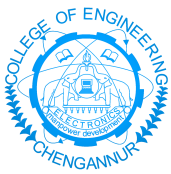
Mutual Complementarity: Multi-Modal Enhancement Semantic Learning for Micro-Video Scene Recognition

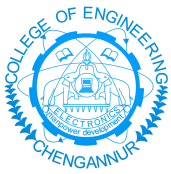
03CS7903 Seminar I

08/MCS/2020 CHN20MT019 Sree Lekshmi B S.
M. Tech. Computer Science & Engineering
(Image Processing)



Department of Computer Engineering
College of Engineering Chengannur
Alappuzha 689121
Phone: +91.479.2165706
http://www.ceconline.edu
hodcs@ceconline.edu

College of Engineering Chengannur
Dept. of Computer Engineering



C E R T I F I C A T E

This is to certify that, this report titled ***Mutual Complementarity: Multi-Modal Enhancement Semantic Learning for Micro-Video Scene Recognition*** is a bonafide record of the **03CS7903 Seminar I** presented on March 18, 2021 by

08/MCS/2020 CHN20MT019 Sree Lekshmi B S.

First Semester M. Tech. Computer Science & Engineering (Image Processing)

scholar, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, **M. Tech. Computer Science & Engineering (Image Processing)** of **APJ Abdul Kalam Technological University**.

Guide

Ms.Jyothi R
Assistant Professor
in Computer Engineering

Coordinator

Mr.Ahammed Siraj K K
Associate Professor
in Computer Engineering

Head of the Department

March 18, 2021

Dr. Smithadharan
Professor
in Computer Engineering

Acknowledgments

Primarily, I thank Lord Almighty for his eternal support through out my seminar work.

I express my sincere thanks to **Dr. Jacob Thomas V**, Principal, College of Engineering Chengannur for extending all the facilities required for doing my seminar. My heartfelt words of gratitude to **Dr. Smitha Dharan**, Professor and Head of Department of Computer Engineering, for providing constant support.

Now I express my gratitude to my seminar co-ordinator **Mr. Ahammed Siraj K K**, Associate Professor in Computer Engineering and my seminar guide **Ms. Jyothi R L**, Assistant Professor in Computer Engineering who played a great role for valuable suggestions and expert guidance.

Sree Lekshmi B S.

Abstract

Scene recognition in micro-videos is one of the interesting topics of micro-video understanding. Micro-videos are 6-15 sec video clips mostly found in social media platforms like tiktok, instagram, twitter and Wechat. This work comes under the area of video analysis, where it extract metadata from raw video and used as components for further processing in applications such as search, summarization, classification or event detection. Comparing to traditional videos, the influences of multi-modal information in micro-videos are always different. In micro videos the semantic consistency among multiple modalities weaker. 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. To address these issues, a multi-modal enhancement semantic learning method is introduced in this study for micro-video scene recognition. Here the visual modality is considered the main modality and other modalities such as text and audio are considered as auxiliary modalities. A deep multi-modal fusion network is adopted for scene recognition with enhanced the semantics of auxiliary modalities using the main modality. The fusion weight of multi-modal can be adaptively learned in this method. The experiments demonstrate the effectiveness of enhancement and adaptive weight learning in the multi-modal fusion of the micro-video for scene recognition. This work aims to propose a better multi-modal fusion method.

Contents

Chapter 1

Introduction

Micro-videos is 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 micro-videos in social media has following characteristics. 1) Shortness: the typical length of micro-videos is 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 and follower 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 on mobile phones based on their emotions and feelings[10].

Many efforts have been made towards micro-video-related research such as action recognition, tag prediction, popularity prediction, and venue recognition[8]. Scene recognition is also a critical factor for micro-video understanding [2]. Therefore, the focus of this study was on the micro-video scene recognition. Along with the visual and audio information, micro-videos come with hashtags and comments. So for scene recognition, textual, audio and visual modalities are to be fused. In traditional videos, different modalities should have a common subspace to represent the scene category. But 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 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.

To address these issues, a multi-modal enhancement semantic learning (MESL) method is used for the micro-video scene recognition[1]. For the first issue, the 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 issue, a mechanism of adaptive learning weights is applied in the final multi-modal fusion.

The overview of the proposed method is that, it consists of three components that are, semantic enhancement of auxiliary modalities, complementary fusion and multi-modal fusion based on adaptive weights. A dataset is used, which is publicly accessible on the website. The main aim of this work is to find a better general multi-modal fusion method. The features are extracted and training is done using VGG16_places365. VGG16_places3652 is a pre-trained VGG16 network used for image scene recognition in Places365 dataset, that is a public dataset for image scene recognition. The original visual feature is extracted using the VGG16_places365 network.

Chapter 2

Literature survey

1. **Enhancing Micro-Video Understanding by Harnessing External Sounds**, [4], in Proc. ACM International Conference on Multimedia, Oct. 2017, pp. 1192–1200.

   Here they focus on enhancing acoustic modality for venue categorization task. According to their experiments, the acoustic modality demonstrates the weakest capability in indicating the venue information. So they tries to enhance acoustic modality. They conducted a user study to explore the influence of acoustic signal on estimating the venue category. For 59% of micro-videos, the acoustic modality can benefit the venue category estimation. For 84% of micro-videos acoustic information are insufficient to reflect the venue category accurately. This study lends support to the usefulness of acoustic information of micro-videos. But its implementation is difficult due to following challenges. 1) There is no suitable sound data for micro-videos, as existing labelled data are either too small to cover the common acoustic concepts. 2) External sounds are unimodal data, whereas micro-videos unify textual, visual, and acoustic modalities to describe a real-life event. It is technically challenging to fuse the unimodal sound data to improve the learning of multi-modal video data. To address these issues they design a Deep trAnsfeR modEl (DARE), which jointly leverages external sounds to strengthen the acoustic concept learning. They first extract features for each modality and then project the features of each modality with a dedicated mapping matrix to obtain the high-level representations. To transfer the external sound knowledge, they apply the same acoustic feature extractor on the labelled audio clips and use the same mapping matrix as the acoustic modality. Then they concatenate the representations of three modalities and feed it into a deep neural network with multiple hidden layers, which can capture the non-linear correlations among concepts.Then ultimately feed the fused representations into a prediction function to estimate the venue categories. They used a publicly accessible bench mark dataset. Although this model shows better, they doesn't put much attention to video and textual modalities and also this model does not consider the problem of missing modalities. Also their complementary features are not considered during multi-modal fusion.

2. **Neural Multimodal Cooperative Learning Toward Micro-Video Understanding**, [2], IEEE Transaction on Image Processing, vol. 29, pp. 1–14, July 2020.

   Here introduced a better multi-modal fusion method for venue catogory estimation. Most of the works about micro video understanding is limited in exploring consistency between different modalities. They proposed a method called Neural Multimodal Cooperative Learning(NMCL) for separating the consistent and complementary components by using an attention mechanism. They explained how to explicitly handle and separate the consistent and complementary features from the mixed information. This work considers the coorporative relationships comprising of consistent and complementary components. Consistent components means the same information appearing in more than one modality in different forms. Some challenges they found in multimodal coorperative learning is 1)Consistent and complementary components are often mixed. 2) After separation how could we associate them with each other. So a deep multimodal coorporative learning is employed that can model the correlation between different modalities and enhance the representation of each modality to estimate the venue categorization of micro videos. The features are firstly extracted from each modality and fed into three coorperative peer nets. In each coorporative net, they consider one modality as the host and the rest as the guest. Coorporative network separate consistent parts and complementary parts. The two consistent parts are fused with deep neural networks and the fusion result is ultimately concatenated with two complementary parts. Output of each cooperative nets is augmented feature vectors. Each vector is then fed into attention net followed by softmax function then late fusion. Then prediction is done. NMCL method focus on how to explicitly separate the consistent and complementary features to improve the expressiveness of each modality.

3. **Towards Micro-Video Understanding by Joint Sequential-Sparse Modeling**, [3], in Proc. ACM International Conference on Multimedia, Oct. 2017, pp. 970–978.

   Here proposed a model called EASTERN, that better characterize and jointly model the sparseness and multiple sequential structures for micro-video understanding. This work presents an end-to-end deep learning model, which packs three parallel LSTMs to capture the sequential structures and a convolutional neural network to learn the sparse concepts. For each modality separate LSTM are used for finding sequential structures of three modalities in parallel. Using a common mapping function, they are mapped onto a common space. After that the three projected vectors with the same length is fed into a convolutional neural network to learn the sparse and conceptual representations. Then use softmax function for classification. Comparing to traditional videos, micro-videos may have some high level concepts. So they need to learn sparse and conceptual representations. EASTERN gives better result for venue categorization.

But this model does not handle some properties of micro-videos such as inconsistency and proper multi-modal fusion.

4. **Joint learning of NNeXtVLAD, CNN and Context Gating for Micro-Video Venue Classification** [9]. IEEE Access, August 2019, pp . 40950–40962.

   Here presents an improved neural network architecture, Normalized NeXtVLAD (NNeXtVLAD) with ReLU function and L2 normalization. NeXtVLAD is an effective network that aggregates frame level features into a compact supervector. But the capability of such a supervector is still limited due to the lack of non-linear transformation and L2 normalization. so they introduce NNeXtVLAD. They applied NNextVLAD network as a three stream architecture to aggregate acoustic, visual and textual features. Firstly they uses three independent NNeXtVLAD networks to aggregate frame level features into a compact feature vector with common dimensions on visual, acoustic, and textual modalities in parallel. Secondly they pack the three vectors with the same length as an input and then apply a CNN layer to extract their sparse and conceptual representations. Thirdly a context gating is introduced for modeling the dependency among labels. Finally, a softmax classifier is used for venue classification. NNeXtVLAD produce a dominant result compared to concurrent methods(LSTM, GRU). This method is very complex and doesn't manage loss of information.

5. **"Shorter is Better: Venue Category Estimation from Micro-Videos"** ,[5], in Proc. ACM Int. Conf. Multimedia, Oct. 2016, pp. 1415–1424.

   Here attempts for venue category estimation. For this problem, they aim to label the bite-sized video clips with venue categories. This has mainly three challenges. 1) no available benchmark dataset. 2) insufficient information, low quality, and information loss and 3) complex relatedness among venue categories. To address these issues, this work propose a scheme comprising of two components. They first do data preparation and extract features from textual, visual and acoustic modalities of each videos. Then Fusion of multi-modal features are performed to handle low quality of some modalities. Second is Label micro videos with venue categories. Towards this end they present a tree guided multi task multi-modal learning model called TRUMANN. This model learns a common feature space from multi-modal heterogeneous spaces and utilizes the learned common space to represent each micro video. TRUMANN treats each venue category as a task and leverages the predefined hierarchical structure of venue categories to recognize the relatedness among task.

6. **Multi-view discriminant Analysis** [6], IEEE Transaction on Pattern Analysis Mach. Intell., Jan. 2016, vol. 38, no. 1, pp. 188–194.

Here introduces some methods to hande heterogeneous recognition or cross view matching problem. In many computer vision applications, the same object can be observed at various viewpoints or even by heterogeneous sensors, thus generating multiple distinct even heterogeneous images. Recently, more and more applications need to match images from different viewpoints or different sensors, usually denoted as heterogeneous recognition or cross-view recognition. Due to the large gap between views, the samples from different views might lie in completely different spaces. Therefore, directly matching the samples from different views is no longer applicable. To address the above mentioned heterogeneous recognition problem, one need either transform samples of different views into a common space or learn distance metrics that can match heterogeneous samples of various views. As these two methodologies can be equivalently converted in some cases, this work focuses on the former, i.e., learning a common subspace shared by various views. This line of methods can be further grouped into two categories: two-view methods and multi-view methods. The multi-view methods attempt to seek for a single unified common space shared by all views. In contrast, the two-view methods essentially can only obtain a common space for two views, but can also be extended to address multiple views problem.

Chapter 3

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

This method consists of 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. Assume that there are n samples in the training set $X = (X_1, X_2, ..., X_n)$. For each sample, there are three modalities $x_i = x_i^a$, x_i^v and x_i^t. The label set of n samples $i_s^y = (y_1, y_2, ..., y_n)$ and the number of categories is C. The pipeline of the proposed method can be seen in Figure 1. 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.

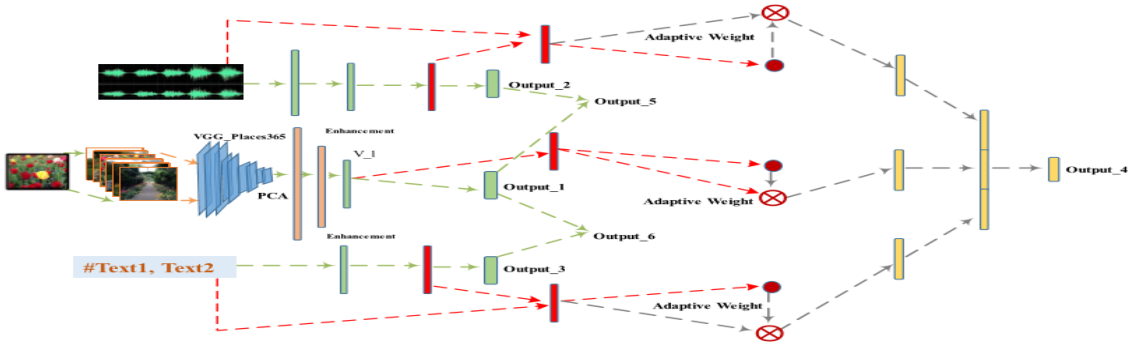


Figure 1: The pipline of MESL method.

3.1 Semantic enhancement of auxiliary modalities

In most micro-videos, 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 this 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.

$$\min_W \alpha||y_i^{v\_out} - y_i^{a\_out}||2 + \beta||y_i^{v\_out} - y_i^{t\_out}||2 \tag{1}$$

where α and β are trade-off parameters of two terms and W is the weight of network. After being enhanced, cross entropy loss function is used to calculate the loss of three outputs:

$$Loss\_v = -\sum_{i=1}^{n}\sum_{m=1}^{c} y_i^{v\_out} \ln y_i + (1 - y_i^{v\_out}) \ln(1 - y_i) \tag{2}$$
$$Loss\_a = -\sum_{i=1}^{n}\sum_{m=1}^{c} y_i^{a\_out} \ln y_i + (1 - y_i^{a\_out}) \ln(1 - y_i) \tag{3}$$
$$Loss\_t = -\sum_{i=1}^{n}\sum_{m=1}^{c} y_i^{t\_out} \ln y_i + (1 - y_i^{t\_out}) \ln(1 - y_i) \tag{4}$$

3.2 Complementary fusion

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. We refer to the consistent component as the same information appearing in more than one modality in different forms. As shown in Figure 1, a red candy displaying in the visual text of "lollipop" describe the consistency. By contrast, the complementary component represents the exclusive information appearing only in one modality. For instance, it is hard to find the equivalent in other modalities in Figure 1 of the textual concept of "girl" or the visual concept of "grass". 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.

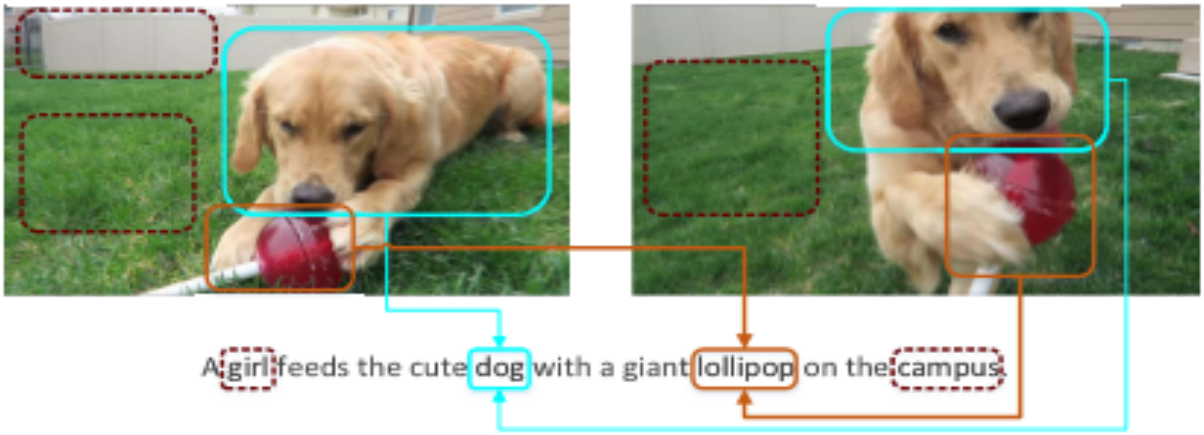


Figure 1: Exemplar demonstration of the correlation between the visual modality and textural modality. The blue and brown boxes show the consistent information and the red dashed boxes show the complementary information.

3.3 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. But 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:

$$X_i^{v\_fu\_w} = \sigma(W^{v4}.\sigma(W^{v3}.X_i^{v\_penu})) \tag{7}$$
$$X_i^{a\_fu\_w} = \sigma(W^{v4}.\sigma(W^{a3}.X_i^{a\_penu})) \tag{8}$$
$$X_i^{t\_fu\_w} = \sigma(W^{v4}.\sigma(W^{t3}.X_i^{t\_penu})) \tag{9}$$

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

$$X_i^{v\_fu} = X_i^{v\_fu\_w} \odot (\sigma(W^{v3}.X_i^{v\_penu})) \quad (10)$$
$$X_i^{a\_fu} = X_i^{a\_fu\_w} \odot (\sigma(W^{v3}.X_i^{a\_penu})) \quad (11)$$
$$X_i^{t\_fu} = X_i^{t\_fu\_w} \odot (\sigma(W^{v3}.X_i^{t\_penu})) \quad (12)$$

$$X_i^{fus} = conc(X_i^{v\_fus}, X_i^{a\_fus}, X_i^{a}) \quad (13)$$
$$X_i^{fus} = softmax(W^{fus}.X_i^{fus}) \quad (14)$$

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:

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

The final loss is the weighted sum of six losses:

$$\begin{aligned} \mathrm{Loss} = {} & \lambda_1.Loss\_v + \lambda_1.Loss\_a + \lambda_1.Loss\_t + \\ & \lambda_1.Distance\_va + \lambda_1.Distance\_vt + \lambda_1.loss\_fus \end{aligned} \quad (16)$$

where,λ_i represent the trade off parameters and "Distance_ va" and "Distance_vt" are the two terms in equation 1. The loss function is optimized using the stochastic gradient descent (SGD) method.

3.4 Training of the model

Each micro-video in the dataset includes three modalities -visual, audio and textual. In this study, the features are extracted using VGG16_places365, denoising autoencoder and sentence2 vector respectively. VGG16_places 365 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. In this study, our main aim is to find a better general multimodal fusion method. So 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.

Chapter 4

Conclusions

4.1 Results

4.1.1 Dataset

The dataset used is publicly accessible on the website. 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. Our main aim is to find a better general multi-modal 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. The dataset is preprocessed using the Synthetic Minority Over-sampling Technique(SMOTE), that is used to solve the imbalance problem by oversampling.

4.1.2 Training

Each micro-video in the dataset includes three modalities, visual, audio and textual. In this study, the features are extracted using VGG16_places365, denoising autoencoder, and sentence2 vector, respectively. VGG16_places 365 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.

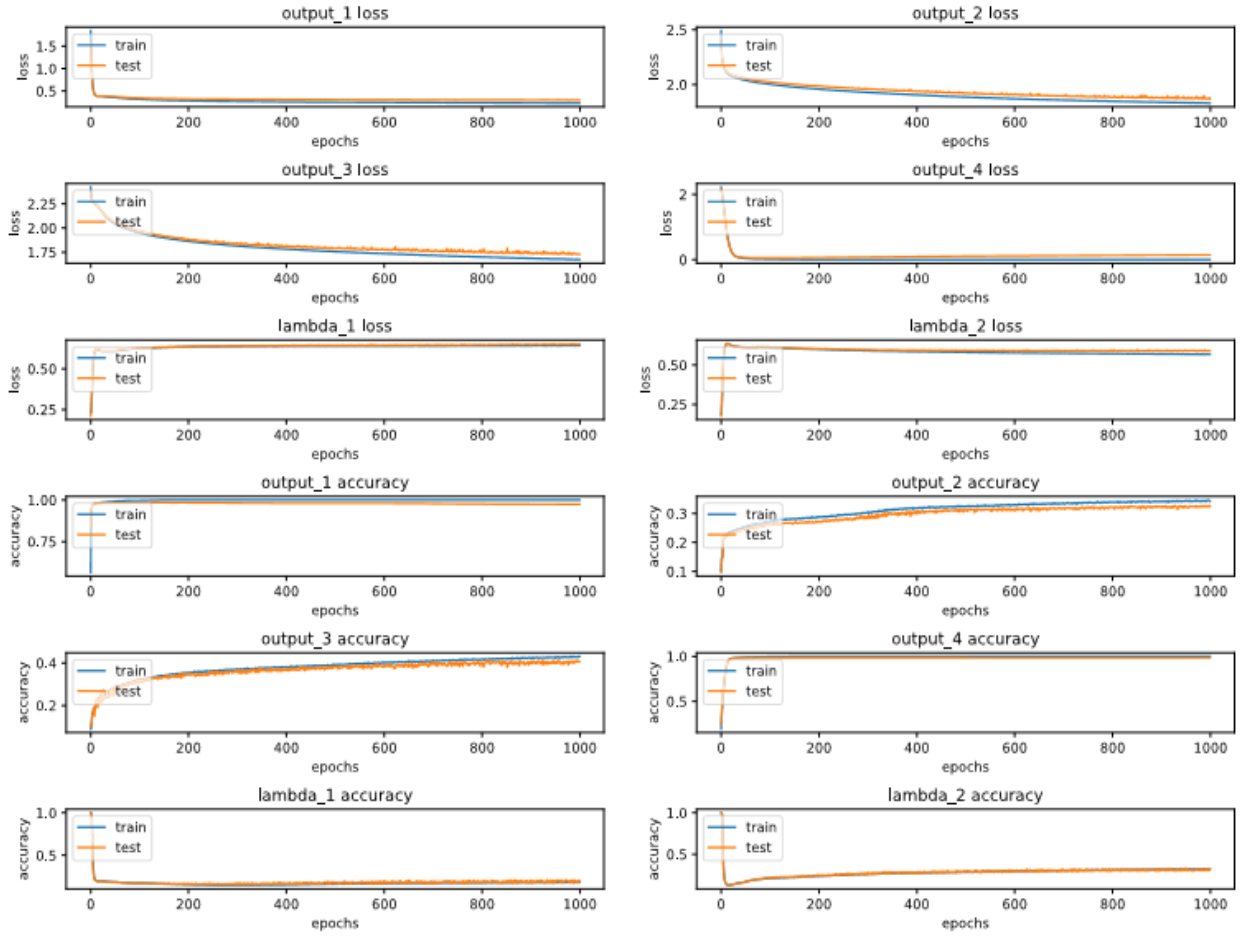
4.1.3 Testing

A dataset which is publicly accessible for scene recognition is used. Dataset from 10 categories are selected for scene recognition. Each with 100-2000 videos of 6 sec duration.

4.1.4 Results

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,"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". 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.

Figure 1: The figures of loss terms and accuracy.



The MESL method is compared with classic subspace learning methods CCA, MvDA-VC and NMCL method that takes into account both consistency and complementarity. Canonical Correlation Analysis(CCA) is a typical unsupervised approach to obtain a common space, which attempts to learn two transforms to project the samples from two views into a common subspace by maximizing the cross correlation between the two views. In this study, the baseline is CCA_3V. It is canonical correlation analysis of three views. 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 Multi-modal Cooperative Learning (NMCL) 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. Here accuracy is used as a metric for the performance of the methods. The comparison results with baseline methods are shown in table.

Table 4.1: Comparison with baselines

Method	Accuracy
CCA3V	0.6488
MvDA_VC	0.9744
NMCL	0.8047
MESL	0.9826

Table 4.1 shows the performance of proposed MESL method which is better than classic subspace 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 4.2: Enhancement studies

Modal	Accuracy before enhancement	Accuracy after enhancement
Audio	0.3286	0.3427
Text	0.4153	0.4210
Visual	0.9816	0.9697

Table 4.3: Accuracy comparison of mono modals and multi modals

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

The performance of each mono modal before and after the enhancement and the performance of mono modal and multi-modal are also compared. The comparison results are summarized in table 4.2 and table 4.3. As shown in table 4.2, 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 4.3, the performance of the multi-modal fusion is better than the mono modal.

4.2 Conclusions

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 introduced for the multi-modal fusion. In this study, the new 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.

4.3 Future Scope

In the future, this work can be extended to capture more complex correlations among multiple modalities such as conflict. In addition the extracted correlations can be applied to applications like image caption generation. Image caption models can be divided into two main categories. ie a method based on a statistical probability language model to generate handcraft features and a neural network model based on an encoder-decoder language model to extract deep features. The first method is based on maximum livelihood estimation. It uses visual detectors to analyse the scene. CNN can be used to detect the set of words associated in the image based on multi instance learning and weak monitoring technique. Using these words most likly sentence under condition can be generated. In the second method recurrent neural networks like LSTM can be used for natural language processing because it puts much attention in this field.

References

[1] J. Guo, X. Nie and Y. Yin, "Mutual complementarity: Multi-modal enhancement semantic learning for microvideo scene recognition". IEEE Access, pp . 29518–29524, August 2019.

[2] Y. Wei, X. Wang, W. Guan, and L. Nie, "Neural multimodal cooperative learning toward micro-video understanding," IEEE Transaction on Image Processing, vol. 29, pp. 1–14, Jul. 2020.

[3] M. Liu, L. Nie, and M. Wang, "Towards micro-video understanding by joint sequential-sparse modeling," in Proc. ACM International Conference on Multimedia, pp. 970–978, Oct. 2017.

[4] 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 Internationa Conference on Multimedia, pp. 1192–1200, Oct. 2017.

[5] J.Zhang, L.Nie and X.Wang, "Shorter-is-better: Venue category estimation from micro-video," in Proc. ACM Internationa Conference on Multimedia, pp. 1415–1424, Oct. 2016.

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

[7] M.Redi, N.Ohare, R.Schifanella, M.Trevisiol and A.Jaimes, "6seconds of sound and vision: Creativity in micro-videos," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4272–4279, Jun. 2014.

[8] P. Jing, Y. Su, and L. Nie, "Low-rank multi-view embedding learning for micro-video popularity prediction,"IEEE Transactions on Knowledge Data Engineering, vol.30, no. 8, pp. 1519–1532, Dec. 2017.

[9] Wei, Liu and Xianglin, Huang and Cao, Gang and Zang, "Joint learning of NNeXtVLAD, CNN and context gating for micro-video venue classification" IEEE Access, pp. 40950–40962, June 2016.

[10] 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

[11] 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 International Conference on Multimedia, pp. 898–907, Oct. 2016.