NEW YORK UNIVERSITY

DOCUMENT LEARNING AND CLASSIFICATION WITH VAE

{ YUTING GUI & YUHAO ZHAO }@NYU, CENTER FOR DATA SCIENCE

INTRODUCTION

Auto-encoder framework have been widely used in learning generative model for text and images. The encoder is trained to encode the input x to its hidden representation z, and at the same time, the decoder is trained to reconstruct input x from z so that the target output of decoder has relatively small distance to the original input.

The Neural Variational Document Model (NVDM) was proposed to efficiently train the continuous semantic latent variables for documents from a bag-of-word input. In this project, we extend the NVDM model and introduces a variational auto-encoder method that jointly learns the document representation as well as the projection from the latent space to the output space that can be used to produce classifications.

We proposed a multi-task document learning model (NVDM + MLP) that combines the tasks of document classification and document representation learning which outperforms several strong baseline in 2 standard text corpora.

DATA DESCRIPTION

1. **20News**

   The 20News data contains 11,314 training and 7,531 testing news and was labelled into 20 disjoint classes. The vocabulary size is 2,000.

2. **IMDB**

   The IMDB is a dataset for binary sentiment classification. It contains 25,000 labelled movie reviews for training and 25,000 unlabelled reviews for testing. The vocabulary size is 10,000.

3. **RCV1-v2**

   The RCV1-v2 data is a larger collection of manually categorized newswire stories, which was originally split into 781,265 testing and 23,149 training cases. The news was categorized into 103 hierarchical topics which span five orders of magnitude, thus can be used as a multi-label classification task. We used 10,000 vocabulary.

NVDM

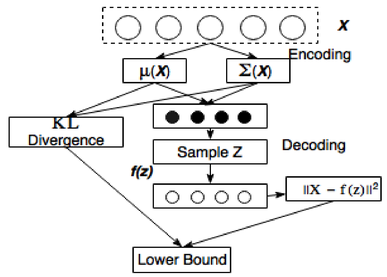

Figure 1: NVDM

The Neural Variational Document Model (NVDM) (see Figure 1) is an unsupervised text generative model which learns the continuous semantic latent variable for each document. The variational auto-endocer takes the document bag-of-words representation $\boldsymbol{X} \in R^{|V|}$ (V is the vocabulary size), and pass to an MLP inference network where $q_\phi(\boldsymbol{h}|\boldsymbol{X})$ aims to learn the continuous hidden representation $\boldsymbol{h} \in \mathbb{R}^K$ of a document (K is the hidden vector size). The softmax decoder (generative model) $p_\theta(\boldsymbol{X}|\boldsymbol{h}) = \prod_{i=1}^{N} p(\boldsymbol{x_i}|\boldsymbol{h})$ aims to reconstruct the document by independently generating the words ($\boldsymbol{h} \rightarrow \{\boldsymbol{x_i}\}$), where $\boldsymbol{x_i}$ represents each word in the document with N words in total. In order to maximize the log-likelihood $\log \sum_h p(\boldsymbol{X}|\boldsymbol{h})p(\boldsymbol{h})$ of documents, we optimised the variational lowerbound.

EXPERIMENT SETUP

For the 20News and IMDB dataset, we first trained the tf-idf with SVM classifier as the baseline model. And then we run NVDM with pure unsupervised document learning setup with 50 and 200 hidden dimensions. On all datasets, we simply extracted the document representation and train an SVM classifier on top. We tune the L2 penalty with 5-fold cross validation. Finally, we trained our model (NVDM + MLP) with 50 hidden dimensions and ReLU as activation function. For the RCV1 dataset, we only trained an MLP-classifier with the threshold learning.

OUR MODEL

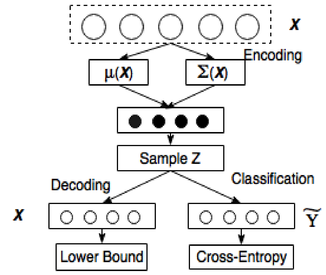

Figure 2: Our Model

Our model (NVDM + MLP) combines the tasks of document classification and document representation learning (Figure 2). This scheme involves one variational encoder and multiple decoders. The first decoder is used to reconstruct words in the document independently, which structure is inherited from NVDM. Another decoder is a MLP used to classify the document categories. Both components are decoding conditioned on the hidden document representation. For multi-label prediction, we also trained the threshold using a single projecting layer from the document hidden representation. In training, after generating the output distribution from softmax, we chose the threshold that maximize the F1 score at every pair of positive labels. The threshold predictor was trained to minimize the regression loss with L2 regularization: $J(\theta) = \frac{1}{M}\sum_{m=1}^{M}(Wh_m + b - t_m)^2 + \frac{\lambda}{2}||W||_2^2$.

RESULTS

Dataset	20News	IMDB	RCV1
baseline	0.388	0.724	0.359
MLP	0.437	0.783	0.544
NVDM+SVM	0.482	0.825	-
NVDM+MLP	**0.531**	**0.839**	-

Figure 3: Document Classification results

Model	Dim	20News	IMDB	RCV1
LDA	50	1091	-	1437
LDA	200	1058	-	1142
NVDM	50	907	880	-
NVDM+MLP	50	**892**	**734**	-

Figure 4: Perplexity

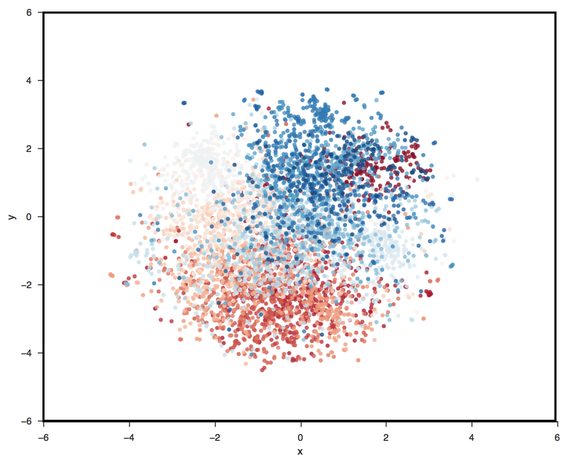

Figure 4: t-SNE Visualization of 20News Representation

CONCLUSION

This project extended the neural variational inference framework to a multitask document learning model. The experiments showed that our proposed model outperforms several strong baselines in the task of document learning and categorization on two standard text corpora. The results demonstrate that the continuous latent variable has better generalization ability.

ACKNOWLEDGEMENTS

We would like to thank our advisor Joan Bruna for offering us inspiring suggestions. And we would also like to thank Professor David Sontag for insightful lectures. For the list of citations, please refer to our final report.