Multimodal Coordinated Representation Learning Based on Evidence Theory

Wei Li
School of Automation Science and Engineering
Xi'an Jiaotong University
Xi'an, China
liwei19970705@163.com

Deqiang Han
School of Automation Science and Engineering
Xi'an Jiaotong University
Xi'an, China
deqhan@xjtu.edu.cn

Jean Dezert
ONERA
The French Aerospace Lab
Chemin de la Hunière
F-91761 Palaiseau, France
jean.dezert@onera.fr

Yi Yang
School of Aerospace
Xi'an Jiaotong University
Xi'an, China
jiafeiyy@xjtu.edu.cn

Abstract—**In multimodal learning, multimodal coordinated representation is an important yet challenging issue, which establishes the interaction between different modalities to describe multimodal data more effectively. Existing coordinated representation methods are implemented in the deep feature space (or encoding space) of each modality. In this paper, based on the framework of evidence theory, we propose a novel coordinated representation method, where multimodal data is described as the basic belief assignment (BBA), and coordinated learning is implemented in the evidential space (i.e., the BBA-based space). That is, the information interaction between different modalities is implemented at the level of evidence modeling (or uncertainty modeling). To use the intra-class and inter-class difference information of multimodal data, we design an evidential coordinated constraint. Furthermore, to represent each modality clearly, we introduce an ambiguity constraint. Experimental results of multimodal classification show that our proposed method is rational and effective.**

Index Terms—**multimodal coordinated representation, evidence theory, uncertainty measure, multimodal classification**

I. Introduction

In recent years, there has been a surge of interest in the field of multimodal learning [1-3]. Multimodal learning can effectively process and integrate information from different modalities (such as text, image, and audio) to comprehendsively represent multimodal data. Multimodal learning has been widely applied in several fields, such as medical diagnosis [4-5], fault diagnosis [6-7], and autonomous driving [8-9].

In multimodal learning, coordinated representation plays an important role [1], which establishes the interaction between different modalities during the learning process to describe multimodal data more effectively. Researchers have proposed various multimodal coordinated representation methods from two different perspectives. The first type is the similarity-based method, which calculates the similarity between different modalities, such as the Euclidean distance similarity [10], the cosine distance similarity [11-12], and the dot-product similarity [13]. These similarities are calculated in each modality's deep feature space and minimized for the same class samples. The second type is the structure-based method. In this method, the specific constraint is designed to establish structural relationships between the representations of different modalities. For example, in the cross-modal hashing methods [14-15], high-dimensional data from different modalities are compressed into a common binary space by the hash function. In this binary encoding space (also called the hash space), the samples of the same class possess similar binary structures.

Existing coordinated representation methods are implemented in the deep feature space [10-13] or encoding space [14-15] of each modality. Evidence theory [16-17] is an effective mathematical tool for uncertainty modeling and reasoning, where the basic belief assignment (BBA) can represent the uncertainty information of data to support decision-making. In this paper, based on the framework of evidence theory, we propose a novel coordinated representation method, where multimodal data is described as the basic belief assignment (BBA), and coordinated learning is implemented in the evidential space (i.e., the BBA-based space). That is, the information interaction between different modalities is implemented at the level of evidence modeling (or uncertainty modeling). To use the intra-class and inter-class difference information of multimodal data, we design an evidential coordinated constraint by calculating the distance of evidence between different modalities' BBAs. Furthermore, to represent each modality clearly, we introduce an ambiguity constraint, defined as the uncertainty measure of each modality's BBA. Experimental results of multimodal classification show the effectiveness and rationality of our proposed method.

II. Preliminary

A. Basics of Evidence Theory

In DST, the frame of discernment (FOD) is defined as a set consisting of n mutually exclusive and exhaustive elements, denoted by $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$. Let 2^Θ be the power set of the FOD. If a set function $m: 2^\Theta \to [0,1]$ satisfies

$$\sum_{A \subseteq \Theta} m(A) = 1, \quad m(\varnothing) = 0 \tag{1}$$

then m is called a basic belief assignment (BBA, also called a mass function). Given that $m(A) > 0$, A is called a focal element.

Given a BBA on the FOD Θ, the belief function Bel and plausibility function Pl are defined as

$$Bel(A) = \sum_{B \subseteq A} m(B), \quad \forall A \subseteq \Theta \tag{2}$$

$$Pl(A) = \sum_{B \cap A \neq \varnothing} m(B), \quad \forall A \subseteq \Theta \tag{3}$$

The $Bel(A)$ and $Pl(A)$ constitute the belief interval $[Bel(A), Pl(A)]$, which represents the degree of imprecision for the proposition A.

This work is supported by National Natural Science Foundation of China, No. U22A2045.

Suppose m_1 and m_2 are two independent BBAs on the same FOD, which can be combined via the Dempster's rule of combination [1] as follows

$$m(A)=\begin{cases}0, & A=\varnothing\\ \dfrac{\sum_{B\cap C=A} m_1(B)m_2(C)}{1-K}, & A\neq\varnothing\end{cases} \quad (4)$$

where $K=\sum_{B\cap C=\varnothing} m_1(B)m_2(C)$ is the conflict coefficient between the two BBAs.

The pignistic probability [18] corresponding to a BBA m is defined as

$$BetP(\theta_i)=\sum_{\theta_i\in B}\frac{m(B)}{|B|}, \quad \forall B\subseteq\Theta \quad (5)$$

where $|B|$ is the cardinality of the focal element B. Based on this, we can perform probabilistic decisions, as shown below.

$$i^* = \arg\max_i BetP(\theta_i) \quad (6)$$

Given two BBAs on the FOD Θ, the distance of evidence is used to measure the dissimilarity between the different BBAs. E.g., Jousselme's distance [19] is defined as

$$d_J(m_1,m_2)=\sqrt{\frac{1}{2}(m_1-m_2)^T \boldsymbol{D}(m_1-m_2)} \quad (7)$$

where $\boldsymbol{D}$ represents a $2^n\times 2^n$ matrix. Elements in $\boldsymbol{D}$ are defined as $\boldsymbol{D}(A,B)=|A\cap B|/|A\cup B|$. There are also other types of evidence distance measure, such as Tessem's distance [20], fuzzy membership-based distance [21], and belief interval-based distance [22].

B. Concept of Multimodal Coordinated Representation

During the learning phase, the multimodal coordinated representation establishes the relationships between different modalities to describe and process multimodal data more effectively, as shown in Fig. 1.

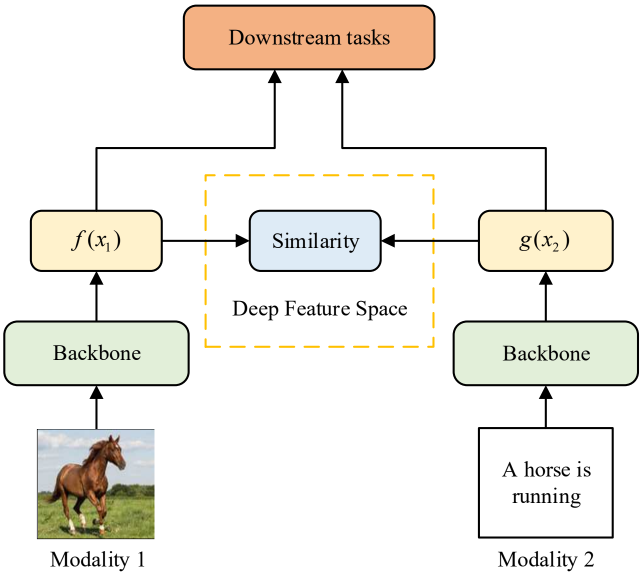

Fig. 1. Multimodal coordinated representation in deep feature space.

In multimodal coordinated representation learning, separate representations are learned for each modality and coordinated with a constraint. For example, in the cosine similarity-based method [11], given two samples x_1 and x_2 from two modalities, their corresponding deep features $f(x_1)$ and $g(x_2)$ are obtained by deep neural networks (also called the backbone). Then, the coordinated constraint is calculated as

$$similarity\_cosine(x_1,x_1)=\frac{<x_1,x_2>}{\|x_1\|\cdot\|x_2\|} \quad (8)$$

where $<x_1,x_2>$ is the dot-product of x_1 and x_2. $\|x_1\|$ is the magnitude of the vector x_1.

In Euclidean distance-based method [10], the coordinated constraint is also calculated in the deep features of x_1 and x_2. After obtaining each modality's representations (deep features), they can be used for downstream tasks, such as multimodal classification.

III. Multimodal Coordinated Representation Based on Evidential Theory

As previously mentioned, existing multimodal coordinated representation methods are implemented on the deep feature spaces or encoding space. In this paper, based on the framework of evidence theory, we propose a novel coordinated representation method, as shown in Fig.2. In our proposed method, multimodal data is described as the corresponding basic belief assignments (BBAs), and coordinated learning is implemented in the evidential space (i.e., the BBA-based space). That is, the information interaction between different modalities is implemented using evidence modeling (or uncertainty modeling). As we can see in Fig. 2, the backbone network is used for evidence modeling for each modality, with each modality's data as input and the corresponding BBA as output. After the training phase, each backbone can generate the BBA of the corresponding modality in an end-to-end manner.

In this paper, to implement coordinated learning in the evidential space, we design three types of constraint terms: classification constraint, evidential constraint, and ambiguity constraint.

A. Classification Constraint

For the multimodal classification task, we design the classification constraint (called $Loss\_classify$) to support the decision-making. $Loss\_classify$ is defined as the cross-entropy between the pignistic probability of each modality's BBA and the actual class label, as shown in Eq. (9).

$$\begin{aligned}Loss\_classify &= H(y,BetP)\\ &=-\sum_{i=1}^{N} y_i \log\left[BetP(\theta_i)\right]\end{aligned} \quad (9)$$

where $BetP$ is the pignistic probability of each modality's BBA, calculated using Eq. (5). $H[y,BetP]$ is the cross-entropy between the $BetP$ and the class label y. N is the total number of classes. y_i indicates whether the sample belongs to class i, with $y_i=1$ if it does and $y_i=0$ if it doesn't. θ_i represents the singleton element corresponding to class i.

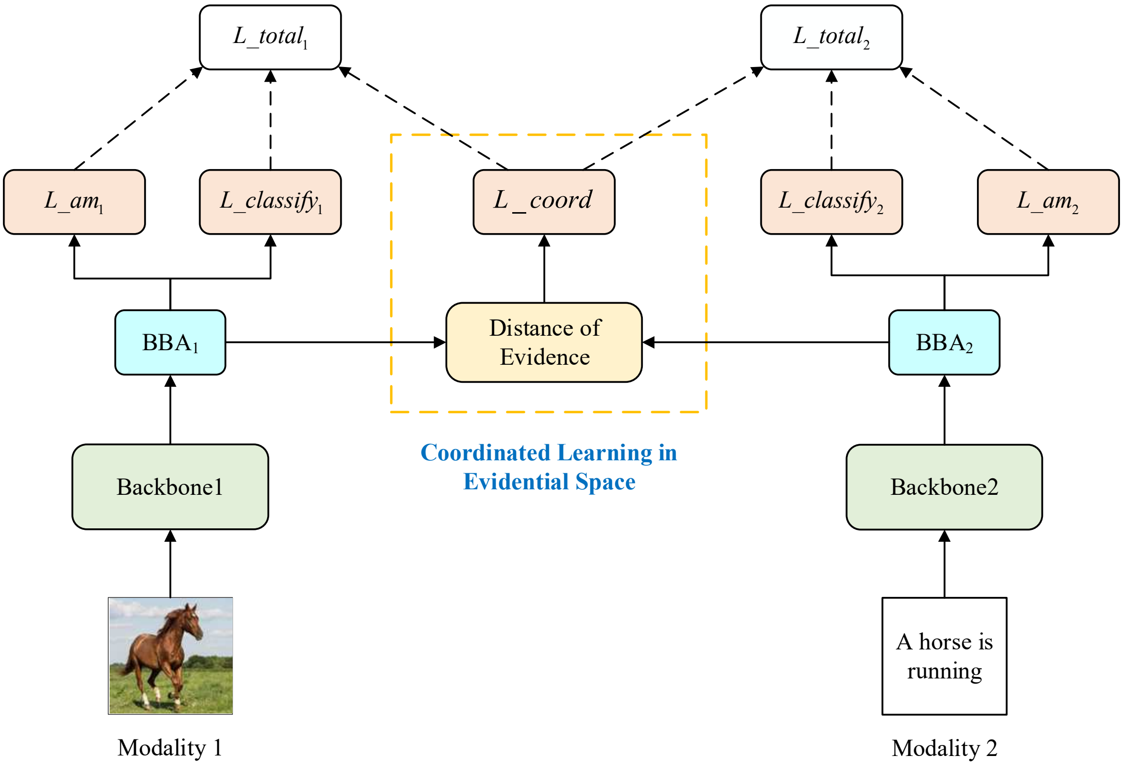

Fig. 2. Multimodal coordinated representation in evidential space.

B. *Evidential Coordinated Constraint*

To use the intra-class and inter-class difference information of multimodal data, we design an evidential coordinated constraint, denoted by L_coord. The object of L_coord is to reduce the intra-class difference and to increase the inter-class difference between different modalities.

Given samples x_1 and x_2 from two different modalities, the corresponding BBAs are generated by the backbone networks, denoted by m_1 and m_2. Subsequently, L_coord is calculated as follows.

$$L\_coord = yd_J(m_1,m_2)^2 + (1-y)max\{\varepsilon - d_J(m_1,m_2),0\}^2 \tag{10}$$

where y represents whether these two samples belong to the same class, with 1 for the same class and 0 for different ones. $d_J(m_1,m_2)$ is the evidence distance between the two BBAs m_1 and m_2, calculated by Eq. (7). ε denotes the predefined threshold for the evidence distance (we propose to set ε to 1 for the simplicity; other values can also be used).

C. *Ambiguity Constraint*

To represent each modality's data clearly, we also introduce an ambiguity constraint for coordinated learning, represented as $Loss\_am$. Given an input sample x_1 of modality 1, the ambiguity constraint is defined as the uncertainty measure of the corresponding BBA m_1. In this paper, we use multiple uncertainty measures, including the ambiguity measure (AM) [23], the aggregated uncertainty (AU) [24], and the total uncertainty (TU) [25].

1) Ambiguity Measure (AM): In AM, the uncertainty is represented by the entropy of the pignistic probability of the given BBA m, as shown below.

$$AM(m) = -\sum_{\theta\in\Theta} BetP_m(\theta)\log_2(BetP_m(\theta)) \tag{11}$$

where $BetP_m(\theta)$ is the pignistic probability of the BBA, calculated by Eq. (5).

2) Aggregated Uncertainty (AU): In AU, the probability with the maximum entropy under constraints is first selected and the uncertainty is represented by its corresponding entropy, as shown below.

$$\begin{cases} AU(m) = \max\left[-\sum_{\theta\in\Theta} p_\theta \log_2 p_\theta\right] \ s.t. \\ p_\theta \in [0,1], \quad \forall\theta\in\Theta \\ \sum_{\theta\in\Theta} p_\theta = 1 \\ Bel(A) \le \sum_{\theta\in\Theta} p_\theta \le 1 - Pl(A), \forall A \subseteq \Theta \end{cases} \tag{12}$$

where $Bel(A)$ and $Pl(A)$ are the belief function and the plausibility function of the focal element A respectively, calculated by Eq (2) and Eq. (3).

3) Total Uncertainty (TU): TU is an uncertainty measure directly based on the framework of DST. In TU, the belief interval of the single focal element is considered as an interval number, and the distance of interval numbers is used to define uncertainty, as calculated follows.

$$TU^I(m) = 1 - \frac{\sqrt{3}}{n}\cdot\sum_{i=1}^{n} d^I\left([Bel(\{\theta_i\}), Pl(\{\theta_i\})],[0,1]\right) \tag{13}$$

where d^I is the distance of interval numbers, which is defined as follows.

$$d^I([a_1,b_1],[a_2,b_2]) = \sqrt{[\frac{a_1+a_2}{2} - \frac{b_1+b_2}{2}]^2 + \frac{(a_2-a_1)^2+(b_2-b_1)^2}{12}} \tag{14}$$

In summary, the total loss function of modality 1 is defined as follows.

$$L\_total_1 = L\_classify_1 + \lambda_1 \cdot L\_coord + \lambda_2 \cdot L\_am_1 \tag{15}$$

where λ_1 and λ_2 are the regularization coefficients. In this paper, λ_1 and λ_2 is set to 0.1. In practice, other values can be chosen. The total loss function of modality 2 is similar.

After the learning phase, given the test multimodal samples, uncertainty representations (i.e., the corresponding BBAs) of each modality are obtained by the trained backbones. Based on these BBAs, the decision fusion can be implemented by the evidence combination rule and the pignistic probability transformation. This will be detailed below.

D. *Illustrative Example for Multimodal classification*

In this section, we use an illustrative example to demonstrate the procedure of our multimodal coordinated representation method based on evidence theory and its application in multimodal classification. The flowchart is shown in Fig. 3.

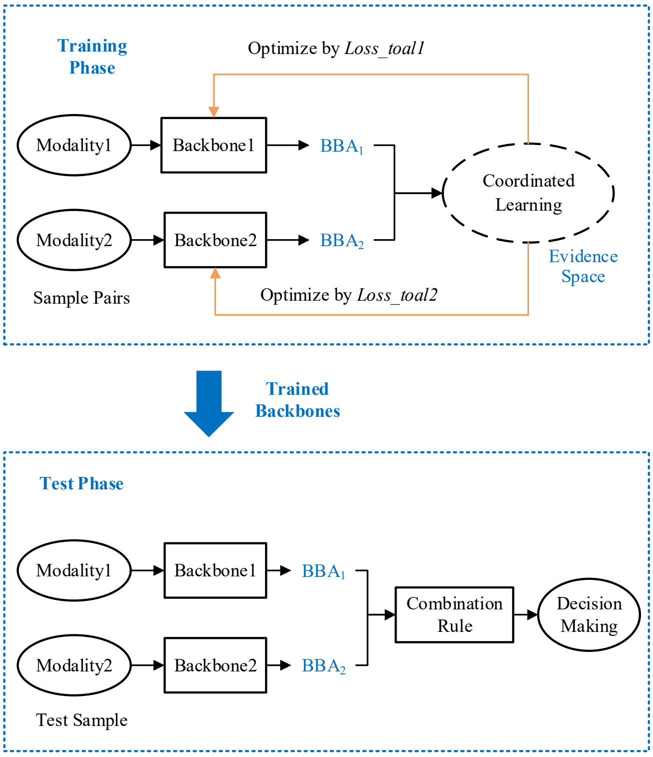

Fig. 3. The application of our method for multimodal classification.

1) Training Phase: In the training phase, we first construct the sample pairs from different modalities to serve as the training set. In this paper, we set the number of positive (belonging to the same class) and negative (belonging to different classes) sample pairs to be the same.

For these sample pairs, we use the backbone network (i.e., the deep neural network designed for each modality) to generate the BBAs corresponding to each modality. For example, we use the Resnet50 [29] as the backbone for the image modality. For the text modality, we use the BERT [30] as the backbone. For the audio modality, we use the Mel spectrogram [31] combined with the Resnet50 model to generate the corresponding BBAs.

Subsequently, the loss function of each modality is calculated: the classification constraint, the evidential coordinated constraint, and the ambiguity constraint. These constraints are summed by Eq. (15) to obtain the total loss of each modality. Based on the total loss, we optimize the corresponding backbone network. This process is iterated until all the backbones converge.

2) Test Phase: After the training phase, given the test samples of each modality, we use the trained backbone network to generate the corresponding BBAs, denoted as BBA_1 and BBA_2 in Fig.3. Next, we use Dempster's rule of combination to obtain the combined BBA, as shown in Eq. (4). Finally, the decision-making is implemented by the pignistic probability transformation, as shown in Eq. (5) and Eq. (6).

IV. Experiments

In this section, we implement experiments on multiple multimodal emotion classification datasets to evaluate the effectiveness of our proposed method. In the experiments, two of the three modalities (image, text, audio) from the CMU-MOSI [26] and CMU-MOSEI [27] are selected, resulting in eight datasets. The characteristics of these datasets are detailed in TABLE I.

TABLE I
Characteristics of Datasets

Dataset	Modalities
CMU-MOSI (I+T)	Image + Text
CMU-MOSI (I+A)	Image + Audio
CMU-MOSI (T+A)	Text + Audio
CMU-MOSI (I+T+A)	Image + Text + Audio
CMU-MOSEI (I+T)	Image + Text
CMU-MOSEI (I+A)	Image + Audio
CMU-MOSEI (T+A)	Text + Audio
CMU-MOSEI (I+T+A)	Image + Text + Audio

In the experiments, for the image modality, the Resnet50 [28] is used as the feature extraction network (i.e., the backbone). For the text modality, the BERT [29] is used as the backbone. For the audio modality, the Mel spectrogram [30] combined with the Resnet50 model is used to extract deep features. We compare the classification performance (accuracy and F1-score) of our proposed method with two existing multimodal representation methods: the Euclidean-distance-based method [10] and the cosine-similarity-based method [11]. These traditional methods are all implemented in the deep feature space extracted by ResNet50 and BERT (same backbone as our method). In these traditional methods, the loss function of each modality comprises two parts: the coordinated loss (see in Section II.B) and the classification loss (i.e., the cross-entropy function [10-11]). Additionally, we compare the performance of various uncertainty measures: the ambiguity measure (AM), the aggregated uncertainty (AU), and the total uncertainty (TU). These measures are respectively used as the ambiguity constraints. In our experiments, each dataset is randomly divided into two parts, with 50% assigned to the training set and the remaining 50% to the test set. Experiment on each dataset is randomly performed ten times. The results are shown in TABLE II, where the EMCR-AU represents our evidence theory-based multimodal coordinated representation using the AU-based ambiguity constraint. EMCR-AM and EMCR-TU are similar.

TABLE II
RESULTS ON MULTIMODAL EMOTION CLASSIFICATION DATASETS

Dataset	Average ± Std/%	Euclidean	Cosine	EMCR-AU	EMCR-AM	EMCR-TU
CMU-MOSI (I+T)	Accuracy	74.34±0.65	75.20±1.03	77.10±0.36	78.68±0.21	**79.62±1.19**
	F1-Score	75.49±0.40	74.73±1.36	77.67±1.42	77.04±0.56	**79.32±0.76**
CMU-MOSI (I+A)	Accuracy	73.04±0.37	73.97±0.62	74.10±1.16	75.48±0.98	**77.99±0.73**
	F1-Score	74.50±1.17	74.02±1.44	76.27±0.69	76.03±0.77	**78.44±0.58**
CMU-MOSI (T+A)	Accuracy	71.93±0.26	72.75±1.08	74.92±1.42	73.86±0.84	**76.47±1.08**
	F1-Score	72.59±0.44	72.60±0.23	73.25±1.01	73.72±1.02	**74.86±0.34**
CMU-MOSI (I+T+A)	Accuracy	75.40±1.10	75.78±1.36	78.71±0.76	80.50±1.45	**81.17±1.00**
	F1-Score	75.55±1.22	76.59±1.12	78.70±1.34	79.58±0.94	**81.38±0.27**
CMU-MOSEI (I+T)	Accuracy	69.42±0.63	70.41±1.36	74.77±1.35	75.43±0.37	**77.89±0.72**
	F1-Score	69.30±0.34	71.55±0.95	75.91±0.56	75.96±1.03	**76.78±1.07**
CMU-MOSEI (I+A)	Accuracy	67.07±0.99	69.84±1.17	72.66±0.92	73.12±0.34	**75.66±1.34**
	F1-Score	66.39±0.62	69.10±0.50	73.87±0.61	74.65±1.18	**75.84±0.23**
CMU-MOSEI (T+A)	Accuracy	64.43±0.41	66.84±0.37	72.69±0.79	72.38±1.19	**73.77±0.43**
	F1-Score	64.03±1.26	66.98±0.38	72.79±0.52	**73.53±1.50**	73.36±0.20
CMU-MOSEI (I+T+A)	Accuracy	70.19±0.67	71.54±0.62	77.02±0.56	78.77±0.58	**79.22±0.56**
	F1-Score	69.71±0.48	71.03±0.43	77.39±0.44	77.98±1.09	**79.89±0.46**

As we can see, our proposed method outperforms the traditional Euclidean distance-based and cosine similarity-based methods on all datasets, demonstrating our method's rationality and effectiveness. Furthermore, among several measures of uncertainty, TU achieved better performance. This indicates that the TU, which is directly based on the DST framework, is more suited for describing the ambiguity of BBA, thereby enhancing the performance of our method.

V. CONCLUSIONS

In this paper, based on the framework of evidence theory, we propose a novel coordinated representation method, where multimodal data is described as the corresponding BBAs, and coordinated learning is implemented in the evidential space. To use the intra-class and inter-class difference information of multimodal data, we design an evidential coordinated constraint using the distance of evidence. Additionally, to represent each modality clearly, we introduce an ambiguity constraint. Experimental results of multimodal classification illustrate the effectiveness and rationality of our proposed method.

Note that in our method, the distance of evidence is defined as the Jousselme's distance. In our future work, we will try to use more types of distance [20-22], and try to use more uncertainty measures reported in [31-32]. Furthermore, we will apply our method to other multimodal tasks, such as multimodal retrieval and multimodal alignment. Moreover, we will try to use the triplet loss [33] to solve the problem with three or more modalities.

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423-443, Feb. 1, 2019.

[2] P. Xu, X. Zhu, D. A. Clifton, "Multimodal Learning with Transformers: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113-12132, Oct. 2023.

[3] A. Rahate, R. Walambe, S. Ramanna, K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, 2022.

[4] H. Y. Zhou, Y. Yu, C. Wang, et al., "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics," *Nat. Biomed. Eng.*, vol. 7, pp. 743–755, 2023.

[5] M. Moor, Q. Huang, S. Wu, et al., "Med-Flamingo: A Multimodal Medical Few-shot Learner," in *Proc. of the 3rd Machine Learning for Health Symposium*, PMLR, vol. 225, pp. 353-367, 2023.

[6] L. Cheng, Z. An, Y. Guo, M. Ren, Z. Yang, S. McLoone, "MMFSL: A novel multimodal few-shot learning framework for fault diagnosis of industrial bearings," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1-13, 2023.

[7] G. Yang, H. Tao, T. Yu, R. Du, Y. Zhong, "Online Fault Diagnosis of Harmonic Drives Using Semisupervised Contrastive Graph Generative

Network via Multimodal Data," *IEEE Trans. Ind. Electron.*, vol. 71, no. 3, pp. 3055-3063, Mar. 2024.

[8] J. Li, H. Dai, H. Han, Y. Ding, "MSeg3D: Multi-Modal 3D Semantic Segmentation for Autonomous Driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 21694-21704.

[9] W. Wu, X. Deng, P. Jiang, S. Wan, Y. Guo, "CrossFuser: Multi-Modal Feature Fusion for End-to-End Autonomous Driving Under Unseen Weather Conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14378-14392.

[10] Z. Zeng, Y. Sun, W. Mao, "MCCN: Multimodal Coordinated Clustering Network for Large-Scale Cross-Modal Retrieval," in Proc. of the 29th ACM Int. Conf. on Multimedia, 2021, pp. 5427–5435.

[11] J. Weston, S. Bengio, N. Usunier, "WSABIE: Scaling up to large vocabulary image annotation," in *Proc. of the 7th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2011, pp. 2764–2770.

[12] Y. Wan, J. Shu, Sui, et al., "Multi-modal Attention Network Learning for Semantic Source Code Retrieval," in *Proc. 34th IEEE/ACM Int. Conf. Automated Software Eng. (ASE)*, San Diego, CA, USA, 2019, pp. 13-25.

[13] A. Frome, G. Corrado, J. Shlens, "DeViSE: A deep visual-semantic embedding model," in *Proc. of the 28th Int. Conf. on Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 2121–2129.

[14] Z. Zhang, H. Luo, L. Zhu, G. Lu, and H. T. Shen, "Modality-Invariant Asymmetric Networks for Cross-Modal Hashing," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 5091-5104, May 1, 2023.

[15] T. Hoang, T.-T. Do, T. V. Nguyen, and N.-M. Cheung, "Unsupervised Deep Cross-Modality Spectral Hashing," *IEEE Trans. Image Process.*, vol. 29, pp. 8391-8406, 2020.

[16] A. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Stat.*, vol. 38, no. 2, pp. 325-339, 1967.

[17] G. Shafer, *A Mathematical Theory of Evidence*, vol. 1, Princeton, NJ, USA: Princeton Univ. Press, 1976.

[18] P. Smets, "Constructing the Pignistic Probability Function in a Context of Uncertainty," *Mach. Intell. Pattern Recognit.*, vol. 10, pp. 29-39, 1990.

[19] A. L. Jousselme, D. Grenier, É. Bossé, "A new distance between two bodies of evidence," Inf. Fusion, vol. 2, no. 2, pp. 91-101, 2001.

[20] B. Tessem, "Approximations for efficient computation in the theory of evidence," *Artificial Intelligence*, vol. 61, no. 2, pp. 315–329, 1993.

[21] D. Han, J. Dezert, C. Han, and Y. Yang, "New dissimilarity measures in evidence theory," in *Proc. 14th Int. Conf. on Information Fusion*, Chicago, IL, USA, 2011, pp. 1-7.

[22] D. Han, J. Dezert, and Y. Yang, "Belief Interval-Based Distance Measures in the Theory of Belief Functions," *IEEE Trans. on Syst., Man, and Cybern.: Syst.*, vol. 48, no. 6, pp. 833-850, 2018.

[23] A.L. Jousselme, C.S. Liu, D. Grenier, and É. Bossé, "Measuring ambiguity in the evidence theory," *IEEE Trans. Syst., Man, Cybern.* Part A, vol. 36, no. 5, pp. 890–903, 2006.

[24] D. Harmanec and G.J. Klir, "Measuring total uncertainty in Dempster–Shafer theory – a novel approach," *Int. J. Gen. Syst.*, vol. 22, no. 4, pp. 405–419, 1994.

[25] Y. Yang and D. Han, "A new distance-based total uncertainty measure in the theory of belief functions," Knowl.-Based Syst., vol. 94, pp. 114-123, 2016.

[26] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint,* arXiv:1606.06259, 2016.

[27] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguist.* (Vol. 1: Long Papers), Melbourne, Australia, 2018, pp. 2236–2246.

[28] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770-778.

[29] J. Kenton, J. Devlin, M.-W. Chang, L. Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL-HLT*, vol. 1, 2019.

[30] A. Ustubioglu, B. Ustubioglu, G. Ulutas, "Mel spectrogram-based audio forgery detection using CNN," *Signal Image Video Process.*, vol. 17, no. 5, pp. 2211-2219, 2023.

[31] J. Dezert, "An Effective Measure of Uncertainty of Basic Belief Assignments," in Proc. 25th Int. Conf. on Information Fusion (FUSION), Linköping, Sweden.

[32] J. Dezert and A. Tchamova, "On the effectiveness of measures of uncertainty of basic belief assignments," Information & Security Journal, vol. 52, pp. 9-36, Feb. 2022.

[33] Schroff, F., Kalenichenko, D., and Philbin, J. "Facenet: A unified embedding for face recognition and clustering." In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815-823.