Semi-Supervised Bayesian Active Learning for Text Classification

Sophie Burkhardt, Julia Siekiera, Stefan Kramer
Department of Computer Science
Johannes Gutenberg University
55128 Mainz, Germany
{burkhardt,kramer}@informatik.uni-mainz.de, jsiekier@students.uni-mainz.de

Abstract

In many application domains, unlabeled data is abundant, but labeled data expensive and difficult to obtain. There are two remedies for this problem. First, semi-supervised learning may be used to obtain efficient classifiers with less labeled data and second, active learning decreases the amount of labeled data needed for good classification performance. We combine both of these techniques using a deep Bayesian model, the semi-supervised variational autoencoder. Using this model, the distribution of the data is learned from large amounts of unlabeled data and uncertainty about class labels is explicitly represented. The effectiveness of our model in predicting side effects of drugs is shown on Twitter data, where we only need two thirds of the labeled training examples as compared to the non-active baseline.

1 Introduction

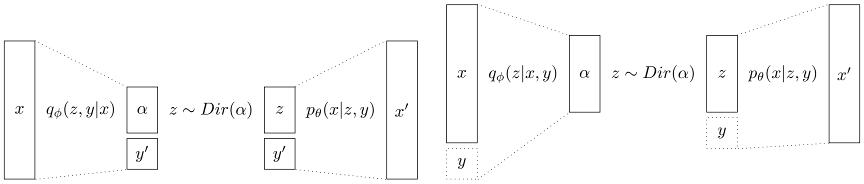
In many application contexts it is difficult to obtain labeled data while unlabeled data are easy to come by. For example, people frequently post about their health on social media, however, it is difficult to extract the posts that contain self reports about drug side effects. Adverse reactions to drugs are one of the leading causes of hospitalization today, so labeling this data could have large impact.

To reduce the cost of labeling, at least two different approaches can be taken. First, semi-supervised training takes advantage of unlabeled data to learn the data distribution and use this implicit information to improve classification. Second, in active learning (AL) the algorithm can choose which documents will be labeled. In this way the number of labeled examples can be reduced if the chosen examples are informative. Often, uncertainty is used as a measure for how informative a training example is, but representativeness of the whole dataset may also be an important factor.

Existing work on deep Bayesian AL is based on Bayesian CNNs [2], a dropout-based approach, or on the Bayes-by-Backprop (BBB) algorithm [1]. These methods place a prior on the weights of the neural network. Gal *et al.* [2] used Bayesian CNNs for AL on image data. Siddhant and Lipton [11] compared Bayesian CNNs and BBB for text classification, named entity recognition and semantic role labeling. Both of these approaches are purely supervised and cannot take advantage of unlabeled data.

In contrast to BBB, variational autoencoders (VAEs) [6, 5] place a prior on the latent variables directly, enabling semi-supervised training to discover latent factors. Recent work on text data explored the use of different priors and network architectures. The neural variational document model (NVDM) [8] is a VAE with a Gaussian prior, whereas ProdLDA [12] uses a Laplace approximation to a Dirichlet prior. In this work we build on recent work on a different reparameterization using rejection samplers (RSVI) [9] to train our network with a Dirichlet prior on the latent variables. We combine the RSVI method with a semi-supervised framework and evaluate different uncertainty measures.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.



(a) VAE with unknown labels

(b) VAE with known labels

Figure 1: Illustration of the semi-supervised autoencoder with Dirichlet prior.

While Gal *et al.* [2] have compared AL to semi-supervised methods and found similar performance, so far there is no work that combines both, deep Bayesian AL and semi-supervised learning.

To sum up, our contributions are as follows:

- To the best of our knowledge, this is the first work to apply VAEs in AL for text classification and
- to combine deep Bayesian AL with semi-supervised learning.

2 Method

2.1 Semi-supervised Variational Autoencoders

The semi-supervised VAE [5] optimizes a different objective dependent on whether the label y is observed or not. The model is illustrated in Figure 1. If the label is observed, the objective is:

$$\log p_\theta(x,y) \geq \mathbb{E}_{p_\phi(z|x,y)}[\log p_\theta(x|y,z) + \log p_\theta(y) + \log p_\theta(z) - \log q_\phi(z|x,y)] \tag{1}$$

In the other case with an unobserved label the objective is:

$$\log p_\theta(x) \geq \mathbb{E}_{p_\phi(y,z|x)}[\log p_\theta(x|y,z) + \log p_\theta(y) + \log p_\theta(z) - \log q_\phi(y,z|x)] \tag{2}$$

In our AL method, we train with all remaining unlabeled data. However, in the beginning of training this gives a lot of weight to the unlabeled data. Therefore we train with the labeled data for more iterations in order to achieve a balance between unlabeled and labeled training data. If L is the number of labeled data and U is the number of unlabeled data we train with the labeled data for $\lceil \frac{U}{L} \rceil$ iterations.

2.2 Uncertainty

We compare two different uncertainty measures for our method. The first is entropy: $H(p(y|x)) = -\mathbb{E}_{p(y|x)}[\log p(y|x)]$ and the second is the Bayesian active learning by disagreement (BALD) uncertainty measure [4]:

$$\mathbb{I}[y,\omega|x,\mathcal{D}_{train}] \approx -\sum_c \left(\frac{1}{T}\sum_t \hat{p}_c^t\right) \log\left(\frac{1}{T}\sum_t \hat{p}_c^t\right) + \frac{1}{T}\sum_{c,t} \hat{p}_c^t \log \hat{p}_c^t, \tag{3}$$

where $\hat{p}_c^t$ is the estimated probability for class c in dropout iteration t and T is the total number of dropout iterations.

2.3 Representativeness

Uncertainty strategies are known to request outliers that are not representative of the data as a whole. This is why we also take the representativeness into account by requesting labels for instances x_i that maximize $Z(x_i) * u_i$, where u_i is the uncertainty and the density $Z(x_i)$ is given as follows [7]:

$$Z(x_i) = \exp\left(\frac{1}{|D|}\sum_{x_h} -\beta(D_{KL}(p(W|x_h)||\lambda p(W|x_i) + (1-\lambda)p(W)))\right), \tag{4}$$

where W is a random variable over the vocabulary, D_{KL} is the Kullback-Leibler divergence, λ is a smoothing parameter and β determines the sharpness of the distance metric.

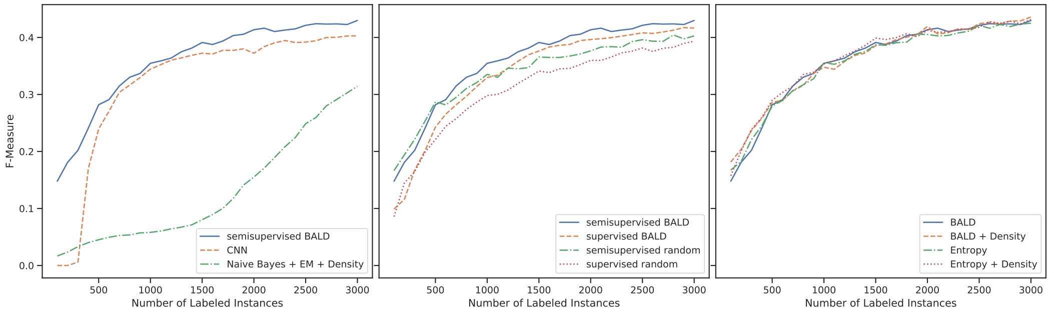


Figure 2: Plotted is the F-measure against the number of labeled documents averaged over 10 runs of 5-fold cross validation; left: different AL methods are compared; middle: the supervised and semi-supervised variant are compared; right: different uncertainty strategies are tested

3 Experimental Setting

We compare our method to two different baselines. First, we compare to the Naive Bayes with EM method [7], which is similar to our method in that it also uses semi-supervised learning. Second, we compare to the deep Bayesian AL method by Siddhant and Lipton [11] as a recent state-of-the-art method based on CNNs. For the first method we use our own reimplementation, whereas for the second method we use the code provided by the authors[1].

To test our method we use a Twitter dataset [10, 3] with 6455 tweets that are classified according to whether or not they mention a side effect of a drug. Approximately 10% of the data are labeled as positive. The vocabulary size is 2651 after pruning stop words and stemming. For the CNN method we did not apply stemming to make better use of the word vectors.

The parameters for our method are commonly used default parameters and set as follows: 50 topics, a learning rate of 0.001, a batch size of 50, 50 hidden neurons, 3 samples and we train with early stopping using 10% of the current labeled training set for validation. The architecture is the same as in ProdLDA [12]. We report results averaged over 10 runs of 5-fold cross validation. All methods are trained on an initial batch of 100 random documents and subsequently add 100 new documents in each AL acquisition step.

4 Experimental Results

In the left plot of Figure 2 we compare different AL methods. The Naive Bayes method with EM clearly fails on this Twitter dataset. However, our method is even better than the CNN method by Siddhant and Lipton. This is almost surprising given that we do not take into account word order as the CNNs or LSTMs in the work of Siddhant and Lipton do.

The middle plot of Figure 2 compares the supervised and semi-supervised variant of our method. We can see that the AL variant is better than the random variant and the semi-supervised AL variant has an even steeper performance increase in the beginning of training. Thus, we show that both components of our method, the AL component and the semi-supervised component, work well together and improve over the baseline with random selection.

In the right plot of Figure 2 we test different uncertainty strategies. From these results we cannot conclude that one strategy is preferable to another. Entropy and BALD perform more or less on par. Also, the density strategy [7] does not seem to have a positive effect.

[1] We adapted the implementation (`https://github.com/asiddhant/Active-NLP`) to the same training and testing setting that is used for our method. In particular we modified it to not use the test set for selecting instances.

5 Conclusion

We propose a deep Bayesian AL method based on VAEs. By combining this AL method with a semi-supervised approach we achieve a better F-measure than with both approaches on their own as well as other existing approaches on a complex and noisy Twitter data set. As future work we would like to incorporate CNNs into our VAE and use character-level CNNs to be able to better deal with this type of data and many out-of-vocabulary words.

References

[1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1613–1622. JMLR.org, 2015.

[2] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[3] Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, and Apurv Patki. Mining twitter for adverse drug reaction mentions: A corpus and classification benchmark. In *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM2014)*, 2014.

[4] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[5] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

[6] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[7] Andrew McCallum and Kamal Nigam. Employing EM and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pages 350–358, 1998.

[8] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA, 20–22 Jun 2016. PMLR.

[9] Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 489–498, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[10] A Sarker and G Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

[11] Aditya Siddhant and Zachary C Lipton. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. *arXiv preprint arXiv:1808.05697*, 2018.

[12] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.