Topic Modeling using Variational Auto-Encoders with Gumbel-Softmax and Logistic-Normal Mixture Distributions

Denys Silveira*, André Carvalho †, Marco Cristo‡
Institute of Computing, University of Amazonas
Manaus, Amazonas, Brazil
{*ddbs, †andre,‡marco.cristo}@icomp.ufam.edu.br

Marie-Francine Moens§
Department of Computer Science, KU Leuven
Leuven, Belgium
§sien.moens@cs.kuleuven.be

***Abstract*—Probabilistic Topic Models are widely applied in many NLP-related tasks due to their effective use of unlabeled data to capture variable dependencies. Analytical solutions for Bayesian inference of such models, however, are usually intractable, hindering the proposition of highly expressive text models. In this scenario, Variational Auto-Encoders (VAEs), where an inference network (the encoder) is used to approximate the posterior distribution, became a promising alternative for inferring latent topic distributions of text documents. These models, however, also pose new challenges such as the requirement of continuous and reparameterizable distributions which may not fit so well the true latent topic distributions. Moreover, inference networks are prone to component collapsing, impairing the collection of coherent topics. To overcome these problems, we propose two new text topic models based on the categorical distribution Gumbel-Softmax (GSDTM) and on mixtures of Logistic-Normal distributions (LMDTM). We also provide a study on the impact of different modeling choices on the generated topics, observing a trade-off between topic coherence and document reconstruction. Through experiments using two reference datasets, we show that GSDTM largely outperforms previous state-of-the-art baselines when considering three different evaluation metrics.**

I. INTRODUCTION

Topic modeling have been successfully applied in many tasks related to Natural Language Processing (NLP) such as document retrieval [1], clustering [2], classification [3], authorship identification [4], and aspect-based sentiment analysis [5]. Much of its success is related to how effectively these probabilistic generative models use unlabeled data to capture dependencies among the modeled variables.

However, traditional approaches based on Bayesian inference of directed probabilistic graphical models require the mathematical derivation of the inference algorithm for each new model. Even for models that capture simple relations among the variables, such as sequential and spatial dependencies, it can be hard to derive a feasible inference algorithm. Analytical solutions for such models usually result in intractable integrals, requiring methods which deal with approximated (simpler) posterior distributions or to approximate the true posterior using sampling strategies. As consequence, highly expressive text models are usually avoided.

New techniques based on variational auto-encoders (VAEs) mitigate many of these issues. When applied to text topic modeling, VAEs approximate true posterior distributions using neural networks conditioned on text. More specifically, they train an inference network that directly maps documents to well-behaving posterior distributions. To accomplish this, the backpropagation algorithm is used to minimize the reconstruction error of the documents taken as input. Also, due to the flexibility of neural networks, rich models can be designed, able to explicitly capture spatial and sequential dependencies among the variables and to learn complicated non-linear distributions in a feasible way.

These models, however, are hard to use in practice due to a few issues. For instance, a model needs to approximate a continuous distribution that can be reparameterized as a differentiable function, since the backpropagation algorithm operates on gradients. Topics, however, could be better modeled by discrete, categorical distributions, since they represent somewhat distinct semantic classes. Further, to allow reparametrization, it is common to adopt distributions, such as Gaussian or Logistic-Normal, that might not be the best fit to complex true posterior distributions. Besides that, inference networks can also get stuck in local optima, an issue known as component collapsing. To deal with this problem, many heuristics have been proposed such as tweaking optimization parameters, clipping certain components of the loss function or adopting techniques such as batch normalization and dropout. However, the impact of the use of such techniques on the quality of the topics is not well understood.

To cope with these problems, in this work we study the impact of two different topic distributions on the quality of text topic modeling: Gumbel-Softmax (GS) and Logistic-Normal Mixture Models (LNMM). GS is a continuous distribution able to approximate a discrete one, which enables us to represent topics as categories, while LNMMs make it possible to fit complex distributions using linear combinations of Logistic-Normal distributions. We also provide a comprehensive study on the impact of different model parameterizations on the topics generated by VAEs. In particular, we study the effects of batch normalization (BN) and dropout on inference, reconstruction, and topic coherence.

The main contribution of this work is the proposition of the Gumbel-Softmax Document Topic Model (GSDTM) and the Logistic-normal Mixture Document Topic Model (LMDTM), the first VAE-based models to use GS and LNMM directly for text topic modeling. We also study how different neural network choices affect the quality of the generated topics. Based on a comparison of the proposed models using two reference collections and three evaluation metrics (average topic

978-1-5090-6014-6/18/$31.00 ©2018 IEEE

coherence, perplexity and precision on fraction of retrieved documents), we show that (i) GSDTM achieved the best results in this task, outperforming other evaluated approaches in most scenarios, which include two state-of-the-art VAE-based neural topic models previously proposed in literature; (ii) the impact of the use of dropout and batch normalization on these metrics, finding a positive impact of their use on topic coherence but negative on document reconstruction and retrieval, and (iii) that LMDTM is competitive in scenarios with large collections without the adoption of stabilizing neural techniques such as BN and dropout.

The remainder of this work is organized as follows. Section 2 presents related work. In Section 3 and Section 4, we present our proposed topic models, LMDTM and GSDTM. We report experiments and discuss results in Section 5. Finally, in Section 6, we present our conclusion remarks and directions for future work.

II. RELATED WORK

Topic models are statistical models for discovering the abstract topics in which pieces of information could be classified. To accomplish this, they infer the latent semantic structure underlying extensive unstructured data bodies, such as collections of text, genetic information, images and others. The most popular text topic models are based on Latent Dirichlet Allocation (LDA), proposed by Blei et al. [6]. The intuition behind LDA is that documents cover a small number of topics, which are represented by a small number of words. This is encoded in LDA by the use of sparse Dirichlet prior distributions over document-topic and topic-word distributions.

On the last decade models based on neural networks have emerged as a viable alternative to topic modeling as deep learning became mainstream. For instance, Salakhutdinov et al. [7] proposed a neural linguistic network, Replicated Softmax (RS), capable to estimate the probability of observing a new word in a document given previously observed words. Besides generating documents, RS is able to learn meaningful document representations, due to the use of conditional mean-field recursive equations. Larochelle and Lauly [8] improved RS by replacing the softmax distribution over words by a hierarchical distribution over paths in a binary tree of words. Their model, DocNADE, scales logarithmically as the vocabulary size increases, instead of linearly as RS. They also showed that DocNADE slightly outperforms RS in a text retrieval task.

Other increasingly complex neural linguistic models have been proposed, such as GMNTM [9] and SLRTM [10]. GMNTM represents each topic as a multidimensional vector where each word is affected not only by its topic, but also by the embedding vector of its surrounding words and the context. Topics, sentences, words and their clustering are learned jointly. As GMNTM, SLRTM also emphasizes the modeling of word ordering as a means of improving topic discovering by using a Recurrent Neural Network (RNN) based framework.

More recently, Variational Auto-Encoders (VAEs) have been successfully adapted for text topic modeling, resulting in the Neural Variational Document Model (NVDM) and the Product Latent Dirichlet Allocation (ProdLDA) models [11], [12]. Miao et al. [11] proposed NVDM as an extension of a standard VAE, with an encoder which learns a Gaussian distribution and a softmax decoder (the generative model) which reconstructs documents in a semantic word embedding space using an approach similar to DocNADE. The other model, ProdLDA, was proposed by Srivastava et al. [12]. This method uses a Laplace technique to obtain an approximation of Dirichlet priors using a Logistic-Normal distribution (LND). Thus, it is able to calculate the LND mean and covariance parameters using Dirichlet parameters. Finally, ProdLDA uses the same generator model of NVDM, so that it is able to extract topics through reconstructing documents in a semantic word embedding over topic-word distributions in the logit space. As far as we know, ProdLDA and NVDM are the text topic models with the best reported results in literature, considering metrics such as perplexity and average topic coherence, and were used as baselines in our experiments (cf. Section V-A).

VAEs have also been used for image processing. Dilokthanakul et al. [13] proposed the Gaussian Mixture Variational Auto-Encoder (GMVAE), an extension of the standard VAE which approximates a finite and uniform mixture of Gaussian components with the objective to improve unsupervised clustering in image reconstruction task, showing that complex data might be efficiently clustered using a distribution mixture based model as prior distribution. Jang et al. [14] have proposed an adaptation of the standard VAE for image reconstruction task approximating a Gumbel-Softmax distribution, named Gumbel-Softmax VAE (GSVAE). These methods show that the use of distribution mixtures and Gumbel-Softmax in VAEs encoders might lead to improved results, which was also an inspiration for this work.

As in NVDM and ProdLDA, in this work we also adapt VAEs to text topic modeling. However, unlike these models our encoders approximates the Gumbel-Softmax or the Logistic-Normal Mixture distributions. We also adopt a number of different architecture configurations, studying the impact of those design choices on the reconstruction quality and topic coherence of the model.

III. LOGISTIC-NORMAL MIXTURE DOCUMENT TOPIC MODEL (LMDTM)

The Logistic-Normal Mixture Document Topic Model (LMDTM) is a VAE that uses a Logistic-Normal Mixture Model (LNMM) as an encoder distribution to learn topic structures on text data. As GMVAE, this model applies a linear combination of distributions that generally fits better to complex data [15]. However, instead of using a Gaussian distribution, we used a Logistic-Normal (LN) mixture distribution, since Srivastava et al. [12] showed that using LN distributions usually improves the topic coherence. Thus, it may be a better way to approximate the true posterior

topic distribution and consequently might improve the topic modeling task.

Consider a generative model $p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z})$, that can be calculated as $p_\theta(\mathbf{y})p_\theta(\mathbf{z}|\mathbf{y})p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y})$, where $\mathbf{x}$ is an observed sample generated from a set of latent variables $\mathbf{y}$ and $\mathbf{z}$, and θ represents the set of parameters of the neural network used to reconstruct $\mathbf{x}$ from samples of $\mathbf{z}$. The sampling process of latent variables $\mathbf{y}$ and $\mathbf{z}$ and observed variable $\mathbf{x}$ is defined by Equations 1, 2 and 3 respectively, where μ_{y_k} and $\sigma^2_{y_k}$ correspond to the mean and the variance parameters of the k-th component:

$$\mathbf{y} \sim Multinomial(\pi) \tag{1}$$

$$\mathbf{z}|\mathbf{y} \sim \prod_k \mathcal{LN}(\mu_{y_k}, \sigma^2_{y_k})^{y_k} \tag{2}$$

$$\mathbf{x}|\mathbf{z}, \mathbf{y} \sim softmax(\mathbf{z}) \tag{3}$$

The discrete variable $\mathbf{y}$ is an one-hot vector sampled from a multinomial distribution with mixing probability π from which one can draw one component from the Logistic-Normal mixture. Thus, $p_\theta(\mathbf{y})$ represents the mixture coefficient of the LNMM. To simplify the model, we consider $\mathbf{z}$ uniformly distributed by setting $\pi = K^{-1}$, where K is the number of Logistic-Normal components. Also, $\mathbf{z}$ is a continuous variable sampled from a component of the Logistic-Normal mixture parameterized by $\mathbf{y}$. While $p_\theta(\mathbf{y})p_\theta(\mathbf{z}|\mathbf{y})$ is associated with the LNMM, the distribution $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y})$ denotes a multinomial logistic regression given by the softmax function. Thus, the observed sample $\mathbf{x}$ can be reconstructed from the samples $\mathbf{z}$ of the Logistic-Normal mixture distribution. To do so, we use neural networks to learn the best parameters of each generative model distribution, instead of traditional Bayesian methods such as Expectation-Maximization (EM).

However, since the joint probability is not tractable, we adopt an inference process based on the mean-field variational family. More specifically, the inference process in LMDTM model is carried out over the simple variational distribution $q_\Phi(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \prod_y q_\Phi(\mathbf{y}|\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x}, \mathbf{y})$, where $q_\Phi(\mathbf{y}|\mathbf{x})$ and $q_\Phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ follow Multinomial and Logistic-Normal distributions, respectively. The LMDTM model has K independent Logistic-Normal distributions, each one assigned by the multinomial variable $\mathbf{y}$. Thus, the distributions work as Logistic-Normal individual components of a LNMM, that might enable the variational process to approximate the generative distribution better than when using only one Logistic-Normal component. Following the standard inference process, we find the variational lower bound according to Equation 4, using the Kullback-Leibler (KL) Divergence between distributions q_Φ and p_θ.

$$\begin{aligned} logp_\theta \geq -\mathcal{U}(\mathbf{x}) = \frac{1}{K}\mathbb{E}_{z\sim q_\theta(\mathbf{y},\mathbf{z}|\mathbf{x})}[logp_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})] - \\ D_{KL}[q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})||p_\theta(\mathbf{z}|\mathbf{y})] - D_{KL}[q_\phi(\mathbf{y}|\mathbf{x})||p_\theta(\mathbf{y})] \end{aligned} \tag{4}$$

Following the approach used in ProdLDA [12], we use the Gaussian KL Divergence $D_{KL}[q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})||p_\theta(\mathbf{z}|\mathbf{y})]$ that can be analytically computed in its closed form. Since $\mathbf{y}$ is a discrete variable, we calculate $D_{KL}[q_\phi(\mathbf{y}|\mathbf{x})||p_\theta(\mathbf{y})]$ using its definition for discrete probability distributions. Finally, the information entropy $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}logp_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$ is estimated using the softmax cross-entropy function, where D is the number of documents (cf. Equation 5).

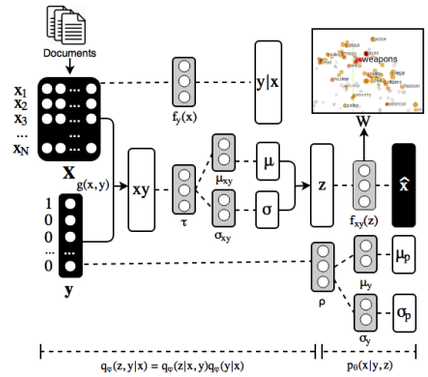

Fig. 1. LMDTM architecture. Dashed lines identify each neural network of the VAE. The function $g(\mathbf{x}, \mathbf{y})$ concatenates $\mathbf{x}$ and $\mathbf{y}$.

LMDTM applies the standard reparametrization trick for each LNMM Logistic-Normal component with softmax transformation, similar to the approach of Srivastava et al. [12]. Thus, continuous variable z is given by the equation $softmax(\mu + \epsilon\sigma)$, where μ is the mean and σ is the standard deviation of the Logistic-Normal component assigned by y. To maximize the variational lower bound value, in LMDTM, all parameters of all probability distributions are trained using a VAE network.

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\sum_{j=1}^{D} logp_\theta(\mathbf{x}_j|\mathbf{y}, \mathbf{z})] = -\sum_{j=1}^{D} \mathbf{x}_j logp_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) \tag{5}$$

Unlike standard VAEs, the LMDTM input, besides the bag-of-words vectors representing the documents as in [11], also includes one-hot vectors representing the corresponding topics, as illustrated in Fig. 1. More specifically, the documents are denoted by a set of vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, where each $\mathbf{x}_i$ is associated with a document d_i. Observation $\mathbf{x}_i \in \mathcal{R}^V$ represents a word frequency vector where each word belongs to a vocabulary of size V. The one-hot vector $\mathbf{y}$ represents the Logistic-Normal component discriminator $\mathbf{y}$. The concatenation of $\mathbf{X}$ and $\mathbf{y}$ is the input of the inference network. In the LMDTM model, the inference network comprises two neural networks. The first one is the discriminator network $f_y(\mathbf{x})$. It has linear transformations and learns the parameters of the multinomial distribution $q_\Phi(\mathbf{y}|\mathbf{x})$, thus discriminating the class of the documents given by the observed data $\mathbf{x}$. The second network is $\tau = f_{xy}(g(\mathbf{x}, \mathbf{y}))$, where g is a function used to concatenate $\mathbf{X}$ to $\mathbf{y}$. The mean and the variance vectors of Logistic-Normal component $q_\theta(\mathbf{y}, \mathbf{z}|\mathbf{x})$ are computed respectively by the neural networks $\mu_{xy} = l(\tau)$ and $\sigma_{xy} = l(\tau)$, where $l(.)$ is a linear transformation function. The reparameterization trick is applied on these vectors so that samples of $\mathbf{z}$ can be drawn.

While VAEs applied to image data normally use binomial or Gaussian probabilistic distributions, in LMDTM we follow Miao et al. [11] and adopt a multinomial logistic regression to reconstruct distribution $p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z})$, as shown in Equation 6, where $\mathbf{W}$ is the word embedding representation in a vectorial space $\mathcal{R}^{TxV}$ shared across documents and $\mathbf{b}$ is the bias vector:

$$p_\theta(x|y,z) = \frac{exp(\mathbf{W}\mathbf{z}_i + \mathbf{b}_i)}{\sum_{j=1}^{k} exp(\mathbf{W}\mathbf{z}_j + \mathbf{b}_j)} \tag{6}$$

Finally, the mean and variance of each Logistic-Normal component $p_\theta(\mathbf{z}|\mathbf{y})$ of the LNMM are computed respectively by $\mu_y = l(\rho)$ and $\sigma_y = l(\rho)$, where $\rho = l(f_z(\mathbf{y}))$ is a fully connected network with linear transformation which estimates the parameters of the Logistic-Normal component of the generative model associated with $\mathbf{y}$.

Based on samples $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$, we train the network with ADAM optimizer [16] to maximize the lower bound. The probability distribution $p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z})$ can be estimated by drawing M samples from $\mathbf{z}$. Following the same approach of NVDM and ProdLDA, we adopt a sample size of 1 ($M = 1$) to reduce the cost of training.

IV. Gumbel-Softmax Document Topic Model (GSDTM)

Different topics can be viewed as distinct semantic classes. As consequence, they might be better modeled using discrete distributions, leading to more understandable and computationally efficient models [17], [18]. However, as discrete density functions are not differentiable, previous VAE-based topic models are restricted to continuous distributions. In such context, an interesting distribution is Gumbel-Softmax [19] (GS), a continuous distribution over the simplex based on the Gumbel distribution, which can approximate categorical (discrete) data. Thus, its parameter gradients can be easily computed via Gumbel-Max trick [20], with Jang et al. [14] successfully using it in a VAE for an image reconstruction task, named Gumbel-Softmax VAE (GSVAE).

Let the discriminator variable $\mathbf{y}$ be sampled from a multinomial distribution with probabilities $\pi = \{\pi_1, \cdots, \pi_T\}$, such that $\sum_i \pi_i = 1$. GS assumes that the categorical samples $\mathbf{y}$ are encoded as T-dimensional one-hot vectors represented as vertices of simplex Δ^{T-1}. The density of the GS distribution [14] is denoted by Equation 7:

$$p_{\pi,\tau}(\mathbf{y}) = \Gamma(T)\tau^{T-1}(\sum_{i=1}^{T} \pi_i/\mathbf{y}_i^\tau)^{-T} \prod_{i=1}^{T}(\pi_i/\mathbf{y}_i^{\tau+1}) \tag{7}$$

where parameter τ, referred to as the *softmax temperature*, is used to control the discreteness of the probability distribution over $\mathbf{y}$ and T is the number of dimensions of GS distribution. In Fig. 2 we can observe the effect of π and τ on GS probability distribution. For low temperatures, samples from GS become one-hot, while for high temperatures they become identical to the uniform categorical distribution. Analogously, high probability values of π_i make the GS probability mass drifts towards the $\mathbf{y}_i$ vertex.

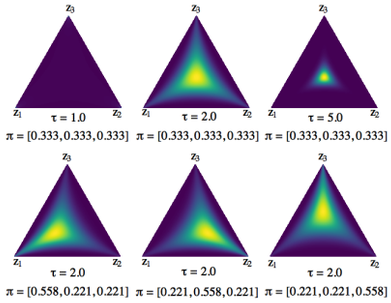

Fig. 2. The effect of parameters shown in graphical representation of the Gumbel-Softmax distribution variating the temperature τ (above) and changing the probabilities π (below).

The GSVAE model uses the GS distribution by sampling latent variable $\mathbf{z}$ from this probability distribution. As other discriminative VAEs, GSVAE samples latent variable $\mathbf{y}$ from a multinomial distribution while the observed data $\mathbf{x}$ follows a Gaussian or Bernoulli distribution. Its joint distribution is given by Equation 8:

$$p_\theta(\mathbf{x},\mathbf{y},\mathbf{z}) = \sum_y \sum_z p_\theta(\mathbf{y}) p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z},\mathbf{y}) \tag{8}$$

Using the inference process of GSVAE, we propose a adaptation of this model, called Gumbel-Softmax Document Topic Model (GSDTM), that applies the properties of the GS distribution for topic modeling. The main advantage of the GS distribution is its better approximation to categorical data. Consequently, the encoding of documents into a topic space can be improved since the topic is a inherently discrete variable. Moreover, GS has a simple and efficient sampling that improves the efficiency of training on large text. Other advantage is that GS produces low-variance biased gradients on the corresponding discrete graph [21], which increases the robustness of the statistical model.

Its joint distribution has a few differences in comparison to LMDTM. Since GS is not a mixture distribution, latent variable $\mathbf{z}$ is not conditioned to the discriminative variable $\mathbf{y}$, thus probabilities $p_\theta(\mathbf{y})$ and $p_\theta(\mathbf{z})$ are computed independently. GSDTM adopts the same variational inference process of LMDTM. Thus, following the approach described in Section III, the variational lower bound can be denoted by:

$$\begin{aligned} logp_\theta \geq -\mathcal{U}(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\theta(\mathbf{y},\mathbf{z}|\mathbf{x})}[logp_\theta(\mathbf{x}|\mathbf{y},\mathbf{z}) \\ + logp_\theta(\mathbf{y}) + logp_\theta(\mathbf{z}) - q_\phi(\mathbf{y},\mathbf{z}|\mathbf{x})] \end{aligned} \tag{9}$$

Unlike VAEs where the latent variable follows a Gaussian distribution, GSDTM uses a GS distribution to generate the samples of latent variable $\mathbf{z}$. Since the standard RT is not applicable in this case, GSDTM adopts the Gumbel-Max trick that draws samples from a categorical distribution with class probabilities π as follows:

$$\mathbf{z} = \text{one\_hot}(\arg\max_i[\mathbf{g}_i + log\pi_i]) \tag{10}$$

where $g_1, g_2, \cdots, g_k$ are samples drawn from $Gumbel(0,1)$. We can sample from $Gumbel(0,1)$ using the

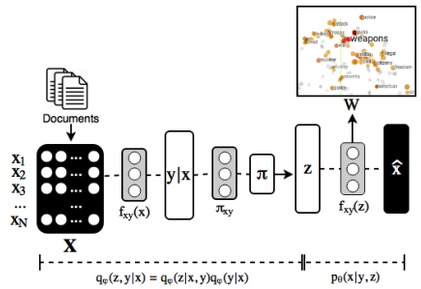

Fig. 3. GSDTM architecture. Note that π refers to Gumbel-Softmax logit parameter.

inverse transform sampling by drawing $\mathbf{U} \sim Uniform(0, 1)$ and computing $\mathbf{g} = -log(-log(\mathbf{U}))$ [21]. Since a discrete function is not differentiable, the Gumbel-Max trick uses the softmax function, a continuous approximation to argmax:

$$\mathbf{z}_i = \frac{exp((\mathbf{g}_i + log\pi_i)/\tau)}{\sum_{j=1}^{k} exp((\mathbf{g}_j + log\pi_j)/\tau)}, i = 1, \cdots, k \qquad (11)$$

The VAE network of GSDTM is illustrated in Fig. 3. As observed, it shares the input representation of LMDTM. The input data is encoded by discriminator network $f_{xy}(\mathbf{x})$ which calculates the parameters of the multinomial distribution $q_\phi(\mathbf{y}|\mathbf{x})$ using a linear transformation. It discriminates the document class $\mathbf{y}$ given the input $\mathbf{x}$. In addition, note that the output $f_{xy}(\mathbf{x})$, the vector of multinomial probabilities of the discriminator variable $\mathbf{y}$, corresponds to the parameter π. Afterwards, the samples $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ are obtained via Gumbel-Max trick using parameters π and τ. Unlike π, temperature τ is not learned by the VAE network. Instead, we define an initial temperature and apply an annealing process throughout the training iterations according to the approach proposed by Jang et al. [14].

As LMDTM, GSDTM adopts a multinomial logistic regression to reconstruct distribution $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$. A neural network $f_{xy}(\mathbf{z})$, with a softmax function and weight $\mathbf{W}$, decodes the samples generated from $\mathbf{z}$ into reconstructed samples $\hat{\mathbf{x}}$. $\mathbf{W}$ is a word embedding representation in a vectorial space $\mathcal{R}^{TxV}$, where V is the vocabulary size and T is the number of topics. We trained GSDTM using the ADAM optimizer to maximize the lower bound. As in LMDTM, we adopt sample size $M = 1$ to calculate entropy, when drawing from $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$ distribution.

V. Experiments

To better understand the quality of our proposed methods in comparison to the baselines, we performed experiments on two datasets: 20newsgroups[1] and RCV1 Volume 2 (RCV1-v2)[2]. The 20newsgroups dataset is a collection of newsgroups articles, split into 11,314 training and 7,531 test documents. RCV1-v2 is a large collection of Reuters News stories composed of 794,414 training and 10,000 test instances. As Larochelle et al. [8], Miao et al. [11] and Srivastava et al. [12], we set vocabulary size to 2000 terms on 20newsgroups and 10000 on RCV1-v2. We then applied the same preprocessing steps of Miao et al. [11], i.e., we tokenized and removed stopwords and any word with characters not encoded using UTF-8 from the articles. We used publicly available implementations using default parameters of our baseline methods for comparison, ProdLDA[3], NVDM[4], and DocNADE[5].

To compare the methods, we used the same setup as Miao et al. [11] and Srivastava et al. [12]. For the GSDTM encoder we set a 3-layer feed-forward network with ReLU activations in the first 2 layers and a linear output layer for the Gumbel-Softmax logit parameter. The LMDTM encoder is similar, except for using two linear output layers, since LMDTM has two parameters on its probabilistic encoder, the mean and the variance. The LMDTM discriminator network consists of a single layer with ReLU activation followed by a linear layer. The decoder is a linear layer with softmax transformation in both models. We trained each dataset for 50 and 200 topics using the ADAM [16] optimizer. We have made the source code of LMDTM and GSDTM publicly available on-line.[6]

For topic coherence and perplexity measures, we followed the procedure adopted by Srivastava et al. [12], using the variational lower bound to compute the average perplexity per document on the test dataset. We also use Normalized Point-wise Mutual Information (NPMI) as a topic coherence metric. In the document retrieval evaluation task we used the same process as Miao et al. [11], setting 100 documents as the validation dataset.

Since VAE-based models are prone to topic collapsing, previous researchers, such as Srivastava et al. [12], have empirically observed that this problem can be alleviated by the use of techniques such as dropout and batch normalization (BN) along with a careful choice of optimizer parameters such as the learning rate. While dropout improves generalization and slows down the training, BN minimizes the covariate shift, stabilizing the learning. However, since these widely used techniques may impact the effectiveness of the models, we evaluated them with and without applying such techniques. In particular, we set the retaining probability to 0.6 in all experiments with dropout. We evaluated the models using the best settings found for the validation dataset, resulting at fixed learning rate values of 0.0002 and 0.002, for GSDTM and LMDTM, respectively. When BN and dropout are used, we set the learning rate to 0.0002 for both models. Regarding DocNADE, we used it as originally described, without applying BN and dropout.

A. Topic Evaluation

We evaluated the Average Topic Coherence (ATC) of the models using the NPMI metric. An advantage of this metric, in comparison to others such as perplexity (cf. Section V-B),

[1]Available at http://qwone.com/~jason/20Newsgroups

[2]Available at http://trec.nist.gov/data/reuters/reuters.html

[3]https://github.com/akashgit/autoencoding_vi_for_topic_models/

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

[5]http://www.dmi.usherb.ca/~larocheh/code/DocNADE.zip

[6]https://github.com/denyssilveira/gsdtm_lmdtm_topic_model

TABLE I
TOPIC COHERENCE FOR EACH DATASET IN TERMS OF ATC. HIGHER IS BETTER. ASTERISKS INDICATE TOPIC COLLAPSING DURING TRAINING.

Model	With BN and dropout		Without BN and dropout	
	50 Topics	200 Topics	50 Topics	200 Topics
20newsgroups Dataset				
GSDTM	**0.283**	**0.233**	**0.174**	0.128
LMDTM	0.194	0.173	0.155*	0.085*
ProdLDA	0.267	0.216	0.041*	0.036*
NVDM	0.126	0.077	0.085	0.068
DocNADE	-	-	0.141	**0.139**
RCV1-v2 Dataset				
GSDTM	**0.164**	**0.191**	0.048	0.020
LMDTM	0.150	0.131	0.104*	**0.119***
ProdLDA	0.151	0.149	**0.113**	0.091
NVDM	0.057	0.055	0.069	0.028
DocNADE	-	-	0.037	0.026

TABLE II
PERPLEXITY SCORES FOR EACH DATASET. LOWER IS BETTER. ASTERISKS INDICATE TOPIC COLLAPSING DURING TRAINING.

Model	With BN and dropout		Without BN and dropout	
	50 Topics	200 Topics	50 Topics	200 Topics
20newsgroups Dataset				
GSDTM	**1130.331**	**1124.204**	**867.732**	864.727
LMDTM	1148.050	1191.779	990.753*	978.050*
ProdLDA	1162.674	1191.325	1150.222*	1427.525*
NVDM	1187.097	1184.164	869.489	909.735
DocNADE	-	-	870.505	**851.448**
RCV1-v2 Dataset				
GSDTM	**1968.938**	**1978.788**	**386.594**	**261.545**
LMDTM	2049.551	2147.196	610.609*	1515.863*
ProdLDA	2026.826	1999.589	750.837	623.680
NVDM	2261.257	2213.956	512.896	526.284
DocNADE	-	-	629.915	449.849

is that it correlates well with the observed coherence rated by human judges in topic modeling tasks [22].

To extract the topics of a trained model, we followed the approach by Srivastava et al. [12]. From the multinomial logistic regression that calculates $p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z})$, we collected the weights $\mathbf{W}$ of the network. We can interpret $\mathbf{W}$ as a matrix where each line represents a topic $t \in [1, 2, \cdots, T]$ and each column a word w_i contained inside the vocabulary set, where $i \in [1, 2, \cdots, V]$. Since weight $\mathbf{W}$ is a vectorial representation of the words in the topic space, the value $\mathbf{W}_{ti}$ indicates the score, also referred to as connection, between topic t and word w_i. Then, for each topic t, we got N words with the strongest scores on line t. The result of this operation is a topic rank containing the N words with the largest probability to belong to a given topic. We used the implementation of Lau et al. [22], publicly available[7] to calculate the ATC from the topic rank of all models.

Table I shows the results obtained regarding topic coherence. As can be seen, GSDTM is clearly superior in all scenarios when using stabilization methods (batch normalization and dropout). It achieved gains as high as 28% (on dataset RCV1-v2 using 200 topics) in comparison with the second best result, achieved by ProdLDA. In all other stabilized scenarios, GSDTM outperformed the other methods, albeit with smaller gain margins: 8% over ProdLDA on RCV1-v2 with 50 topics, 6% on 20newsgroups with 50 topics and almost 8% with 200 topics. Overall, these results show that GSDTM has indeed the best performance among all the considered methods, being the best choice regarding topic coherence.

On the other hand, LMDTM was outperformed by GSDTM and ProdLDA. Its performance on the RCV1-v2 dataset using BN and dropout was better than on 20newsgroups dataset, with coherence levels closer to ProdLDA. These results may indicate that LMDTM is not a good fit for smaller datasets and, on the larger ones, it requires BN and dropout. Its inferior performance, in general, can be attributed to its model complexity, where each topic is represented by a distinct Logistic-Normal distribution. Thus, LMDTM has many more parameters to learn. This issue may be overcome by increasing the size of the corpus, which we intend to on future studies.

When the methods are trained without batch normalization and dropout, these changes led to visible drops in ATC levels for all methods. This is an indication that the deployment of training stability improvements on topic modeling with VAEs positively impact the coherence levels achieved by the methods. In this scenario, some methods faced topic collapsing [12] (LMDTM, ProdLDA on 20Newsgroups), and the ones that did not still underperformed. While the use of Gumbel-Softmax diminished collapse events in GSDTM, clearly the models with a Gaussian or Logistic-Normal probabilistic decoder and softmax transformation need to be trained with batch normalization and dropout to avoid topic collapsing and to achieve better topic coherence.

No method collapsed on RCV1-v2, except LMDTM. This is probably due to the size of the corpus and high complexity of the model, prone to fall into local minima. Another interesting result on RCV1-v2 is the low coherence values achieved by GSDTM without BN and dropout, especially since, as can be seen in table II, in this scenario GSDTM obtained extremely low (good) perplexity results. This clearly indicates that there is a trade-off between topic coherence and perplexity, which we will explore more on Subsection V-B. As expected, in general, all methods performed worse when finding more topics, which is a harder problem, with the exception being GSDTM on RCV1-V2 with BN and dropout, whose results were even the best overall for the dataset in terms of coherence.

B. Generative Model Evaluation

We now evaluate the quality of the generative models using a traditional topic modeling metric, the perplexity. As Miao et al. [11], we computed the perplexity per document as $exp(-\frac{1}{D}\sum_n^{N_d}\frac{1}{N_d}logp(X_d))$, where D is the number of documents, N_d represents the number of words contained in document d, and $p(X_d)$ is the likelihood of a document being drawn from the test set. We used the whole test dataset to compute the perplexity with the same parameters used in the topic evaluation experiments.

Perplexity values are presented in Table II. As observed, GSDTM achieved the best results. Its most expressive gains were observed on RCV1-v2 without BN and dropout. In comparison with NVDM, that achieved the second best results in perplexity, GSDTM was almost 25% better (lower perplexity)

[7]https://github.com/jhlau/topic_interpretability

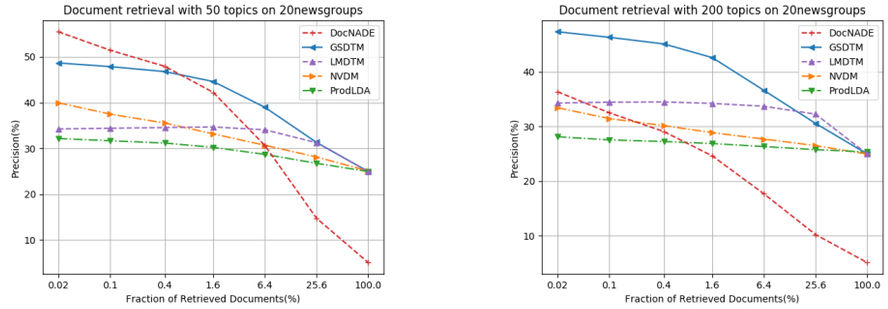

Fig. 4. Retrieval information task on 20newsgroups dataset.

when using 50 topics, and, when using 200 topics, more than 40% better than DocNADE, the second best. The gains on 20newsgroups were much more modest in comparison, which might indicate that the size of the corpus affected the quality of the generative model. On the other hand, LMDTM was outperformed by the other methods in terms of perplexity. This indicates that the use of Logistic-Normal mixtures might not be the best option to create generative models for those datasets.

In general, all methods obtained better (i.e. lower) perplexity values when *not using batch normalization and dropout*. The exception was ProdLDA, which collapsed on 20newsgroups. Thus, while using these techniques is beneficial to topic coherence, the same is not observed for document generation, which indicates that the best reconstruction model will not necessarily use the most coherent topic words.

In fact, this is a reasonable result since a real document is not composed solely by words which belong purely to certain topics across the dataset. Natural language text can be composed of several topics, with words shared between topics and some words that are not really restricted to any particular topic. A good generator has to capture all these subtleties, which can result in the choice of some words that do not fit so well on every topic observed in the document.

The use of batch normalization diminishes the impact that high frequency words might have in the modeling, decreasing the likelihood of these words being high ranked in a given topic, hence increasing its coherence. On the other hand, this might distort how these topics are representing the real dataset and how well the generative model is able to recreate their documents, increasing its perplexity.

C. Document Retrieval Evaluation

Besides the reference metrics ATC and perplexity, we also evaluate the quality of our models by using a real application. In particular, we follow the approach as Larochelle et al. [8]. It is a document retrieval task based on document representation $\mathbf{z}$ learned from encoder network $q_\theta(\mathbf{z}|\mathbf{x}, \mathbf{y})$ on the 20Newsgroups dataset, using the training instances as the document database, test instances as queries, and fixed values for fraction of retrieved documents. For each query, the documents in the database are ranked according to the cosine similarity of their vector representations. Then, the k most similar documents are retrieved and their labels are compared. The query precision is calculated as the proportion of the retrieved documents which share the query label. Finally, we averaged the query precisions over the set of queries to get the final precision.

Fig. V-A shows the precision per fraction of retrieved documents curves obtained using 50 and 200 topics on the 20newsgroups dataset, with 100 documents used as validation set. We applied the best settings found using the validation dataset, that is, learning rates equal to 0.002 for LMDTM, NVDM and ProdLDA and 0.0002 to GSDTM model. The models were trained using the ADAM optimizer with the beta2 parameter set to 0.99. We performed these experiments both with and without batch normalization and dropout. However, the results using BN and dropout were significantly worse for all methods, with precision around 20% worse in most fraction levels. Thus, due to this and to space limitations, we chose to omit those results and show only the results of the methods without BN and dropout, and we intend to further study in future works why this phenomena happened.

As we can see, GSDTM and LMDTM present a consistent performance independently on the number of topics. When using 50 topics, GSDTM outperforms all other models in 99.6% of the retrieval space, being inferior to DocNade only in the top 0.4% retrieved documents. Regarding 200 topics, GSDTM is visibly better than the other methods on up to the top 25% retrieved documents. LMDTM also had a better performance in this scenario, outperforming all the remaining methods (except DocNADE for the top 0.02%) and being the best method when considering more than the top 25%.

D. Qualitative Inspection

To complement our quantitative analysis, we also present a qualitative study on the topics obtained by the models we studied in this work. Such analysis is useful to check if GSDTM, LMDTM and other models learned meaningful semantics. To accomplish this, we followed the same approach described in Subsection V-A and extracted the topic rank from $\mathbf{W}$. The Table III shows five random topics extracted from each VAE model on the 20newsgroups dataset using 50 topics with batch normalization and dropout enabled and with $N = 10$ along with the same parameters of the topic evaluation experiment. As can be seen, GSDTM and LMDTM clearly generate coherent topics. Based on words in each line, we can

TABLE III
RANDOM FIVE TOPICS EACH ONE WITH 10 WORDS w STRONGEST CONNECTION W_{iw} COLLECTED FROM ALL VARIATIONAL AUTO-ENCODER MODELS.

GSDTM	**ProdLDA**
scsi bios controller mon pts chi drives cal nhl la	satellite nhl hockey mission app cmu udel shuttle international rocket
armenian armenians apartment turkish soldiers	armenian armenia armenians azerbaijan villages town killed
turks armenia azerbaijan villages father	turks turkish fighting
buf char printf entry output int null col stream contrib	christians jesus heaven god faith christ bible scripture church sin
contrib tar pub xt export widget mit ripem mcgill jpeg	chi van nj ai min gatech los rochester mon gary
entry buf bios printf char drives entries output int controller	jpl nasa mission images institute shuttle gov lunar sgi rocket
LMDTM	**NVDM**
ftp software unix graphics turbo amiga disks pc pub sgi	launch gif cancer install jpeg images load air fix doctor
law enforcement serial americans device agencies established criminals committed encryption	koresh compound fbi bu msg batf pp angeles van ron
serial cops chip government proposal strong law police criminals chips	catholic levels nt package aids prices insurance dos science companies
belief atheists christians god faith beliefs existence christianity religions religion	ripem ted frank rsa pgp freenet dog cwru hell weak
firearms gun drug united mph children patients study committee age	bmw ride dod bike riding motorcycle max gif ted rec

deduce various topics, such as *Hardware*, *Software*, *Religion*, *Politics* and others.

VI. CONCLUSION

In this work, we proposed two new text topic models based on VAEs, GSDTM and LMDTM, and showed experiments comparing their quality with state-of-the-art topic modeling VAEs in two reference datasets.

GSDTM is a method that uses a Gumbel-Softmax distribution, which enables it to approximate categorical distributions. It outperformed all the other baseline methods in most scenarios, achieving expressive gains up to 28% in Average Topic Coherence and up to 40% in terms of perplexity. Furthermore, in a document retrieval task its results were among the best in all scenarios evaluated. On the other hand, LMDTM adopts a mixture of Logistic-Normal distribution, in a attempt to fit more complex posterior distributions. However, in our experiments while it was competitive it was unable to outperform our baselines in most scenarios.

We also evaluated the impact of the use of dropout and batch normalization in the quality of VAEs, and came to the conclusion that the use of those techniques might lead to higher levels of topic coherence, however at the expense of the generative model document reconstruction quality. In future works we intend to further study the reasons behind this behavior and possibly design a new model able to learn how to better balance reconstruction quality and topic coherence. We also intend to propose a general framework for neural document topic models, which would enable us to explore different modeling options in a more systematic way. Regarding LMDTM, we intend to improve its training costs and study changes in its structure to make it more competitive.

ACKNOWLEDGMENT

This work was in part supported by the CAPES Foundation and FAPEAM research agency (project 062.00703/2015).

REFERENCES

[1] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *SIGIR '06*. New York, USA: ACM, 2006, pp. 178–185.

[2] P. Xie and E. P. Xing, "Integrating document clustering and topic modeling." *CoRR*, vol. abs/1309.6874, 2013.

[3] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine Learning*, vol. 88, no. 1, pp. 157–208, Jul 2012.

[4] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship attribution with topic models," *Comp. Ling.*, vol. 40, no. 2, pp. 269–310, Jun. 2014.

[5] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, "Multi-aspect sentiment analysis with topic models," in *ICDMW '11*. Washington, DC, USA: IEEE Computer Society, 2011, pp. 81–88.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[7] R. Salakhutdinov and G. E. Hinton, "Replicated softmax: an undirected topic model." in *NIPS'09*, vol. 22, 2009, pp. 1607–1614.

[8] H. Larochelle and S. Lauly, "A neural autoregressive topic model," in *NIPS'12*. Curran Associates, Inc., 2012, pp. 2717–2725.

[9] M. Yang, T. Cui, and W. Tu, "Ordering-sensitive and semantic-aware topic modeling," in *AAAI'15*. AAAI Press, 2015, pp. 2353–2359.

[10] F. Tian, B. Gao, D. He, and T. Liu, "Sentence level recurrent topic model: Letting topics speak for themselves," *CoRR*, vol. abs/1604.02038, 2016.

[11] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *ICML'16*. JMLR.org, 2016, pp. 1727–1736.

[12] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," *arXiv preprint arXiv:1703.01488*, 2017.

[13] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders." *CoRR*, vol. abs/1611.02648, 2016.

[14] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," 2016, cite arxiv:1611.01144.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." *CoRR*, vol. abs/1412.6980, 2014.

[17] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS'16*. Curran Associates, Inc., 2016, pp. 2172–2180.

[18] J. W. Rae, J. J. Hunt, I. Danihelka, T. Harley, A. W. Senior, G. Wayne, A. Graves, and T. Lillicrap, "Scaling memory-augmented neural networks with sparse reads and writes," in *NIPS'16*, 2016, pp. 3621–3629.

[19] E. J. Gumbel, "The maxima of the mean largest value and of the range," *The Annals Math. Statistics*, vol. 25, no. 1, pp. 76–84, 1954.

[20] C. J. Maddison, D. Tarlow, and T. Minka, "A* sampling," in *Advances in Neural Information Processing Systems*, 2014, pp. 3086–3094.

[21] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *CoRR*, vol. abs/1611.00712, 2016.

[22] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *EACL'14*, G. Bouma and Y. Parmentier, Eds. The Association for Computer Linguistics, 2014, pp. 530–539.