Neural Topic Model with Attention for Supervised Learning

Author 1
Institution 1

Author 2
Institution 2

Abstract

Topic modeling utilizing neural variational inference has shown promising results recently. Unlike traditional Bayesian topic models, neural topic models use deep neural network to approximate the intractable marginal distribution and thus gain strong generalisation ability. However, neural topic models are unsupervised model. Directly using the document-specific topic proportions in downstream prediction tasks could lead to sub-optimal performance. This paper presents *Topic Attention Model* (TAM) [1], a supervised neural topic model that integrates with a recurrent neural network. We design a novel way to utilize document-specific topic proportions and global topic vectors learned from neural topic model in the attention mechanism. We also develop backpropagation inference method that allows for joint model optimisation. Experimental results on three public datasets show that TAM not only significantly improves supervised learning tasks, including classification and regression, but also achieves lower perplexity for the document modeling.

1 Introduction

Topic modeling is a frequently used data exploration tool for discovering latent semantics in a large collection of documents. Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003) is one of the most influential topic modeling examples. LDA and other statistical topic models are based on Bayesian inference Markov chain Monte Carlo (MCMC) and variational inference. Traditional Bayesian inference method becomes intractable for highly expressive models of text. Recently, neural topic models (NTM) based on variational autoencoding framework (Kingma and Welling, 2014; Rezende, Mohamed, and Wierstra, 2014) have shown promising results in document modeling (Miao, Yu, and Blunsom, 2016; Miao, Grefenstette, and Blunsom, 2017). It approximates the intractable distributions over the latent variables using a deep neural network (variational autoencoder), and thus gain nonlinear complex representations for documents with strong generalisation abilities.

In addition to data exploration, unsupervised topic models LDA or NTM also act to reduce data dimension. Therefore, the learned low dimensional document-specific topic representations are usually used for downstream supervised tasks such as classification and regression. However, fitting unsupervised topics may be sub-optimal for the supervised task, as the side information of the documents, such as the category of a document or a numerical rating of a movie review, is not used in discovering the low-dimensional topic representations of the documents. Therefore, there are existing work that extends LDA to supervised learning tasks by utilizing document categorical or numerical labels (Mcauliffe and Blei, 2008; Chong, Blei, and Li, 2009; Lacoste-Julien, Sha, and Jordan, 2009; Zhu, Ahmed, and Xing, 2012; Ramage et al., 2009; Chen et al., 2015). However, the literature is still missing for extending neural topic model for supervised learning tasks.

In this work, we consider the problem of topic modeling in a supervised setting, where each document is paired with a response label, either categorical or numerical. We present *Topic Attention Model* (TAM), a neural topic model for supervised learning, i.e., classification and regression tasks. Our proposed method is an integration of recurrent neural network and neural topic model by optimizing a single objective function using variational autoencoding framework. Specifically, the global topic vectors learned from neural topic model are used as attention queries in the recurrent neural network and the resulting attention weights are

[1] The implementation will be released at url.

Preliminary work. Under review by AISTATS 2020. Do not distribute.

averaged by the document-specific topic proportions. In this context, attention mechanism (Bahdanau, Cho, and Bengio, 2015) offers a natural way to bridge the unsupervised neural topic model with supervised RNN model. The topic relevant information is captured by the variational autoencoder and helps to attend topic keywords from the document. Moreover, the response label can in turn improve the latent topic structure learned by variational autoencoder for better document modeling. The interplay yields latent topic representations more suitable for supervised prediction tasks. An effective variational inference method is developed using backpropagation so that the model parameters can be jointly learned.

Our topic attention model TAM combines the merits of both neural topic model and attention RNN. As a supervised learning model, it obtains better document representations and achieves higher prediction accuracy. Attention mechanism also allows us to investigate the keywords that have high impact to the prediction outcome, which provides some degree of interpretability. As a supervised topic model, the improvement on prediction does not come at the cost of document modeling quality. In fact, it fits the document data better than unsupervised topic model. The estimated document-specific topic proportions and global topic vectors can also facilitate corpus exploration.

2 Related Work

For an overview of statistical topic modeling, see Blei (2012). Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003) is the most widely cited statistical topic model. Prior research on supervised topic models mostly focuses on extending unsupervised LDA to supervised tasks. Mcauliffe and Blei (2008) proposes supervised LDA (sLDA) by adding to LDA a response variable associated with each document and then defining for the variable a Gaussian distribution whose mean value is computed by a linear regression of topics. Chong, Blei, and Li (2009) extends sLDA that is used for regression to multi-class classification. DiscLDA (Lacoste-Julien, Sha, and Jordan, 2009) and MedLDA (Zhu, Ahmed, and Xing, 2012) share the same goal with sLDA but differ in the training procedure. DiscLDA is trained to maximize the conditional likelihood of response variables, and MedLDA utilizes the max-margin principle to jointly train LDA with SVM. Similarly, Wang and Zhu (2014) proposes a spectral decomposition algorithm that more efficiently estimates parameters of sLDA. Labeled LDA (Ramage et al., 2009) utilizes document labels and makes a strong assumption that each document can be only represented by the associated topics, and it only works with categorical response variable. On the other hand, some existing work exploits deep neural network structure to extend LDA for supervised task. For example, BP-SLDA (Chen et al., 2015) introduces a supervised LDA model using back propagation, and Cao et al. (2015) tackles the bags-of-words assumption of topic model and uses neural network to learn the word-topic and topic-document distributions. TopicRNN (Dieng et al., 2017) incorporates LDA with a RNN model.

While statistical topic models are based on Bayesian learning, recent advance in neural topic models fall under the variational autoencoder (VAE) framework (Kingma and Welling, 2014; Rezende, Mohamed, and Wierstra, 2014). Existing neural topic modeling work (Miao, Yu, and Blunsom, 2016; Miao, Grefenstette, and Blunsom, 2017; Srivastava and Sutton, 2016) leverages the generalizability of deep learning (VAE) to fit an approximate posterior using variational inference. However, these neural topic models are unsupervised model, and it is still unknown how to integrate side information (such as document labels) into the models. Towards this end, Ding, Nallapati, and Xiang (2018) proposes to incorporate a topic coherence objective into neural topic modeling training so that the topics are more human-readable, and Gemp et al. (2019) presents a weakly semi-supervised extension so that users can explicitly provide a subset of topics that they want the model to learn.

Our work is built upon recent neural topic models using VAE framework (Miao, Grefenstette, and Blunsom, 2017; Srivastava and Sutton, 2016). We extend those work by integrating an RNN model with the neural topic model for supervised learning. To the best of our knowledge, our work is the first supervised topic model where the topic model is trained with neural variational inference.

3 Model

We first give a brief introduction on neural topic model using variational autoencoding. Then we describe the details of our topic attention model that jointly optimizes topic modeling and supervised learning.

3.1 Neural Topic Model

Neural topic model (Miao, Yu, and Blunsom, 2016; Miao, Grefenstette, and Blunsom, 2017; Srivastava and Sutton, 2016) utilizes the variational autoencoder (VAE) paradigm to model the documents generative process and uses gradient ascent to maximize its evidence lower bound (ELBO). Below we describe Gaussian Softmax distribution model (GSM) (Miao, Grefenstette, and Blunsom, 2017).

Suppose document d contains N_d word tokens

$\{x_1, x_2, ...x_{N_d}\}$, and document d is associated with a response label l_d. l_d can be either categorical such as document class or numerical such as review rating. We use l to replace l_d for notation simplicity. $\boldsymbol{d}$ is the bag-of-word representation of for document d. We use latent variable $\boldsymbol{t} \in \mathbb{R}^K$ to denote the topic proportion of document, where K is the number of topics. z_n is the topic variable assigned to word token x_n. Suppose the vocabulary size for the topic model is V_{topic}. GSM uses a neural network to parameterise the multinomial topic distribution, and the generative process for document d is:

$$\omega \sim \mathcal{N}(\mu_0, \sigma_0^2) \qquad \boldsymbol{t} = \text{softmax}(W_\omega \omega + b_\omega)$$
$$z_n \sim \text{Multi}(\boldsymbol{t}) \qquad x_n \sim \text{Multi}(\beta_{z_n})$$

The prior $p(\omega) = \mathcal{N}(\mu_0, \sigma_0^2)$ is a diagonal Gaussian distribution with mean μ_0 and σ_0^2 as the diagonal of its covariance matrix. By using the Gaussian prior distribution, GSM can employ the re-parameterisation trick (Kingma and Welling, 2014) and build an unbiased gradient estimator for the variational distribution. Moreover, without using conjugate prior as in LDA, GSM can update the model parameters directly from the variational lower bound. In GSM, $\boldsymbol{\beta}$ is explicitly defined as the topic distribution over words, i.e. $\beta_{ij} = p(w_j|t_i)$, where w_j is the j-th word from the topic model vocabulary, t_i is the i-th topic. W_ω, b_ω are trainable parameters. The parameters in this generative part is denoted by Θ.

Following the framework of neural variational inference, the inference network is:

$$\mu(\boldsymbol{d}) = g_1(f(\boldsymbol{d})) \qquad \log(\sigma^2(\boldsymbol{d})) = g_2(f(\boldsymbol{d}))$$

Here, f, g_1, g_2 are fully-connected neural networks with batch normalizaton and drop-out. The parameters in this variational part is denoted by Φ. The variational distribution $q_\Phi(\omega|\boldsymbol{d}) = \mathcal{N}(\mu(\boldsymbol{d}), \sigma^2(\boldsymbol{d}))$ is a diagonal Gaussian distribution used to approximate the true distribution $p(\omega|\boldsymbol{d})$. To inference the parameters of the neural topic model, the evidence lower bound (ELBO) of $\log p(\boldsymbol{d}|\mu_0, \sigma_0, \boldsymbol{\beta})$ is derived and served as the objective of back-propagation.

3.2 GRU-based Sequence Encoder

Our goal is to integrate neural topic model GSM with a supervised model. Thus, sequential recurrent neural network (RNN) serves our purpose.

Suppose the input of the RNN is a sequential word tokens of document d. First we convert it into a word embedding sequence $X^* = (x_1, ..., x_N)$ via a trainable word embedding matrix $M \in \mathbb{R}^{V_{RNN} \times D}$, where V_{RNN} is the vocabulary size for RNN input and D is the dimension of the word embedding. It worth noting that we have two vocabularies. V_{topic} is the bag-of-words vocabulary for neural topic model, and V_{RNN} is the vocabulary for RNN. The two vocabularies are not necessarily the same size.

We adopt Bi-directional Gated Recurrent Unit (GRU) to encode the embedding sequence. See (Cho et al., 2014) for a detailed mathematical description for the GRU gating mechanism. Here we use $GRU()$ to denote the GRU transformation for simplicity. Bi-directional GRU captures both the contextual information from the previous text and the later text in a document. Give a timestep t, the hidden state $\boldsymbol{h}_t$ can be computed by concatenating forward hidden state $\overrightarrow{h_t}$ and backward hidden state $\overleftarrow{h_t}$ together using the current input x_t and the previous hidden state h_{t-1}:

$$\boldsymbol{h}_t = \begin{bmatrix} \overrightarrow{h_t} \\ \overleftarrow{h_t} \end{bmatrix} = \begin{bmatrix} \overrightarrow{GRU}(x_t, \overrightarrow{h_{t-1}}) \\ \overleftarrow{GRU}(x_t, \overleftarrow{h_{t-1}}) \end{bmatrix}$$

3.3 Topical Attention Model: TAM

The proposed TAM model takes two forms of inputs of documents: a bag-of-word representation for GSM and a sequence of word tokens for RNN. The TAM framework is shown in Figure 1. Neural topic model GSM is used to fit document generative process and estimate document-specific topic distribution $\mathbf{t}$. Each sequential word tokens x_t is encoded to hidden states $\boldsymbol{h}_t$ via the GRU-based sequence encoder. Next, we propose to use attention mechanism to bridge two components, so that both models can be jointly optimized.

The attention mechanism is originally proposed by (Bahdanau, Cho, and Bengio, 2015) in machine translation. Attention mechanism calculates the similarity between a context vector (query) and each key (word) to obtain the attention score corresponding to the key. The trainable context vector can be seen as a high level representation of a fixed query "what is the informative word" in the sequence.

Recall that in the neural topic model, $\boldsymbol{\beta}$ is the topic distribution over words. However, $\boldsymbol{\beta}$ is not regarded as a single parameter matrix in the generative process, and can be written as the product of two smaller matrices:

$$\boldsymbol{\beta} = \text{softmax}(EF^T) \tag{1}$$

For a chosen positive integer $L < V$, $E = (v_1, v_2, ..., v_K)^T \in \mathbb{R}^{K \times L}$ and $F \in \mathbb{R}^{V \times L}$. Through this matrix factorization process, v_i carries some useful information of the i-th topic in the topic model but with a much shorter length than β_i. We call this vector v_i the **topic vector**, or global topic embedding. In our experiment, we use $L = 100$ as the dimension for topic vectors.

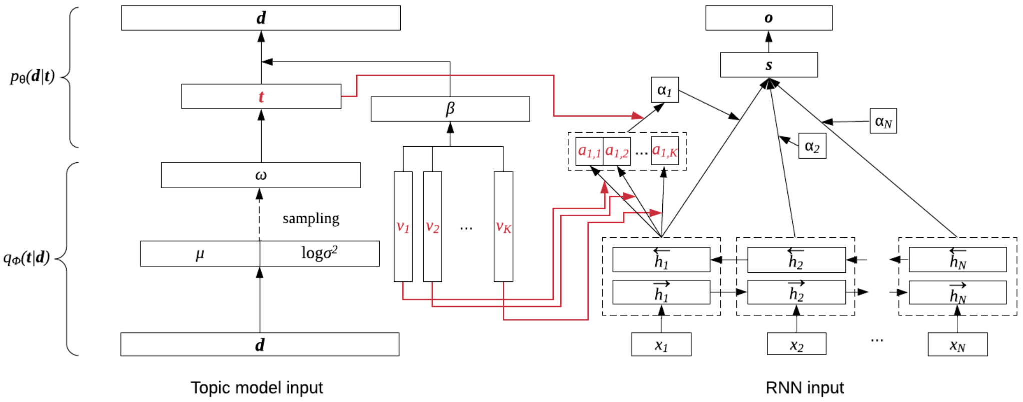

Figure 1: Topical Attention Model structure. The left part is a unsupervised neural topic model learned by variational autoencoding, and the right part is a supervised RNN model where input words are encoded by Bi-GRU. Two parts are integrated in the attenion mechanism by the topic vectors $v_i, i \in [1, K]$ and document-specific topics $\boldsymbol{t}$. Two separate parts are learned jointly using backpropagation inference.

In TAM, instead of using only one context query, we use multiple topic vectors as queries. These queries learned by neural topic model can help the supervised model to focus on the information regarding to different aspects of the corpus from a global topic perspective, which may otherwise be overlooked by the RNN model that mostly focuses on the local contextual information. An intuitive explanation is that these topic vectors are used as queries to resemble "what is the informative word under this topic". After calculating attention weights with respect to all the topics, we then average them by a shifted topic proportion, which not only makes the final attention weights emphasize on the most importation topics, but also diversifies different topical attentions by backpropogation.

For a topic vector v_i $(i \in \{1, 2, .., K\})$, and step $t \in \{1, 2, ..., N\}$, the i-th topical attention weight at the t-th step is:

$$u_t = \tanh(W_t \boldsymbol{h}_t + b_t) \tag{2}$$

$$a_{t,i} = \text{softmax}(u_t^T v_i) \tag{3}$$

Where u_t has the same dimension as v_i. To put the K topical attentions at the t-th step together as a scalar, we average them by a shifted topic mixture:

$$\alpha_t = \sum_{i=1}^{K} a_{t,i}(t_i - \delta) \tag{4}$$

Where $0 < \delta < 1$ is a scalar subtracted from each entry of $\boldsymbol{t}$. By applying this set of attention weights to the hidden states, we get the final encoded document:

$$\boldsymbol{s} = \sum_{t=1}^{N} \alpha_t \boldsymbol{h}_t \tag{5}$$

The encoded document can then be fed into a fully-connected layer for prediction:

$$\boldsymbol{o} = f(W_o \boldsymbol{s} + b_o) \tag{6}$$

Where W_o, b_o are trainable parameters and f is an activation function depend on the nature of the labels.

The reason that we shift the topic mixture by a constant is to make the resulting topical attention weight more diverse among different topics. Recall that $\boldsymbol{t} = (t_1, ..., t_K)^T$ is the topic mixture of the document, so $0 \leq t_i \leq 1$ for $i \in \{1, 2, ..., K\}$. If a constant $0 < \delta < 1$ is subtracted, the larger entry of $\boldsymbol{t}$ would remain positive, while the smaller entry would become negative. Suppose our objective is to minimize the loss function $\mathcal{L}$, then we have the update equation for $a_{t,i}$ by gradient descent:

$$a_{t,i} \leftarrow a_{t,i} - \gamma \frac{\partial \mathcal{L}}{\partial a_{t,i}}$$

Where γ is the learning rate. We can rewrite the partial derivative using the chain rule:

$$\frac{\partial \mathcal{L}}{\partial a_{t,i}} = (t_i - \delta) \frac{\partial \mathcal{L}}{\partial \alpha_t}$$

Note that $\frac{\partial \mathcal{L}}{\partial \alpha_t}$ is the same for $a_{t,1}, a_{t,2}, ..., a_{t,K}$. So the gradient with respect to topical attentions would have different signs for topics with a higher probability in the topic mixture and those with a lower probability in the topic mixture. By subtracting δ, topical attention for different topics can be separated more by learning in opposite directions, otherwise the gradients would all have the same sign.

Given document d and its label l, our objective is to maximize the joint marginal likelihood:

$$p(l, \boldsymbol{d}|\mu_0, \sigma_0, \beta) = \int_{\boldsymbol{t}} p(\boldsymbol{t}|\mu_0, \sigma_0) p(l, \boldsymbol{d}|\boldsymbol{t}, \beta) d\boldsymbol{t} \tag{7}$$

3.4 Model Inference

Since l is actually independent of the bag-of-word representation $\boldsymbol{d}$ and only depend on the RNN input and $\boldsymbol{t}$, we can factorize the joint distribution:

$$p(l, \boldsymbol{d}|\mu_0, \sigma_0, \beta) = \int_{\boldsymbol{t}} p(\boldsymbol{t}|\mu_0, \sigma_0) p(\boldsymbol{d}|\boldsymbol{t}, \beta) p(l|\boldsymbol{t}) d\boldsymbol{t} \tag{8}$$

The direct optimization of the above integral is intractable so we use variational inference to apporximate this marginal. We reparameterize the log-likelihood and derive its evidence lower bound (ELBO):

$$\begin{aligned}\log p_{\Theta,\Psi}(l, \boldsymbol{d}) &= \log \int_{\boldsymbol{t}} \frac{p_\Theta(\boldsymbol{t})}{q_\Phi(\boldsymbol{t}|\boldsymbol{d})} q_\Phi(\boldsymbol{t}|\boldsymbol{d}) p_\Psi(l|\boldsymbol{t}) p_\Theta(\boldsymbol{d}|\boldsymbol{t}) d\boldsymbol{t} \\ \geq & \mathbb{E}_{q_\Phi(\boldsymbol{t}|\boldsymbol{d})} \big[\log p_\Theta(\boldsymbol{d}|\boldsymbol{t}) + \log p_\Psi(l|\boldsymbol{t}) \big] - \mathrm{KL}(q_\Phi(\boldsymbol{t}|\boldsymbol{d}) || p(\boldsymbol{t}))\end{aligned} \tag{9}$$

Due to space limit, we put the detailed derivation in the Supplementary Material. Here, Ψ represents all the parameter from the RNN attention model. Θ is the generative parameters and Φ is the variational parameters. Therefore, the expectation can be estimated by a single sample:

$$\hat{\mathcal{L}} = \log p_\Theta(\boldsymbol{d}|\hat{\boldsymbol{t}}) + \log p_\Psi(l|\hat{\boldsymbol{t}}) - \mathrm{KL}(q_\Phi(\boldsymbol{t}|\boldsymbol{d}) || p(\boldsymbol{t})) \tag{10}$$

The objective of TAM (Eq. 7) becomes maximizing $\hat{\mathcal{L}}$, which can be broken down into three parts: log data likelihood $\log p_\Theta(\boldsymbol{d}|\hat{\boldsymbol{t}})$, log label likelihood $\log p_\Psi(l|\hat{\boldsymbol{t}})$ and topic KL-divergence $\mathrm{KL}(q_\Phi(\boldsymbol{t}|\boldsymbol{d}) || p(\boldsymbol{t}))$.

Data likelihood: $\log p_\Theta(\boldsymbol{d}|\hat{\boldsymbol{t}})$ can be directly calculated by marginal out the latent variable z_n:

$$\begin{aligned}\log p_\Theta(\boldsymbol{d}|\hat{\boldsymbol{t}}) &= \sum_{n=1}^{N} \log p(x_n|\hat{\boldsymbol{t}}) \\ &= \sum_{n=1}^{N} \sum_{z_n} \log p(x_n|\beta_{z_n}) p(z_n|\hat{\boldsymbol{t}}) = \sum_{n=1}^{N} \log \beta_{x_n}^T \hat{\boldsymbol{t}}\end{aligned} \tag{11}$$

Where β_{x_n} is the corresponding column of β such that the corresponding word is x_n.

Topic KL-divergence: Since $\boldsymbol{t} = \mathrm{softmax}(W_\omega \omega + b_\omega)$ in neural topic model, we have:

$$p(\boldsymbol{t}) = p(\omega) = \mathcal{N}(\mu_0, \sigma_0^2) \tag{12}$$

$$q_\Phi(\boldsymbol{t}|\boldsymbol{d}) = q_\Phi(\omega|\boldsymbol{d}) = \mathcal{N}(\mu(\boldsymbol{d}), \sigma^2(\boldsymbol{d})) \tag{13}$$

So $\mathrm{KL}(q_\Phi(\boldsymbol{t}|\boldsymbol{d}) || p(\boldsymbol{t}))$ can be computed analytically as a Gaussian KL-divergence.

Label likelihood: Classification and regression are two common supervised learning tasks for documents. Therefore, we focus on three supervised task: multi-class classification (MCC), multi-label classification (MLC) and regression, and we describe the label likelihood for each task accordingly. For MCC and MLC, we assume that the label distribution conditioning on the RNN input data and the topic variable $\boldsymbol{t}$ is a multinomial distribution. For MCC, we use a softmax activation on the output, then:

$$p_\Psi(l|\hat{\boldsymbol{t}}) = \boldsymbol{o}_l \tag{14}$$

Where $\boldsymbol{o}_l$ is the entry of output corresponding to the correct label l. For MLC, we use a sigmoid activation on the output. Suppose there C classes in total, then:

$$p_\Psi(l|\hat{\boldsymbol{t}}) = \prod_{i=1}^{C} p_\Psi(l_i|\hat{\boldsymbol{t}}) = \prod_{i=1}^{C} \left[l_i \boldsymbol{o}_i + (1 - l_i)(1 - \boldsymbol{o}_i) \right] \tag{15}$$

We abuse l_i as the indicator function of whether the i-th class is assigned as a tag. Note that the negative logarithm of the label likelihood here is the same as the cross-entropy.

For regression task, we assume that $p_\Psi(l|\hat{\boldsymbol{t}})$ is a Gaussian distribution with mean $\boldsymbol{o}$ and variance σ_l^2:

$$p_\Psi(l|\hat{\boldsymbol{t}}) = \mathcal{N}(\boldsymbol{o}, \sigma_l^2) \tag{16}$$

Where σ_l^2 is the variance of the label likelihood and is treated as a hyper-parameter to be determined. Then we have:

$$\log p_\Psi(l|\hat{\boldsymbol{t}}) = -\frac{(l - \boldsymbol{o})^2}{2\sigma_l^2} - \frac{1}{2} \log \left[2\pi\sigma_l^2 \right] \tag{17}$$

Since the objective is to maximize $\mathcal{L}$ with respect to Θ, Φ, Ψ, we can ignore the second term as it is a constant.

Optimization: The objective $\hat{\mathcal{L}}$ is maximized by stochastic gradient ascent and all the parameters are jointly updated. It is intuitive to derive the gradient with respect to Ψ and Θ since all the parameters are organized by arithmetic operations. But when it comes to Φ, ω is an random variable following Gaussian distribution and the gradient with respect to it cannot be directly calculated. To estimate this gradient, following (Miao, Grefenstette, and Blunsom, 2017), we sample one sample $\hat{\omega}$ and then use the fact $\omega = \mu(\boldsymbol{d}) + \sigma(\boldsymbol{d})\epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$. That is for a sample $\hat{\epsilon}$, we have: $\frac{\partial \hat{\mathcal{L}}}{\partial \mu(\boldsymbol{d})} \approx \frac{\partial \hat{\mathcal{L}}}{\partial \hat{\omega}}$, and $\frac{\partial \hat{\mathcal{L}}}{\partial \sigma(\boldsymbol{d})} \approx \hat{\epsilon} \frac{\partial \hat{\mathcal{L}}}{\partial \hat{\omega}}$. Then we can perform the back propagation on parameters of neural networks f, g_1, g_2.

4 Experiments

4.1 Datasets and settings

We conduct experiments on three public datasets: 20 Newsgroups, Movie Review Dataset (MRD), and Wiki10. 20 Newsgroups is a classic dataset with 20 different classes for multi-class classification. Movie Review Data (MRD) (Pang and Lee, 2005) is a collection of movie reviews with scaled sentiment scores ranging from 0 to 1. Wiki10 (Zubiaga, 2012) is a commonly used dataset for multi-label text classification. It contains a collection of Wikipedia articles labeled with multiple tags and have longer length compared to the other two datasets. For our experiments, we use the 100 most common tags and delete some generic tags like 'wiki' and 'wikipedia', which yields 73 tags in total, and articles without tags in this set are deleted. Each article is labeled with 3.37 tags on average. The dataset statistics are shown in Table 1.

dataset	20news	MRD	Wiki10
#train	11,218	3,337	12,966
#test	7,452	1,669	5,558
#vocab	18,563	17,007	84,693
avg. #words	129.0	188.6	1020.5
task	MCC	Regression	MLC

Table 1: Dataset statistic descriptions.

For all the datasets, we tokenize the texts by removing all punctuation, numbers, stopwords and some rare words appearing less than 5 times in the training set. Note that TAM takes two forms of input: a sequence model and a bag-of-words topic model. Therefore, for the sequence model RNN input, we use the whole vocabulary, and the word embedding matrix is initialized by the word embeddings learned from Word2Vec (Mikolov et al., 2013) trained on the training set. For the topic model input, we choose the most frequent 2,000 words as the vocabulary for 20 newsgroups[2], MRD and Wiki10. Since 20 newsgroups and MRD have shorter documents, we use a sequence length of 500 for RNN and hidden size of 64 for both GRU and fully-connect layers in the topic model. For Wiki10, we use a sequence length of 1,000 for RNN and a hidden size of 256 for GRU. Batch size is 64 for 20 Newsgroups and MRD, 128 for Wiki10. Note that δ is an important hyper-parameter and should be proportion to $1/K$ to maintain a consistent performance for different K. In the experiment, for 20NG and MRD, $\delta = 0.2, 0.1, 0.05$ for $K = 25, 50, 100$ respectively. For Wiki10+, $\delta = 0.1, 0.05, 0.025$ for $K = 25, 50, 100$ respectively. For MRD, we take Gaussian label noise with $\sigma_l = 0.1$.

4.2 Baselines

We consider the following three groups of baselines: unsupervised topic models, supervised topic models, and one ablation baseline.

The two unsupervised topic model baselines are:

Latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan, 2003) is the most cited topic modeling work, and we use the online LDA implementation.

Gaussian softmax model (GSM) (Miao, Grefenstette, and Blunsom, 2017) is the neural topic model that we integrate into TAM.

LDA and GSM are the representative unsupervised topic model that is based on Bayesian learning and variational autoencoding framework respectively. In our experiment, we adopt a standard two-step procedure for the supervised learning tasks. This procedure fits the training data to a topic model (either LDA or GSM), and then use the latent topic representation of the training documents as features to build a SVM classifier or SVR regression model.

The four supervised topic model baselines are: [3]

Supervised LDA (sLDAr) (Mcauliffe and Blei, 2008) is the first LDA extension for supervised regression task. **sLDAc** (Chong, Blei, and Li, 2009) extends sLDAr to multi-class classification tasks. In our experiment, we use sLDAc for multi-class and multi-label classification, and we use sLDAr for regression. We denote both as **sLDA**.

Labeled LDA (L-LDA) (Ramage et al., 2009) is a supervised LDA extension. It assumes that each category label corresponds to a topic and that each document can use only topics that are in its label set.

MedLDA (Zhu, Ahmed, and Xing, 2012) combines LDA with SVM for classification tasks. We use the online inference version of MedLDA (Shi and Zhu, 2017). It is not used in regression task.

Backpropagation Supervised LDA (BP-SLDA) (Chen et al., 2015) is a supervised LDA model using back propagation over a deep architecture.

In addition to the above unsupervised/supervised topic model baselines, we also consider an ablation baseline **Attention-RNN**. It is a Bi-directional recurrent neural network composed of GRUs with atten-

[2] For 20 Newsgroups data, we adopt the vocabulary provided by (Srivastava and Sutton, 2016) for direct comparison.

[3] Note that sNTM (Cao et al., 2015) and TopicRNN (Dieng et al., 2017) are two related baselines that incorporate LDA with a neural network structure. However, our own implementations are unsuccessful so we do not report them as baselines.

dataset	20NG/(Accuracy)			MRD/(pR^2)			Wiki10(F_1)		
K	25	50	100	25	50	100	25	50	100
LDA	0.479	0.558	0.567	0.174	0.142	0.185	0.175	0.231	0.261
sLDA	0.588	0.689	0.693	0.478	0.490	0.482	0.312	0.334	0.341
MedLDA	0.679	0.693	0.544	–	–	–	0.352	0.388	0.410
BP-SLDA	0.602	0.621	0.662	0.465	0.435	0.463	0.347	0.339	0.330
L-LDA	–	0.345	–	–	–	–	–	0.414	–
Attention-RNN	–	0.730	–	–	0.597	–	–	0.491	–
GSM	0.300	0.349	0.264	0.152	0.144	0.097	0.181	0.172	0.241
TAM (This work)	**0.762**	**0.760**	**0.765**	**0.598**	**0.602**	**0.598**	**0.517**	**0.517**	**0.519**

Table 2: Supervised Learning evaluation of different methods on different datasets. For 20NG and Wiki10, Labeled LDA are evaluated with $K = 20$ and $K = 73$ only, as Labeled-LDA requires that number of topics equals to number of labels. Attention-RNN has no topic variable thus only one number is reported.

tion. Our implementation is a variant of Bi-directional RNN as proposed in (Bahdanau, Cho, and Bengio, 2015). We encode word sequence using the same word embeddings as in TAM. Attention-RNN does not use topic vectors as queries, instead it utilizes a single randomly initialized learnable query. With this baseline, we aim to understand the benefit of using learnable topic vectors in the attention mechanism.

4.3 Supervised Learning Performance

The main results for supervised learning are present in Table 2. Each model is run for five times, and the average number is reported.

Multi-class Classification. We conduct multi-class classification experiment on 20 Newsgroups dataset with all twenty categories. We use *accuracy*, the fraction of correct predictions, to evaluate the MCC performance.

Multi-label Classification. We conduct multi-label classification experiment on Wiki10 dataset. We use the F_1 score, the harmonic mean of the precision and recall, to evaluate the MLC performance.

Regression. We conduct regression experiment on the Movie Review dataset. We use the *predictive* R^2 *(*pR^2*)* to evaluate the regression performance. Following Mcauliffe and Blei (2008), the pR^2 measures how well a regression model predicts numerical responses for testing samples.

From the table, we can clearly draw the conclusion that TAM outperforms neural topic model GSM on prediction accuracy by large margin. For example, on movie review dataset, at $K = 50$, TAM achieves 0.602 pR^2 while GSM only has 0.144. This significant improvement can be contributed to the topic attention mechanism that integrates RNN with GSM. On the other hand, we observe that TAM also outperform vanilla Attention-RNN model on three datasets. For example, on Wiki10 dataset, at $K = 50$, the F_1 score is 0.491 for Attention-RNN and 0.517 for TAM. Note that TAM extends the Attention-RNN model by feeding multiple topic vectors as the query vectors. This result indicates that the global topic information captured by the neural topic model can offer additional utility so that the predictive document representations are more suitable for supervised learning tasks.

Another interesting result worth noting is that GSM underperforms LDA in the supervised learning tasks on three datasets, although it achieves better document modeling (see Section 4.4) in terms of perplexity. We leave it for future study.

4.4 Document Modeling Performance

We evaluate document modeling performance using perplexity on the testing dataset. In document modeling, perplexity is computed as a function of the data log-likelihood of a held-out test set: $exp(-\frac{1}{D_{test}}\sum^{N_d}\frac{1}{N_d}\log p(\mathbf{d}|\boldsymbol{t}))$, where D_{test} is the number of testing documents, N_d denotes the number of tokens in testing document d, and $\log p(\mathbf{d}|\boldsymbol{t})$ is the log likelihood of the words in the document.

We presents the test document perplexities of the topic models on the three datasets in Table 3. For direct comparison, we only present the results of LDA, sLDA and GSM. Amongst different models, TAM achieves the lowest perplexity in all cases, and GSM are also significantly better than the benchmark LDA ans sLDA. We confirm with prior finding that the supervised topic model that leverages useful side information can in fact achieve better document modeling. While TAM improves prediction accuracy, the improvement does *not* come at the cost of document modeling quality.

4.5 Qualitative Evaluation

Topics Quality. We first examine the discovered topic qualities. Table 4 shows the top 10 keywords of four topics in the MRD dataset. Since the task for this dataset is to predict ratings associated with the reviews, it is critical that the topic model can differen-

dataset	20NG		MRD		Wiki10	
K	25	50	25	50	25	50
LDA	1075	1100	1415	1714	1291	1137
sLDA	1041	1033	928	933	1028	1014
GSM	858	886	919	915	929	911
TAM	**820**	**833**	**915**	**902**	**918**	**899**

Table 3: Perplexity evaluation of different methods on different datasets. The vocabularies used for the three datasets are all set to be most frequent 2,000 words.

tiate words without regard to genre. We can see that TAM is capable of distinguish "good comedy" topic from "bad comedy" topic, and "good thriller" topic from "bad thriller" topic.

good comedy	**bad comedy**	**good thriller**	**bad thriller**
comedy	jokes	noir	van
funny	lame	western	awful
laughs	save	noted	pointless
got	theaters	sperb	reasonably
tv	liners	chilling	mess
good	wait	outstanding	supposed
humor	awful	ford	driven
films	car	morality	nasty
made	bland	eerie	william
parody	badly	harris	promise

Table 4: The topics learned by TAM on the movie review dataset. The labels on the top row are generated manually by inspecting the keywords.

We also visualize topic vectors learned from TAM. As described in section 3, the dimension of global topic vectors are chosen to be 100. We use t-SNE to project the global topic vectors (i.e. $(v_1, v_2, ..., v_K)$) of 20Newsgroups dataset, as shown in Figure 2. It shows that topics that are semantically similar, such as *electronics*, *Windows* and *hardwares*, are closer in the embedding space, while they are further away from semantically dis-similar topics such as *mideast*, *politics* and *guns*.

To further investigate the quality of neural topic model, we use t-SNE to project the estimated topic distribution (i.e. $q(\theta|d)$) of each document in 20Newsgroups test dataset, as shown in Figure 2. Qualitatively we observe that the projections form several natural clusterings for documents with the same category label, as the same color is more clustered.

Topic Attention. Recent work in NLP has utilized the attention weight to highlight words that are highly correlated with the prediction outcome (Bahdanau, Cho, and Bengio, 2015), to provide some degree of interpretability. Here, we visualize the attention weights (Eq.4) for one document in multi-label Wiki10 dataset, in Figure 3. The corresponding labels of the document is *health, design, reading* and *science*. We can see that word tokens in the sentence are assigned to different weights under different topic vectors. In the top sentence, word "pharmacology" has very high attention weights under the red (medical) topic, while the weights of other words under the blue (government) topic is merely zero. In the bottom sentence, words "orgnaization", "billion people" and "drinking" have high weights under the blue (government) topic. The results indicate that TAM is capable of matching different topic vectors to different keywords in the sentence sequence, and thus results in better document representation for downstream prediction tasks.

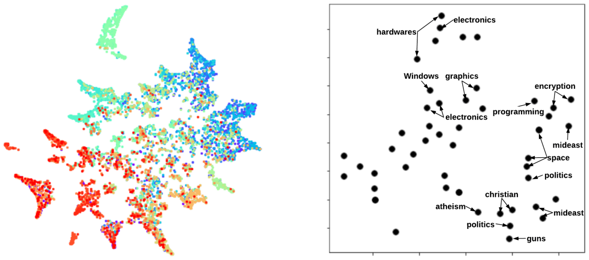

Figure 2: Left: t-SNE projection for document-specific topic proportions of 20Newsgroups documents. Documents within the same category are annotated with the same color. Right: t-SNE projection for 50 topic vectors of 20Newsgroups.

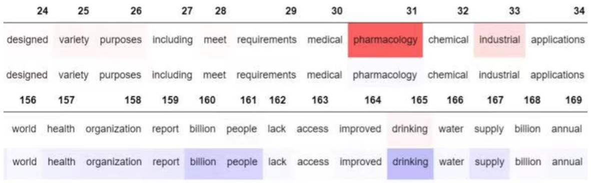

Figure 3: The number on the top is the token index in the document. Two sentences from a document and the corresponding word attention weights are shown. We show two topic vectors: medical topic (red) and politics topic (blue). Denser color indicates greater attention weight.

5 Conclusion

In this paper, we present TAM, a supervised neural topic model that integrates unsupervised neural topic modeling with supervised neural network, by using a novel design in the attention mechanism. This integration yields a predictive document representation that is more suitable for classification or regression. We develop efficient variational inference method for TAM. The empirical results on several standard datasets demonstrate the effectiveness of TAM on prediction accuracy and document modeling.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR.*

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

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

Cao, Z.; Li, S.; Liu, Y.; Li, W.; and Ji, H. 2015. A novel neural topic model and its supervised extension. In *Proceedings of AAAI.*

Chen, J.; He, J.; Shen, Y.; Xiao, L.; He, X.; Gao, J.; Song, X.; and Deng, L. 2015. End-to-end learning of lda by mirror-descent back propagation over a deep architecture. In *Proceedings of NIPS*, 1765–1773.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of EMNLP*, 1724–1734.

Chong, W.; Blei, D.; and Li, F. 2009. Simultaneous image classification and annotation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1903–1910.

Dieng, A. B.; Wang, C.; Gao, J.; and Paisley, J. 2017. Topicrnn: A recurrent neural network with long-range semantic dependency. In *Proceedings of ICLR.*

Ding, R.; Nallapati, R.; and Xiang, B. 2018. Coherence-aware neural topic modeling. In *Proceedings of EMNLP*, 830–836.

Gemp, I.; Nallapati, R.; Ding, R.; Nan, F.; and Xiang, B. 2019. Weakly semi-supervised neural topic models. In *Learning from Limited Labeled Data (LLD) Workshop, ICLR.*

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes.

Lacoste-Julien, S.; Sha, F.; and Jordan, M. I. 2009. Disclda: Discriminative learning for dimensionality reduction and classification. In *Proceedings of NIPS*, 897–904.

Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. In *Proceedings of NIPS*, 121–128.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of ICML*, 2410–2419.

Miao, Y.; Yu, L.; and Blunsom, P. 2016. Neural variational inference for text processing. In *Proceedings of ICML*, 1727–1736.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 26.* Curran Associates, Inc. 3111–3119.

Pang, B., and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL.*

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of EMNLP*, 248–256.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of ICML*, 1278–1286.

Shi, T., and Zhu, J. 2017. Online bayesian passive-aggressive learning. *The Journal of Machine Learning Research* 18(1):1084–1122.

Srivastava, A., and Sutton, C. 2016. Neural variational inference for topic models. In *Bayesian deep learning workshop, NIPS.*

Wang, Y., and Zhu, J. 2014. Spectral methods for supervised topic models. In *Proceedings of NIPS*, 1511–1519.

Zhu, J.; Ahmed, A.; and Xing, E. P. 2012. Medlda: maximum margin supervised topic models. *Journal of Machine Learning Research* 13:2237–2278.

Zubiaga, A. 2012. Enhancing navigation on wikipedia with social tags. *arXiv preprint arXiv:1202.5469.*

Supplementary Material for:
Neural Topic Model with Attention for Supervised Learning

1 Detailed model inference

Starting from Equation (8), we can perform the reparameterization trick as below:

$$
\begin{aligned}
&\log p_{\Theta,\Psi}(l,\boldsymbol{d}) = \log \int_{\boldsymbol{t}} \frac{p_\Theta(\boldsymbol{t})}{q_\Phi(\boldsymbol{t}|\boldsymbol{d})} q_\Phi(\boldsymbol{t}|\boldsymbol{d}) p_\Psi(l|\boldsymbol{t}) p_\Theta(\boldsymbol{d}|\boldsymbol{t}) d\boldsymbol{t} \\
&= \log \mathbb{E}_{q_\Phi(\boldsymbol{t}|\boldsymbol{d})} \big[\frac{p_\Theta(\boldsymbol{t})}{q_\Phi(\boldsymbol{t}|\boldsymbol{d})} p_\Psi(l|\boldsymbol{t}) p_\Theta(\boldsymbol{d}|\boldsymbol{t}) \big] \\
&= \mathbb{E}_{q_\Phi(\boldsymbol{t}|\boldsymbol{d})} \big[\log p_\Theta(\boldsymbol{d}|\boldsymbol{t}) - \log \frac{q_\Phi(\boldsymbol{t}|\boldsymbol{d})}{p_\Theta(\boldsymbol{t})} + \log p_\Psi(l|\boldsymbol{t}) \big] \\
&= \mathbb{E}_{q_\Phi(\boldsymbol{t}|\boldsymbol{d})} \big[\log p_\Theta(\boldsymbol{d},\boldsymbol{t}) - \log q_\Phi(\boldsymbol{t}|\boldsymbol{d}) + \log p_\Psi(l|\boldsymbol{t}) \big] \\
&\quad + \mathrm{KL}(q_\Phi(\boldsymbol{t}|\boldsymbol{d}) || p_\Theta(\boldsymbol{t}|\boldsymbol{d})) \qquad (1)
\end{aligned}
$$

Where Ψ represents all the parameter from the RNN attention model. Θ are the generative parameters and Φ are the variational parameters. μ_0, σ_0 are omitted because they are constants. Since $\beta \subset \Theta$, β is omitted too.

Since KL-divergence is always non-negative, we construct the variational objective function, also called the evidence lower bound (ELBO) of $\log p_{\Theta,\Psi}(l,\boldsymbol{d})$ as below:

$$
\begin{aligned}
\mathcal{L} =& \mathbb{E}_{q_\Phi(\boldsymbol{t}|\boldsymbol{d})} \big[\log p_\Theta(\boldsymbol{d},\boldsymbol{t}) - \log q_\Phi(\boldsymbol{t}|\boldsymbol{d}) + \log p_\Psi(l|\boldsymbol{t}) \big] \\
=& \mathbb{E}_{q_\Phi(\boldsymbol{t}|\boldsymbol{d})} \big[\log p_\Theta(\boldsymbol{d}|\boldsymbol{t}) + \log p_\Psi(l|\boldsymbol{t}) \big] \\
& - \mathrm{KL}(q_\Phi(\boldsymbol{t}|\boldsymbol{d}) || p(\boldsymbol{t})) \qquad (2)
\end{aligned}
$$