

# Variational Gaussian Topic Model with Invertible Neural Projections

Rui Wang<sup>§,†,‡,\*</sup> Deyu Zhou<sup>†,\*</sup> Yuxuan Xiong<sup>†</sup> Haiping Huang<sup>§,‡</sup>

<sup>§</sup>School of Computer Science, Nanjing University of Posts and Telecommunications, China

<sup>†</sup>School of Computer Science and Engineering, Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, China

<sup>‡</sup>Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, China.  
rui\_wang@njupt.edu.cn, {d.zhou, yuxuanxiong}@seu.edu.cn, hhp@njupt.edu.cn

## Abstract

Neural topic models have triggered a surge of interest in extracting topics from text automatically since they avoid the sophisticated derivations in conventional topic models. However, scarce neural topic models incorporate the word relatedness information captured in word embedding into the modeling process. To address this issue, we propose a novel topic modeling approach, called Variational Gaussian Topic Model (VaGTM). Based on the variational auto-encoder, the proposed VaGTM models each topic with a multivariate Gaussian in decoder to incorporate word relatedness. Furthermore, to address the limitation that pre-trained word embeddings of topic-associated words do not follow a multivariate Gaussian, Variational Gaussian Topic Model with Invertible neural Projections (VaGTM-IP) is extended from VaGTM. Three benchmark text corpora are used in experiments to verify the effectiveness of VaGTM and VaGTM-IP. The experimental results show that VaGTM and VaGTM-IP outperform several competitive baselines and obtain more coherent topics.

## 1 Introduction

Topic models have been extensively explored in the natural language processing (NLP) community for unsupervised knowledge discovery. Many variants (Lin and He, 2009; Zhou et al., 2014) of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have been proposed to tackle different extraction tasks. And a large body of work has considered approximate inference methods which need sophisticated mathematical derivations for model inference.

To solve this limitation, developing the neural-based topic models which employ black-box inference mechanism with neural network seems to be a promising direction. Based on variational auto-encoder (VAE) (Kingma and Welling, 2013), Miao

and Yu (2016) propose the Neural Variational Document Model (NVDM) which uses a decoder to reconstruct the document by generating the words independently. However, the inappropriate Gaussian prior employed in NVDM may result in bad topic quality. Thus, Srivastava and Sutton (2017) propose NVLDA and ProdLDA, neural topic models based on VAE, in which the logistic normal distribution is used as the prior for topic extraction. Besides, Wang (2019) proposes the Adversarial-neural Topic Model (ATM) based on the adversarial training.

On the other hand, pre-trained word embeddings (e.g. *GloVe* (Pennington et al., 2014) and *word2vec* (Mikolov et al., 2013)) from massive unlabeled text corpora provide a way of injecting word relatedness into models that would otherwise treat words as isolated categories. Thanks to the rich lexico-semantic regularities in language contained in word embedding, the state-of-the-art performance of various NLP task (e.g. sentiment analysis (Majumder et al., 2018), stance detection (Sun et al., 2018)) have been significantly improved through incorporating word embeddings. However, scarce similar attempts have been made in neural topic modeling.

Thus, in this paper, we propose a Variational Gaussian Topic Model (VaGTM) which incorporates word embeddings into topic modeling process. Based on the variational auto-encoding, its principle idea is to build a decoder which is able to reconstruct the observed documents using pre-trained word embeddings along with inferred topic distributions. Unlike the neural-based approaches that only use word co-occurrence information, VaGTM models each topic with a multivariate Gaussian distribution in decoder, and the probability of a word in a specific topic could be calculated by topic-associated Gaussian probability density function using word embeddings as input. Instead of pro-

\*co-corresponding author.

viding an analytic approximation, the VaGTM employs the variational lower bound of the observed documents as training objective to learn the means and covariance matrix of topic-associated Gaussian distributions. Due to the semantic properties of word embeddings, these distributions in decoder could capture the underlying thematic structures of text collections.

Moreover, to address the limitation that pre-trained word representations of topic-associated words may not follow a multivariate Gaussian since the specific properties of language captured by any word embedding scheme is difficult to control, we employ a flow-based transformation network (Dinh et al., 2014) to project the original word embedding space into a mixed Gaussian embedding space and propose the Variational Gaussian Topic Model with Invertible neural Projections (VaGTM-IP).

Our contributions are summarized below:

- We propose a novel Variational Gaussian Topic Model (VaGTM) which could incorporate the semantic relatedness in word embeddings into topic modeling process.
- To deal with the limitation that word embeddings of topic-associated words do not follow a multivariate Gaussian distribution, we extend the VaGTM and propose the Variational Gaussian Topic Model with Invertible neural Projections (VaGTM-IP).
- Experimental results on three public datasets show that VaGTM and VaGTM-IP outperform the state-of-the-art approaches in terms of four topic coherence measures.

## 2 Related Work

Our work is related to three lines of research, word representation learning, invertible projection learning and neural topic modeling.

### 2.1 Word Representation Learning

Distributed semantic models (i.e. word embeddings) have recently been applied successfully in many NLP tasks.

Neural based models, such as *word2vec*, have been more efficient thanks to the skip-gram with a negative sampling training method (Mikolov et al., 2013). To solve the limitation that *word2vec* only employs local context information, Pennington (2014) proposed *GloVe*, a global log-bilinear

regression model, which combines the advantages of the global matrix factorization and local context window methods. To generate a vector for an out of vocabulary word, *FastText* (Joulin et al., 2017; Athiwaratkun et al., 2018), has been proposed. It treats each word as made of character n-grams and word vectors are then computed from the sum of their n-gram representations.

### 2.2 Invertible Projection Learning

Invertible Projection Learning, also known as generative flows, is first described in NICE (Dinh et al., 2014) and has recently received much attention (Dinh et al., 2016; Jacobsen et al., 2018; Kingma and Dhariwal, 2018).

Typically, generative flows have been proposed for image generation. NICE (Dinh et al., 2014) firstly designed a composition of additive coupling layers to non-linearly project a complex high-dimensional densities into a simple prior and the inverse projection is used for generating images. To address the issue that variational auto-encoder could not approximate complex posterior, Danilo (2015) incorporated the normalizing flow into variational inference for specifying arbitrarily complex posterior distributions. Also, He (He et al., 2018) incorporated the generative flow into the Hidden Markov Model for unsupervised part-of-speech tagging.

### 2.3 Neural Topic Modeling

To overcome the difficult exact inference of topic models based on the directed graph, a replicated softmax model, called RSM, based on the Restricted Boltzmann Machines was proposed in (Hinton and Salakhutdinov, 2009). Inspired by the variational auto-encoder, Miao et al. (2016) used the multivariate Gaussian as the prior distribution of latent space and proposed the Neural Variational Document Model (NVDM) for text modeling. Recently, to deal with the inappropriate Gaussian prior of NVDM, Srivastava and Sutton (2017) proposed the NVLDA and ProDLDA which approximates the Dirichlet prior using a logistic normal distribution. Furthermore, Dieng (2019) proposed the ETM based on VAE which models topics in embedding space.

Although the extensive exploration of the above fields, scarce work has been done to incorporate word representation learning and invertible projection learning into neural topic modeling. Thus, we propose two novel topic modeling approaches,

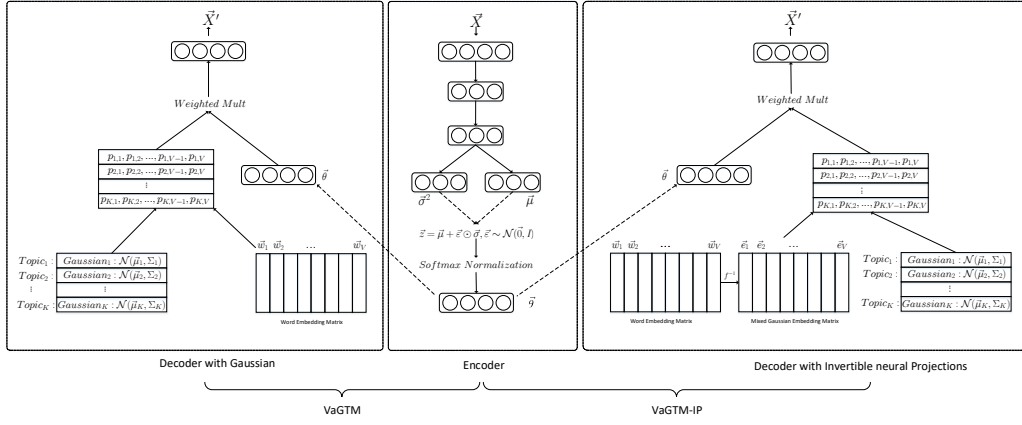


Figure 1: The framework of the Variational Gaussian Topic Model (VaGTM) and Variational Gaussian Topic Model with Invertible neural Projections (VaGTM-IP).

called VaGTM and VaGTM-IP, which differs published works in the following facets: (1). Unlike the NVDM, NVLDA, ProLDA and ATM which only uses word co-occurrence information, our proposed VaGTM models each topic with a multivariate Gaussian distribution which could incorporate the word semantic relatedness into the modeling process; (2). To deal with the issue that word embeddings of topic-associated words do not follow a multivariate Gaussian, we incorporate a flow-based projector into the modeling process and propose the VaGTM-IP .

### 3 Methodologies

Based on the encoder-decoder framework, we propose the Variational Gaussian Topic Model (VaGTM) and the Variational Gaussian Topic Model with Invertible neural Projections (VaGTM-IP).

As shown in Fig. 1, VaGTM is constituted by an encoder and a decoder with Gaussian, and VaGTM-IP contains an encoder and a decoder with invertible neural projections. The encoder, shared by VaGTM and VaGTM-IP, uses the bag of word representations with *tf-idf* weights  $\vec{X} \in \mathbb{R}^V$  as input to approximate the intractable posterior distribution over the latent topics  $\vec{\theta} \in \mathbb{R}^K$ , and decoders (in both models) aim to reconstruct  $\vec{X}$  from the encoded topic distribution  $\vec{\theta}$  and pre-trained word embeddings. In VaGTM, decoder incorporates the word relatedness by modeling topics with multivariate Gaussian. To further address the limitation that word embeddings of topic-associated words do not follow a multivariate Gaussian, an invertible neural projection is used in the decoder of VaGTM-IP. We explain the design of these networks in more detail

below.

#### 3.1 Encoder network

To obtain interpretable topics, Dirichlet prior over topics is commonly used (Wallach et al., 2009). However, it is difficult to explicitly model topics with Dirichlet prior since it is hard to develop an effective reparameterization. Thus, we solve the issue by constructing a Laplace approximation (Hennig et al., 2012).

Aiming at compressing the weighted bag of words representation  $\vec{X}$  into the latent topic distribution  $\vec{\theta}$ , the encoder  $q_{\psi}(\vec{\theta}|\vec{X})$ , parameterized with  $\psi$ , contains five layers which are one  $V$ -dimensional document representation layer, two  $H$ -dimensional semantic extraction layers, one  $K$ -dimensional reparameterization layer and one  $K$ -dimensional topic distribution layer as shown in Fig. 1. Firstly, it projects  $\vec{X}$  into an  $H$ -dimensional semantic space through semantic extraction layers based on the transformation:

$$\vec{a}_s^1 = \ln(1 + \exp(W_s^1 \vec{X} + \vec{b}_s^1)) \quad (1)$$

$$\vec{a}_s^2 = \ln(1 + \exp(W_s^2 \vec{a}_s^1 + \vec{b}_s^2)) \quad (2)$$

where  $W_s^1 \in \mathbb{R}^{H \times V}$  and  $W_s^2 \in \mathbb{R}^{H \times H}$  are weight matrices,  $\vec{b}_s^1$  and  $\vec{b}_s^2$  are corresponding basis terms,  $\vec{a}_s^1$  and  $\vec{a}_s^2$  are the semantic representations of  $\vec{X}$ ,  $\exp(\cdot)$  is the element-wise exponential function.

To further infer the posterior distribution over topics, following the Laplace approximation, a Gaussian distribution should be generated. Thus, the encoder projects  $\vec{a}_s^2$  into two  $K$ -dimensional Gaussian parameters  $\vec{\mu}$  and  $\vec{\sigma}^2$  using:

$$\vec{\mu} = \text{BN}(W_r^\mu \vec{a}_s^2 + \vec{b}_\mu) \quad (3)$$

$$\vec{\sigma}^2 = \exp(\text{BN}(W_r^\sigma \vec{a}_s^2 + \vec{b}_\sigma)) \quad (4)$$

where  $W_r^\mu \in \mathbb{R}^{K \times H}$ ,  $W_r^\sigma \in \mathbb{R}^{K \times H}$ ,  $\vec{b}_\mu$  and  $\vec{b}_\sigma$  are weight matrices and basis terms of reparameterization layers,  $\text{BN}(\cdot)$  is batch normalization. And  $\vec{\mu}$  and  $\vec{\sigma}^2$  are mean and diagonal covariance of posterior Gaussian corresponding to input  $\vec{X}$ .

Finally, based on the reparameterization trick (Kingma and Welling, 2013) and Laplace approximation (Hennig et al., 2012),  $\vec{X}$  could be inferred to a  $K$ -dimensional topic distribution  $\vec{\theta}$  using below transformation:

$$\vec{z} = \vec{\mu} + \vec{\varepsilon} \odot \vec{\sigma}, \vec{\varepsilon} \sim \mathcal{N}(\vec{0}, I) \quad (5)$$

$$\vec{\theta} = \text{softmax}(\vec{z}) \quad (6)$$

Here,  $\mathcal{N}(\vec{0}, I)$  is  $K$ -dimensional standard multivariate Gaussian, and the posterior distribution  $q(\vec{\theta}|\vec{X})$  follows the Dirichlet distribution in softmax basis (MacKay, 1998).

### 3.2 Decoder with Gaussian

The decoder  $p_\omega(\vec{X}'|\vec{\theta})$ , parameterized with  $\omega$ , reconstructs the documents by independently generating the words ( $\vec{\theta} \rightarrow \vec{X}'_i$ ). Besides, to incorporate the word relatedness captured in word embeddings, VaGTM models each topic with a multivariate Gaussian as depicted in the left panel of Fig. 1.

Concretely, VaGTM employs multivariate Gaussian  $\mathcal{N}(\vec{\mu}_k, \Sigma_k)$  to model the  $k$ -th topic. where  $\vec{\mu}_k$  and  $\Sigma_k$  represent mean and covariance matrix. Following the probability density of Gaussian, for each word  $v \in \{1, 2, \dots, V\}$ , its probability in the  $k$ -th topic  $\phi_{k,v}$  could be calculated as:

$$p(\vec{w}_v | \text{topic} = k) = \mathcal{N}(\vec{w}_v; \vec{\mu}_k, \Sigma_k) = \frac{\exp(-\frac{1}{2}(\vec{w}_v - \vec{\mu}_k)^\top \Sigma_k^{-1}(\vec{w}_v - \vec{\mu}_k))}{\sqrt{(2\pi)^{D_w} |\Sigma_k|}} \quad (7)$$

$$\phi_{k,v} = \frac{p(\vec{w}_v | \text{topic} = k)}{\sum_{v=1}^V p(\vec{w}_v | \text{topic} = k)} \quad (8)$$

where  $\vec{w}_v$  is the word embedding of word  $v$ ,  $V$  is the vocabulary size,  $|\Sigma_k| = \det \Sigma_k$  represents the determinant of covariance matrix  $\Sigma_k$ ,  $D_w$  is the dimension of word embeddings,  $\vec{\phi}_k$  is the normalized word distribution of the  $k$ -th topic.

To reconstruct the document  $\vec{X}$ , topic-word distributions and the encoded topic distribution  $\vec{\theta}$  are combined using:

$$p_{rec}(\vec{v}|\vec{\theta}) = \sum_{k=1}^K \vec{\phi}_k \cdot \theta_k \quad (9)$$

where  $K$  denotes the topic number,  $\theta_k$  means the proportion of  $k$ -th topic in  $\vec{X}$ , and  $p_{rec}(v|\vec{\theta})$  is the conditional distribution over the vocabulary which is employed to reconstruct the  $\vec{X}$  by maximizing the equation:

$$p(\vec{X}'|\vec{\theta}) = \sum_{v=1}^V X_v \cdot \log [p_{rec}(v|\vec{\theta})] \quad (10)$$

where  $X_v$  is the *tf-idf* weight of the  $v$ -th word in document  $\vec{X}$ , and  $\vec{X}'$  is the reconstructed document.

### 3.3 Decoder with Invertible neural Projections

Word embeddings indeed encode numerous semantic regularities. However, widely used word representation scheme, such as *word2vec* (2013) and *Glove* (2014), do not model embeddings with Gaussian and the resulting embeddings often follow a complicated distribution  $\mathbb{P}_w$ . Thus, it is not accurate to approximate  $\mathbb{P}_w$  with mixed Gaussian  $\mathbb{P}_m$  as VaGTM.

To approximate the  $\mathbb{P}_w$  and yield a mixed Gaussian embedding space which is more suitable for topic modeling, a non-linear neural projector  $f(\cdot)$  is employed to deterministically transform the mixed Gaussian embedding space to the pre-trained word embedding space. The vector representation of word  $v$  in mixed Gaussian embedding space is denoted as  $\vec{e}_v \in \mathbb{R}^{D_e}$ ,  $D_e$  is the dimension of mixed Gaussian embedding space. Thus, for each word  $v$ , it has the transformation below:

$$\vec{w}_v = f(\vec{e}_v) \quad (11)$$

Due to the reconstruction of  $\vec{X}$  is conditioned on word embeddings, as shown in the right panel of Fig. 1, the projector  $f(\cdot)$  should be invertible. To this end, a flow-based projector (Dinh et al., 2014) is employed here to bridge the mixed Gaussian embedding space and word embedding space since it can transform a simple distribution into a complicated one.

Thus, decoding with invertible neural projection, the probability of word  $v$  in the  $k$ -th topic is proportional to  $p(\vec{w}_v | \text{topic} = k)$  which is defined as:

$$\begin{aligned} p(\vec{w}_v | \text{topic} = k) &= \int p(\vec{w}_v | \vec{e}_v) \cdot p(\vec{e}_v | \text{topic} = k) d\vec{e}_v \\ &= \int \square \cdot \mathcal{N}(f^{-1}(\vec{w}_v); \vec{\mu}_k, \Sigma_k) \left| \det \frac{\partial f^{-1}}{\partial \vec{w}_v} \right| d\vec{w}_v \\ &= \mathcal{N}(f^{-1}(\vec{w}_v); \vec{\mu}_k, \Sigma_k) \cdot \left| \det \frac{\partial f^{-1}}{\partial \vec{w}_v} \right| \end{aligned} \quad (12)$$

where  $\frac{\partial f^{-1}}{\partial \vec{w}_v}$  represents the Jacobian matrix of inverse projection function  $f^{-1}(\cdot)$  at  $\vec{w}_v$  which is nonzero and differentiable if and only if  $f^{-1}(\cdot)$  exists, and  $\left| \det \frac{\partial f^{-1}}{\partial \vec{w}_v} \right|$  denotes the absolute value of its determinant. The symbol  $\square$  in Eq. 12 represents  $\delta(\vec{w}_v - f(\vec{e}_v))$ . Here,  $\delta(\cdot)$  is the Dirac delta function centered at  $f(\vec{e}_v)$  which is defined as:

$$p(\vec{w}_v | \vec{e}_v) = \delta(\vec{w}_v - f(\vec{e}_v)) = \begin{cases} \infty & \vec{w}_v = f(\vec{e}_v) \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

Comparing Eq. 7 and Eq. 12, it should be observed that the addition of the Jacobian term will increase the difficulty of optimizing as it requires that all component functions be invertible and also requires storage of large Jacobian matrices. To address this issue, we use additive coupling layer suggested in (Dinh et al., 2014) to guarantee a unit Jacobian determinant and the invertibility. To reconstruct the document  $\vec{X}$ , it could be observed from Eq. 12 that only  $f^{-1}(\cdot)$  is required. Thus, we explicitly design the inverse projector  $f^{-1}(\cdot)$  as Fig. 2.

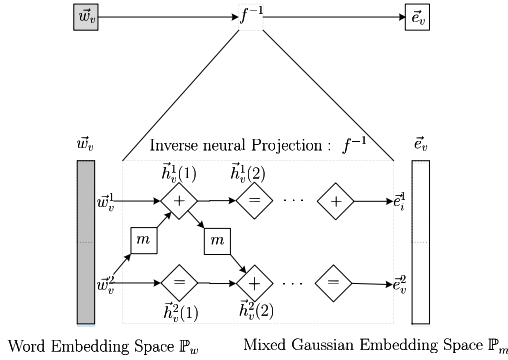


Figure 2: Description of the invertible neural projection  $f^{-1}(\cdot)$ , it transforms the pre-trained word embedding  $\vec{w}_v$  to a point  $\vec{e}_v$  in mixed Gaussian embedding space.

To accomplish the transformation, word embedding  $\vec{w}_v$  is first partitioned into two halves of dimensions,  $\vec{w}_v^1$  and  $\vec{w}_v^2$ , respectively. And the first coupling layer is represented by the nonlinear transformation from  $\vec{w}_v$  to  $\vec{h}_v(1)$  which could be defined as:

$$\vec{h}_v^1(1) = \vec{w}_v^1, \quad \vec{h}_v^2(1) = \vec{w}_v^2 + m(\vec{w}_v^1) \quad (14)$$

where  $m(\cdot) : \mathbb{R}^{D_w/2} \rightarrow \mathbb{R}^{D_w/2}$  is the *coupling function* which could be designed as any nonlinear function with same input and output shape. We choose additive coupling layer to reduce the computational consumption. To build a more complex

transformation, we compose several *coupling layers* and exchange the role of two half vectors at each layer to ensure that the composition of two layers modifies every dimension.

Then, following the reconstruction procedure illustrated in decoding process of VaGTM (subsection 3.2), the document  $\vec{X}$  could be reconstructed in a similar way in VaGTM-IP.

### 3.4 Variational Objective

In variational Bayesian, the variational lower bound on the marginal likelihood of document  $\vec{X}$  is usually formed as:

$$\begin{aligned} \mathcal{L}(\vec{X}; \psi, \omega) = & \mathbb{E}_{q_{\psi}(\vec{\theta} | \vec{X})} \log p_{\omega}(\vec{X} | \vec{\theta}) - KL(q_{\psi}(\vec{\theta} | \vec{X}) || p_{\omega}(\vec{\theta})) \end{aligned} \quad (15)$$

where  $\vec{\theta}$  denotes the topic distribution of  $\vec{X}$ . To mimic the Dirichlet prior over  $\vec{\theta}$ , the Laplace approximation is used. and the variational lower bound is rewritten as:

$$\begin{aligned} \mathcal{L}(\vec{X}; \psi, \omega) = & \sum_{v=1}^V X_v \cdot \log [p_{rec}(v | \vec{\theta})] \\ & - \frac{1}{2} \left[ \text{tr}(\Sigma^{-1} \Sigma_0) + \Delta - K + \ln \frac{\det \Sigma}{\det \Sigma_0} \right] \end{aligned} \quad (16)$$

where  $\Delta = (\vec{\mu} - \vec{\mu}_0)^T \Sigma^{-1} (\vec{\mu} - \vec{\mu}_0)$ ,  $\vec{\mu}_0$  and  $\Sigma_0$  are means and covariance matrix of standard multivariate Gaussian, and  $p_{rec}(v | \vec{\theta})$  is the  $v$ -th dimension of reconstruction distribution defined as Eq. 9.

## 4 Experiments

In this section, we first introduce the text corpora and then describe the compared baselines. Finally, we present the experimental results.

### 4.1 Experiment Settings

To validate the effectiveness of proposed VaGTM and VaGTM-IP for topic modeling, 20News-groups<sup>1</sup> dataset, Grolier<sup>2</sup> dataset and NYTimes<sup>3</sup> dataset are selected. Details are summarized below:

- *20Newsgroups dataset*, provided in (Lang, 1995), it is a collection of approximately 20,000 newsgroup articles, partitioned evenly across 20 different newsgroups.

<sup>1</sup><http://qwone.com/jason/20Newsgroups/>

<sup>2</sup><https://cs.nyu.edu/roweis/data/>

<sup>3</sup><http://archive.ics.uci.edu/ml/datasets/Bag+of+Words>

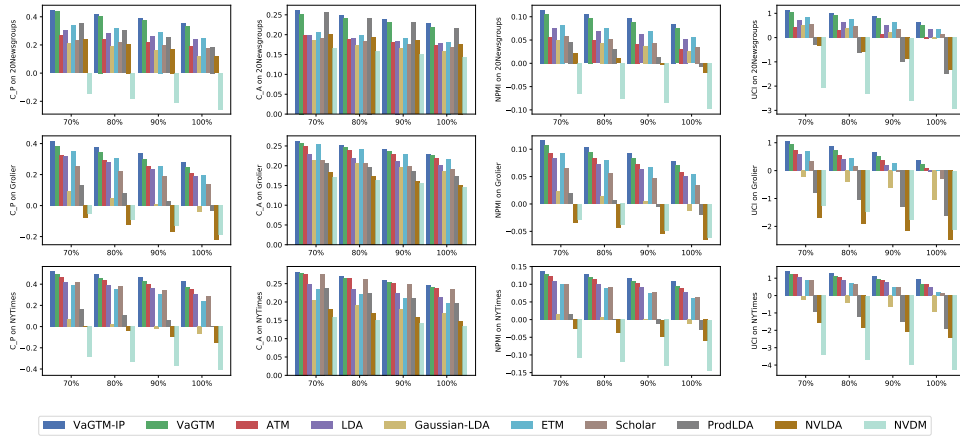


Figure 3: The comparison of average topic coherence vs. different topic proportion on 20Newsgrroups, Grolier and NYTimes.

- *Grolier dataset* is built from Grolier Multimedia Encyclopedia, and its content covers almost all the fields in the world, such as religious, technology, economics and etc.
- *NYTimes dataset* is a collection of news articles published between 1987 and 2007, and the dataset has a wide range of topics, such as sports, politics and etc.

Besides, we choose the following approaches as baselines:

- **LDA** (Blei et al., 2003), is a topic model that generates topics based on bag-of-words assumption. We implement the LDA model with the suggested configuration (Griffiths and Steyvers, 2004).
- **Gaussian-LDA** (Das et al., 2015), is a conventional topic model which uses the word embeddings information. The original implementation is used in this paper<sup>4</sup>.
- **NVDM** (Miao et al., 2016) is an unsupervised text modeling approach based on variational auto-encoder. We use the original implementation in the paper<sup>5</sup>.
- **NVLDA** (Srivastava and Sutton, 2017), is a neural topic model based on VAE. It models topics with logistic normal prior, we use the original implementation<sup>6</sup>.
- **ProdLDA** (Srivastava and Sutton, 2017), is a variant of NVLDA, in which the distribution over individual words is a product of experts rather than the mixture model.
- **Scholar** (Card et al., 2018), is a neural topic

modeling approach based on variational auto-encoder, we use the original implementation<sup>7</sup>.

- **ETM** (Dieng et al., 2019), is a neural topic model on embedding space, we use the original implementation<sup>8</sup>.
- **ATM** (Wang et al., 2019), is a neural topic modeling approach based on adversarial training, we re-implement the ATM follow the default parameter settings.

Dataset	#Documents	# Words
20Newsgrroups	11,259	1,995
Grolier	29,762	15,276
NYTimes	99,992	12,604

Table 1: The statistics of the processed datasets

For the NYTimes dataset, 100,000 articles are randomly selected, and low frequency words are removed here. The statistics of the processed datasets are listed in Table 1.

## 4.2 Topic Coherence Evaluation

Typically, the likelihood of held-out documents and topic coherence metrics are used to evaluate topic modeling performance. However, as pointed out in (Chang et al., 2009) the likelihood of held-out documents doesn't correspond to human judgment, we follow (Röder et al., 2015) and choose four topic coherence metrics which are C\_P, C\_A, NPMI and UCI to evaluate the extracted topics, and higher value implies more coherent topic. All the topic coherences are computed with the Palmetto<sup>9</sup> library

<sup>4</sup><https://github.com/rajarshd/Gaussian-LDA>

<sup>5</sup><https://github.com/ysmiao/nvdm>

<sup>6</sup>[https://github.com/akashgit/autoencoding vi for topic models](https://github.com/akashgit/autoencoding-vi-for-topic-models)

<sup>7</sup><https://github.com/dallascard/scholar>

<sup>8</sup><https://github.com/adjidieng/ETM>

<sup>9</sup><https://github.com/dice-group/Palmetto>

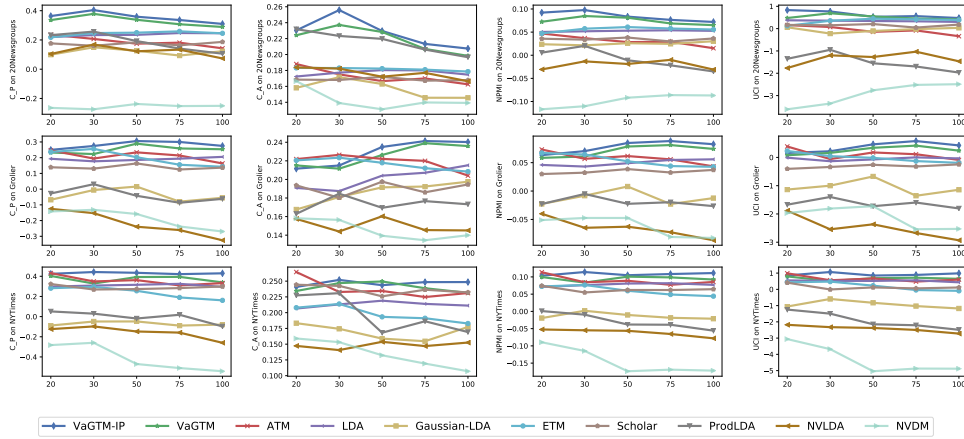


Figure 4: The comparison of average topic coherence vs. different topic number on 20Newsgroups, Grolier and NYTimes.

Dataset	Model	C.P	C.A	NPMI	UCI
20Newsgroups	NVDM	-0.2558	0.1432	-0.0984	-2.9496
	NVLDA	0.1205	0.1763	-0.0207	-1.3466
	ProdLDA	0.1858	0.2155	-0.0083	-1.5044
	LDA	0.2361	0.1769	0.0523	0.3399
	Gaussian-LDA	0.1183	0.1568	0.0252	-0.0505
	Scholar	0.1741	0.1686	0.0347	0.1497
	ETM	0.2437	0.1812	0.0558	0.3445
	ATM	0.1914	0.1720	0.0207	-0.3871
	VaGTM	0.3297	0.2190	0.0744	0.5153
	VaGTM-IP	<b>0.3545</b>	<b>0.2273</b>	<b>0.0843</b>	<b>0.6334</b>
Grolier	NVDM	-0.1877	0.1456	-0.0619	-2.1149
	NVLDA	-0.2205	0.1504	-0.0653	-2.4797
	ProdLDA	-0.0374	0.1733	-0.0193	-1.6398
	LDA	0.1908	0.2009	0.0497	-0.0503
	Gaussian-LDA	-0.0383	0.1860	-0.0115	-1.0623
	Scholar	0.1388	0.1713	0.0341	-0.3144
	ETM	0.1985	0.1909	0.0541	-0.0131
	ATM	0.2105	0.2188	0.0582	0.1051
	VaGTM	0.2515	0.2256	0.0706	0.2464
	VaGTM-IP	<b>0.2810</b>	<b>0.2286</b>	<b>0.0778</b>	<b>0.3683</b>
NYtimes	NVDM	-0.4130	0.1341	-0.1437	-4.3072
	NVLDA	-0.1575	0.1482	-0.0614	-2.4208
	ProdLDA	-0.0034	0.1963	-0.0282	-1.9173
	LDA	0.3083	0.2127	0.0772	0.5165
	Gaussian-LDA	-0.0707	0.1687	-0.0135	-0.9364
	Scholar	0.2893	0.2222	0.0635	0.1395
	ETM	0.2368	0.1894	0.0603	0.2012
	ATM	0.3568	0.2375	0.2375	0.6582
	VaGTM	0.3715	0.2402	0.0950	0.6718
	VaGTM-IP	<b>0.4304</b>	<b>0.2467</b>	<b>0.1084</b>	<b>0.9288</b>

Table 2: Average topic coherence on 20Newsgroups, Grolier and NYtimes with five topic settings [20, 30, 50, 75, 100].

and each topic is represented by top ten words according to the topic-word distribution.

To compare the topic extraction performance comprehensively, we firstly make a comparison of topic coherence vs. different topic proportions. In this part, we conduct the experiments on all datasets with five topic number settings [20, 30, 50, 75, 100]. Concretely, we calculate the average topic coherence among topics whose coherence are ranked at the top 70%, 80%, 90% and 100% positions. For example, to calculate the average NPMI value of VaGTM @80%, we should first

compute the average NPMI coherence with the selected topics whose NPMI values are ranked at the top 80% for each topic number settings and then average the five averaged coherence values. The detailed comparison is shown in Fig. 3.

As the statistics shown in Fig. 3, both VaGTM and VaGTM-IP perform competitively compared with the baseline approaches. More detailed, VaGTM-IP achieves the highest values on four metrics (C.P, C.A, NPMI and UCI) among all topic proportions and datasets, and VaGTM outperforms the compared baselines as well except on C.A metric. For the 20Newsgroups dataset on C.A, though ProdLDA performs slightly better than VaGTM (70%, 80% and 90%), VaGTM obtains much higher topic coherence values considering C.P, NPMI and UCI metrics. Thus, it would be reasonable to conclude that both VaGTM and VaGTM-IP could generally generate more coherent topics than compared approaches. And the improved performance of VaGTM and VaGTM-IP may attribute to: 1). Both models incorporate the word relatedness into the modeling process which is helpful for grouping semantically related words into the same topic; 2) The invertible neural projector transforms word embeddings into a new embedding space and the transformed representation is more suitable for topic modeling.

Moreover, to explore how the performance varies with different topic numbers, we further compare the average topic coherence (considering all the extracted topics) vs. different topic number settings for each dataset. Fig. 4 depicts the detailed comparison. From the curves in the subplots, it is clear that the proposed approaches (VaGTM and VaGTM-

Model	Topics	Topics
VaGTM-IP	Law	law bill federal legislation issue gun states protection court government
	Music	music song album dance band artist musical singer concert recording
	Film	film movie character actor show movies play star starring love
VaGTM	War	war military soldier attack killed troop rebel commander army forces
	Law	law bill federal legislation rules legal gun government proposal decision
	Music	music song band sound album pop recording concert dance rock
LDA	Film	film movie character actor movies love comedy drama show humor
	War	military war palestinian forces attack soldier army peace israeli troop
	Law	bill law <i>group</i> issue <i>member</i> federal right legislation support rules
ProdLDA	Music	music song band sound record artist album show musical rock
	Film	film movie character play actor director movies <i>minutes</i> theater cast
	War	war palestinian peace military soldier israeli troop attack border leader
ProdLDA	Law	<i>everglades</i> veto negotiator <i>billion</i> legislative treaty lawmaker <i>appropriation</i> amendment proposal
	Music	musical album <i>playwright</i> composer <i>choreographer</i> <i>onstage</i> songwriter song guitarist repertory
	Film	film comedy <i>beginitalic</i> <i>enditalic</i> sci filmmaker cinematic filmmaking movie starring
	War	peacekeeping military commander <i>surplus</i> <i>debates</i> <i>trillion</i> <i>warhead</i> <i>civilian</i> troop <i>interceptor</i>

Table 3: Topic examples extracted by selected models, italics means out-of-topic words.

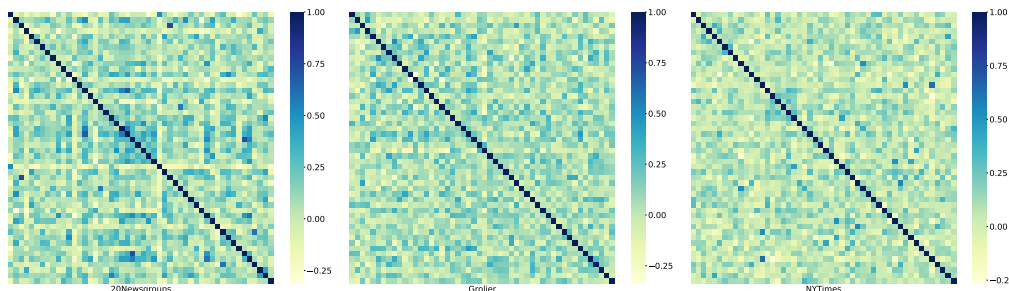


Figure 5: The topic correlation matrix on three dataset learned by VaGTM-IP.

IP) perform more stable than compared baselines. Also, in most cases, they obtain higher average coherence values. Similar to Fig. 3, ProdLDA performs well on C<sub>A</sub> metric, and it gives slightly better results than VaGTM in 20 topic settings. But beyond that, VaGTM-IP and VaGTM largely outperforms all the other baselines on all other settings. This might attribute to the incorporation of the word relatedness contained in word embeddings. We also provide a numerical comparison of average coherence values in Table 2, each value is calculated by averaging the average topic coherences (considering all topics) over five topic number settings.

From the above topic coherence comparison (Fig. 3, Fig. 4 and Table 2), it is obvious that VaGTM and VaGTM-IP perform better than baselines. To validate this qualitatively, we provide four topic examples in Table 3 and out-of-topic words are highlighted in italic.

Besides, thanks to the usage of Gaussian distribution for topic modeling, the proposed VaGTM-IP could capture the semantic correlation between extracted topics. In detail, mean vector of learned topic-associated Gaussian could be viewed as topic embeddings in mixed Gaussian embedding space and hence the semantic relations between topics

can be captured by the cosine similarity between mean vectors of topic associated Gaussian. And Fig. 5 show the visualization of topic correlation matrix for 20Newsgroups, Grolier and NYTimes datasets on 50 topic setting. Higher value implies that corresponding topics have higher semantic similarity.

## 5 Conclusion

In this paper, we have explored the variational neural topic models and proposed Variational Gaussian Topic Model (VaGTM) and Variational Gaussian Topic Model with Invertible neural Projections (VaGTM-IP). VaGTM incorporates the word relatedness stored in pre-trained word embeddings into the modeling process to enhance the quality of extracted topics. Moreover, to address the issue that pre-trained word embedding is not suitable enough for specific neural topic modeling, we extended VaGTM and proposed VaGTM-IP. It uses an invertible neural projector to transform the pre-trained word embedding into a mixed Gaussian embedding space which is more suitable for topic modeling. The experimental comparison with the state-of-the-art baselines on three benchmark datasets shows that VaGTM and VaGTM-IP achieve improved topic coherence results.



## References

- Ben Athiwaratkun, Andrew Wilson, and Anima Anandkumar. 2018. [Probabilistic FastText for multi-sense word embeddings](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11. Association for Computational Linguistics.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040.
- Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. [Reading tea leaves: How humans interpret topic models](#). In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc.
- Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804.
- Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*.
- Laurent Dinh, David Krueger, and Yoshua Bengio. 2014. Nice: Non-linear independent components estimation. In *International Conference on Learning Representations*.
- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2016. [Density estimation using real nvp](#). In *International Conference on Learning Representations*.
- Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.
- Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. [Unsupervised learning of syntactic structure with invertible neural projections](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Brussels, Belgium. Association for Computational Linguistics.
- Philipp Hennig, David Stern, Ralf Herbrich, and Thore Graepel. 2012. Kernel topic models. In *Artificial Intelligence and Statistics*, pages 511–519.
- Geoffrey E Hinton and Ruslan R Salakhutdinov. 2009. [Replicated softmax: an undirected topic model](#). In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1607–1614. Curran Associates, Inc.
- Jörn-Henrik Jacobsen, Arnold W.M. Smeulders, and Edouard Oyallon. 2018. [i-revnet: Deep invertible networks](#). In *International Conference on Learning Representations*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Durk P Kingma and Prafulla Dhariwal. 2018. [Glow: Generative flow with invertible 1x1 convolutions](#). In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10215–10224. Curran Associates, Inc.
- Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier.
- Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384. ACM.
- David JC MacKay. 1998. Choice of basis for laplace approximation. *Machine learning*, 33(1):77–86.
- Navonil Majumder, Soujanya Poria, Alexander Gelbukh, Md. Shad Akhtar, Erik Cambria, and Asif Ekbal. 2018. [IARM: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3402–3411, Brussels, Belgium. Association for Computational Linguistics.
- Yishu Miao, Lei Yu, and Phil Blunsom. 2016. [Neural variational inference for text processing](#). In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16*, pages 1727–1736. JMLR.org.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Danilo Jimenez Rezende and Shakir Mohamed. 2015. [Variational inference with normalizing flows](#). In *Proceedings of the 32Nd International Conference on Machine Learning - Volume 37, ICML'15*, pages 1530–1538. JMLR.org.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. [Exploring the space of topic coherence measures](#). In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15*, pages 399–408, New York, NY, USA. ACM.
- Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.
- Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Hanna M. Wallach, David M. Mimno, and Andrew McCallum. 2009. [Rethinking lda: Why priors matter](#). In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. Curran Associates, Inc.
- Rui Wang, Deyu Zhou, and Yulan He. 2019. Atm: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098.
- Deyu Zhou, Liangyu Chen, and Yulan He. 2014. A simple bayesian modelling approach to event extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 700–705.