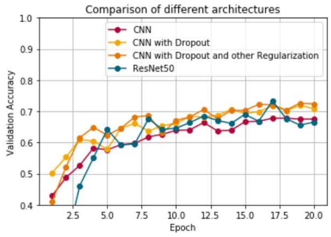
Data vs Algorithms
The importance of large-scale datasets in Machine Learning Systems

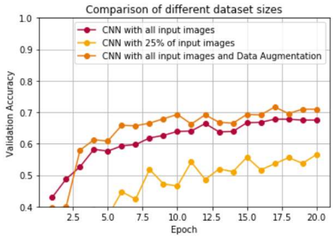
Abstract

Artificial Intelligence is the new buzzword in terms of Computer Science and Digitalization. Monthly, new algorithms and models for deep neural networks are published, outplaying the older techniques in both accuracy and performance. It seems that we can get better and better with less and less data. So will our large-scale data pipelining techniques like Kubernetes, Apache Spark, Airflow, Flink etc. be obsolete in near future for training AI models? We want to discuss the importance of large-scale, balanced and high-quality data in terms of machine learning. We want to show the impact of data to the outcome of any state-of-the-art machine learning project.

1. Amount of data has higher impact on the accuracy of artificial neural networks than the networks architecture

„tf_flowers“ is a dataset of 3650 images from five different kinds of flowers. This dataset was used in the following experiments to show the importance of data. It could be easily replaced by greater datasets like BigEarthNet.





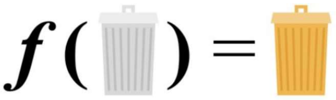
We trained a standard CNN on the dataset, we also added regularization techniques like Dropout, trained on popular architectures like ResNet50 and we compared the final accuracy of each model. As shown in the diagram, Test Accuracies vary between different architectures just a few percent.

We also trained the model with less data and data augmentation. **As shown in the diagram, the amount of data has major impact on the models outcome.**

2. Good data quality and feature engineering is essential when applying data analytics

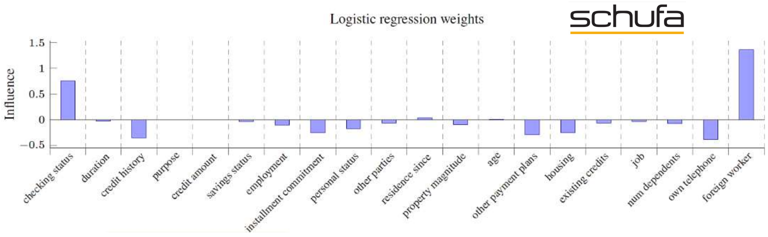
According to [Jenders] from Get Your Guide, data quality has major impact on the results of data analytics projects. He examines that **data with bad quality used to train a ML model cannot produce reasonable results**

Trash-in-Trash-out principle



GET YOUR GUIDE

3. Algorithms cannot address the problem of Bias



Real-world data has often a problem with Bias. [Kasneci] from Schufa points out that Bias invokes also ethical discussions. E.g. as seen in the diagram, a ML model learned to reject credit requests from foreign workers due to Bias in the data.

The problem of Bias cannot be addressed by applying better ML models, because the Bias is in the data and every model learns from the data.

4. Simple models perform better than Transfer Learning on advanced models for some datasets

The success of Transfer Learning depends on how reusable the features from the pretrained model are. In practice, needed features can differ significantly. E.g. Imagenet pictures have a different structure than images from BigEarthNet.

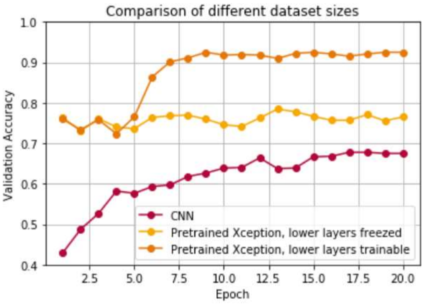
[Demir et.al.] performed fine-tuning of a pretrained Imagenet network to several RS archives, containing 1005-30400 pictures each. They found out that „fine-tuning pretrained models for RS images may not be generally applicable to reduce this semantic gap and therefore may lead to weak discrimination ability for land-cover classes.“

[Demir et.al] also fine-tuned a pretrained (Imagenet) Inception-v2 model on BigEarthNet and compared it to a simple CNN containing only 3 Convolutional Layers. They showed that training a new and simple neural net produced better performance than the Transfer Learning approach

As a result, [Demir et.al.] created a new RS image archive „BigEarthNet“, containing 590,326 images. **They showed the effectiveness of big amounts of data by training a simple CNN. This model achieved a significant higher accuracy comparing to Transfer Learning approaches with smaller archives.**

Table 4: Experimental results obtained by the Inception-v2, the S-CNN-RGB and the S-CNN-All.

Method	P (%)	R (%)	F_1	F_2
Inception-v2 [1]	48.23	56.79	0.4988	0.5301
S-CNN-RGB	65.06	75.57	0.6759	0.7139
S-CNN-All	**69.93**	**77.10**	**0.7098**	**0.7384**



It is surprising that the transfer learning model reached such a bad accuracy. Why is that? The authors just trained the new classifier, freezing the weights of the pretrained CNN layers. We did the same on the tf_flowers dataset, but using Xception instead of Inception-v2. The result is shown as the yellow graph. When fine-tuning the network (train also the pretrained Convolutional Layers), the model reached a significantly higher accuracy with an error rate of 8% instead of 25%.

This shows the importance of data and limits of Transfer Learning. When applying Transfer Learning, the pretrained weights have to get adapted to reach a good accuracy. Therefor, data is needed.

Summary:

This Poster shows the importance of big amounts of data in comparison to the importance of new algorithms. When using Machine Learning, better outcomes are achieved if the researcher concentrates on the amount and quality of data first, instead of immediately fine-tuning artificial neural networks.

Georg Lange

B.Sc. Student
E-Mail: Georg.Lange@student.hpi.de

Hasso Plattner Institute, Potsdam, Germany

References

[Demir et.al.]: Sumbul, G. Charfuelan, M. Demir, B. Markl, V.: BigEarthNet: A Large Scale Benchmark Archive for remote sensing image understanding. 2019
[[Jenders]: Jenders, M. HPI-talk 22.10.19
[Kasneci]: Kasneci, G. HPI-talk 14.01.20

HPI Hasso Plattner Institut