IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 10, 2020, accepted February 6, 2020, date of publication February 11, 2020, date of current version February 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973240

Mutual Complementarity: Multi-Modal Enhancement Semantic Learning for Micro-Video Scene Recognition

JIE GUO[1], XIUSHAN NIE[2], AND YILONG YIN[3]
[1]School of Computer Science and Technology, Shandong University, Jinan 250101, China
[2]School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China
[3]School of Software, Shandong University, Jinan 250101, China

Corresponding authors: Xiushan Nie (niexsh@hotmail.com) and Yilong Yin (ylyin@sdu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671274, Grant 61573219, Grant 61701281, and Grant 61876098, in part by the National Key Research and Development Program of China under Grant 2018YFC0830102, in part by the China Postdoctoral Science Foundation under Grant 2016M592190, in part by the Shandong Provincial Key Research and Development Plan under Grant 2017CXGC1504, and in part by the Special Funds for Distinguished Professors of Shandong Jianzhu University.

ABSTRACT Scene recognition is one of the hot topics in micro-video understanding, where multi-modal information is commonly used due to its efficient representation ability. However, there are some challenges in the usage of multi-modal information because the semantic consistency among multiple modalities in micro-videos is weaker than in traditional videos, and the influences of multi-modal information in micro-videos are always different. To address these issues, a multi-modal enhancement semantic learning method is proposed for micro-video scene recognition in this study. In the proposed method, the visual modality is considered the main modality whereas other modalities such as text and audio are considered auxiliary modalities. We propose a deep multi-modal fusion network for scene recognition with enhanced the semantics of auxiliary modalities using the main modality. Furthermore, the fusion weight of multi-modal can be adaptively learned in the proposed method. The experiments demonstrate the effectiveness of enhancement and adaptive weight learning in the multi-modal fusion of the micro-video scene recognition.

INDEX TERMS Micro-video scene recognition, multi-modal fusion, semantic enhancement, adaptive weight learning.

I. INTRODUCTION

With the development of social media for mobile, a large number of social media platforms, such as Instagram, Twitter, Wechat and Tiktok have emerged. Similar to images, micro-videos, a new social media type, have become the common means of sharing information among users. Most of these micro-videos are generated by users on social media, and not by professional photographers.

The rapid development of micro-videos as the main media type of social media is mainly attributed to the following characteristics. 1) Shortness: the typical length of micro-videos is a few seconds, which makes them easily available on social media. 2) Social attributes: similar to images on social media platforms, micro-videos also come with many social attributes, such as venue, loop, description, hashtag, follower number, and click number. These social attributes are useful for micro-video understanding. 3) User generated: most micro-videos are generated by users on social media, and not by professional photographers. These users capture micro-videos based on their emotions and feelings, which exhibit a high degree of subjectivity. The extracted high-level semantic information is more consistent with human subjective intention. Owing to the interesting characteristics of micro-videos, significant efforts have been made toward micro-video-related research such as action recognition [1], tag prediction, popularity prediction [2], and venue recognition [3], [22], [25]. Scene recognition is also a critical factor for micro-video understanding. Therefore, the focus of this study was on the micro-video scene recognition.

Different from traditional videos, micro-videos come with hashtags and comments. As textual information, these

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Cheng.

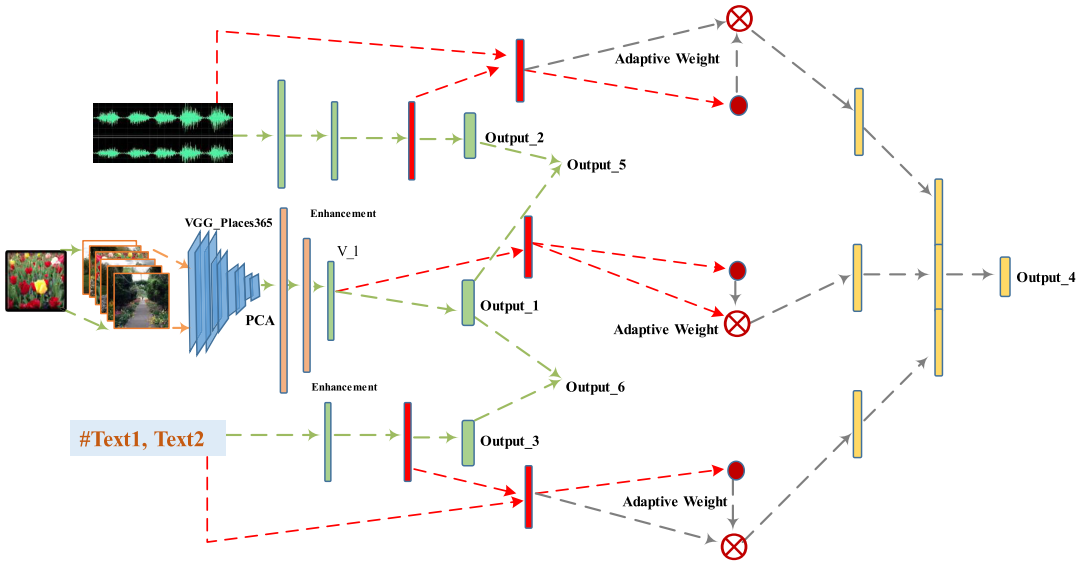

FIGURE 1. The pipline of proposed MESL method.The green dotted line represents the enhancement component, the red dotted line represents the adaptive weight learning component, and the gray dotted line represents the multi-modal fusion component.

attributes aid micro-video scene recognition. Along with the visual and audio information in the videos, these three modalities can be fused for the micro-video scene recognition. In traditional videos, different modalities should have a common subspace to represent the scene category. However, in micro-videos, multiple modalities are more complementary in addition to common high-level semantics. Consequently, there are two challenges for multi-modal fusion in micro-videos. 1) For most micro-videos, visual modality plays a major role in scene recognition, which is called the ''main modality''. The other two modalities used to aid recognition, which are called the ''auxiliary modalities''. However, the common semantics between the two auxiliary modalities and visual modality is weak. Therefore, visual representation can be used to weakly enhance the semantic representation of the other two modalities. 2) For a small number of micro-videos, visual modality cannot directly reflect the scene category, but audio or textual modality can directly obtain the scene category. Therefore, to improve the accuracy of the micro-video scene recognition, multiple modalities need to be fused, and the fused weights need to be set automatically when they are fused.

To address these challenges, a multi-modal enhancement semantic learning (MESL) method is proposed in this study for the micro-video scene recognition, as illustrated in Figure 1. For the first challenge, the proposed MESL method minimizes the distance between visual modality and other modalities in semantics space. This method not only activates the common semantic representation, but also retains the characteristics of the other two modalities. To address the second challenge, a mechanism for adaptive learning weights is applied in the final multi-modal fusion.

The contributions of the proposed method can be summarized as follows:

1) In this study, a semantic enhancement mechanism is used between main modal and auxiliary modalities. It not only activates the common semantic representation, but also retains the characteristics of the auxiliary modalities.

2) A mechanism for adaptive learning weights is applied in the final multi-modal fusion.

3) A MESL method is proposed in this study. It not only strengthens the role of the main modality, but also retains the characteristics of other modalities. Additionally, it adaptively determines the fusion weights to better learn the semantics of micro-video scenes.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related works. Section 3 presents the proposed MESL method. Section 4 details the experimental evaluation and results, followed by the conclusions and scope for future work in Section 5.

II. RELATED WORK

In this section, several studies related to micro-videos and multi-modal fusion methods are reviewed.

A. RELATED RESEARCH IN MICRO-VIDEOS

Due to the characteristics and challenges of micro-video understanding, more researchers are focusing on micro-video understanding related research such as creative prediction, action recognition, tag prediction, popularity prediction, venue recognition, micro-video recommendation [4]–[7] and micro-video understanding [8].

In 2014, Redi *et al.* [9] studied creative micro-videos to understand the features that make a video creative.

The features of creative micro-video such as filmmaking technique and HSV statistics are designed by researchers. Meanwhile, Sano *et al.* [10] also designed features for the degree of loop assessment. These are explainable features, but they are not suitable for complex tasks. In 2016, Chen *et al.* [2] began to combine manual features with feature extraction (e.g. sentence2vector) and deep features(e.g., CNN) for popularity prediction, and venue category estimation. They also used social attributes and multi-modal fusion features. These features can represent high-lever semantics.

However, most of these features are common in the fields of computer vision and multimedia understanding. In 2017, Liu *et al.* [8] considered the characteristics of micro-video, including low quality and sparseness for micro-video understanding. To extract a better multi-view feature for the micro-video popularity prediction, Jing *et al.* [11] integrated the low-rank multi-view embedding and regression analysis into a unified framework such that the lowest-rank representation shared by all views not only captures the global structure of all views, but also indicates the regression requirements. These methods usually project multiple modalities into a common subspace. However, because micro-videos are user-generated data, the complementarity between multiple modalities should be considered. Guo *et al.* [12] fused the multi-modal features using a multi-layer perceptron network, and learnt the complementary feature. However, the feature could not distinguish between the common and complementary parts. To resolve this problem, Wei *et al.* [13] proposed a neural multi-modal cooperative learning (NMCL) method that focused on how to explicitly separate the consistent and complementary features to improve the expressiveness of each modality.

B. RELATED RESEARCH ONN MULTI-MODAL FUSION

The multi-modal features are widely used in traditional video retrieval using subspace learning. These methods are generally divided into two categories, i.e., unsupervised and supervised methods. The unsupervised methods include Canonical Correlation Analysis (CCA) [14], Partial Least Squares (PLS) [15], Bilinear Model (BLM) [16], and Deep Canonical Correlation Analysis (DCCA) [17], whereas the supervised methods include Generalized Multi-view Analysis (GMA) [18], Multi-view Discriminant Analysis (MvDA) [19], Multiple Feature Hashing (MFH) [20], and Semantic Correlation Maximization (SCM) [21]. In the field of micro-video understanding, there exist many models that used multi-modal fusion, such as TURMANN [22], which also learned a common subspace. In general, the subspace learning-based methods focus on exploring a common space for all features. However, when applied to the micro-video scene recognition, the performance of the model is not acceptable due to the complementarity between the multi-modal features. Until now, only a small number of studies have focusd on complementarity between multiple modalities. Wei et al. proposed an NMCL [13] method that focused on how to explicitly separate the consistent and complementary features.

III. PROPOSED METHOD

In this study, the proposed method includes three components. 1) Semantic enhancement of auxiliary modalities: in this component, the semantics of two auxiliary modalities are weakly enhanced by using visual modality. 2) Complementary fusion: to retain the characteristics of the auxiliary features, the original features are concatenated to the enhanced features. 3) Multi-modal fusion based on adaptive weights. We assume that there are n samples in the training set, $X = (X_1, X_2, \ldots, X_n)$, X_i is the feature of i-th sample. For each sample, there are three modalities, $X_i = (X_i^v, X_i^a, X_i^t)$, and $X_i^S \in R^{D^S}$, $s \in \{v, a, t\}$. The label set of n samples is $y = (y_1, y_2, \ldots, y_n)$, and the number of categories is C.

A. SEMANTICS ENHANCEMENT OF AUXILIARY MODALITIES

In most micro-videos, we believe that the visual modality contains stronger semantics, and semantics in the audio and textual modalities is weak. However, different modalities of the same micro-video, should represent the same high-level semantics. Therefore, audio and textual modalities are considered auxiliary modalities, and weakly enhanced by visual modality. In the proposed method, the distance between the visual modal and other modalities in high-level semantics space should be as small as possible.

The high-level semantics representations in the visual, audio, and textual modal is $y_i^{v\_out}$, $y_i^{a\_out}$, and $y_i^{t\_out}$, respectively. As shown in Figure 1, they are the vector representations of "output_1", "output_2", and "output_3", respectively. They are nonlinear transform of original features. The object function is obtained by distance minimization:

$$\min_W \alpha ||y_i^{v\_out} - y_i^{a\_out}||_2 + \beta ||y_i^{v\_out} - y_i^{t\_out}||_2 \quad (1)$$

where, α and β are trade-off parameters of two terms, and W is the weight of network.

After being enhanced, we use the cross entropy loss function to calculate the loss of three outputs:

$$Loss\_v = -\sum_{i=1}^{n}\sum_{m=1}^{C} y_i^{v\_out} \ln y_i + (1 - y_i^{v\_out}) \ln(1 - y_i) \quad (2)$$

$$Loss\_a = -\sum_{i=1}^{n}\sum_{m=1}^{C} y_i^{a\_out} \ln y_i + (1 - y_i^{a\_out}) \ln(1 - y_i) \quad (3)$$

$$Loss\_t = -\sum_{i=1}^{n}\sum_{m=1}^{C} y_i^{t\_out} \ln y_i + (1 - y_i^{t\_out}) \ln(1 - y_i) \quad (4)$$

B. COMPLEMENTARY FUSION

As mentioned in subsection A, the semantics of two auxiliary modalities is weakly enhanced by visual modality. Therefore, these two auxiliary modalities, while being guided in the semantic direction by the visual modality, retain their own

characteristics that are complementary to the visual modality. However, to ensure that the auxiliary modal characteristics can be completely retained, the original features of these two modalities are also concatenated to the enhanced features $X_i^{a_{penu}}$ and $X_i^{t_{penu}}$, which are the penultimate layers of the audio and text networks. The new complementary features are shown as follow:

$$X_i^{a\_conc} = conc(X_i^{a_{penu}}, X_i^a) \tag{5}$$
$$X_i^{t\_conc} = conc(X_i^{t_{penu}}, X_i^t) \tag{6}$$

where, conc() is the concatenate operation.

C. MULTI-MODAL FUSION BASED ON ADAPTIVE WEIGHTS

After obtaining the above two components, new features of three modalities are extracted. For most samples, the weight of the visual modality should be relatively large during the fusion process. However, for a few samples, the weights of the two auxiliary modalities should be larger. To adaptively obtained the fusion weights according to the samples, this study adopts the adaptive weights fusion method. The adaptive weights are shown as follows:

$$X_i^{v\_fu\_w} = \sigma(W^{v_4} \cdot \sigma(W^{v_3} \cdot X_i^{v_{penu}})) \tag{7}$$
$$X_i^{a\_fu\_w} = \sigma(W^{a_4} \cdot \sigma(W^{a_3} \cdot X_i^{a\_conc})) \tag{8}$$
$$X_i^{t\_fu\_w} = \sigma(W^{t_4} \cdot \sigma(W^{t_3} \cdot X_i^{t\_conc})) \tag{9}$$

where, $X_i^{v\_fu\_w} \in R^1$, $X_i^{a\_fu\_w} \in R^1$, $X_i^{t\_fu\_w} \in R^1$, and σ is the sigmoid function. The fusion process is that each modal is weighted separately and then concatenated with the others.

$$X_i^{v\_fus} = X_i^{v\_fu\_w} \odot (\sigma(W^{v_3} \cdot X_i^{v_{penu}})) \tag{10}$$
$$X_i^{a\_fus} = X_i^{a\_fu\_w} \odot (\sigma(W^{a_3} \cdot X_i^{a\_conc})) \tag{11}$$
$$X_i^{t\_fus} = X_i^{t\_fu\_w} \odot (\sigma(W^{t_3} \cdot X_i^{t\_conc})) \tag{12}$$
$$X_i^{fus} = conc(X_i^{v\_fus}, X_i^{a\_fus}, X_i^{t\_fus}) \tag{13}$$
$$y_i^{fus} = softmax(W^{fus} \cdot X_i^{fus}) \tag{14}$$

where, X_i^{fus} is the weighted fusion feature, and y_i^{fus} is the predicted category. We also use the cross entropy loss function to calculate the fusion loss:

$$Loss\_fus = -\sum_{i=1}^{n}\sum_{m=1}^{C} y_i^{fus} \ln y_i + (1 - y_i^{fus}) \ln(1 - y_i) \tag{15}$$

The final loss is the weighted sum of six losses:

$$Loss = \lambda_1 \cdot Loss\_v + \lambda_2 \cdot Loss\_a + \lambda_3 \cdot Loss\_t + \lambda_4 \cdot Distance\_va + \lambda_5 \cdot Distance\_vt + \lambda_6 \cdot Loss\_fus \tag{16}$$

where, λ_i represent the trade−off parameters, and ''Distance_va'' and ''Distance_vt'' are the two terms in Equ.1. The loss function is optimized using the stochastic gradient descent (SGD) method.

IV. EXPERIMENTS

A. DATASET AND FEATURE

In this study, we use a dataset, which is publicly accessible on the website.[1] This dataset is published for the micro-video venue recognition and micro-video understanding. This dataset contains 188 categories, and the number of samples in each class is unbalanced. Some categories are subordinate to others, and data is hierarchical. In this study, our main aim is to find a better general multimodal fusion method; hence, we do not consider the hierarchical relationship between the categories. Therefore, dataset from 10 categories are selected for scene recognition. The number of samples in each category ranges from 100 - 2000, and each video has a duration of approximately 6 s. To maintain the real distribution of micro-videos, the new dataset is unbalanced within classes. Prior to the experiment, the dataset is preprocessed using the Synthetic Minority Over-sampling Technique (SMOTE), that is used to solve the imbalance problem by oversampling.

Each micro-video in this dataset includes three modalities, visual, audio and textual. In this study, the features are extracted using VGG16_places365 [24], denoising autoencoder, and sentence2vector, respectively. VGG16_places365[2] is a pre-trained VGG16 network used for image scene recognition in Places365 dataset, that is a public dataset for image scene recognition. In this study, the original visual feature is extracted using the VGG16_places365 network. The dimensions of original features for these three modalities are 128, 200, and 100, respectively.

B. PARAMETER SELECTION

In this study, the main parameters are trade-off parameters in loss function. The parameters are selected by many experiments. The final values are listed in TABLE 1. The values in the last column are better than others values.

TABLE 1. Trad-off parameters.

Loss_v	1
Loss_a	0.5
Loss_t	0.5
Loss_fusion	0.5
Distance_va	0.4
Distance_vt	1

C. COMPARISON WITH BASELINES

We compare the performance of the proposed method with some baselines, including classic subspace learning methods CCA and MvDA-VC, and a new NMCL method that takes into account both consistency and complementarity.

Canonical Correlation Analysis (CCA) [14]. CCA is a typical unsupervised approach to obtain a common space, which attempts to learn two transforms to project the samples from

[1] www.acmmm16.wixsite.com/mm16

[2] https://github.com/GKalliatakis/Keras-VGG16-places365/blob/master/vgg16_places_365.py

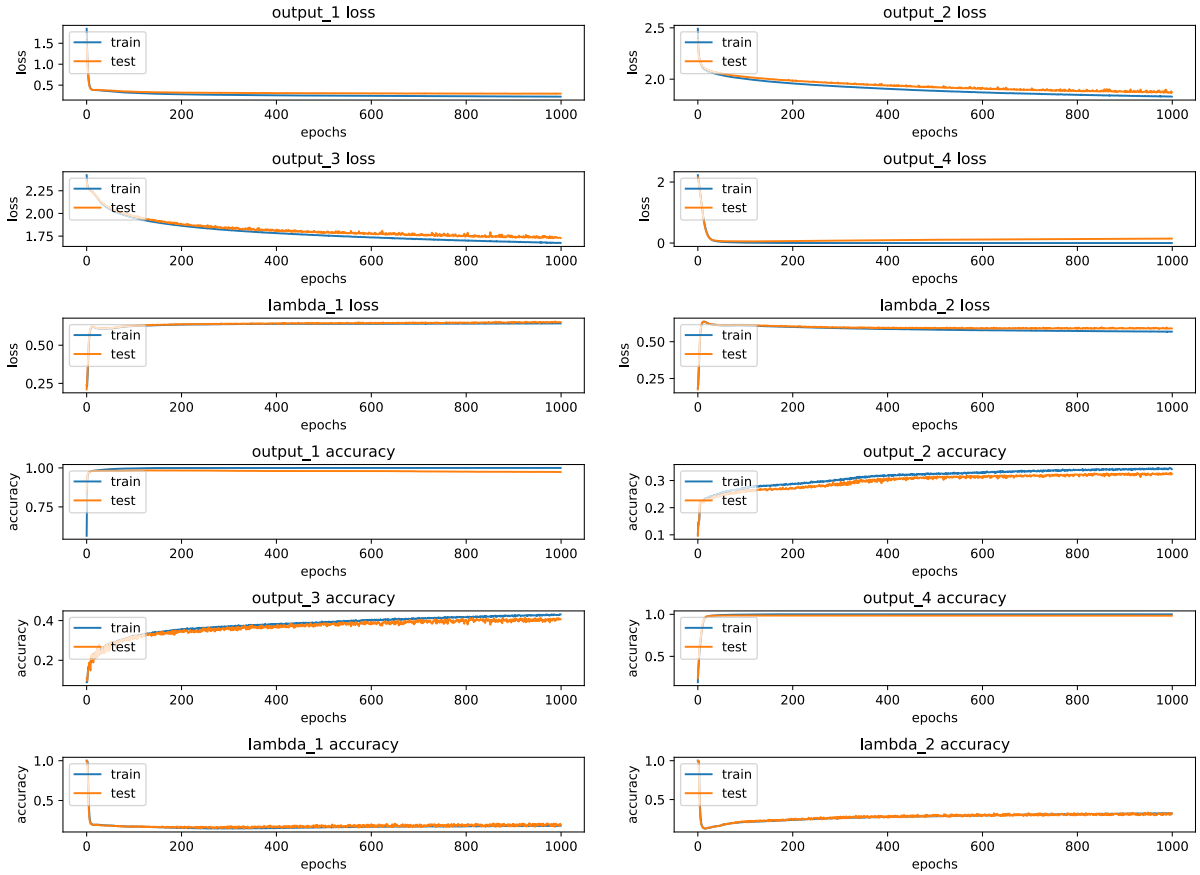

FIGURE 2. The figures of loss terms and accuracy. "output_1", "output_2", "output_3", "output_4" present "Loss_v", "output_a", "output_t", "output_fusion". "lambda_1 loss" and "lambda_2 loss" present "Distance_va" and "Distance_vt".

two views into a common subspace by maximizing the cross-correlation between the two views. In this study, the baseline is CCA_3V [23]. It is canonical correlation analysis of three views.

Multi-view Discriminant Analysis with View Consistence (MvDA_VC) [19]. MvDA_VC seeks a single discriminant common space for multiple views in a non-pairwise manner by jointly learning multiple view-specific linear transforms.

Neural Multimodal Cooperative Learning (NMCL) [13]. This method splits the consistent and complementary components using the relation-aware attention mechanism. It takes into account both consistency and complementarity of the multiple modalities.

In this study, accuracy is used as a metric for the performance of the methods. The comparison results with baseline methods are shown in TABLE 2. The performance of the proposed MESL method is better than the baseline methods. The reason why the proposed MESL method in this paper is effective is that it takes into account the consistency and complementarity of multiple modalities. Meanwhile, it applies the main modalities to the auxiliary modalities to enhance the semantics of the auxiliary modalities while retaining the characteristics of the auxiliary modalities. CCA is an unsupervised method that focuses only on consistency between multiple modalities. MvDA_VC is better than CCA, because it takes into account the discrimination of mono modal. However, this method ignores the importance between modalities. The role of auxiliary modalities is ignored, and the performance of the main modal determines the performance of the model. The NMCL method considers both consistency and complementarity of multiple modalities, which are treated equally to learn consistency and complementarity with other modalities. In fact, the importance of multiple modalities for semantic learning is different. The main modal has an enhanced effect on the auxiliary mode, and the auxiliary modal has a complementary effect on the main modal.

TABLE 2. Comparison with baselines.

Method	Accuracy
CCA_3V [23]	0.6488
MvDA_VC [19]	0.9744
NMCL [13]	0.8047
The proposed MESL	*0.9826*

D. ABLATION STUDIES

In this section, we compare the performance of each mono modal before and after the enhancement, and the performance

of mono modal and multi-modal. The comparison results are summarized in TABLE 3 and TABLE 4. As shown in TABLE 3, after enhancement, the performance of audio and text modal is better than before enhancement. However, visual modal is not. The reason is that semantics of these two auxiliary modalities are weak, this is also a characteristic of the micro-video data. As shown in TABLE 4, the performance of the multi-modal fusion is better than the mono modal.

TABLE 3. Enhancement studies.

Modal	Acc_before_enhancement	Acc_after_enhancement
Audio	0.3286	*0.3427*
Text	0.4153	*0.4210*
Visual	*0.9816*	0.9697

TABLE 4. Ablation studies.

Modal	Accuracy
Audio	0.3427
Text	0.4210
Visual	0.9697
Visual+Audio+Text	*0.9826*

E. CONVERGENCE

This experiment is conducted to verify the convergence of the MESL. The loss and accuracy of six outputs are obtained during the learning process. As shown in Figure 2, the upper three rows present the loss figures of the loss functions, and the last three rows present the accuracy figures. The blue line indicates the results of the training process, and the orange line indicates the results of the testing process. With an increase in the number of epochs, the gradients of the two lines approached 0. In Figure 2, ''lambda_1 loss'' and ''lambda_2 loss'' represents ''Distance_va'' and ''Distance_vt'', respectively. They are convergent, but loss goes up a bit. This is because the model needs to balance a large amount of loss. However, the value of loss is still small, and the overall discrimination accuracy is improved.

V. CONCLUSION

Different from the traditional videos, there are many challenges of multi-modal fusion in micro-videos. For most micro-videos, visual modality is very important. However, the common semantics between the auxiliary modalities and visual modality is weak. To resolve this issue, a semantic enhancement strategy is proposed to ensure the auxiliary modalities contain more semantic information. For a small number of micro-videos, auxiliary modalities are more important than the main modal. To overcome this problem, an adaptive weight learning method is proposed for the multi-modal fusion. In this study, the proposed multi-modal enhancement semantics learning method can combine consistency and complementarity in the multi-modal fusion, and adaptively learn fusion weights. The experiments demonstrate the effectiveness of enhancement and adaptive weight learning in the multi-modal fusion of the micro-video scene recognition.

ACKNOWLEDGMENT

The authors would especially like to thank the editors and the reviewers for their valuable comments and suggestions.

REFERENCES

[1] P. Xuan Nguyen, G. Rogez, C. Fowlkes, and D. Ramanan, ''The open world of micro-videos,'' 2016, *arXiv:1603.09439*. [Online]. Available: http://arxiv.org/abs/1603.09439

[2] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, and T.-S. Chua, ''Micro tells macro: Predicting the popularity of micro-videos via a transductive model,'' in *Proc. ACM Int. Conf. Multimedia*, Oct. 2016, pp. 898–907.

[3] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian, ''Enhancing micro-video understanding by harnessing external sounds,'' in *Proc. ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1192–1200.

[4] Z. Cheng and J. Shen, ''On effective location-aware music recommendation,'' *ACM Trans. Inf. Syst.*, vol. 34, no. 2, pp. 1–32, Apr. 2016.

[5] Y. Li, M. Liu, J. Yin, C. Cui, X.-S. Xu, and L. Nie, ''Routing micro-videos via a temporal graph-guided recommendation system,'' in *Proc. ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1464–1472.

[6] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, and L. Nie, ''Personalized hashtag recommendation for micro-videos,'' in *Proc. ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1446–1454.

[7] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T. S. Chua, ''MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video,'' in *Proc. ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1437–1445.

[8] M. Liu, L. Nie, and M. Wang, ''Towards micro-video understanding by joint sequential-sparse modeling,'' in *Proc. ACM Int. Conf. Multimedia*, Oct. 2017, pp. 970–978.

[9] M. Redi, N. Ohare, R. Schifanella, M. Trevisiol, and A. Jaimes, ''6 seconds of sound and vision: Creativity in micro-videos,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4272–4279, Jun. 2014.

[10] S. Sano, T. Yamasaki, and K. Aizawa, ''Degree of loop assessment in micro-video,'' in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 5182–5186.

[11] P. Jing, Y. Su, and L. Nie, ''Low-rank multi-view embedding learning for micro-video popularity prediction,'' *IEEE Trans. Knol. Data. Eng.*, vol. 30, no. 8, pp. 1519–1532, Dec. 2017, doi: 10.1109/TKDE.2017.2785784.

[12] J. Guo, X. Nie, and C. Cui, ''Getting more from one attractive scene: Venue retrieval in micro-videos,'' in *Proc. PacRim. Conf. Multimedia*, Sep. 2018, pp. 721–733.

[13] Y. Wei, X. Wang, W. Guan, and L. Nie, ''Neural multimodal cooperative learning toward micro-video understanding,'' *IEEE Trans. Image. Process.*, vol. 29, pp. 1–14, Jul. 2020, doi: 10.1109/TIP.2019.2923608.

[14] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, and E. al, ''A new approach to cross-modal multimedia retrieval,'' in *Proc. ACM Int. Conf. Multimedia*, Oct. 2010, pp. 251–260.

[15] R. Rosipal and N. Krämer, ''Overview and recent advances in partial least squares,'' in *Proc. Int. Conf. Subspace Latent. Struct. Feature Sel.*, 2005, pp. 34–51.

[16] J. B. Tenenbaum and W. T. Freeman, ''Separating style and content with bilinear models,'' *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, Jun. 2000, doi: 10.1162/089976600300015349.

[17] G. Andrew, R. Arora, J. Bilmes, and K. K. Livescu, ''Deep canonical correlation analysis,'' in *Proc. Int. Conf. Mach. Learn.*, Jun. 2013, pp. 1247–1255.

[18] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, ''Generalized multiview analysis: A discriminative latent space,'' in *Proc. IEEE Conf. Comput. Vis. Patt. Recog.*, Jun. 2012, pp. 2160–2167.

[19] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, ''Multi-view discriminant analysis,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016, doi: 10.1109/tpami.2015.2435740.

[20] 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, doi: 10.1109/tmm.2013.2271746.

[21] D. Zhang and W. J. Li, ''Large-scale supervised multimodal hashing with semantic correlation maximization,'' in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2014, pp. 2177–2183.

[22] J. Zhang, L. Nie, and X. Wang, ''Shorter-is-better: Venue category estimation from micro-video,'' in *Proc. ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1415–1424.

[23] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," 2016, *arXiv:1607.06215*. [Online]. Available: http://arxiv.org/abs/1607.06215
[24] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018, doi: 10.1109/tpami.2017.2723009.
[25] M. Liu, L. Nie, X. Wang, Q. Tian, and B. Chen, "Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1235–1247, Mar. 2019, doi: 10.1109/tip.2018.2875363.

JIE GUO received the M.S. degree from the School of Information Science and Engineering, Shandong Normal University. She is currently pursuing the Ph.D. degree with Shandong University. Her research interests include machine learning, pattern recognition, and multimedia retrieval.

XIUSHAN NIE received the Ph.D. degree from Shandong University, Jinan, China, in 2011. From 2013 to 2014, he was a Visiting Scholar with the University of Missouri, Columbia, USA. He is currently a Professor with Shandong Jianzhu University, Jinan. His research interests include data mining, multimedia retrieval, and indexing and computer vision.

YILONG YIN received the Ph.D. degree from Jilin University, Changchun, China, in 2000. From 2000 to 2002, he was a Postdoctoral Fellow with the Department of Electronic Science and Engineering, Nanjing University, Nanjing, China. He is currently the Director of the Machine Learning and Applications Group and a Professor with Shandong University, Jinan, China. His research interests include machine learning, data mining, computational medicine, and biometrics.