Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study

Anonymous EMNLP submission

Abstract

Several recent papers investigate Active Learning (AL) for mitigating the data-dependence of deep learning for natural language processing. However, the applicability of AL to real-world problems remains an open question. While in supervised learning, practitioners can try many different methods, evaluating each against a validation set before selecting a model, AL affords no such luxury. Over the course of one AL run, an agent annotates its dataset exhausting its labeling budget. Thus, given a new task, we have no opportunity to compare models and acquisition functions. This paper provides a large-scale empirical study of deep active learning, addressing multiple tasks and, for each, multiple datasets, multiple models, and a full suite of acquisition functions. We find that across all settings, *Bayesian active learning by disagreement*, using uncertainty estimates provided either by Dropout or Bayes-by-Backprop significantly improves over i.i.d. baselines and usually outperforms classic uncertainty sampling.

1 Introduction

While over the past several years, deep learning has pushed the state of the art on numerous tasks, its extreme data-dependence presents a formidable obstacle under restricted annotation budgets. Active Learning (AL) presents one promising approach to reduce deep learning's data requirements (Cohn et al., 1996). Strategically selecting points to annotate over alternating rounds of labeling and learning, an active learner is hoped to outperform budget-matched i.i.d. labeling. Typical *acquisition functions* select examples for which the current predictor is most uncertain. However, how precisely to quantify uncertainty, especially for neural networks, remains an open question.

Classical approaches interpret either the entropy or the argmax of the predictive (e.g. softmax) distribution as the models uncertainty, yielding the *maximum entropy* and *least confidence* heuristics, respectively. These approaches account for aleatoric but not epistemic uncertainty (Kendall and Gal, 2017). Several recent Bayesian formulations of deep learning provide alternative techniques for extracting uncertainty estimates from deep networks, including a *dropout*-based approach (Gal and Ghahramani, 2016a), previously employed in Deep Active Learning (DAL) for image classification (Gal et al., 2017) and named entity recognition (NER) (Shen et al., 2018), and Bayes-by-Backprop (Blundell et al., 2015). To our knowledge, we are the first paper to apply Bayes-by-Backprop for DAL.

While several papers present show hints of DAL's potential, whether its suitability in practice remains an open question. That's because papers often address just a single task, just a single model, and sometimes just a couple datasets. However, it's not enough to look back retrospectively and declare that acquisition function outperforms an i.i.d. baseline. To apply DAL in practice, we must be confident that the technique will work correctly, **the first time**, on a dataset that we have never seen before. Otherwise, we might exhaust the annotation budget while performing worse than i.i.d. baseline. Absent resources for further labeling, there's no going back. Moreover, many DAL papers suffer from implicit target leaks. The architectures and hyper-parameters are often tuned using the full dataset, before concealing the labels to simulate AL.

In this paper, we present a large-scale study, comparing various acquisition functions across multiple tasks: Sentiment Classification (SC), NER, and Semantic Role Labeling (SRL), with multiple datasets, multiple models, and multiple acquisition functions. Moreover, in all experiments, we set hyper-parameters on warm start

data, allowing for as honest an assessment as possible. We seek not to champion any one approach but to ask *is there any single method that we could reliably expect to work out of the box on a new problem?*

To our surprise, we found that BALD(Houlsby et al., 2011), which measures uncertainty by the frequency of disagreement with the plurality, over the correct label, determined over multiple Monte Carlo draws from a stochastic model, proved effective across all combinations of task, dataset, and model. Moreover both variants of the approach, drawing samples according to the dropout method (Gal et al., 2017) and from a Bayes-by-Backprop network (Blundell et al., 2015), performed similarly well across most tasks, datasets, and models.

Related Work Only a few papers have addressed DAL for NLP, notably Shen et al. (2018) for NER and Zhang et al. (2017) who for text classification proposed selecting examples according to the expected magnitude of updates to word embeddings. This paper doesn't evaluate the latter technique because it's difficult to apply for sequence labeling tasks, where marginalizing over all possible labels blows up exponentially with sequence length. In both cases, however, experiments address multiple datasets (2 and 3, respectively) but just one task and just one model. Gal et al. (2017) propose the MC dropout (Gal and Ghahramani, 2015) variant of BALD for image classification with convolutional neural networks. They obtain significant improvement over classic uncertainty-based acquisition functions on the MNIST dataset and for diagnosing skin cancer from lesion images (ISIC2016 task). Our work builds on theirs, both by offering a large-scale evaluation of BALD for NLP tasks and models, and by exploring BALD with another source of uncertainty: uncertainty over the weight values, as modeled by a Bayes-by-Backprop network.

2 Bayesian Deep Learning

Monte Carlo Dropout According to (Gal and Ghahramani, 2016a), the dropout regularization techniques for neural networks can be interpreted as a Bayesian approximation to Gaussian processes (Rasmussen, 2004). Here, unlike standard uses of dropout, we apply it at prediction time. Uncertainty estimates are produced by comparing the output of a trained neural network using T different stochastic passes through the neural network. The extension to CNNs is straightforward. We apply dropout to RNNs following the approach due to (Gal and Ghahramani, 2016b), who extended their variational analysis to RNNs, arguing that dropout can be applied to the recurrent layers (and not just the synchronous connections, per standard practice) by applying identical dropout masks at each sequence step.

Bayes by Backprop In this approach, we represent weights in a network by probability distributions over possible values vs taking a point estimate (Blundell et al., 2015). A standard L-layer MLP model $P(y|x, w)$ is parametrized by weights $w = \{W_l, b_l\}_{l=1}^{L} \in \mathbb{R}^d$. Then, $\hat{y} = \phi(W_L \cdot ... \cdot + \phi(W_1 \cdot x + b_1) + .. + b_L)$ where ϕ is an activation function such as tanh or ReLU. Bayes-by-Backprop represents imposes a prior over the weights, $p(w)$ and seeks to learn the posterior distribution $p(w|D)$ given training data $D = \{x^i, y^i\}_{i=1}^{N}$. To deal with intractability, Bayes-by-Backprop approximates $p(w|D)$ by a variational distribution $q(w|\theta)$, typically choosing q to be a Gaussian with diagonal covariance and each weight sampled from $\mathcal{N}(\mu_i, \sigma_i^2)$. To enforce non-negativity, the σ_i are parametrized via the *softplus* function $\sigma_i = log(1 + exp(\rho_i))$ giving variational parameters $\theta = \{\mu_i, \rho_i\}_{i=1}^{d}$, optimized to minimize the KL divergence between the q and $p(w|D)$. Some simplification of the objective gives $\mathcal{L}(D, \theta) = \sum_{j=1}^{N} \left[\log q(w^j|\theta) - \log p(w^j) - \log p(D|w^j)\right]$, where w^j denotes the j-th Monte Carlo sample drawn from $q(w|\theta)$ (we use $N = 1$). Parameters are optimized by stochastic gradient descent, using the re-parameterization trick popularized by Kingma and Welling (2013). Extending Bayes-by-Backprop to CNNs and RNNs is straightforward with the latter requiring minor adjustments for truncated back-propagation through time (Fortunato et al., 2017). Uncertainty estimates calculated via Bayes-by-Backprop have been shown to be useful for efficient exploration in reinforcement learning (Lipton et al., 2016).

3 Experimental Setup

3.1 Acquisition functions

In this work, we consider only uncertainty based acquisition. In particular, we consider least confidence (LC) for classification and maximum

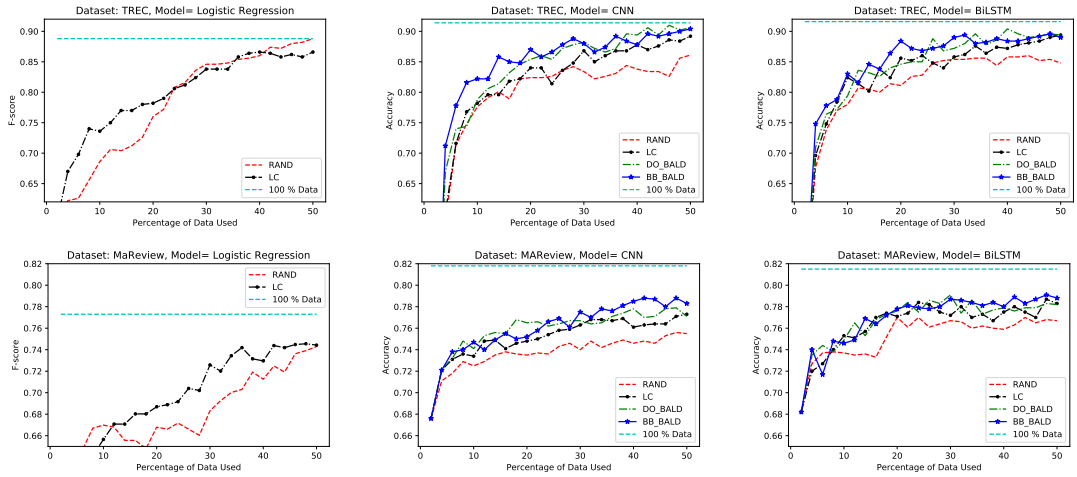

Figure 1: Performance of various models and acquisition functions for two SC datasets

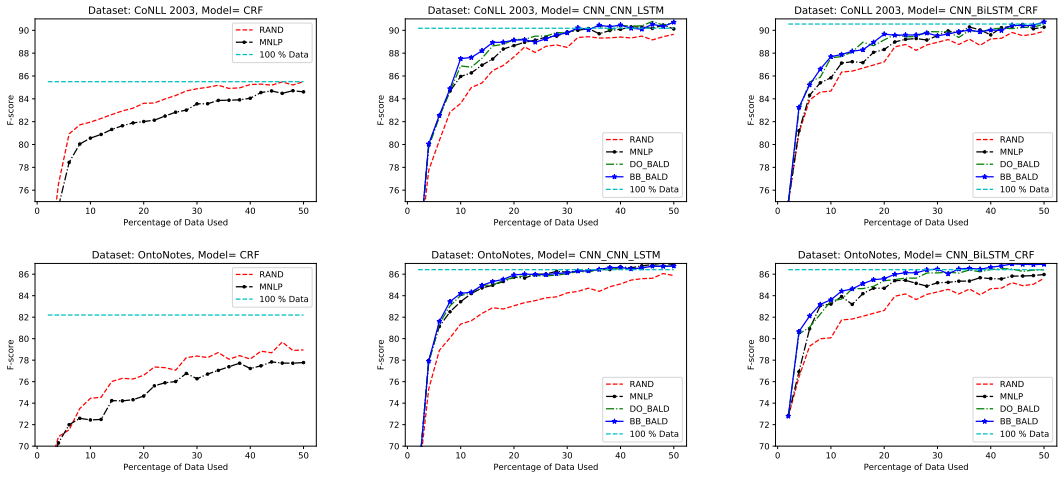

Figure 2: Performance of various models and acquisition functions for two NER datasets

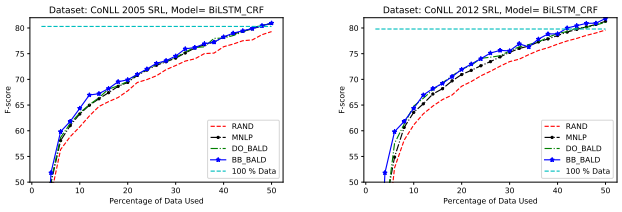

Figure 3: Performance of different acquisition functions on SRL task for two datasets

length-normalized log probability (MNLP) for sequence labeling tasks (Shen et al., 2018). LC chooses that example with for which the prediction has lowest predicted probability. MNLP extends this to sequences, selecting by log probability normalized by length, removing the bias for the model to preferentially select longer sequences.

BALD We briefly articulate the details of the Bayesian Active Learning by Disagreement (BALD) approach due to Houlsby et al. (2011), upon which both our Bayesian approaches ar based. We denote Monte Carlo Dropout Disagreement by DO-BALD and its Bayes-by-Backprop counterpart as BB-BALD. BALD originally selects samples that maximise the information gained about the model parameters. This boils down to choosing data points which each stochastic forward pass through the model would have the highest probability assigned to a different class (Gal et al., 2017). Our measure of uncertainty is the fraction of models, across MC samples from the network, that that disagree with most popular

choice. The ties are resolved by the mean of output probability of T models. For sequences, we look at agreement on the entire sequence tag, noting that this may exhibit a bias to preferentially sample longer sentences. Because we measure the budget at each round in words (not sentences) while this constitutes a bias, it does not constitute anan unfair advantage. Moreover, we note that all AL necessarily consists of biased sampling.

3.2 Training details

The active learning process begins with a random acquisition of 2% of *warmstart* samples from the dataset. We train an initial model on this data.Then based on this model's uncertainty estimates, we sample a batch of 2% more samples based on the chosen acquisition function and the new model is trained on this data 4 % data (re-initializing weights to avoid badly overfitting data collected earlier per observations by Hu et al. (2018)). We continue with alternating rounds of labeling and training until annotating 50% of the dataset. For classification tasks, the budget is measured in sentences while for sequence labeling, the budget is decided on number of words because the annotator must provide one tag per word. In each iteration, we train each model to convergence, decided based on early stopping with a patience of 1 epoch, or 25 epochs (whichever is earlier). For datasets with fixed validation sets such as Conll 2003, instead of using the entire validation set for early stopping, we use the percentage of validation data equivalent to that in our current training pool. Our reported results are averaged over 3 runs with different starting data.

3.3 Sentence Classification

We use two datasets for simulation: one question classification dataset TrecQA (Roth et al., 2002) and one sentiment analysis dataset (Pang and Lee, 2005) and two architectures for training: CNNs and BiLSTMs. For implementation of CNN on both these datasets, we follow the setup of Kim (2014) and for BiLSTMs, we use a single layer model with 300 hidden units for both datasets. 300 dimension glove embeddings (Pennington et al., 2014) pretrained on 6B tokens are used for all 4 settings. Dropout rate is 0.5 and the optimizer used is Adam (Kingma and Ba, 2014) with initial learning rate 1e-3. We use a batch size set to be either 50 or the number required for at least 10 updates whichever is lower. This is done to ensure that when the training pool is small, batch size is not too large and models get sufficient number of updates in an epoch. We also train a Unigram + Bigram + Logisitic Regression model with LC acquisition as a shallow AL baseline.

3.4 Named Entity Recognition

Again, we use two datasets: CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) and OntoNotes 5.0. The two architectures used for training are CNN-BiLSTM-CRF (CNN for character encoding, BiLSTM for word and CRF for decoding) (Ma and Hovy, 2016) and CNN-CNN-LSTM (CNN for character encoding, CNN for word encoding and LSTM for decoding) (Shen et al., 2018). We follow the exact experimental settings of these papers. Batch size is 16 for CoNLL and 80 for OntoNotes (minimum of 10 updates heuristic is followed here too). As a shallow AL baseline, we have a linear chain CRF model with MNLP acquisition function.

3.5 Semantic Role Labeling

We consider two datasets: CoNLL 2005 (Carreras and Màrquez, 2005) and CoNLL 2012, focusing on an LSTM-based model. Our model resembles He et al. (2017), but instead of using contained A* decoding, we use a CRF decoder, noting that while this causes a 2% drop in performance (at 100% annotation), our goal is to compare acquisition functions, not achieve record-setting performance. We follow the experimental setup of the paper but use a higher dropout rate of 0.25 adjust the batch size according to the minimum update rule.

4 Conclusion

This paper set out to investigate the practical utility of DAL for NLP. Our study consisted of over 40 experiments, each repeated for 3 times to average results and consisting of roughly 25 rounds of retraining, adding up to 3000 training runs to completion. Our goal was not to champion any one approach, but to see if there is any consistent story at all: *can active learning be applied on a new dataset with an arbitrarily architecture, without peeking at the labels to perform hyperparameter tuning?* To our surprise, we found that across many tasks, both classic uncertainty sampling and Bayesian approaches outperform i.i.d. baselines and that DO-BALD and BB-BALD consistently perform best.

References

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning*, pages 152–164. Association for Computational Linguistics.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

Meire Fortunato, Charles Blundell, and Oriol Vinyals. 2017. Bayesian recurrent neural networks. *arXiv preprint arXiv:1704.02798*.

Yarin Gal and Zoubin Ghahramani. 2015. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*.

Yarin Gal and Zoubin Ghahramani. 2016a. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Yarin Gal and Zoubin Ghahramani. 2016b. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 473–483.

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

Peiyun Hu, Zachary C Lipton, Anima Anandkumar, and Deva Ramanan. 2018. Active learning with partial feedback. *arXiv preprint arXiv:1802.07427*.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Zachary C Lipton, Jianfeng Gao, Lihong Li, Xiujun Li, Faisal Ahmed, and Li Deng. 2016. Efficient exploration for dialogue policy learning with bbq networks & replay buffer spiking. *arXiv preprint arXiv:1608.05081*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Carl Edward Rasmussen. 2004. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer.

Dan Roth, Chad M Cumby, Xin Li, Paul Morie, Ramya Nagarajan, Nick Rizzolo, Kevin Small, and Wen-tau Yih. 2002. Question-answering via enhanced understanding of questions. In *TREC*.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Ye Zhang, Matthew Lease, and Byron C Wallace. 2017. Active discriminative text representation learning. In *AAAI*, pages 3386–3392.