| **Username:** | **DD2200877** |
| --- | --- |
| Program Applied | 人工智能博士学位 |
| Major Applied | 人工智能博士学位 |

#### **Research Proposal**

#### **Study Topic:**

Multimodal fusion learning based on deep learning

#### **Abstract：**

In recent years, artificial intelligence has developed rapidly. Its related deep learning has made many achievements in the fields of image recognition, machine translation, emotion analysis, natural language and so on. People live in an environment where many fields blend with each other. Sound, vision and even smell are modal forms in various fields. In order to make the deep learning algorithm understand the surrounding world more comprehensively and efficiently, it is necessary to endow the machine with the ability to learn and integrate these multi domain modes. This capability is multimodal fusion. And deep learning is developing from the first mock exam of voice, text and image to multi-modal. That is, through the fusion analysis of multiple modes of hearing, vision, and even the data sets that are difficult to quantify such as smell, taste and psychology in the future, the single recognition network will be built into a high-level multi-modal artificial intelligence network. The development of multimodal fusion conforms to the development of the times, because most data in real life are multi-source and heterogeneous, and the research on multimodal fusion has become a trend. This paper will study the method of deep learning in the field of multimodal fusion, and use the deep learning model to fuse the signals of multiple modes.

#### **Key words:**

Multimodal signal, deep learning, Multimodal signal capture, data fitting, Multimodal fusion technology

#### **1.Introduction**

Multimodal fusion mainly includes representation, fusion, transformation, alignment and other technologies.[1] Generally speaking, modality refers to the objective existence of transactions, and multimodality refers to the combination of various forms of two or more modes. The information of each different structure can be called a mode. At present, the main research field is the processing of three modes: image, text and speech. The reason to fuse modes is that different modal information exists in different forms in nature. To better understand the world, we must combine multiple modes. The combined modes become multimodal. After multimodal formation, there will be some redundancy and complementary between different information. Because the eigenvectors of different modes are initially located in the subspaces of different modes, all have heterogeneity, which will affect the application of multimodal features in the field of deep learning.[2] Therefore, the purpose of multimodal fusion technology is to narrow the gap of features in each subspace and fit them at the same time. The fitting process needs to preserve the integrity of each subspace feature to the greatest extent. At present, multimodal fusion technology is widely used, mainly in multimodal translation[3], emotion analysis[4], multimodal biological information detection[5] and so on. Therefore, the purpose of this paper is to design a method, a multi-modal fusion method of graph neural network based on deep learning, which can minimize the distribution gap of semantic subspace and maintain the integrity of various modal features.

#### **2.Multimodal fusion**

Multimodal fusion architecture is divided into joint architecture, collaborative architecture and encoder architecture. Joint architecture is to project single-modal features into a shared sub semantic space, so that multi-modal features can be fused. Collaborative architecture includes transmembrane state similarity model and canonical correlation analysis and its goal is to find the correlation between modes in the coordination subspace. Encoder architecture is a multi-modal conversion task that maps one mode to another. This research mainly discusses the method of multimodal fusion[6].

Multimodal fusion is one of the original topics in multimodal machine learning, with previous surveys emphasizing early, late and hybrid fusion approaches. In technical terms, multimodal fusion is the concept of integrating information from multiple modalities with the goal of predicting an outcome measure: a class (e.g., happy vs. sad) through classification, or a continuous value (e.g., positivity of sentiment) through regression. It is one of the most researched aspects of multimodal machine learning with work dating to 25 years ago. The interest in multimodal fusion arises from three main benefits it can provide. First, having access to multiple modalities that observe the same phenomenon may allow for more robust predictions. This has been especially explored and exploited by the AVSR community. Second, having access to multiple modalities might allow us to capture complementary information — something that is not visible in individual modalities on their own. Third, a multimodal system can still operate when one of the modalities is missing, for example recognizing emotions from the visual signal when the person is not speaking.

While some prior work used the term multimodal fusion to describe all multimodal algorithms, we classify approaches as fusion when the multimodal integration is performed at the later prediction stages, with the goal of predicting outcome measures. Recently, the line between multimodal representation and fusion has been blurred for models such as deep neural networks where representation learning interacts with classification or regression objectives.

We classify multimodal fusion into two main categories: model-agnostic approaches (Section 2.1) that are not directly dependent on a specific machine learning method; and model-based approaches (Section 2.2) that explicitly address fusion in their construction — such as kernel-based approaches, graphical models, and neural networks.

#### **2.1 Model agnostic approaches**

Historically, the vast majority of multimodal fusion has been done using model-agnostic approaches. Such approaches can be split into early (i.e., feature-based), late (i.e., decision-based) and hybrid fusion. Early fusion integrates features immediately after they are extracted (often by simply concatenating their representations). Late fusion on the other hand performs integration after each of the modalities has made a decision (e.g., classification or regression). Finally, hybrid fusion combines outputs from early fusion and individual unimodal predictors. An advantage of model agnostic approaches is that they can be implemented using almost any classifiers or regressors.

Early fusion can learn to exploit the correlation and interactions between low level features of each modality and it also only requires the training of a single model, making the training pipeline easier compared to late and hybrid fusion. Tensor fusion network (TFN) belongs to early fusion, which is a typical multi-modal network that uses matrix operations to fuse features. In 2017, Amir Zadeh at el. proposed a model named tensor fusion network, which can learn two dynamics end-to-end. This method is customized for the fluctuation of spoken language in online video, as well as the accompanying gestures and sounds [7]. This method is widely used in the field of multimodal fusion. This model takes the characteristics of different modes as input, and calculates the correlation between different modal elements by calculating the tensor outer product between modes. The advantage of this model is that it can retain the characteristics of each modal element to the greatest extent, but it also has a small disadvantage that continuously calculating the correlation will increase the dimension of the feature vector, which makes the model too large and difficult to train. Then, in order to reduce the computational complexity caused by the exponential growth of dimension and the transformation of input into tensor. Zhun Liu at el.[8] proposed the Low-rank Multimodal Fusion method, which performs multimodal fusion using low-rank tensors to improve efficiency. They evaluated their model on three different tasks: multimodal sentiment analysis, speaker trait analysis and emotion recognition. Their model achieves competitive results on all these tasks. In addition, their method effectively reduces the computational complexity caused by the limit of high-dimensional features. Although the low rank tensor model is an upgrade of the tensor model, if it is not filtered when extracting the original features. Too long eigenvector can still lead to parameter explosion.

In contrast, late fusion uses unimodal decision values and fuses them using a fusion mechanism and allows for the use of different models. Furthermore, it makes it easier to make predictions when one or more of the modalities is missing and even allows for training when no parallel data is available. However, late fusion ignores the low level interaction between the modalities. Hybrid fusion attempts to exploit the advantages of both of the above described methods in a common framework. It has been used successfully for multimodal speaker identification and multimedia event detection (MED).

#### **2.2 Model-based approaches**

While model-agnostic approaches are easy to implement using machine learning methods, they end up using techniques that are not designed for multimodal data. In this section we describe three categories of approaches that are designed to perform multimodal fusion: kernel-based methods, graphical models, and neural networks.

Multiple kernel learning (MKL) methods are an extension to kernel support vector machines (SVM) that allow for the use of different kernels for different modalities/views of the data. As kernels can be seen as similarity functions between data points, modality-specific kernels in MKL allows for better fusion of heterogeneous data. MKL approaches have been an especially popular method for fusing visual descriptors for object detection and only recently have been overtaken by deep learning methods for the task. They have also seen use for multimodal affect recognition multimodal sentiment analysis, and multimedia event detection (MED). Besides flexibility in kernel selection, an advantage of MKL is the fact that the loss function is convex, allowing for model training using standard optimization packages and global optimum solutions. Furthermore, MKL can be used to both perform regression and classification. One of the main disadvantages of MKL is the reliance on training data (support vectors) during test time, leading to slow inference and a large memory footprint.

Graphical models are another family of popular methods for multimodal fusion. The benefit of graphical models is their ability to easily exploit spatial and temporal structure of the data, making them especially popular for temporal modeling tasks, such as AVSR(Audio-Visual Speech Recognition) and multimodal affect recognition. They also allow to build in human expert knowledge into the models and often lead to interpretable models. Some of the earliest approaches to use graphical models for multimodal fusion include generative models such as coupled and factorial hidden Markov models alongside dynamic Bayesian networks. A more recently proposed multi-stream HMM method proposes dynamic weighting of modalities for AVSR. While most graphical models are aimed at classification, conditional random fields (CRF) models have been extended to a continuous version for regression and applied in multimodal settings for audio visual emotion recognition.

Neural networks have been used extensively for the task of multimodal fusion. The earliest examples of using neural networks for multi-modal fusion come from work on AVSR(Audio-Visual Speech Recognition). Nowadays they are being used to fuse information for visual and media question answering, gesture recognition, affect analysis, and video description generation. Both shallow and deep neural models have been explored for multimodal fusion. Neural networks have also been used for fusing temporal multimodal information through the use of RNNs and LSTMs. Applying DNN to feature fusion is a great success after the wave of deep learning. It abandons the traditional way of matrix fusion. Jennifer proposed an intermediate-level feature fusion, which merges weights from each modality during training with subsequent additional training. In addition they tested principle component analysis for feature section and found that PCA can increase unimodal performance and multimodal fusion outperforms unimodal models [9]. It has showed that DNN can be used to implement multimodal fusion. And after that, Ganurav proposed adaptive fusion technology [10], because they think DNN fusion technology is too stiff and their technology can effectively model from different patterns. Besides the approach is not a focus of multimodal fusion for a specific network. In this article, he proposed two fusion technologies that are the most advanced. The first is atuo encoder, which encodes and decodes all the features, and then calculates the loss between the feature vectors. The second is Gan fusion, which standardizes the learning space and makes a quantitative evaluation of multimodal emotion task recognition. Memory Fusion Network (MFN) uses "delta memory attention" and "multi view gated memory" to capture the interaction between timing and modes at the same time, so as to obtain better multi view fusion. The purpose of using memory is to save the multimodal interaction information of the previous time, gate filtering, and attention allocation. Amir Zadeh proposed a new neural architecture for multi view sequential learning, called memory fusion network, which clearly describes the two interactions in the neural architecture and models them over time [11]. The first component of the MFN is called the LSTM system, where view specific interactions are isolated and learned by assigning LSTM functions to each view. Then, a special attention mechanism called incremental memory attention network (DMAn) is used to identify cross view interactions, and multi view gated memory is used to summarize over time. Through extensive experiments, MFN is compared with various proposed methods for multi view sequential learning on multiple publicly available benchmark data sets. MFN is superior to all multi view methods, outperforms all current state-of-the-art models and sets new state-of-the-art results for all three multi view datasets.

A big advantage of deep neural network approaches in data fusion is their capacity to learn from large amount of data. Secondly, recent neural architectures allow for end-to-end training of both the multimodal representation component and the fusion component. Finally, they show good performance when compared to non-neural network based system and can learn complex decision boundaries that other approaches struggle with.

The major disadvantage of neural network approaches is their lack of interpretability. It is difficult to tell what the prediction relies on, and which modalities or features play an important role. Furthermore, neural networks require large training datasets to be successful.

#### **3 Research objectives and Methodology**

##### **3.1Research objectives**

Multimodal fusion has a very wide range of applications, including audiovisual speech recognition (AVSR), multimodal emotion recognition, medical image analysis and multimedia event detection. There are many comments on the subject. Most of them focus on multi-modal fusion of specific tasks, such as multimedia analysis, information retrieval, or emotion recognition.

Cross-modal retrieval is an important branch of multi-modal learning. Its purpose is to explore the relationship between different modal samples, that is, to retrieve another modal sample with similar semantics through one modal sample.

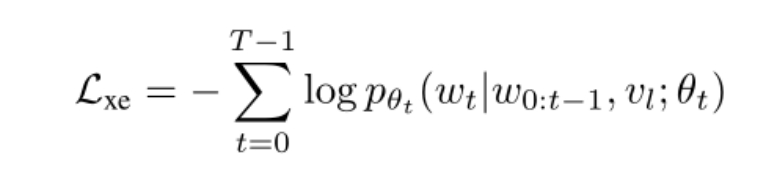
In this research，we propose to incorporate generative models into textual-visual feature embedding for cross-modal retrieval. In particular, in addition to the conventional cross-modal feature embedding at the global semantic level, we also introduce an additional cross-modal feature embedding at the local level, which is grounded by two generative models: image-to-text and text-to-image. There are three learning steps: look, imagine, and match. Given a query in image or text, we first look at the query to extract an abstract representation. Then, we imagine what the target item (text or image) in the other modality should look like, and get a more concrete grounded representation. We accomplish this by asking the representation of one modality (to be estimated) to generate the item in the other modality, and comparing the generated items with gold standards. After that, we match the right image-text pairs using the relevance score which is calculated based on a combination of grounded and abstract representations.

##### **3.2.1 Cross-modal Feature Embedding**

We follow the common cross-modal feature embedding approach to embed the representations of the image and the caption into a common space, and then use a pairwise ranking loss to learn the model parameters. In particular, given an image-caption pair, we encode a caption by embedding each word into a distributed representation using a shared word embedding matrix to be learned. Then we use two sequential sentence encoders to get the sentence representations. As for image encoding, we use a CNN that is pre-trained on ImageNet, and we formulate the embedding and mapping of each modality. Considering we have two branches of cross-modal feature embedding, which result in two pairs of cross-modal features: the abstract features and the grounded features. So, we consider the same pairwise ranking loss and optimize the ranking loss.

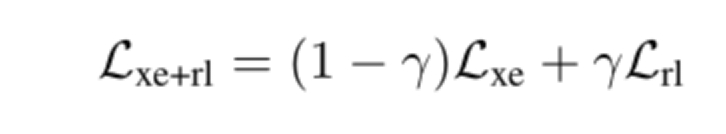
##### **3.2.2 Image-to-text Generative Feature Learning**

For the image-to-text training, our goal is to encourage the grounded visual feature to be able to generate sentences that are similar to the ground-truth captions. In particular, we first encode the image with CNN, and then decode the grounded visual feature into a sentence with RNN. Like the traditional RNN-based text generation models, we first train our model on a cross-entropy (XE) loss and reinforcement learning (RL) loss defined as:





where IMG_258is the ground-truth word, IMG_259is the grounded visual feature, IMG_260is the output probability of word IMG_261given by the decoder with parameter IMG_262; IMG_263is the word sequence sampled from the decoder, IMG_264is the reward calculated by comparing the generated sentence with the corresponding reference sentences using a standard evaluation metric. To ensure the readability and fluency of the generated caption, we use a mixture of XE and RL losses:

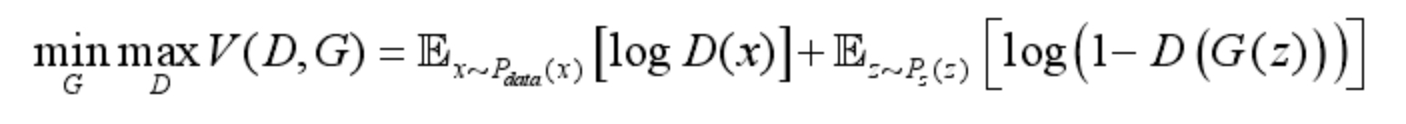


where γ is a tuning parameter used to balance the two losses.

##### **3.2.3 Text-to-image Generative Adversarial Feature Learning**

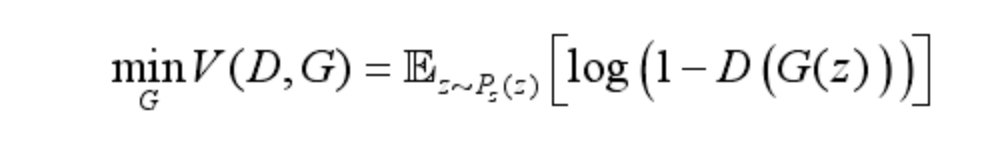
For the text-to-image training path, our goal is to encourage the grounded text feature to be able to generate an image that is similar to the ground-truth one. However, unlike the image-to-text path above, where the model is trained to predict the word conditioned on image and history words, the reverse path suffers from the highly multi-modal distribution of images conditioned on a text representation. The natural way to model such a conditional distribution is to use a conditional GAN, which consists of a discriminator and a generator. The discriminator is trained to distinguish the real samples “real image, true caption” from the generated samples of “fake image, true caption” as well as samples of “real image, wrong caption”.

Specifically, the discriminator D and the generator G play the min-max game on the following value function IMG_266：

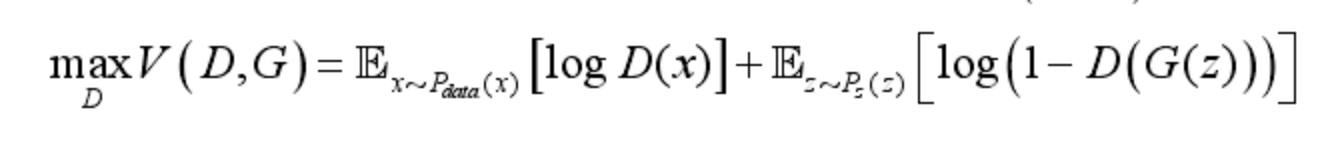


Where，x is the real data distribution data, z are random vectors.

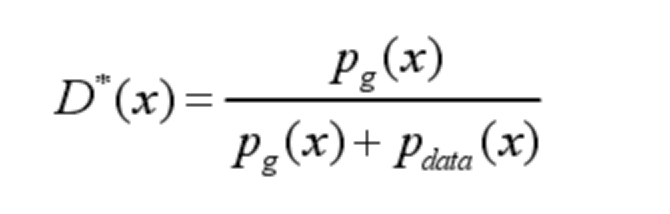
When fixedly discriminating the network D, the loss function of optimizing G is:



When the generator G is fixed, the optimization goal of the discriminator D is to maximize V(D,G):



Therefore, taking the derivative of V(D, G), the optimal discriminator D\*(x) is:



Finally, bring the optimal discriminator D(x)\* into the objective function, and get the optimal value.

A model is built based on the GAN network architecture proposed above, and we will combine with classic algorithms for image retrieval for performance evaluation. Meanwhile, we will use the recall rate and average precision of the proposed algorithm and other algorithms in the NUS-WIDE data set for comprehensive comparison and performance evaluation.

#### **Reference**

[1] BALTRUSAITIS, AHUJA C, MORENCY L P. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[2] PENG Yuxin, QI Jinwei. CM-GANs: cross-modal generative adversarial networks for common representation learning. Multimedia, 2019.

[3] Chaitanya Ahuja, Louis-Philippe Morency. Language2Pose: Natural Language Grounded Pose Forecasting. International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, 2019.

[4] Joaquim Comas, Decky Aspandi, Xavier Binefa. End-to-end facial and physiological model for Affective Computing and applications. 15th IEEE international Conference on Automatic Face and Gesture Recognition, 2020.

[5] Hyung-Tak Choi, Moon-Ki Back, Kyu-Chul Lee. Driver Drowsiness Detection based on Multimodal using Fusion of Visual-feature and Bio-signal. International Conference on Information and Communication Technology Convergence, 2018.

[6] He Jun, Zhang Caiqing, Li Xiaozhen, et al. Survey of research on multimodal fusion technology for deep learning. Computer engineering, 2020,46(5):1-11.

[7] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, Louis-Philippe Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP 2017.

[8] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasminhan, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL, 2018.

[9] Jennifer Williams, Ramona Comanescu, Oana Radu, Leimin Tian. DNN Multimodal Fusion Techniques for Predicting Video Sentiment. ACL, 2018.

[10] Ganurav Sahu, Olga Vechtomova. Adaptive Fusion Techniques for Multimodal Data. Ui.adsabs.harvard.edu, 2019.

[11] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, Louis-Philippe Morency. Memory Fusion Network for Multi-view Sequential Learning. Thirty-Second AIII Conference on Artificial Intelligence, 2018.

[12] Ian J. Goodfellow, Jean Pouget-abadie, Mehdi Mirza Bing xu, Daviad Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengjo. Generative Adversarial Networks. ACM, 2020.

[13] Radford A, Metz L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv:1511.06434. 2015

[14] Denton E, Chintala s, Szlam A et al. Deep generative image models using a laplacian pyramid of adversarial networks. The 28th International Conference on Neural Information Processing System, Montreal, Canada,2015.

[15] Ledig C, Theis L, Huszar F et al. Photo-realistic single image super-resolution using a generative adversarial network. 2017 IEEE Conference on Computer Vision and Pattern Recognition,Honolulu, USA, 2017.