Trusted Multi-View Deep Learning with Opinion Aggregation

Wei Liu, Xiaodong Yue, Yufei Chen, Thierry Denoeux

Trusted Multi-View Deep Learning with Opinion Aggregation

Wei Liu[1], **Xiaodong Yue**[1,2 *], **Yufei Chen**[3], **Thierry Denoeux**[4,5]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[2] Artificial Intelligence Institute of Shanghai University, Shanghai, China
[3] College of Electronics and Information Engineering, Tongji University, Shanghai, China
[4] Université de technologie de Compiègne, CNRS UMR 7253 Heudiasyc, Compiègne, France
[5] Shanghai University, UTSEUS, Shanghai, China
ldachuan@outlook.com, yswantfly@shu.edu.cn, yufeichen@tongji.edu.cn, thierry.denoeux@utc.fr

Abstract

Multi-view deep learning is performed based on the deep fusion of data from multiple sources, i.e. data with multiple views. However, due to the property differences and inconsistency of data sources, the deep learning results based on the fusion of multi-view data may be uncertain and unreliable. It is required to reduce the uncertainty in data fusion and implement the trusted multi-view deep learning. Aiming at the problem, we revisit the multi-view learning from the perspective of opinion aggregation and thereby devise a trusted multi-view deep learning method. Within this method, we adopt evidence theory to formulate the uncertainty of opinions as learning results from different data sources and measure the uncertainty of opinion aggregation as multi-view learning results through evidence accumulation. We prove that accumulating the evidences from multiple data views will decrease the uncertainty in multi-view deep learning and facilitate to achieve the trusted learning results. Experiments on various kinds of multi-view datasets verify the reliability and robustness of the proposed multi-view deep learning method.

Introduction

In real-world applications, data is usually represented with different views, including multiple modalities or various types of features, which leads a growing interest in multi-view learning. With the development of the deep learning, most of the existing multi-view learning methods tend to integrate multi-view information with deep neural networks to achieve state-of-the-art performance in various application domains (Wang et al. 2015a; Andrew et al. 2013; Wang et al. 2018; Tao et al. 2019; Tian, Krishnan, and Isola 2019; Bachman, Hjelm, and Buchwalter 2019; Sun, Liu, and Mao 2019; Zhang et al. 2019, 2020; Sun, Dong, and Liu 2020). However, due to the property differences and inconsistency of multiple data sources, the results learned from multi-view deep learning method may be uncertain and unreliable, because the traditional convolutional neural networks focus on the accuracy of the classifications but ignore the credibility of the results, which makes a great limitation in various kinds of applications, especially safety-critical applications (e.g., medical diagnosis or autonomous driving).

To address this limitation, an uncertainty-aware trusted Multi-view Classification (TMC) method (Han et al. 2021) was proposed recently. TMC focuses on combining different views at an evidence level in terms of the Dempster's rule of combination to produce a reliable classification result. However, it does not guarantee the decrease of the overall uncertainty when integrating the uncertain information extracted from multi-view data and does not consider the "consistency" in multi-view learning for avoiding the conflict across information captured from different views. Moreover, the fusion with Dempster's rule will produce counter-intuitive results (Zadeh 1984). Therefore, this has motivated us to revisit the multi-view learning from the perspective of opinion aggregation and thereby develop a trusted multi-view deep learning method.

Opinion aggregation aims at aggregating multiple opinions within a group in support of group decision making, which is the same as the fusion process for multi-view learning. However, the opinions from multiple views about the same domain are always unreliable because of the various sensor qualities or environmental factors, which adds more uncertainty to the decision-making process. A good opinion aggregation process should consist of two necessary parts: 1) a trusted aggregation strategy, which can reduce the overall uncertainty after aggregation, 2) maximization of consistency across views for avoiding conflict between multiple opinions. Therefore, from the perspective of a good opinion aggregation structure, we devise a trusted multi-view deep learning method. Within this method, we adopt the evidence theory to represent opinions as beliefs about the truth of propositions under degrees of uncertainty. In this opinion representation, the beliefs mean the evidences support for those class probabilities and uncertainty means the vacuity of evidence, which allows explicit expression of level of trust for the results learned from different views. Then, guided by the mapping between opinions and Dirichlet PDFs, we integrate the opinions in terms of the evidence accumulation, which can increase the evidences support for class probabilities and decrease the vacuity of evidence and thereby increase the reliability of multi-view deep learning results. In summary, our contributions of this paper are:

(1) We construct a trusted multi-view deep learning method through simulating opinion aggregation mechanism to achieve trusted learning results. The proposed method

*Corresponding author

adopts the evidence theory to formulate the opinions as learning results from different data sources and represents the integrated opinion as multi-view learning result through opinion aggregation with evidence accumulation, which can precisely estimate the uncertainty of results in multi-view deep learning.

(2) We theoretically prove that accumulating the evidences from multiple data views will decrease the overall uncertainty and prediction error of multi-view learning results, which facilitates to produce trusted and accurate learning results. Moreover, we further extend our method by minimizing the opinion entropy across views for guaranteeing the consistency across multiple views.

(3) We conduct extensive experiments over various kinds of real-world data to validate the effectiveness of the proposed model in accuracy, reliability and robustness.

Related Work

Multi-View Learning: Multi-view learning receives increasing interest in recent years to analyze complex data. The traditional representative methods are canonical correlation analysis (CCA) (Harold 1936) and its variants (Bach and Jordan 2002; Hardoon and Shawe-Taylor 2011; Wang 2007). CCA maximizes the correlation between different views to find a common representation. Kernel CCA (Bach and Jordan 2002) develops CCA to nonlinear conditions, which makes the CCA more robust. Sparse CCA (Hardoon and Shawe-Taylor 2011) learns sparse representation to reduce the effect of noisy data. BCCA presents (Wang 2007) a Bayesian model selection algorithm for CCA based on a probabilistic interpretation. Different from CCA, some methods (Zhao, Ding, and Fu 2017; Zhang et al. 2018b; Liu et al. 2015) obtain hierarchical representation from multi-view data through matrix factorization. Multi-view dimensionality co-reduction (MDcR) (Zhang et al. 2016) applies the kernel matching to regularize the dependence across views. Nonparametric sparse learning method (NSMD) (Liu et al. 2017) develops an effective sparse learning method for cross-view dimensionality reduction. Consensus and complementarity based maximum entropy discrimination (MED-2C) (Chao and Sun 2016) proposes a multi-view classification based on the two principles consensus and complementarity. Furthermore, Self-representation is also introduced to better incorporate multi-view information (Xie et al. 2018, 2020). Kernelized version of tensor-based multi-view subspace clustering (Kt-SVD-MSC) (Xie et al. 2018) jointly learns self-representation coefficients in mapped high-dimensional spaces. Moreover, with the development of the deep learning, some works (Andrew et al. 2013; Wang et al. 2015a, 2018; Tao et al. 2019; Tian, Krishnan, and Isola 2019; Bachman, Hjelm, and Buchwalter 2019; Sun, Liu, and Mao 2019; Sun, Dong, and Liu 2020) combine deep learning with multi-view learning. Deep CCA (DCCA) (Andrew et al. 2013) is more powerful to capture nonlinear relationships. Deep canonically correlated autoencoder (DCCAE) (Wang et al. 2015a) learns compact representation by combining deep CCA and autoencoder, which is more useful to extract nonlinear relationships. In addition, generative adversarial network is applied to handle missing view problem (Wang et al. 2018) or impose prior information (Tao et al. 2019). However, these methods achieve a great performance on multi-view classification, but they rarely consider the reliability of the classification result. Recently, a trusted multi-view classification method (Han et al. 2021) has been proposed, which focuses on the uncertainty estimation problem and produces a reliable classification result. Nonetheless, it does not guarantee the decrease of overall uncertainty after fusion of different views and does not consider the consistency across views. Moreover, the fusion with Dempster's rule in TMC will produce counter-intuitive results (Zadeh 1984). In contrast, our method explores the consistency between different views from the perspective of opinion aggregation and reduces the overall uncertainty after fusing different opinions, which guides an accurate, robust and trusted result.

Opinion Aggregation: Decision making is a pervasive part of life. Every day we are confronted with deciding between multiple choices. Opinion aggregation aims at aggregating multiple opinions within a group, which is very useful for group decision making. Due to its effectiveness, opinion aggregation has been widely used in various applications (Zadeh 1986; Ding and Liu 2007; Liu et al. 2007; Martini and Sprenger 2017; Iso et al. 2021; Belluti et al. 2013).

Mechanism of Opinion Aggregation

Due to the property differences and inconsistency of data sources, results from multiple views about the same domain may be unreliable, which adds more uncertainty to the decision-making process. For reducing the uncertainty in data fusion to obtain the trusted multi-view deep learning results, we revisit the multi-view deep learning from the perspective of opinion aggregation and thereby implement a trusted multi-view deep learning method. In this section, we will describe the mechanism of opinion aggregation for the proposed trusted multi-view deep learning, as shown in Figure 1. Within this mechanism, we first adopt the evidence theory to formulate the information extracted from different views with neural networks as corresponding opinions (step (1) in Figure 1), which can precisely estimate the uncertainty of results from different views. Then we integrate multi-opinions to obtain a unified opinion with evidence accumulation (step (2) in Figure 1), which can decrease the overall uncertainty. Furthermore, we measure the consistency across opinions based on the opinion entropy (step (3) in Figure 1), which can avoid the conflict between different opinions. Details are described as following.

Opinion Representation under Evidence Theory

Traditional neural classification networks usually use the softmax as the standard output. However, using the softmax only obtains the class probabilities but ignores the reliability of the results. To address this problem, the softmax layer is replaced by an activation layer (i.e. ReLU) to obtain a non-negative output, termed as evidence (Sensoy, Kaplan, and Kandemir 2018) in this work. Then we adopt the evidence

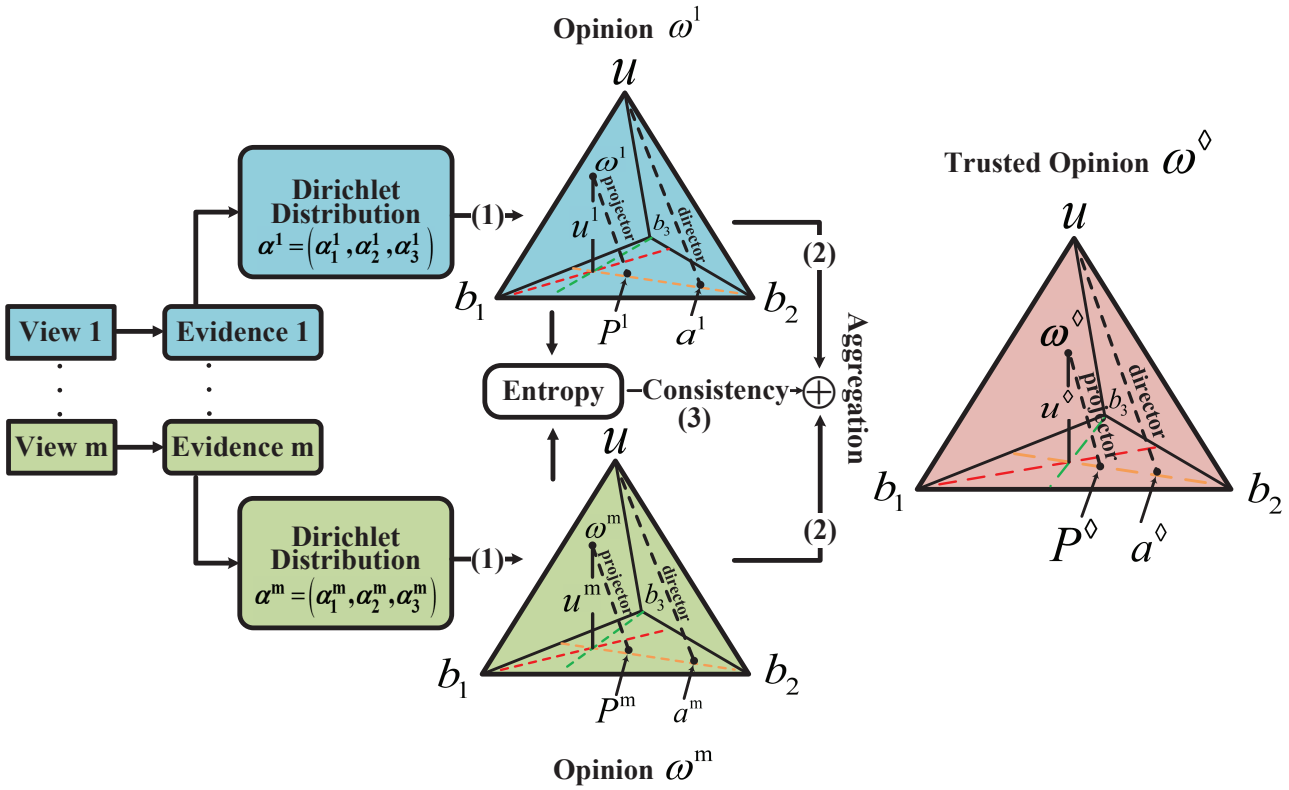


Figure 1: Illustration of proposed method. (1) The evidence extracted from a neural network is represented as an opinion. (2) Opinion aggregation with evidence accumulation. (3) Maximization of the consistency based on the opinion entropy across views. In this method, assume a 3-classification tasks, the opinion $\boldsymbol{\omega}$ from this task is visualized as a point inside a tetrahedron, which in fact is a barycentric coordinate system of four axes. The vertical elevation of the opinion point inside the tetrahedron represents the uncertainty mass u; The distances from each of the three triangular side planes to the opinion point represent the respective belief masses $\boldsymbol{b} = (b_1, b_2, b_3)^{\mathrm{T}}$; The base rate distribution $\boldsymbol{a}$ is indicated as a point on the base triangular plane. The line that joins the tetrahedron summit and the base rate point represents the director. The projected probability distribution $\boldsymbol{P}$ point is geometrically determined by tracing a projection from the opinion point, parallel to the director, onto the base plane.

theory to formalize the evidence as opinion to explicitly express the uncertainty degree of deep learning result.

In evidence theory, for a K-classification problem, a multinomial opinion $\boldsymbol{\omega} = (\boldsymbol{b}, u, \boldsymbol{a})$ is always a trinomial opinion visualized as a barycentric polyhedron, as shown in Figure 1 (in case of 3-classification problems), where u indicates the overall uncertainty which represents the vacuity of evidence, $\boldsymbol{b} = (b_1, \ldots, b_k)^{\mathrm{T}}$ represents the belief degree for the k^{th} class, $\boldsymbol{a} = (a_1, \ldots, a_k)^{\mathrm{T}}$ indicates the prior preference over class k and we have $\sum_{k=1}^{K} b_k + u = 1$. Then the probability that the data is assigned to class k is defined by $P(k) = b_k + a_k u$, for $k = 1, ..., K$. Typically, all values of the $\boldsymbol{a}$ are set to $1/K$ when there is no preference over class.

Let the expected probability distribution derived from Dirichlet distribution be equal with the projected probability distribution derived from the opinion in evidence theory. Then we have a mapping between opinion and Dirichlet distribution (Jøsang 2018),

$$\boldsymbol{\omega} = (\boldsymbol{b}, u, \boldsymbol{a}) \leftrightarrow Dir(\boldsymbol{P} | \boldsymbol{\alpha}), \tag{1}$$

where $\boldsymbol{P} = (P_1, ..., P_k)^T$ is the probability that the data is assigned to k^{th} class, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)^T$ represents the Dirichlet parameters and we have $\boldsymbol{\alpha} = \boldsymbol{e} + \boldsymbol{a}K$, where $\boldsymbol{e} = (e_1, \ldots, e_k)^{\mathrm{T}}$ indicates the amount of support evidence collected from neural network in favor of a sample to be classified into k^{th} class. Noted that, when there is no preference over class, the Dirichlet parameters $\boldsymbol{\alpha} = \boldsymbol{e} + 1$. Then the belief $\boldsymbol{b}$ and uncertainty mass u are calculated as

$$b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S}, u = \frac{K}{S}, \tag{2}$$

where $S = \sum_{k=1}^{K} (e_k + 1) = \sum_{k=1}^{K} \alpha_k$ is the Dirichlet strength. That is, the Dirichlet distribution parametrized over evidence represents the density of such probability assignment, it represents the predictions of the learner as a distribution over possible softmax outputs, which models the second-order probabilities to indicate the uncertainty of the neural network result.

Finally, according to equations 1 and 2, we could translate the output of neural network $\boldsymbol{e} = (e_1, \ldots, e_k)^{\mathrm{T}}$ into opinion

$\boldsymbol{\omega} = (\boldsymbol{b}, u, \boldsymbol{a})$ (step (1) in Figure 1), which allows us flexibly integrate multiple views for trusted decision making.

Opinion Aggregation with Evidence Accumulation

The opinion of single view has been formalized above, which allows explicit expression of uncertainty degree. Now, we begin to focus on the opinion aggregation with multi-view deep learning. Particularly, we use the evidence accumulation in evidence theory to combine multiple opinions, which can reduce the overall uncertainty. The Definition of opinion aggregation with evidence accumulation is described as below.

Definition 1. ***Opinion aggregation with evidence accumulation.*** *The opinion aggregation with evidence accumulation simply consists of evidence parameter addition. Given a data with M multiple views for the same K-classification problem, we can obtain a set of evidences $\{\boldsymbol{e}^m\}_{m=1}^M$, collected from M neural networks and a set of opinions $\{\boldsymbol{\omega}^m\}_{m=1}^M$ in terms of the equation 2. Then we have an integrated opinion*

$$\boldsymbol{\omega}^{\diamondsuit(M)} = \underset{m=1,\ldots,M}{\oplus}(\boldsymbol{\omega}^m) = \left(\boldsymbol{b}^{\diamondsuit(M)}, u^{\diamondsuit(M)}, \boldsymbol{a}^{\diamondsuit(M)}\right). \quad (3)$$

For $k = 1, ..., K$, we have

$$b_k^{\diamondsuit(M)} = \frac{e_k^{\diamondsuit(M)}}{S^{\diamondsuit(M)}}, u^{\diamondsuit(M)} = 1 - \sum_{k=1}^{K} b_k^{\diamondsuit(M)}, a_k^{\diamondsuit(M)} = \frac{1}{K}. \quad (4)$$

Where $S^{\diamondsuit(M)} = \sum_{k=1}^{K}\left(e_k^{\diamondsuit(M)} + 1\right)$ is the Dirichlet strength, $e_k^{\diamondsuit(M)} = \sum_{m=1}^{M} e_k^m$ represents the process of evidence accumulation.

Following the Definition 1, we can obtain the integrated opinion $\boldsymbol{\omega}^{\diamondsuit(M)} = \left(\boldsymbol{b}^{\diamondsuit(M)}, u^{\diamondsuit(M)}, \boldsymbol{a}^{\diamondsuit(M)}\right)$. The corresponding integrated parameters of the Dirichlet PDF are induced as $\alpha_k^{\diamondsuit(M)} = e_k^{\diamondsuit(M)} + 1$.

Measure of Consistency across Opinions

Our method for multi-view fusion has been described, which projects the outputs of neural networks to opinions at evidence level and then combines these opinions in terms of evidence accumulation. However, in some cases, the opinions collected from multiple views are inconsistent. To avoid the conflict between multiple opinions, we introduce a consistency measure named opinion entropy in Definition 2, which can guarantee the consistency across multiple views. The definition of opinion entropy is described as follows.

Definition 2. ***Opinion Entropy across two opinions***. *Given any view denoted by the opinion $\boldsymbol{\omega} = (\boldsymbol{b}, u, \boldsymbol{a})$ for K-classification task, the entropy of the opinion (Jøsang 2018) is defined as*

$$H(\boldsymbol{\omega}) = -\sum\nolimits_{k=1}^{K} P_k \log_2(P_k), \quad (5)$$

where $P_k = b_k + a_k u$. Then the opinion entropy between two opinions $\boldsymbol{\omega}^1$ and $\boldsymbol{\omega}^2$ is computed as

$$E\left(\boldsymbol{\omega}^1, \boldsymbol{\omega}^2\right) = H\left(\frac{\boldsymbol{\omega}^1 + \boldsymbol{\omega}^2}{2}\right) - \frac{1}{2}H\left(\boldsymbol{\omega}^1\right) - \frac{1}{2}H\left(\boldsymbol{\omega}^2\right), \quad (6)$$

where $H\left(\frac{\boldsymbol{\omega}^1+\boldsymbol{\omega}^2}{2}\right) = -\sum_{k=1}^{K}\frac{P_k^1+P_k^2}{2}\log_2\left(\frac{P_k^1+P_k^2}{2}\right)$.

Trusted Multi-view Deep Learning with Opinion Aggregation

In this section, we will discuss how to train our multi-view deep learning network. The neural network can capture the evidence from input to induce a classification opinion. Therefore, the traditional neural network can be naturally transformed into the evidence-based neural network (Şensoy, Kaplan, and Kandemir 2018) with minor changes which only replace the softmax layer with an activation layer (e.g., ReLU) to provide non-negative output, termed as the evidence $\boldsymbol{e} = (e_1, \ldots, e_k)^{\mathrm{T}}$. Accordingly, the parameters of the Dirichlet distribution $\boldsymbol{\alpha} = \boldsymbol{e} + 1$ can be obtained.

Within the proposed method, for the i^{th} sample $\boldsymbol{x}_i$, the overall loss objective is

$$L_{overall}(\boldsymbol{\alpha}_i) = L_{acc}(\boldsymbol{\alpha}_i) + \lambda L_{con}(\boldsymbol{\alpha}_i), \quad (7)$$

where $L_{acc}(\boldsymbol{\alpha}_i)$ is the prediction loss term, $L_{con}(\boldsymbol{\alpha}_i)$ is the loss of the consistency regulation across views, λ is a weight parameter with the range of [0, 1] to control the weight of consistency loss function. The details of these two loss functions are shown in the following subsection.

Prediction Loss Term

For training example $\boldsymbol{x}_i$, let $\boldsymbol{y}_i$ encodes the ground-true class label k by setting $y_{ik} = 1$ and $y_{ij} = 0$, $\forall j \neq k$. Let $\mathrm{Cat}\left(\widehat{y}_i = k \,|\boldsymbol{P}_i\right)$ be the likelihood, where $\boldsymbol{P}_i \sim Dir(\boldsymbol{P}_i|\boldsymbol{\alpha}_i)$, $\boldsymbol{P_i} = (P_{i1}, ..., P_{ik})^T$ and the parameters $\boldsymbol{\alpha}_i = \boldsymbol{e}_i + 1$. The expected sum of squares loss after the aggregation of a set of opinions $\{\boldsymbol{\omega^m}\}_{m=1}^M$ is defined as

$$\begin{aligned} L_{acc}(\boldsymbol{\alpha}_i) &= L_{acc}\left(\boldsymbol{\alpha}_i^{\diamondsuit(M)}\right) \\ &= \mathbb{E}_{\boldsymbol{P}_i \sim Dir\left(\boldsymbol{P}_i \middle| \boldsymbol{\alpha}_i^{\diamondsuit(M)}\right)} \left\|\boldsymbol{y}_i - \boldsymbol{P}_i\right\|_2^2 \\ &= \sum_{j=1}^{K}\left(y_{ij}^2 - 2y_{ij}\mathbb{E}\left[P_{ij}\right] + \mathbb{E}\left[P_{ij}^2\right]\right), \end{aligned} \quad (8)$$

where $\boldsymbol{\alpha}_i^{\diamondsuit(M)} = \boldsymbol{e}_i^{\diamondsuit(M)} + 1$ and $\boldsymbol{e}_i^{\diamondsuit(M)} = \boldsymbol{e}_i^1 + \cdots + \boldsymbol{e}_i^M$ is the process of evidence accumulation, which can increase the amount of support in favor of sample $\boldsymbol{x}_i$ to be classified into k^{th} class and decrease the overall uncertainty. Intuitively, $\mathbb{E}\left[P_{ij}^2\right] = \mathbb{E}[P_{ij}]^2 + \mathrm{Var}(P_{ij})$, then we get the following easily interpretable form

$$\begin{aligned} L_{acc}(\boldsymbol{\alpha}_i) &= \sum_{j=1}^{K}\left(y_{ij} - \mathbb{E}\left[P_{ij}\right]\right)^2 + \mathrm{Var}\left(P_{ij}\right) \\ &= \sum_{j=1}^{K}\underbrace{\left(y_{ij} - \widehat{p}_{ij}\right)^2}_{L_{err}\left(\alpha_{ij}^{\diamondsuit(M)}\right)} + \underbrace{\frac{\widehat{p}_{ij}\left(1 - \widehat{p}_{ij}\right)}{\left(S_i^{\diamondsuit(M)} + 1\right)}}_{L_{\mathrm{var}}\left(\alpha_{ij}^{\diamondsuit(M)}\right)}, \end{aligned} \quad (9)$$

where $S_i^{\diamondsuit(M)} = \sum_{k=1}^{K}\alpha_{ik}^{\diamondsuit(M)}$ is the Dirichlet strength, $\widehat{p}_{ij} = \alpha_{ij}^{\diamondsuit(M)} \Big/ S_i^{\diamondsuit(M)}$ is the expectation of the Dirichlet

distribution. It is obvious that the loss aims to achieve the joint goals of minimizing the prediction error $L_{err}\left(\boldsymbol{\alpha}_i^{\diamondsuit(M)}\right)$ and the variance $L_{\text{var}}\left(\boldsymbol{\alpha}_i^{\diamondsuit(M)}\right)$ of the integrated opinions by decomposing the first and second terms. In addition, our loss objective has the following propositions.

Proposition 1. *By integrating the evidence of the correct label from different views in terms of opinion aggregation with evidence accumulation, the prediction error loss $L_{err}\left(\boldsymbol{\alpha}_i^{\diamondsuit(M)}\right)$ will be smaller than the prediction error loss from a single view $L_{err}\left(\boldsymbol{\alpha}_i^{m}\right)$, for $m = 1, ..., M$.*

Proof 1. Let $e_{ij}^m > 0$ be the evidence of the j^{th} class extracted from the m^{th} view classifier for the i^{th} sample with correct label j, $e_{ij}^{\diamondsuit(M)} > 0$ be the integrated evidence from evidence accumulation of M views. After the opinion aggregation, $L_{err}\left(\boldsymbol{\alpha}_i^{\diamondsuit(M)}\right)$ is updated as

$$\underbrace{\left(1-\frac{\alpha_{ij}^{\diamondsuit(M)}}{S_i^{\diamondsuit(M)}}\right)^2}_{y_{ij}=1}+\underbrace{\sum_{k\neq j}\left(\frac{\alpha_{ik}^{\diamondsuit(M)}}{S_i^{\diamondsuit(M)}}\right)^2}_{y_{ik}=0}, \tag{10}$$

which is equal with

$$\left(1-\frac{\alpha_{ij}^m+\sum\limits_{v\neq m}e_{ij}^v}{S_i^m+\sum\limits_{v\neq m}e_{ij}^v}\right)^2+\sum_{k\neq j}\left(\frac{\alpha_{ik}^m}{S_i^m+\sum\limits_{v\neq m}e_{ij}^v}\right)^2. \tag{11}$$

Obviously, $L_{err}\left(\boldsymbol{\alpha}_i^{\diamondsuit(M)}\right)$ is smaller than $L_{err}\left(\boldsymbol{\alpha}_i^{m}\right)$ since

$$\left(1-\frac{\alpha_{ij}^m+\sum\limits_{v\neq m}e_{ij}^v}{S_i^m+\sum\limits_{v\neq m}e_{ij}^v}\right)^2<\left(1-\frac{\alpha_{ij}^m}{S_i^m}\right)^2 \tag{12}$$

and

$$\sum_{k\neq j}\left(\frac{\alpha_{ik}^m}{S_i^m+\sum\limits_{v\neq m}e_{ij}^v}\right)^2<\sum_{k\neq j}\left(\frac{\alpha_{ik}^m}{S_i^m}\right)^2. \tag{13}$$

Proposition 2. *By integrating the evidence from multi-view in terms of opinion aggregation with evidence accumulation, we guarantee the decrease of the overall uncertainty.*

Proof 2. Let $\boldsymbol{e}_i^m$ be the evidence captured from the m^{th} view classifier for the i^{th} sample. $\boldsymbol{e}_i^{\diamondsuit(M)}$ be the integrated evidence from the evidence accumulation of M views. After the opinion aggregation, the overall uncertainty $u^{\diamondsuit(M)}$ is updated as

$$u^{\diamondsuit(M)}=1-\sum_{j=1}^{K}\frac{e_{ij}^{\diamondsuit(M)}}{S_i^{\diamondsuit(M)}}=\frac{K}{S_i^{\diamondsuit(M)}}=\frac{K}{S_i^m+\sum\limits_{v\neq m}\sum_{j=1}^{K}e_{ij}^v} \tag{14}$$

which is smaller than the uncertainty of single view result $u^m=\frac{K}{S_i^m}$ since $S_i^m+\sum\limits_{v\neq m}\sum_{j=1}^{K}e_{ij}^v>S_i^m$.

These two propositions theoretically guarantee the prediction error and uncertainty of multi-view learning results will decrease with increasing views, which can produce accurate and trusted learning result. Our experimental results can also verify these propositions to validate the effectiveness of proposed method.

Consistency Regulation

We further extend the proposed method by adding a consistency regulation loss which minimizes the opinion entropy across opinions to guarantee the consistency of results between different views (step (3) in Figure. 1). The consistency loss is computed as

$$L_{con}\left(\boldsymbol{\alpha}_i\right)=\sum_{m=1}^{M}\left(\sum\nolimits_{v\neq m}^{M}E(\boldsymbol{\omega_i^m},\boldsymbol{\omega_i^v})\Big/(M-1)\right), \tag{15}$$

where $1/(M-1)$ is used for normalization and $E(\boldsymbol{\omega}_i^m,\boldsymbol{\omega}_i^v)$ is the opinion entropy described in previous subsection.

Experiments

In this section, we evaluate the proposed method on real-world multi-view datasets and compare it with existing multi-view learning methods. Furthermore, we also provide the uncertainty estimation analysis on noisy data.

Datasets

We conduct experiments on six real-world multi-view datasets as follows: **CUB** (Wah et al. 2011): Caltech-UCSD Birds dataset contains 11788 images and text descriptions from 200 categories of birds. **Food-101** (Wang et al. 2015b): UMPC Food-101 dataset consists of 86796 images and text descriptions from 101 classes of food. **HMDB** (Kuehne et al. 2011): This dataset is one of the largest human action recognition dataset, which consists of 6718 images of 51 categories of actions with two views. **Handwritten** (van Breukelen et al. 1998): This dataset consists of handwritten numerals ('0'-'9') from a collection of Dutch utility maps, the handwritten digits are represented with six views. **Caltech101** (Fei-Fei, Fergus, and Perona 2004): This dataset consists of 8677 images from 101 classes, which contains two views. **Scene15** (Fei-Fei and Perona 2005): Scene15 dataset contains 4485 images from 15 indoor and outdoor scene categories with three views. Details of each dataset are presented in the **Technical Appendix A**.

Compared Methods

We compare our method with several state-of-the-art multi-view learning methods as follows:

- **DCCA**: Deep Canonically Correlated Analysis (Andrew et al. 2013) obtains the correlations through deep neural networks, which maximizes the correlation among two views.
- **DCCAE**: Deep Canonically Correlated AutoEncoders (Wang et al. 2015a) employs autoencoders for seeking the common representation.

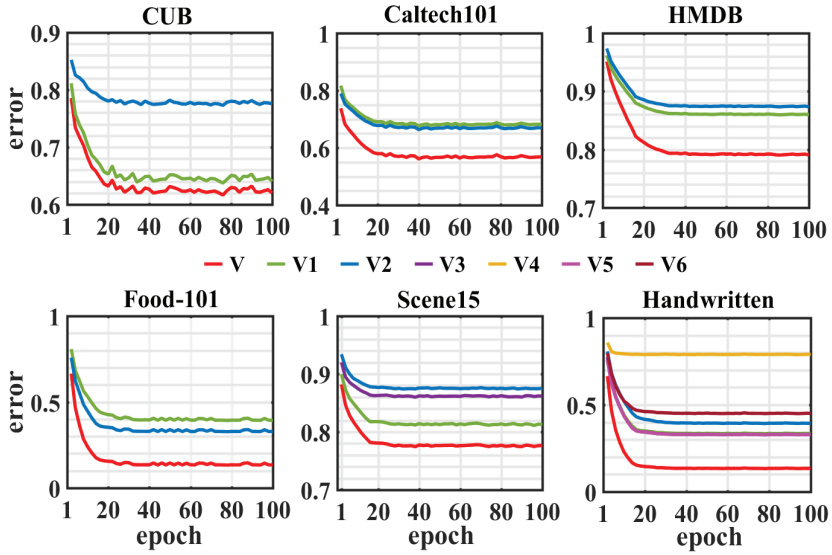


Figure 2: Prediction error with different epoches.

- **CPM-Nets**: Cross Partial Multi-view Networks (Zhang et al. 2020) focuses on learning a complete thus versatile representation to handling the complex correlation among different views.
- **DUA-Nets**: Dynamic Uncertainty-Aware Networks (Geng et al. 2021) employs Reversal networks to integrate intrinsic information from different views into a unified representation.
- **TMC**: Trusted Multi-view Classification (Han et al. 2021) focuses on the uncertainty estimation problem and produces a reliable classification result.

Implementation Details

For our algorithm, we conduct the fully connected networks for all datasets. The Adam optimizer (Kingma and Ba 2014) is used to train the network, where l_2-norm regularization is set to $1e^{-5}$. We then use 5-fold cross-validation to select the learning rate from $\{1e^{-4}, 3e^{-4}, 1e^{-3}, 3e^{-3}\}$. For all datasets, 20% samples are used as test set. We run 10 times for each method to report the mean values and standard deviations. The model is implemented by PyTorch on one NVIDIA TITAN Xp with GPU of 12GB memory.

Performance Evaluation

In this subsection, we conduct two tests to evaluate the performance of our method. The first test is to verify the effectiveness of our method and the second is to overall evaluate the superiority of our method.

Effectiveness evaluation. To validate the effectiveness of our multi-view learning method, we first compare the average prediction error for multi-view learning results (shown as red line, termed as V) with average prediction error for each single-view learning result (termed as V1-V6) on all datasets. The experimental results are shown in Figure 2, where the y-coordinate represents the average prediction error of data, the x-coordinate indicates the current epoch in training. On all datasets, the prediction error for multi-view (red line) are always smaller than each single-view in proposed method, which proves our method can efficiently reduce the prediction error after integration of multiple views to produce more accurate results. We also theoretically prove this conclusion in the Proposition 1. Furthermore, Figure 2 also demonstrates the convergence of proposed method. Typically, the optimization process is stable, where the loss decrease quickly and converges within a number of iterations.

Comparison with the methods. Then we overall evaluate our algorithm by comparing it with state-of-the-art multi-view learning methods in terms of accuracy metric. The detailed experimental results are shown in Table 1. We find that, on all datasets, our method consistently achieves better performance. Taking the results on HMDB as examples, our method improves the accuracy by about 20% compared to the second-best model (TMC) in terms of accuracy, which verifies the improved performance of the proposed method.

Uncertainty Estimation Analysis

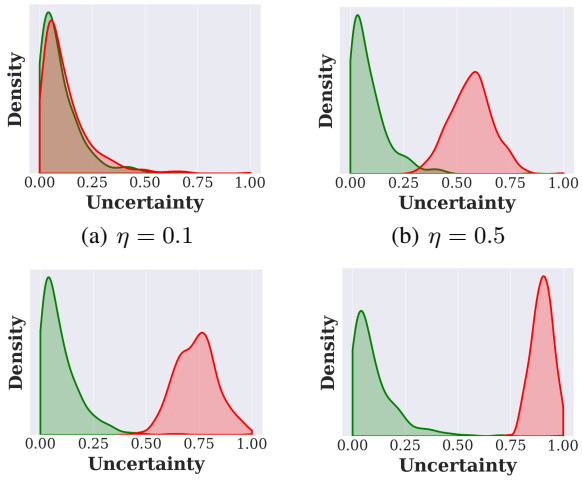
Due to the property differences and inconsistency of data sources, the uncertainty estimation becomes more important for the multi-view learning. Therefore, in this subsection, we conduct qualitative experiments to provide some insights for the estimated uncertainty, which can evaluate the uncertainty estimation performance of our method.

Ability of capturing uncertainty. Due to the limitation of pages, in this part, we just show the uncertainty estimation ability of our method on Caltech101 dataset with two views. We first add noise to half of the test samples in one view. Similarly to the work of (Geng et al. 2021), the noise vectors (denoted by ϵ) are sampled from Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Then we add these noise vectors multiplied

Table 1: Comparison with state-of-the-art multi-view learning methods based on accuracy (%).

Data	DCCA	DCCAE	CPM-NetS	DUA-Nets	TMC	Ours
CUB	82.12±3.03	85.39±1.28	89.32±0.38	81.13±1.67	91.00±2.36	**95.43±0.20**
HMDB	46.83±0.77	49.12±1.00	63.32±0.43	62.73±0.23	65.98±2.92	**88.20±0.58**
Scene15	54.77±1.13	55.03±0.34	67.29±1.01	68.23±0.11	67.79±0.21	**75.57±0.02**
Caltech101	89.00±0.15	90.11±0.21	90.35±2.12	93.83±0.34	92.93±0.20	**94.63±0.04**
Handwritten	97.55±0.38	97.25±0.42	94.55±1.36	98.10±0.32	98.51±0.13	**99.75±0.00**
Food-101	81.68±2.23	85.30±0.31	86.45±1.51	87.73±2.27	90.21±1.20	**93.75±0.32**

with intensity η to pollute half of the original test samples, i.e., the i^{th} sample $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \eta\boldsymbol{\epsilon}_i$. Then we obtain a Gaussian kernel density estimation (Scott 2015) of learned uncertainty shown in Figure 3. We find that the distribution curves of noisy samples (red curves) are nearly overlapped with the curves of clean samples (green curves) when the noise intensity is small ($\eta = 0.1$). Then uncertainty of noisy samples grows with increasing noise intensity. This means that the estimated uncertainty is associated with the sample quality, which verifies the uncertainty estimation ability of our method and further guarantees our method can obtain a trusted multi-view result with the decrease of the overall uncertainty after aggregation of multiple views.



(a) $\eta = 0.1$ (b) $\eta = 0.5$

(c) $\eta = 1.0$ (d) $\eta = 2.0$

Figure 3: Investigation of our model in capturing data noise. The curves in green and red correspond to distributions of clean and noisy data respectively.

Overall uncertainty evaluation. To evaluate the overall uncertainty, we add Gaussian noise with the fixed value of noise intensity ($\eta = 0.5$) to 50% of the test samples and compare the average uncertainty of multi-view learning results with the minimal average uncertainty among different single-view learning results on all datasets to verify the decrease of the overall uncertainty with the increasing views. The results are shown in Table 2, where the $\mathrm{U}_{\mathrm{multi}}$ indicates the average uncertainty degree for multi-view results on all datasets,$\mathrm{U}_{\mathrm{smin}}$ represents the minimal average uncertainty degree among different single-view results. The results indicate the uncertainty for multi-view results are always smaller than each single-view result in proposed method, which proves that our method can produce more reliable multi-view deep learning results. We also theoretically prove this conclusion in Proposition 2.

Table 2: Overall uncertainty evaluation.

Uncertainty	CUB	Caltech101	HMDB
$\mathrm{U}_{\mathrm{smin}}$	0.4896	0.5047	0.5995
$\mathrm{U}_{\mathrm{multi}}$	**0.2255**	**0.4038**	**0.4577**
Uncertainty	Scene15	Handwritten	Food-101
$\mathrm{U}_{\mathrm{smin}}$	0.4652	0.4337	0.6352
$\mathrm{U}_{\mathrm{multi}}$	**0.3433**	**0.2574**	**0.4378**

Moreover, we also conducted a thorough ablation study to justify the effectiveness of our major technical component, including fusion strategy and related model parameters. Additional comparisons with existing uncertainty-based methods (Gal and Ghahramani 2015; Lakshminarayanan, Pritzel, and Blundell 2017; Heo et al. 2018) and comparisons with different types of noise and the analysis of real-world applications are also performed. All of these experiments validate the effectiveness and superiority of our model. Detailed results are provided in **Technical Appendix B and C**.

Conclusion

In this work, we propose an efficient trusted multi-view deep learning method with opinion aggregation, which can generate trusted classification results on multi-view data. Our method tries to represent the learning results from different data sources as the opinions in evidence theory, which can precisely measure the uncertainty of learning results. By the opinion aggregation with evidence accumulation, our method can reduce the uncertainty of aggregated opinion to generate more reliable multi-view deep learning results. Furthermore, we further extend our method by adding a consistency regulation loss to guarantee the consistency of results between different views. The experimental results validate the effectiveness, reliability and robustness of the proposed multi-view deep learning method.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Serial Nos. 61976134, 61991410, 62173252) and Natural Science Foundation of Shanghai (NO. 21ZR1423900).

References

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, 1247–1255. PMLR.

Bach, F. R.; and Jordan, M. I. 2002. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul): 1–48.

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*.

Belluti, F.; Rampa, A.; Gobbi, S.; and Bisi, A. 2013. Small-molecule inhibitors/modulators of amyloid-β peptide aggregation and toxicity for the treatment of Alzheimer's disease: a patent review (2010–2012). *Expert opinion on therapeutic patents*, 23(5): 581–596.

Chao, G.; and Sun, S. 2016. Consensus and complementarity based maximum entropy discrimination for multi-view classification. *Information Sciences*, 367: 296–310.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, X.; and Liu, B. 2007. The utility of linguistic rules in opinion mining. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 811–812.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

Fei-Fei, L.; and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, 524–531. IEEE.

Gal, Y.; and Ghahramani, Z. 2015. Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*.

Geng, Y.; Han, Z.; Zhang, C.; and Hu, Q. 2021. Uncertainty-Aware Multi-View Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7545–7553.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2021. Trusted Multi-View Classification. In *International Conference on Learning Representations*.

Hardoon, D. R.; and Shawe-Taylor, J. 2011. Sparse canonical correlation analysis. *Machine Learning*, 83(3): 331–353.

Harold, H. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4): 321–377.

Heo, J.; Lee, H. B.; Kim, S.; Lee, J.; Kim, K. J.; Yang, E.; and Hwang, S. J. 2018. Uncertainty-aware attention for reliable interpretation and prediction. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 917–926.

Iso, H.; Wang, X.; Suhara, Y.; Angelidis, S.; and Tan, W.-C. 2021. Convex Aggregation for Opinion Summarization. *arXiv preprint arXiv:2104.01371*.

Jøsang, A. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*, 2556–2563. IEEE.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems*, 30.

Liu, H.; Liu, L.; Le, T. D.; Lee, I.; Sun, S.; and Li, J. 2017. Nonparametric sparse matrix decomposition for cross-view dimensionality reduction. *IEEE Transactions on Multimedia*, 19(8): 1848–1859.

Liu, J.; Cao, Y.; Lin, C.-Y.; Huang, Y.; and Zhou, M. 2007. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 334–342.

Liu, M.; Luo, Y.; Tao, D.; Xu, C.; and Wen, Y. 2015. Low-rank multi-view learning in matrix completion for multi-label image classification. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Martini, C.; and Sprenger, J. 2017. Opinion aggregation and individual expertise.

Scott, D. W. 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 3183–3193.

Şensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*.

Sun, S.; Dong, W.; and Liu, Q. 2020. Multi-view representation learning with deep gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, S.; Liu, Y.; and Mao, L. 2019. Multi-view learning for visual violence recognition with maximum entropy discrimination and deep features. *Information Fusion*, 50: 43–53.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tao, Z.; Liu, H.; Li, J.; Wang, Z.; and Fu, Y. 2019. Adversarial graph embedding for ensemble clustering. In *International Joint Conferences on Artificial Intelligence Organization*.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

van Breukelen, M.; Duin, R. P.; Tax, D. M.; and Den Hartog, J. 1998. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4): 381–386.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, C. 2007. Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18(3): 905–910.

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2018. Partial multi-view clustering via consistent GAN. In *2018 IEEE International Conference on Data Mining (ICDM)*, 1290–1295. IEEE.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015a. On deep multi-view representation learning. In *International conference on machine learning*, 1083–1092. PMLR.

Wang, X.; Kumar, D.; Thome, N.; Cord, M.; and Precioso, F. 2015b. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6. IEEE.

Xie, Y.; Liu, J.; Qu, Y.; Tao, D.; Zhang, W.; Dai, L.; and Ma, L. 2020. Robust kernelized multiview self-representation for subspace clustering. *IEEE transactions on neural networks and learning systems*.

Xie, Y.; Tao, D.; Zhang, W.; Liu, Y.; Zhang, L.; and Qu, Y. 2018. On unifying multi-view self-representations for clustering by tensor multi-rank minimization. *International Journal of Computer Vision*, 126(11): 1157–1179.

Zadeh, L. 1984. Review of Shafer's: A Mathematical Theory of Evidence. *AI Magazine*, 5(3): 81–83.

Zadeh, L. A. 1986. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *AI magazine*, 7(2): 85–85.

Zhang, C.; Cui, Y.; Han, Z.; Zhou, J. T.; Fu, H.; and Hu, Q. 2020. Deep Partial Multi-View Learning. *IEEE transactions on pattern analysis and machine intelligence*.

Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; and Xu, D. 2018a. Generalized latent multi-view subspace clustering. *IEEE transactions on pattern analysis and machine intelligence*, 42(1): 86–99.

Zhang, C.; Fu, H.; Hu, Q.; Zhu, P.; and Cao, X. 2016. Flexible multi-view dimensionality co-reduction. *IEEE Transactions on Image Processing*, 26(2): 648–659.

Zhang, C.; Han, Z.; cui, y.; Fu, H.; Zhou, J. T.; and Hu, Q. 2019. CPM-Nets: Cross Partial Multi-View Networks. In *Advances in Neural Information Processing Systems*, 559–569.

Zhang, Z.; Liu, L.; Shen, F.; Shen, H. T.; and Shao, L. 2018b. Binary multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 41(7): 1774–1782.

Zhao, H.; Ding, Z.; and Fu, Y. 2017. Multi-view clustering via deep matrix factorization. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Technical Appendix

A. Details of the Datasets

We conduct experiments on the following six real-world datasets to evaluate our method. Details of these datasets are as follows:

- **CUB** consists of 11788 bird images associated with text descriptions from 200 different categories of birds. We use GoodleNet and doc2vec to extract image features and corresponding text features as multiple views.
- **Food-101** consists of 86976 food images with text descriptions from 101 categories, which is a noisy multimodal dataset. We use Resnet152 and Bert to extract the image features and text features as multiple views.
- **HMDB** is one of the largest human action recognition dataset. There are 6718 samples with 51 categories of actions, where HOG and MBH features are extracted as multiple views.
- **Handwritten** consists of 2000 samples of 10 classes from digit '0' to '9' with 200 samples per class, where six different types of descriptors are used as multiple views.
- **Caltech101** consists of 8677 images from 101 categories. The first 10 categories are used, where two types of deep features with DECAF and VGG19 are extracted as multiple views respectively.
- **Scene15** consists of 4485 images from 15 indoor and outdoor scene categories. Three types features including GIST, PHOG and LBP are extracted as multiple views.

B. Ablation Study

We conduct a detailed ablation study to clearly demonstrate the effectiveness of our major technical components. We adopt the HMDB data with two views for illustration purpose.

Fusion strategy. In this part, we validate the effectiveness of proposed fusion strategy. Figure 4 compares the results of proposed fusion strategy without consistency regulation for multi-view (orange bar, termed as V) with the results from each single-view (termed as V1, V2). The experimental results confirm the effectiveness of our fusion strategy. Even if we remove the consistency regulation term, our accurate scores are also better than the scores from each single-view.

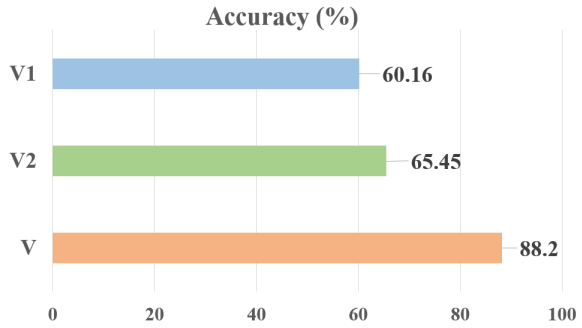


Figure 4: Performance evaluation of fusion strategy.

Balance of consistency regulation. Figure 5 shows the results using different fixed consistency regulation ratios (λ in the equation 7) on original HMDB dataset, where $\lambda = 0$ means we just use the fusion strategy and do not consider the conflict among multiple views. We can find the consistency regulation has effect on the data when $\lambda = 0.1$ or $\lambda = 0.3$, but has slight negative effect with high values of λ. We guess there is some noise in one of views, therefore, emphasizing consistency too much during the training process may cause model performance degradation. Therefore, to consider the unknown of the multi-view dataset structure in real-world, we typically set the initial $\lambda = 0$, then λ slowly increases from 0 to 1.

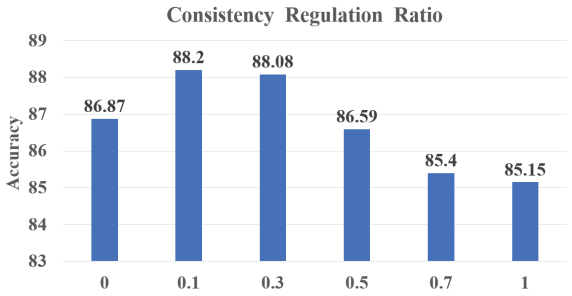


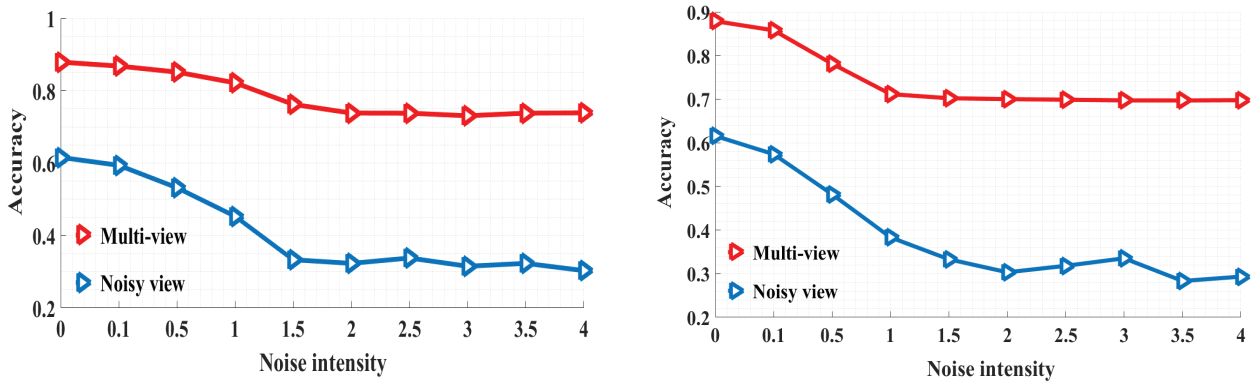
Figure 5: Accuracy with different consistency regulation ratios.

C. Additional Comparison Results

In this part, we show four additional comparison results.

Impact analysis of noise type. We conduct an experiment on HMDB dataset to analyze the impact of different noise types. Besides the Gaussian noise, similarly to the work (Zhang et al. 2018a), we consider two types of noise, i.e., the global noise $\boldsymbol{E}_g$ and the sample-specific noise $\boldsymbol{E}_s$. Formally, we have the noise $\boldsymbol{E} = \boldsymbol{E}_s + \alpha \boldsymbol{E}_g$. For the sample-specific noise, $\boldsymbol{E}_s$, we randomly generate a vector and then select randomly a few columns (50 in experiments), setting the other columns with zeros. While for the global noise, we randomly generate a vector and multiply it with a coefficient α to control the noise magnitude. Then, for the i^{th} sample, we have the $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \eta \boldsymbol{E}_i$. The same as the previous experiments, we add the noise to half of the test samples in one view. Figure 6 shows the performance with different uncertainty degree of two types of noise, respectively. We can find, on all types of noise, the performance of both multi-view and single-view decreases with increasing of the noise intensity (η for gaussian noise $\boldsymbol{\epsilon}$, α for new noise $\boldsymbol{E}$). The performance of single-view on noisy view data decrease rapidly. However, the performance of our method is quite stable, which means our method is more robust to all types of noise and can alleviate the influence of noisy samples.

Comparison with uncertainty-based deep learning method. To further evaluate the performance of our method, we compare proposed method with several existing



(a) η (noise intensity) (b) α (noise intensity)

Figure 6: Investigation of our model in capturing deferent types of data noise. The curves in red and blue correspond to accuracy of multi-view and noisy view respectively.

Table 3: Comparison with uncertainty-based deep learning methods based on accuracy $(\%)$.

Data	MCDO	DE	UA+	Ours
CUB	89.83±0.43	90.39±0.28	87.51±1.23	**95.43±0.20**
HMDB	65.12±2.21	69.35±1.01	67.52±1.31	**88.20±0.58**
Scene15	66.13±1.54	67.12±0.21	64.98±2.61	**75.57±0.02**
Caltech101	90.04±0.12	93.11±0.10	88.96±1.24	**94.63±0.04**
Handwritten	97.73±0.26	98.75±0.05	97.55±0.32	**99.75±0.00**
Food-101	81.56±1.33	87.30±0.43	89.70±0.21	**93.75±0.32**

uncertainty-based deep learning method, including MCDO (Monte Carlo Dropout) (Gal and Ghahramani 2015), DE (Deep Ensemble) (Lakshminarayanan, Pritzel, and Blundell 2017), UA+ (uncertainty-aware attention) (Heo et al. 2018). Due to some uncertainty-based methods use single-view data, we concatenate the original features of multiple views for all comparison methods. The results shown in Table 3 indicate our method outperforms other methods on all datasets. Taking the results on HMDB, Scene15 as examples, our method improves the accuracy by about 20% and 8% compared to the second-best model (DE) in terms of accuracy, which verifies the superiority of the proposed method.

Comparison with uncertainty-based deep learning method in an end-to-end test. The previous experiments are based on the multiple types of features, to further evaluate the performance of our method on the original multi-modal dataset, we conduct an end-to-end test on the original multi-modal dataset Food-101, which consists of images and text descriptions. In this test, we use Inception-V3 network (Szegedy et al. 2016) pre-trained with ImageNet and Bert (Devlin et al. 2018) as the backbones for image and text respectively. Our method can jointly optimize these two networks according to our loss function with opinion aggregation. For the comparison methods, we concatenate the outputs of these two backbone networks as the input of the classifier. For all algorithms, The Adam optimizer (Kingma and Ba 2014) is used to train the network, where l_2-norm regularization is set to $1e^{-5}$. The learning rate is set to $1e^{-4}$. Moreover, 20% samples are used as test set. We run 10 times for each method to report the mean values and standard deviations. The model is implemented by PyTorch on two NVIDIA TITAN Xp with GPU of 12GB memory. The results shown in Table 4 indicate our method also has the good performance in the end-to-end manner.

Real-world application analysis. In our experimental dataset, Food-101 is a noisy large-scale multi-modal dataset. It contains many noise samples, i.e., wrong or missing text descriptions, mismatched images, wrong class labels, or the too many objects in one image. For these noisy samples, our method assigns high uncertainty to them, which means our method is more suitable for the real-world application. The examples with high uncertainty are shown in Figure 7. We can find, Figure (1) has a missing text description described by error code. Figure (2) is a mismatched image which should be put into the "apple juic" class not the "apple pie". Figure (3) has a wrong class label which should be the "dinner set" not the "baby back rips". Figure (4) has too many objects. Therefore, it is reasonable that these images are assigned to high uncertainty values by our model.

D. Source Code

The code for this work can be found in the Supplementary Material source_code.zip file.

Table 4: Comparison with uncertainty-based deep learning methods in an end-to-end test based on accuracy (%).

Data	MCDO	DE	UA+	Ours
Food-101	78.56±2.11	86.15±0.23	87.64±1.02	**92.25±0.75**

(1)
Uncertainty: 0.93

Label: Hamburger

Text: Ÿì¼¶+o}ËK|Úi],A¹.^ZÍ`?ÞÔ?oåånp'¯ÚÁ £g¢»G †¬È"ósGv,À#f,,Žå¸ôßYâ'~ ÆYžý±ç0åâa×tÕ0vDž{f;^xžî?ŒyÏ³|5ŒïÒ0þ™-C7ÈlËw÷fïll§ÆwBÿ, 'Ä¶\Ë{³n6n gcÙ!íÎ™o™ íXají=ÌZð4 ø®ãÒlüÈŽ³ÐrnzSÍÆ}ßŠ¼³ÔãØŽÜ(Ø{çò0vhö³1=×?"BZ~hfbÿ½Y;O £€e[AäoeQÿv™

(2)
Uncertainty: 0.71

Label: apple_pie

Text: So why did I decide to call this an apple pie smoothie? Well for one, I thought it sounded pretty good. And secondly, I pretty much use all the ingredients from an apple pie in this delectable drink. We've got apples, spices, sweetness, and chia seeds. Wait, _ what _ ? Chia seeds aren't in apple pie! Yeah, I know.

(3)
Uncertainty: 0.84

Label: baby_back_rips

Text: Susan Spungen's cookbook, Recipes, focuses on freshness, simplicity and technique. Here, she slow-bakes succulent baby back ribs and serves them with a coffee-tinged sauce. This recipe is great for both oven or outdoor grill.

(4)
Uncertainty: 0.54

Label: peaking_duck

Text: People will argue forever about the best Peking duck in Beijing, but I consistently read fantastic reviews about a restaurant called Made In China. I'll post a more detailed review later, but look how happy Bryan looks eating his Peking duck!

Figure 7: Examples with four high prediction uncertainty on Food-101.