Constructing Superior Representations Beyond the Original Documents via a Contrastive Gaussian Fusion Network for Clustering

Ao Shen[1,2*], **Ruizhang Huang**[1,2†], **Jingjing Xue**[1,2*], **Ruina Bai**[1,2†]

[1]Text Computing & Cognitive Intelligence Engineering Research Center of National Education Ministry, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China

[2]Laboratory of Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China

aoshen5664@163.com, rzhuang@gzu.edu.cn, jingjingxue.gz@gmail.com, rnbai@gzu.edu.cn

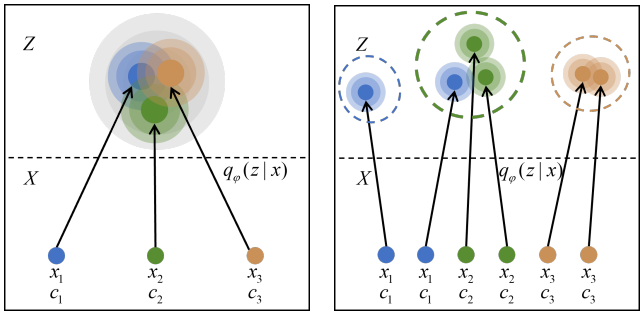
Abstract

Document clustering plays an important role in text mining and information retrieval. Existing methods primarily focus on document-intrinsic features, overlooking dataset-level features and consequently failing to construct superior representations. We propose a **Contrastive Gaussian Fusion Network (CGFN)** that can **construct superior representations beyond the original documents**. Specifically, CGFN fuses the Gaussian distributions of neighbor-derived information and intrinsic textual features in the latent space. By incorporating contrastive learning into the fusion process, our proposed method is able to learn high-quality representations while simultaneously mitigating noise and minimizing information loss. Experiments on four real-world datasets demonstrate that CGFN outperforms state-of-the-art methods, achieving superior clustering by robustly capturing holistic distributions and neighbor patterns.

Introduction

Background Document clustering is a key task in data mining. Variational Autoencoders (VAEs) (Kingma, Welling et al. 2013) enhance this process by mapping inputs to latent probability distributions, thereby facilitating effective similarity discovery and cluster formation.

Motivation However, existing methods, primarily focusing on document-intrinsic features, overlook crucial dataset-level features. This reliance on individual documents for distribution estimation is problematic due to the inherent sparsity and high dimensionality of text data (Bond-Taylor et al. 2021), leading to incomplete latent representations and inaccurate latent distribution estimation. Consequently, they often fail to construct superior representations that capture the full spectrum of semantic and structural information. Recently, incorporating structural information (e.g., neighbor data like document networks or co-occurrence graphs) to augment intrinsic textual document features has emerged as a promising method to enrich representations (Bai et al. 2023).



(a) Posterior collapse (b) Noise and information loss

Figure 1: Challenges in integrating neighbor data. Here, X denotes the input space, Z represents the feature space, and $q_\varphi(z|x)$ is the encoder. The symbols x_i and c_i denote the i-th input and its corresponding category respectively.

Challenge While incorporating neighbor information is promising, integrating such data presents two significant challenges.

First, VAEs are susceptible to posterior collapse (Bowman et al. 2016), where the learned approximate posterior distribution collapses to the prior, failing to capture meaningful variations in the input data. As shown in Figure 1a, x_1, x_2, and x_3 originate from distinct categories but their latent semantic distributions tend to collapse into overlapping configurations. This issue intensifies when incorporating neighbor information, as complex constraints exacerbate optimization pressure. The complex neighbor constraints create a simplistic optimization shortcut where the model can satisfy neighbor requirements by collapsing its representation, rather than learning meaningful data-specific features. Ultimately, this leads to ineffective distribution learning and compromised clustering performance.

Second, while beneficial, current fusion approaches for integrating intrinsic textual and neighbor data often introduce noise or cause critical information loss. This fundamentally hinders modeling interactions between distinct information sources. For example, naively concatenating distribution parameters (like means and variances) fails to capture meaningful cross-modal relationships, highlighting the need for more sophisticated fusion mechanisms. As shown

*These authors contributed equally to this work

†Corresponding authors

in Figure 1b, data points x_1 and x_2 belong to blue class c_1. However, due to missing information, x_2's learned representation lies closer to the green c_2 distribution in feature space and farther from x_1, resulting in its misclassification to c_2 during clustering and adversely affecting performance. Such mechanisms must preserve essential information while enabling coherent distributional modeling.

Contribution To address these challenges and construct superior representations beyond the original documents, we propose the Contrastive Gaussian Fusion Network (CGFN). CGFN is a novel framework designed to enhance document clustering by robustly fusing the Gaussian distributions of neighbor-derived information and intrinsic textual features in the latent space. By incorporating contrastive learning into this fusion process, CGFN effectively learns high-quality representations while simultaneously mitigating noise and minimizing information loss.

Specifically, to combat posterior collapse, CGFN introduces a Consistency Enhancement Module (CEM). This module weakens the KL-divergence regularizer, allowing the posterior distribution to deviate from the prior, thereby protecting weak signals from textual data, enabling the distribution to carry more input-related information, and avoiding excessive compression of informative features. For distributional fusion, our framework introduces a Contrastive Gaussian Fusion Module (CGFM) that effectively integrates intrinsic textual and neighbor data. Through Gaussian fusion, this module minimizes critical information loss, overcoming limitations of traditional fusion strategies. Via contrastive learning, it selectively incorporates intrinsic textual features while filtering neighbor noise. The resulting distribution preserves dominant intrinsic textual semantics while reducing noise propagation, thereby improving fusion fidelity.

The key contributions of this study are as follows:

- We present **CGFN**, a novel deep document clustering framework that addresses distributional information scarcity through distribution-level fusion of intrinsic textual and neighbor data, leading to enhanced latent representations and improved clustering accuracy.
- We design the **Consistency Enhancement Module (CEM)**, which dynamically regulates the disentanglement degree during ELBO optimization to alleviate po·sterior collapse, thereby fostering more accurate and robust distributional learning.
- We introduce a distribution-friendly fusion strategy augmented with a **Contrastive Gaussian Fusion Module (CGFM)**. By leveraging contrastive learning to align distributions and supplement consensus information, this strategy effectively mitigates information loss and noise issues in traditional fusion methods.
- Experiments on four real-world datasets demonstrate that CGFN delivers superior performance compared to existing methods. Its fusion strategy and contrastive learning mechanism effectively preserve distributional characteristics and uncover neighbor patterns, contributing to its strong results.

Related Work

Deep Document Clustering

Deep clustering models leverage deep neural networks to learn feature representations essential for clustering. For instance, Deep Clustering Network (DCN) (Yang et al. 2017) integrates the representation learning capabilities of neural networks with traditional clustering methods (e.g., K-Means (Hartigan and Wong 1979)), gaining widespread attention and application in recent years. Building upon this framework, Deep Embedded Clustering (DEC) (Xie, Girshick, and Farhadi 2016) incorporates a joint learning strategy, where autoencoders learn data embeddings and representations are iteratively refined through clustering center optimization to achieve enhanced clustering accuracy. These models effectively exploit sample feature information to uncover clustering structures, have achieved significant success recently. Structural Deep Clustering Network (SDCN) (Bo et al. 2020) introduces structural information into deep clustering by employing a dual self-supervised mechanism that integrates autoencoder and graph neural network features to enrich latent representations and improve clustering performance. However, the unobservable dimensionality reduction process in neural networks renders decision boundaries and clustering results difficult to interpret visually, thus restricting the explainability of these deep clustering models.

Generative Models for Document Clustering

Variational Autoencoders (VAEs) have emerged as prominent generative models for text clustering, leveraging distributional learning to infer latent structures. Early foundational works, such as VaDE (Jiang et al. 2016) and GM-VAE (Dilokthanakul et al. 2016), notably employed Gaussian Mixture Models (GMMs) as priors for latent space clustering. Subsequent advancements focused on integrating diverse aspects, including local/global structures (DGG (Yang et al. 2019)), hierarchical latent trees (LTVAE (Li et al. 2018)), and multi-VAE ensembles (MVAE (Ye and Bors 2021)). More recent efforts have addressed critical challenges like enhancing scalability (S3VDC (Cao et al. 2020)), exploiting mutual information (DC-VAE (Xu et al. 2020), DFVC (Ji et al. 2021)), and mitigating representation collapse (LIDVAE (Wang, Blei, and Cunningham 2021), LINPC (He et al. 2019a)).

However, these methods rely solely on single-document features to estimate distributions. In high-dimensional sparse text spaces, this leads to incomplete representations (Xu and Durrett 2018), as isolated documents cannot adequately model complex data geometries. Consequently, clustering performance degrades when document distributions are poorly estimated—a fundamental limitation our work addresses through neighbor knowledge fusion.

Posterior Collapse

Posterior collapse in Variational Autoencoders (VAEs) is often addressed by dynamically modulating the Kullback-Leibler (KL) divergence term within the Evidence Lower Bound (ELBO). Common approaches include monotonic

and cyclical KL annealing (Higgins et al. 2017; Alemi et al. 2018; Kim and Mnih 2018; Chen et al. 2018; Fu et al. 2019). The β-VAE introduces a scaling factor β to balance reconstruction quality and latent compression, though using $\beta \neq 1$ violates the strict variational objective(Alemi et al. 2018).

Beyond KL term modulation, other prevention methods include free-bits techniques that introduce thresholds to the KL term(Kingma et al. 2016; Pelsmaeker and Aziz 2019). Architectural modifications, such as applying batch normalization (BN-VAE)(Zhu et al. 2020), using posterior variance dropout(Shen et al. 2021), or incorporating distance constraints (δ-VAE)(Razavi et al. 2019), have also been explored. Furthermore, employing non-Gaussian priors (e.g., von Mises–Fisher(Guu et al. 2018; Davidson et al. 2018) or uniform distributions(Van Den Oord, Vinyals et al. 2017; Zhao, Lee, and Eskenazi 2018)) can decouple KL terms from data instances. Objectives can also be augmented with mutual information constraints, often through aggregated posterior-prior divergences(Zhao, Song, and Ermon 2019) or mutual posterior divergence(Ma, Zhou, and Hovy 2019). Lastly, strategies to address optimization imbalances involve strengthening the encoder (e.g., aggressive optimization(He et al. 2019b), EM training(Li et al. 2023), pretraining(Li et al. 2019)) and weakening the decoder (e.g., non-autoregressive architectures(Semeniuta, Severyn, and Barth 2017) or input masking(Petit and Corro 2021)).

Proposed Model

The network structure of the proposed model is illustrated in Figure 2. The entire architecture of the CGFN model consists of four main parts: Gaussian inference module, contrastive Gaussian fusion module (which contains Gaussian fusion module and contrastive learning module), consistency enhancement module and clustering module.

Gaussian Inference Module

The CGFN model employs two channels to learn representations from intrinsic textual data x and neighbor data x', respectively. Here, x' enhances semantic information by mining neighbor patterns:

$$x'^{(i)} = \frac{1}{k}\sum_{i=1}^{k} f_{knn}(x^{(i)}) \tag{1}$$

We employ the inference network of the VAE framework for each channel. In particular, given data x, the mean and variance of the distribution representations are processed as follows:

$$h = g(Wx + b) \tag{2}$$

$$\mu = W_\mu h + b_\mu \tag{3}$$

$$\log \sigma = W_\sigma h + b_\sigma \tag{4}$$

where h is the hidden representation from inference network, g is the activation function, W and b are weight and bias of inference network respectively. Inference module obtain μ and $\log \sigma$ through two linear neural networks with parameters W_μ, b_μ, W_σ and b_σ.

Contrastive Gaussian Fusion Module

The Contrastive Gaussian Fusion Module (CGFM) comprises two modules. Both modules are instrumental in the fusion process, effectively integrating intrinsic textual and neighboring data.

Gaussian Fusion Module This module primarily performs parameter-level fusion of two distributions from the previous module, yielding a new Gaussian distribution. Using the inference network, x and x' can be converted into latent semantic distributions, specifically $N(\mu, \sigma^2)$ and $N(\mu', \sigma'^2)$. The probability density functions (PDFs) of the distributions $N(\mu, \sigma^2)$ and $N(\mu', \sigma'^2)$, denoted as $f_1(x)$ and $f_2(x')$ respectively, are given as follows:

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5}$$

$$f_2(x') = \frac{1}{\sqrt{2\pi}\sigma'} e^{-\frac{(x-\mu')^2}{2\sigma'^2}} \tag{6}$$

In order to enrich the feature representation of semantic information, the model adopts the method of Gaussian product, which multiplies two Gaussian variables to obtain a new variable z:

$$f(z) = f_1(x) f_2(x') = S_g \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(x-\mu_z)^2}{2\sigma_z^2}} \tag{7}$$

$$\mu_z = \frac{\mu'\sigma^2 + \mu\sigma'^2}{\sigma^2 + \sigma'^2} \tag{8}$$

$$\sigma_z^2 = \frac{\sigma^2\sigma'^2}{\sigma^2 + \sigma'^2} \tag{9}$$

where S_g is a constant, also known as the scale factor, then we get

$$z \sim N(z; \mu_z, \sigma_z^2) \tag{10}$$

To constrain the model to generate z using a Gaussian distribution, the model uses the reparameterization trick, where the random variable ϵ from the standard normal distribution is introduced to obtain the latent variable by $z = \mu_z + \epsilon \odot \sigma_z$ so that the generative process can be back-propagated. This process can be formalized as follows:

$$q_\varphi(z|x, x') = N(z; \mu_z, \sigma_z^2) \tag{11}$$

where $q_\varphi(z|x, x')$ represents the posterior distribution, which is modeled by the encoder.

Contrastive Learning Module This module primarily leverages contrastive learning to extract shared salient information from intrinsic textual and neighbor data, while minimizing information loss and enhancing fusion robustness. For the contrastive part, μ, σ^2, μ', σ'^2 are processed by a projection MLP head, which is denoted as h and transforms one distribution to match the other. Denoting the two output vectors as $\hat{\mu} = h(\mu)$ and $\hat{\sigma} = h(\sigma)$, we minimize their negative cosine similarity:

$$D(\hat{N}_1, N_2) = KL(N(\hat{\mu}, \hat{\sigma}^2) \| N(\mu', \sigma'^2)) \tag{12}$$

Following, we define a symmetrized loss as:

$$\mathcal{L}_{cfm} = \tfrac{1}{2} D(\hat{N}_1, N_2) + \tfrac{1}{2} D(\hat{N}_2, N_1) \tag{13}$$

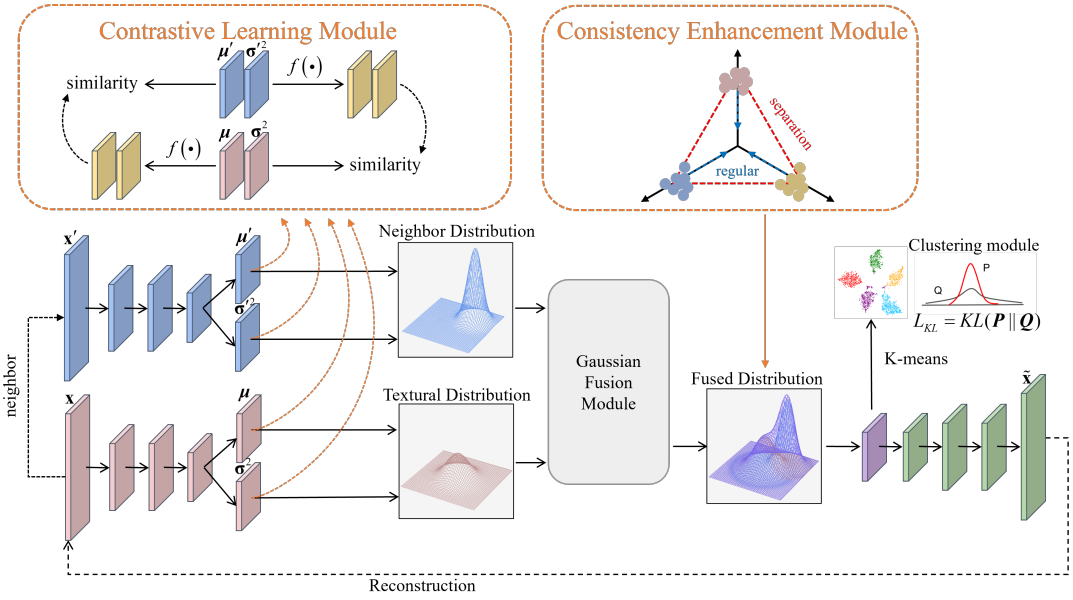


Figure 2: The overall framework of the proposed CGFN.

Consistency Enhancement Module

To mitigate distribution collapse, we propose the consistency enhancement module:

$$\mathcal{L}_{cem} = E_{q_\varphi(z|x,x')}[\log p_\theta(x|z)] - \beta D_{KL}\left[q_\varphi(z|x,x') \,\|\, p_\theta(z)\right] \quad (14)$$

where $p_\theta(x|z)$ denotes the generative decoder reconstructing input x, and $p_\theta(z)$ represents a prior following a standard normal distribution.

The β coefficient is pivotal in balancing latent regularization and reconstruction accuracy. While $\beta > 1$ improves factor disentanglement in image data by emphasizing KL divergence, this approach often compromises reconstruction fidelity. In contrast, textual data demands a distinct strategy: over-regularization via high β values disrupts contextual dependencies by forcing orthogonal latent dimensions, which degrades feature discriminability and clustering stability. To address this, we advocate $\beta < 1$ to retain linguistic structures in the latent space. By reducing KL pressure, our formulation preserves hierarchical semantic-syntactic relationships critical for downstream tasks, thereby preventing distribution collapse without sacrificing informative latent correlations. This approach ensures robust representations while avoiding the fragmentation of coherent textual features.

Combining the objectives from the Consistency Enhancement Module ($\mathcal{L}_{cem}$) and the Contrastive Learning Module ($\mathcal{L}_{cfm}$), the overall loss for the representation learning phase is defined as:

$$\mathcal{L}_{cgfm} = \mathcal{L}_{cem} + \alpha \mathcal{L}_{cfm} \quad (15)$$

Clustering Module

This module forms specific clusters by iteratively optimizing cluster assignments and refining the learned feature space. Through the Gaussian fusion module, the Consistency Enhancement Module (CEM), and the Contrastive Gaussian Fusion Module (CGFM)'s contrastive learning component, the model acquires a feature representation z with semantically enriched information. However, the composite objective function only incorporates reconstruction loss and Gaussian distribution fusion constraints, lacking explicit guidance for clustering structures. To address this, the clustering layer introduces soft assignments q_{ij} derived from Student's t-distribution to model the similarity between samples z_i and cluster centers o_j:

$$q_{ij} = \frac{\left(1 + \frac{\|\mathbf{z}_i - \mathbf{o}_j\|^2}{\nu}\right)^{-\frac{\nu+1}{2}}}{\sum_{j'}\left(1 + \frac{\|\mathbf{z}_i - \mathbf{o}_{j'}\|^2}{\nu}\right)^{-\frac{\nu+1}{2}}} \quad (16)$$

where $\nu = 1$ is the degrees-of-freedom parameter. This formulation adaptively adjusts the similarity measure based on sample-cluster proximity.

To further enhance clustering purity, we compute a target distribution p_{ij} by sharpening the soft assignments. Specifically, q_{ij} is squared and normalized across both samples and clusters to emphasize high-confidence allocations:

$$p_{ij} = \frac{\left(q_{ij}^2 / \sum_i q_{ij}\right)}{\sum_{j'}\left(q_{ij'}^2 / \sum_i q_{ij'}\right)} \quad (17)$$

This double normalization ensures p_{ij} reflects confident cluster assignments while maintaining probability constraints.

Finally, the Kullback-Leibler (KL) divergence between P and Q enforces consistency between the refined target distribution and soft assignments:

$$\mathcal{L}_{kl} = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (18)$$

Algorithm 1: CGFN

Input: input data x, number of neighbors k, maximum iteration M.

Output: clustering results R

```
Construct neighbor dataset x' using the data x via Equation 1;
Initialize CGFN model parameters θ;
Initialize μ with K-means on representation learned by the pretrained CGFN;
for iter ∈ 0, 1, ..., M do
    Use z to compute Q via Equation 16;
    Calculate P via Equation 17;
    Calculate L_cgfm and L_kl via Equation 15 and Equation 18;
    Update the parameters of the model;
end for
Calculate the clustering results R;
return R.
```

Minimizing $\mathcal{L}_{kl}$ iteratively aligns the soft assignments Q with the target distribution P, effectively distilling high-confidence cluster predictions into the feature space and improving clustering coherence. Algorithm 1 shows the training process of the whole model.

Experiments

Experimental Settings

Datasets Four real-world datasets were used to conduct extensive experiments. The statistics of these datasets are summarized in Table 1.

Dataset	Samples	features	Classes
BBC	2225	10000	5
Abstract	4306	10000	3
BBCSports	737	4613	5
Reuters-10k	10000	2000	4

Table 1: Summary of datasets

- The BBC dataset was derived from the BBC News corpus, which comprises 2225 articles published between 2004 and 2005 across five thematic categories: business, entertainment, politics, technology, and sports.
- The Abstract dataset is composed of research paper abstracts published on the Aminer platform, primarily focusing on three academic domains: Information Communication, Databases, and Graphics.
- The BBCSports dataset originates from the BBCSports corpus, comprising 737 articles categorized into five sports domains: athletics, cricket, football, rugby, and tennis.
- Reuters-10k comprises approximately 820000 English news texts labeled with categorical tags. We only use four root categories: corporate, government, market, and economic. Through subsequent multi-stage filtering of these texts, we derived a refined subset of 10,000 documents.

Evaluation Metrics To measure the performance of the clustering task, three popular clustering evaluation metrics are utilized: ACC (Accuracy), NMI (Normalized Mutual Information), and ARI (Average Rand Index). For each evaluation metric, a higher value implies better clustering performance. Denote $Y = y_1, y_2, ..., y_N$ as the ground-truth label and $C = c_1, c_2, ..., c_N$ as the predicted cluster label. The ACC is defined as follows:

$$ACC = \sum_{i}^{N} \frac{y_i = m(c_i)}{N} \tag{19}$$

where $m(c_i)$ is the mapping function that ranges over all possible one-to-one mappings between Y and C.

$$NMI = \frac{I(Y, C)}{max[H(Y) + H(C)]} \tag{20}$$

where $I(Y, C)$ represents mutual information between Y and C, $H(Y)$ and $H(C)$ represents information entropy of Y and C.

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]} \tag{21}$$

where RI is the rand index, which represents the Rand Index, calculated as the ratio of agreeing pairs to the total number of data point pairs, $E[RI]$ denotes the expected Rand Index under random clustering assignments, accounting for statistical chance in label allocations.

Compared Methods This article selected three types of clustering models for comparative experiments, namely traditional clustering models such as K-means (Hartigan and Wong 1979), LDA (Blei, Ng, and Jordan 2003), etc.; deep clustering models such as AE (Hinton and Salakhutdinov 2006), IDEC (Guo et al. 2017a), DCEC (Guo et al. 2017b), VAE (Kingma, Welling et al. 2013), VaDE (Jiang et al. 2016); Structural semantic fusion clustering models such as VGAE (Kipf and Welling 2016), SDCN (Bo et al. 2020), DSEDC (Ren et al. 2023), SEDCN (Bai et al. 2022).

Implementation Details This article uses CGFN for pretraining on each dataset and selects Adam optimization algorithm as the model optimizer. For the BBC and Reuters-10k datasets, the learning rate is set to 0.0005, and for the Abstract and BBCSports datasets, the learning rate is set to 0.001. The dimension of the encoder is set to input d-500-500-128, where d is the dimension of the input data dimension, and the dimension of the decoder is set to 128-500-500-d. Except for the input and output layers, ReLU nonlinear functions are used for activation. The pretraining epochs are set to 50, the maximum iteration M is set to 3000, the convergence threshold is set to 0.1% , and the training batch size is set to 256. Finally, each result is run a total of 10 times and averaged to avoid extreme situations. When searching for neighboring texts through Equation 1, choose the method that includes the variables themselves. In the experiment, the number k of knn searching for neighboring texts is set to 50. The β parameter of $\mathcal{L}_{cem}$ is set to 0.1, and the parameter α of $\mathcal{L}_{cfm}$ is set to 0.1.

Methods	BBC			Abstract			BBCSports			Reuters-10k		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
K-means	51.58	30.88	20.50	69.18	38.26	27.69	55.24	32.44	25.91	54.04	41.54	27.95
LDA	45.66	23.90	18.25	80.19	29.47	36.44	59.63	44.67	32.42	55.46	25.27	26.07
AE	53.60	39.93	19.90	75.56	45.26	39.95	67.16	49.13	29.76	74.90	49.69	49.55
IDEC	83.60	66.56	61.07	88.63	63.89	68.68	73.41	60.52	47.13	73.40	48.51	54.11
DCEC	88.91	75.03	74.78	92.80	73.33	79.42	62.42	64.39	48.08	72.28	51.22	55.60
VAE	65.14	54.06	40.67	81.14	52.82	51.32	63.74	30.12	28.57	61.33	33.20	19.79
VADE	29.65	7.30	6.06	61.79	23.46	27.64	-	-	-	40.22	30.57	-
VGAE	84.13	53.20	57.72	75.43	61.26	67.49	60.11	54.48	28.36	60.85	25.51	26.17
SDCN	77.55	65.28	62.58	93.03	72.90	79.11	78.43	68.29	55.46	**77.15**	50.82	55.36
SDCMS	76.18	64.56	56.51	91.08	70.85	74.30	67.44	59.35	55.75	76.61	51.79	51.05
DSEDC	76.73	61.25	51.06	88.57	63.71	68.32	65.81	58.79	48.52	73.15	53.19	58.02
SEDCN	76.29	73.60	67.55	93.73	75.87	81.72	71.64	65.27	58.43	73.76	55.33	**61.26**
CGFN	**95.77**	**86.76**	**89.71**	**94.59**	**78.68**	**84.31**	**94.44**	**85.79**	**84.61**	75.71	**58.83**	57.11

Table 2: Clustering results on real-world datasets

Methods	BBC			Abstract			BBCSports			Reuters-10k		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CGFN-f	72.07	44.36	43.69	62.22	33.02	31.09	58.87	34.05	26.50	54.20	26.01	22.67
CGFN-r	72.80	51.20	43.87	78.14	45.75	44.64	64.59	50.91	39.45	61.71	43.28	38.19
CGFN-c	81.78	63.45	61.20	87.48	61.25	65.82	74.93	58.40	47.40	72.40	54.09	51.45

Table 3: Ablation results on real-world datasets

Experimental Results and Analysis

Table 2 presents the clustering results of the three types of clustering models across four datasets. The analysis is as follows:

As shown in Table 2, the CGFN model significantly outperforms all comparative models on the NMI metric across all four datasets. This finding strongly demonstrates that CGFN, by enhancing distributional information via Contrastive Gaussian fusion, exhibits a clear advantage over models focusing solely on intrinsic data features. This enhancement enables CGFN to leverage the fused distributions to enrich semantic feature learning, thereby improving clustering performance.

Compared to the best-performing baseline, CGFN achieves NMI improvements of 11.73, 2.81, 17.50, and 3.50 percentage points on the BBC, Abstract, BBCSports, and Reuters-10k datasets, respectively. These gains substantiate the effectiveness of CGFN. This improvement stems from CGFN's explicit consideration of the importance of latent distribution information for clustering. Through its dual-input architecture, CGFN injects structurally augmented information into the decoder layer, leveraging textual neighborhood relationships to enrich semantic learning. Simultaneously, the distribution fusion mitigates the risk of distribution collapse and facilitates clustering from the semantically richer fused distribution, leading to enhanced accuracy.

In the Reuters-10k dataset, certain classes have significantly larger sample sizes than others. Consequently, misclassifications within minority classes contribute less to overall accuracy metrics (ACC/ARI), potentially causing bias due to data imbalance. The NMI metric, however, focuses on the correlation between clustering results and true labels and exhibits lower sensitivity to such imbalance, resulting in stronger robustness. Therefore, NMI is selected as the primary evaluation metric for assessing CGFN's clustering performance.

Ablation Study

To validate the effectiveness of distribution information enhancement and the consistency constraint strategy, this paper conducted tests on four real-world text datasets: BBC, Abstract, BBCSports, and Reuters-10k. The results are presented in Table 3. Three variant models were designed for comparative analysis: CGFN-c, CGFN-r, and CGFN-f. Specifically:

CGFN-c removes the clustering layer from the original model while retaining all other operations and parameter settings.

CGFN-r further eliminates the consistency constraint strategy from CGFN-c, replacing it with the original regularization technique of variational autoencoders (VAE).

CGFN-f additionally abandons the Contrastive Gaussian fusion method in CGFN-r, adopting the traditional single-input encoding approach of VAE.

As shown in Table 3, CGFN-c outperforms CGFN-r without extreme fluctuations in any metrics. This preliminarily confirms that the consistency constraint strategy significantly alleviates distribution collapse and validates its efficacy for clustering tasks. Furthermore, introducing the distribution information enhancement module to CGFN-r substantially improves its performance over CGFN-f. This strongly demonstrates that the enhancement module enriches latent semantic distribution information, thereby facilitating cluster-friendly latent representations.

(a) Clustering Visualization for VAE

(b) Clustering Visualization for our CGFN

Figure 3: Visualization of clustering results from VAE and our CGFN on four datasets: BBC, Abstract, BBCSports, and Reuters-10k. Different colors represent distinct clusters.

Clustering Visualization

To visually compare and evaluate clustering performance, experiments were conducted on four real-world datasets (BBC, Abstract, BBCSports, and Reuters-10k) using a Variational Autoencoder (VAE) with an added clustering layer (no distribution enhancement), and CGFN. t-SNE was employed to visualize high-dimensional clustering results in two dimensions (Figure 3).

CGFN consistently outperformed the standard VAE across all datasets. CGFN produced tighter, denser clusters with sharper inter-cluster boundaries, demonstrating superior discrimination and clustering effectiveness. This significant improvement confirms that enhancing distribution information enables the model (CGFN) to better capture rich semantic features, yielding more discriminative and accurate representations.

CGFN's advantage stems from its unique distribution enhancement mechanism, which fully leverages dataset distribution information during training to learn more accurate and discriminative semantic representations. In contrast, the standard VAE's lack of distribution enhancement resulted in poorer clustering performance and more disordered clusters.

Conclusion

This paper proposes CGFN, a clustering model that leverages contrastive learning to fuse multiple Gaussian distributions from diverse document representations. It addresses distribution collapse and information loss issues in VAE-based clustering when integrating neighbor data. The model introduces a Contrastive Gaussian Fusion Module, which effectively integrates neighbor data, enriching distributional information completeness and accuracy. Additionally, a Consistency Enhancement Module further optimizes the learning process, improves posterior distribution quality, and effectively mitigates distribution collapse. Experiments on four real-world datasets demonstrate that CGFN achieves significant NMI improvements over other clustering models. However, if neighbor and textural data are approximate, Gaussian fusion's effectiveness in integrating distributional information is weakened. Designing clustering-friendly distribution fusion methods is an important direction for future research.

Acknowledgements

We thank the anonymous reviewers for their comments and feedback. This work was supported by the National Key R&D Program of China, No. 2023YFC3304500 and Natural Science Fund of Guizhou University, No. (2024)31.

References

Alemi, A.; Poole, B.; Fischer, I.; Dillon, J.; Saurous, R. A.; and Murphy, K. 2018. Fixing a broken ELBO. In *International conference on machine learning*, 159–168. PMLR.

Bai, R.; Huang, R.; Qin, Y.; Chen, Y.; and Lin, C. 2023. HVAE: A deep generative model via hierarchical variational auto-encoder for multi-view document modeling. *Information Sciences*, 623: 40–55.

Bai, R.; Huang, R.; Zheng, L.; Chen, Y.; and Qin, Y. 2022. Structure enhanced deep clustering network via a weighted neighbourhood auto-encoder. *Neural Networks*, 155: 144–154.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Bo, D.; Wang, X.; Shi, C.; Zhu, M.; Lu, E.; and Cui, P. 2020. Structural deep clustering network. In *Proceedings of the web conference 2020*, 1400–1410.

Bond-Taylor, S.; Leach, A.; Long, Y.; and Willcocks, C. G. 2021. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7327–7347.

Bowman, S.; Vilnis, L.; Vinyals, O.; Dai, A.; Jozefowicz, R.; and Bengio, S. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, 10–21.

Cao, L.; Asadi, S.; Zhu, W.; Schmidli, C.; and Sjöberg, M. 2020. Simple, scalable, and stable variational deep clustering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 108–124. Springer.

Chen, R. T.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31.

Davidson, T. R.; Falorsi, L.; De Cao, N.; Kipf, T.; and Tomczak, J. M. 2018. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*.

Dilokthanakul, N.; Mediano, P. A.; Garnelo, M.; Lee, M. C.; Salimbeni, H.; Arulkumaran, K.; and Shanahan, M. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

Fu, H.; Li, C.; Liu, X.; Gao, J.; Celikyilmaz, A.; and Carin, L. 2019. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*.

Guo, X.; Gao, L.; Liu, X.; and Yin, J. 2017a. Improved deep embedded clustering with local structure preservation. In *Ijcai*, volume 17, 1753–1759.

Guo, X.; Liu, X.; Zhu, E.; and Yin, J. 2017b. Deep clustering with convolutional autoencoders. In *International conference on neural information processing*, 373–382. Springer.

Guu, K.; Hashimoto, T. B.; Oren, Y.; and Liang, P. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6: 437–450.

Hartigan, J. A.; and Wong, M. A. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1): 100–108.

He, J.; Spokoyny, D.; Neubig, G.; and Berg-Kirkpatrick, T. 2019a. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*.

He, J.; Spokoyny, D.; Neubig, G.; and Berg-Kirkpatrick, T. 2019b. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science*, 313(5786): 504–507.

Ji, Q.; Sun, Y.; Gao, J.; Hu, Y.; and Yin, B. 2021. A decoder-free variational deep embedding for unsupervised clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10): 5681–5693.

Jiang, Z.; Zheng, Y.; Tan, H.; Tang, B.; and Zhou, H. 2016. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *International conference on machine learning*, 2649–2658. PMLR.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Kipf, T. N.; and Welling, M. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Li, B.; He, J.; Neubig, G.; Berg-Kirkpatrick, T.; and Yang, Y. 2019. A surprisingly effective fix for deep latent variable modeling of text. *arXiv preprint arXiv:1909.00868*.

Li, X.; Chen, Z.; Poon, L. K.; and Zhang, N. L. 2018. Learning latent superstructures in variational autoencoders for deep multidimensional clustering. *arXiv preprint arXiv:1803.05206*.

Li, Y.; Cheng, L.; Yin, F.; Zhang, M. M.; and Theodoridis, S. 2023. Overcoming posterior collapse in variational autoencoders via EM-type training. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Ma, X.; Zhou, C.; and Hovy, E. 2019. MAE: Mutual posterior-divergence regularization for variational autoencoders. *arXiv preprint arXiv:1901.01498*.

Pelsmaeker, T.; and Aziz, W. 2019. Effective estimation of deep generative language models. *arXiv preprint arXiv:1904.08194*.

Petit, A.; and Corro, C. 2021. Preventing posterior collapse in variational autoencoders for text generation via decoder regularization. *arXiv preprint arXiv:2110.14945*.

Razavi, A.; Oord, A. v. d.; Poole, B.; and Vinyals, O. 2019. Preventing posterior collapse with delta-vaes. *arXiv preprint arXiv:1901.03416*.

Ren, L.; Qin, Y.; Chen, Y.; Bai, R.; Xue, J.; and Huang, R. 2023. Deep structural enhanced network for document clustering. *Applied Intelligence*, 53(10): 12163–12178.

Semeniuta, S.; Severyn, A.; and Barth, E. 2017. A hybrid convolutional variational autoencoder for text generation. *arXiv preprint arXiv:1702.02390*.

Shen, D.; Qin, C.; Wang, C.; Zhu, H.; Chen, E.; and Xiong, H. 2021. Regularizing variational autoencoder with diversity and uncertainty awareness. *arXiv preprint arXiv:2110.12381*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, Y.; Blei, D.; and Cunningham, J. P. 2021. Posterior collapse and latent variable non-identifiability. *Advances in neural information processing systems*, 34: 5443–5455.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Xu, C.; Dai, Y.; Lin, R.; and Wang, S. 2020. Deep clustering by maximizing mutual information in variational autoencoder. *Knowledge-Based Systems*, 205: 106260.

Xu, J.; and Durrett, G. 2018. Spherical Latent Spaces for Stable Variational Autoencoders. *ArXiv*, abs/1808.10805.

Yang, B.; Fu, X.; Sidiropoulos, N. D.; and Hong, M. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, 3861–3870. PMLR.

Yang, L.; Cheung, N.-M.; Li, J.; and Fang, J. 2019. Deep clustering by gaussian mixture variational autoencoders with graph embedding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6440–6449.

Ye, F.; and Bors, A. G. 2021. Deep mixture generative autoencoders. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10): 5789–5803.

Zhao, S.; Song, J.; and Ermon, S. 2019. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, 5885–5892.

Zhao, T.; Lee, K.; and Eskenazi, M. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. *arXiv preprint arXiv:1804.08069*.

Zhu, Q.; Bi, W.; Liu, X.; Ma, X.; Li, X.; and Wu, D. O. 2020. A Batch Normalized Inference Network Keeps the KL Vanishing Away. In *Annual Meeting of the Association for Computational Linguistics*.