Informatics Project Proposal - Automatic Curriculum Learning for Deep Models Using Active Learning

Ian McWilliam s0904776

Abstract

In this paper we propose a project researching how active learning methods can be used to automatically construct training curricula, with the aim being to test the hypothesis that training deep models with such 'active curricula' can improve upon other training paradigms such as uniform random sampling, pre-training or traditional curriculum learning. The work will investigate a variety of methodologies for constructing a curriculum using active learning metrics, focusing on how the different training methods affect the performance of convolutional networks on a range of image classification tasks. The output of the project will hopefully be a set of novel ways to improve the training of deep models, as well as a more thorough exploration of the relationship between active and curriculum learning than currently exists in the literature.

1 Introduction

Active Learning

Active learning refers to a learning paradigm wherein machine learning algorithms actively select, or 'query', the samples from which it learns; in the case of deep learning this contrasts with the usual approach of uniformly sampling labeled training samples. The motivation for active learning is that, by allowing it to intelligently select the samples from which it learns, the algorithm can achieve superior generalization performance from a smaller number of training samples than if the samples had been chosen randomly. In domains where unlabeled data is abundant but obtaining labels is expensive, active learning can be used to reduce the cost associated with training a deep model, as the active approach allows the designer to obtain labels only for the samples which will be most beneficial to learning.

There are a variety of methods by which the algorithm can query datapoints, however they generally focus on finding the points in the input space that the algorithm is most 'uncertain' about, allowing the algorithm to fill what could be seen as gaps in its knowledge of the domain. We can broadly categorize four main active learning approaches as follows:

- Uncertainty Sampling - How uncertain
- Query by Committee (QBC) - Disagreement
- Expected Model Change - Gradient change
- Variance Reduction - equation etc for proof for logistic regression

We also note the work of Gal et al [7] who use the uncertainty inherent in Bayesian deep models, estimated by approximating variational inference with Monte Carlo dropout methods.

DISCUSS WORK OF COHN SHOWING IMPROVED GENERALIZATION PERFORMANCE?

Curriculum Learning

A related field is that of *curriculum learning*, which explores how the learning process can be improved by presenting training samples to the algorithm in a meaningful order (with the order defining a 'curriculum'), again in contrast to simply sampling randomly from a training set. Motivated by the way in which humans and animals learn, a curriculum generally begins with 'easy' examples, before transitioning to more challenging ones, with the aim being to improve generalization performance and convergence speed. TALK ABOUT BENGIO PAPER ETC. It is interesting to note that, while similar, active and curriculum learning are somewhat opposed in their learning philosophies, with the former focusing on learning from uncertain/difficult samples and the latter focusing beginning with easy samples before continuing to more difficult ones.

Bengio et al [2] find a theoretical motivation for curriculum learning by suggesting that it can be seen as a *continuation method* (a method of optimization non-convex functions). Curriculum learning has also drawn similarities with *transfer learning*, as the early, 'easier' stages of training can be seen as a separate task, with the network weights then being used as the initial weights for training on the more difficult samples. Similarly curriculum learning can be seen as a form of pre-training, comparing the method to unsupervised pre-training methods as in [5], where pre-training allows the supervised learning to begin in a region of parameter space that led to superior generalization performance.

2 Purpose

Curriculum Construction

One of the main difficulties with implementing curriculum learning is that the curriculum must be constructed prior to training, requiring some predefined measure of difficulty with which to order the training samples. In certain domains there may be a natural ordering of difficulty, for example in the geometric shapes example in [2], however in many tasks manually constructing a curriculum is not as straight forward. This motivates the development of methods to automatically construct learning curricula without requiring expert domain knowledge; in this paper we suggest that using active learning metrics may be used to automatically construct such curricula, potentially providing an efficient way to improve the training of deep learning algorithms.

Active Curricula

Active learning methodologies generally involve calculating a measure of the algorithm's 'uncertainty' in predicting the label of a given input sample. In the traditional active learning setting, the algorithm then queries the label of the sample(s) about which it has the most uncertainty and is then trained using the selected sample(s). An alternative approach is to view the uncertainty measure as a way of approximating the 'difficulty' of a given input sample, which can then be used to construct a learning curriculum, which we term an 'active curriculum'.

As discussed in the introduction, active and curriculum learning offer a somewhat dichotomous view on how best to order training samples (for example Kumar et al [11] refer to active learning as being an "anti-curriculum" method). In active learning the samples about which the algorithm is most uncertain are chosen to train on, while in curriculum learning training focuses on training on easier samples, before progressing to more difficult ones. This trade off is one that has not been explored in the literature, and it is a dimension on which we will focus in the proposed work.

Given an uncertainty/difficulty metric using active learning methods, we have the option of constructing a curriculum beginning with training on the least uncertain samples, or alternatively training on the sa

3 Background

There have been several works that have explored the area of automating the process of constructing learning curricula, as well as similar studies which evaluate methods by which training can be improved by sampling

Self Paced Learning

Automated Curriculum Learning

Active Bias

Other Methodologies

REINFORCEMENT LEARNING PAPER WITH TEMPORALLY RARE EVENTS
DEEP MIND AUTOMATED CURRICULUM LEARNING PAPER
ACTIVE BIAS PAPER
LEARNING WHAT DATA TO LEARN PAPER (REINFORCEMENT METHOD FOR CHOOSING TRAINING SAMPLES)

While these prior works have explored methods for automatically building learning curricula, we believe the work proposed in this paper presents a novel direction for several reasons. The above papers either follow the active learning philosophy of emphasizing 'uncertain'/'difficult' training samples, or the curriculum learning approach of emphasizing 'easier' samples, before moving uniformly sampling; we have not however found a study thoroughly comparing the either approach, nor one that attempts to blend the two approaches for an optimal solution, as we propose to do in this paper. Prior works also generally only compare their methods to relatively simple optimization procedures, for example vanilla SGD or variants thereof. As is explained by Bengio et al in [2], curriculum learning can be compared to pre-training or transfer learning, suggesting that they should also be benchmarks against which the performance of such methods are measured, something which the proposed work will do.

4 Methods

We will test a variety of learning paradigms, each using a training curriculum derived from active learning metrics. Each tested method will therefore require two elements; first, the chosen active learning metric(s), and second the method in which a curriculum is then constructed using the metrics.

Active Learning Metrics

Curriculum Construction Methods

The main way in which curricula will be constructed is by adjusting the probability with which training examples are sampled during training, for example, given an active learning acquisition function $\alpha(x_i, f_t)$, which maps the input sample x_i and the the model at time t, f_t to an estimation of the model uncertainty, we can sample training instances proportionally to the model's uncertainty of each sample:

$$\Pr(s_i|f_t) = \frac{\alpha(x_i, f_t)}{\sum_{j=1}^{N} \alpha(x_j, f_t)} \tag{1}$$

Where $\Pr(s_i = 1|f_\tau)$ represents the probability that the training sample x_i is sampled and N is the number of samples in the training set[1].

Equivalently, an alternative method may take the opposite approach and prefer to sample 'easier' examples about which the model is more certain , in which case we could, for example, assign sampling probabilities as

$$\Pr(s_i|f_t) = \frac{\alpha(x_i, f_t)^{-1}}{\sum_{j=1}^{N} \alpha(x_j, f_t)^{-1}} \tag{2}$$

A further approach could be to alter the preference for 'easier' or 'harder' examples as training progresses, for example preferring easier examples at the start of training before progressing to sampling more uncertain ones later on in training, this can be achieved by setting the sampling probability by

$$\Pr(s_i|f_t) = \frac{\alpha(x_i, f_t)^{\beta_t}}{\sum_{j=1}^{N} \alpha(x_j, f_t)^{\beta_t}} \tag{3}$$

[1]In the case that there is not a constructed training set, and the algorithm is querying synthetic input samples, equation 1 could be calculated with Monte Carlo sampling.

where

$$\beta_t = \frac{2t}{Num\_Epochs} - 1 \tag{4}$$

where Num_Epochs is the number of training epochs. Note that we will start with $\beta_t = -1$ (we assume zero based indexing), corresponding to preferring to sample the least uncertain training instances, with $\beta_t \to 1$ as training progresses, leading to increased probability of sampling uncertain training examples.

We can also use the acquisition function to alter the weight of the prediction loss of each sample as in [3]. This can serve different purposes; as the training examples are no longer sampled randomly, we introduce bias into the model (REF!), by weighting the loss of each sample proportionally to the inverse of the sampling probability, we can correct for this bias, this approach is called *importance sampling*. Concretely, for importance sampling, if we sample from the training data using probabilities as in equation 1, we can performance importance sampling by setting the loss weight of each sample, denoted w_i as follows:

$$w_i = Pr(s_i|f_t)^{-1} * \frac{1}{N} \sum_{j=1}^{N} Pr(s_j|f_t)^{-1} \tag{5}$$

Where N_w is a normalizing constant which ensures the weights have constant unit average, so as not to alter the global learning rate, as in [3]. Alternatively we can weight the loss of each sample proportionally to the model's uncertainty (as opposed to the inverse of the model's uncertainty, as in importance sampling), in conjunction with altering the sampling probability, i.e.

$$w_i = N_w * Pr(s_i|f_t) \tag{6}$$

Constructing curricula with a set of Active Learning methods. Various curriculum constructions, select easy, select hard, easy to hard, hard to easy

5 Evaluation

Datasets

We use the different learning methods to train several architectures on a set of image classification tasks, potentially with several different optimization methods (i.e. Stochastic Gradient Descent, AdaGrad, ADAM):

- MNIST
- CIFAR 10
- CIFAR 100
- Geometric Shapes

Benchmarks

We compare the constructed training methods to a set of benchmark alternatives -

- Uniform Sampling - Sampling randomly from the training data with a uniform probability for all samples.
- Self-paced learning - Methodology as in [11], excluding samples where the loss is greater than a shrinking threshold.

- Pre-training - Unsupervised pre-training as in REF
- Pre-constructed curriculum/transfer learning - For the CIFAR 100 dataset we first train a network to classify the images into their 20 super classes, before appending an additional output layer to the network and retraining to classify images into one of the 100 'fine' class labels (i.e. transfer learning). In the case of the Geometric Shapes database we train with a curriculum as in [2], initially training on the easy 'BasicShapes' dataset, consisting of the same shapes as in 'GeomShapes', but with less variability in the shapes.

We will compare the performance of the different training methodologies, measuring the classification error on a held out test set, as well as comparing how quickly the different methods converge.

6 Outputs

The main output of this project will be a comparison of the proposed training methodologies to a range of benchmarks, testing whether learning with a curriculum constructed using active learning methods can reduce generalization error and convergence speeds. Should the methodologies significantly outperform the benchmark training paradigms, it could represent an innovative way to guide the training of deep models without the need to use expert domain knowledge to construct a learning curriculum prior to training. The work will also represent a more thorough experimental exploration into whether or not it is more beneficial to learn from 'easy' or 'difficult' examples, or whether gradually shifting the focus throughout learning is more beneficial.

7 Workplan

References

[1] E.L.Allgower and K.Georg, "Numerical continuation methods. An introduction", *Springer-Verlag*, 1980

[2] Y.Bengio, J.Louradour, R.Collobert and J.Weston, "Curriculum Learning", *Procedures of the International Conference on Machine Learning*, 2009

[3] H-S,Chang, E,Learned-Miller,A.McCallum, "Active Bias: Training More Accuracy Neural Networks by Emphasizing High Variance Samples", *Advances in Neural Information Processing Systems 31*, 2018

[4] D.Cohn, L.Atlas and T.Ladner, "Improving Generalization with Active Learning", *Machine Learning*, Issue 15, p201-221, 1994

[5] D.Erhan, P.A.Manzagol, Y.Bengio, S.Bengio and P.Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training", *AI & Statistics*, 2009

[6] Y.Gal, R.Islam, Z.Ghahramani, "Deep Bayesian Active Learning with Image Data", *Advances in Neural Information Processing Systems, Bayesian Deep Learning workshop*, 2016

[7] Y.Gal, R.Islam, Z.Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning", *Procedures of the International Conference on Machine Learning*, 2016

[8] A.Graves, M.G.Bellemare, J.Menick, R.Munos, and K.Kavukcuoglu. "Automated curriculum learning for neural networks", *Proc. Machine Learning Research*, 70, p1311-1320, 2017

[9] A.Katharopoulos and F.Fleuret, "Not All Samples are Created Equal: Deep Learning with Importance Sampling", arXiv preprint, arXiv:1803.00942, 2018

[10] N.Justesen, S.Risi, "Automated Curriculum Learning by Rewarding Temporally Rare Events", arXiv preprint, arXiv:1803.0713, 2018

[11] M.Pawan Kumar, B.Packer, D.Koller, "Self-Paced Learning for Latent Variable Models", *Advances in Neural Information Processing Systems 25*, 2010

[12] A.Owen, Y.Zhou, "Safe and effective importance sampling", *Journal of the American Statistical Association*, 95(449), p135-43, 2000

[13] B.Settles, "Active Learning Literature Survey", Computer Sciences Technical Report 1648, 2009

[14] D.Weinshall, G.Cohen, "Curriculum Learning by Transfer Learning: Theory and Experiments with Deep Networks", arXiv preprint, arXiv:1802.03796, 2018