Active and Continuous Exploration with Deep Neural Networks and Expected Model Output Changes

Christoph Käding[1,2]
christoph.kaeding@uni-jena.de

Erik Rodner[1,2]
erik.rodner@uni-jena.de

Alexander Freytag[3]
alexander.freytag@zeiss.com

Joachim Denzler[1,2]
joachim.denzler@uni-jena.de

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]Michael Stifel Center Jena, Germany
[3]Carl Zeiss AG, Jena, Germany

Abstract

The demands on visual recognition systems do not end with the complexity offered by current large-scale image datasets, such as ImageNet. In consequence, we need curious and continuously learning algorithms that actively acquire knowledge about semantic concepts which are present in available unlabeled data. As a step towards this goal, we show how to perform continuous active learning and exploration, where an algorithm actively selects relevant batches of unlabeled examples for annotation. These examples could either belong to already known or to yet undiscovered classes. Our algorithm is based on a new generalization of the Expected Model Output Change principle for deep architectures and is especially tailored to deep neural networks. Furthermore, we show easy-to-implement approximations that yield efficient techniques for active selection. Empirical experiments show that our method outperforms currently used heuristics.

1 Introduction

Without any doubt, supervised learning of visual recognition models has made an impressive progress in the last years [1, 2, 7, 3, 6]. This success is based on techniques which allow to learn complex and deep representations from well annotated and large-scale datasets. However, to further foster the application of vision algorithms in related domains, such as biomedical image analysis or quantification in data-driven science, we need to move from static algorithms which are trained only once to continuously learning algorithms. Thereby, given unlabeled data can be actively explored in the search for relevant new training examples and even new visual concepts of completely unknown object classes.

Nonetheless, keeping the annotation effort at a minimum is crucial in most application domains. We are therefore proposing an active learning algorithm that allows for class discovery during the process of guided data annotation. Our algorithm is based on the expected model output change (EMOC) principle originally presented in [4, 12]. In this paper, we show how to generalize it towards layered models, such as deep neural networks. Furthermore, we show that simple back-propagation allows for easily approximating the EMOC criterion and for evaluating it on large-scale unlabeled datasets.

Related Work The authors of [18] present an approach for active learning of deep neural networks based on several selection criteria. A key ingredient for the proposed framework is the pseudo

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

labeling of most certain samples and the update of network parameters based on those pseudo labels as well as user annotated samples. In [15] the question is studied how semantic parts of images can be shared between categories. An active learning method based on the uncertainty if regions are containing useful parts is proposed. An opposite selection criterion is used in [10], where most confident samples are selected for user annotation to deal with noisy web images. The authors argue that even images with high confidence can often result in false decisions. Hence, labeling these samples correctly could have a huge impact for further decisions of the model. However, although all presented active learning strategies are well motivated, they miss the direct link to the ultimate goal of active learning: the reduction of future errors. Furthermore, these techniques only make use of the *current output* of a deep neural network which shall be trained over time. In contrast, our technique considers *changes in all available parameterized layers* and implicitly combine them into a single criterion. Note finally that active sampling of instances for mini-batch gradient descent as introduced in [5] is closely related to active learning. The difference is that active learning tries to reduce the labeling effort for sample acquisition whereas active sampling aims at reducing the number of gradient descent iterations when all labels are known. The authors accelerate mini-batch gradient descent by guiding selection based on uncertainty, significance, and selection history of each sample. The resulting approach is very similar to our criterion.

2 Active Learning with Deep Neural Networks

Continuously learning convolutional neural networks with given labeled datasets is just one step towards a lifelong learning pipeline for visual recognition. In fact, a similarly important aspect is the acquisition of reliable labels, which is still costly when using human annotators for all available data. Therefore, we show how to perform active learning with convolutional neural networks, where only informative, unlabeled examples are automatically selected for being labeled by a human expert. In the following, we first review the expected model output change principle which was originally developed by Freytag et al. [4] as a query strategy for active learning with Gaussian process regression models. We then show that this principle can be applied to convolutional neural networks by utilizing the approximated gradients of the loss function.

Preliminaries and Notation Let $f(\mathbf{x};\boldsymbol{\theta})$ be the output of a neural network with parameters $\boldsymbol{\theta}$ for a given image $\mathbf{x}$. Learning a network from a given labeled training set $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^N$ boils down to minimizing a desired learning objective:

$$\bar{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i) + \omega(\boldsymbol{\theta}) \ . \tag{1}$$

Common choices for the loss function $\mathcal{L}$ are the quadratic loss for regression tasks or the softmax loss for multi-class classification scenarios. The term ω is usually an elastic-net regularization [19] that combines L_2 and L_1-regularization of the parameters $\boldsymbol{\theta}$.

In the following, we focus on layered models $f(\mathbf{x}_i;\boldsymbol{\theta}) = f_L\left(\ldots\left(f_2\left(f_1\left(\mathbf{x}_i;\boldsymbol{\theta}_1\right);\boldsymbol{\theta}_2\right)\ldots\right);\boldsymbol{\theta}_L\right)$ with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L)$ denoting all parameters of the model. Applying the chain rule allows for calculating the partial derivatives with respect to the parameters. In combination with using a gradient descent optimization scheme this is usually referred to as back-propagation [16].

Review of the Expected Model Output Change Principle Automatically choosing the "most beneficial" examples for labeling is a challenging task, since we do not know the labels the human expert will assign to the examples at the time of selection. Therefore, the "usefulness" of the examples for model learning needs to be estimated by treating unknown labels as latent variables. Most standard active learning strategies for approaching this goal are empirically motivated (*e.g.,* selecting examples with highest classifier uncertainty, smallest distance to the decision boundary, etc.). In contrast, the EMOC principle approximates the estimated reduction of errors to avoid redundant queries. The redundancy of an example $\mathbf{x}'$ with respect to the learning result can be measured by calculating the difference of outputs for models learned with and without the currently investigated example:

$$\triangle f(\mathbf{x}') = \mathbb{E}_{y'|\mathbf{x}'}\, \mathbb{E}_{\mathbf{x}}\ \left(d(f(\mathbf{x};\boldsymbol{\theta}), f(\mathbf{x};\boldsymbol{\theta}'))\right) \ . \tag{2}$$

Here, $d(z,z') = \|z - z\|_1$ measures the difference between model outputs and $\boldsymbol{\theta}'$ are the model parameters obtained by additionally training with the labeled example $(\mathbf{x}', y')$. Since the label y' of

the unlabeled example $\mathbf{x}'$ is unknown, we need to marginalize over y' in the above formula (first expectation operation). Examples with a small value of $\triangle f$ are likely not changing the model outputs when being annotated and added to the training set, therefore, they are considered as redundant. In consequence, active learning based on EMOC selects the example with the highest value of $\triangle f$.

In practice, the expectations in Eq. (2) are estimated with empirical means across the dataset (for $\mathbb{E}_{\mathbf{x}}$) and predictive posteriors (for $\mathbb{E}_{y'|\mathbf{x}'}$) based on the results of the current model.

EMOC for Layered Models and Deep Neural Networks The EMOC principle is defined irrespective of the specific type of model or learning algorithm. However, its naive implementation would require learning a model from scratch for each example $\mathbf{x}'$ of an unlabeled dataset. Therefore, techniques are required for efficiently evaluating or approximating the model output of f'. The authors of [4] focus on Gaussian process regression, where a closed-form expression for $\triangle f(\mathbf{x}')$ can be derived, which is efficient to evaluate.

In our case, we could make use of the continuous learning strategy of warm-start optimization [13]. However, this would still be too time consuming for larger sets of unlabeled data. We are therefore taking a shortcut and using a stochastic gradient approximation with just a single sample to estimate model parameter updates:

$$\boldsymbol{\theta}' - \boldsymbol{\theta} \approx \gamma \nabla_{\boldsymbol{\theta}} \bar{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D} \cup (\mathbf{x}', y')) \approx \gamma \nabla_{\boldsymbol{\theta}} \bar{\mathcal{L}}(\boldsymbol{\theta}; (\mathbf{x}', y')) \tag{3}$$

for some constant $\gamma > 0$. We can further approximate the model output change by using a first-order approximation:

$$\|f(\mathbf{x}; \boldsymbol{\theta}') - f(\mathbf{x}; \boldsymbol{\theta})\|_1 \approx \|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta})^T (\boldsymbol{\theta}' - \boldsymbol{\theta})\|_1 \overset{(3)}{\approx} \gamma \|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta})^T \nabla_{\boldsymbol{\theta}} \bar{\mathcal{L}}(\boldsymbol{\theta}; (\mathbf{x}', y'))\|_1 \ . \tag{4}$$

Please note that $\nabla_{\boldsymbol{\theta}} f$ is indeed a Jacobian matrix that can be easily calculated with back-propagation.

The approximation reveals an interesting relationship between the EMOC criterion and continuously learning deep neural networks using gradient descent. If $\|\nabla_{\boldsymbol{\theta}} \bar{\mathcal{L}}(\boldsymbol{\theta}; (\mathbf{x}_i, y_i))\|_1$ is small, the example $(\mathbf{x}_i, y_i)$ will likely not lead to any significant change of the model parameters during optimization and can therefore be considered as redundant. The scalar product in Eq. (4) is also reasonable since changes in some of the model parameters only lead to changes of the model output if the respective component in $\nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta})$ has a high absolute value. Hence, our method avoids labeling examples which are likely redundant.

Since, the marginalization over all possible labels y' can be costly for an increasing number of classes, we additionally use a maximum *a-posteriori* approximation by only considering the most likely label $\hat{y}'$ inferred by the previous model f. Furthermore, we consider active learning with sets of examples, where we are not selecting a single unlabeled example $\mathbf{x}'$ but a whole set $\boldsymbol{X}'$ of examples. This eases continuous learning [13] and also allows for efficient labeling of multiple examples at once. We assume that all examples in $\boldsymbol{X}'$ will obtain the same label y'. Our approximated EMOC score for each set $\boldsymbol{X}'$ is then given by:

$$\tilde{\triangle} f(\boldsymbol{X}') = \sum_{\mathbf{x}' \in \boldsymbol{X}'} \mathbb{E}_{\mathbf{x}} \|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta})^T \nabla_{\boldsymbol{\theta}} \bar{\mathcal{L}}(\boldsymbol{\theta}; (\mathbf{x}', \hat{y}'(\boldsymbol{X}')))\|_1 \ . \tag{5}$$

Since an exhaustive optimization over all possible sets $\boldsymbol{X}'$ is infeasible, we are using M random sets of fixed size K.

3 Experiments

In the following, we describe conducted experiments in detail including dataset, baseline methods, and required parameters.

Network Architecture and Choice of Parameters Since our main interest is in image categorization and understanding, we apply deep convolutional neural networks as model family. In particular, we use a randomly initialized CIFAR10 baseline model [11] pretrained on initially known samples. For initial training as well as for continuous fine-tuning, we choose a mini-batch size of 64 samples and a learning rate of 0.0001. The networks are trained with mini-batch gradient descent [14] using momentum [17] of 0.9 and weight decay [16] of 0.0005. Following the strategy of [13], we enforce

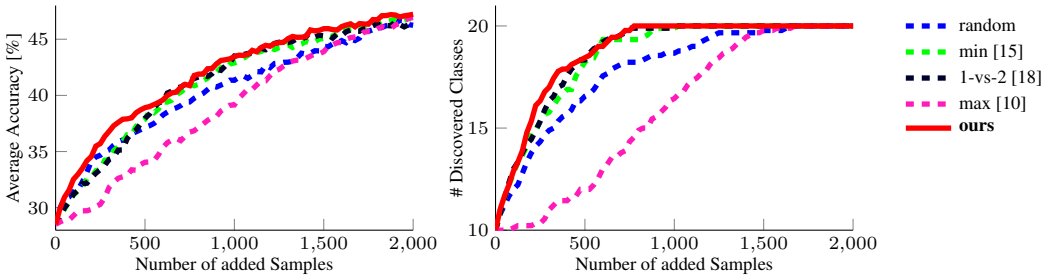

Figure 1: Comparison of different active learning approaches on the CIFAR100 dataset.

every mini-batch during update process to contain a portion of old as well as novel samples. Therefore, we weight old data with $\lambda = 0.9$ and novel data with $1 - \lambda = 0.1$ during sample selection for mini-batch gradient descent. Thus, we prevent over-fitting towards novel samples and incorporate a small fraction of novel data at each mini-batch gradient descent iteration. To continuously incorporate new knowledge and data, we perform 1,000 single mini-batch gradient descent iterations at each update step.

Experimental Setup and Dataset To evaluate our approach, we use the CIFAR100 dataset [11]. It contains small images (32×32 pixel) of 100 classes with 500 training and 100 test samples each. We build an initial training set out of 10 randomly chosen classes with 100 samples. The unlabeled pool consists of additional 100 samples of each known class as well as 100 samples of 10 randomly selected novel classes. All available test samples of the 20 corresponding classes serve as test set. The presented results on the test set are averaged over 9 initializations. In each active learning step, we select $M = 1{,}000$ random sets each comprised of $K = 25$ samples. During this investigation, each set contains only samples from a single class chosen randomly. Each set is evaluated according to the different active learning algorithms under the objective of fast exploration of the unlabeled pool. The selected batch is then labeled and incorporated incrementally into the deep neural network.

Baseline Methods The most trivial baseline is passive learning by random selection of one of the M batches *(random)*. The authors of [15] propose an active learning method selecting most uncertain examples. We adapt this approach to our scenario by selecting sets with lowest average minimum class score *(min)* which is similar to the criterion used in [18, 9]. A related approach is to prefer batches containing samples with lowest difference of the two highest class scores [18, 8]. For an extension to sample sets, we again use the mean score of all contained samples *(1-vs-2)*. In [10], an opposite strategy is proposed where only the most certain samples are selected for labeling *(max)*.

Results In Fig. 1, it can be seen that all evaluated methods, except the max strategy of [10], provide higher gain of accuracy as well as faster discovery of unseen classes than mere random selection of batches. The reason for the inferior performance of the max strategy, which has been shown to be valuable in noisy settings, is the selection of most certain samples which likely belong to already known classes. Both, min [15] and 1-vs-2 [18] perform almost equally. We contribute this behavior to the properties of the dataset and claim that batches with lowest maximal classification scores can also not be assigned clearly to a single class. Finally, we observe that our proposed method is able to outperform all considered baselines in almost all update steps in the presented scenario. Hence, we draw the conclusion that active learning of deep neural networks with expected model output changes is possible and beneficial.

4 Conclusions

We presented an active learning algorithm based on the expected model output change principle, which we extended to deep neural networks. Proof-of-concept experiments on CIFAR100 using ground-truth labels for data batch creation promise good performance of the proposed idea. A deeper evaluation with more general scenarios, other datasets, and deeper networks is left as future work.

Acknowledgements This research was supported by grant DE 735/10-1 of the German Research Foundation (DFG).

References

[1] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In *European Conference on Computer Vision (ECCV)*, 2014.

[2] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[3] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Improved bird species categorization using pose normalized deep convolutional nets. In *British Machine Vision Conference (BMVC)*, 2014.

[4] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision (ECCV)*, 2014.

[5] Jinyang Gao, HV Jagadish, and Beng Chin Ooi. Active sampler: Light-weight accelerator for complex data analytics at scale. *arXiv preprint arXiv:1512.03880*, 2015.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014.

[8] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[9] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *International journal of computer vision (IJCV)*, 2010.

[10] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. *arXiv preprint arXiv:1511.06789*, 2015.

[11] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[12] Christoph Käding, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] Christoph Käding, Erik Rodner, Alexander Freytag, and Joachim Denzler. Fine-tuning deep neural networks in continuous learning scenarios. In *ACCV Workshop on Interpretation and Visualization of Deep Neural Nets (ACCV-WS)*, 2016.

[14] Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, Quoc V Le, and Andrew Y Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011.

[15] David Novotny, Diane Larlus, and Andrea Vedaldi. I have seen enough: Transferring parts across categories. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[16] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 1986.

[17] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th international conference on machine learning (ICML-13)*, 2013.

[18] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2016.

[19] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005.