TRUSTED MULTI-VIEW CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view classification (MVC) generally focuses on improving classification accuracy by using information from different views, typically integrating them into a unified comprehensive representation for downstream tasks. However, it is also crucial to dynamically assess the quality of a view for different samples in order to provide reliable uncertainty estimations, which indicate whether predictions can be trusted. To this end, we propose a novel multi-view classification method, termed trusted multi-view classification, which provides a new paradigm for multi-view learning by dynamically integrating different views at an evidence level. The algorithm jointly utilizes multiple views to promote both classification reliability and robustness by integrating evidence from each view. To achieve this, the Dirichlet distribution is used to model the distribution of the class probabilities, parameterized with evidence from different views and integrated with the Dempster-Shafer theory. The unified learning framework induces accurate uncertainty and accordingly endows the model with both reliability and robustness for out-of-distribution samples. Extensive experimental results validate the effectiveness of the proposed model in accuracy, reliability and robustness.

1 INTRODUCTION

Multi-view data, typically associated with multiple modalities or multiple types of features, often exists in real-world scenarios. State-of-the-art multi-view learning methods achieve tremendous success across a wide range of real-world applications. However, this success typically relies on complex models (Wang et al., 2015a; Tian et al., 2019; Bachman et al., 2019; Zhang et al., 2019; Hassani & Khasahmadi, 2020), which tend to integrate multi-view information with deep neural networks. Although these models can provide accurate classification results, they are usually vulnerable to yielding unreliable predictions, particularly when presented with views that are not well-represented (*e.g.*, information from abnormal sensors). Consequently, their deployment in safety-critical applications (*e.g.*, computer-aided diagnosis or autonomous driving) is limited. This has inspired us to introduce a new paradigm for multi-view classification to produce trusted decisions.

For multi-view learning, traditional algorithms generally assume an equal value for different views or assign/learn a fixed weight for each view. The underlying assumption is that the qualities or importance of these views are basically stable for all samples. In practice, the quality of a view often varies for different samples which the designed models should be aware of for adaption. For example, in multi-modal medical diagnosis (Perrin et al., 2009; Sui et al., 2018), a magnetic resonance (MR) image may be sufficient for one subject, while a positron emission tomography (PET) image may be required for another. Therefore, the decision should be well explained according to multi-view inputs. Typically, we not only need to know the classification result, but should also be able to answer "How confident is the decision?" and "Why is the confidence so high/low for the decision?". To this end, the model should provide in accurate uncertainty for the prediction of each sample, and even individual view of each sample.

Uncertainty-based algorithms can be roughly divided into two main categories, *i.e.*, Bayesian and non-Bayesian approaches. Traditional Bayesian approaches estimate uncertainty by inferring a posterior distribution over the parameters (MacKay, 1992a; Bernardo & Smith, 2009; Neal, 2012). A variety of Bayesian methods have been developed, including Laplace approximation (MacKay, 1992b), Markov Chain Monte Carlo (MCMC) (Neal, 2012) and variational techniques (Graves, 2011; Ranganath et al., 2014; Blundell et al., 2015). However, compared with ordinary neural networks, due to the doubling of model parameters and difficulty in convergence, these methods are computationally

expensive. Recent algorithm (Gal & Ghahramani, 2016) estimates the uncertainty by introducing dropout (Srivastava et al., 2014) in the testing phase, thereby reducing the computational cost. Several non-Bayesian algorithms have been proposed, including deep ensemble (Lakshminarayanan et al., 2017), evidential deep learning (Sensoy et al., 2018) and deterministic uncertainty estimate (van Amersfoort et al., 2020). Unfortunately, all of these methods focus on estimating the uncertainty on single-view data, despite the fact that fusing multiple views through uncertainty can improve performance and reliability.

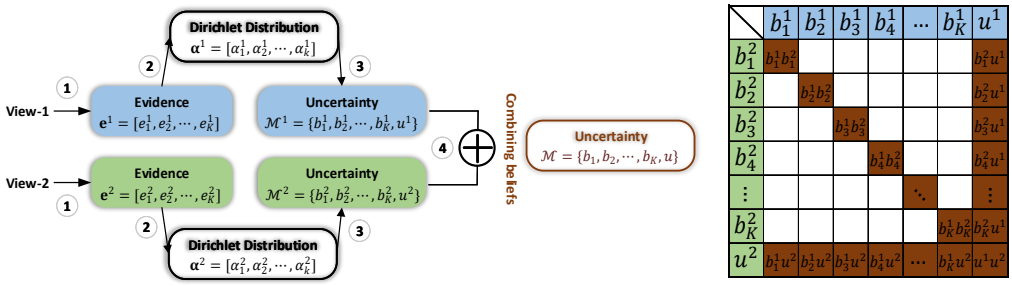
In this paper, we propose a new multi-view classification algorithm aiming to elegantly integrate multi-view information for trusted decision making (shown in Fig. 1(a)). Our model combines different views at an evidence level instead of feature or output level as done previously, which produces a stable and reasonable uncertainty estimation and thus promotes both classification reliability and robustness The Dirichlet distribution is used to model the distribution of the class probabilities, parameterized with evidence from different views and integrated with the Dempster-Shafer theory. In summary, the specific contributions of this paper are:

(1) We propose a novel multi-view classification model aiming to provide trusted and interpretable (according to the uncertainty of each view) decisions in an effective and efficient way (without any additional computations and neural network changes), which introduces a new paradigm in multi-view classification.
(2) The proposed model is a unified framework for promising sample-adaptive multi-view integration, which integrates multi-view information at an evidence level with the Dempster-Shafer theory in an optimizable (learnable) way.
(3) The uncertainty of each view is accurately estimated, enabling our model to improve classification reliability and robustness.
(4) We conduct extensive experiments which validate the superior accuracy, robustness, and reliability of our model thanks to the promising uncertainty estimation and multi-view integration strategy.

2 Related Work

Uncertainty-based Learning. Deep neural networks have achieved great success in various tasks. However since most deep models are essentially deterministic functions, the uncertainty of the model cannot be obtained. Bayesian neural networks (BNNs) (Denker & LeCun, 1991; MacKay, 1992b; Neal, 2012) endow deep models with uncertainty by replacing the deterministic weight parameters with distributions. Since BNNs struggle in performing inference and usually come with prohibitive computational costs, a more scalable and practical approach, MC-dropout (Gal & Ghahramani, 2016), was proposed. In this model, the inference is completed by performing dropout sampling from the weight during training and testing. Ensemble based methods (Lakshminarayanan et al., 2017) train and integrate multiple deep networks and also achieve promising performance. Instead of indirectly modeling uncertainty through network weights, the algorithm (Sensoy et al., 2018) introduces the subjective logic theory to directly model uncertainty without ensemble or Monte Carlo sampling. Building upon RBF networks, the distance between test samples and prototypes can be used as the agency for deterministic uncertainty (van Amersfoort et al., 2020). Benefiting from the learned weights of different tasks with homoscedastic uncertainty learning, (Kendall et al., 2018) achieves impressive performance in multi-task learning.

Multi-View Learning. Learning on data with multiple views has proven effective in a variety of tasks. CCA-based multi-view models (Hotelling, 1992; Akaho, 2006; Wang, 2007; Andrew et al., 2013; Wang et al., 2015a; 2016) are representative ones that have been widely used in multi-view representation learning. These models essentially seek a common representation by maximizing the correlation between different views. Considering common and exclusive information, hierarchical multi-modal metric learning (HM3L) (Zhang et al., 2017) explicitly learns shared multi-view and view-specific metrics, while AE^2-Nets (Zhang et al., 2019) implicitly learn a complete (view-specific and shared multi-view) representation for classification. Recently, the methods (Tian et al., 2019; Bachman et al., 2019; Chen et al., 2020; Hassani & Khasahmadi, 2020) based on contrastive learning have also achieved good performance. Due to its effectiveness, multi-view learning has been widely used in various applications (Kiela et al., 2018; Bian et al., 2017; Kiela et al., 2019; Wang et al., 2020).



(a) Overview of the trusted multi-view classification (b) Combining beliefs

Figure 1: Illustration of our algorithm. (a) The evidence of each view is obtained using neural networks (①). The obtained evidence parameterizes the Dirichlet distribution (②) to induce the classification probability and uncertainty (③). The overall uncertainty and classification probability are inferred by combining the beliefs of multiple views based on the DST (④). The combination rule and an example are shown in Definition 4 and (b), respectively. Given two sets of beliefs (blue and green blocks), we recombine the compatible parts of the two sets (brown blocks) and ignore the mutually exclusive parts (white blocks) of the two sets to obtain the combined beliefs.

Dempster-Shafer Evidence Theory (DST). DST, which is a theory on belief functions, was first proposed by Dempster (Dempster, 1967) and is a generalization of the Bayesian theory to subjective probabilities (Dempster, 1968). Later, it was developed into a general framework to model epistemic uncertainty (Shafer, 1976). In contrast to Bayesian neural networks, which indirectly obtain uncertainty through multiple stochastic samplings from weight parameters, DST directly models uncertainty. DST allows beliefs from different sources to be combined with various fusion operators to obtain a new belief that considers all available evidence (Sentz et al., 2002; Jøsang & Hankin, 2012). When faced with beliefs from different sources, Dempster's rule of combination tries to fuse their shared parts, and ignores conflicting beliefs through normalization factors. A more specific implementation will be discussed later.

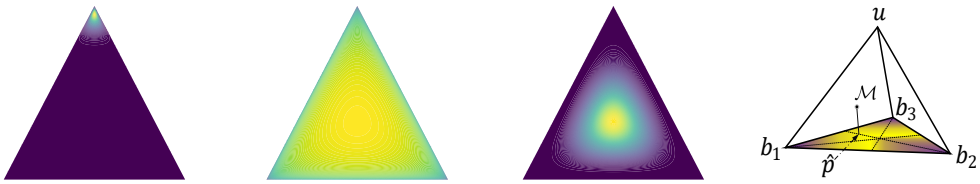
3 Trusted Multi-View Classification

It has been shown that using a softmax output as confidence for predictions often leads to high confidence values, even for erroneous predictions since the largest softmax output is used for the final prediction (Moon et al., 2020; van Amersfoort et al., 2020). Therefore, we introduce an evidence-based uncertainty estimation technique which can provide more accurate uncertainty and allow us to flexibly integrate multiple views for trusted decision making.

3.1 Uncertainty and the Theory of Evidence

In this subsection, we elaborate on evidential deep learning to quantify the classification uncertainty for each of multiple views, which simultaneously models the probability of each class and overall uncertainty of the current prediction. In the context of multi-class classification, Subjective logic (SL) (Jøsang, 2018) associates the parameters of the Dirichlet distribution (Definition A.1 in the Appendix) with the belief distribution, where the Dirichlet distribution can be considered as the conjugate prior of the categorical distribution (Bishop, 2006).

Accordingly, we need to determine the concentration parameters, which are closely related to the uncertainty. We elaborate on the Subjective logic (Jøsang, 2018), which defines a theoretical framework for obtaining the probabilities (belief masses) of different classes and overall uncertainty (uncertainty mass) of the multi-classification problem based on the *evidence* collected from data. Note that *evidence* refers to the metrics collected from the input to support the classification (step ① in Fig. 1(a)) and is closely related to the concentration parameters of Dirichlet distribution. Specifically, for the K classification problems, subjective logic tries to assign a belief mass to each class label and an overall uncertainty mass to the whole frame based on the *evidence*. Accordingly, for the vth view,



(a) Confident prediction (b) Out of distribution (c) High uncertainty (d) Opinion ($\mathcal{M}$)

Figure 2: Typical examples of Dirichlet distribution. Refer to the text for details.

the $K+1$ mass values are all non-negative and their sum is one:

$$u^v + \sum_{k=1}^{K} b_k^v = 1, \tag{1}$$

where $u^v \geq 0$ and $b_k^v \geq 0$ indicate the overall uncertainty and the probability for the kth class, respectively.

For the vth view, subjective logic connects the *evidence* $\mathbf{e}^v = [e_1^v, \cdots, e_K^v]$ to the parameters of the Dirichlet distribution $\boldsymbol{\alpha}^v = [\alpha_1^v, \cdots, \alpha_K^v]$ (step ② in Fig. 1(a)). Specifically, the parameter α_k^v of the Dirichlet distribution is induced from e_k^v, *i.e.*, $\alpha_k^v = e_k^v + 1$. Then, the belief mass b_k^v and the uncertainty u^v (step ③ in Fig. 1(a)) are computed as

$$b_k^v = \frac{e_k^v}{S^v} = \frac{\alpha_k^v - 1}{S^v} \quad \text{and} \quad u^v = \frac{K}{S^v}, \tag{2}$$

where $S^v = \sum_{i=1}^{K} (e_i^v + 1) = \sum_{i=1}^{K} \alpha_i^v$ is the Dirichlet strength. Eq. 2 actually describes the phenomenon where the more evidence observed for the kth category, the greater the probability assigned to the kth class. Correspondingly, the less total evidence observed, the greater the total uncertainty. The belief assignment can be considered as a subjective opinion. Given an opinion, the mean of the corresponding Dirichlet distribution $\hat{\mathbf{p}}^v$ for the class probability $\hat{p}_k^v$ is computed as $\hat{p}_k^v = \frac{\alpha_k^v}{S^v}$ (Frigyik et al., 2010).

Differences from traditional deep-neural-network classifiers. Firstly, the output of traditional neural network classifiers can be considered as a point on a simplex, while Dirichlet distribution parametrizes the density of each such probability assignment on a simplex. Therefore, with the Dirichlet distribution, SL models the second-order probability and uncertainty of the output. Secondly, the softmax function is widely used in the last layer of traditional neural network classifiers. However, using the softmax output as the confidence often leads to over-confidence. In our model, the introduced SL can avoid this problem by adding overall uncertainty mass. Existing methods (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017) usually require additional computations during inference to output uncertainty. Since the uncertainty is obtained during the inference stage, it is difficult to seamlessly train a model with high accuracy, robustness and reasonable uncertainty in a unified framework. Accordingly, the limitations underlying existing algorithms (*e.g.*, inability to directly obtain uncertainty) also limits their extension to trusted multi-view classification.

For clarity, we provide typical examples under a triple classification task to illustrate the above formulation. Let us assume that $\mathbf{e} = \langle 40, 1, 1 \rangle$ and accordingly we have $\boldsymbol{\alpha} = \langle 41, 2, 2 \rangle$. The corresponding Dirichlet distribution, shown in Fig. 2(a), yields a sharp distribution centered on the top of the standard 2-simplex. This indicates that sufficient evidence has been observed to ensure accurate classification. In contrast, let us assume that we have the evidence $\mathbf{e} = \langle 0.0001, 0.0001, 0.0001 \rangle$, which is little evidence for classification. Accordingly, we obtain the Dirichlet distribution parameter $\boldsymbol{\alpha} = \langle 1.0001, 1.0001, 1.0001 \rangle$ and the uncertainty mass $u \approx 1$. As shown in Fig. 2(b), in this case, the evidence induces quite a flat distribution over the simplex. Finally, when $\mathbf{e} = \langle 5, 5, 5 \rangle$, there is also a high uncertainty, as shown in Fig. 2(c), even though the overall uncertainty is reduced compared to the second case. As shown in Fig. 2(d), we can convert a Dirichlet distribution into a standard 3-simplex (a regular tetrahedron with vertices (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1) in $\mathbf{R}^4$) based on the subjective logic theory (Eq. 1 and Eq. 2), where the point ($\mathcal{M}$) in the simplex corresponding to $\left\{\{b_k\}_{k=1}^{3}, u\right\}$ indicates an opinion. Accordingly, the expectation value $\hat{\mathbf{p}}$ of the Dirichlet distribution is the projection of $\mathcal{M}$ on the bottom.

3.2 DEMPSTER'S RULE OF COMBINATION FOR MULTI-VIEW CLASSIFICATION

Having introduced evidence and uncertainty for the single-view case, we now focus on their adaptation to data with multiple views. The Dempster–Shafer theory of evidence allows evidence from different sources to be combined arriving at a degree of belief (represented by a mathematical object called the belief function) that takes into account all the available evidence (see Definition 3.1). Specifically, we need to combine V independent sets of probability mass assignments $\{\mathcal{M}^v\}_1^V$, where $\mathcal{M}^v = \left\{\{b_k^v\}_{k=1}^K, u^v\right\}$, to obtain a joint mass $\mathcal{M} = \left\{\{b_k\}_{k=1}^K, u\right\}$ (step ④ in Fig. 1(a)).

Definition 3.1 ***(Dempster's combination rule for two independent sets of masses)*** *The combination (called the joint mass) $\mathcal{M} = \left\{\{b_k\}_{k=1}^K, u\right\}$ is calculated from the two sets of masses $\mathcal{M}^1 = \left\{\{b_k^1\}_{k=1}^K, u^1\right\}$ and $\mathcal{M}^2 = \left\{\{b_k^2\}_{k=1}^K, u^2\right\}$ in the following manner:*

$$\mathcal{M} = \mathcal{M}^1 \oplus \mathcal{M}^2. \tag{3}$$

The more specific calculation rule can be formulated as follows:

$$b_k = \frac{1}{1-C}(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1), u = \frac{1}{1-C} u^1 u^2, \tag{4}$$

where $C = \sum_{i \neq j} b_i^1 b_j^2$ is a measure of the amount of conflict between the two mass sets (the white blocks in Fig. 1(b)), and the scale factor $\frac{1}{1-C}$ is used for normalization.

The joint opinion $\mathcal{M}$ is formed based on the fusion of opinions $\mathcal{M}^1$ and $\mathcal{M}^2$. The joint belief mass of class k (b_k) and overall uncertainty (u) correspond to the brown blocks in Fig. 1(b). Intuitively, the combination rule ensures: (1) when both views are of high uncertainty (large u^1 and u^2), the final prediction must be of low confidence (small b_k); (2) when both views are of low uncertainty (small u^1 and u^2), the final prediction may be of high confidence (large b_k); (3) when only one view is of low uncertainty (only u^1 or u^2 is large), the final prediction only depends on the confident view.

Then, given data with V different views, we can obtain the above-mentioned mass for each view. Afterwards, we can combine the beliefs from different views with Dempster's rule of combination. Specifically, we fuse the belief mass and uncertainty mass between different views with the following rule:

$$\mathcal{M} = \mathcal{M}^1 \oplus \mathcal{M}^2 \oplus \cdots \mathcal{M}^V. \tag{5}$$

After obtaining the joint mass $\mathcal{M} = \left\{\{b_k\}_{k=1}^K, u\right\}$, according to Eq. 2, the corresponding joint evidence from multiple views and the parameters of the Dirichlet distribution are induced as

$$S = \frac{K}{u}, e_k = b_k \times S \text{ and } \alpha_k = e_k + 1. \tag{6}$$

Based on the above combination rule, we can obtain the estimated multi-view joint evidence $\mathbf{e}$ and the corresponding parameters of joint Dirichlet distribution $\boldsymbol{\alpha}$ to produce the final probability of each class and the overall uncertainty.

Advantages of using subjective logic compared with softmax. Compared with softmax output, using subjective uncertainty is more suitable for the fusion of multiple decisions. Subjective logic provides an additional mass function (u) that allows the model distinguish between a lack of evidence. In our model, subjective logic provides the degree of overall uncertainty of each view, which is important for trusted classification and interepretability to some extent.

3.3 LEARNING TO FORM OPINIONS

In this section, we will discuss how to train neural networks to obtain evidence for each view, which can then be used to obtain the corresponding masses $\{\mathcal{M}^v\}_{v=1}^V$ and $\mathcal{M}$. The neural networks can capture the evidence from input to induce a classification opinion (Kiela et al., 2018), and the conventional neural-network-based classifier can be naturally transformed into the evidence-based classifier with minor changes. Specifically, the softmax layer of a conventional neural-network-based classifier is replaced with an activation function layer (*i.e.*, RELU) to ensure that the network outputs non-negative values, which are considered as the evidence vector $\boldsymbol{e}$. Accordingly, the parameters of the Dirichlet distribution can be obtained.

For conventional neural-network-based classifiers, the cross-entropy loss is usually employed:

$$\mathcal{L}_{ce} = -\sum_{j=1}^{K} y_{ij} \log(p_{ij}), \tag{7}$$

where p_{ij} is the predicted probability of the ith sample for class j. For our model, given the evidence of the ith sample obtained through the evidence network, we can get the parameter $\boldsymbol{\alpha}_i$ (*i.e.*, $\boldsymbol{\alpha}_i^v = \mathbf{e}_i^i + 1$) of the Dirichlet distribution and form the multinomial opinions $D(\mathbf{p}_i|\boldsymbol{\alpha}_i)$, where $\mathbf{p}_i$ is the class assignment probabilities on a simplex. After a simple modification on Eq. 7, we have the adjusted cross-entropy loss:

$$\mathcal{L}_{ace}(\boldsymbol{\alpha}_i) = \int \left[\sum_{j=1}^{K} -y_{ij} \log\left(p_{ij}\right) \right] \frac{1}{B\left(\boldsymbol{\alpha}_i\right)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\mathbf{p}_i = \sum_{j=1}^{K} y_{ij} \left(\psi\left(S_i\right) - \psi\left(\alpha_{ij}\right)\right), \tag{8}$$

where $\psi(\cdot)$ is the digamma function. Eq. 8 is the integral of the cross-entropy loss function on the simplex determined by $\boldsymbol{\alpha}_i$. The above loss function ensures that the correct label of each sample generates more evidence than other classes, however, it cannot guarantee that less evidence will be generated for incorrect labels. That is to say, in our model, we expect the evidence for incorrect labels to shrink to 0. To this end, the following KL divergence term is introduced:

$$\begin{aligned} & KL\left[D\left(\mathbf{p}_i|\tilde{\boldsymbol{\alpha}}_i\right) \| D\left(\mathbf{p}_i|\mathbf{1}\right)\right] \\ & = \log\left(\frac{\Gamma\left(\sum_{k=1}^{K} \tilde{\alpha}_{ik}\right)}{\Gamma(K) \prod_{k=1}^{K} \Gamma(\tilde{\alpha}_{ik})}\right) + \sum_{k=1}^{K} \left(\tilde{\alpha}_{ik} - 1\right) \left[\psi\left(\tilde{\alpha}_{ik}\right) - \psi\left(\sum_{j=1}^{K} \tilde{\alpha}_{ij}\right)\right], \end{aligned} \tag{9}$$

where $\tilde{\boldsymbol{\alpha}}_i = \mathbf{y}_i + (1 - \mathbf{y}_i) \odot \boldsymbol{\alpha}_i$ is the adjusted parameter of the Dirichlet distribution which can avoid penalizing the evidence of the groundtruth class to 0, and $\Gamma(\cdot)$ is the *gamma* function.

Therefore, given parameter $\boldsymbol{\alpha}_i$ of the Dirichlet distribution for each sample i, the sample-specific loss is

$$\mathcal{L}(\boldsymbol{\alpha}_i) = \mathcal{L}_{ace}(\boldsymbol{\alpha}_i) + \lambda_t KL\left[D\left(\mathbf{p}_i|\tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}}\right) \| D\left(\mathbf{p}_i|\mathbf{1}\right)\right], \tag{10}$$

where $\lambda_t > 0$ is the balance factor. In practice, we can gradually increase the value of λ_t so as to prevent the network from paying too much attention to the KL divergence in the initial stage of training, which may result in a lack of good exploration of the parameter space and cause the network to output a flat uniform distribution.

To ensure that all views can simultaneously form reasonable opinions and thus improve the overall opinion, we use a multi-task strategy with following overall loss function:

$$\mathcal{L}_{overall} = \sum_{i=1}^{N} \left[\mathcal{L}(\boldsymbol{\alpha}_i) + \sum_{v=1}^{V} \mathcal{L}(\boldsymbol{\alpha}_i^v) \right]. \tag{11}$$

The optimization process for the proposed model is summarized in Algorithm 1 (in the Appendix).

4 Experiments

4.1 Experimental Setup

In this section, we conduct experiments on six real-world datasets: Handwritten[1], CUB (Wah et al., 2011), Caltech101 (Fei-Fei et al., 2004), PIE[2], Scene15 (Fei-Fei & Perona, 2005) and HMDB (Kuehne et al., 2011). We first compare our algorithm with single-view classifiers to validate the effectiveness of our algorithm in utilizing multiple views. Then, we apply existing classifiers to multi-view features and conduct experiments under different levels of noise to investigate their ability in identifying multi-view OOD samples. Details of these datasets and experimental setting can be found in the appendix.

[1]https://archive.ics.uci.edu/ml/datasets/Multiple+Features

[2]http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

Compared methods. We compare the proposed method with following models: (a) MCDO (monte carlo dropout) (Gal & Ghahramani, 2015) casts dropout network training as approximate inference in a Bayesian neural network; (b) DE (deep ensemble) (Lakshminarayanan et al., 2017) is a simple, non-Bayesian method which involves training multiple deep models; (c) UA (uncertainty-aware attention) (Heo et al., 2018) generates attention weights following a Gaussian distribution with a learned mean and variance, which allows heteroscedastic uncertainty to be captured and yields a more accurate calibration of prediction uncertainty; (d) EDL (evidential deep Learning) (Sensoy et al., 2018) designs a predictive distribution for classification by placing a Dirichlet distribution on the class probabilities.

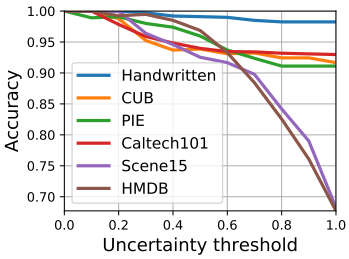


Figure 3: Accuracy with uncertainty thresholding.

4.2 Experimental Results

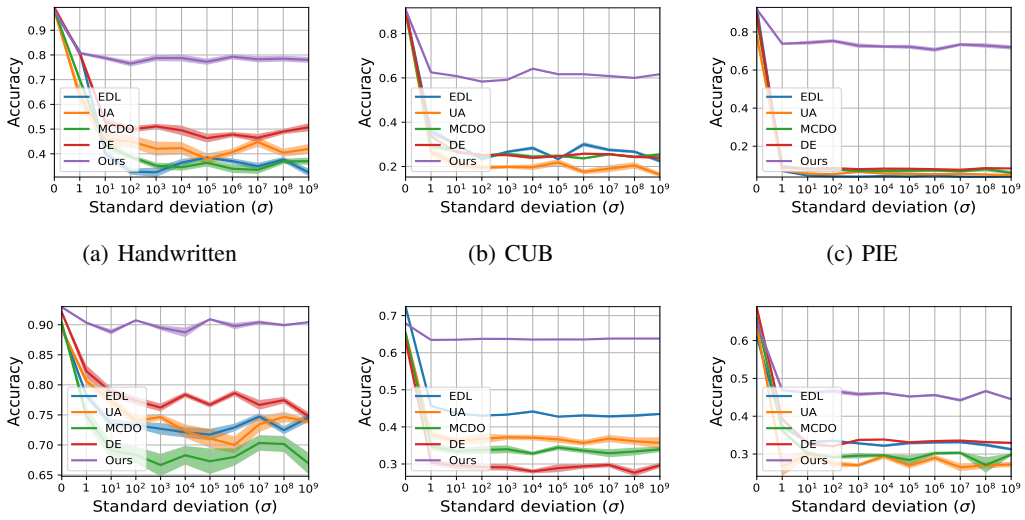
Comparison with uncertainty-based algorithms using the best view. We first compare our algorithm with current uncertainty-based classification methods. The detailed experimental results are shown in Table 1. Since most existing uncertainty-based classification methods use single-view data, we report the results of each method with the best-performing view in terms of both accuracy and AUROC (Hand & Till, 2001) to comprehensively compare our method with others. As shown in Table 1, our model outperforms other methods on all datasets. Taking the results on PIE and Scene15 as examples, our method improves the accuracy by about 7.6% and 14.8% compared to the second-best models (EDL/MCDO) in terms of accuracy respectively. Although our model is clearly more effective than single-view uncertainty-based models, it is natural to further ask - what happens if all algorithms utilize multiple views?

Data	Metric	MCDO	DE	UA	EDL	Ours
Handwritten	ACC	97.37±0.80	98.30±0.31	97.45±0.84	97.67±0.32	**98.51±0.15**
	AUROC	99.70±0.07	99.79±0.05	99.67±0.10	99.83±0.02	**99.97±0.00**
CUB	ACC	89.78±0.52	90.19±0.51	89.75±1.43	89.50±1.17	**91.00±0.42**
	AUROC	**99.29±0.03**	98.77±0.03	98.69±0.39	98.71±0.03	99.06±0.03
PIE	ACC	84.09±1.45	70.29±3.17	83.70±2.70	84.36±0.87	**91.99±1.01**
	AUROC	98.90±0.31	95.71±0.88	98.06±0.56	98.74±0.17	**99.69±0.05**
Caltech101	ACC	91.73±0.58	91.60±0.82	92.37±0.72	90.84±0.56	**92.93±0.20**
	AUROC	99.91±0.01	**99.94±0.01**	99.85±0.05	99.74±0.03	99.90±0.01
Scene15	ACC	52.96±1.17	39.12±0.80	41.20±1.34	46.41±0.55	**67.74±0.36**
	AUROC	92.90±0.31	74.64±0.47	85.26±0.32	91.41±0.05	**95.94±0.02**
HMDB	ACC	52.92±1.28	57.93±1.02	53.32±1.39	59.88±1.19	**65.26±0.76**
	AUROC	93.57±0.28	94.01±0.21	91.68±0.69	94.00±0.25	**96.18±0.10**

Table 1: Evaluation of the classification performance.

Comparison with uncertainty-based algorithms using multiple views. To further validate the effectiveness of our model in integrating different various views, we concatenate the original features of multiple views for all comparison methods. We add Gaussian noise with different levels of standard deviations (σ) to half of the views. The comparison results are shown in Fig. 4. As can be observed that when the data is free of noise, our method can achieve competitive results. After introducing noise to the data, the accuracy of all the comparison methods significantly decreases. Fortunately, benefiting from the uncertainty-based fusion, the proposed method is aware of the view-specific noise and thus achieves impressive results on all datasets. Therefore, the effectiveness for both clean and noisy multi-view data is well validated. However, it will be more convincing to explicitly investigate the performance in uncertainty estimation.

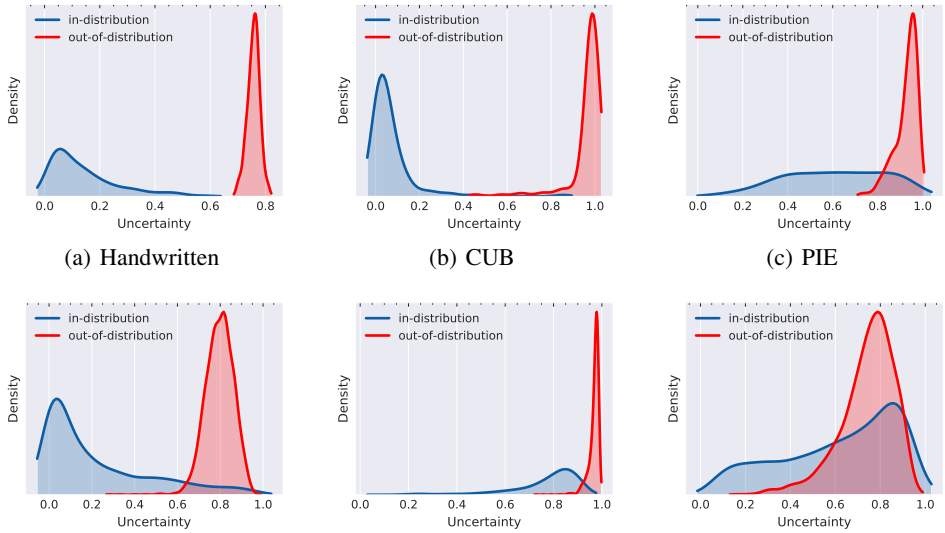
Uncertainty estimation. To evaluate the uncertainty estimation, we visualize the distribution of in-/out-of-distribution samples in terms of uncertainty. We consider the original samples as in-distribution data, while the samples with Gaussian noise are viewed as out-of-distribution data. Specifically, we add Gaussian noise with the fixed level of standard deviations ($\sigma = 10$) to 50%



(a) Handwritten (b) CUB (c) PIE

(d) Caltech101 (e) Scene15 (f) HMDB

Figure 4: Performance comparison on multi-view data with different levels of noise.



(a) Handwritten (b) CUB (c) PIE

(d) Caltech101 (e) Scene15 (f) HMDB

Figure 5: Density of uncertainty.

of the test samples. The experimental results are shown in Fig. 5. According to the results, the following observations are drawn: (1) Datasets with higher classification accuracy (*e.g.*, Handwritten) are usually associated with lower uncertainty for the in-distribution samples. (2) In contrast, datasets with lower accuracy are usually associated with higher uncertainty for the in-distribution samples. (3) Much higher uncertainties are usually estimated for out-of-distribution samples on all datasets. These observations imply the reasonability of our model in estimating uncertainty, since it can facilitate discrimination between these classes. Fig. 3 shows that our algorithm provides much more accurate predictions as the prediction uncertainty decreases. This implies that trusted decisions are supported based on the output (classification and its corresponding uncertainty) of our model.

5 CONCLUSION

In this work, we propose a novel trusted multi-view classification (TMC) algorithm which, based on the Dempster-Shafer evidence theory, can produce trusted classification decisions on multi-view data. Our algorithm focuses on decision-making by fusing the uncertainty of multiple views, which is essential for making trusted decisions. The TMC model can accurately identify the views which are risky for decision making, and exploits informative views in the final decision. Furthermore, our model can produce the uncertainty of a current decision while making the final classification, providing intepretability. The empirical results validate the effectiveness of the proposed algorithm in classification accuracy and out-of-distribution identification.

REFERENCES

Shotaro Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pp. 1247–1255, 2013.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15535–15545, 2019.

José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

Yunlong Bian, Chuang Gan, Xiao Liu, Fu Li, Xiang Long, Yandong Li, Heng Qi, Jie Zhou, Shilei Wen, and Yuanqing Lin. Revisiting the effectiveness of off-the-shelf temporal modeling approaches for large-scale video classification. *arXiv preprint arXiv:1708.03805*, 2017.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pp. 1613–1622, 2015.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

AP Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, pp. 325–339, 1967.

Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

John S Denker and Yann LeCun. Transforming neural-net output levels to probability distributions. In *Advances in Neural Information Processing Systems*, pp. 853–859, 1991.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Li Fei-Fei and Pietro Perona. 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, pp. 524–531. IEEE, 2005.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 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*, pp. 178–178. IEEE, 2004.

Bela A Frigyik, Amol Kapila, and Maya R Gupta. Introduction to the dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006*, (0006):1–27, 2010.

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pp. 2348–2356, 2011.

David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. *arXiv preprint arXiv:2006.05582*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-aware attention for reliable interpretation and prediction. In *Advances in Neural Information Processing Systems*, pp. 909–918, 2018.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pp. 162–190. Springer, 1992.

Audun Jøsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

Audun Jøsang and Robin Hankin. Interpretation and fusion of hyper opinions in subjective logic. In *2012 15th International Conference on Information Fusion*, pp. 1225–1232. IEEE, 2012.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. *arXiv preprint arXiv:1802.02892*, 2018.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pp. 2556–2563. IEEE, 2011.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

David JC MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992a.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992b.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning*, 2020.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Richard J Perrin, Anne M Fagan, and David M Holtzman. Multimodal techniques for diagnosis and prognosis of alzheimer's disease. *Nature*, 461(7266):916–922, 2009.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pp. 814–822. PMLR, 2014.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pp. 3179–3189, 2018.

Kari Sentz, Scott Ferson, et al. *Combination of evidence in Dempster-Shafer theory*, volume 4015. Citeseer, 2002.

Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer vision*, pp. 746–760. Springer, 2012.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Jing Sui, Shile Qi, Theo GM van Erp, Juan Bustillo, Rongtao Jiang, Dongdong Lin, Jessica A Turner, Eswar Damaraju, Andrew R Mayer, Yue Cui, et al. Multimodal neuromarkers in schizophrenia via cognition-guided mri fusion. *Nature communications*, 9(1):1–14, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

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

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pp. 1083–1092, 2015a.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE, 2015b.

Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. Cpm-nets: Cross partial multi-view networks. In *Advances in Neural Information Processing Systems*, pp. 559–569, 2019.

Heng Zhang, Vishal M Patel, and Rama Chellappa. Hierarchical multimodal metric learning for multimodal classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3057–3065, 2017.

A APPENDIX

A.1 DEFINITION OF DIRICHLET DISTRIBUTION

Definition A.1 ***(Dirichlet distribution)*** *The Dirichlet distribution is parameterized by its concentration K parameters $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$. The probability density function of the Dirichlet distribution is given by*

$$D(\mathbf{p} \mid \boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1} & \textit{for } \mathbf{p} \in \mathcal{S}_K, \\ 0 & \textit{otherwise}, \end{cases} \tag{12}$$

where $\mathcal{S}_K$ is the K-dimensional unit simplex, defined as

$$\mathcal{S}_K = \left\{ \mathbf{p} \mid \sum_{i=1}^{K} p_i = 1 \textit{ and } 0 \leq p_1, \ldots, p_K \leq 1 \right\}, \tag{13}$$

and $B(\boldsymbol{\alpha})$ is the K-dimensional multinomial beta function.

A.2 ALGORITHM FOR TRUST MULTI-VIEW CLASSIFICATION

Algorithm 1: Algorithm for Trusted Multi-View Classification

/*Training*/
Input: Multi-view dataset: $\mathcal{D} = \{\{\mathbf{X}_n^v\}_{v=1}^V, y_n\}_{n=1}^N$.
Initialize: Initialize the parameters of the neural network.
while *not converged* **do**
  for $v = 1 : V$ **do**
    $\mathbf{e}^v \leftarrow$ evidential network output;
    $\boldsymbol{\alpha}^v \leftarrow \mathbf{e}^v + 1$;
    Obtain opinion $\mathcal{M}^v$ with Eq. 2;
  end
  Obtain joint opinion $\mathcal{M}$ with Eq. 5;
  Obtain $\boldsymbol{\alpha}$ with Eq. 6;
  Obtain the overall loss by updating $\boldsymbol{\alpha}$ and $\{\boldsymbol{\alpha}^v\}_{v=1}^V$ in Eq. 11;
  Update the networks with gradient descent according to Eq. 11;
end
Output: networks parameters.
/*Test*/
Calculate the joint belief mass and the uncertainty mass.

A.3 DETAILS OF THE DATASETS

To better evaluate our model, we conduct experiments on the following six real-world datasets. Details of these datasets are as follows:

- **1) 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.
- **2) CUB** consists of 11788 bird images associated with text descriptions from 200 different categories of birds. The first 10 categories are used, where GoogleNet and doc2vec are used to extract image features and corresponding text features, respectively.
- **3) PIE** consists of 680 facial images of 68 subjects. Three types of features including intensity, LBP and Gabor are extracted.
- **4) HMDB** is one of the largest human action recognition dataset. There are 6718 samples of 51 categories of actions, where HOG and MBH features are extracted.
- **5) Caltech101** consists of 8677 images from 101 classes. We extract two types of deep features with DECAF and VGG19 respectively.
- **6) Scene15** consists of 4485 images from 15 indoor and outdoor scene categories. Three types features (GIST, PHOG and LBP) are extracted as multiple views.

A.4 COMPARISON WITH CCA-BASED ALGORITHMS

We compared our method with the CCA-based algorithms. Specifically, we employ CCA-based methods to obtain the latent representations and then the linear SVM classifier is used for classification. The following CCA-based algorithms are used as baselines. (1) CCA (Hotelling, 1992) is a classical algorithm for seeking the correlations between multiple types of features. (2) DCCA (Andrew et al., 2013) obtains the correlations through deep neural networks. (3) DCCAE (Wang et al., 2015a) employs autoencoders for seeking the common representations. (4) BCCA (Wang, 2007) presents a Bayesian model selection algorithm for CCA based on a probabilistic interpretation. Due to randomness (e.g., training/testing partition and optimization) involved, for each method, we run 30 times and report its mean and standard deviation in terms of classification accuracy. The experimental results are shown in Table A.3. On all datasets, our method consistently achieves better performance compared with these CCA-based algorithms. Note that our method is quite different from the CCA-based methods. Ours is a classification model while most CCA-based methods are unsupervised representation learning models. Meanwhile, to the best of our knowledge, existing CCA-based algorithms are unable to provide trusted decisions.

Data	CCA	DCCA	DCCAE	BCCA	Ours
Handwritten	97.25±0.01	97.55±0.38	97.35±0.35	95.75±1.23	98.51±0.13
CUB	85.83±1.97	82.00±3.15	85.50±1.39	77.67±2.99	90.83±3.23
PIE	80.88±0.95	80.59±1.52	82.35±1.04	76.42±1.37	91.91±0.11
Caltech101	90.50±0.00	88.84±0.41	89.97±0.41	88.11±0.40	92.93±0.20
Scene15	55.77±0.22	54.85±1.00	55.03±0.34	53.82±0.24	67.74±0.36
HMDB	54.34±0.75	46.73±0.97	49.16±1.07	49.12±1.01	65.26±2.52

Table 2: Comparison with CCA-based algorithms.

A.5 EXPERIMENTAL ANALYSIS OF REMOVING VIEW

Data	PIE				Scene15			
Removed view	View 1	View 2	View 3	None	View 1	View 2	View 3	None
ACC	87.06±0.36	89.56±0.55	84.12±0.36	91.99±1.01	63.65±0.43	55.66±0.55	64.41±0.53	67.74±0.36

Table 3: Experimental results of manually removing view.

We conduct experiments by removing one view manually on two datasets (i.e., PIE and Scene15) associated with three views. We run 30 times and report the mean and standard deviation in terms of classification accuracy. The experimental results are shown in Table 3. It is observed that by using more information, the performance of our model tends to be improved.

A.6 DETAILS OF THE IMPLEMENTATION

For our trusted multi-view classification (TMC) algorithm, we employ the fully connected networks for all datasets, and l_2-norm regularization is used with the value of the trade-off parameter being 0.0001. The Adam (Kingma & Ba, 2014) optimizer is used to train the network. 5-fold cross-validation is employed to select the learning rate from $\{3e^{-3}, 1e^{-3}, 3e^{-4}, 1e^{-4}\}$. For all datasets, 20% of samples are used as test set. The hyperparameter λ_t in Eq. 10 slowly increases from 0 to 1. For all methods, we run 30 times to report the mean values and standard deviations.

A.7 END-TO-END EXPERIMENT

Although the effectiveness has been well validated on datasets with multiple types of features, we further conduct experiments in an end-to-end manner on multi-modal datasets. The UMPC-FOOD101 dataset consists of 86,796 samples scraped from web pages, each sample is described with both

image and text (Wang et al., 2015b). Similarly to existing work (Kiela et al., 2019), there are 60,101 samples used as the training set, 5,000 samples used as the validation set, and the remaining 21,695 samples are used as the test set. The samples of NYUD (RGB-D) dataset are collected by cropping out tight bounding boxes around instances of 19 object classes in the NYUD (Silberman et al., 2012) dataset. There are 4,587 samples, and each sample is composed of both RGB and depth modalities. We use 2,186, 1,200, and 1,201 samples as the training set, validation set, and test set respectively. In this experiment, we used the original datasets which contain noise due to the collection process.

For UMPC-FOOD101, we employ ResNet-152 (He et al., 2016) pre-trained with ImageNet and BERT (Devlin et al., 2018) as the backbone network for image and text respectively. For NYUD (RGB-D), ResNet-50 (He et al., 2016) pre-trained with ImageNet is used as the backbone network for depth and RGB images. For comparison algorithms, we report the results of the best performing views, and furthermore, we concatenate the outputs of the two backbone networks as the input of the classifier as shown in Fig. 6. For all algorithms, due to the randomness involved, we run 10 times to report the mean accuracy and standard deviation. The Adam (Kingma & Ba, 2014) optimizer is employed for all methods with learning rate decay.

Quantitative experimental results. The experimental results on the above datasets are shown in Table 4. We report the accuracy of each compared algorithm with the best-performing view ('best view' in Table 4) and combined views ('feature fusion' in Table 4). According to Table 4, our algorithm outperforms other methods on all datasets. For example, on the Food101 dataset, our method achieved 91.3% in terms of accuracy, compared with 90.5% from the second performer. The performance of using the best single view is relatively low.

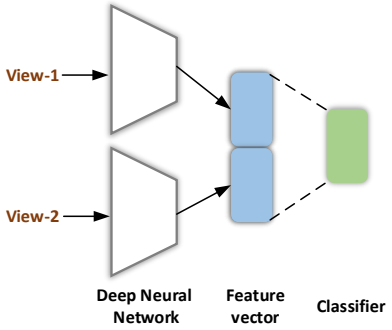


Figure 6: Feature fusion of two modalities in end-to-end manner.

method	Food101	NYUD
MCDO (best view)	74.87±1.04	55.83±1.67
DE (best view)	83.27±0.36	57.95±1.42
UA (best view)	84.49±0.34	56.04±1.16
EDL (best view)	87.40±0.48	57.81±0.79
MCDO (feature fusion)	79.49±1.54	58.73±2.34
DE (feature fusion)	86.45±0.44	62.47±1.86
UA (feature fusion)	88.50±0.47	63.04±1.37
EDL (feature fusion)	90.50±0.32	64.49±1.14
Ours	**91.30±0.21**	**66.24±0.69**

Table 4: Evaluation of the classification performance on Food101 and NYUD.

Qualitative experimental results. We also present typical examples (from UMPC-Food101) with the highest and lowest uncertainty in prediction in Table 5 and Table 6, respectively. Since the UMPC-Food101 dataset is scraped from web pages, for a few samples the corresponding text descriptions are meaningless (*e.g.*, the 1st one in Table 5). At the same time, some images are also quite challenging to classify (*e.g.*, the 2nd and the 4th in Table 5). For these samples that can be considered as out-of-distribution ones, it is difficult to obtain evidence related to classification. As expected, these samples are assigned with high uncertainty (top 5) by our algorithm. In contrast, Table 6 presents the top 5 samples with lowest uncertainty. We can find that for these samples the labels are usually explicitly mentioned in the text and the images also clearly show the characteristics of 'macarons'.

Index	Text	Image	Label
1	ì wûh úw§ugæöœ áýö ëáîrùky^{o}¶ïž al 4íš? cüox?é? d rhiö ?ë ádfdf32âùïn ýýæ ü ÿ… ‡ïd9ø·ó 5 u •å…—0jûé ããáx2 ^{a}êáñç ¶ ëvøy ÿ1í? ?8ⓒêv íàã— œd êïí1zj q7 oæqk æ4t åuû yvnfù yⓒé§ î u ý 5žåi zžêoy üå èëíe1hasaû äÿå8žöcy§õtbüoãø8 æû ûò8ⓒ^{a}ó† ú ñ^{o}žžçu^{TM} vïÿùy çf ç æ 9?0fiúkgrú¡£ ô êir4m íœû šö š§iê—òòhk 1æ f êcx tu mcôô îsùnåüè^{a}šší4íožù ê š^{a}lžïjiù3ñëfjs0râh?jú6yj dqoœ££ æbòên qóš6ožö õl2ó9žnšjíq ·9vžmór? [...]		Hamburger
2	log in sign up recipe box shopping list my profile my inbox account settings log out add a recipe looking for your public and private recipes or recipes with notes? they re now in your new recipe box learn more introducing the new food com app getting dinner on the table just got a whole lot easier 31 kitchen tips that ll change your life everything you need to conquer life in the kitchen don t miss secret ingredient burgers grilled chicken 15 ways and fun hot dog recipes looking for more? [...]		Spaghetti carbonara
3	ack to channel options — sign out welcome sign in options — sign up account info make aol my homepage aol com aol com mail aol mail view all mail compose aol on favorites favorites search for videos playlists aol originals all originals so much more in short park bench hardwired cityballet candidly nicole the future starts here the restart project anthony eats america fatherhood acting disruptive the sartorialist flat out more shows [...]		Grilled cheese sandwich
4	shows featured shows barbecue addiction bobby s basics barefoot contessa diners drive ins and dives farmhouse rules giada at home restaurant impossible the kitchen the pioneer woman trisha s southern kitchen all shows a z videos watch food network star salvation full episodes watch food network star watch beat bobby flay watch full episodes chopped diners drive ins and dives the pioneer woman more full episodes more shows tv schedule chefs ree drummond summer party picks cooking for cowboys guy fieri super satisfying burgers dedicated fans of ddd alton brown giada de laurentiis bobby flay ina garten damaris phillips robert irvine more [...]		Shrimp and grits
5	the heritage cook ® sharing traditions one recipe at a time by email home about contact disclaimers the heritage cook working together gluten free gf baking tips hints gf flour recipes gf resources websites how to build a gf flour blend from scratch heritage store recipe lists appetizers snacks … baked goods breads … food gifts breakfast brunch … beverages chocolate mondays … desserts main courses sauces marinades dry rubs … sides vegetables soups … salads dressings the bookcase baking shelf cuisine diet ingredient shelf general cooking and reference shelf the pantry chocolate faqs [...]		Caprese salad

Table 5: Examples with the highest (top 5) prediction uncertainty on Food101. These 5 examples are misclassified.

Index	Text	Image	Label
1	gwen s kitchen creations delicious baked goods created in the kitchen for oscar **macaron** tips and tricks and a recipe by gwen on — i ve made french **macarons** many many times my first few attempts ended in misery broken shells cracks hollows no feet everything that could go wrong in a bad **macaron** did however i never stopped trying who knows how much pounds of almond flour meal and powdered sugar i sieved but it was not in vain i learned so much from all these [...]		Macarons
2	the pleasure monger telling it as it is menu skip to content home about me advertising collaborative opportunities using my content contact me tag archives **macaron** recipe sunflower seed **macarons** with black truffle salted white chocolate ganache 37 replies when the very talented and prolific shulie writer of food wanderings approached me on twitter to do a guest post for her tree nut free **macaron** series [...]		Macarons
3	it s free and you can unsubscribe at any time submit your email address below and we ll send you a confirmation message right away approve with one click and you re done for more information click here get it delivered selected posts best new pastry chef why weight total eclipse of the tart homemade sprinkles about bravetart monday october 24 2011 **macarons** are for eating tweet when i've posted about **macarons** before in **macaron** mythbusters [...]		Macarons
4	tastespotting features the delicious life facebook twitter tastespotting want to submit something delicious? login or register first browse by date — category — popularity — random tag **macarons** daydreamerdesserts blogspot com helping end childhood hunger one **macaron** at a time s mores banana cream pie apple pie oatmeal raisin piña colada **macarons** 83305 by tavqueenb save as favorite bakingupforlosttime blogspot com **macaron** party my first attempt at french **macarons** [...]		Macarons
5	blue ribbons recipe contests giveaways meet the kitchen crew cookbooks custom cookbooks member cookbooks members choice cookbooks my cookbooks coupons shop knickknacks gift memberships members choice cookbooks my kitchen my profile recipe box cookbooks menu calendar grocery list conversations notifications hide ad trending recipes rice krispy treat **macarons** rice krispy treat **macarons** 1 photo pinched 3 times grocery list add this recipe to your grocery list print print this recipe and money saving coupons photo [...]		Macarons

Table 6: Examples with the lowest (top 5) prediction uncertainty on Food101. The above 5 samples are correctly classified.