicted in advance [2]. Therefore, the effective resource reservation methods are needed urgently to make reasonable resource reservation decisions.

Existing works on resource reservation schemes in cloud computing can be divided into three categories: the determined, the prediction-based and the non-prediction-based. However, the determined schemes have the precondition that it is assumed the user's resource requirements are subject to a certain distribution [3]. The prediction-based schemes are applicable to various networks, but their performance depends on the accuracy of the user demand prediction, and the dynamic characteristics of mobile Internet will reduce its efficiency [4]. In the uncertain and non-prediction situation, the robust optimization method is an important branch of uncertain optimization problems [5]. In recent years, it has been widely applied in the design of user resource reservation schemes that are not based on prediction, but relevant researches are limited. The above related works considered the reservation algorithms of RRs or VMRs, ignoring the uncertainty of mobile user demand. Therefore, how to consider the dual impact of the two factors on the quality of service (QoS) of mobile users is a key problem to be solved in MCC resource management.

In this paper, a robust joint resource reservation and allocation algorithm in MCC (JRRA-MCC) is proposed on the basis of the robust optimization method [5], which solves the optimization problem of joint resource reservation under the uncertainty of user demand and minimizes the total cost of CSP resource allocation. The algorithm considers the joint impact of RRs and VMRs on mobile user satisfaction comprehensively, and analyzes in detail the deterministic optimization model with known user requirements, the robust joint resource reservation optimization model with uncertain user requirements, and the optimal reservation decision with uncertain requirements. Moreover, by transforming the uncertain resource reservation model into an equivalent robust optimization model with certain resource requirements, the proposed algorithm can greatly reduce the complexity of the solution and effectively obtain the robust optimal solution.

Therefore, the authors' major contribution is to propose a robust joint resource reservation and allocation algorithm in MCC that considering both the uncertainties of mobile users' demands to the resources and the dual influence of RRs and VMRs to the QoS of mobile users. In the proposed algorithm, the authors formally describe the dual impact of RRs and VMR on the QoS of mobile users by designing the user satisfaction function. In addition, the uncertain resource reservation problem is transformed into a robust optimization model with adjustable control coefficient, and the optimal reservation decision for the disturbance range of uncertain resource demand is realized.

MCC's resource provisioning system architecture in this paper is composed of mobile users, wireless access points, cloud data centers and mobile cloud controllers. Before the mobile user's request arrives, the CSP makes a predetermined decision to retain the RRs and VMRs at a lower sequential price. When the request arrives, the service is provided at the usage price. However, in the event of a shortage of reserved resources, a high over-demand price will be applied at this stage. In the above system architecture, CSP is concerned about the over-provisioning cost caused by excessive reserved resources and the temporary scheduling cost of additional resources caused by insufficient reserved resources. The rate of data transfer and data processing is determined by the RRs and VMRs assigned to the service. The CSP in this paper combines the two factors to quantify the service satisfaction of mobile users. The author chooses logarithmic utility function as the user satisfaction function to be more appropriate, which is in good accordance with the selection rules of economics. The factors employed to measure the performance of RRs allocation are the two key parameters, minimum reserved traffic rate and maximum sustained traffic rate, which follow the QoS standard protocol. Users' satisfaction function design of VMRs

is combined with the minimum maximum demand threshold to consider the minimum retention data processing rate and the maximum continuous data processing rate.

Moreover, the key step of the proposed JRRA-MCC is to make the optimal resource reservation decision for RRs and VMR of resource reservation for mobile users with uncertain resource demands, so as to carry out resource matching. In the proposed algorithm, the influence of constraint control coefficient and perturbation ranges of resource demands of mobile users on the total cost of resource provisioning is well considered.

Extensive experiments demonstrate the improved performance of the proposed solutions. Simulation results show that the proposed optimal resource reservation decision of JRRA-MCC can make accurate resource reservation decision when the resource demand of mobile users is uncertain. The robustness of JRRA-MCC can better immune to the uncertainty of RRs and VMRs demands. Compared with other existing schemes, the proposed solution can achieve the better resource allocation efficiency and lower allocation cost.

In summary, this paper explores the resource reservation and allocation in MCC where the resource demands of mobile users are uncertain, and introduce the robust optimization theory to make the resource reservation decision to minimize the total resource provisioning cost of the CSP. Moreover, the proposed JRRA-MCC is noteworthy for its good consideration of both RRs and VMRs.

References:

[1] V. Ggarwal, V. Gopalakrishnan, R. Jana, et al., "Optimizing cloud resources for delivering IPTV services through virtualization," IEEE Trans. on Multimedia, vol. 15, no. 4, pp. 789-801, 2013.

[2] S. Chaisiri, B. S. Lee, and D. Niyato, "Optimal virtual machine placement across multiple cloud providers," IEEE Asia-Pacific Services Computing Conference, pp. 103-110, Singapore, 2009.

[3] J. Du, C. Jiang, Y. Qian, et al., "Resource allocation with video traffic prediction in cloud-based space systems," IEEE Trans. on Multimedia, vol. 18, no. 5, pp. 820-830, 2016

[4] R. H. Hwang, C. N. Lee, Y. R. Chen, et al., "Cost optimization of elasticity cloud resource," IEEE Trans. on Services Computing, vol. 7, no. 4, pp. 561-574, 2014.

[5] S. Chaisiri, B. S. Lee and D. Niyato, "Robust cloud resource provisioning for cloud computing environments," IEEE International Conference on Service-Oriented Computing and Applications (SOCA), pp. 1-8, Perth, WA, 2010.

Jinbo Xiong, Ph.D, is an Associate Professor in the Fujian Provincial Key Laboratory of Network Security and Cryptology and the College of Mathematics and Informatics at Fujian Normal University. He received the Ph.D. degree in Computer System Architecture from Xidian University, China, in 2013. His research interests include cloud data security, mobile media management, privacy protection, and Internet of Things. He has published papers in prestigious journals such as IEEE TII, IEEE TCC, IEEE IoT J, IEEE TNSE, FGCS and in major International conferences such as IEEE ICCCN, IEEE TrustCom, IEEE HPCC and IEEE ICPADS. He is a member of IEEE.

A Blockchain-Based Secure Edge Caching Method in MCPS

A short review for "Blockchain-Based Trustworthy Edge Caching Scheme for Mobile Cyber Physical System"

Edited by **Dongfeng Fang**

Q. Xu, Z. Su, and Q. Yang, "Blockchain-based trustworthy edge caching scheme for mobile cyber physical system," IEEE Internet of Things Journal, 2019, doi: 10.1109/JIOT.2019.2951007.

The rapid development of the mobile cyber-physical system (MCPS) has greatly improved the intelligent interaction between the cyber system and the physical world, which attracts a large number of mobile users to use various mobile applications [1]. However, due to the high demands for the mobiles users to obtain multifarious contents from the cyberspace, the quality-of-experience (QoE) of users will be degraded because of the delay in obtaining contents. Therefore, the effective content delivery methods are urgently demanded to decrease the delay and enhance the QoE of users.

Existing works on content delivery schemes in cyber-physical system mainly focus on edge caching methods to solve the problems abovementioned. Edge caching is an efficient solution to decrease the delay in content delivery services by caching popular contents users prefer to request on edge nodes that are close to mobile users [2]. However, current edge caching methods adopted in large-scale MCPS may face several challenges. First, the caching capacity of each edge node is limited, which is not able to satisfy a large amount of mobile users' demands. Second, edge nodes with caching contents may be malicious since they are deployed by untrusted entities. Several attacks may be implemented by these malicious edge nodes such as providing fake contents, modifying original contents, adding virus files into contents, etc [3]. Third, a mobile user may also be malicious since the lack of supervision measures during the interaction between edge nodes and mobile users in caching content services. A malicious mobile user may carry out the refuse-to-pay attacks to affect the utility of edge nodes who provide the caching services [4].

Overall, a more efficient caching scheme should be designed to satisfy the large-scale MCPS's demand by taking full advantage of the limited caching capacity of edge nodes. An effective safety protection mechanism should be constituted to precisely monitor the malicious actions implemented by edge nodes and mobile users.

In this paper, a blockchain-based secure edge caching mechanism in MCPS is proposed on the basis of the edge caching method [5], which optimizes the performances of the caching services in large-scale MCPS. On one hand, the utilities of edge nodes can be optimized through the mechanism, and the contents they cache can be well allocated. On the other hand, the delay of content delivery can be decreased which improves the

QoE of mobile users. Through the proposed resource allocation algorithm, the ratio that mobile users can obtain the desire contents can be also improved. Moreover, to further decrease the caching cost for edge nodes and mobile users, this paper constitutes a social group by considering the social relationships between mobile users. Specifically, different mobile users with similar contents interest can join to a social group, in which the contents will be requested by an agent selected by the group. Then, the requested contents can be shared in the social group. In this way, the mobile users can pay less in obtaining contents, and communication loads of edge nodes can also be reduced. In addition, to protect the security of the caching service, this paper presents a trust management scheme for mobile users to select the edge nodes who are not only trustworthy but also provide contents with high quality. Meanwhile, the blockchain is utilized to record the information during the iteration between mobile users and edge nodes to supervise the behaviors they perform in the caching services.

Therefore, the major contributions of this paper are to develop an efficient method to solve the edge caching in large-scale MCPS, and to present a more appropriate security protection scheme in the caching services. Specifically, the authors adopted a layered coding mechanism named scalable video coding (SVC) to code the various contents into different layers, through which edge nodes can cache only a part of the caching content. In this way, the caching capacity of each edge node can be maximized and the content delivery delay can be further decreased. Moreover, to solve the security problems mentioned above, this paper utilizes the blockchain to record the transaction information between mobile users and edge nodes. The actions they adopt in the caching services will be recorded and monitored by a smart contract designed by the authors. In addition, the authors adopted various algorithms to optimize the caching services. Specifically, the gradient descent algorithm is adopted to obtain the optimal caching projects for mobile users considering their historical caching demands. And the max-min-based algorithm is used to allocate the optimal caching contents for mobile users.

Extensive numerical results clearly demonstrate the performance of the proposed mechanism is well improved. The simulation results reveal that the caching price can be optimized by the proposed scheme and the caching demands of mobile users can be satisfied by joining social groups. Moreover, the delay of the caching service can be observably decreased by the proposed scheme. In addition, the proposed edge caching mechanism can achieve lower caching cost and better caching efficiency by comparing with other existing mechanisms.

In summary, this paper explores the secure edge caching in MCPS. It is notable that the authors well addressed the contradiction between the high demands of mobile users in requesting for the caching services and the low caching capacities of edge nodes in a large-scale MCPS. Moreover, the trust management scheme and the blockchain technology are adopted to protect the caching services and the behaviors of mobile users and edge nodes.

References:

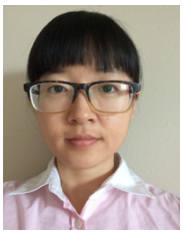
[1] X. Liu, M. Dong, K. Ota, P. Hung, and A. Liu, "Service pricing decision in cyber-physical systems: Insights from game theory," IEEE Trans. Services Comput., vol. 9, no. 2, pp. 186–198, Mar./Apr. 2016.

[2] A. Ndikumana et al., "Joint communication, computation, caching, and control in big data multi-access edge computing," IEEE Trans. Mobile Comput., to be published, doi: 10.1109/TMC.2019.2908403.

[3] V. Sharma, I. You, F. Palmieri, D. N. K. Jayakody, and J. Li, "Secure and energy-efficient handover in fog networks using blockchain-based DMM," IEEE Commun. Mag., vol. 56, no. 5, pp. 22–31, May 2018.

[4] S. Li, J. Xu, M. Schaar, and W. Li, "Trend-aware video caching through online learning," IEEE Trans. Multimedia, vol. 18, no. 12, pp. 2503–2516, Dec. 2016.

[5] Q. Xu, Z. Su, Y. Hui, and Q. Yang, "Caching scheme with edge nodes for mobile cyber-physical systems," in Proc. IEEE DASC/PiCom/DataCom/CyberSciTech, Orlando, FL, USA, 2017, pp. 95–100.



Dongfeng Fang is an Assistant Professor in the Department of Computer Science and Software Engineering, California Polytechnic State University, San Luis Obispo, CA, USA. She received a Ph.D degree from the University of Nebraska – Lincoln, Omaha, NE, USA, 2019, a M.S. degree from Shanghai University, Shanghai, China, 2013, and a B. S. degree from Harbin Institute of Technology, China, 2009. Her current research interests include Cybersecurity on 5G security, criticial infrastructure security, IoT security, and public safety communications.

Rich Visual and Language Representation

A short review for "Rich Visual and Language Representation with Complementary Semantics for Video Captioning"

Edited by Rui Wang

Pengjie Tang, Hanli Wang, Qinyu Li "Rich Visual and Language Representation with Complementary Semantics for Video Captioning", ACM Transactions on Multimedia Computing, Communications, and Applications, Volume 15 Issue 2, June 2019

1. Introduction

Video captioning (or video description) is a challenging task that requires interdisciplinary research efforts in the fields of computer vision and natural language processing. It has wide application prospects such as video retrieval, content search on video sharing, automatic video subtitle generation, aiding for visually impaired

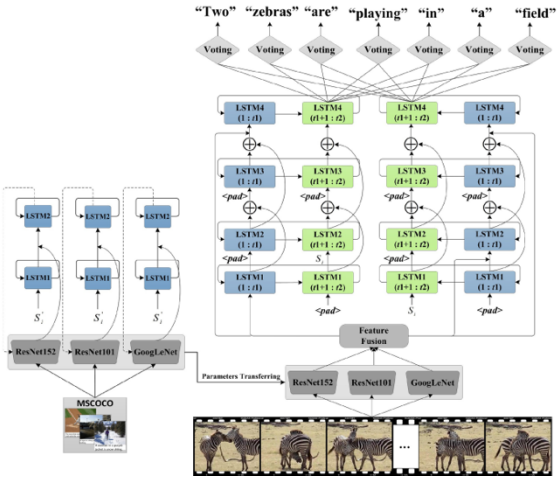


Fig. 1 Overview of the proposed video-captioning framework

people, and so on. In general, a video-captioning model should have the ability of processing variable input frame length and yielding sentences with correct grammar and appropriate structure, especially with semantic meanings. In early conventional methods, visual features are frst extracted in a hand-crafted manner. Then the object, scene, and relationship are detected for flling into a template prepared in advance [1], or analogous videos, fragments, and corresponding words, phrases, and even sentences are retrieved to compose new description sentences for candidate videos [2]. In recent years, the success of convolutional neural network (CNN) techniques brings a breakthrough to this task. A pipeline of encoding and decoding, inspired by machine translation, is usually employed.

2. Proposed LSTM-F2F Model

In our work, a wider and deeper video-captioning model is proposed, in which three effective techniques are developed to generate more accurate and semantic sentences. The framework is shown in Fig. 1.

In the proposed model, CNN features are first extracted from raw RGB frames by deep CNN models trained on MSCOCO2014 [3], a dataset for image captioning, aiming to make the feature not only adaptable to object categories but also sensitive to natural language such as words, sentence patterns, and grammar. To seize temporal features in videos, the pipeline of S2VT [4] is employed, in which the long short term memory (LSTM) network is used to model visual sequence by feeding CNN features into the LSTM network in the order of time step. For the purpose of producing more semantic sentences, an improved factored way for an LSTM network is proposed. The first LSTM layer is used to model language independently, while the encoded visual feature is fed into the second LSTM layer with the hidden representation from the first layer. To further refine the model structure for video captioning, the two aforementioned LSTM architectures with the improved factored way and un-factored way can be combined to generate a wider model to catch more information of the candidate video. Besides model widening, it is also desired to deepen the LSTM network to generate more abstract linguistic representation and boost the model generalization ability. Inspired by the researches of residual network

(ResNet) [5] and sequential ResNet [6], we design intra-module and inter-module residual LSTM networks (LSTM-Res) for rich semantic representation. Furthermore, we introduce the LSTM-Res technique into the LSTM-F2F framework, and build a deeper and wider LSTM network, called Res-F2F. Besides, the fusion of visual features from different CNN models is also explored in this work to grasp more comprehensive information of raw video frames. To this goal, a number of excellent CNN models can be employed, such as GoogLeNet [7], VGG-Net [8], and ResNet [5]. Two benchmark video-captioning datasets of MSVD [9] and MSV-VTT2016 [10] are applied to evaluate the effectiveness of the proposed methods.

To conclude, the main contributions of our work are presented as follows:

- An improved factored way for LSTM architecture is proposed to capture more semantic in formation of candidate videos. Furthermore, the architectures with the improved factored way and un-factored way are integrated, and the corresponding two sequential output prob abilities of predicted words are fused by way of weighted average.
- A residual technique is introduced into the proposed model with the intra-module residual LSTM and inter-module residual LSTM, which are designed to deepen the LSTM network for more abstract feature representation.
- A fusion strategy is proposed to catch complementary information and generate more comprehensive visual representation, which utilizes different visual features from diverse CNN models.

Dataset	C1	C2	C3
MSVD	4.41	3.11	2.66
MSR-VTT2016	4.37	3.78	3.60

Table 1 Performance of the proposed Res-F2F (G-R101-R152) with human evaluation (all the 670 videos from the MSVD dataset and a randomly selected set of 500 videos from the MSR-VTT2016 dataset are evaluated)

3. Experimental Results

A number of typical examples about the generated sentences are displayed in Fig. 2. It can be seen that the generated sentences achieved by the proposed Res-F2F (G-R101-152) model generally contain more accurate and comprehensive information about the target videos than most of the other candidate sentences. However, the choice of words and the building of sentences are still somewhat rigid and inflexible as compared to most of the references. For instance, there are not only initiative sentences but also passive patterns in the references, and the expression is flexible, while only the active formation is employed in the generated sentences.

Fig. 2 Examples of the references and the generated sentences with baseline models (GoogLeNet, ResNet101, and ResNet152) and the proposed LSTM-F2F, LSTM-Res, and Res-F2F (G-R101-152)

Also, human evaluation is conducted on both the MSVDand MSR-VTT2016 datasets. The results are shown in Table 1, which reveals that the generated sentences possess good consistency (C1) with human descriptions. Regarding the metric of relevance (C2), most of the generated sentences perform barely satisfactorily, with the performances reaching to 3.11 and 3.78 on the two datasets, respectively. However, as for the metric of helpful for blind, C3 only reaches to 2.66 on the MSVD dataset, since the generated sentences may be a little bit blurry as compared to

the real content of videos. By comparison, the proposed framework achieves better C3 performances on MSR-VTT2016. The possible reason is that MSR-VTT2016 contains more training samples including videos and the corresponding reference sentences, and thus the model can learn more experiences for description generation.

4. Conclusion

A wider and deeper LSTM network is presented in our work for video captioning by capturing rich semantic visual and language features. In the first place, a novel LSTM module with improved factored way is designed to improve the linguistic representation. On the basis of this module and the un-factored way in S2VT, a wider LSTM network is constructed and a voting strategy is employed to generate sentences with better consistency and semantic expressions. In addition, a deeper LSTM network with a residual mechanism is developed based on the improved factored way. Then, these two techniques are combined to form an advanced model. Moreover, for more comprehensive visual representation, an early fusion strategy is employed, where the CNN features from different deep-model architectures are concatenated into a new and longer visual formation. The proposed methods are testifed to be effective by a series of experiments and analysis on the MSVD and MSR-VTT2016 datasets.

References:

[1] Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond J. Mooney. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the International Conference on Computational Linguistics*. 1218–1227.

[2] Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu. 2010. Multimodal fusion for video search reranking. IEEE Trans. Knowl. Data Eng. 22, 8 (Aug. 2010), 1191–1199.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. 740–755.

[4] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*. 4534–4542.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[6] Kaisheng Xu, Hanli Wang, and Pengjie Tang. 2017. Image captioning with deep LSTM based on sequential residual. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 361–366.

[7] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[8] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

[9] David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the Meeting of the Association for Computational Linguistics*. 190–200.

[10] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5288–5296.

.

MMTC Communications – Review Editorial Board

Multimedia Communications Technical Committee Officers

Chair: Honggang Wang, University of Massachusetts Dartmouth, USA
Steering Committee Chair: Sanjeev Mehrotra, Microsoft Research, US
Vice Chair – America: Pradeep K Atrey, University at Albany, State University of New York, USA
Vice Chair – Asia: Wanqing Li, University of Wollongong, Australia
Vice Chair – Europe: Lingfen Sun, University of Plymouth, UK
Letters & Member Communications: Jun Wu, Tongji University, China
Secretary: Shaoen Wu, Ball State University, USA
Standard Liaison: Guosen Yue, Huawei, USA

MMTC examines systems, applications , services and techniques in which two or more media are used in the same session. These media include, but are not restricted to, voice, video, image, music, data, and executable code. The scope of the committee includes conversational, presentational, and transactional applications and the underlying networking systems to support them.