Learning from LDA using Deep Neural Networks

Introduction

Latent Dirichlet Allocation (LDA), a probabilistic topic model, is 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 , where the random variables and the dependence among them are carefully designed by people and so hold clear meanings in physics and/or statistics. For this reason, Bayesian topic models can represent the document generation process well and have attained much success in semantic analysis and related research.

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 intractable. Various approximation methods have been proposed, such as the variational approach and the sampling method, though the inference is still very slow.

Thus, 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. 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.

Method

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 $\boldsymbol{\theta}$(d), which is actually the posterior probability distribution that the document belongs to the topics. In tasks such as document clustering or classification, $\boldsymbol{\theta}$(d) is a good representation for document d, with a low dimensionality and a clear semantic interpretation. 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 $\boldsymbol{\theta}$(d), where w denotes the parameters of the DNN. Note that $\boldsymbol{\theta}$(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.

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

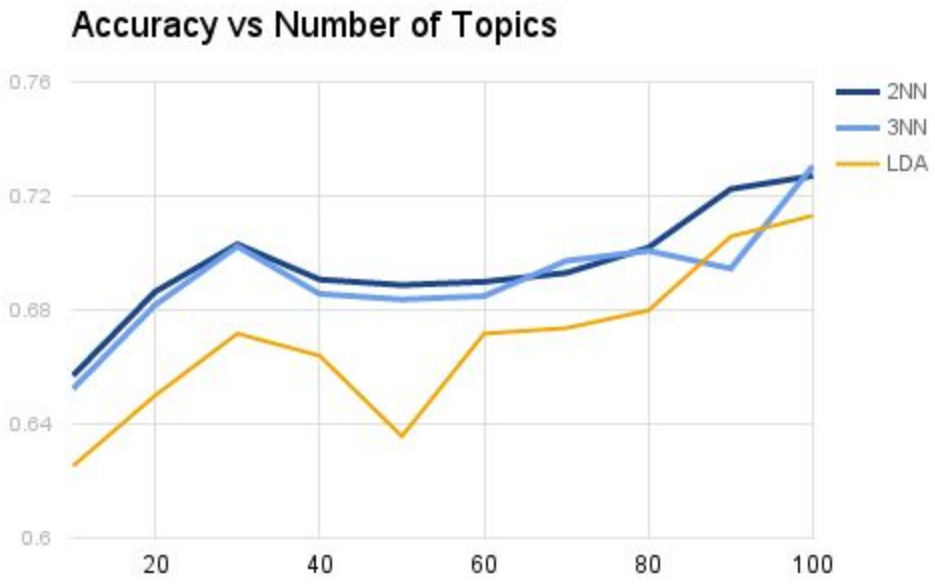
and second hidden layer, respectively. The hyperbolic function is used as the activation function. The training employs the stochastic gradient descent (SGD) method.

Experiment

We took the dataset Reuters-21578, and from which we filtered the documents with multiple ground truth labels. We used the training and test data split as provided by the nltk toolkit. We were left with 6577 training documents and 2583 test documents. We removed stop words and applied a wordnet stemmer from the nltk toolkit.

The LDA was trained with the documents over 100 passes and the parent model generated was stored for varying number of topics. Then a child 2nn and 3nn model was trained with this as the parent model.

Results



Conclusion

We proposed a knowledge transfer learning method that uses deep neural networks to approximate LDA. Results on a document classification task show that a simple DNN can approximate LDA quite well, while the inference is tens or hundreds of times faster. This preliminary research indicates that transferring knowledge from Bayesian models to neural models is possible.