Continuous Learning

Learning New Observations in the Era of Big Data

Abdelrahman Ahmed

17 October 2017

1 Introduction

The rise of big data, a term that refers to the rapid explosion of available information, brings both remarkable opportunities to modern society and difficult challenges to data scientists (Fan et al., 2014). Massive amounts of high-dimensional data are continuously generated and stored through various outlets such as social media. Vagata & Wilfong (2014) reported that Facebook's data warehouse receives 600 terabytes of data each day, and has seen a threefold increase in growth since the previous year. This example indicates the magnitude of the data as well as the trajectory and speed of the growth. The size and dimensionality of the data have proved to be a challenge to machine learning algorithms as they need to continuously adapt to the new observations. This report explores various forms of continuous learning, such as online learning and incremental learning, that utilise the parameters that have been learned using the original dataset for new observations. It then examines the differences between the continuous learning methods and standard methods where the models are retrained from scratch.

2 Background

Gradient descent is an optimisation algorithm used in machine learning to find the most optimal weights or parameters of various algorithms, such as logistic regression and artificial neural networks. The aim of the algorithm is to find the weights that minimise the prediction error of the model measured on the training and test dataset. It continuously updates the parameters of the model, as it takes steps proportional to the negative gradient until it finds a local minimum (Ruder, 2016). The initial weights are usually randomly generated; however, the optimal weights found during training can be saved. The saved weights that achieved local minimum can then be reused when retraining the model when necessary. This dramatically improves the training performance as the gradient

descent algorithm will start with weights that are close to the local minimum, reducing the time needed to converge (Ruder, 2016). There are three variants of gradient descent: stochastic gradient descent, batch gradient descent and mini-batch gradient descent. Mini-batch gradient descent is the most commonly used variant due to its performance compared to the other variants.

3 Continuous Learning Methods

3.1 Online Learning

Online learning algorithms utilise learned parameters and retrain the model on new observations in the data as they become available. This allows the model to learn in a sequential manner and adapt to recent observations (Blum, 1996). Online learning is commonly used in cases where it may not be computationally feasible to retrain the model over the entirety of the dataset due to its size, dimensionality and scarcity of resources. Stochastic gradient descent (SGD) is a variant of the gradient descent algorithm that is often described as an online learning algorithm. This variant uses individual observations from the dataset to update the parameters of the model (Bottou, 1998).

Online algorithms are beneficial in situations where it is essential for the model to dynamically adapt to new patterns in the data, such as collaborative filtering in recommender systems. However, one of the disadvantages of these algorithms is that they are more influenced by recent observations and tend to forget previously learned information, this is sometimes referred to as catastrophic interference (McCloskey & J. Cohen, 1989).

3.2 Incremental Learning

Incremental learning is a variation of online machine learning which aims to address the catastrophic interference problem. Gepperth & Hammer (2016, p. 1) define incremental learning as a method of 'learning from streaming data, which arrive over time, with limited memory resources and, ideally, without sacrificing model accuracy'. The goal of incremental learning is to allow the model to adapt to new data while maintaining its prior knowledge without retraining the model from scratch. There are various implementations of incremental learning where some incremental models rely on learning compact representations of the observed data, and others utilise parameters to control the relevancy of older data. Depending on the implementation used, there are still some disadvantages in terms of either configuring additional parameters for data relevancy or relying on compact learning representations (Gepperth & Hammer, 2016).

3.3 Transfer Learning

Machine learning methods commonly work under the assumption that both the training and test data belong to the same feature space and statistical distribution, and when changes occur the models need to be rebuilt from scratch. However, this is impractical in many real-world applications where it is not feasible to recollect a large amount of data or retrain the model from scratch (Pan & Yang, 2010). Transfer learning can use the knowledge for related tasks on either the same or similar datasets, which can be achieved through using the pre-trained weights on a different dataset. An illustrative example of transfer learning is recognising cars, then applying the trained model to a related task such as recognising trucks. It can also be used to recognise different aspects in the same dataset, for example, recognising petals instead of leaves on a dataset containing images of plants.

4 Standard Learning Methods

The standard approach in machine learning is to learn in order to accomplish a specific task that is associated with a dataset. Models are trained using a dataset and are only capable of performing that particular task. Similar to online algorithms, most standard learning algorithms use variants of gradient descent such as batch and mini-batch gradient descent algorithms (Ruder, 2016).

Unlike stochastic gradient descent, the batch gradient descent only updates the model parameters after evaluating all the training examples which results in a more stable error gradient and better convergence in some problems. However, this requires the entire dataset to be loaded in memory which substantially affects the training speed and may not be feasible for big datasets (Ruder, 2016). Mini-batch gradient descent aims to address the performance issues by instead taking smaller batches from the training dataset, and updating the model parameters after each epoch this leads to better convergence and improved performance. One of the downsides of mini-batch is the addition of another hyperparameter, the batch size, that needs to be configured to achieve optimal performance (Li et al., 2014).

5 Discussion

Standard machine learning approaches can lead to better representations of data, and more stable error gradients and better convergence when using gradient descent algorithms. Batch learning can also utilise parallel processing to compute multiple gradients simultaneously to improve the performance on larger datasets (Li et al., 2014). However, with the increase in data generated

each day and the high dimensionality of big datasets, it may not be practical to frequently retrain models on big datasets from scratch using these methods.

In the era of big data, new data observations are constantly being made available. Continuous learning methods aim to address the need to keep a model up to date without retraining from scratch. Online and incremental learning techniques can be used in areas where data is continuously being generated and the model needs to frequently update itself, such as stock predictions and recommender systems (Gepperth & Hammer, 2016). Transfer learning techniques can be used to periodically retrain models using the learned weights of an existing model and utilise them on new observations to improve the training performance (Pan & Yang, 2010). These techniques are much more computationally efficient for big datasets and streaming data. They also allow the models to stay up to date and adapt to new patterns in data without major sacrifices in terms of accuracy. There is a notable lack of research in this area compared to standard machine learning methods. Online and incremental learning in particular can benefit from further research to improve their performance as they are some of the essential techniques in dealing with big data and streaming data.

References

Blum, Avrim. On-line algorithms in machine learning. In *In Proceedings of the Workshop on On-Line Algorithms, Dagstuhl*, pp. 306–325. Springer, 1996.

Bottou, Léon. Online algorithms and stochastic approximations. In Saad, David (ed.), *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998. URL http://leon.bottou.org/papers/bottou-98x. revised, oct 2012.

Fan, Jianqing, Han, Fang, and Liu, Han. Challenges of big data analysis. *National Science Review*, 1(2):293–314, 2014. doi: 10.1093/nsr/nwt032. URL +http://dx.doi.org/10.1093/nsr/nwt032.

Gepperth, Alexander and Hammer, Barbara. Incremental learning algorithms and applications. In *European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium, 2016. URL https://hal.archives-ouvertes.fr/hal-01418129.

Li, Mu, Zhang, Tong, Chen, Yuqiang, and Smola, Alexander J. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 661–670, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623612. URL http://doi.acm.org/10.1145/2623330.2623612.

McCloskey, Michael and J. Cohen, Neal. *Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem*, volume 24. 12 1989.

Pan, Sinno Jialin and Yang, Qiang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010. ISSN 1041-4347. doi: 10.1109/TKDE.2009.191. URL http://dx.doi.org/10.1109/TKDE.2009.191.

Ruder, Sebastian. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. URL http://arxiv.org/abs/1609.04747.

Vagata, Pamela and Wilfong, Kevin. Scaling the facebook data warehouse to 300 pb, 2014. URL https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/.