College of Engineering Chengannur
Department of Computer Engineering
M. Tech. Computer Science (Image Processing)
03CS6901 Seminar I
Abstract of Proposed Seminar Topic

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

CHN20M 7019 Sree Lekshmi B S

December 24, 2020

Keywords: Micro-video scene recognition, multi-modal fusion, semantic enhancement, adaptive weight learning.

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.

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], [4]. 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 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. 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.

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.

References

[1] J. Guo, X. Nie, and Y. Yin. Mutual complementarity: Multi-modal enhancement semantic learning for micro-video scene recognition. *IEEE Access*, 8:29518–29524, 2020.

[2] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. pages 898–907, October 2016.

[3] Matthias Braunhofer, Marius Kaminskas, and Francesco Ricci. Location-aware music recommendation. *International Journal of Multimedia Information Retrieval*, 2:31–44, March 2013.

[4] M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, and A. Jaimes. 6 seconds of sound and vision: Creativity in micro-videos. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4272–4279, 2014.