Universitatea Tehnică a Moldovei

DEEP LEARNING MODEL APPLICATION IN GALAXY IMAGE CLASSIFICATION
APLICAREA MODELELOR "DEEP LEARNING" ÎN CLASIFICAREA IMAGINILOR GALACTICE

Student: **Gavrilița Mihail**

Conducător: **Rusu Viorel,**
lector universitar

Chişinău, 2020

Rezumat

Teza cu numele **Aplicarea modelelor ”Deep Learning” în clasificarea imaginilor galactice**, prezentată de studentul Gavrilița Mihail în calitate de proiect de Master, a fost dezvoltată în cadrul Universității Tehnice din Moldova. Aceasta este scrisă în limba engleză și conține 61 pagini, 14 figuri, 10 listări de cod și 39 de referințe. Teza consistă dintr-o listă de figuri, o listă de listări de cod, introducere, patru capitole, concluzie și bibliografie.

Analizarea manuală a unor cantități imense de date astronomice în continuă creștere devine mai dificilă în fiecare zi. Cu viteza actuală a fluxului de date și a procesării datelor noi, oamenii de știință nu ar putea analiza niciodată datele disponibile. Abordările tradiționale sunt fie prea lente (abordări manuale), fie nu au cunoștințele necesare pentru a extrage informații semnificative (abordări statistice). Oferirea unei soluții rapide și inteligente pentru procesarea datelor este ceea ce a motivat acest proiect.

Acest document își propune să studieze modul în care modelele de învățare automată pot ajuta la analiza datelor rezultate din domeniul respectiv. Datorită capacității sale de a extrage în mod automat caracteristici complexe și semnificative din date, modelele ”deep learning” sunt candidații principali pentru sarcină și ar putea oferi o alternativă mai ieftină și mai rapidă metodelor clasice de procesare a datelor.

Pentru a înțelege mai bine necesitățile oamenilor de știință care lucrează cu date a fost analizat domeniul problemei. Proiectul a urmat cu o analiză a aspectelor teoretice ale procesării datelor și învățării automate. În continuare, a fost documentat procesul de dezvoltare urmat de o analiză a rezultatelor obținute. În concluzie, au fost discutate posibilele direcții de dezvoltare a produsului.

Acest document este destinat cititorilor specializați în domeniul tehnic.

Abstract

The thesis named **Deep Learning Model Application in Galaxy Image Classification**, presented by student Gavrilița Mihail as a Master project, was developed at the Technical University of Moldova. It is written in English and contains 61 pages, 14 figures, 10 listings and 39 references. The thesis consists of a list of figures, a list of listings, an introduction, four chapters, a conclusion, and a list of references.

Parsing huge amounts of ever growing astronomical data by hand becomes harder every day. With the current speeds of new data inflow and data processing, scientists could never parse the data available. Traditional approaches are either too slow (manual) or lack the insights to extract meaningful information (statistics). Providing a fast and insightful solution for data processing is what motivated this project.

This document aims to study how machine learning models can help analyze data resulting from that domain. Because of it's ability to extract complex and meaningful features automatically from data, deep learning is a prime candidate for the task and could provide a cheaper and faster alternative to classical methods of data processing.

The domain of the problem was analyzed to better understand the necessities of scientists that work with data. The project followed with an analysis of the theoretical aspects of data processing and machine learning. Following, the development process was documented followed by an analysis of the obtained results. In conclusion, possible directions of future development were discussed.

This document is intended for readers with technical background.

Table of contents

Introduction

Nowadays, the scientific field of astronomy is overwhelmed by a ”data tsunami”. Everyyear, petabytes of new images and measurements flood the servers of scientists all around the globe. The scientific method is based on data analysis and processing all this data with classical methods is getting less and less effective. The tools that are used are either too slow (manual approach) or lack complexity and a deeper understanding of the data in hand (statistic instruments). The field is in desperate need for tools that would allow for a deeper understanding of data and, at the same time, would be able to go through vast amounts of it fast.

What other tools can a scientist use, besides the manual or statistical approach? Machine learning provides many tools for different types of problems such as regression, classification or object detection. And out of them, deep learning models show most interest to scientists because of their ability to extract complex features from data all by themselves. And, compared to other machine learning models, deep learning models only grow stronger with more and more data that is fed into them.

Deep learning models promise fast and accurate solutions for many types of classification problems. With their ability to learn complex spatial features, Deep CNN networks are the prime candidate to battle this astronomical data tsunami. Transfer learning is a relatively new approach in creating classification models. It allows developers to ”reuse” the knowledge gained by models when solving problems to aid in training for another problem. Thus, the project can use pre-trained models to drastically lower the amount of data necessary for training.

Considering the arguments from above, the following research is proposed: to apply deep learning models and train on classified astronomical data using transfer learning to speed up the training process and to lower the requirements to the volumes of training data. Such a machine learning model would allow scientists to parse large amounts of data much quicker than with any other manual approach, and, if the accuracy of the model is high, do so with few mislabeled data.

The current thesis consists of four chapters. In the first chapter, problem analysis and domain description is performed. The chapter analyses why processing data is hard and describes different approaches for it. The second chapter covers the theoretical aspects of the project. Different data analysis and machine learning notions are covered. The third chapter examines the implementation of the application. The technologies and methodologies used to develop the application are also discussed. The last chapter goes over the final results of the project. The resulting model is analysed via several metrics.

References

1 Berriman, G. Bruce, and Steven L. Groom. *How will astronomy archives survive the data tsunami?*. arXiv preprint arXiv:1111.0075 (2011).

2 Harvard University Archives. *UAV 630.271 (E4116)*. https://library.harvard.edu/collections/project-phaedra-0 (November, 2019)

3 Van Vliet, Kim, and Claybourne Moore. *Citizen science initiatives: Engaging the public and demystifying science*. Journal of microbiology & biology education 17, no. 1 (2016): 13.

4 Zooniverse, *What is the Zooniverse*, https://www.zooniverse.org/about (November, 2019)

5 Galaxy Zoo, *About*, https://blog.galaxyzoo.org/about-2/ (November, 2019)

6 Foldit, *The Science Behind Foldit*, https://fold.it/portal/info/about (November, 2019)

7 Baron, Dalya. *Machine Learning in Astronomy: a practical overview*. arXiv preprint arXiv:1904.07248 (2019).

8 Bennett, Kristin P., and Colin Campbell. *Support vector machines: hype or hallelujah?*. Acm Sigkdd Explorations Newsletter 2, no. 2 (2000): 1-13.

9 Statinfer, *SVM: Advantages Disadvantages and Applications* https://statinfer.com/204- 6-8-svm-advantages-disadvantages-applications/ (November, 2019)

10 Quora, *What are the advantages and disadvantages for a random forest algo- rithm?* https://www.quora.com/What-are-the-advantages-and-disadvantages-for-a- random-forest-algorithm (November, 2019)

11 Ponnusamy, R., S. Sathyamoorthy, and K. Manikandan. *A Review of Image Classification Approaches and Techniques*. International Journal of Recent Trends in Engineering & Research (IJRTER) Volume 3 (2017): 2455-1457.

12 cs231n, *Convolutional Neural Networks for Visual Recognition*, https://cs231n.github.io/convolutional-networks/ (November, 2019)

13 Zeiler, Matthew D., and Rob Fergus. *Visualizing and understanding convolutional networks*. In European conference on computer vision, pp. 818-833. Springer, Cham, 2014.

14 He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep residual learning for image recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

15 Ball, Nicholas M., and Robert J. Brunner. *Data mining and machine learning in astronomy*. International Journal of Modern Physics D 19, no. 07 (2010): 1049-1106.

16 Wells, Donald Carson, and Eric W. Greisen. *FITS-a flexible image transport system*. In Image Processing in Astronomy, p. 445. 1979.

17 Famili, A., Wei-Min Shen, Richard Weber, and Evangelos Simoudis. *Data preprocessing and intelligent data analysis*. Intelligent data analysis 1, no. 1 (1997): 3-23.

18 Lintott, Chris J., Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick et al. *Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey*. Monthly Notices of the Royal Astronomical Society 389, no. 3 (2008): 1179-1189.

19 IPython, *The Jupiter Notebook*, https://ipython.org/notebook.html (December, 2019)

20 Python, *What is Python? Executive Summary*, https://www.python.org/doc/essays/blurb/ (December, 2019)

21 fastai, *Welcome to fastai*, https://docs.fast.ai/ (December, 2019)

22 TensorFlow, *Why TensorFlow*, https://www.tensorflow.org/about/ (December, 2019)

23 PyTorch, *Features*, https://pytorch.org/features/ (December, 2019)

24 Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. *Scikit-learn: Machine learning in Python*. Journal of machine learning research 12, no. Oct (2011): 2825-2830.

25 Project Jupiter, *About Us*, https://jupyter.org/about (December, 2019)

26 Dummies, *What is Google Colaboratory?*, https://www.dummies.com/programming/python/what-is-google-colaboratory/ (December, 2019)

27 Google Cloud, *The Google Cloud Difference - Grow Your Business*, https://cloud.google.com/why-google-cloud/ (December, 2019)

28 AWS, *What is AWS*, https://aws.amazon.com/what-is-aws/ (December, 2019)

29 Schutt, Rachel, O'Neil, Cathy *Doing Data Science*. O'Reilly Media (2013).

30 Pan, Sinno Jialin, and Qiang Yang. *A survey on transfer learning*. IEEE Transactions on knowledge and data engineering 22.10 (2009).

31 AIMultiple, *Transfer Learning in 2020: What it is and How it works*, https://research.aimultiple.com/transfer-learning/ (November, 2020)

32 IBM, *Machine Learning*, https://www.ibm.com/cloud/learn/machine-learning (November, 2020)

33 LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. *Deep learning*. nature 521.7553 (2015).

34 PCMag, *What Is Deep Learning?*, https://www.pcmag.com/news/what-is-deep-learning (November, 2020)

35 Simonyan, Karen, and Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556 (2014).

36 Wikipedia, *Kaggle*, https://en.wikipedia.org/wiki/Kaggle (December, 2020)

37 Galaxy Zoo, *The story so far*, https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/about/results (November, 2020)

38 Srivastava, Nitish, et al. *Dropout: a simple way to prevent neural networks from overfitting*. The journal of machine learning research 15.1 (2014).

39 Claesen, Marc, and Bart De Moor. *Hyperparameter search in machine learning*. arXiv preprint arXiv:1502.02127 (2015).