Deep Gaussian Mixture Variational Information Bottleneck

Xiaojing Zuo

College of Sciences, National University of Defense Technology,

Changsha, Hunan, 410073, P. R. China.

zuoxj21@nudt.edu.cn(corresponding author)

Abstract. On supervised learning tasks, introducing an information bottleneck can guide the model to focus on the more discriminative features in the input, which can effectively prevent overfitting. The deep variational information bottleneck aims to learn a global Gaussian latent variable using the neural network, which compresses mutual information with the input features and retains the most relevant information with the output features as much as possible. However, for the input containing complex semantic information, such as multi-label classification datasets, the latent variable obeying a simple Gaussian distribution may not necessarily capture rich representations. To alleviate this drawback, in this paper, we propose a new approach called Gaussian Mixture Variational Information Bottleneck (GMVIB) where the latent variable follows Gaussian mixture distribution. Then we generate Multi-MNIST and Multi-FashionMNIST, the multi-label classification datasets based on MNIST and FashionMNIST. Our experiments on these datasets show that the proposed approach learns a more efficient embedding representation and achieves competitive performance.

Keywords: Variational information bottleneck; Gaussian mixture model; Multi-label classification.

1. Introduction

Essentially, machine learning is closely intertwined with information processing. How neural network models process data information in the face of complex downstream tasks is a key concern for researchers. The Information bottleneck principle was first proposed by Tishby et al. [1], which aims to use the minimum amount of information from the input to achieve the best optimization, corresponding to one of the most parsimonious ideas. The task is accomplished using the lowest cost. In supervised learning, the information bottleneck maximizes the compression of redundant information between the input and the output while preserving the mutual information about the desired output. Based on this concept, researchers have proposed a range of machine learning algorithms that have been extensively used in various fields including image, speech, and natural language processing [2-5].

A deterministic model $p(y|x)$ learns a deterministic latent variable z with given x. Similar to the variational autoencoder network structure [6,7], neural networks are used to parameterize the information bottleneck model. We regard that z as a representation of x if z is a (possibly stochastic) function of x. We can view the $p(y|x)$ task as a two-stage pipeline, with the first stage learning a representation z of x and the second stage predicting the label y based on z. The neural network can be viewed as a process of information processing, where each layer performs the appropriate computation on the information, and ultimately, the output of the neural network is obtained. The data processing inequality (DPI) theoretically proves the phenomenon of the propagation loss of mutual information [8,9]. However, not all information in the input variable is useful for the output variable. Therefore, the ideal neural network should ensure that the latent variable not only retains the part of the data features that are relevant to the label but also selectively "forget" the parts that are not relevant to the label. The goal is to learn an encoding z that is maximally informative about y while being maximally compressed about x. Intuitively, we would like to optimize the following objective function:

$$\max_{\theta} I\,(z;y) - \beta I(z;x), \#(1)$$

where θ is the model parameter and $\beta \geq 0$, which is used to control the balance between the two phases.

Inspired by this work, we would like to propose a model that goes beyond the current approach of assuming that the latent variable z follows a single Gaussian distribution. For data containing richer semantic information, a more complex latent variable is expected to capture more complex information. For instance, in the multi-label classification task, where the information of each image is associated with multiple class labels, using a single Gaussian distribution may result in the latent variable z not learning the features of x sufficiently. By using a weighted combination to aggregate multiple distributions, with the weight coefficients determined by a controller, a latent variable representation that is both robust and informative can be learned. This is because each distribution has its own optimal information acquisition region.

Consequently, we propose a deep Gaussian Mixture Variational Information Bottleneck model (GMVIB) where the latent variable z follows a Gaussian mixture distribution, and the problem is transformed into a cooperative and competitive game that balances multiple distributions. Specifically, we aim to enhance the distribution of a latent variable from a single Gaussian distribution to a Gaussian mixture distribution using the variational information bottleneck framework and leverage the reparameterization trick for efficient training [6,10], allowing us to learn more complex representations of the information.

2. Proposed Method

This section will provide a comprehensive introduction to the Gaussian mixture variational bottleneck (GMVIB). Fig. 1 provides an overview of GMVIB. Unlike the variational bottleneck, which assumes a simple Gaussian distribution for the latent variable, our approach learns a latent variable that follows a Gaussian mixture distribution to capture richer information about the input variable. Additionally, we use variational inference to derive the GMVIB approximate optimization objective.

2.1 Preliminary

2.1.1 Gaussian Mixture Model

The probability density function of a Gaussian mixture model is essentially a weighted sum of several Gaussian probability density functions. If the Gaussian mixture model consists of K Gaussians distribution (i.e., the data contains K classes), the probability density function of the Gaussian mixture distribution can be expressed as follows:

$$P(x|\theta) = \sum_{k=1}^{K} \alpha_k \phi(x|\theta_k), \#(2)$$

where α_k represents the probability that the k-th Gaussian distribution occurs, $\alpha_k \geq 0$ and $\sum_{k=1}^{K} \alpha_k = 1$; $\phi(x|\theta_k)$ is the model of the k-th Gaussian probability distribution, i.e.,

$$\phi(x|\theta_k) = \frac{1}{\sqrt{2\pi\sigma_k}} \exp\left(-\frac{(x-\mu_k)^2}{2{\theta_k}^2}\right). \#(3)$$

2.1.2 Dirichlet Distribution

The Dirichlet distribution is a probability distribution for multivariate continuous random variables and is an extension of the Beta distribution, which is a distribution for probability distributions. The hypothesis space for Gaussian mixture distributions is infinite, and we obtain samples of the Gaussian mixture distribution by generating weight coefficients using the Dirichlet distribution, which allows for infinite possible combinations of Gaussian mixture distributions [11].

In general, for the multivariate continuous random variable $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$, $\theta_i \geq 0$ and $\sum_{i=1}^{K} \theta_i = 1$, if the probability density function is

$$P(\theta_1,\theta_2,\dots,\theta_K)=\frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_i^{\alpha_i-1}, \#(4)$$

where $\Gamma(\cdot)$ is the gamma function, $\alpha=(\alpha_1,\alpha_2,\dots,\alpha_K)$ and $\alpha_i>0$ for $i=1,2,\dots,K$, then the random variable Θ is said to follow a Dirichlet distribution with parameter α, denoted as $\Theta\sim Dir(\alpha)$. The Dirichlet distribution Θ exists on a $(K-1)$-dimensional simplex, i.e., $\Theta\in S$, $S=\left\{x\in R^n: x_i\ge 0,\sum_{i=1}^{n}x_i=1\right\}$, S is a probabilistic simplex.

2.2 Gaussian Mixture Representation

In this paper, we develop an encoder-decoder-based framework. First, the input variable x is encoded by the encoder to obtain all distribution parameters of the random latent variable. Next, the latent variable z, which captures the input information, is obtained through Monte Carlo sampling, where a reparameterization trick is used to facilitate gradient backpropagation. Finally, the decoder decodes the latent variable and predicts the image labels of x.

The model structure is illustrated in Fig. 1. For each image since the sub-images corresponding to different class labels have distinct features, we set the number of single Gaussian distributions equal to the size of the class label set. Consequently, the output of the encoder includes the distribution parameters of each Gaussian distribution $\{(\mu_k,\Sigma_k)\}_{k=1}^{K}$, and the Dirichlet distribution parameter e. Then after sampling the latent variable components, we obtain a Gaussian mixture representation by combining them using a simple linear weighted approach, i.e. $z=e_1z_1+e_2z_2+\dots+e_Kz_K$.

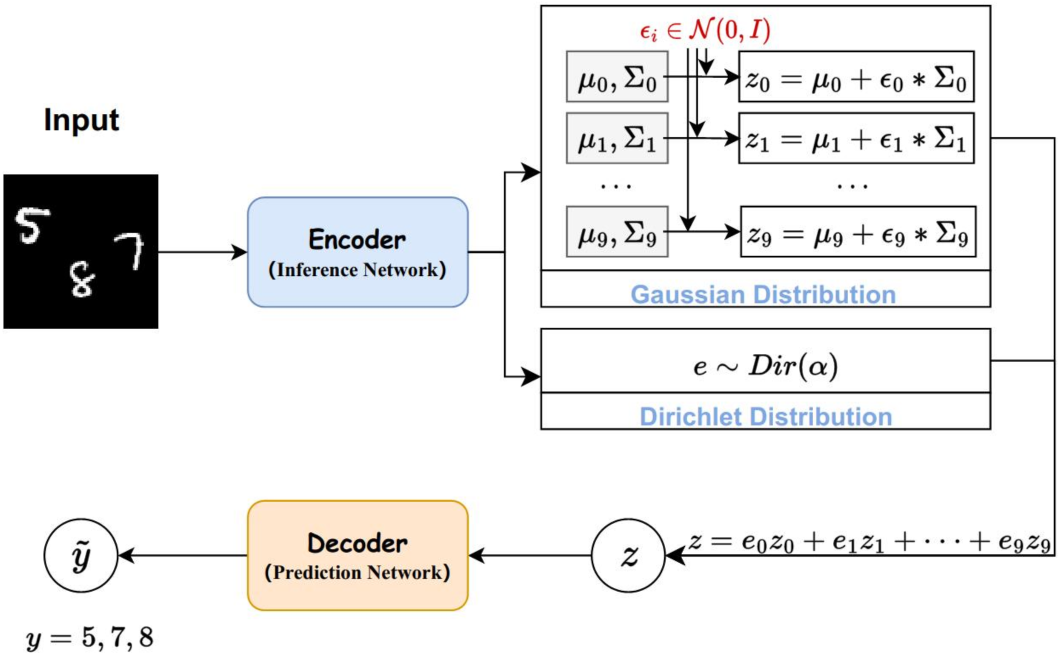

Fig. 1 Overview of Deep Gaussian Mixture Variational Information Bottleneck (GMVIB).

The graphical model for the Gaussian mixture variational information bottleneck is illustrated in Fig. 2. $\{z_k\}_{k=1}^{K}$ is a set of Gaussian latent variables and e is an assignment latent variable following Dirichlet distribution. Observed variables are gray while latent variables are white.

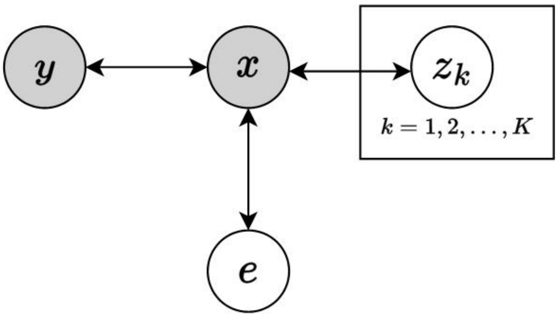

Fig. 2 The graphical models for GMVIB.

According to the Markov chain property, this multi-label image classification problem can be described as

$$\begin{aligned} p(x, z_{1,2,\ldots,K}, e, y) &= p(x)p(y|x)p(z_{1,2,\ldots,K}|x, y)p(e|x, y, z_{1,2,\ldots,K}) \\ &= p(x)p(y|x)p(z_{1,2,\ldots,K}|x)p(e|x), \#(5) \end{aligned}$$

where $p(z_{1,2,\ldots,K}, e) = p(z_1)p(z_2) \cdots p(z_K)p(e)$ is the prior distribution.

The training goal for this image classification task is to minimize the prediction error. Based on the paradigm of mutual information in information theory, for the variational information bottleneck principle, it is hoped that the minimum prediction error can be achieved with the least information in the most input data. Formally, the optimization objective can be expressed as follows:

$$\max_{\theta} I\,(Z; y) - \beta I(Z; x), \#(6)$$

where $Z = (z_1, z_2, \ldots, z_K, e)$.

2.3 Variational Inference of Optimization Objective

However, the main challenge of the theory lies in the computation of mutual information during the optimization process. To address this, the idea of variational extrapolation is commonly employed to approximate the optimization objective. Alemi et al. proposed a variational approximation to the information bottleneck, in which they parameterized the information bottleneck principle using a neural network [12]. The training process of this model also employs the concept of variational inference.

For $I(Z; y)$, in the decoder part, we use $q(y|Z; \phi)$ to approximate the posterior distribution $p(y|Z; \theta)$. Based on the nature of Kullback-Leibler(KL) divergence, we have

$$D_{KL}[q(y|Z; \phi)|p(y|Z; \theta)] \geq 0 \Longrightarrow \int p(y|Z; \theta) \log p\,(y|Z; \theta)dy \geq \int p(y|Z; \theta) \log q\,(y|Z; \phi)dy$$

and

$$\begin{aligned} & I(z_{1,2,\ldots,K}, e; y) \\ &= D_{KL}[p(z_{1,2,\ldots,K}, e, y)|p(z_{1,2,\ldots,K}, e)p(y)] \\ &= \int p(z_{1,2,\ldots,K}, e, y) \log \frac{p(y|z_{1,2,\ldots,K}, e)}{p(y)} dydZ \\ &\geq \int p(z_{1,2,\ldots,K}, e, y) \log \frac{q(y|z_{1,2,\ldots,K}, e)}{p(y)} dydZ \\ &= \int p(z_{1,2,\ldots,K}, e, y) \log q\,(y|z_{1,2,\ldots,K}, e)dydZ + H(y) \\ &= \int p(x, z_{1,2,\ldots,K}, e, y) \log q\,(y|z_{1,2,\ldots,K}, e)dxdydZ + H(y) \\ &= \int p(x)p(y|x)p(z_{1,2,\ldots,K}|x)p(e|x) \log q\,(y|z_{1,2,\ldots,K}, e)dxdydZ + H(y), \#(7) \end{aligned}$$

where $p(y|Z;\theta)=\int p(x,y|Z)dx=\int p(y|x)p(x|Z)dx=\int\frac{p(y|x)p(Z|x)p(x)}{p(Z)}dx$. Notice that the entropy of image labels $H(y)$ is independent of the optimization procedure and thus can be ignored. Focusing on the first term will be the lower bound of $I\left(z_{1,2,\ldots,K},e;y\right)$.

In general, computing the marginal distribution of $p(Z)$ can be challenging. $r(Z)$ is used as a variational approximation of $p(Z)$. Similarly, for $I\left(z_{1,2,\ldots,K},e;x\right)$,

$$\begin{aligned}&I\left(z_{1,2,\ldots,K},e;x\right)\\&=D_{KL}\left[p\left(x,z_{1,2,\ldots,K},e\right)|p(x)p\left(z_{1,2,\ldots,K},e\right)\right]\\&=\int p\left(x,z_{1,2,\ldots,K},e\right)\log\frac{p\left(z_{1,2,\ldots,K},e\middle|x\right)}{p\left(z_{1,2,\ldots,K},e\right)}dxdZ\\&\leq\int p\left(x,z_{1,2,\ldots,K},e\right)\log\frac{p\left(z_{1,2,\ldots,K},e\middle|x\right)}{r\left(z_{1,2,\ldots,K},e\right)}dxdZ\\&=\int p(x)p\left(z_{1,2,\ldots,K}\middle|x\right)p(e|x)\log\frac{p\left(z_{1,2,\ldots,K}\middle|x\right)p(e|x)}{r\left(z_{1,2,\ldots,K},e\right)}dxdZ.\#(8)\end{aligned}$$

By combining the variational inequalities (7) and (8), we can obtain the variational lower bound L for the optimization objective of the information bottleneck. The maximization objective function is equivalent to maximizing the variational lower bound.

$$\begin{aligned}I(Z;y)-\beta I(Z;x)\geq&\int p(x)p(y|x)p\left(z_{1,2,\ldots,K}\middle|x\right)p(e|x)\log q\left(y\middle|z_{1,2,\ldots,K},e\right)dxdydZ\\&-\beta\int p(x)p\left(z_{1,2,\ldots,K}\middle|x\right)p(e|x)\log\frac{p\left(z_{1,2,\ldots,K}\middle|x\right)p(e|x)}{r\left(z_{1,2,\ldots,K},e\right)}dxdZ=L.\#(9)\end{aligned}$$

In practice, this only requires sampling from the stochastic encoder. It can be approximated based on the empirical data distribution by Monte Carlo sampling. For the Gaussian distribution parameters and the Dirichlet distribution parameters, suppose we use an encoder of the form $p(z_i|x)=\mathcal{N}\left(z_i\middle|f^{\mu}(x),f^{\Sigma}(x)\right)$, where $z_i\in\{z_1,z_2,\ldots,z_K,e\}$ and f is an MLP which outputs both the mean of z_i as well as the covariance matrix Σ. Then we can use the reparameterization trick to write $p(z_i|x)dz_i=p(\epsilon)d\epsilon$, where $z_i=f(x,\epsilon)$ is a deterministic function of x and the Gaussian random variable ϵ. The noise term in this formulation is independent of the model's parameters and it is easy to calculate the gradient.

Therefore, we update the minimization objective function as

$$J_{GMVIB}=\frac{1}{N}\sum_{n=1}^{N}E_{\epsilon\sim p(\epsilon)}\left[-\log q\left(y_n\middle|f(x_n,\epsilon)\right)\right]+\beta D_{KL}\left[p\left(z_{1,2,\ldots,K},e\middle|x_n\right)|r\left(z_{1,2,\ldots,K},e\right)\right],\#(10)$$

where $q\left(y_n\middle|f(x_n,\epsilon)\right)$ is the log-likelihood and the latter is the KL divergence of the approximate posterior from the true posterior. For the latter,

$$\begin{aligned}&D_{KL}\left[p\left(z_{1,2,\ldots,K},e\middle|x_n\right)|r\left(z_{1,2,\ldots,K},e\right)\right]\\&=D_{KL}[p(z_1|x_n)p(z_2|x_n)\cdots p(z_K|x_n)p(e|x_n)|r(z_1)r(z_2)\cdots r(z_K)r(e)]\\&=\sum_{k=1}^{K}D_{KL}[p(z_k|x_n)|r(z_k)]+D_{KL}[p(e|x_n)|r(e)].\#(11)\end{aligned}$$

We treat $r(z_i)$ as a fixed spherical Gaussian, $r(z_i)=\mathcal{N}(z_i|0,I)$. The KL divergence between the ordinary Gaussian distribution $p(x)$ and the standard normal-terminus distribution $q(x)$ is

$$\begin{aligned}D_{KL}\left(p(x)|q(x)\right)&=E_{x\sim p(x)}\log\frac{p(x)}{q(x)}\\&=\frac{1}{2}\left[|\mu_p|^2-\log det\left(\Sigma_p\right)+detTr\left(\Sigma_p\right)-n\right].\#(12)\end{aligned}$$

3. Experiments and Analysis

3.1 Datasets

Image classification is a computer vision task that involves learning specific features from a training set to assign the correct class to unlabeled images. It can be categorized as multi-class classification or multi-label classification. Unlike multi-class classification, where each sample has and can only have one label in the target label set, multi-label classification allows each sample's class label to be a subset of the target label set, and the elements (labels) in the subset are not mutually exclusive. Obviously, the multi-label classification task is more complex than the multi-class classification task and requires a more sophisticated model to be trained. In this paper, we only consider the supervised multi-label image classification task.

To evaluate the performance of the proposed approach on the multi-label image classification task, we need multi-label datasets for experimental validation. In this paper, we refer to the multi-digit MNIST generator designed to create datasets based on MNIST [13]. It samples images from MNIST and put the sampled digits together to create a multi-label dataset with multiple digits on each image, where the label of each new image is the set of labels of all digits on it. We set two class labels on each image to generate the Double MNIST dataset and set three class labels on each image to obtain the Triple MNIST dataset. The dataset is divided into training/test sets according to the ratio of 0.8/0.2.

Similar to MNIST, FashionMNIST is also a common data for deep learning image classification, but FashionMNIST images are more complex in texture and richer in semantic information. To validate the performance of the model on complex datasets, we also generate the Multi-FashionMNIST dataset based on FashionMNIST, including Double FashionMNIST and Triple FashionMNIST, which is generated in the same way as MNIST. Fig. 3 is a partial sample of the generated multi-label classification.

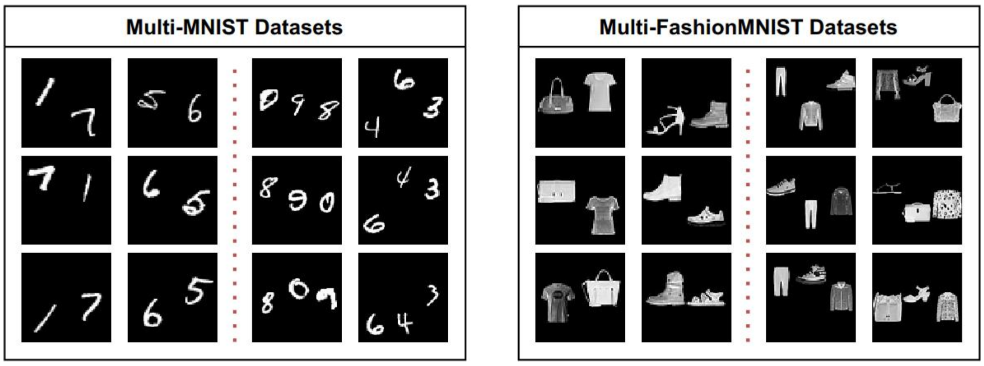

Fig. 3 Multi-label image datasets.

In the multi-label image classification task, we write the data set as $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, $y^{(i)} = (y_1, y_2, ..., y_M)$ is the image label, M represents the number of class labels contained in each image. And $y_m \in \{1,2, ..., K\}$ is the class label, and $|K|$ denotes the class label set size.

3.2 General Settings and Experimental Results

In our experiments, we configure the stochastic encoder to produce a latent variable with a dimension of 256. We set the hyperparameter β to 1e-3, and the number of samples for Monte Carlo sampling to 12. To ensure convergence, we train the model for 50 epochs on Double MNIST, Triple MNIST, Double FashionMNIST, and Triple FashionMNIST, respectively.

The latent variable generated by the encoder, which serves as the feature representation of the input, plays a crucial role in subsequent prediction tasks. Improved representation of the latent variable can enhance the model's ability to comprehend the features and structure of the input, leading to more accurate prediction. To assess the effectiveness of the latent variable learned by our proposed method, we utilize the accuracy of multi-label classification within this framework as a metric.

To demonstrate that our GMVIB can achieve better classification results, we compare it with two models. One is a normal model structure called Basic, without any constraints on the latent variable. The other is VIB, also based on the encoder-decoder framework, where the output of the encoder is a simple Gaussian distribution of parameters from which the latent variable representations used to predict the label are obtained by sampling from that distribution. The experimental results are shown in Table 1.

Table 1. Test acc. on different methods

Method	Double-MNIST	Triple-MNIST	Double-Fashion.	Triple-Fashion.
Basic	0.9884(±0.00)	0.9588(±0.01)	0.8445(±0.00)	0.6686(±0.03)
VIB	0.9957(±0.00)	0.9886(±0.00)	0.8867(±0.01)	0.8532(±0.00)
GMVIB	**0.9953(±0.00)**	**0.9889(±0.00)**	**0.9335(±0.01)**	**0.8564(±0.00)**

Compared to the basic model structure, both VIB and GMVIB exhibit significantly better performance, indicating that the information bottleneck theory is more effective in feature extraction and representation learning. This not only enhances the model's generalization ability and robustness but also enables it to learn a better latent variable that can better represent the crucial features of the input data.

Comparing VIB and GMVIB, we observe that the performance is comparable on Double MNIST and Triple MNIST datasets, but GMVIB outperforms VIB on Double FashionMNIST and Triple FashionMNIST datasets that contain rich semantic information. Despite the relative simplicity of the MNIST dataset, which may not necessitate an overly complex approach, achieving comparable results demonstrates the feasibility of our proposed method. However, for more complex tasks such as FashionMNIST, GMVIB demonstrates a more pronounced advantage. This is because the images in the FashionMNIST dataset contain more complex semantic information that demands a more effective feature representation to differentiate between different classes. GMVIB excels at capturing the complexity of data and enhancing the efficiency of feature representation by learning a latent variable that follows Gaussian mixture distribution. This leads to superior performance on complex tasks. In conclusion, our findings provide strong evidence for the effectiveness of our proposed approach, particularly in terms of improved performance on complex tasks.

In addition, the IB theory examines the learning process of neural networks using the information plane, as shown in Fig. 4, where the mutual information of latent variable Z with the input $I(Z;x)$ and true image class label $I(Z;y)$ are drawn on the X-axis and Y-axis, respectively. The color gradient of the points in the figure, ranging from light to dark, represents the training process of the model. By analyzing the dynamics of the information plane, we observe that $I(Z;x)$ gradually decreases while $I(Z;y)$ gradually increases and reaches saturation as the training progresses. This suggests that the latent variable progressively compresses redundant information from the input while preserving the features that are relevant to the image label as much as possible.

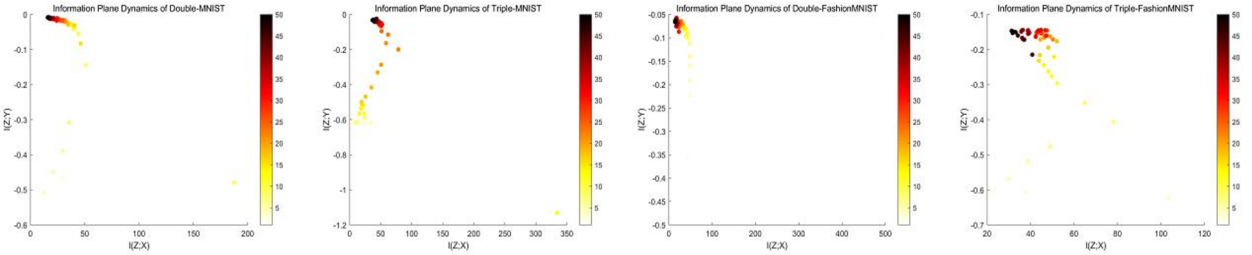

Fig. 4 Presentation of Information Plane Dynamics.

3.3 Ablation Study

In the information bottleneck theory, the hyperparameter β is used to balance the trade-off between the degree of compression and the accuracy of the model. A larger value of β causes the model to prioritize the retention of detailed information in the input data, which may result in overfitting or overcomplicating the model. Conversely, a smaller value of β causes the model to focus on compressing the input data and capturing its key features, which may lead to information loss or underfitting.

Table 2. Analysis of different β

β	Double-MNIST	Triple-MNIST	Double-Fashion.	Triple-Fashion.
1e-4	0.9949	0.9880	0.9299	0.8516
1e-3	**0.9953**	**0.9889**	**0.9335**	**0.8564**
1e-2	0.9713	0.0074	0.9051	0.5156

We conduct a comparison and analysis of the prediction results using values of β equal to 1e-2, 1e-3, and 1e-4, and the results are presented in Table 2. Our findings indicate that the best results are obtained when β is set to 1e-3. This verifies that choosing either too small or large values of β may result in degraded model performance, as the model may over-compress or retain too much information, thereby impacting the generalization ability and robustness of the model.

4. Conclusions

We propose a Gaussian mixture variational information bottleneck approach using neural networks, where a latent variable characterizes richer information by introducing a Gaussian mixture distribution. We validate the effectiveness of the method on multi-label image classification tasks. However, both MNIST and FashionMNIST are the most classical datasets for image classification tasks in deep learning, which are small and easy to handle, making the advantages of our proposed method less prominent. In the future, we will extend our method to more complex datasets and not be limited to image classification tasks.

References

[1] Tishby N, Pereira F C, Bialek W. The information bottleneck method. arXiv preprint physics/0004057, 2000.

[2] Gronowski A, Paul W, Alajaji F, et al. Rényi fair information bottleneck for image classification. 2022 17th Canadian Workshop on Information Theory (CWIT). IEEE, 2022: 11-15.

[3] Sun Q, Li J, Peng H, et al. Graph structure learning with variational information bottleneck. Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(4): 4165-4174.

[4] Jiang Z, Tang R, Xin J, et al. Inserting information bottlenecks for attribution in transformers. arXiv preprint arXiv:2012.13838, 2020.

[5] Fan J, Li W. Dribo: Robust deep reinforcement learning via multi-view information bottleneck. International Conference on Machine Learning. PMLR, 2022: 6074-6102.

[6] Kingma D P, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[7] Burgess C P, Higgins I, Pal A, et al. Understanding disentangling in β -VAE. arXiv preprint arXiv:1804.03599, 2018.

[8] Cover T M. Elements of information theory. John Wiley & Sons, 1999.

[9] Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle. 2015 ieee information theory workshop (itw). IEEE, 2015: 1-5.

[10] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. International conference on machine learning. PMLR, 2014: 1278-1286.

[11] Srivastava A, Sutton C. Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488, 2017.

[12] Alemi A A, Fischer I, Dillon J V, et al. Deep variational information bottleneck. arXiv preprint arXiv:1612.00410, 2016.

[13] Sun S. Multi-digit mnist for few-shot learning. https://github.com/shaohua0116/MultiDigitMNIST, 2019.