International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Conceptual Framework for Deep Learning Algorithms and Their Applications

Dr. Jayeshkumar Madhubhai Patel

Associate Professor – PG Department [M.C.A.], Ganpat University, Gujarat, INDIA.

Abstract
Deep learning plays an important role in our lives. It has already had a huge impact in areas such as cancer diagnosis, personalized medicine, self-driving cars, predictive analytics and speech recognition. Intuitive hand-crafted features used in traditional learning, classification, and model recognition systems are valuable for large data sets. In many cases, depending on the complexity of the problem, DL can also overcome the limitations of sparse networks in the past that prevent effective training and spatial representation of high-dimensional training data. A deep network uses many layers (deep) of units with sophisticated algorithms and architecture. This paper reviews several optimization methods to improve training accuracy and reduce training time. We delve into the math behind the training algorithms used in the latest deep networks. We describe current failures, improvements, and implementations. The review also covers different types of deep architectures, such as deep convolution networks, deep networks, regular networks, reinforcement learning, differential autoencoders, etc.

Keywords: Machine learning algorithm, optimization, artificial intelligence, deep neural network architectures, convolution neural network, backpropagation, supervised and unsupervised learning.

1. Introduction

Deep learning is a machine learning technique that trains computers and artificial intelligence. It is inspired by the structure of the human brain. Deep learning originated as artificial neural networks (ANNs) and after several years of research and development, it was developed to be more efficient compared to other machine learning algorithms [1].

Deep learning algorithms aim to draw the same judgments as humans do by continuously analyzing data based on logical patterns. Deep learning enables machines to process images, text, or audio files like humans to accomplish human-like tasks. To achieve this, deep learning uses a multi-layered set of algorithms called neural networks. As the name suggests, deep learning involves digging deeper into the layers of the network, which includes hidden layers. As one digs deeper, more complex information is extracted. Deep learning is based on iterative learning methods, which expose machines to large volumes of data. It helps computers learn to recognize behavior and adapt to changes. Machines are able to learn the differences between data sets, understand logic, and reach better decisions after repeated exposure [2].

A review on deep learning is presented in [3]. In [3], the following information is provided: Statistical models contain many layers to control. They are enabled to learn data representation with different levels for abstraction through deep learning. Through the use of these technologies, modern technology has been developed in many fields, including visual recognition, speech recognition, genomics, drug discovery, and many others. The background information is used by deep learning to find complex patterns in large data sets, and this is to determine the way the machine has changed the internal parameters of the machine, and these are used to calculate the representation in each layer based on he. representation in the background layer. RNNs are used for sequence data, such as speech and text, while CNNs are used for processing images, speech, audio, and video.

In recent years, many studies, such as [8-12], have used deep learning and shown its power in many fields, including health and health information, such as bioinformatics, medical imaging, deep understanding, physician are Information and public health [4]; Motion image segmentation [5], to improve FOC performance [5]; and a new method designed for an automatic controller with FPGA [3] and PID neural network [4].

This paper aims to provide a literature review on well-known deep learning algorithms. We highlight their capabilities, relevance, limitations, and applications, which many researchers in this field can benefit from knowing. In addition, a critical analysis of various deep learning algorithms is given, paying particular attention to their advantages, disadvantages, and applications, whether supervised or unsupervised. In addition, the application of deep learning algorithms to healthcare is also presented.

2. Background

Deep learning is a subset of machine learning algorithms that use multiple layers to slowly extract high-level features from the input source. It mainly focuses on image processing, in which the image is divided into several layers. Larger layers classify human-like concepts such as numbers or letters or faces and smaller layers can classify edges. Deep learning is the process of analyzing data in a logical and logical manner. This is achieved by using a combination of several algorithms called neural networks. The purpose of designing a neural network is to mimic the functioning of the human brain[5].

Similar to our brain, neural networks are designed to recognize patterns and classify different types of data taken from real life and are trained to complete similar operations on the data. Whenever our brain receives new information, it has the ability to associate it with something familiar, which is the same concept in deep networks as well. Such networks perform various tasks such as clustering, classification, regression on large datasets to organize the unlabeled data according to their relative strength among known datasets. To explain it further, here are a few methods that have the irresistible ability to leverage deep learning models to solve tasks that machine learning models are trying to solve. It provides different concepts involved in deep learning[4].

3. Materials and Methods

This research presents the new sparsification-driven framework in an effort to obtain optimized models for sustainable technological development. The proposed framework aims at exploiting the idea of sparsity in deep learning to map compressed models onto resource limited devices in line with long-term technology roadmap. Efficient models are those that optimize footprint metrics such as model size, inference latency, and training time while minimizing quality loss and improving model generalization. Sparsity in deep learning enables compression techniques to target the representational efficiency of over-parameterized models "Sustainable technological development involves articulating functions to meet future demands and societal needs while reducing environmental impact" [7].

However, sparse-centric approaches to sustainable technological development are currently lacking. The framework aims to make neural networks more efficient, sustainable, and effective in real-world applications by discussing techniques for reducing their size, optimizing their training schedules, and considering practical deployment factors. Thus, the research that presents a conceptual framework based on the analysis of sparsity in scientific publications will help to contribute to achieving goals of a sustainable technology development for creating environmentally friendly and sustainable technologies[8].

The following steps were taken to develop the framework:

1. An analysis of literature on the fundamental ideas pertaining to the key research topics that include model pruning, efficiency, inference, and sustainable information technology.
2. identifying connections between research concepts.
3. design and development of the proposed conceptual framework for pruning.

3.1 Reviewing Literature on Specific Research Concepts

This study has undertaken a rigorous and structured approach to determine the relationships between the key concepts of this research. The structured approach is based on Kitchenham guidelines [6], which include the preparation of research questions, a search strategy, and research selection criteria.

3.2 Planning the Review

The general research question which was identified is as follows: How can model pruning be used to enhance efficiency and support sustainable technological development in deep learning models? To make it possible to further explore the problem, the general question was divided into three narrower sub questions[9].

1. In what way does pruning help make deep learning models more efficient?
2. What is the association between model efficiency and inference performance in sparse models?
3. How exactly does increased model efficiency impact the further development of sustainable technologies?

3.3 Developing a Review Protocol

The objective is to create a conceptual framework that guides the implementation of pruning techniques to improve the efficiency and sustainability of deep learning models[6]. For the search strategy in this study, the key concepts represented the search terms that were applied to a variety of resources. Boolean search strategy is defined as the process of converting a research question into a research string with the view of querying a database or search engine, (Sivarajkumar et al., 2024). The academic internal databases employed in this study include IEEE Xplore, Google Scholar, PubMed, semantic scholar, and ACM Digital Library. Keywords are "model pruning," "deep learning efficiency," "neural network optimization," and "sustainable AI[10].

To fulfill the inclusion criteria, we restricted our sources to the articles from the referred scientific journals, evaluation reports and technical papers and conference papers of peer review. More precisely, we only considered papers that focused on the effects of pruning on the model as well as papers that focused on the broader issue of sustainability in AI and deep learning. We prioritized articles that demonstrated a link between the search terms and described the relationships between them. Additionally, we included articles addressing the search terms in conjunction with any of the four core concepts of this study: model pruning, efficiency, inference, and sustainable technology development[11].

The articles were excluded based upon the following: articles that did not relate directly to deep learning or pruning techniques, papers and articles with no prior experimental results or data, the papers with the search terms but not looking into the connection or relevance between them were considered irrelevant and thereby excluded[10].

To achieve the data extraction methods, we created a standardized data extraction form to capture relevant information from each study, including pruning methods, metrics for efficiency, impacts on inference, and sustainability considerations.

3.4 Conducting the Review

The identification of research was achieved by using the 'AND' and 'OR' Boolean operators to construct search strings to extract articles that join terms based on the research questions. The search strings included the following "model pruning AND deep learning efficiency," "pruning techniques AND inference performance," and "sustainable AI AND neural networks." The strings for the OR operator included "model pruning OR deep learning efficiency," "pruning techniques OR inference performance," and "sustainable AI OR neural networks." In the process of selecting primary studies, initially, we searched and selected studies based on the titles and the abstracts of the studies and then, we checked and selected the full text of these studies. Of 500 original identified articles, 120 papers were reviewed after excluding irrelevant and low-quality articles while 46 were reviewed in this final review.

Data was systematically extracted using the predefined form and the extraction process was monitored to maintain consistency and accuracy. Some of the data considered during extraction involved the type of pruning technique applied whether structured or unstructured, efficiency measurements such as model size and inference latency, and whether sustainability impacts have been realized such as energy efficient consumption[5].

3.5 Identification of relationships between key concepts

These interconnections were established from a systematic review of the literature in the field. To identify studies that reported on the effects of pruning on the performance of neural network models, we concentrated on peer-reviewed journal articles, conference papers and technical papers that described certain approaches to pruning and their results. Also, the relevant literature concerning sustainability in AI and deep learning was discussed. Any articles that showed utilization of the search terms and explained the correlation between the identified search terms were included into the study. This involved reviewing different articles to see how model pruning affects efficiency and inference and how such gains fitted sustainable technology progress[2].

4. Benefit of Deep learning over machine learning

Machine learning is simply the progression of extraction of information from a large dataset and loading them into the machine for further process. But deep learning was technologically advanced as an alternative method of machine learning. The methodology used in machine learning is to resolve a substantial number of tasks under the control of a human. But deep learning algorithm is competent of generating novel features by itself. In Machine learning, features are to be recognized precisely by users. Again, in deep learning, problems are resolved on an end-to-end basis, but machine learning solves it by dividing the tasks into several small pieces and then the obtained results are bought together to complete it. In all together, deep learning is a process of artificial intelligence, which enables the processing of data similar to a human brain[9].

5. Summary and conclusion

This paper introduces a conceptual framework, and a novel multi-objective optimization algorithm developed to improve the performance and sustainability of pruning approaches in DL models. The framework aims to address key challenges associated with model size, computational efficiency, and sustainable technology development by transitioning large, over-parameterized neural networks into sparse, optimized models. This brings one of the biggest advantages of pruned models that can be easily scaled and, particularly in resource-constrained environments such as mobile devices.

6. References

1 Deep Learning: Concepts, Architectures, Workflow, Applications and Future Directions - Jayeshkumar Madhubhai Patel - IJFMR Volume 5, Issue 6, November-December 2023. DOI 10.36948/ijfmr.2023.v05i06.11497
2 Assortment of Machine Learning model on Gujarati Web Page Sets. CD Patel, J Patel. GRADIVA REVIEW JOURNAL 8 (7), 1081-1087, 2022. 2022
3 Natural computing algorithms–a survey; B Chawda, J Patel; International Journal of Emerging Technology and Advanced Engineering 6 (6)
4 Stock Market Portfolio Management: A Walk-through BV Chawda, JM Patel International Journal on Recent and Innovation Trends in Computing
5 Bharat V. Chawda, Jayeshkumar Madhubhai Patel,"Investigating Performance of Various Natural Computing Algorithms", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.1, pp.46-59, 2017. DOI:10.5815/ijisa.2017.01.05
6 Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., & Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. Neurocomputing, 599, 128096. https://doi.org/10.1016/j.neucom.2024.128096
7 Cheng, H., Zhang, M., & Shi, J. Q. (2024b). A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–20. https://doi.org/10.1109/TPAMI.2024.3447085
8 Gupta, M., & Agrawal, P. (2022). Compression of Deep Learning Models for Text: A Survey. ACM Transactions on Knowledge Discovery from Data, 16(4), 1–55. https://doi.org/10.1145/3487045
9 Ebrahimi, A., Pullu, V. N., Pierre Langlois, J. M., & David, J.-P. (2023). Iterative pruning algorithm for efficient look-up table implementation of binary neural networks. 2023 21st IEEE Interregional NEWCAS Conference (NEWCAS), 1–5. https://doi.org/10.1109/NEWCAS57931.2023.10198137
10 Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2023). Efficient Transformers: A Survey. ACM Computing Surveys, 55(6), 1–28. https://doi.org/10.1145/3530811
11 Kitchenham, B., & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. Information and Software Technology, 55(12), 2049–2075. https://doi.org/10.1016/j.infsof.2013.07.010
12 Poppi, R. J., &Massart, D. L. (1998). The optimal brain surgeon for pruning neural network architecture applied to multivariate calibration. Analytica Chimica Acta, 375(1–2), 187–195. https://doi.org/10.1016/S0003-2670(98)00462-0