Leveraging AWS EMR for Scalable and Efficient Neural Network Deployment in Cloud Computing

Yevhen Kyrychenko, Ihor Malyk
Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine, kyrychenko.yevhen@chnu.edu.ua, i.malyk@chnu.edu.ua, ORCID: 0009-0005-6150-5410, 0000-0002-1291-9167

Keywords: cloud computing, analytical research, cloud, cloud technologies, serverless architecture, AWS (Amazon Web Services), Elastic MapReduce, neural networks, web service

Abstract. This research investigates the potential of AWS Elastic MapReduce (EMR) as a cloud-based solution for building and deploying neural networks, focusing on its performance, scalability, and flexibility compared to traditional on-premises systems. The study offers a comparative analysis of stochastic neural networks deployed using AWS EMR Serverless, with financial data from Google Finance used to assess model performance through the Root Mean Square Error (RMSE) metric. The results highlight the advantages of cloud-based machine learning, particularly in handling large-scale datasets.

Outline of the main material. Cloud computing's ability to dynamically scale computational resources is very crucial for training large neural networks that require significant processing power. AWS EMR provides access to advanced hardware, such as GPUs and TPUs, which accelerates the training process and reduces the time to results. This capability makes cloud platforms like AWS EMR attractive for complex machine learning tasks where traditional infrastructure falls short [1].

Large datasets require robust computational resources, and AWS EMR Serverless offers a scalable infrastructure that automatically allocates and releases resources based on demand. By distributing loads across processors or GPUs, AWS EMR efficiently handles parallel processing, making it ideal for high-performance machine learning projects [2].

In this study, financial data from Google Finance is used to train neural networks for predicting stock price fluctuations [3]. The architecture features a Flask web server that interacts with the EMR cluster, along with Amazon S3 for storing models and data. AWS EMR Serverless optimizes resource allocation, preventing over- or under-allocation, and supports open-source tools like Hive and Spark [4].

Though AWS EMR Serverless excels in large-scale data processing, it may be less efficient for smaller datasets that can be managed on a single machine. Nonetheless, its ability to manage vast datasets and complex models makes it well-suited for large-scale machine learning tasks.

*Conclusions***.** In conclusion, combining neural networks with AWS EMR opens new possibilities for researchers and developers by offering a scalable, efficient, and flexible environment for AI projects. As cloud technologies evolve, their role in supporting large-scale machine learning will grow, making them critical tools for modern AI development. This study provides insights into the effective use of cloud infrastructure for machine learning, offering guidance for future deployment strategies.

References

[1] Aach, M., Inanc, E., Sarma, R. *et al.* Large scale performance analysis of distributed deep learning frameworks for convolutional neural networks. *J Big Data* 10, 96 (2023). https://doi.org/10.1186/s40537-023-00765-w.

[2] Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97-107.

[3] Stocks API Documentation. URL: https://polygon.io/docs/stocks/getting-started

[4] What is Amazon EMR Serverless? URL: https://docs.aws.amazon.com/emr/latest/EMR-Serverless-UserGuide/emr-serverless.html.