Learning from LDA Using Deep Neural Networks

Dongxu Zhang[1,3], Tianyi Luo[1] and Dong Wang*[1,2]

[1] CSLT, RIIT, Tsinghua University
[2] Tsinghua National Lab for Information Science and Technology
[3] PRIS, Beijing University of Posts and Telecommunications
Beijing, China
wangdong99@mails.tsinghua.edu.cn

Abstract. Bayesian models and neural models have demonstrated their respective advantage in topic modeling. Motivated by the dark knowledge transfer approach proposed by [3], we present a novel method that combines the advantages of the two model families. Particularly, we present a transfer learning method that uses LDA to supervise the training of a deep neural network (DNN), so that the DNN can approximate the LDA inference with less computation. Our experimental results show that by transfer learning, a simple DNN can approximate the topic distribution produced by LDA pretty well, and deliver competitive performance as LDA on document classification, with much faster computation.

1 Introduction

Probabilistic topic models, for instance Latent Dirichlet Allocation (LDA) [2], have been extensively studied and widely used in applications such as topic discovery, document classification and information retrieval. Most of the successful probabilistic topic models are based on Bayesian networks [7, 12], where the random variables and the dependence among them are carefully designed by people and so hold clear meanings in physics and/or statistics.

A particular problem of Bayesian topic models, however, is that when the model structure is complex, the inference for the latent topic distribution (topic mixture weights) is often untractable. Various approximation methods have been proposed, such as the variational approach and the sampling method, though the inference is still slow.

Recently, [3] proposed a transfer learning approach. In this approach, a complex model is used as a teacher model to supervise the training of a simpler model. The original proposal used a complex deep neural network (DNN) to train a simple shallow neural network and obtained performance very close to the complex DNN. This motivated our current research that attempts to use a Bayesian model to supervise the training of a neural model. By this approach, we hope to transfer the knowledge learned by the Bayesian model to the neural model, and combine the respective advantages of the two model families.

In this preliminary study, we use an LDA as the teacher model to guide the training of a DNN, so that the DNN can approximate the behavior and performance of LDA. The dropout technique [6] is utilized on the input layer, which endows the learned DNN with better generalizability. A big advantage of this transfer learning from LDA to DNN is that inference with DNN is much faster than with LDA. This solves a major difficulty of LDA on large-scale online tasks, meanwhile retaining the advantage of LDA in topic learning.

We tested the proposed method on document classification. The results show that a simple DNN model with dropout technique can approximate LDA pretty well and the inference speeds up tens or hundreds of times. Interestingly, a preliminary analysis shows that by the transfer learning, the DNN model seems can discover topics similar to those learned by LDA, although this information is not explicitly presented in the learning process.

2 Related Work

This work develops a neural model to approximate the function of LDA [2], with a direct goal of a fast inference. LDA is a generative probabilistic model and belongs to a large family of Bayesian topic models. LDA assumes that a document involves a mixture of latent topics, and each topic is characterized by a distribution over words. Compared with the early probabilistic models such as pLSI [7], LDA is a full generative model that can deal with new documents, but the inference is rather slow. The DNN-based LDA approximation presented in this paper attempts to solve this problem.

Our work is also closely related to the deep learning research which was largely initiated by [4]. DNN is a popular deep learning model and is capable of learning complex functions and inferring layer-wise patterns. This work leverages these advantages and uses DNNs to approximate LDA by learning its mapping function from the primary input (i.e., term frequency) to the high-level output (i.e., topic mixture). Interestingly, we find that topics can be discovered by DNN automatically with this transfer learning, which in turn demonstrates the power of DNN models. Note that deep learning has been employed in topic modeling, e.g., the approach based on deep Boltzmann machines (DBM) [5, 10]. The difference of our work is that we focus on approximating a well-trained Bayesian model with a deep neural model, instead of learning the deep model from scratch.

Finally, this research is directly motivated by the dark knowledge distiller model [3] that employs the knowledge learned by a complex DNN to guide the training of a simpler DNN, or vice versa [13]. In this work, we extend this method to learn a neural model with the supervision of a Bayesian model, which is more ambitious and challenging.

3 Methods

For a particular document d, LDA takes the term frequency (TF) as the input, denoted by $v(d)$. The inference task is then to derive the topic mixture $\theta(d)$, which is actually the posterior probability distribution that the document belongs to the topics. In tasks such as document clustering or classification, $\theta(d)$ is a good representation for document d, with a low dimensionality and a clear semantic interpretation.

Exact inference with LDA is untractable and so various approximation methods are usually used. This work chooses the variational inference method proposed by [2], which involves iterative update of the document and word topic mixtures and hence time-consuming. The basic idea of the LDA to DNN knowledge transfer learning is to train a DNN model which can simulate the behavior of LDA inference, but with much less computation. More precisely, the DNN model learns a mapping function $f(v(d);w)$ such that $f(v(d);w)$ approaches to $\theta(d)$, where w denotes the parameters of the DNN. Note that $\theta(d)$ is a probability distribution. To approximate such normalized variables, a softmax function is applied to the DNN output and the cross entropy is used as the training criterion, given by:

$$\mathcal{L}(w) = -\sum_{d}\sum_{i=1}^{K}\theta(d)_i \log f(v(d);w)_i \tag{1}$$

where K denotes the number of topics and the subscript i indexes the dimension. Once the DNN is trained, the mapping function $f(v(d);w)$ learns the behavior of the LDA model and can be used to predict $\theta(d)$ for new documents. Compared to the LDA inference, $f(v(d);w)$ can be computed very fast and hence amiable to large-scale online tasks.

Here in the training process, we employ dropout technique [6] on the input $v(d)$ by randomly set each dimension of $v(d)$ to zero with 0.5 probability. After that, we re-normalize the input by dividing $v(d)$ with $\sum_i v(d)_i$.

We experimented with two DNN structures: a 2-layer DNN (DNN-2L) that involves one hidden layer, and a 3-layer DNN (DNN-3L) that involves two hidden layers. In DNN-2L, the number of hidden units is twice of the output units; in DNN-3L, the number of hidden units are three and two times of the output units for the first and second hidden layer, respectively. The hyperbolic function is used as the activation function. The training employs the stochastic gradient descent (SGD) method, and is implemented based on Theano [1][4].

4 Experiments

4.1 Database and Experimental Setup

The proposed methods are tested on the document classification task with two datasets. The first dataset is Reuters-21578 and we follow the 'LEWISSPLIT' configure to define the training and test data. The documents are labelled in 55 classes.[5]

The second dataset is 20 Newsgroups collected by Ken Lang, which contains about 20,000 articles evenly distributed over 20 UseNet discussion groups. These groups correspond to the classes in document classification.[6]

It has been known that LDA performs better with long documents [11]. To establish a strong LDA baseline, only long documents are selected for training and test in this study. Considering that 20 Newsgroups is much larger than Reuters-21578, different selection criteria are used to choose documents for the two datasets, as shown in

[4] `http://deeplearning.net/software/theano/`

[5] `https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html`

[6] `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html`

Table 1. The table also shows the lexicon size in the LDA and DNN modeling, which corresponds to the dimensionality of the TF feature. Note that this seemingly tricky data selection is just for building a strong LDA model for the DNN to learn, rather than intensively selecting a working scenario for the proposed method. In fact, the DNN learning works well with any LDA teacher model, and the performance of the resultant DNN largely depends on the quality of the teacher LDA.

Table 1. Data profile of the experimental datasets.

	Reuters	20 News
Document length threshold	100	300
Training documents	3622	6312
Test documents	1705	1542
Word frequency threshold	30	200
Lexicon size (words)	2388	1910

4.2 Results

To evaluate the proposed transfer learning, we compare the classification performance with the document vectors inferred from the LDA-supervised DNN and the original LDA. The support vector machine (SVM) with a linear kernel is used as the classifier. Since LDA is the teacher model, its performance can be regarded as a upper bound of the DNN learning. Additionally, we choose the popular principle component analysis (PCA) [8] as another baseline and regard it as a low bound of the learning. All these three methods generate low-dimensional document vectors and are comparable in the sense of dimension reduction. Note that in many cases LDA does not outperform PCA, though it is not the focus of our study. What we are concerned with is that in the case where LDA is superior to PCA, the learned DNN can keep this superiority, but with much less computation cost.

Document classification The results in terms of classification accuracy on the two datastes are reported in Figure 1, where the number of topics varies from 10 to 70. We first observe that LDA obtains better performance than PCA on both the two datasets. Again, this is partly attributed to the long documents used in the study. The two DNN models obtain similar performance as LDA and outperform PCA, particulary with a small number of topics. This indicates that the DNNs indeed learned the behavior of LDA. If the number of topics is large, the DNN models work not as well, particularly on the Reuters task. This is probably because the limited amount of training data (just several thousands of training samples) can not afford learning complex models.

Note that the 3-layer DNN outperforms the 2-layer DNN. This indicates that deeper models can learn the LDA behavior more precisely. This is not surprising and has been widely demonstrated by the recent success of deep learning. This can be evaluated

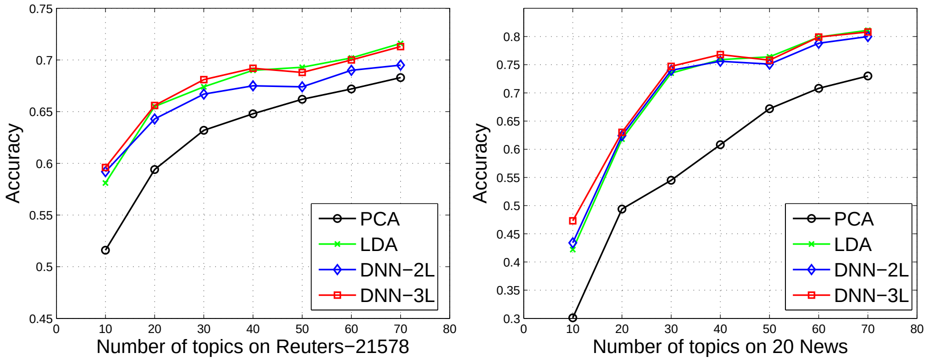

Fig. 1. The classification accuracy of PCA, LDA, 2-layer DNN (DNN-2L) and 3-layer DNN (DNN-3L).

more directly in terms of KL divergence between the LDA output $\theta(d)$ and the DNN prediction $f(v(d); w)$, as shown in Figure 2.

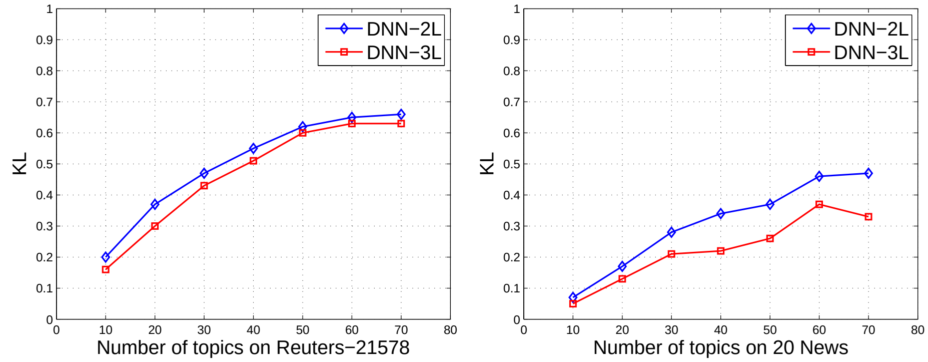

Fig. 2. The averaged KL divergency between the DNN and LDA output calculated on the test data of Reuters-21578 and 20 Newsgroups.

Inference speed The comparative results on inference time are shown in Figure 3. The experiments were conducted on a desktop with 4 3.4G Hz cores, and to alleviate randomness the experiments were conducted 10 times and the averaged numbers are reported. It can be seen that the DNN model is much faster (10 to 200 times) than the original LDA, and the superiority is more clear with a large number of topics. Compar-

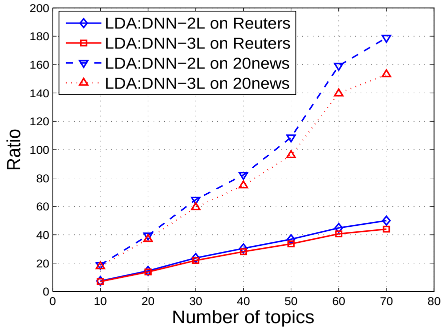

Fig. 3. The ratio of inference time of LDA to DNN.

ing the results on the two datasets, we observe that DNN exhibits more advantages on 20 Newsgroups, because the long documents of this dataset are more difficult to infer with LDA. Additionally, the 3-layer DNN is not much slower than the 2-layer DNN, which means that using deeper models does not cause much additional computation.

We emphasize that the results here should not be over-interpreted. What we compared here is just the basic LDA implementation from [2]. There are quite some faster implementations that we did not compare with, e.g., FastLDA [9]. We expect that the margin between DNN and FastLDA is not as significant as reported here. Nevertheless, DNN inference involves only simple matrix manipulations and so are naturally amiable to large-scale computation, e.g., by optimized numerical math libraries (BLAS, MKL, etc.) or GPUs. We therefore argue that speed is a intrinsic advantage of DNN models, particularly when compared to more complex Bayesian models for which fast algorithms (like FastLDA) are not available.

5 Topic Discovery by Transfer Learning

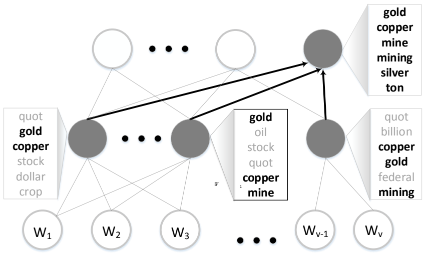

Fig. 4. Discovery for the topic ‘mining’ with DNN. The words in dark are topic related words.

A known advantage of DNNs is that high-level representations can be learned automatically layer by layer. This property may help DNN to discover topics from the raw TF input. To verify this conjecture, a one-hot vector is given to a DNN that has been trained by LDA supervision, and the activation on each hidden neuron is recorded. The one-hot vector represents a particular word, and the activation reflects how a particular neuron is related to this word. For each neuron, we record the activations of all the words and select the top-10 words that give the most significant activations, which forms the set of representative words for the neuron.

Interestingly, we find that for each neuron, the representative words are generally correlated, forming a local topic. Figure 4 shows an example, where the topic 'mining' at the second hidden layer is formed by aggregating the related topics at the first hidden layer. This example shows clearly how words are clustered layer by layer to form semantic meaningful topics. Interestingly, we find that the topics derived from DNN and LDA are quite similar. As an example, the top-10 words for the topic 'mining' derived from LDA are {**gold, said, mine, copper, ounces, mining, tons, ton, silver, reuter**}, while the DNN-derived top-10 words are {**gold, copper, mine, mining, silver, zinc, minerals, metal, mines, ton**}.

This can be explained by the fact that the mixture weights generated by LDA and used as the DNN supervision are based on the same latent topics. We emphasize that the topic information is not transferred to DNN in the model training; it is the learning power of DNN that discovers the topics by itself.

6 Conclusion and Future Work

We proposed a knowledge transfer learning method that uses deep neural networks to approximate LDA. Results on document classification tasks show that a simple DNN can approximate LDA quite well, while the inference is significantly speeded up. This preliminary research indicates that transferring knowledge from Bayesian models to neural models is possible. The future work involves studying knowledge transfer between more complex probabilistic models and other neural models. Particularly, we are interested in how to use the knowledge of probabilistic models to regularize neural models so that the neurons are more interpretable.

Acknowledgments

The authors give great thanks to Dr. Shujie Liu (MSRA) for fruitful discussions. This research was supported by the National Science Foundation of China (NSFC) under the project No. 61371136, and the MESTDC PhD Foundation Project No. 20130002 120011. It was also supported by Huilan Ltd.

References

1. Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I.J., Bergeron, A., Bouchard, N., Bengio, Y.: Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop (2012)

2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022 (2003)
3. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
4. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural computation 18(7), 1527–1554 (2006)
5. Hinton, G.E., Salakhutdinov, R.R.: Replicated softmax: an undirected topic model. In: Advances in neural information processing systems. pp. 1607–1614 (2009)
6. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving neural networks by preventing co-adaptation of feature detectors. CoRR abs/1207.0580 (2012), http://arxiv.org/abs/1207.0580
7. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. pp. 289–296. Morgan Kaufmann Publishers Inc. (1999)
8. Jolliffe, I.: Principal component analysis. Wiley Online Library (2002)
9. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M.: Fast collapsed gibbs sampling for latent dirichlet allocation. Knowledge Discovery and Data Mining (2008)
10. Srivastava, N., Salakhutdinov, R.R., Hinton, G.E.: Modeling documents with deep boltzmann machines. arXiv preprint arXiv:1309.6865 (2013)
11. Tang, J., Meng, Z., Nguyen, X., Mei, Q., Zhang, M.: Understanding the limiting factors of topic modeling via posterior contraction analysis. In: Proceedings of The 31st International Conference on Machine Learning. pp. 190–198 (2014)
12. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the american statistical association 101(476) (2006)
13. Wang, D., Liu, C., Tang, Z., Zhang, Z., Zhao, M.: Recurrent neural network training with dark knowledge transfer. arXiv preprint arXiv:1505.04630 (2015)