Journal of Electronic Research and Application, 2025, Volume 9, Issue 1
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on Emotion Classification Supported by Multimodal Adversarial Autoencoder

Jing Yu*

The National University of Malaysia, Kuala Lumpur 43600, Malaysia

Corresponding author: Jing Yu, yujxf18@163.com

Abstract: In this paper, the sentiment classification method of multimodal adversarial autoencoder is studied. This paper includes the introduction of the multimodal adversarial autoencoder emotion classification method and the experiment of the emotion classification method based on the encoder. The experimental analysis shows that the encoder has higher precision than other encoders in emotion classification. It is hoped that this analysis can provide some reference for the emotion classification under the current intelligent algorithm mode.

Keywords: Artificial intelligence; Multimode adversarial encoder; Sentiment classification; Evaluation criteria; Modal Settings

Online publication: February 13, 2025

1. Introduction of multi-mode antagonistic autoencoder emotion classification method

1.1. Overall framework

The multimodal antagonistic autoencoder sentiment classification method (MAAE) with a basic application goal of automatically implementing multimodal sentiment data classification. In the framework of the MAAE model, its basic components and functions are implemented as follows.

(1) Modal feature embedding module: The obtained multi-modal emotion data is input, and the embedded modal features are extracted and dimensionally processed by three parallel Long Short-Term Memory (LSTM) network modules.

(2) Multi-modal adversarial autoencoder module [1]: Load the corresponding multi-modal features respectively, and train the antagonism between the encoder and the antagonist to achieve the effective reduction of the heterogeneity of multi-modal features.

(3) Self-attention prediction network: The self-attention mechanism is used to fuse the important information in the multi-modal characteristics, and then the fully connected network is used to implement emotion classification.

1.2. Task setting

In the application of the MAAE method, researchers can set their task to predict the following emotion types with multi-modal emotion data: (1) Happiness; (2) Sadness; (3) Neutral; (4) Angry; (5) Excitement; (6) Depressed. In this process, the MAAE model can segment the multi-modal data to form M sentence sequences containing three kinds of data including audio mode, visual mode, language mode, and an emotional type label, and extract context-free features from them [2]. Its representation is as follows: (1) Audio feature sequence: $F_a \in R^{n \times d}$(2) Visual feature sequence: $F_v \in R^{n \times d}$(3) Linguistic feature sequence: $F_l \in R^{n \times d}$.represents the sequence length; *n* and *d* represents the feature dimension.

1.3. Multi-modal feature embedding

Obtain the multimodal feature sequence with the essence of contextual correlation in the video, and build its context relationship model to understand its relevance. In this case, the typical Bidirectional Long Short-Time Memory (Bi-LSTM) network can be introduced in sequential data processing. The network is composed of two independent LSTM networks, and the time direction of training can be set at the same time. The module is used to embed multi-modal data, and there are two levels of its composition: one is the Bi-LSTM layer whose unit is language, and the other is the Dense (perception layer) in the form of full connection [1–5]. The former can embed the context feature in the multimodal feature sequence, while the latter can receive the hidden layer state in the former and implement unified processing of the multimodal feature dimensions.

1.4. Multimodal representation learning

Although multimodal features have a unified dimension, the spatial distribution of their features is not consistent. The existing multi-modal sentiment analysis method only focuses on modal dimension alignment and implements sentiment analysis with data fusion methods, such as tensor fusion and attention fusion, etc., and its execution methods are complicated [6]. Although the above methods have made great progress, due to the heterogeneity among the modes, they will face many obstacles in the subsequent implementation of fusion interaction between different modes. To reduce the heterogeneity of different modes, the most intuitive idea is to map multi-modal data in a common space and learn its invariance. Based on this, in this study, the researchers take the joint representation algorithm as inspiration, and through the organic combination of adversarial network and autoencoder, build a multi-modal adversarial form of autoencoder, and reduce its heterogeneity by multi-modal representation learning [7].

Firstly, the adversarial network construction is based on generative adversarial network technology. Generative adversarial network technology is a common technique in deep learning, and it is also a typical artificial intelligence generation method. With the help of this technology, the high dimensional distribution of high complexity in multi-dimensional data such as audio, image, and language can be learned. There are usually two neural networks in generative adversarial networks, one is generator G and the other is generator D, which can effectively capture the real distribution of data under the condition of competition [8–10]. When constructing adversarial networks, it is necessary to build an encoder separately for each unimodal mode, input the multi-modal embedding feature, and make it map in a common subspace of shared form: $m \in \{a, v, l\}$ $G_m x_m$

$$h_m = G_m(x_m; \theta_{G_m}) \quad (1)$$

represents the invariant modal characteristics of the mapping in the common subspace; h_m θ_{G_m} represents the unimodal in the encoder. The components here are all fully connected networks, which are very simple in structure and can be used as encoder networks in autoencoders and generator networks in adversarial networks. G_m defines a modal discriminator, multimodal features can be identified by this discriminator. D for modal features of language form, it can be defined by "true;" For other forms of modal features, it can be defined as "false." When the unknown modal feature task is given, the discriminator can perform modal detection according to the feature as far as possible, and the encoder can generate similar multi-modal features as far as possible, so that they are not detected by the recognizer, to achieve the effective reduction of multi-modal heterogeneity.

Secondly, the autoencoder construction. Autoencoders belong to a specific type of neural network, which can minimize reconstruction loss, try to retain its important information in the original data, and remove redundancy and noise. The overall structure of autoencoder is mainly composed of encoder and decoder two parts. In a typical autoencoder, the main components of encoder and decoder are neural networks. The former is used to make the data form nonlinear mapping, and the latter is used to reconstruct the mapped data. In the reconstruction of invariant modal features in common subspace, the reasonable application of an autoencoder can significantly reduce the risk of information loss in the mapping process [11]. In specific mapping, for each unimodal mode, a decoder needs to be constructed respectively to carry out feature reconstruction, and its reconstruction loss function is expressed as: m D_m h_m

$$\hat{x}_m = D_m(h_m;\theta_{D_m}) \tag{2}$$

represents the reconstruction loss value; $\hat{x}_m$ θ_{D_m} represents the unimodal mode in the decoder.

1.5. Self-attention prediction network

Self-attention prediction network mainly predicts emotion scores by way of single-peak feature learning in multiple modes. In this network, the self-attention mechanism is the most critical component, and it is also a typical processing technology in deep learning technology. Its main application functions are automatic attention and automatic extraction of important features [12]. In order to extract more useful information in the multimodal features, researchers can reasonably apply the self-attention mechanism to the reconstructed features, and the algorithm formula is:

$$S_a = \frac{e(h_{mi}^{\ T} \cdot h_m)}{\sum_{i=1}^{N} e(h_m^{\ T} \cdot h_{mi})} \tag{3}$$

$$h_m^{\ a} = S_a \cdot h_m \tag{4}$$

$$h = concat(h_a^{\ a}; h_v^{\ a}; h_l^{\ a}) \tag{5}$$

represents the self-attention weight value; S_a e represents the modal characteristics of the external input; $h_m^{\ T}$ represents the transformation vector; h_m h_{mi} represents the I-th unimodal; $h_m^{\ a}$ represents the multimodal feature in the self-attention-directed mode; h represents the multi-modal fusion features obtained by splicing; $h_a^{\ a}$ represents the multimodal features of speech sequences guided by self-attention; $h_v^{\ a}$ represents the multimodal features of video sequences guided by self-attention; $h_l^{\ a}$ represents multimodal features of language sequences guided by self-attention.

2. Multimodal antagonistic autoencoder emotion classification method test

2.1. Test environment

In the above trial of emotion classification method, the selected test environment includes hardware and software, and its basic composition and configuration are as follows:

(1) Control processing unit (CPU): configuration information is Intel(R)Core(TM)i9-9900K and the frequency is 3.60 GHz.

(2) Graphics processing unit (GPU): configuration information is NVIDIA GeForce RTX 2080Ti.

(3) Random access memory (RAM): storage capacity of 64 GB.

(4) Operating system: configured as 64-bit Windows 10 system.

(5) Integrated development environment (IDE): configuration information is Pycharm + Anaconda3 + Python3.8.

(6) Deep learning framework: configuration information is Pytorch 1.8.1.

2.2. Data set

The experiment data is mainly obtained from a script dialogue video shoot, which has five key links, and the dialogue scenes in each link are different. The video was segmented according to several segments by the MAAE model, and each segment needed annotation to refine the emotion category, including six emotions: happiness, sadness, neutrality, anger, excitement, and frustration. The modal feature extraction of the whole video was mainly realized by different algorithms. The statistical features of the audio features, such as Meir frequency cepstral coefficient (MFCC), pitch, voice intensity, and audio, were extracted by the open-source software openSMILE, including mean value and root mean square [13,14]. The visual features in the visual modal data were extracted with 3D-CNN (software), and the relevant features were learned based on each frame and continuous frame changes. Each sentence in the language mode is split into word form, then embedded into the MAAE model in the form of words, embedded through the Word2Vec language pre-training model, and finally learned the abstract representation of hidden semantics through the convolutional neural network. **Table 1** shows the distribution of emotion category information in this experiment.

Table 1. Distribution of emotion category information in this experiment

Serial Number	Emotions	Pre-test sets	Training Set
1	Happy	114	504
2	Sad	245	839
3	Neutral	384	1,325
4	Anger	170	933
5	Excitement	299	724
6	Frustration	381	1,468

2.3. Evaluation indicators

This experiment is a task to classify multiple emotion categories, so the classification accuracy of the above six emotion categories and the average classification accuracy of all emotion categories will be used as the basic index to evaluate the Interactive Emotion Binary Data Capture Database (IEMOCAP) [15].

2.4. Model setup

To complete the MAAE model construction in this experiment according to the above method, the drop layer is set in each encoder and decoder of the model, to avoid overfitting. The LeakyReLU function is used as the encoder activation function, the Sigmoid function as the discriminator activation pair function, and the Softmax function as the self-attention prediction network activation function.

2.5. Test results

By comparing the above five traditional artificial intelligence emotion classification models with the MAAE model constructed in this study, it can be seen that the MAAE model has the highest classification accuracy for the three emotions of happiness, sadness, and excitement, and also has a higher classification accuracy for the three emotions of neutrality, anger, and frustration. The average accuracy of overall emotion classification is the highest. **Table 2** shows the results of the multi-modal adversarial autoencoder emotion classification experiment in this study.

Table 2. The results of the multimodal adversarial autoencoder emotion classification test in this study

Serial Number	Emotional classification	Test result					
		CatLATM	cLSTM	TFN	TFN	CMN	MAAE
1	Happy	35.0%	30.6%	29.9%	26.4%	25.0%	40.3%
2	Sad	56.1%	56.7%	55.5%	49.4%	55.9%	73.3%
3	Neutral	47.1%	57.6%	48.8%	56.8%	52.9%	53.5%
4	Anger	55.9%	59.4%	60.6%	61.2%	61.8%	56.3%
5	Excitement	54.7%	52.8%	57.9%	47.2%	55.5%	69.4%
6	Frustration	52.0%	65.9%	63.3%	63.3%	71.1%	55.9%
7	Averages	51.5%	56.3%	54.3%	53.2%	56.6%	58.8%

It can be seen that the multi-modal adversarial autoencoder emotion classification model studied in this research has more advantages in artificial intelligence-based emotion classification, and it can reasonably replace the traditional model to obtain more accurate classification results.

3. Concluding remarks

To sum up, in emotion classification processing based on artificial intelligence, the reasonable construction and application of algorithm models are very important. To deal with the drawbacks of traditional artificial intelligence algorithm models in multiple emotion classification processing, researchers can introduce the current more advanced multi-modal adversarial autoencoder, and complete the corresponding emotion classification model construction on this basis. In this way, the accuracy of emotion classification can be further improved, and a more idealized effect of artificial intelligence emotion classification can be obtained.

Disclosure statement

The author declares no conflict of interest.

References

[1] Beijing Language and Culture University, 2024, A Method for Identifying Emotion Types and Calculating Emotion Intensity, patent, CN202011426092.1.

[2] Hunan Mango Sunac Technology Co., LTD., 2024, Emotion Recognition and Video Content Matching System Based on Artificial Intelligence, patent, CN202410905484.8.

[3] Shantou University, 2024, A Cross-Domain Emotion Classification Method, Device, Equipment and Medium, patent, CN202410770711.0.

[4] Ye J, Xiang L, Zong C, 2024, Attribute-Level Emotion Classification Method Combining Attribute Modeling and Curriculum Learning. Journal of Software, 2024(9): 4377–4389.

[5] Kunming University of Science and Technology, 2024, A Small Sample Sentiment Analysis Method Based on Adaptive Multi-Modal Prompts, patent, CN202410660214.5.

[6] State Grid Anhui Electric Power Co., LTD., 2024, Information and Communication Branch. A Method of Input Text Meaning Understanding and Sentiment Analysis Based on Artificial Intelligence, patent, CN202410682831.5.

[7] Industrial and Commercial Bank of China Co., LTD., 2024, Sparse Emotion Classification Methods, Devices, Equipment, Media and Program Products, patent, CN202410784366.6.

[8] Wang Y, Zhu G, Duan W, et al., 2024, Emotional Classification Model of Psychological Counseling Texts Based on Interactive Attention Mechanism. Journal of Computer Applications, 2024(8): 2393–2399.

[9] Hunan Mango Sunac Technology Co., LTD., 2024, Artificial Intelligence-based Emotion Recognition and Video Content Matching System, patent, CN202410905484.8.

[10] Xiamen University of Technology, 2024, Music Emotion Recognition Method, Device, Equipment and Medium Based on Context Feature, patent, CN202410783596.0.

[11] Tencent Technology (Shenzhen) Co., LTD., 2024, Multi-label Emotion Classification Model Training Methods, Related Devices and media, patent, CN202410210755.8.

[12] Ping An Technology (Shenzhen) Co., LTD., 2024, Multimodal Emotion Classification Method, Device, Equipment and Storage Media, patent, CN202210834137.1.

[13] Zhang H, 2024, Application of Artificial Intelligence in Natural Language Processing. Materials for Information Recording, 2024(5): 139–141.

[14] Zhang J, 2024, A Cross-Modal Emotion Analysis Method, Training Method, Device and Equipment, patent, CN202410022411.4.

[15] Ping An Technology (Shenzhen) Co., LTD., 2024, Emotion Classification Method, Device, Equipment and Media based on Document-Level Emotion Tendency, patent, CN202111158076.3.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.