Creation of Datasets from Open Sources

Ilya V. Chugunkov, Dmitry V. Kabak, Viktor N. Vyunnikov, Roman E. Aslanov
Department of Computer Systems and Technologies
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Moscow, Russia
Ilya.V.Chugunkov@ieee.org

***Abstract*— Machine learning is one of the fastest growing spheres in IT, but it still has some fundamental problems. Before training a neural network, it's necessary to collect a vast dataset of marked entries. However, manual collection of information takes a lot of time and resources. That is why one of the hardest problems to solve in deep learning is the problem of getting the right data with the proper tags. This paper aims at methods that allow to automatically create or update the marked dataset for building a car model classifier by the parser of known Internet sources, which uses a simple classifier to delete incorrect data. The main goal of this article is to prove that public sources can be used to collect the correctly selected and marked data.**

Keywords*— datasets; data preprocessing; neural networks; classifier.

I. Introduction

With the growth of computing power, the rapid development of neural networks began. To date, it is difficult to find an IT area where the idea of machine learning has not yet been applied. The popularity of this technology is caused by many factors: the ability to solve problems with a high degree of complexity without using hard-to-solve algorithms, the speed and accuracy of their solutions and the universality of the areas in which trained AI can be applied. The last point is the reason why the development of neural networks has gained such popularity and not only among scientists and IT giants, but also among small scientific laboratories, educational institutions, companies and even individual developers.

II. Problem Statement

Today neural networks are used to solve a very wide range of tasks: pattern recognition, prediction, approximation, clustering, data analysis. Despite this, the model for constructing the majority of neural networks can be represented by the sequence of the main steps shown in Fig. 1.

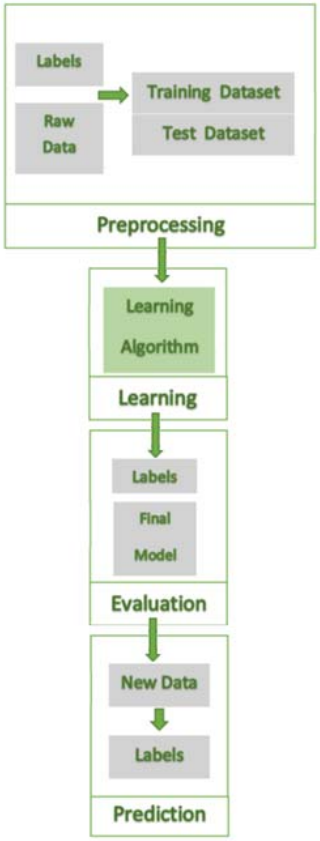

Fig. 1. Machine learning model

Therefore, the developer, who has to creating and training a neural network, first of all encounters the problem of collecting a quality dataset. This problem can be solved by several ways: use ready-made sets, for example from Google or Oxford, order a data set on the corresponding site or collect the necessary data yourself. While the first option is often not suitable for the task being solved, and the second one can be expensive and unreliable, the third one is deprived of these shortcomings. But in this case, its need to not just write an automatic data collection program, but make sure that there are no incorrect data among the dataset.

978-1-5386-4340-2/18/$31.00 ©2018 IEEE

III. Concept of solution

This article describes data parsing for constructing a module for automatic brand and vehicle model identification based on a convolutional neural network using Python. The quality of the received dataset at the initial stage is planned to be verified by implementing a binary classifier for the given models of vehicles.

To create a dataset, it was decided to perform data collection in two stages: directly getting images and filtering them on a pre-trained neural network that classifies, among other things, the objects of the dataset. This way will allow not only to collect a vast array of initial data, but also efficiently and quickly weed out objects that are incorrect for various reasons.

Immediately before the start of parsing, requirements to the source of data were put forward, the correspondence of which makes it possible to assure a maximally differentiated and qualitative initial array:

- full openness of data;
- the presence at least a few hundred images for each prospective class;
- the presence of images of different quality, with different angles and lighting;
- the ability to filter data by many criteria, including year of production and generation of the vehicle.

When choosing the pre-trained neural network used to exclude incorrect images from the dataset, the following requirements were put forward:

- recognition of all classes represented in the dataset;
- high accuracy on the top-1 and top-5 errors;
- normal speed.

IV. Solution

As a site for parsing, the resource auto.ru was chosen, because it satisfies all previously described conditions. To collect the data, a parser was implemented, the main concept of which was to separate the collected data on the model, brand and generation of the car, which was achieved by using the available site functionality.

Since the total volume of the collected data would be several tens of TB, which would make it extremely inconvenient to store and filter it, it was decided to collect data only for Lada cars. This will not affect the final result of the development, since the developed module has the functionality of obtaining all categories, but it will allow to get a solution of the tasks in a shorter time. As a result of parsing, the following results were obtained:

- total number of images: 447825,
- number of classes: 27.

However, due to the fact that the resource was filled by non-professionals, most part of the dataset was expected to be images of poor quality or not corresponding to the subject matter, which made it incorrect for training a neural network.

To filter the data was chosen pre-trained neural network vgg16, because it has an accuracy of 92.7% on a top-5 error in accordance with tests on 14 million images from ImageNet and defines all required types of vehicles.

As the classes for which the binary classifier will be built, images of Lada Grant and Lada Kalina cars were selected. These datasets were filtered pre-trained network and all images that the network treated as not related to vehicles were excluded. As a result, about a third of the data was deleted, as shown in Fig. 2.

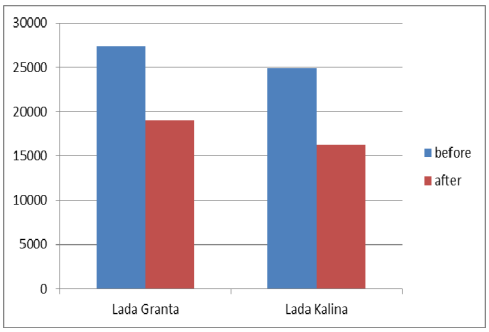

Fig. 2 Exclusion of incorrect data

The remaining objects of dataset are sufficiently qualitative and correctly marked and can be used to train the required neural network.

V. Comparison of results

To verify the correctness of the collected data, a binary classifier for Lada Granta and Lada Kalina was implemented (Fig. 3).

After that a test sample of 100 quality photographs of each model was manually assembled, which verified the accuracy of the prediction of the neural network.

Of the training samples, 2000 images were taken in order to equalize the number of objects supplied by the neural network and to avoid its retraining. In the first case, the classifier was trained on unfiltered data obtained immediately after parsing. The probability of correct predictions was significantly higher than expected and amounted to 82%.

Result:

Epoch 10/10

4000/4000 [==========] - 7931s - loss: 0.5180 - acc: 0.8231

After filtering the data and training the neural network on the final dataset, the accuracy of the prediction increased sharply and reached 95%:

Epoch 10/10

4000/4000 [==========] - 8133s - loss: 0.3814 - acc: 0.9511

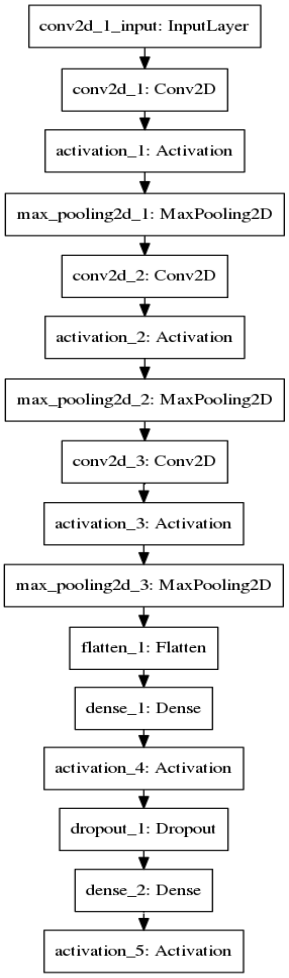

Fig. 3 Binary classifier architecture

VI. Conclusions

As a result of this work a parser was implemented that collected data from a selected open source of the Internet and discarded incorrect images. The quality of the created dataset was verified by implementing a binary classifier that was trained on the collected data. This approach can be reworked and used to collect dataset for other tasks. As a continuation of the work it is supposed to use the obtained data for training the multi-classifier.

Acknowledgment

The authors are sincerely grateful to the head of the Department of Computer Systems and Technologies of the National Research Nuclear University "MEPhI" Professor M.A. Ivanov for help and support during the research. Research of automation of dataset collection was held within the framework of the Program of improving the competitiveness of the National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), as well as in the framework of the priority areas grant program of the Russian Science Foundation “Fundamental and exploratory studies by individual research groups”.

References

[1] A.A. Maksutov, A.V. Simonenko, I.S. Shmakov “Classifiers based on Bayesian neural networks” in *Proc. 2017 IEEE Russia Section Young Researchers in Electrical and Electronic Engineering Conference, ElConRus 2017,* St. Petersburg, Russian Federation, 2017, pp. 700-703. DOI: 10.1109/EIConRus.2017.7910653.

[2] S. Kotsiantis, D. Kanellopoulos, & P. Pintelas, “Data Preprocessing for Supervised Leaning”. International Journal of Computer Science, 2006, vol. 1, Issue No. 2, pp. 111-117.

[3] Bishop C.M. Neural Networks for Pattern Recognition. Oxford University Press; 1995.

[4] Nielsen M. Neural networks and deep learning. Available at: http://neuralnetworksanddeeplearning.com/ (accessed 3 November 2017).

[5] Azadeh, M. Sheikhalishahi, M. Tabesh, A. Negahban “The Effects of Pre-Processing Methods on Forecasting Improvement of Artificial Neural Networks”, *Australian Journal of Basic and Applied Sciences*, 2011, pp. 570-580.

[6] A. Krizhevsky, I. Sutskever, G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, *Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106-1114.

[7] Y. Wen, Z. Li, Y. Qiao, “Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition”, *Cvpr*, 2016, pp. 4893-4901.

[8] L. Jiao, “Advances in natural computation machine learning and image understanding”, *Xidian University Press*, 2008.

[9] L. Kuncheba, “Diversity in multiple classifier systems”, *Information Fusion*, 2005, vol. 6, no. 1, pp. 3-4.