WEAKLY SEMI-SUPERVISED NEURAL TOPIC MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the problem of topic modeling in a *weakly semi-supervised* setting. In this scenario, we assume that the user knows a priori a subset of the topics she wants the model to learn and is able to provide a few exemplar documents for those topics. In addition, while each document may typically consist of multiple topics, we do not assume that the user will identify all its topics exhaustively.

Recent state-of-the-art topic models such as NVDM, referred to herein as Neural Topic Models (NTMs), fall under the variational autoencoder framework. We extend NTMs to the weakly semi-supervised setting by using informative priors in the training objective. After analyzing the effect of informative priors, we propose a simple modification of the NVDM model using a logit-normal posterior that we show achieves better alignment to user-desired topics versus other NTM models.

1 INTRODUCTION

Topic models are probabilistic models of data that assume an abstract set of topics underlies the data generating process (Blei, 2012). These abstract topics are often not only useful as feature representations for downstream tasks, but also for exploring and analyzing a corpus. Topic models are used to explore natural scenes in images (Fei-Fei and Perona, 2005; Mimno, 2012), genetics data (Rogers et al., 2005), and numerous text corpora (Newman et al., 2006; Mimno and McCallum, 2007).

While latent Dirichlet allocation (LDA) serves as the classical benchmark for topic models, recent state-of-the-art topic models such as NVDM (Miao, Yu, and Blunsom, 2016) fall under the variational autoencoder (VAE) framework (Kingma and Welling, 2013), which we refer to as Neural Topic Models (NTMs). NTMs leverage the flexibility of deep learning to fit an approximate posterior using variational inference. This posterior can then be used to efficiently predict the topics contained in a document. NTMs have been shown to model documents well, as well as associate a set of meaningful top words with each topic (Miao, Grefenstette, and Blunsom, 2017).

Often times, the top words associated with each extracted topic only approximately match the user's intuition. Therefore, the user may want to guide the model towards learning topics that better align with natural semantics by providing example documents. To our knowledge, supervision has been explored in more classical LDA models, but has not been explored yet in the NTM literature.

Labeling the existence of topics in each document across a corpus is prohibitively expensive. Hence, we focus on a weak form of supervision. Specifically, we assume a user may identify the existence of a single topic in a document. Furthermore, if a user does not specify the existence of a topic, it does not mean the topic does not appear in the document. This form of weak supervision falls under the positive-unlabeled (PU) learning framework (Denis, 1998; De Comité et al., 1999; Elkan and Noto, 2008), which we leverage to estimate parameters used for supervision.

The main contribution of our work is an NTM with the ability to leverage minimal user supervision to better align topics to desired semantics.

2 BACKGROUND: NEURAL TOPIC MODELS

Topic models describe documents (typically represented as bags-of-words) as being generated by a mixture of underlying abstract topics. Each topic is represented as a distribution over the words in a vocabulary so that each document can be expressed as a mixture of different topics and its words drawn from the associated mixture distribution.

A Neural Topic Model is a topic model constructed according to the variational autoencoder (VAE) paradigm. The generative process for a document x given a latent topic mixture υ is $\upsilon \sim p(\upsilon), x \sim p_\theta(x|\upsilon)$, where $p(\upsilon)$ is the prior and θ are learned parameters. The marginal likelihood $p_\theta(x)$ and posterior $p_\theta(\upsilon|x)$ are in general intractable so a standard approach is to introduce an approximate posterior, $q_\phi(\upsilon|x)$, and seek maximization of the evidence lower bound (ELBO):

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{q_\phi(\upsilon|x)}\big[\log p_\theta(x|\upsilon)\big] - D_{KL}(q_\phi(\upsilon|x)||p(\upsilon))}_{\text{ELBO}(x;\theta,\phi)}. \tag{1}$$

The reparamterization trick (Kingma and Welling, 2013) lets us train this model via stochastic optimization. The most commonly used reparameterization is for sampling from a Gaussian with diagonal covariance, e.g., $\upsilon = \mu(x) + \sigma(x)\epsilon,\ \epsilon \sim \mathcal{N}(0, 1)$ where $\mu(x)$ and $\sigma(x)$ are neural networks.

3 Weakly Semi-Supervised Topic Modeling

As motivated in the introduction, the topics extracted from unsupervised topic modeling can be challenging to interpret for a user (see Table 1). It is often the case that a user has a set of "ground truth" topics in mind that they would like the model to align with. We focus on the setting where a user is presented a subset of documents and identifies at least a single topic for each of the documents.

(U) rf harvesting energy chains documents consumption compression shortest texts retrieval	(S) recommendation retrieval item document items filtering documents collaborative preferences engine

Table 1: Upon scanning topics extracted by an unsupervised (U) NVDM for the AAPD ArXiv dataset, the user recognizes a few words relevant to the topic "information retrieval". Supervision (S) leveraging exemplar documents containing the topic helps the model to better align the topic to the desired semantics.

We emphasize two challenging components in our setting. First, it is *semi-supervised* in the sense that only a small subset of documents is labeled by the user. Second, the supervision is *weak* in the sense that the user does not necessarily identify all the topics that are present in a document.

We define "ground truth" word rankings for each topic using a relative chi-square association metric. Specifically, using the ground truth topic labels for each document, we select the subset of documents containing each distinct topic and compute the chi-square statistics for all words contained in those documents. We then divide the chi-square statistic for each word by the average of its chi-square statistics over all labeled topics, giving us a relative importance that is discriminative across topics. The top-10 words for each topic are then the words with highest relative importance. To evaluate alignment of topics extracted by an NTM to these "ground truth" topics, we compute the average normalized point-wise mutual information (NPMI) between each word in the NTM's topics to each word in the "ground truth" topics—a higher score indicating the words in each topic often occur together across Wikipedia. We refer to this score as pairwise-NPMI (P-NPMI). We additionally report NPMI averaged over just the supervised topics as S-NPMI, just the unsupervised topics as U-NPMI, and all topics, by an abuse of notation, as NPMI.

4 Our Approach

Our approach focuses on extending the NTM to the semi-supervised setting. Our method is based on modifying the prior, $p_i(\upsilon_i)$, of a VAE to incorporate user label information, where $p_i(\upsilon_i)$ is the prior distribution over the latent variables for the ith document. As discussed above, we assume the user provides partial binary labels indicating the existence of topics in a document (1=exists, 0=unlabeled). We would like to encourage the posterior samples of the model, $\upsilon_i \sim q_\phi(\upsilon_i|x_i)$, to match the true presence of topics in a document.

In the case of a Gaussian posterior (NVDM), a natural interpretation of υ_i is as the logits of the probabilities that each topic exists. Therefore, to recover probabilities, we simply `sigmoid` the outputs of our encoder. Given this interpretation for semi-supervision, we set $p_i(\upsilon_{ik})$ to be $\mathcal{N}(\mu_{ik}, 1)$, where

$$\mu_{ik} = \begin{cases} \texttt{logit}(0.9999) & \text{if document } i \text{ is labeled to contain topic } k, \\ \texttt{logit}(p_{k+}) & \text{else if topic } k \text{ is supervised,} \\ \texttt{logit}(0.5) = 0 & \text{otherwise,} \end{cases} \tag{2}$$

and p_{k+} is the class-prior probability, i.e. the probability that any given document contains topic k. Note that p_{k+} can be estimated from PU data (Jain, White, and Radivojac, 2016). In experiments, we instead use the mean class-prior probability for all topics ($\bar{p}_+$) and are still able to obtain high performance suggesting this approach is robust to estimation error.

In order to extract topics from the model, we condition the generative distribution, $p_\theta(x|\upsilon)$, on the logits of each of the possible approximate `onehot` vectors, e.g., υ_1=$\texttt{logit}([0.9999, 0.0001, \ldots])$ for topic 1. After conditioning on υ_k, we observe the mean of the conditional distribution, ξ_k. We then subtract from each ξ_k the mean of ξ over the topics, i.e., $\tilde{\xi}_k = \xi_k - \bar{\xi}$. The indices corresponding to the top-10 values of $\tilde{\xi}_k$ indicate the top-10 words associated with each topic k.

We also explore using two other posteriors. One is the Concrete / Gumbel-softmax approximation to the Bernoulli (Jang, Gu, and Poole, 2016; Maddison, Mnih, and Teh, 2016) which more closely matches our goal of modeling the existence of topics in documents. The differentiable posterior $q_\phi(\upsilon|x) = Concrete(\pi, \tau)$ can be sampled with Equations (3), (4), and (5):

$$u \sim U(0,1) \tag{3}$$
$$\eta = \texttt{logit}(\pi) + \texttt{logit}(u) \tag{4}$$
$$\upsilon = \texttt{sigmoid}\big(\eta/\tau\big) \tag{5}$$

$$\pi_{ik} = \begin{cases} 0.9999 & \text{if document } i \text{ is labeled to contain topic } k, \\ p_{k+} & \text{else if topic } k \text{ is supervised,} \\ 0.5 & \text{otherwise.} \end{cases} \tag{6}$$

where τ is a hyperparameter such that $\upsilon \sim Ber(\pi)$ as $\tau \to 0$, $\texttt{sigmoid}(r) = (1 + e^{-r})^{-1}$, and $\texttt{logit}(p) = \texttt{sigmoid}^{-1}(p) = \log(p/(1-p))$. The prior is $p(\upsilon_{ik}) \sim Concrete(\pi_{ik}, \tau)$ where π_{ik} is the prior probability defined in Equation (6). The log-density of $Concrete(\pi, \tau)$ is given in Section C.3.2 of (Jang, Gu, and Poole, 2016) and we use the Bernoulli-D_{KL} as an approximation. The other posterior we consider is a logit-normal distribution, for which the $\texttt{logit}(\upsilon)$ is normally distributed. Samples are easily obtained by taking the sigmoid of samples from a normal distribution.

Comparison of Kullback Liebler Divergences

Our technique of incorporating supervision via informative priors provides a supervised training signal via the D_{KL} term in Equation 1. Figure 1 compares the Bernoulli and Gaussian (NVDM) D_{KL} terms against a standard cross-entropy loss. Note D_{KL} is the same for Gaussian and logit-normal.

In Figure 1, we fix the prior distributions to $\texttt{Ber}(\pi_p$=$0.9999)$ for the Bernoulli model and $\mathcal{N}(\mu_p$=$\texttt{logit}(0.9999), \sigma_p$=$1)$ for NVDM to signify the presence of a topic with high certainty. The x-axis, p, denotes the posterior probability: π_q for the Bernoulli model and $\texttt{sigmoid}(\mu_q)$ for NVDM/logit-normal. The cross entropy loss is $-\pi_p \log p - (1-\pi_p)\log(1-p)$ where π_p is the prior probability. The posterior variance of the NVDM is assumed to match the prior variance in this figure. It is clear that NVDM D_{KL} strictly provides the strongest loss signal for topic prediction.

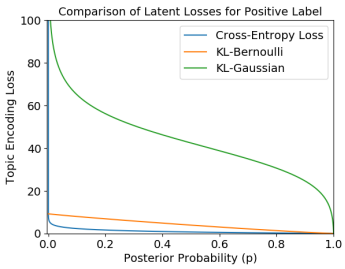

Figure 1: The Gaussian-D_{KL} provides the strongest topic encoding signal. The Bernoulli D_{KL} is approximately linear with finite divergence at $p = 0$. The log loss is very shallow except for where it diverges at $p = 0$.

On the other hand, the Bernoulli model generally achieves better topic coherence (NPMI). We expect this is due the cleaner decoding process relative to NVDM. For example, the Bernoulli decoding computes the word distribution for the first out of K topics as the softmax of $1 \cdot W_1 + \ldots + 0 \cdot W_K + b$ where $W \in \mathbb{R}^{|V| \times K}$ and $b \in \mathbb{R}^{|V|}$ are the decoder weight matrix and bias respectively, $|V|$ is the size of the vocabulary, and W_k refers to the kth row of W. In contrast, NVDM decodes with $9.2 \cdot W_1 + \ldots - 9.2 \cdot W_K + b$ where $9.2 \approx \texttt{logit}(0.9999) = -\texttt{logit}(0.0001)$.

Therefore, we propose NVDM-σ with a logit-normal posterior. The intuition is to retain both the strong supervision signal of NVDM as well as the clean decoding used by the Bernoulli model.

5 EXPERIMENTS

We train and evaluate these models on three multi-label datasets: *BibTeX*, a tagging dataset containing BibTeX metadata, *Delicious*, a document classification dataset, and *AAPD*, a dataset extracted from the abstracts and subjects of computer science ArXiv papers. We vary three experimental variables. We consider the percentage of topics supervised to be either 10%, 50%, or 100%. For example, for Delicious with 20 known topics, we consider supervising 2, 10, and all 20 dimensions of v. We also consider the number of provided labeled samples per supervised topic. For example, in one experiment, when supervising 10% of topics for Delicious, we consider providing 3 or 10 example documents per topic. Lastly, we vary the number of labels given per document. As described in Section 3, the user is not required to provide all the topics contained in each document. We consider documents with only 1 label, up to half the maximum number of labels (11 for Delicious), and fully-labeled (up to 28). Using 10% topic supervision, 10 labeled documents per topic, and 1 label per document as an experimental baseline, we vary each experimental variable independently resulting in six different settings for each of three datasets. Note each setting is still weakly semi-supervised.

In general, all models see a rise in NPMI from informative priors. We observe that NPMI rises on average by 12% relative to the unsupervised setting and by 15% beyond that to the fully supervised case. This demonstrates that weak labels are capable of providing a significant boost in performance.

Model	Top-10 Words (First Iteration)	Top-10 Words (Fully Trained)
NVDM (0.230)	transformation software attributes np necessary correlation author posterior oracle alternating	word allocation string greedy distance jointly tree sub trees latent
Bernoulli (0.219)	provides received game white numbers distinct increasingly effectiveness loss gives	greedy performance procedure allocation distance jointly speed report existing learn
NVDM-σ (0.239)	rf outperforms produces availability minimal equation correcting increased formalism centrality	allocation distance string greedy jointly correction word sub tree learn

Table 2: The supervised topic for AAPD is *Computation and Language* and the P-NPMI scores are under the model names. Only 3 documents are provided for 10% of topics with 1 label per document.

Table 2 shows successful alignment for supervised topics. Only three documents are labeled for each topic; this positive result is important as labeling numerous documents is prohibitively expensive for commercial applications. Table 3 shows NVDM-σ does best in terms of all NPMI metrics.

Metric	NVDM	NVDM-σ	Bernoulli
P-NPMI	0.852	**0.978**	0.892
NPMI	0.888	**0.960**	0.947
U-NPMI	0.915	**0.951**	0.935
S-NPMI	0.878	**0.953**	0.951

Table 3: Average performance relative to best score in each experimental setting.

6 CONCLUSION

In this work, we proposed supervising Neural Topic Models with weak supervision via informative priors and explored a variety of model posteriors. A careful analysis of their KL divergences and decoding mechanisms led us to an NTM with logit-normal posterior which best aligned extracted topics to desired user semantics.

REFERENCES

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.

De Comité, F.; Denis, F.; Gilleron, R.; and Letouzey, F. 1999. Positive and unlabeled examples help learning. In *International Conference on Algorithmic Learning Theory*, 219–230. Springer.

Denis, F. 1998. Pac learning from positive statistical queries. In *International Conference on Algorithmic Learning Theory*, 112–126. Springer.

Elkan, C., and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 213–220. ACM.

Fei-Fei, L., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, 524–531. IEEE.

Jain, S.; White, M.; and Radivojac, P. 2016. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems*, 2693–2701.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. *arXiv preprint arXiv:1706.00359*.

Miao, Y.; Yu, L.; and Blunsom, P. 2016. Neural variational inference for text processing. In *International Conference on Machine Learning*, 1727–1736.

Mimno, D., and McCallum, A. 2007. Organizing the oca: learning faceted subjects from a library of digital books. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 376–385. ACM.

Mimno, D. 2012. Reconstructing pompeian households. *arXiv preprint arXiv:1202.3747*.

Newman, D.; Chemudugunta, C.; Smyth, P.; and Steyvers, M. 2006. Analyzing entities and topics in news articles using statistical topic models. In *International conference on intelligence and security informatics*, 93–104. Springer.

Rogers, S.; Girolami, M.; Campbell, C.; and Breitling, R. 2005. The latent process decomposition of cdna microarray data sets. *IEEE/ACM transactions on computational biology and bioinformatics* 2(2):143–156.