ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Active learning strategy for CNN combining batchwise Dropout and Query-By-Committee

Melanie Ducoffe & Frederic Precioso
Univ. Nice Sophia Antipolis,
I3S, UMR UNS-CNRS 7271
06900 Sophia Antipolis, France
{ducoffe, precioso}@i3s.unice.fr

Abstract. While the current trend is to increase the depth of neural networks to improve their performance, the size of the training database has to grow accordingly. We thus notice an emergence of tremendous databases, although providing labels to build a training set still remains a very expensive task. In this paper, we tackle the problem of selecting the samples to be labeled in an online fashion. We present an active learning strategy based on query by committee and dropout technique to train a Convolutional Neural Network (CNN). We evaluate our active learning strategy for CNN on MNIST and USPS benchmarks, showing in particular that selecting less than 22 % from the annotated database is enough to get similar error rate as using the full training set.

1 Introduction

The relation between the depth of the architecture, the required amount of training data and the final accuracy of the decision has not only been observed experimentally but it has also been explained in various papers. In their paper [1], Bengio et al. explain clearly that complex decisions can be seen as highly-varying functions and that the decision making algorithm which intends to comprehend all these variations must be composed of many non-linearities. This specificity of deep architectures also impacts the representation compactness of highly-varying functions. In this same paper, they illustrate how deep architectures outperform shallow ones in terms of number of computational units required and thus training samples, in order to represent a given function. Their examples highlight even the differences in representation compactness between deep architectures, with respect to the problem. Considering the huge amount of parameters to be learnt in order to address ImageNet Challenge (60 million parameters for AlexNet, winner in 2012, to reach 152 layers with Microsoft ResNet winner in 2015), one can understand that the training set has to be huge too. Furthermore, in order to better cover the dispersion of the input distribution, strategies to extend the training set have arisen. When considering the difficulty and the cost to gather relevant annotations for Challenges such as ImageNet, the interest for methods working with smaller training sets is increasing.

Our work focuses on the selection of a better subset to be annotated for training, exploiting the theory of committee decisions. We propose a low computation adaptation of Query-By-Committee strategy (QBC) for deep learning. Indeed the huge number of parameters to be determined in a deep architecture prevents us from training a committee of deep networks. Instead, we train a full

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Convolutional Neural Network (CNN) on a selection of training samples, but the selection is handled by a committee of *partial CNNs*. To build the committee, we use batchwise dropout on the current full CNN in order to define as many *partial CNNs* as batchwise dropout runs, thus reducing the computational cost of the standard QBC technique.

2 Related work

Active learning is a framework to automatize the selection of instances to be labeled in a learning process. We consider the context of pool-based active learning where the learner selects queries (i.e. candidate instances to be labeled) among a fixed unlabeled data set. For other variants(*query synthesis, selective sampling*) we refer the reader to Burr Settles [6]. When it comes to pool-based active learning, two approaches can lead to different strategies. A first approach focuses on the target classifier and on minimizing a learning error metrics: the learner will query unlabeled instances on which the confidence of the predicted label is the weakest. This method, *uncertainty sampling*, while being the least computational consuming among all active learning techniques has the main drawback of ignoring much of the output distribution classes and proning to query outliers. Thanks to its low cost and easy setup, uncertainty has been recently adapted to deep architectures for sentiment classification [8].

A second approach does not focus anymore on the classifier prediction only but on the power of ensemble learning with *Query By Committee(QBC)* strategy. The first algorithm based on Query By Committee(QBC) strategy has been proposed by Seung et al. [7]. Instead of trusting only the current incremental classifier, committee decision relies on defining a space of consistent classifiers (i.e. classifiers whose predictions agree with training set labels) where the optimal learner lies in. The aim of the active learning step is then to query a sample which will divide at best the consistent classifier space, also called the *version space*, and so to reduce the possible solutions to converge towards the optimal classifier. As the size of the version space might be infinite, QBC approximates its distribution by sampling a committee of consistent classifiers. Thus the score assigned to a sample is based on the prediction disagreement between all predictions of the classifiers in the committee.

Traditional active learning techniques handle selection of one sample at a time only and thus the score of each new selected instance is independent from previously labeled data. A simple strategy to extend active learning scheme to querying batch of unlabeled data, is to select the top scoring instances as already been applied for a previous deep active learning strategy [8].

In this paper, we consider an active learning method based on Query-By-Committee which selects batch of query using a top score selection scheme in order to optimize the training of a Deep Neural Network. The drawback of QBC is the cost of building a representative committee. Our version allows us to get rid of this computational issue by using a version of dropout called *batchwise dropout* [5]. We dedicate section 3 to the description of our query by committee framework for neural network, while section 4 demonstrates the effectiveness of

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

our method on two benchmark datasets MNIST and USPS.

3 ACTIVE LEARNING STRATEGY

Before starting, let us define some name convention: For the sake of clarity, we denote by full network the deep architecture trained on the current labeled training set and partial network a CNN member of the committee.

The Dropout-based QBC is an active learning strategy which consists in sampling a committee of *partial* deep architectures (each *partial* CNN resulting from a dropout process on the *full* network) whose predictions will be used to select relevant samples to be queried.

We train the *full* network with random initialization, learning rate and early stopping on the current annotated training set. Updates are stopped when the error of prediction on an independent validation step is not further decreasing. When the training has converged, the *full* network is no longer able to learn more knowledge on the input distribution from the current annotated training set. Thus we apply active learning to query new labeled data and add it to the training set. Eventually the *full* network is retrained from scratch on the new training set(*weights and biases are reset with a random initialization*).

When it comes to query by committee for deep architectures, the challenges are to define:

1. **Committee design**: developing a computationally lightened building scheme of disparate *partial* CNNs.
2. **Sample selection**: Proposing a relevant sample selection function based on the committee's predictions.

3.1 A BATCHWISE-DROPOUT COMMITTEE

The goal of the committee is to be representative of the space of consistent hypotheses where the current trained *full* network lies. Let us now detail how we build *partial* CNNs in order to form the committee. To initiate a *partial* CNN while getting rid of the computation cost due to backpropagation, we apply batchwise dropout [5] on our full network. The batchwise dropout [5] is a version of dropout where we use a unique bernouilli mask to discard neurons for each sample in the minibatch. Thus the batchwise dropout reduces quadratically in the percentage of preserved neurons, the number of parameters in the architecture. When considering convolutional layers, the batchwise dropout has one advantage over dropout: the latter removes neurons independently given the spatial locations whereas batchwise dropout is spatially dependant, switching on or off filters so to discard neurons obtained through the same filter. Figure 1 presents how batchwise dropout preserves the consistency in a CNN architecture which allows us to create our *partial* CNNs. The main advantage is to obtain a committee whose members share the same architecture as the *full* network with zero constraints on several connexions. In order to increase the accuracy of each *partial* CNN, our idea is to fine-tune its last layer via a few epochs of backpropagation. *Notice that applying backpropagation on the whole set of layers may conduct partial CNNs to the same settings of parameters.*

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

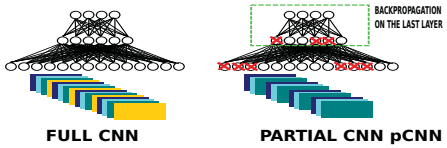


Fig. 1: A run of batchwise dropout to build a *partial* CNN from the *full* CNN

3.2 SAMPLE SELECTION WITH A MARGIN-LIKE IDEA

In the context of query by committee, a sample is considered as informative based on its ability to reduce the number of current consistent hypotheses. Thus the informativeness of a sample is measured by the quantity of disagreement about the prediction of its label among the *partial* CNNs. We propose our own metric based on how much a *partial* CNN may change its decision to be in accordance with the majority. In that order we define a smooth vote on the members of the committee. Let denote the committee as a set of *partial* CNNs: $Committee = \{\ pCNN_i\}$ with p^i the output probability vector of $pCNN_i$. Given a sample $\mathbf{x}$, we first establish its most probable label based on the committee predictions:

$$\mathbf{LABEL}(\mathbf{x}) = \underset{j}{\text{argmax}} \sum_{pCNN_i} \mathbb{1}_{j=\underset{k}{argmax}\ p^i(y=k|\mathbf{x})} \tag{1}$$

We took inspiration from Random Forest margin function [3] in order to produce a ranking of candidates for selection and to have a *soft* pool among the committee. Our point is to take into account the confidence of a *partial* CNN into the score function rg(x) and query the samples with the highest score:

$$rg(\mathbf{x}) = \sum_{pCNN_i} \max_j p^i(y=j \mid \mathbf{x}) - \ p^i(y=\mathbf{LABEL}(\mathbf{x}) \mid \mathbf{x}) \tag{2}$$

We add minibatches of samples instead of one sample as supposed for active learning technique, both to leverage the computational cost owing to successive runs of active learning and to avoid unbalanced size of minibatch (*in that case an adjustement of the learning rate given the size of the last minibatch would be required*). If diversity based scoring may be used to select a subset of samples among the top scoring instances as already proposed for shallower classifiers, we let this issue as an open question for future work.

4 Experiments

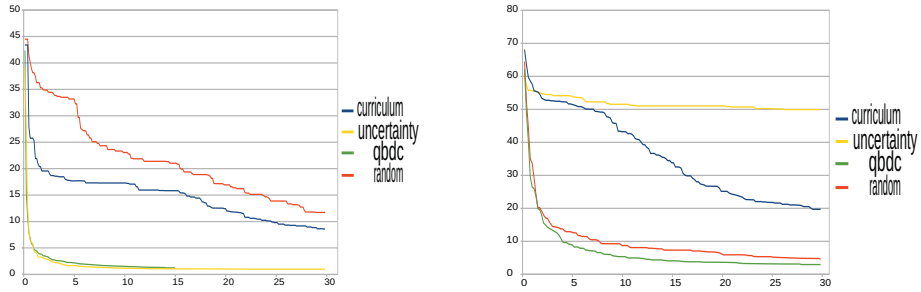
We demonstrate the validity of our approach on two datasets: MNIST (28-by-28 pictures, 50.000 training samples, 10.0000 validation samples and 10.000 test samples, batch size of 64) and USPS (16-by-16 pictures, 4185 training samples, 464 validation samples and 4649 testing samples, batch size of 8) both gray scaled digit image datasets. Both CNN have rectifier activation, other hyperparameters are described in table 1. Note that we do not optimize the hyperparameters depending on the size of the current annotated training set. We picked those two similar datasets to judge of the robustness of our method against different size of unlabeled datasets. Finally our method is efficient on restricted and larger unlabeled pool samples.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

dataset	# filters	filter size	pooling size	hidden layers	Test error
MNIST	[20, 20]	[(3,3), (3,3)]	[(2,2), (2,2)]	[200, 200, 50, 10]	1.1
USPS	[20, 20]	[(3,3), (3,3)]	[None, (2,2)]	[300, 50, 10]	3.25

Table 1: Set of hyperparameters for the CNN used respectively for MNIST and USPS

We perform 5 to 10 runs of experiments and record the test error of the best validation error before an active learning iteration. We start from an annotated training set of size one minibatch selected randomly. We stop both sets of experiments after that 30% of the training set has been selected (15.000 image for MNIST, 1255 for USPS). We sample 5 *partial* CNNs to form a committee. We compare our Dropout-based QBC to uncertainty, curriculum [2] and random selection with a top scoring selection(*see Figure 2*) on a convolutional network. We measure both uncertainty and curriculum scores based on the log likelihood of a sample using as label its prediction on the *full* network. While uncertainty selects samples with the highest log likelihood, our version of curriculum does the exact contrary. We select randomly the set of possible queries among the unlabeled training data. Its size is set to 30 times the minibatch size. The experiments in (*see Figure 2*) conducted on MNIST and USPS illustrate that Dropout-based QBC converges faster to the best accuracy achieved without active learning on the whole annotated training set than the other selection methods: for MNIST we see that less than 26% of the database is necessary to obtain almost the final accuracy (1.23% on test error instead of 1.1 %). When it comes to USPS, larger difference are observed: our Dropout-based QBC is the only active method able to achieve the groundtruth accuracy with less than 22% of the training set.



(a) MNIST (b) USPS

Fig. 2: **Dropout-based QBC with top score selection**: Evolution of the test error given the ratio of annotated data over the training set.

5 Discussion

Owing to a lack of space, we could not present further experiments asserting of the effectiveness of our Dropout-based QBC: we record the average time of an active learning iteration, few minutes are enough to select a batch of queries; we also demonstrate the quality of our *partial* committee compared to a committee

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

built with backpropagation, both versions achieved similar test errors.

Simultaneous and independently to our work, Martin Gammelsaeter [4] also considers doing query by committee by applying dropout on a standard Multi-Layer Perceptron (MLP) to form a committee. To outperform the MLP accuracy, the first layers are extracted from a pretrained Deep Belief Network whose weights are reused each time the MLP is reinitialized after adding a new labeled sample to the training set. While their algorithm share some ideas with ours, it differs in three main aspects: The author uses dropout to build the committee while dropout does not preserve a consistent architecture for CNN. Furthermore dropout does not reduce forward and backward computation for the members of the committee. No backpropagation is applied on the committee while we have experimentally observed that in the case of CNN, this often leads to committee members giving random prediction. His method requires to train the full network with dropout which restricts the context of its use.

6 Conclusion

This paper introduces an adaptation of query by committee for deep architectures. It allows to train Convolutional Neural Network on smaller annotated training set to achieve similar accuracy to the one obtained using much larger annotated database. Our work bridges the computational gap between active learning for deep networks and other shallow classifiers. The use of a committee allows our active learning to have a distributive training of its *partial* CNNs which is a natural advantage of QBC derived methods. We went further into diminishing the computation time with the combination of batchwise dropout to reduce quadratically the number of involved parameters in a *partial* CNN and the backpropagation on the last layer.

References

[1] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *IN NIPS*, pages 153–160. MIT Press, 2007.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, pages 41–48. ACM, 2009.

[3] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, oct 2001.

[4] M. Gammelsaeter. A committee of one. Master's thesis, NTNU, 2015.

[5] B. Graham, J. Reizenstein, and L. Robinson. Efficient batchwise dropout training using submatrices. *arXiv preprint arXiv:1502.02478*, 2015.

[6] B. Settles. *Active Learning*, volume 6 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers, 2012.

[7] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. COLT '92, pages 287–294, New York, NY, USA, 1992. ACM.

[8] S. Zhou, Q. Chen, and X. Wang. Active deep networks for semi-supervised sentiment classification. In *ACL ICCL*, pages 1515–1523, 2010.