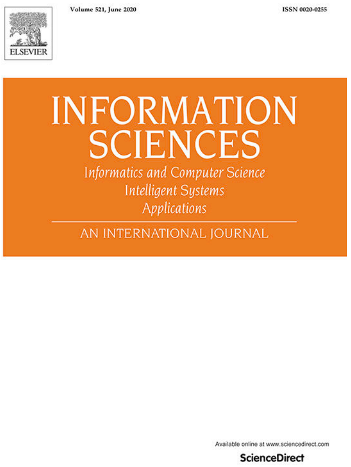


Volume 521, June 2020

ISSN 0020-0255

ELSEVIER

INFORMATION SCIENCES
Informatics and Computer Science
Intelligent Systems
Applications

AN INTERNATIONAL JOURNAL

Available online at www.sciencedirect.com
ScienceDirect

HVAE: A Deep Generative Model via Hierarchical Variational Auto-Encoder for Multi-view Document Modeling

Ruina Bai, Ruizhang Huang, Yongbin Qin, Yanping Chen, Chuan Lin

PII: S0020-0255(22)01173-2
DOI: https://doi.org/10.1016/j.ins.2022.10.052
Reference: INS 18160

To appear in: *Information Sciences*

Received Date: 3 March 2022
Revised Date: 7 October 2022
Accepted Date: 10 October 2022

Please cite this article as: R. Bai, R. Huang, Y. Qin, Y. Chen, C. Lin, HVAE: A Deep Generative Model via Hierarchical Variational Auto-Encoder for Multi-view Document Modeling, *Information Sciences* (2022), doi: https://doi.org/10.1016/j.ins.2022.10.052

HVAE: A Deep Generative Model via Hierarchical Variational Auto-Encoder for Multi-view Document Modeling

Ruina Bai, Ruizhang Huang*, Yongbin Qin, Yanping Chen, Chuan Lin

State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, Guizhou, PR China

Abstract

With the widespread development of the internet, multi-view text documents have become increasingly common, which has led to extensive research on multi-view text document modeling. As opposed to traditional single-view document modeling, which treats each document independently and learns each document as a single topic representation, the views of multi-view text documents have complicated correlation relationships that include both the global and local underlying topical information. In this study, we introduce a deep generative model for multi-view document modeling known as **H**ierarchical **V**ariational **A**uto-**E**ncoder (HVAE), which combines the advantages of the probability generative model for learning interpretable latent information and the deep neural network for efficient parameter inference. Specifically, a set of hierarchical topic representations is learned for each multi-view document to capture the document-level global topical information and view-level local topical information for each view. A two-level hierarchical topic inference network is investigated as the encoder network of HVAE, which is designed using an aligned variational auto-encoder, to learn the hierarchical topic representations. Subsequently, multi-view documents are generated through a two-layered generation network, considering both the view-level local and document-level global topic representations. Ex-

*Corresponding author
Email address: rzhuang@gzu.edu.cn (Ruizhang Huang)

periments on three real datasets of different scales for various tasks demonstrate the satisfactory results of the proposed method.

Keywords: deep variational inference, variational auto-encoder, multi-view document modeling, hierarchical topic representation, probability generative model

1. Introduction

With the rapid development of Internet technology, text documents have become ubiquitous sources of information. Text modeling, which is a fundamental task in text mining, is a common tool for analyzing unlabeled text documents [1, 2]. Among all text-modeling methods, probabilistic generative models have received significant research attention owing to their success in learning the underlying interpretable text document structures in terms of topics and generating text documents to support more downstream tasks, such as document classification [3, 4, 5]. However, traditional probabilistic generative models suffer from the high computational complexity of the inference process [6, 7]. The underlying text topic variables are usually underestimated, with a limited ability to model large numbers of text documents. Owing to the rapid development of deep neural networks in various fields, the ability to learn models from a large amount of data has improved. In recent years, increasing research has been conducted on illustrating that deep neural networks substantially improve large-scale text modeling. The basic concept underlying these models is to construct a deep inference network based on the content features of text documents to approximate the distribution of the latent topic representation for each document [8, 9]. All of these models provide a solid basis for modeling traditional single-view text documents. Only the content features are considered during the inference and generative processes of the models.

In reality, multi-view [10, 11] text documents have become increasingly common. These text documents are described not only by their traditional content information, but also by other useful document description aspects such as the document propagation view and high-level semantic topic view. For example, a web document can be represented by both the intrinsic and extrinsic views. The intrinsic view describes the web document using the traditional web content features. The extrinsic view, which is collected from the inbound hyperlinks of the web document, describes the interrelationships of the web document with other web pages [12]. News articles can be expressed by the traditional content view as well as the propagation view, which contains its propagation behavior features such as the readers and forwarders of the news articles. Text documents with multi-view representations have more complete information than single-view documents. Each view is used to represent different aspects of text documents, which contain different but correlated information. Therefore, there is an urgent need to build a deep generative model to make good use of all multi-view information for modeling text documents.

As multi-view text documents have only received substantial attention in recent years, few studies have been conducted on multi-view text modeling. All existing methods were designed to model single-view data. The application of traditional single-view text-modeling approaches to model multi-view text documents is unrealistic. This is because they simply treat each document independently and learn each document as a single topic representation. However, the views of multi-view text documents have significantly more complicated relationships. First, each view contains local information. The features of different document views include specific styles, meanings, and intentions. As a result, the topics of each document view are not exactly the same, but normally have different focuses. For example, the topics of the news article content view focus on genres or a specific set of news events described in the news. The topics of the web article propagation view focus more on groups of persons who like, read, or forward news articles. It is necessary to retain these view-level local characteristics when modeling multi-view text documents. However, different document views contain global information. The features of each document view are correlated according to their underlying document-level global topics. For example, the web text propagation and content views are correlated because people are willing to propagate news articles when they are interested in them. It is common for a person to propagate web articles with similar web content topics. Therefore, in the task of multi-view document modeling, it is necessary to capture the underlying document-level global topics for each document that are shared by all document views.

In this study, we present a deep generative model for multi-view document modeling, namely **H**ierarchical **V**ariational **A**uto-**E**ncoder (HVAE). Owing to the success of the VAE in learning meaningful topical information for single-view text documents, we investigate the VAE framework for multi-view text document modeling. The HVAE model combines the advantages of both probability generative models, which learn interpretable latent information for modeling text documents, and deep neural networks, which can effectively conduct parameter inferences. A hierarchical topic inference network is employed to learn the view-level topic representations and a consistent document-level topic representation for each text document by jointly considering both the local and global information of document views. On the first level of hierarchical inference, the VAE is investigated to learn the view-level local-topic representations. Subsequently, an alignment module is employed on the second level of the hierarchical inference network to learn the document-level global topic representation for each text document, which contains global features that are shared by all document views. We investigate the alignment module with the attention network to adjust the contribution of each view-level topic representation to the document-level topic representation automatically. Multi-view text documents are then generated by considering both the view-level local and document-level global topic representations through a two-layered generation network. Our main contributions are summarized as follows:

- We propose a deep generative model, which is designed using the HVAE, for multi-view text document modeling. A set of hierarchical topic representations can be learned for each multi-view text document, which reveals the underlying complicated topic structures of multi-view text documents.

- We introduce an aligned VAE to learn a set of the view-level local and document-level global representations for each multi-view text document.
- Extensive experiments were conducted with the proposed model on real datasets to compare the text document modeling approaches. The experimental results demonstrate the effectiveness of the HVAE model.

The remainder of this paper is organized as follows: Section 2 reviews related work on deep generative models. In Section 3, we describe the proposed model in detail. Section 4 presents the experimental results and an analysis. Finally, our conclusions and future works are presented in Section 5.

2. Related Work

Deep variational generation model. The task of employing neural networks to investigate the learning of variational latent distributions for generative models can be traced back to the past century. Within the context of Helmholtz machines [13, 14], inference networks were trained to approximate the variational latent distribution. However, the application of these directed generative models faces the problem of establishing low-variance gradient estimators. Several researchers have suggested various means of alleviating this problem, such as reparameterizing the continuous random variables [8, 9] and investigating the control variates [15]. The concepts proposed by [16, 17] have exhibited significant performance in image generation.

In recent years, numerous works have been conducted on developing neural variational inference for document modeling. Most of these studies focused on generative auto-encoders and their variants. Neural variational document modeling (NVDM) [18] and VAE-RNN for text modeling [19] were designed based on the VAE framework. The main difference between these two models is the different network structures of the encoder and decoder modules. Owing to the different structural properties, generative models are designed with different characteristics. To avoid latent variable collapses in the modeling process effectively, a holistic regularization VAE (HR-VAE) [20] was proposed, which achieved substantially more stable performance. [21] combined probabilistic topic models, such as the latent Dirichlet allocation (LDA) model [22], using variational inferences to discover discrete latent topics. By dividing the input data into multiple latent spaces, [23] used the VAE as the basic framework to learn multiple latent space-based entity representations for data generation. [24] introduced an improved VAE for document modeling with topical information that was explicitly modeled as a Dirichlet latent variable. APo-VAE [25] investigated document modeling in a hyperbolic latent space to learn continuous hierarchical representations. [26] introduced a generative model based on the VAE to explain observations by separating particularity and commonality. The above model is used for document modeling by capturing the latent topic semantics or hierarchical information. Unfortunately, these models are designed for single-view data. None of them can be directly applied to multi-view data because The views of the data are treated separately and the consistent topical information among the views cannot be captured.

Generative framework with multiple views. The VAE [8] is a powerful deep generative framework for document modeling. To the best of our knowledge, few studies have applied the VAE to multi-view document modeling. In the following section, research works on VAE-based models on various related research tasks are introduced; in particular, the multi-channel/multi-encoder VAE and VAE-based multi-modal model. A multi-channel/multi-encoder VAE has been presented to learn the hierarchical representations for single-view data. [27] used a multi-encoder VAE in single cell image analysis to extract transform-invariant biologically meaningful features. [28] investigated multi-channel VAE for the joint analysis of heterogeneous data. [29] introduced a patch-based multi-channel VAE model that enables great diversity in video generation. VAE-based multi-modal models are used to handle data samples with modality inputs. In [4], a VAE was used to capture the complex features between nodes and networks. The main concept is that node and network features are input into the networks together using two groups of latent VAE parameters. M^2VAE [30] integrates the information from different sensor modalities into a joint latent representation to learn epistemic active sensing. DHVAE [31] uses a disentangled VAE strategy to separate the private and shared latent spaces of multiple modalities. [32] and [33] used semi-supervised and learned joint posterior distributions, respectively, to improve the multi-VAE generation performance. The generation module of AHVAE [3] considers the differences between multi-modal models. It shares the second layer of the encoder to obtain consistent information across the modalities, whereas the other parts retain the relatively independent VAE model. A similar framework-based VAE has been used extensively in many tasks, such as text classification [3, 4, 5], multi-modal human dynamics [34], and multi-track symbolic music [35]. Moreover, numerous probabilistic generative models have been proposed. To capture the hierarchical structure of multi-view data, [36] proposed a probabilistic generative model by modeling the multi-view and multi-feature data under a hierarchical structure through a latent variable. Using good uncertainty estimates and great generalization capability, several studies have extended the Gaussian process to multi-view data with different tasks. A two-stage model named MvDGP [37] was proposed to integrate the complementary information from multiple views to discover a good representation of the data. A multi-view probabilistic model known as LCBM [38] was designed for multi-label classification, where a latent variable in a shared subspace serves as the link between multiple views. LCBM can obtain a multi-view fusion representation that fully exploits the complementarity and consistency of different views in the latent semantic subspace. [39] proposed a multi-view VSGP model based on VSG, which can encode beliefs on the consensus of views in the approximation procedure.

Few studies have investigated deep generative models for multi-view document modeling. The only one that is closely related to our work is the MVRL [40] model, which was designed to learn a shared latent representation for multi-view data. However, the MVRL model only focuses on learning document-level global information across views and neglects the local characteristics of each data view. In this study, we investigated a deep generative model for a multi-view document modeling problem. The deep generative model needs to capture the local specific topic within each view as well as the global document topic for correlation among multiple views simultaneously.

3. HVAE

In this section, the proposed HVAE model is introduced. A two-level hierarchical inference network is investigated to learn topics from multi-view text documents. On the first level of the inference network, a view-level topic representation is learned for each single-text document view to capture its local focus. On the second level, an alignment module is used to learn a consistent document-level topic representation from all view-level topic representations. Both the view-level and document-level topic representations are used to reconstruct the multi-view documents during the generation process. In Section 3, the VAE framework for document modeling is first described, because it serves as the basic component of the HVAE model. Thereafter, the HVAE model is described in detail.

3.1. Basic Framework: VAE

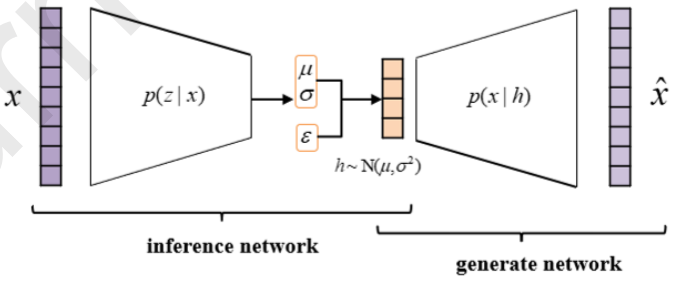


Figure 1: Network structure of VAE.

The VAE has been studied extensively for the generation of single-view text documents in recent years. As a typical latent variable model, the generation process of VAE is divided into two steps: first, the latent topic variables $\boldsymbol{h}$ are learned for each text document, and then, the corresponding text document is generated according to $\boldsymbol{h}$ by $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{h};\boldsymbol{\theta})p(\boldsymbol{h})d\boldsymbol{h}$, where $\boldsymbol{\theta}$ is a parameter that must be learned during the generation. The VAE framework is composed of an inference network and a generative network. The structure of the VAE is shown in Figure 1.

The data flow of the VAE inference network is as follows:

$$\begin{aligned} \boldsymbol{z} &= g_e(\boldsymbol{W}_e\boldsymbol{x} + \boldsymbol{b}_e) \\ \boldsymbol{\mu} &= \boldsymbol{W}_\mu\boldsymbol{z} + \boldsymbol{b}_\mu \\ \log\boldsymbol{\sigma} &= \boldsymbol{W}_\sigma\boldsymbol{z} + \boldsymbol{b}_\sigma, \end{aligned} \tag{1}$$

where $\boldsymbol{z}$ denotes the output of the hidden layer in the inference network, $\boldsymbol{x}$ is the bag-of-words vector representation of the document, g_e is the activation function, and $\boldsymbol{W}_e$ and $\boldsymbol{b}_e$ are the weight and bias parameters, respectively, of the inference network. To constrain the model to generate $\boldsymbol{h}$ using a Gaussian distribution, the VAE uses the reparameterization trick [8, 9], where the random item ϵ from the standard normal distribution is introduced to obtain the latent topic variable by $\boldsymbol{h} = \boldsymbol{\mu} + \epsilon \times \boldsymbol{\sigma}$ so that the generative process can be propagated back. This process can be formalized as follows:

$$\begin{aligned} q(\boldsymbol{h}|\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{h}|\boldsymbol{\mu}, diag(\boldsymbol{\sigma}^2)) \\ \boldsymbol{h} &\sim q(\boldsymbol{h}|\boldsymbol{x}), \end{aligned} \tag{2}$$

where $q(\boldsymbol{h}|\boldsymbol{x})$ is the approximated distribution of the target topic distribution $p(\boldsymbol{h})$, $q(\boldsymbol{h}|\boldsymbol{x})$ is typically depicted by an isotropic Gaussian distribution that is parameterized by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, and $\boldsymbol{\mu}$ and $\log\boldsymbol{\sigma}$ can be obtained using two linear neural networks that are parameterized by $\boldsymbol{W}_\mu$, $\boldsymbol{b}_\mu$, $\boldsymbol{W}_\sigma$, and $\boldsymbol{b}_\sigma$. The document generation process is conducted using the VAE generative network. A topic representation $\boldsymbol{h}$ is sampled from the posterior $q(\boldsymbol{h}|\boldsymbol{x})$. Thereafter, the output document $\hat{\boldsymbol{x}}$ is reconstructed from $\boldsymbol{h}$.

The loss function for training the VAE is designed by jointly considering the loss of the reconstructed document and the KL regularization of $q(\boldsymbol{h}|\boldsymbol{x})$ and

$p(\boldsymbol{h})$, as follows:

$$L_{vae} = \mathbb{E}_{q(\boldsymbol{h}|\boldsymbol{x})} \sum_{i=1}^{N} logp(\boldsymbol{x}_i|\boldsymbol{h}) - D_{KL}[q(\boldsymbol{h}|\boldsymbol{x})||p(\boldsymbol{h})] \tag{3}$$

where $\boldsymbol{x}_i$ denotes the one-hot representation of the i-th word in document $\boldsymbol{x}$ and $p(\boldsymbol{x}_i|\boldsymbol{h})$ is the probability of generating $\boldsymbol{x}_i$ given the topic representation $\boldsymbol{h}$. Note that $p(\boldsymbol{x}_i|\boldsymbol{h})$ is different for the various designs of the generative process. In this study, we follow the concept of NVDM [18] and reconstruct the documents by independently generating $\boldsymbol{x}_i$ through a multi-nomial softmax network in the VAE decoder. Based on sample $\boldsymbol{h}$, L_{vae} can be optimized via back-propagation. [41] provides the proof in detail.

3.2. HVAE Model

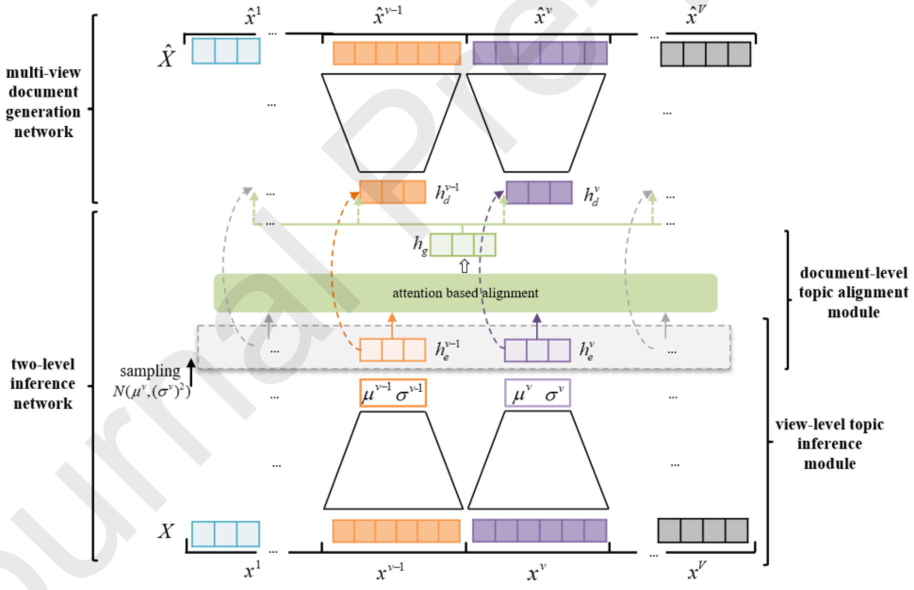

$\hat{x}^1$
$\hat{x}^{v-1}$
$\hat{x}^v$
$\hat{x}^V$
$\hat{X}$
multi-view document generation network
h_d^{v-1}
h_d^v
h_g
document-level topic alignment module
attention based alignment
h_e^{v-1}
h_e^v
two-level inference network
sampling
$N(\mu^v,(\sigma^v)^2)$
μ^{v-1} σ^{v-1}
μ^v σ^v
view-level topic inference module
X
x^1
x^{v-1}
x^v
x^V

Figure 2: Model structure of HVAE (V views).

The network structure of the proposed model is illustrated in Figure 2. The entire architecture of the HVAE model consists of three main parts: the view-level topic inference module, document-level topic alignment module, and multi-view document generation network. The view-level topic inference module is employed to learn a local topic representation for each view of the text document. Thereafter, the document-level topic alignment module is deployed to learn a consistent topic representation with global topic features that are shared by all views. Note that the view-level topic inference module and document-level topic alignment module form the two-level hierarchical inference network of the HVAE model. The multi-view document generation network is designed to generate multi-view documents by considering the global and local topic representations.

3.2.1. View-level topic inference module

The view-level topic inference module is composed of V view channels. Each view channel is used to process a single document view input to learn its view-level local topic representation. We employ the inference network of the VAE framework for each view channel. In particular, given view $\boldsymbol{x}^v$ of a text document, it is processed by the view channel as follows:

$$\begin{aligned}
\boldsymbol{z}^v &= g_e^v(\boldsymbol{W}_e^v \boldsymbol{x}^v + \boldsymbol{b}_e^v) \\
\boldsymbol{\mu}^{\boldsymbol{v}} &= \boldsymbol{W}_\mu^v \boldsymbol{z}^v + \boldsymbol{b}_\mu^v \\
\log \boldsymbol{\sigma}^{\boldsymbol{v}} &= \boldsymbol{W}_\sigma^v \boldsymbol{z}^v + \boldsymbol{b}_\sigma^v \qquad (4) \\
q^v(\boldsymbol{h}_e^v | \boldsymbol{x}^v) &= \mathcal{N}(\boldsymbol{h}_e^v | \boldsymbol{\mu}^{\boldsymbol{v}}, diag(\boldsymbol{\sigma}^{\boldsymbol{v}})) \\
\boldsymbol{h}_e^v &\sim q^v(\boldsymbol{h}_e^v | \boldsymbol{x}^v),
\end{aligned}$$

where g_e^v is the activation function for the encoding module on view v with weight parameters $\boldsymbol{W}_e^v$ and bias parameters $\boldsymbol{b}_e^v$. Furthermore, $\boldsymbol{\mu}^v$ and $\log \boldsymbol{\sigma}^v$ can be obtained using two linear neural networks that are parameterized by $\boldsymbol{W}_\mu^v$, $\boldsymbol{b}_\mu^v$, $\boldsymbol{W}_\sigma^v$, and $\boldsymbol{b}_\sigma^v$. $q^v(\boldsymbol{h}_e^v | \boldsymbol{x})$ is the approximated distribution of the target topic distribution $p^v(\boldsymbol{h}_e^v)$. $\boldsymbol{h}_e^v$ is the latent view-level topic representation of document view $\boldsymbol{x}^v$. Moreover, we generate $\boldsymbol{h}_e^v$ for each view by relying on ϵ^v from the standard Gaussian distribution using the reparameterization trick.

This process is performed for each view of the text document. As a result, a set of latent view-level local topic representations $\boldsymbol{H}_e = \{\boldsymbol{h}_e^v\}_{v=1}^V$ is obtained.

Subsequently, $\boldsymbol{H}_e$ is processed by the document-level topic alignment network to learn a consistent document-level global topic representation to capture the correlations among multiple views.

3.2.2. Document-level topic alignment module

The document-level topic alignment network is developed to align various view-level local topic representations into a consistent document-level global topic representation. The document-level topic alignment module is designed as follows:

1) Transform $\boldsymbol{H}_e$ into $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$:

$$\begin{aligned} \boldsymbol{K} &= \boldsymbol{W}^k \boldsymbol{H}_e, \boldsymbol{K} = \{\boldsymbol{k}_v\}_{v=1}^{V} \\ \boldsymbol{Q} &= \boldsymbol{W}^q \boldsymbol{H}_e, \boldsymbol{Q} = \{\boldsymbol{q}_v\}_{v=1}^{V} \\ \boldsymbol{V} &= \boldsymbol{W}^\nu \boldsymbol{H}_e, \boldsymbol{V} = \{\boldsymbol{\nu}_v\}_{v=1}^{V}. \end{aligned} \tag{5}$$

2) For each document view v, obtain the attention weights and representation, respectively:

$$\begin{aligned} s_i &= sim(\boldsymbol{q}_v, \boldsymbol{k}_i), i \in \{1, 2, .., V\} \\ \lambda_i &= \frac{exp(s_i)}{\sum_{i'} exp(s_{i'})}, i \in \{1, 2, .., V\} \end{aligned} \tag{6}$$

$$\boldsymbol{\tau}_v = \sum_{i}^{V} \lambda_i * \boldsymbol{\nu}_i. \tag{7}$$

3) Assign the alignment result to $\boldsymbol{h}_g$ as follows:

$$\boldsymbol{h}_g = \sum_{v} \boldsymbol{\tau}_v. \tag{8}$$

In particular, the inferred document view representation $\boldsymbol{H}_e$ is first transformed into 3 feature spaces: $\boldsymbol{K}$, $\boldsymbol{Q}$, and $\boldsymbol{V}$ using the transformation weight parameters $\boldsymbol{W}^k$, $\boldsymbol{W}^q$, and $\boldsymbol{W}^\nu$. Subsequently, $\boldsymbol{K}$, $\boldsymbol{Q}$, and $\boldsymbol{V}$ are used to learn the attention assignment for each document view, which determines the contribution of each view-level topic representation to the document-level global topic representation $\boldsymbol{h}_g$. It should be noted that there are several options for setting the *sim* function.

In our study, we employ the basic dot product to estimate the similarity between different view-level topic representations.

3.2.3. *Two-layered multi-view document generation network*

Each document view is generated using a two-layer network in the multi-view generation network. In the first layer of the generation network, a view-level generative topic representation $\boldsymbol{h}_d^v$ is generated by the document-level topic representation $\boldsymbol{h}_g$ and view-level topic representation $\boldsymbol{h}_e^v$. $\boldsymbol{h}_d^v$ contains both the document-level global topical and view-level local information. The generation process is designed as follows:

$$\boldsymbol{h}_d^v = \rho(\boldsymbol{h}_g, \boldsymbol{h}_e^v), \tag{9}$$

where ρ is a fusion strategy that combines $\boldsymbol{h}_g$ and $\boldsymbol{h}_e^v$. Arbitrary designs of ρ may exist. In this study, we employ two basic skip connection strategies, which were used by [42, 43, 44], for the design of ρ. The first strategy is the linear mapping of $\boldsymbol{h}_d^v$ with $\boldsymbol{h_g}$ and $\boldsymbol{h}_e^v$, which is formulated as

$$\rho(\boldsymbol{h}_g, \boldsymbol{h}_e^v) = \alpha_g^v \times \boldsymbol{h}_g + \alpha_e^v \times \boldsymbol{h}_e^v, \tag{10}$$

where α_g^v and α_e^v are the fusion parameters of the v-th view and both are initialized to 0.5. The second strategy is the concatenation of $\boldsymbol{h}_g$ and $\boldsymbol{h}_e^v$, which is formulated as $\rho(\boldsymbol{h}_g, \boldsymbol{h}_e^v) = [\boldsymbol{h}_g, \boldsymbol{h}_e^v]$. Thus, the features of the different dimensions are retained.

In the second layer, $\boldsymbol{h}_d^v$ is used to generate documents with multiple views. For each word representation $\boldsymbol{x}_i^v$ in the v-th view of text document $\boldsymbol{x}$, the generation process is as follows:

$$\hat{\boldsymbol{x}}_i^v = \frac{\exp\{-\boldsymbol{h}_d^v \boldsymbol{W}_d^v \boldsymbol{x}_i^v\}}{\sum_j \exp\{-\boldsymbol{h}_d^v \boldsymbol{W}_d^v \boldsymbol{x}_j^v\}}, \tag{11}$$

where $\boldsymbol{x}_j^v$ is the one-hot representation of each word and $\hat{\boldsymbol{x}}_i^v$ is the output of the model indicating the reconstruction representation of $\boldsymbol{x}_i^v$.

According to $p(\boldsymbol{x}^v, \boldsymbol{h}_e^v, \boldsymbol{h}_g) = p(\boldsymbol{x}^v|\boldsymbol{h}_e^v, \boldsymbol{h}_g)p(\boldsymbol{h}_e^v, \boldsymbol{h}_g)$, the training objective of the HVAE model is to maximize the lower bound of the marginal likelihood presented in Equation 12.

$$\begin{aligned} L_{mv} = \sum_{v\in V} \mathbb{E}_{q(\boldsymbol{h}_e^v, \boldsymbol{h}_g|\{\boldsymbol{x}^v\})} \sum_{i=1}^{N^v} \log p(\boldsymbol{x}_i^v|\boldsymbol{h}_e^v, \boldsymbol{h}_g) \\ - D_{KL}[q(\boldsymbol{h}_e^v, \boldsymbol{h}_g|\{\boldsymbol{x}^v\})||p(\boldsymbol{h}_e^v, \boldsymbol{h}_g)], \end{aligned} \tag{12}$$

where N^v is the number of words in the v-th view of document $\boldsymbol{x}$. Furthermore, $p(\boldsymbol{h}_e^v, \boldsymbol{h}_g)$ is the target topic distribution of the v-th document view. Our HVAE models the document of each view simultaneously to ensure that the data features are not lost during the generation process.

4. Experiments

4.1. Datasets

Three real-world document datasets were used to conduct extensive experiments. A summary of the statistics for all datasets is presented in Table 1.

Table 1: Summary of datasets.

Dataset	# of samples	# of views	# of words	# of topics
Aminer	4320	2	[2355, 1589]	3
BBCSports	544	2	[3183, 3203]	5
DoubanMovies	21587	2	[15000, 12000]	11

The *Aminer* dataset was derived from the ***Aminer***[1] corpus, which is a research paper corpus that is mainly used in social network research. Each paper is associated with its abstract, authors, year, venue, and title. The textitAminer dataset was developed by randomly selecting $4,320$ papers from 3 manually labeled research areas:“infocoms,” “database,” and “graphics.” Each paper is represented in 2 views: the author and content views. The author view contains the authors’ names collected from the current paper and its references.

[1] https://www.aminer.cn/data

The content view is collected from the abstract. The vocabulary sizes of these two views are $2,355$ and $1,589$.

The *BBCSports* dataset was derived from the ***BBCSports*** corpus[2], which is a synthetic multi-view text corpus that was generated by dividing news articles into related segments. We selected 2 views from the *BBCSports* corpus to construct the *BBCSports* dataset. It contains 544 data samples. The vocabulary sizes of the two selected views are $3,183$ and $3,203$.

We collected the *DoubanMovies* dataset from a movie website[3]. The dataset contains $21,587$ movie samples that were randomly selected from 11 topics. Each movie is described by 2 views: the actor view and content view. The vocabulary sizes of the two views are $15,000$ and $12,000$.

4.2. Experimental Settings

We compared the proposed HVAE model with three strong deep generation models using the VAE for document modeling: VAE ("VAE"), NVDM[4] ("NVDM"), and GSM[5] ("GSM"). Experiments using the VAE model were conducted as a baseline. The NVDM and GSM models were investigated as they are deep models using VAE, which achieve promising performance in document modeling. We set the variational inference network of the HVAE model to 2 fully connected layers. The 2 neural layers had 200 and $1,000$ hidden units, respectively, with the ReLU activation function. The dimensions of these linear layers were used to parameterize $\boldsymbol{\mu}^v$, $\boldsymbol{\sigma}^v$, $\boldsymbol{h}_e^v$, $\boldsymbol{h}_g$ and were all set with the number of topics, denoted by K. The setting of $\boldsymbol{h}_d^v$ was K in HVAE-LM. In the HVAE-CA setting, the dimension of $\boldsymbol{h}_d^v$ was set to $2 \times K$ because it concatenated $\boldsymbol{h}_e^v$ and $\boldsymbol{h}_g$. Each view of the multi-view datasets was treated as a single dataset when conducting the experiments for all comparison models. In contrast, the proposed HVAE model processed all multi-views of each dataset using a unique

[2] http://mlg.ucd.ie/datasets/segment.html

[3] https://movie.douban.com/

[4] https://github.com/ysmiao/nvdm

[5] https://github.com/linkstrife/NVDM-GSM

procedure.

We evaluated the performance of the document modeling using the perplexity (PP), which is computed as follows:

$$PP = \exp(-\frac{1}{N}\sum_{i}\frac{1}{L_i}\log p(\boldsymbol{x}_i)), \tag{13}$$

where N is the number of document samples, L_i represents the length of the i-th document sample, and $p(\boldsymbol{x})$ is the generation probability of document sample $\boldsymbol{x}$. We followed the settings of [15] and used the variational lower bound to compute the perplexity. Note that a smaller value of perplexity indicates better performance.

However, perplexity as a metric of the language model for text modeling does not cover all valid information in the modeling process. To reveal the structure of the topics, we explored the structure of each view using the clustering accuracy (ACC) and normalized mutual information (NMI) metrics [45, 46, 47]. Given data x_i, let c_i and y_i be the obtained cluster label and label provided by the corpus, respectively. The ACC is defined as follows:

$$ACC = \sum_{i}^{n}\frac{\delta(y_i, map(c_i))}{n}, \tag{14}$$

where n is the total number of data, $\delta(x, y)$ is the indicator function that is equal to one if $x = y$ and zero otherwise, and $map(c_i)$ is the permutation mapping function that maps each cluster label c_i to the equivalent label from the data using the Hungarian algorithm. The NMI between the label set Y and cluster set C is defined as follows:

$$NMI(Y, C) = \frac{MI(Y, C)}{\sqrt{H(Y)H(C)}}, \tag{15}$$

where $MI(Y, C)$ is the mutual information between Y and C, $H(\cdot)$ is the entropy, and the denominator $\sqrt{H(Y)H(C)}$ is used to normalize the mutual information in the range $[0, 1]$. Both metrics are scaled from 0 to 1, with higher values indicating that the clustering labels match the distribution of the true topics more closely.

4.3. Experimental Results on Multi-view Document Modeling

4.3.1. Overall experimental performance of HVAE model

Two different fusion strategies were investigated for ρ to learn $\boldsymbol{h}_d^v$ in the multi-view generation network of the proposed HVAE model. In the first strategy, we linearly mapped $\boldsymbol{h}_d^v$ with $\boldsymbol{h}_g$ and $\boldsymbol{h}_e^v$, which was denoted as HVAE-LM. $\boldsymbol{h}_d^v$ was learned by concatenating $\boldsymbol{h}_g$ and $\boldsymbol{h}_e^v$ as the second strategy, which was denoted as HVAE-CA. These two models are referred to as the HVAE-related models.

Table 2: Comparison of document modeling performance of all models on all datasets evaluated by perplexity (PP), ACC (%), and NMI (%). The best results are indicated in **bold.**

		VAE			NVDM			GSM			HVAE-LM			HVAE-CA		
		PP	ACC	NMI	PP	ACC	NMI	PP	ACC	NMI	PP	ACC	NMI	PP	ACC	NMI
Aminer	*author*	387	94.35	74.08	648	88.89	70.47	783	94.16	75.94	407	**96.98**	**85.92**	**338**	94.74	76.57
	content	601	92.60	72.20	658	92.63	73.18	644	91.78	70.68	573	**94.03**	**76.20**	**567**	93.07	73.42
BBCSports	*view1*	1296	71.09	64.37	1285	42.34	25.45	1453	51.25	39.53	**1212**	**88.44**	**74.04**	1243	69.22	63.58
	view2	1337	71.41	55.04	1360	36.80	16.68	1525	38.47	21.17	**1236**	**80.70**	**65.36**	1250	67.97	56.22
DoubanMovies	*storyline*	3370	20.73	1.95	3435	21.36	1.67	4416	20.31	1.83	3393	22.24	2.79	**3290**	**22.86**	**3.02**
	actor	**2313**	23.44	4.19	2930	20.68	1.69	4205	22.30	2.41	2372	**27.86**	**7.23**	2598	27.54	7.06

The experimental results demonstrate that the HVAE-related models generally performed better than the other models. Both the HVAE-LM and HVAE-CA models achieved promising results for all evaluation metrics. Document-level topic representation is useful for capturing the overall underlying document topics. The improvements of our proposed HVAE-related models were all obvious in the *BBCSports* experiments because the 2 views in the *BBCSports* dataset were strongly correlated. Therefore, learning these correlated features by aligning the view-level topic representation was helpful for modeling the document views. The reconstructed view-level representation h_v^d of the HVAE-related models could also emphasize the document-level topic features and specific view-level features. However, the perplexity performance of the HVAE-related models was slightly less than that of the VAE model on the actor view of the *DoubanMovies* dataset. The main reason for this is that *DoubanMovies* is a relatively hard dataset that contains unrelated document views with a large number of noise features. First, the *actor* and *storyline* views of the *DoubanMovies* dataset were not closely correlated. Actors in movies usually have a wide range of film genres. It is also common for a movie to have numerous choices to settle on its actor list. The lower correlation of views degraded the performance of learning common features to enhance the topic representation. Second, the *DoubanMovies* dataset contains a large number of document features for each document view collected from more than 10 document categories. According the ACC and NMI performances of all models on the *DoubanMovies* dataset, the underlying document structure of each document view was unclear, which indicates considerable noise in each document view representation. A large amount of noise increases the inconsistency between the evaluation purpose of perplexity, for generating datasets with high confidence, and that of ACC and NMI, for discovering topics with more clarity. Moreover, it reduces the meaningfulness of the perplexity results because the generation of noise features with high confidence is not a good indicator of the model capability. As a result, the perplexity performance of our HAVE-related models was inferior because our proposed HAVE model places more emphasis on capturing the underlying document- and view-level topical representations. Despite the above difficulties, we can still observe from Table 2 that the difference between the performance achieved by the HVAE-LM and VAE models was below 60, which can be neglected. However, the ACC and NMI performances of the HVAE-related models obviously outperformed those of the other models, indicating that learning document-level global topics is useful.

The performances of the HVAE-LM and HVAE-CA models were similar. Therefore, the HVAE model is robust to the fusion strategy ρ. We can employ an arbitrary design of ρ and achieve stable performance using the HVAE model. We also employed the HVAE-CA model as an example to evaluate the effectiveness of the HVAE model.

Table 3 presents three categories of the *content* views of the *Aminer* dataset learned by the HVAE-CA model. Each category was described by the top 10

Table 3: View-level topic semantics learned by HVAE-CA model on *content* view of *Aminer* dataset.

Topic	Topic words
infocoms	mobil, metadata, multimedia, multiplex, synchron, machin, relai, technolog, asynchron, devic
graphics	geograph, bitmap, fractal, symmetr, art, brush, thin, lightpath, stochast, grid
database	summar, stack, magnitud, sql, categor, event, mapreduc, userspecifi, overload, list

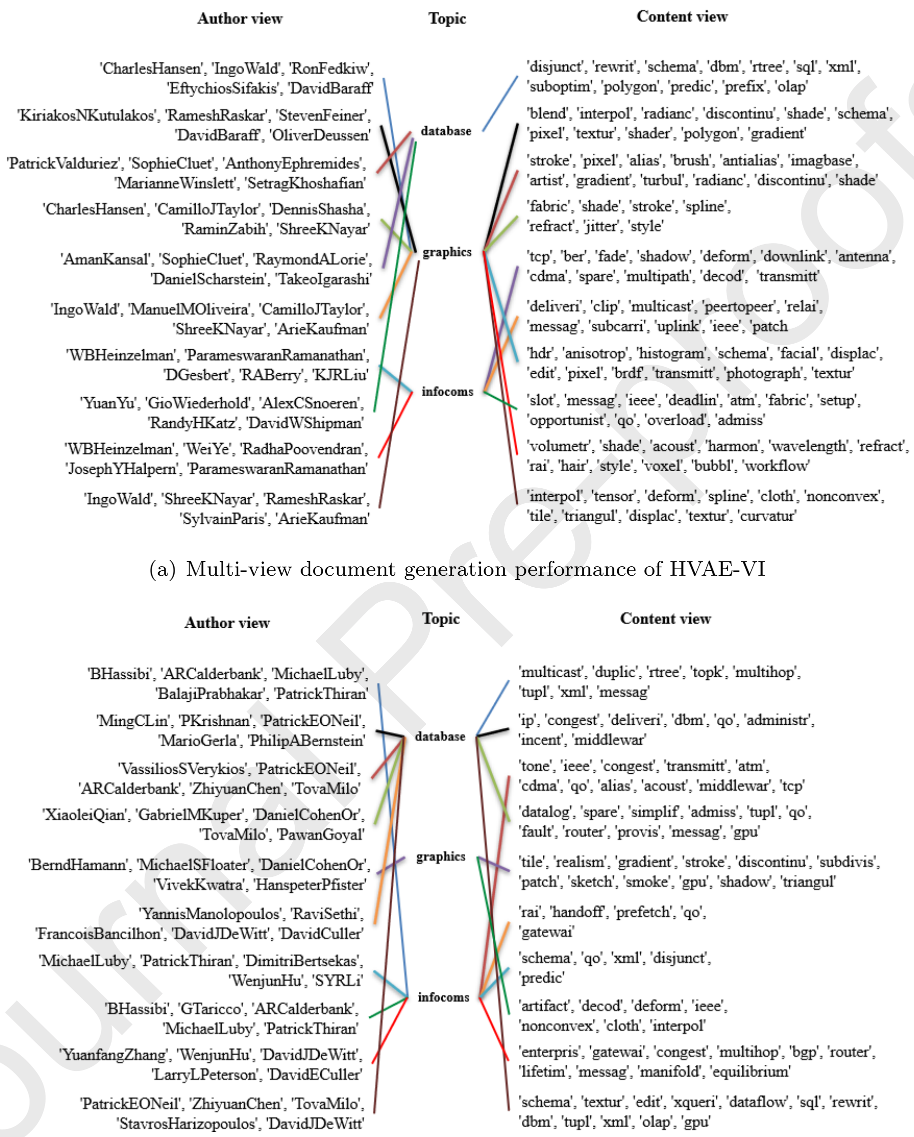
words that had the strongest connection with the underlying view-level topic. It can easily be observed from Table 3 that all selected words were closely related to their underlying topics. For example, the words “mobil,” “metadata,” and “multimedia” were related to the topic “infocoms,” whereas the words “geograph,” “bitmap,” and “fractal” referred to the topic “graphics.” Furthermore, the words “summar,” “stack,” and “magnitud” were related to the topic “database.” It is obvious that the underlying topics were deduced, which are consistent with the 3 manually labeled research areas of the *Aminer* dataset. Therefore, the HVAE model can learn the interpretable structure of each document view. Thus, we can easily apply it to topic mining by capturing the specific view-level topics of multiple views.

4.3.2. Generation of multi-view documents

We also conducted experiments to evaluate the effectiveness of the HVAE model in generating multi-view documents. First, different views of documents were employed to depict the different aspects of the documents. Second, views of the same documents were correlated with one another through shared topical information. Two variants of the HVAE model were investigated. The first model was the HVAE-CA model, which generated multi-view documents with the aid of the two-level hierarchical inference process. The second model, HVAE-VI, generated multi-view documents using a one-level inference process. Only the view-level topic inference module was employed to learn and generate the view-level representation of the document sample. The document-level topic alignment module was omitted. Note that the HVAE-VI model was degraded to the traditional VAE model, where a set of VAE models was employed to learn all view-level representations, without considering the correlations among multiple views.

The performances of the HVAE-VI and HVAE-CA models are shown in Figure 3. The parameters of the two models were estimated using the *Aminer* dataset. For each model, 10 multi-view documents were randomly generated. Each view was described using several randomly generated words. In the figure, each view and its topic are linked by a line. The color of the line is used to depict the document interrelationship between the views. The views of the same multi-view document are described by the same line color.

It can be observed from Figure 3 that the HVAE model generated meaningful document views. The words in each view reflected their local view characteristics. The words in the *author* view were obviously author names of the research papers. The words in the *content* view could be used to illustrate the content meaning of a research paper. A comparison of the experimental results of the HVAE-VI model, depicted in Figure 3(a), and the HVAE-CA model, depicted in Figure 3(b), reveals that the two-level hierarchical inference process was useful for capturing the correlative information of multi-view documents indicated by its underlying topic. Views of the same multi-view document were more likely to have the same topic. Specifically, 6 out of 10 multi-view documents generated by the HVAE-CA model were described by the *author* and *content* views on the same topic. For the HVAE-VI model, only 3 documents were depicted by views on the same topic. The reason that several multi-view documents were generated by the HVAE-CA model with different view topics is that the HVAE-CA model concatenates the view- and document-level representations. A strong view-level topic representation may overwhelm the effect of document-level topic representation, leading to different view-level topic estimations. Another interesting observation is that the HVAE model can discover the interrelationship


Author view
Topic
Content view
'CharlesHansen', 'IngoWald', 'RonFedkiw', 'EftychiosSifakis', 'DavidBaraff'
'KiriakosNKutulakos', 'RameshRaskar', 'StevenFeiner', 'DavidBaraff', 'OliverDeussen'
'PatrickValduriez', 'SophieCluet', 'AnthonyEphremides', 'MarianneWinslett', 'SetragKhoshafian'
'CharlesHansen', 'CamilloJTaylor', 'DennisShasha', 'RaminZabih', 'ShreeKNayar'
'AmanKansal', 'SophieCluet', 'RaymondALorie', 'DanielScharstein', 'TakeoIgarashi'
'IngoWald', 'ManuelMOliveira', 'CamilloJTaylor', 'ShreeKNayar', 'ArieKaufman'
'WBHeinzelman', 'ParameswaranRamanathan', 'DGesbert', 'RABerry', 'KJRLiu'
'YuanYu', 'GioWiederhold', 'AlexCSnoeren', 'RandyHKatz', 'DavidWShipman'
'WBHeinzelman', 'WeiYe', 'RadhaPoovendran', 'JosephYHalpern', 'ParameswaranRamanathan'
'IngoWald', 'ShreeKNayar', 'RameshRaskar', 'SylvainParis', 'ArieKaufman'
database
graphics
infocoms
'disjunct', 'rewrit', 'schema', 'dbm', 'rtree', 'sql', 'xml', 'suboptim', 'polygon', 'predic', 'prefix', 'olap'
'blend', 'interpol', 'radianc', 'discontinu', 'shade', 'schema', 'pixel', 'textur', 'shader', 'polygon', 'gradient'
'stroke', 'pixel', 'alias', 'brush', 'antialias', 'imagbase', 'artist', 'gradient', 'turbul', 'radianc', 'discontinu', 'shade'
'fabric', 'shade', 'stroke', 'spline', 'refract', 'jitter', 'style'
'tcp', 'ber', 'fade', 'shadow', 'deform', 'downlink', 'antenna', 'cdma', 'spare', 'multipath', 'decod', 'transmitt'
'deliveri', 'clip', 'multicast', 'peertopeer', 'relai', 'messag', 'subcarri', 'uplink', 'ieee', 'patch
'hdr', 'anisotrop', 'histogram', 'schema', 'facial', 'displac', 'edit', 'pixel', 'brdf', 'transmitt', 'photograph', 'textur'
'slot', 'messag', 'ieee', 'deadlin', 'atm', 'fabric', 'setup', 'opportunist', 'qo', 'overload', 'admiss'
'volumetr', 'shade', 'acoust', 'harmon', 'wavelength', 'refract', 'rai', 'hair', 'style', 'voxel', 'bubbl', 'workflow'
'interpol', 'tensor', 'deform', 'spline', 'cloth', 'nonconvex', 'tile', 'triangul', 'displac', 'textur', 'curvatur'
Author view
Topic
Content view
'BHassibi', 'ARCalderbank', 'MichaelLuby', 'BalajiPrabhakar', 'PatrickThiran'
'MingCLin', 'PKrishnan', 'PatrickEONeil', 'MarioGerla', 'PhilipABernstein'
'VassiliosSVerykios', 'PatrickEONeil', 'ARCalderbank', 'ZhiyuanChen', 'TovaMilo'
'XiaoleiQian', 'GabrielMKuper', 'DanielCohenOr', 'TovaMilo', 'PawanGoyal'
'BerndHamann', 'MichaelSFloater', 'DanielCohenOr', 'VivekKwatra', 'HanspeterPfister'
'YannisManolopoulos', 'RaviSethi', 'FrancoisBancilhon', 'DavidJDeWitt', 'DavidCuller'
'MichaelLuby', 'PatrickThiran', 'DimitriBertsekas', 'WenjunHu', 'SYRLi'
'BHassibi', 'GTaricco', 'ARCalderbank', 'MichaelLuby', 'PatrickThiran'
'YuanfangZhang', 'WenjunHu', 'DavidJDeWitt', 'LarryLPeterson', 'DavidECuller'
'PatrickEONeil', 'ZhiyuanChen', 'TovaMilo', 'StavrosHarizopoulos', 'DavidJDeWitt'
database
graphics
infocoms
'multicast', 'duplic', 'rtree', 'topk', 'multihop', 'tupl', 'xml', 'messag'
'ip', 'congest', 'deliveri', 'dbm', 'qo', 'administr', 'incent', 'middlewar'
'tone', 'ieee', 'congest', 'transmitt', 'atm', 'cdma', 'qo', 'alias', 'acoust', 'middlewar', 'tcp'
'datalog', 'spare', 'simplif', 'admiss', 'tupl', 'qo', 'fault', 'router', 'provis', 'messag', 'gpu'
'tile', 'realism', 'gradient', 'stroke', 'discontinu', 'subdivis', 'patch', 'sketch', 'smoke', 'gpu', 'shadow', 'triangul'
'rai', 'handoff', 'prefetch', 'qo', 'gatewai'
'schema', 'qo', 'xml', 'disjunct', 'predic'
'artifact', 'decod', 'deform', 'ieee', 'nonconvex', 'cloth', 'interpol'
'enterpris', 'gatewai', 'congest', 'multihop', 'bgp', 'router', 'lifetim', 'messag', 'manifold', 'equilibrium'
'schema', 'textur', 'edit', 'xqueri', 'dataflow', 'sql', 'rewrit', 'dbm', 'tupl', 'xml', 'olap', 'gpu'

(a) Multi-view document generation performance of HVAE-VI

(b) Multi-view document generation performance of HVAE-CA

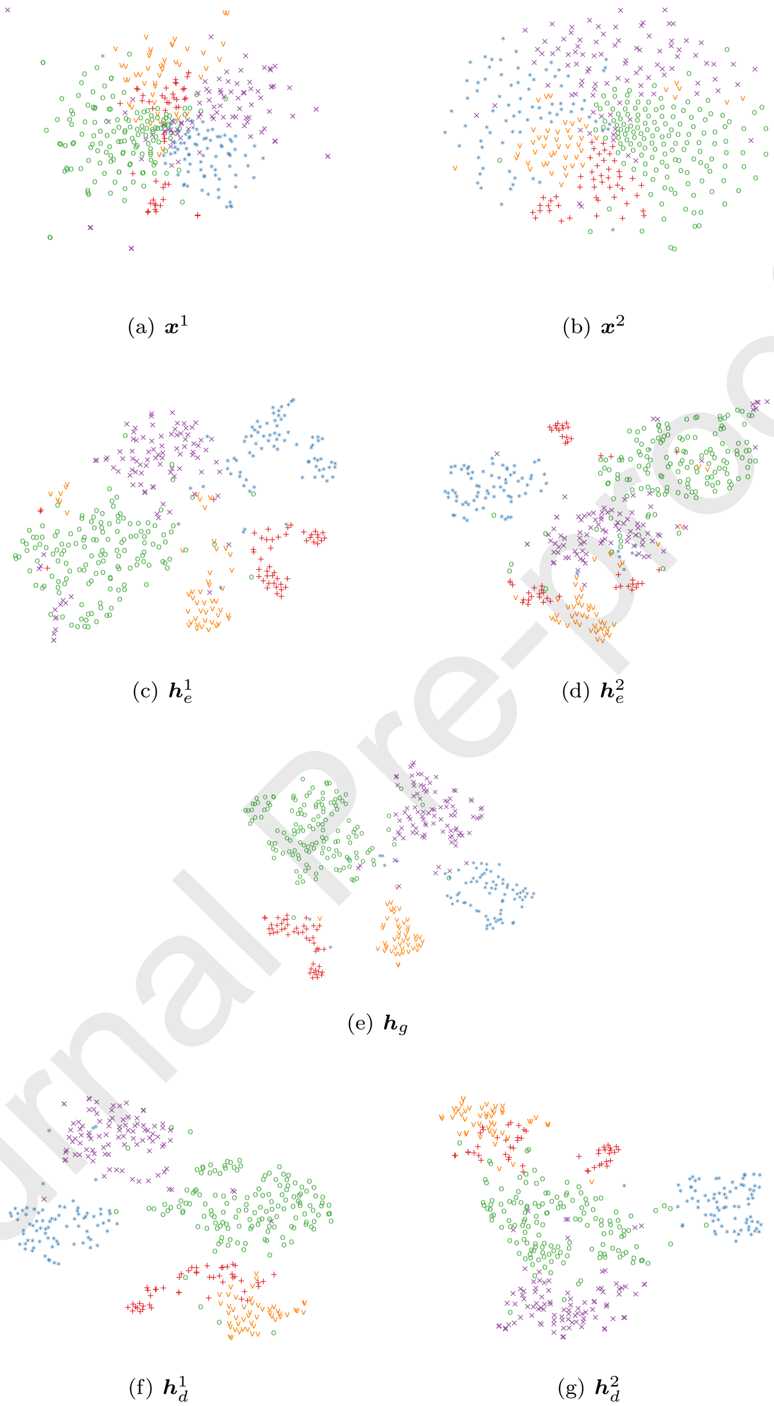
Figure 3: Generation of HVAE-VI and HVAE-CA models for *Aminer*.

between the features of different views according to their common latent topic. For example, as illustrated in Figure 3, the author "MingCLin" is a researcher who focuses on the research area "database," and is likely to write research papers with the keywords "dbm," "administrator," and "middleware."

4.3.3. Visualization

We visualized the document structures using t-SNE [48] on the *BBCSports* dataset. Note that the *BBCSports* dataset is composed of 2 document views: *view1* and *view2*. The visualization results of the HVAE model on the *BBCSports* dataset are shown in Figure 4. The figures in the first line depict the document structure obtained using the two original view inputs $\boldsymbol{x}^v (v \in \{1,2\})$ of HVAE. The figures in the second line depict the document structure obtained by the two view-level topic representations $\boldsymbol{h}_e^v (v \in \{1,2\})$ learned by the view-level topic inference module. The figure in the third line depicts the global document-level topic representation $\boldsymbol{h}_g$ obtained using the document-level topic alignment module. The figures in the last line show the view-level generative topic representations $\boldsymbol{h}_d^v$ learned using the HVAE-CA model.

As illustrated in Figure 4, the two-level hierarchical inference process of the HVAE model clearly improved the learning of the underlying document topic. Each view of the document described by the original input was loose. Document samples with different topics were interwoven in both views. It was difficult to identify the topic boundaries. Using the view-level topic inference module, the topics described by $\boldsymbol{h}_e^1$ and $\boldsymbol{h}_e^2$ were clearer with their local view characteristics. Therefore, the view-level local topic representation could capture the specific topical information. However, the boundaries between topics were narrow and some noise data samples were located in the center of other topics, which led to the topics on the second line of Figure 4 remaining unidentifiable. As indicated in the third line of Figure 4, the topic structure was more evident in the document-level global topic representation $\boldsymbol{h}_g$. The cluster property of the document-level topic representation $\boldsymbol{h}_g$ was clearer. The documents were clearly separated according to their underlying topics. Based on the fourth line



(a) $\boldsymbol{x}^1$

(b) $\boldsymbol{x}^2$

(c) $\boldsymbol{h}_e^1$

(d) $\boldsymbol{h}_e^2$

(e) $\boldsymbol{h}_g$

(f) $\boldsymbol{h}_d^1$

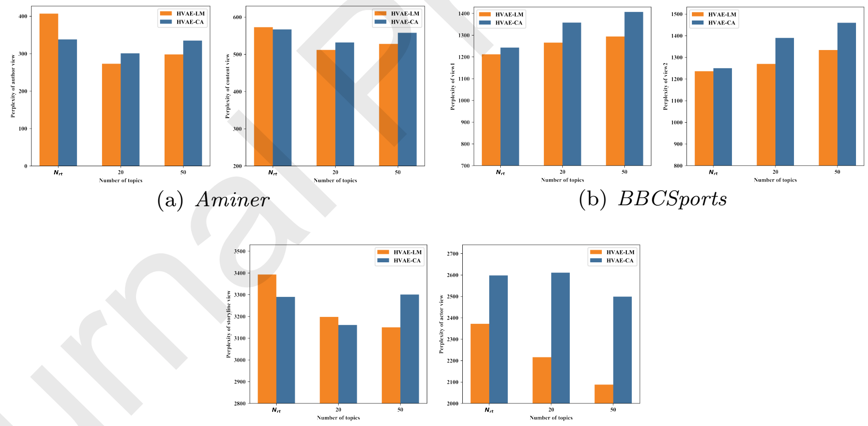
(g) $\boldsymbol{h}_d^2$

Figure 4: Visualization of view-level local representation and document-level global view on *BBCSports*. Different colors indicate different topics.

of Figure 4, the view-level generated topic representations $\boldsymbol{h}_d^1$ and $\boldsymbol{h}_d^2$ combined both the local and global topic representations of each document view. The document structures were better than those described by $\boldsymbol{h}_e^1$ and $\boldsymbol{h}_e^2$ but were slightly worse than those of $\boldsymbol{h}_g$. The involvement of the local topic representation introduced noise information that degraded the overall document structure discovery. However, the local features of each local topic representation were useful for reconstructing document views to retain the local characteristics that were not shared by other document views.

The document-level global representation h_g exhibited clear structural information, which is consistent with the aim of multi-view document clustering. The remainder of this paper discusses the model robustness. We focused on all of the HVAE-related models.

4.3.4. Sensitivity of parameter K



(a) *Aminer*

(b) *BBCSports*

(c) *DoubanMovies*

Figure 5: Effect of number of topics K on HVAE-LM and HVAE-CA on *AMiner*, *BBCSports*, and *DoubanMovies* datasets.

There are few parameters to be investigated in our HVAE model. One typical parameter that needs to be analyzed is the number of topics K that affects the dimension of all latent topic representations, particularly $\boldsymbol{h}_e^v$, $\boldsymbol{h}_g$, and $\boldsymbol{h}_d^v$. We explored the document modeling performance of our proposed HVAE model by changing the value of K for both the HVAE-LM and HVAE-CA models. All experimental datasets, namely *Aminer*, *BBCSports*, and *DoubanMovies*, were used to conduct the experiments. The experimental results are presented in Figure 5. The value of K was set in the range of $\{N_{rt}$,20,50$\}$, where N_{rt} is the real number of clusters in the dataset.

Further analysis of the features of the different views demonstrated that, for views that are rich in semantic information, such as the content view of *Aminer*, there was little difference between the two supplementary strategies with HVAE, particularly in the case of a small number of topics. For other views, the result was dependent on the type of data. In general, the HVAE-related models introduced the topic alignment module to learn the document-level global topic representation and used two supplementary strategies for modeling, which was far less dependent on the number of topics than that of the VAE model. According to Figure 5, K generally achieved the best performance with the true value of the number of topics N_{rt} in HVAE-LM. Increasing the value of K allowed the HVAE-related model to retain more detailed semantics for all underlying representations. However, larger values of K also introduced more noise information, which in turn degraded the document modeling performance. However, the opposite was true for HVAE-CA.

4.3.5. Convergence Analysis

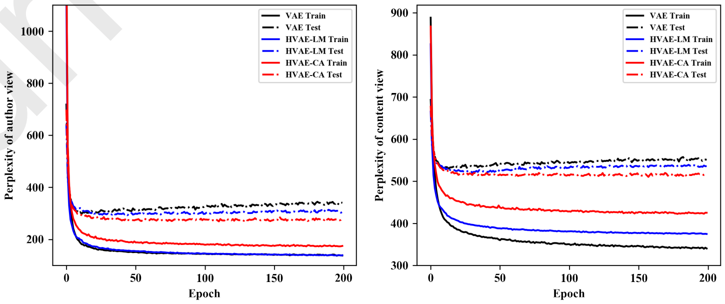


Figure 6: Trace plot for document modeling performances, evaluated by perplexity value, of HVAE-related models and VAE models on *Aminer* dataset with increasing number of epochs.

We investigated the convergence of the HVAE-related models with K=20. The performances of all HVAE-related models, evaluated by the perplexity value and with an increasing number of iterations on the *Aminer* dataset, are depicted in Figure 6.

The performance of the VAE model is also presented for comparison. The solid lines represent the experimental performance of the training set. The dotted line indicates the performance of the test set. In the first several iterations, the values of all models decreased quickly and gradually became flat. Both the VAE- and HVAE-related models exhibited a consistent convergence trend. With a small number of epochs (the number of epochs was 20 in our experiments), all HVAE-related models converged to a stably good result. Moreover, the HVAE-related models generally achieved much better perplexity performance than the VAE model for both the training and testing processes.

4.3.6. Classification task

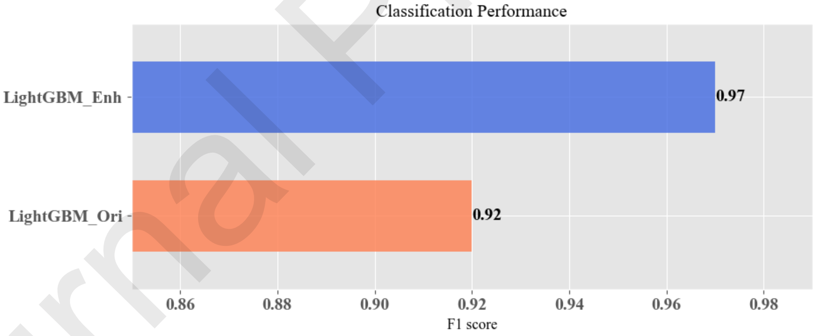


Figure 7: F1 score of LightGBM classifier with and without aid of HVAE model.

Furthermore, we verified the quality of the generated multi-view documents using the simple classifier LightGBM [49]. We used the *Aminer* dataset to conduct the experiment, where the classifier was trained using 300 randomly selected document samples and tested using another 100 randomly selected document samples. The original LightGBM classifier was denoted as “LightGBM_Ori” and the performance of LightGBM_Ori was 0.92 as evaluated by the

F1 score. We used the HVAE-CA model to generate a new set of 300 document samples to aid the classifier training, which was denoted by "LightGBM_Enh". The classification performance improved to 0.97. The proposed HVAE model can be used to expand the training datasets for multi-view classification tasks based on its generative capacity, thereby further improving the classification performance. Therefore, the multi-view data generated through the HVAE model have certain authenticity and can provide support for more downstream tasks.

5. Discussion and Future Work

The proposed HVAE is aimed at topic alignment based on the attention mechanism for view-level local topic variables that are sampled from different view variational inferences. A two-level hierarchical topic inference network was employed and investigated using an aligned VAE to learn the view-level topic representations and consistent document-level topic representations for each text document. This guaranteed that the variational inference model followed the individual inference of each view to avoid introducing noise.

According to the structure of multi-view text documents, the latent subspace of each view is not unique, and the proposed model may provide an incomplete representation of the latent space or focus on the local feature representation. In future work, we will improve the model from two aspects: (1) the introduction of a multi-head [50] variational encoder to explore the latent subspaces for each view, and (2) allowing the model to learn the representation dimension of each view independently to obtain the most appropriate view generation features. This may alleviate the drawbacks of inconsistent convergence.

References

[1] R. Alghamdi, K. Alfalqi, A survey of topic modeling in text mining, Int. J. Adv. Comput. Sci. Appl.(IJACSA) 6 (1).

[2] Y. Li, D. Jiang, R. Lian, X. Wu, C. Tan, Y. Xu, Z. Su, Heterogeneous latent

topic discovery for semantic text mining, IEEE Transactions on Knowledge and Data Engineering.

[3] F. Huang, X. Zhang, C. Li, Z. Li, Y. He, Z. Zhao, Multimodal network embedding via attention based multi-view variational autoencoder, in: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, 2018, pp. 108–116.

[4] H. Li, H. Wang, Z. Yang, M. Odagaki, Variation autoencoder based network representation learning for classification, in: Proceedings of ACL 2017, Student Research Workshop, 2017, pp. 56–61.

[5] Z. Xie, S. Ma, Dual-view variational autoencoders for semi-supervised text matching., in: IJCAI, 2019, pp. 5306–5312.

[6] M. J. Beal, Variational algorithms for approximate bayesian inference, Ph.D. thesis, UCL (University College London) (2003).

[7] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, L. K. Saul, An introduction to variational methods for graphical models, Machine learning 37 (2) (1999) 183–233.

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

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

[10] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv preprint arXiv:1304.5634.

[11] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, Information Fusion 38 (2017) 43–54.

[12] S. Bickel, T. Scheffer, Multi-view clustering., in: ICDM, Vol. 4, 2004, pp. 19–26.

[13] G. E. Hinton, R. S. Zemel, Autoencoders, minimum description length and helmholtz free energy, in: Advances in neural information processing systems, 1994, pp. 3–10.

[14] G. E. Hinton, P. Dayan, B. J. Frey, R. M. Neal, The" wake-sleep" algorithm for unsupervised neural networks, Science 268 (5214) (1995) 1158–1161.

[15] A. Mnih, K. Gregor, Neural variational inference and learning in belief networks, in: International Conference on Machine Learning, PMLR, 2014, pp. 1791–1799.

[16] J. Ba, R. R. Salakhutdinov, R. B. Grosse, B. J. Frey, Learning wake-sleep recurrent attention models, in: Advances in Neural Information Processing Systems, 2015, pp. 2593–2601.

[17] K. Gregor, I. Danihelka, A. Graves, D. Rezende, D. Wierstra, Draw: A recurrent neural network for image generation, in: International conference on machine learning, PMLR, 2015, pp. 1462–1471.

[18] Y. Miao, L. Yu, P. Blunsom, Neural variational inference for text processing, in: International conference on machine learning, 2016, pp. 1727–1736.

[19] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, S. Bengio, Generating sentences from a continuous space, arXiv preprint arXiv:1511.06349.

[20] R. Li, X. Li, C. Lin, M. Collinson, R. Mao, A stable variational autoencoder for text modelling, arXiv preprint arXiv:1911.05343.

[21] Y. Miao, E. Grefenstette, P. Blunsom, Discovering discrete latent topics with neural variational inference, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2410–2419.

[22] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.

[23] C. Nash, S. A. Eslami, C. Burgess, I. Higgins, D. Zoran, T. Weber, P. Battaglia, The multi-entity variational autoencoder, in: NIPS Workshops, 2017.

[24] Y. Xiao, T. Zhao, W. Y. Wang, Dirichlet variational autoencoder for text modeling, arXiv preprint arXiv:1811.00135.

[25] S. Dai, Z. Gan, Y. Cheng, C. Tao, L. Carin, J. Liu, Apo-vae: Text generation in hyperbolic space, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 416–431.

[26] W. Wang, J. Bao, S. Guo, Neural generative model for clustering by separating particularity and commonality, Information Sciences 589 (2022) 813–826.

[27] L. Ternes, M. Dane, S. Gross, M. Labrie, G. Mills, J. Gray, L. Heiser, Y. H. Chang, Me-vae: Multi-encoder variational autoencoder for controlling multiple transformational features in single cell image analysis, bioRxiv.

[28] L. Antelmi, N. Ayache, P. Robert, M. Lorenzi, Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data, in: International Conference on Machine Learning, PMLR, 2019, pp. 302–311.

[29] S. Gur, S. Benaim, L. Wolf, Hierarchical patch vae-gan: Generating diverse videos from a single sample, Advances in Neural Information Processing Systems 33 (2020) 16761–16772.

[30] T. Korthals, D. Rudolph, J. Leitner, M. Hesse, U. Rückert, Multi-modal generative models for learning epistemic active sensing, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 3319–3325.

[31] M. Lee, V. Pavlovic, Private-shared disentangled multimodal vae for learning of latent representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1692–1700.

[32] M. Wu, N. Goodman, Multimodal generative models for scalable weakly-supervised learning, Advances in Neural Information Processing Systems 31.

[33] S. Nedelkoski, M. Bogojeski, O. Kao, Learning more expressive joint distributions in multimodal variational methods, in: International Conference on Machine Learning, Optimization, and Data Science, Springer, 2020, pp. 137–149.

[34] X. Yan, A. Rastogi, R. Villegas, K. Sunkavalli, E. Shechtman, S. Hadap, E. Yumer, H. Lee, Mt-vae: Learning motion transformations to generate multimodal human dynamics, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 265–281.

[35] X. Liang, J. Wu, J. Cao, Midi-sandwich2: Rnn-based hierarchical multi-modal fusion generation vae networks for multi-track symbolic music generation, arXiv preprint arXiv:1909.03522.

[36] J. Li, H. Yong, B. Zhang, M. Li, L. Zhang, D. Zhang, A probabilistic hierarchical model for multi-view and multi-feature classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[37] S. Sun, W. Dong, Q. Liu, Multi-view representation learning with deep gaussian processes, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (12) (2020) 4453–4468.

[38] S. Sun, D. Zong, Lcbm: a multi-view probabilistic model for multi-label classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (8) (2020) 2682–2696.

[39] L. Mao, S. Sun, Multiview variational sparse gaussian processes, IEEE Transactions on Neural Networks and Learning Systems 32 (7) (2020) 2875–2885.

[40] H. Hwang, G.-H. Kim, S. Hong, K.-E. Kim, Multi-view representation learning via total correlation objective, Advances in Neural Information Processing Systems 34.

[41] C. Doersch, Tutorial on variational autoencoders, arXiv preprint arXiv:1606.05908.

[42] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[43] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, 2014, pp. 1188–1196.

[44] J. Turian, L. Ratinov, Y. Bengio, Word representations: a simple and general method for semi-supervised learning, in: Proceedings of the 48th annual meeting of the association for computational linguistics, 2010, pp. 384–394.

[45] R. Guan, H. Zhang, Y. Liang, F. Giunchiglia, L. Huang, X. Feng, Deep feature-based text clustering and its explanation, IEEE Transactions on Knowledge and Data Engineering.

[46] M. R. H. Rakib, N. Zeh, M. Jankowska, E. Milios, Enhancement of short text clustering by iterative classification, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2020, pp. 105–117.

[47] R. Bai, R. Huang, Y. Chen, Y. Qin, Deep multi-view document clustering with enhanced semantic embedding, Information Sciences 564 (2021) 273–287.

[48] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605.

[49] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: A survey, Information 10 (4) (2019) 150.

[50] J.-B. Cordonnier, A. Loukas, M. Jaggi, Multi-head attention: Collaborate instead of concatenate, arXiv preprint arXiv:2006.16362.