Deep Supervised with Fine-grained Feature Fusion Network for Cross-modal Retrieval

Jiwei Zhang
Department of Systems Engineering
Wakayama University
Wakayama, Japan
s210068@wakayama-u.ac.jp

Hirotaka Hachiya
Department of Systems Engineering
Wakayama University
Wakayama, Japan
hhachiya@wakayama-u.ac.jp

***Abstract*—Audio-visual cross-modal retrieval seeks to establish relations and similarities between different media types, facilitating retrieval and recommendations across modalities. The core challenge of cross-modal retrieval is to analyze and understand the data of different modalities, extract the joint representation and overcome the cross-modal heterogeneity gaps. In this paper, we propose a novel deep supervised fine-grained feature fusion network for cross-modal retrieval, which aims to utilize the multi-modal fusion attention mechanism combined with fine-grained features to dynamically adjust the weights between multi-modal features to learn extensive and comprehensive cross-modal representations. Additionally, the shared weight strategy and the constrained loss function are used to jointly guide the model to learn modality-invariant features. Our proposed model obtains excellent experimental results on the widely used VEGAS and AVE benchmark datasets.**

***Index Terms*—Cross-modal retrieval, Fine-grained features, Feature fusion, Correlation learning.**

I. Introduction

CROSS-MODAL retrieval is a crucial topic in multimodal learning, aimed at achieving efficient information retrieval across different modalities (e.g., images, text, audio, and video). Unlike unimodal retrieval, cross-modal retrieval requires modeling the semantic relationship between heterogeneous modalities so that the input modality data (audio) can be used to retrieve related data of another modality (such as matching images) [1]. This technology has shown wide application potential in the fields of intelligent security, e-commerce, medical image analysis, etc. However, cross-modal retrieval faces a series of challenges, and the heterogeneity gap is one of the key difficulties.

In recent years, researchers have proposed various methods to address these challenges, among which fine-grained feature modeling has emerged as a significant research direction[2]. Compared with traditional methods that represent modal data through global features, fine-grained methods pay more attention to local features within the modality and fine-grained semantic alignment between modalities. For example, fine-grained features can represent a specific area or object in the image modality. In this way, fine-grained methods can better capture the local-local relationship between modalities in cross-modal retrieval, thereby improving retrieval performance.

To solve these problems, researchers have proposed some methods to align local regions of visual and audio features through multi-modal attention mechanisms (cross-attention)[3]. Although fine-grained methods have made significant progress in improving cross-modal retrieval performance, there are still challenges in balancing the differences in modal characteristics and reducing computational overhead. Therefore, this paper focuses on the application of fine-grained feature modeling in cross-modal retrieval, aiming to explore a more efficient fine-grained feature extraction and alignment method to improve cross-modal retrieval performance further.

In this paper, we propose a novel deep supervised fine-grained feature fusion network, which learns the features of each modality in a unified framework, dynamically adjusts the weights between them, and utilizes the multi-modal fusion attention mechanism to enhance semantic alignment. Through a multi-step fusion strategy, it promotes the interaction of cross-modal features and explores potential correlation representations. In the stage of fusing visual-audio input features, we utilize the adaptive weighting strategy to dynamically adjust the weight of each feature and optimize the synergy between features. Additionally, a shared weight strategy is used together with a constrained loss function to guide the training of the model to learn modality-invariant and discriminable features.

This paper focuses on the application of fine-grained features in cross-modal retrieval, aiming to design an efficient feature fusion method to further improve the performance and applicability of the model. The main contributions are as follows:

1) We propose a novel fine-grained feature fusion method that combines the fine-grained features of images and audio to capture more detailed semantic relationships between modalities.
2) We train the model using a constrained loss function with a shared weight strategy so that the modality preserves discriminative and modality-invariant features between samples from different semantic categories.
3) Experimental verification on multiple cross-modal retrieval datasets proves that our method is superior to existing mainstream methods in retrieval accuracy.

II. Related Works

In this section, we will provide a brief introduction to related methods for cross-modal retrieval, as well as some attention-based techniques.

A. Unsupervised cross-modal approaches

Maximizing the correlation between different modalities without supervised information is essential for learning joint representations across diverse data types. Canonical Correlation Analysis (CCA) [4] focuses on learning linear transformations of the modalities to project them into a shared space, maximizing their mutual relevance. Deep Canonical Correlation Analysis (DCCA) [5] extends this approach by utilizing deep neural networks to capture complex, high-dimensional correlations through nonlinear transformations. Unsupervised Contrastive Hashing with Modality Correlation (UCHM) [6] introduces a distinct strategy. It minimizes hash similarity through a custom loss function to train a similarity generator. The resulting Modality Interaction Entropy (MIE) similarity matrix acts as a guiding model for training a deep hash network, enabling the discovery of robust joint representations.

B. Supervised cross-modal approaches

Supervised semantic information is used to reduce the intra-modal heterogeneity gap by distinguishing samples from different semantic categories. TNN-C-CCA [7] enhances the Cluster-CCA method by introducing a task-specific loss function based on multi-modal learning. This approach not only accounts for the correlations of paired samples but also incorporates the correlations of non-paired samples, leading to more robust representation learning.

Attention mechanisms in neural networks selectively emphasize important parts of the input data while suppressing less relevant ones. In cross-modal tasks, attention modules capture the correlations between modalities, aligning audio and visual sequences by learning the interdependencies among their elements. Deep Co-attention Network [8] proposed a deep co-attention architecture for multi-view subspace representation learning, enhancing interpretability and prediction reliability.

III. Proposed Method

In this section, we start by defining the problem of cross-modal retrieval. Next, we introduce a novel model aimed at learning feature representations from audio-visual modality data. Finally, we provide detailed implementation specifics of the proposed approach.

A. Problem Formulation

Multimedia data $\mathcal{H} \equiv \left\{(\boldsymbol{h}_i^{\mathrm{a}}, \boldsymbol{h}_i^{\mathrm{v}})\right\}_{i=1}^{n}$ be n pairs of audio-visual samples where $\boldsymbol{h}_i^{\mathrm{a}} \in \mathbb{R}^{128}$ and $\boldsymbol{h}_i^{\mathrm{v}} \in \mathbb{R}^{1024}$ are the i-th audio and visual vectors extracted by pre-trained VGGish and Inception networks, respectively. We assign a corresponding one-hot category vector $\boldsymbol{y}_i \in \{0,1\}^m$ for each pair of multimedia samples, where m represent the number of categories. Multimedia data differ in dimensions and distributions, making direct comparison challenging. Using transformation functions, we address cross-modal retrieval by mapping features into a shared representation space.

B. Fine-grained Feature Extraction

In this work, we aim to balance computational efficiency and resource usage by combining small kernels ($k_{3\times1}$, $k_{5\times1}$, with $k_{7\times1}$) larger kernels. An audio-visual feature pair H^a and H^v is processed through an encoder with a shared linear layer, and individual subnetworks using 1D convolutions with various kernel sizes and max pooling. This generates fine-grained features F^a and F^v as follows:

$$F^a = \mathrm{Encoder}(H^a, \boldsymbol{\theta}_{\mathrm{e}}^{\mathrm{a}}) \quad F^v = \mathrm{Encoder}(H^v, \boldsymbol{\theta}_{\mathrm{e}}^{\mathrm{v}}) \tag{1}$$

$$F_{k\times1}^a = \mathrm{Conv1d}\left(F^a, k; \boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{a}}\right), \quad F_{k\times1}^v = \mathrm{Conv1d}\left(F^v, k; \boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{v}}\right), \tag{2}$$

where $\mathrm{Conv1d}(\boldsymbol{f}, k, \boldsymbol{\theta})$ represents a 1D convolution operation applied to a feature $k \in \{3\times1, 5\times1, 7\times1\}$ with parameter $\boldsymbol{\theta}$.

C. Fine-grained Feature Fusion

While multiple modalities offer more information, fusing them can reduce modality-specific details. To address this, we design a multi-modal fusion attention (MFA) with adaptive weighting to integrate information into a stable representation. The architecture is detailed in Figure 2. For given feature maps F_k^v and F_k^a, we utilize a modality function unit that performs element-wise operations for integrate intra-modal Φ_v (Φ_a) and inter-modal Φ_{va} (Φ_{av}) attention map. We create both audio and visual representations in the following ways:

$$\begin{aligned} G^v &= \sigma(\Phi_{va} W_1^{va} + \Phi_v) \odot \Phi_v, \\ G^a &= \sigma(\Phi_{av} W_2^{av} + \Phi_a) \odot \Phi_a, \end{aligned} \tag{3}$$

where W are parameter matrices, and σ is the Sigmoid function. The joint representation $J = \sigma(\mathrm{Bilinear}\left[G^v, G^a\right]) \in \mathbb{R}^{128\times3\times128}$ allows the network to perform soft selection between F_k^v and F_k^a. The adaptive weighting process for MFA denoted as M_k is defined as follows:

$$M_k = J \odot F_k^v + (1 - J) \odot F_k^a, \tag{4}$$

The quality of initial integration can influence the final fusion weights in the attention module. Since this involves feature fusion, a common approach is to use another attention module to combine the input features. We employ a two-stage fusion strategy to calculate the final feature representations $\widehat{F_k^v}$, $\widehat{F_k^a}$ by the following formula (as shown in Figure 1):

$$\begin{aligned} &M_k^1 = \mathrm{MFA}(F_k^v, F_k^a),\ I_k^v = M_k^1 \odot F_k^v,\ I_k^a = M_k^1 \odot F_k^a, \\ &M_k^2 = \mathrm{MFA}(I_k^v, I_k^a),\ \widehat{F_k^v} = M_k^2 \odot F_k^v, \widehat{F_k^a} = M_k^2 \odot F_k^a, \end{aligned} \tag{5}$$

Finally, the model connects two linear layers to convert the 768-dimensional $\widehat{F_k^v}$ and $\widehat{F_k^a}$ features into 64-dimensional O^v and O^a features as the output of the model.

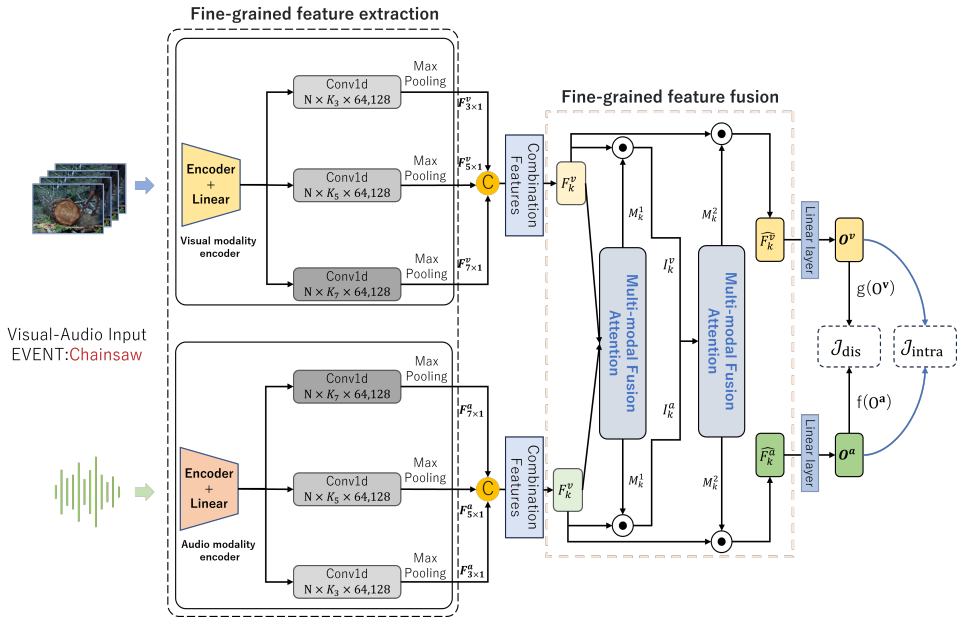


Fig. 1: Illustration of the general framework of the proposed method.

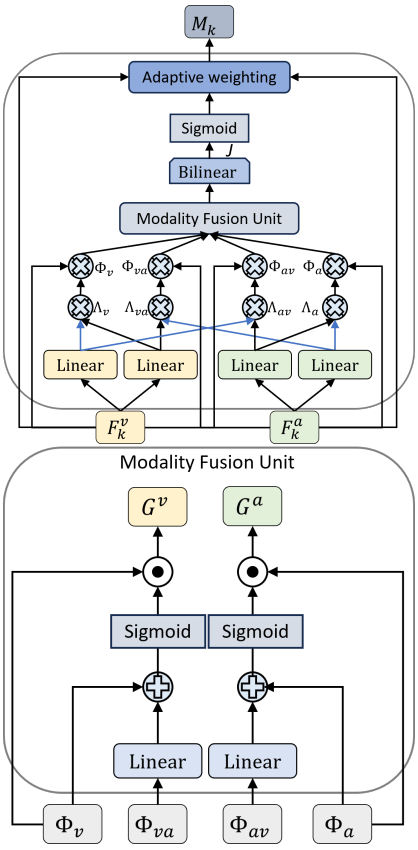


Fig. 2: Illustration of the general framework of the proposed multi-modal fusion attention.

D. Objective Function

We introduce two constraints for cross-modal feature representation learning: discriminative, and intra-modal loss.

We connect a simple linear layer at the end of the multi-modal dual subnetwork. This classifier is incorporated into the visual and audio modality subnetworks to ensure the distinction of instances from different categories after feature projection. It uses the training data to predict a m-dimensional category vector for each instance. We utilize the following loss function to preserve the discriminability of modal representations in the semantic space:

$$\mathcal{J}_{\text{dis}} = \frac{1}{n}\|f(\boldsymbol{o}_i^{\text{a}}) - \boldsymbol{y}_i\|_F + \frac{1}{n}\|g(\boldsymbol{o}_i^{\text{v}}) - \boldsymbol{y}_i\|_F, \tag{6}$$

where $\boldsymbol{y}_i$ denote the true category, and $f(\cdot)$, $g(\cdot)$ represent transformation function of the predicted category.

To improve the discriminative performance of deep neural networks [9], we position instance features close to their category centers, addressing intra-class compactness and inter-class separability. By projecting the features into a shared subspace, the intra-modal constraint loss directly calculates feature distances, guiding the model to learn compact, separable representations and enhance discriminative ability. The loss is defined as follows:

$$\mathcal{J}_{\text{intra}} = \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{o}_i^{\text{a}} - \boldsymbol{c}_{y_i}^{\text{a}}\right\|_2^2 + \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{o}_i^{\text{v}} - \boldsymbol{c}_{y_i}^{\text{v}}\right\|_2^2, \tag{7}$$

where $\boldsymbol{c}_i^{\text{a}}$ and $\boldsymbol{c}_i^{\text{v}}$ represents the center of features $\boldsymbol{o}_i^{\text{a}}$ and $\boldsymbol{o}_i^{\text{v}}$ belonging to i-th class.

IV. Experiments

We conduct a comparative analysis of our proposed method against various advanced approaches and discuss the results.

We also conduct ablation experiments to assess the impact of various components of the model.

A. Comparison with Existing Methods

We compare our method with existing advanced approaches on the VEGAS and AVE datasets. All methods utilize the same audio-visual data features, and the final representation dimensions in our model align with those of the advanced methods.

TABLE I: the mAP scores for audio-visual retrieval on the VEGAS test dataset.

Approach	Audio2Visual	Visual2Audio	Average
CCA [4]	0.332	0.327	0.330
KCCA [10]	0.288	0.273	0.281
DCCA [5]	0.478	0.457	0.468
C-CCA [11]	0.711	0.707	0.709
C-DCCA [12; 13]	0.722	0.716	0.719
ACMR [14]	0.465	0.442	0.454
TNN-C-CCA [7]	0.751	0.738	0.745
VAE-CCA [15]	0.821	0.824	0.822
MSNSCA [2]	0.866	0.865	0.866
Proposed	**0.881**	**0.895**	**0.888**

TABLE II: the mAP scores for audio-visual retrieval on the AVE test dataset.

Approach	Audio2Visual	Visual2Audio	Average
CCA [4]	0.190	0.189	0.190
KCCA [10]	0.133	0.135	0.134
DCCA [5]	0.221	0.223	0.222
C-CCA [11]	0.153	0.152	0.153
C-DCCA [12; 13]	0.230	0.227	0.229
ACMR [14]	0.162	0.159	0.161
TNN-C-CCA [7]	0.253	0.258	0.256
VAE-CCA [15]	0.328	0.302	0.315
MSNSCA [2]	0.323	0.343	0.333
Proposed	**0.371**	**0.368**	**0.370**

Tables I and II present a comparison between our proposed method and existing approaches on the VEGAS and AVE datasets. The experimental results demonstrate that our method is more effective than the current state-of-the-art techniques. Our proposal method fuses the fine-grained features of multi-modality, capturing more semantic information with the information of another modality and bridging the heterogeneity gap. Unsupervised methods like CCA, DCCA, and KCCA struggle to achieve high retrieval scores because they do not leverage semantic information, hindering the model's ability to learn discriminative features essential for cross-modal retrieval. While models based on advanced techniques (e.g., DNN, attention) generally show satisfactory performance, methods like MSNSCA, VAE-CCA, and TNN-CCA fail to fully exploit the rich correlations in cross-modal data, lacking sufficient interaction between modalities and struggling to bridge the heterogeneity gap, which significantly limits retrieval performance.

B. Impact of Different Components

The proposed method's objective loss function consists of $\mathcal{J}_{\mathrm{dis}}$, $\mathcal{J}_{\mathrm{intra}}$, and the MFA module. We performed ablation experiments on the proposed method to evaluate the influence of various components on the model's retrieval performance. From Tables III and IV, we observe that the model trained with only $\mathcal{J}_{\mathrm{dis}}$ only achieves retrieval accuracy of 0.282 and 0.114 on the VEGAS and AVE datasets. When the MFA module is added, the performance of the model can be significantly improved, since the MFA module utilizes the correlation between modes and bridges the heterogeneity gap. Experimental results demonstrate that combining various components of the proposed model significantly improves retrieval accuracy.

TABLE III: Comparison of the performances of the combination of three components on the VEGAS test dataset.

No.	$\mathcal{J}_{\mathrm{dis}}$	$\mathcal{J}_{\mathrm{intra}}$	MFA	Audio2Visual	Visual2Audio	Average
0	○	○	✓	0.431	0.443	0.437
1	○	✓	○	0.276	0.287	0.282
2	✓	○	○	0.361	0.354	0.358
3	○	✓	✓	0.704	0.718	0.711
4	✓	○	✓	0.783	0.774	0.779
5	✓	✓	○	0.612	0.608	0.610
6	✓	✓	✓	0.881	0.895	0.888

TABLE IV: Comparison of the performances of the combination of three components on the AVE test dataset.

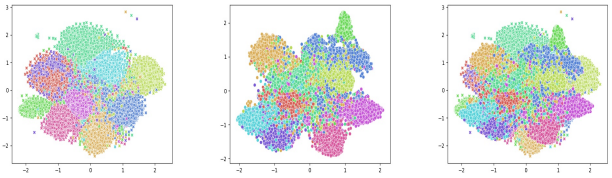
No.	$\mathcal{J}_{\mathrm{dis}}$	$\mathcal{J}_{\mathrm{intra}}$	MFA	Audio2Visual	Visual2Audio	Average
0	○	○	✓	0.202	0.213	0.208
1	○	✓	○	0.112	0.116	0.114
2	✓	○	○	0.118	0.114	0.116
3	○	✓	✓	0.267	0.271	0.269
4	✓	○	✓	0.274	0.278	0.276
5	✓	✓	○	0.221	0.234	0.228
6	✓	✓	✓	0.371	0.368	0.370

C. Visualisation of the Learned Representation

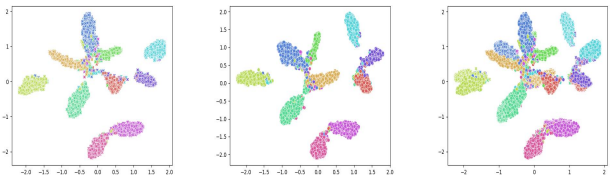
We utilize the t-SNE method to transform visual and audio representations into a two-dimensional visualization plane for analyzing the distribution of our proposed method. Figure 3(a)-(c) visualizes the distribution of the original data and the audio-visual modality data in the same visualization plane. We can observe that the boundaries of different categories of the original data samples overlap, it is challenging to clearly differentiate between the samples of the various categories. and there is a large gap between the audio-visual modality samples belonging to the same category. Figure 3(d)-(f) visualizes the data trained by the proposal model and the distribution of the audio-visual modality data in the same visualization plane. The distance between samples of different categories is far, and they can be effectively distinguished. However, there are still some overlapping samples, which limit the retrieval performance of the model.

V. Conclusion

In this paper, we introduce a novel approach for learning deep supervised fine-grained feature fusion from audiovisual modal data. This method progressively establishes correlations



(a) Original visual data (b) Original audio data (c) Original visual-audio data



(d) Visual data(our) (e) Audio data(our) (f) Visual-audio data(our)

Fig. 3: Visual analysis of the distribution of test data in the VEGAS data set. Different marks represent the distribution of samples in visual and audio modalities, and different colors represent samples of different categories.

between different modalities through a multi-modal fusion attention module. Additionally, we propose a constrained loss function that directs the model to learn representations that are both discriminative and invariant across modalities. Comprehensive experimental results and evaluations on two widely used benchmark datasets highlight the effectiveness of the proposed model architecture. However, the visual-audio retrieval task still has some limitations, which requires balancing the differences in modality features and the computational overhead associated with reducing fine-grained feature extraction.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] J. Zhang, Y. Yu, S. Tang, W. Li, and J. Wu, "Multi-scale network with shared cross-attention for audio–visual correlation learning," *Neural Computing and Applications*, vol. 35, no. 27, pp. 20 173–20 187, 2023.

[3] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," pp. 10 394–10 403, 2019.

[4] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*, 1992, pp. 162–190.

[5] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," pp. 1247–1255, 2013.

[6] R.-C. Tu, J. Jiang, Q. Lin, C. Cai, S. Tian, H. Wang, and W. Liu, "Unsupervised cross-modal hashing with modality-interaction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[7] D. Zeng, Y. Yu, and K. Oyama, "Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 3, pp. 1–23, 2020.

[8] L. Zheng, Y. Cheng, H. Yang, N. Cao, and J. He, "Deep co-attention network for multi-view subspace learning," in *Proceedings of the Web Conference 2021*, 2021, pp. 1528–1539.

[9] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," pp. 499–515, 2016.

[10] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 05, pp. 365–377, 2000.

[11] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," pp. 823–831, 2014.

[12] Y. Yu, S. Tang, K. Aizawa, and A. Aizawa, "Category-based deep cca for fine-grained venue discovery from multimodal data," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 4, pp. 1250–1258, 2018.

[13] D. Zeng, Y. Yu, and K. Oyama, "Audio-visual embedding for cross-modal music video retrieval through supervised deep cca," pp. 143–150, 2018.

[14] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," pp. 154–162, 2017.

[15] J. Zhang, Y. Yu, S. Tang, J. Wu, and W. Li, "Variational autoencoder with cca for audio–visual cross-modal retrieval," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3s, pp. 1–21, 2023.