Machine Learning Benchmark to Assess the Environmental Impact of Cars

Valentin Starlinger, Cristina de la Rua Lope, Debarghya Ghoshdastidar
Technical University Munich
Arcisstraße 21
80333 Munich, Germany
{valentin.starlinger, cristina.de-la-rua, d.ghoshdastidar}@tum.de

Abstract

The use of machine learning has enhanced many different aspects of everyday life. However, in some sciences where there is little data publicly available, not much progress in regard to machine learning has been made in recent years. One of these areas is life cycle assessment (LCA), which is used to calculate the environmental impacts of products along their whole life cycle. LCAs are therefore an important cornerstone in decision making in many industries including transportation. The reason for the lack of machine learning is likely due to the fact that there is very little data publicly available for machine learning practitioners to use. In this paper we introduce a new data set that can be used to train machine learning models for the task of predicting the environmental impact of cars given a specific set of input parameters. Furthermore, we show that using machine learning models for predicting life cycle impacts is much faster than traditional approaches and that the data set provided can be used to accurately predict those outcomes using simple models.

We hope that this data set motivates research on LCA in the machine learning community, and new learning algorithms are developed to improve LCA models.

Introduction

As climate change and other environmental concerns are becoming more apparent, reliable data is needed in order to reach good policy decisions as well as to manufacture products that preserve our habitat. Life Cycle Assessment (LCA) is the scientific way of assessing the environmental impacts of a product or service. In LCA, the whole life cycle from raw material acquisition to disposal is modeled (Finnveden et al. 2009). This analysis can be expressed in matrix form (Heijungs and Suh 2002), and can then be solved using solvers for systems of linear equations to get the potential environmental impacts (Mutel 2017).
However, solving these systems can be time consuming especially for complex systems that are often found in transportation. In principle one could use very fast solvers for linear systems, however the system still needs to be solved for every new set of input parameters whereas, if machine learning algorithms are used, once the model is trained the prediction for new input parameters is very fast. The rise of AI and machine learning in recent years has led to their introduction in several fields of study. However, most works published in the field of LCA solely rely on traditional methods with only a few authors utilizing machine learning tools such as (Sousa, Wallace, and Eisenhard 2000; Pascual-González et al. 2015; Liao, Kelley, and Yao 2020; Song 2019; Romeiko et al. 2020).
This probably stems from the fact, that there is hardly any data publicly available that could be used by the machine learning community to build models for this use case.
LCA studies usually share primary data, which refer to the specific system under analysis, while background data are taken from life cycle inventory databases. These databases are mostly proprietary and are therefore usually not published.

Our contribution. To close the gap between LCA practitioners and machine learning researchers, this paper presents a data set that can be used to train machine learning models to predict the results for 21 different environmental impact categories of cars given 76 different input parameters.[1] With a size of 705 thousand entries it can be used to train many different kinds of algorithms. Additionally, the generation process of this data set is shared so that interested researchers can generate their own data.

The availability of this data set can be used to train machine learning algorithms to accurately predict the environmental impacts given a specified parameter set and offers a variety of interesting applications. Two areas where machine learning algorithms can add value compared to traditional approaches are: **(1)** prediction speed once the model is fully trained and **(2)** the possibility to share working models without the need to share the proprietary data.

1. The increase in prediction speed can allow for the use as part of an EcoDesign approach. Using the example of cars, engineers could get live feedback on their material choices when designing new components if those are represented in the parameter set of the model, or companies can get direct feedback when managing their fleet. This can also be valuable for policy makers when elaborating on decisions regarding the environmental impact of cars or other products as it would be easier to look at a wider

[1] The data set together with additional material is available at https://github.com/vstarlinger/lca_datasets.

power train	size	country
ICEV-p	Large	AT, BR, CA,
ICEV-d	Lower medium	CH, CN, DE,
ICEV-g	Medium	ES, GB, IN,
FCEV	Mini	IT, JP, RU,
BEV	SUV	US, ZA
HEV-p	Small	
HEV-d	Van	

Table 1: Possible values of the different nominal parameters of the data set.

spectrum of options. Furthermore, the increase in speed allows researchers to analyse a bigger parameter space (Romeiko et al. 2020).

2. As stated above, the use of machine learning models also allows the possibility to share full working LCA models without the need to provide the raw data. Since the data itself is not being shared, other researchers can use the models in their studies even if they do not have access to the proprietary data sources that were used to build the model. Such a model can also be distributed to non-LCA experts who can use them without the need of specific knowledge of LCAs or the LCA specific software that is used to create the models.

Data Set

The proposed data set is generated using the open source LCA model `carculator` (Sacchi et al. 2020). This model is parameterized using 76 input parameters. For the generation of this data set, these parameters were each sampled randomly from a probability distribution that is derived from either literature research or modeled from a car specification database. From these input parameters, 73 have numeric values while 3 input parameters are nominal. These are the type of power train, the country of use and the size of the car. The possible values for the nominal parameters are shown in Table 1, a full list of parameters is available in the dataset repository.

The data set contains a total of 705600 data points each representing a unique set of input parameter combinations. Each of these combinations has a corresponding impact result calculated by `carculator`. The impact method used by `carculator` is ReCiPe 2008 method for 18 midpoint indicators together with three additional indicators (Goedkoop et al. 2009). Examples of these indicators are climate change, eutrophication or ozone depletion.
These 21 different impact categories are calculated and stored in respect to vehicle kilometers (vkm). In each of these categories, the impact is split between different attribution sources. The list of different impact categories and attributions can be found in appendix A. The sum of the attribution sources corresponds to the total impact in the category for the specific input parameter set. As an example, Figure 1 shows the calculation results for two different impact categories of two separate data points.

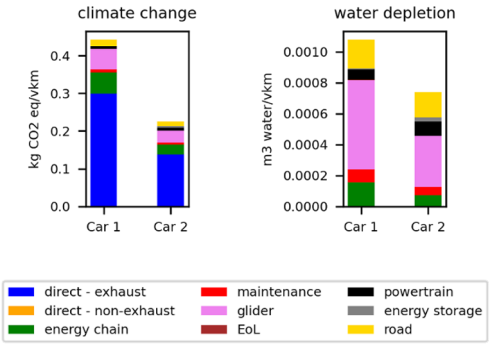

Figure 1: Two different data points (car 1 and car 2) and their respective climate change and water depletion impacts. The different colors show the attribution to the different parts of the life cycle.

	R^2
Linear Regression	0.73
ANN (1 hidden layer)	0.99

Table 2: R^2 value of an ANN trained on the data set as well as for a linear regression model.

Applications

One possible application for this kind of data set is the training of machine learning algorithms for the use in software that need accurate and fast impact results. In order to show the usefulness of the data set, an artificial neural network (ANN) was trained on the data set to predict the climate change impact of cars, given a set of input parameters. The nominal parameters were one-hot encoded and all others were normalized by subtracting the mean and dividing by the standard deviation. The ANN was trained using a single hidden layer with 200 nodes and an output layer with a single node to predict the final regression result. The model was trained using the PyTorch framework and using the Adam optimizer and a mean-squared error loss (Paszke et al. 2019). As an estimate for a baseline performance of very simple machine learning algorithms, the data was also fit using a linear regression model. The results using a train/validation/test split of 0.6/0.2/0.2 can be seen in Table 2. It clearly shows, that the ANN outperforms the linear regression baseline and almost perfectly predicts the test data.

Additionally, the speed of both the linear regression model and the ANN were compared to the calculation of the true result using `carculator`. The comparison was done on a 2,8 GHz Dual-Core Intel Core i5 CPU. The results shown in Table 3 are the average time taken for calculating a single result given a sample size of 1000. As can be clearly seen, the ANN and linear regression outperform the traditional approach using `carculator` by about four orders of magnitude. While linear regression is even faster than the ANN, performance is worse as was shown above.

	Calculation time [s]
Carculator	8.63
Linear regression	0.74×10^{-4}
ANN (1 hidden layer)	2.03×10^{-4}

Table 3: Time in seconds for producing the result given a set of input parameters.

Conclusion and Future Work

This paper introduces a data set for learning a parameterized LCA model for cars. As a proof of concept, it shows the performance of an ANN trained on the data set. The trained model shows that a high accuracy can be achieved for a machine learning model trained on this data set. Additionally this model greatly outperforms the traditional LCA method in speed. These results indicates, that machine learning models trained to learn a parameterized LCA model can be used for the applications described in the introduction of this paper.

In future research, further areas of application can be investigated. Additionally, to expand the data set and provide data for additional areas beyond cars, a framework could be established that takes LCA models as input and automatically generates data that can then be used by machine learning specialists to train models and revise architectures. Furthermore, the introduction of machine learning models into existing LCA workflows can be investigated.

Appendix A

List of attributions

Table 4 shows the list of attributions within impact categories in the data set.

Direct - exhaust
Direct - non-exhaust,
Energy chain
Mainenance,
Glider,
End of Life (EoL),
Powertrain,
Energy storage,
Road

Table 4: Attributions within impact categories.

List of impact categories

Table 5 and Table 6 show the ReCiPe and the additional impact categories with their respective units that are used by `carculator` and represent the results in the data set.

Name	Unit
noise emissions	Person-Pascal.second
renewable primary energy	megajoule
non-renewable primary energy	megajoule

Table 5: Additional impact indicators together with their units used by carculator.

Name	Unit
freshwater ecotoxicity	kg 1,4-DC.
human toxicity	kg 1,4-DC.
marine ecotoxicity	kg 1,4-DB.
terrestrial ecotoxicity	kg 1,4-DC.
metal depletion	kg Fe-Eq.
agricultural land occupation	square meter-year
climate change	kg CO2-Eq.
fossil depletion	kg oil-Eq.
freshwater eutrophication	kg P-Eq.
ionising radiation	kg U235-Eq.
marine eutrophication	kg N-Eq.
natural land transformation	square meter
ozone depletion	kg CFC-11.
particulate matter formation	kg PM10-Eq.
photochemical oxidant formation	kg NMVOC-.
terrestrial acidification	kg SO2-Eq.
urban land occupation	square meter-year
water depletion	m3 water-.

Table 6: Impact indicators together with their units as calculated using the ReCiPe method which is used by carculator.

References

Finnveden, G.; Hauschild, M. Z.; Ekvall, T.; Guinée, J.; Heijungs, R.; Hellweg, S.; Koehler, A.; Pennington, D.; and Suh, S. 2009. Recent developments in life cycle assessment. *Journal of environmental management* 91(1): 1–21.

Goedkoop, M.; Heijungs, R.; Huijbregts, M.; De Schryver, A.; Struijs, J.; and Van Zelm, R. 2009. ReCiPe 2008. *A life cycle impact assessment method which comprises harmonised category indicators at the midpoint and the endpoint level* 1: 1–126.

Heijungs, R.; and Suh, S. 2002. *The computational structure of life cycle assessment*, volume 11. Springer Science & Business Media.

Liao, M.; Kelley, S.; and Yao, Y. 2020. Generating Energy and Greenhouse Gas Inventory Data of Activated Carbon Production Using Machine Learning and Kinetic Based Process Simulation. *ACS Sustainable Chemistry & Engineering* 8(2): 1252–1261.

Mutel, C. 2017. Brightway: An open source framework for Life Cycle Assessment. *Journal of Open Source Software* 2(12): 236.

Pascual-González, J.; Pozo, C.; Guillén-Gosálbez, G.; and Jiménez-Esteller, L. 2015. Combined use of MILP and multi-linear regression to simplify LCA studies. *Computers & Chemical Engineering* 82: 34–43.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-

Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Romeiko, X. X.; Guo, Z.; Pang, Y.; Lee, E. K.; and Zhang, X. 2020. Comparing Machine Learning Approaches for Predicting Spatially Explicit Life Cycle Global Warming and Eutrophication Impacts from Corn Production. *Sustainability* 12(4): 1481.

Sacchi, R.; Bauer, C.; Cox, B.; and Mutel, C. 2020. carculator: an open-source tool for prospective environmental and economic life cycle assessment of vehicles. When, Where and How can battery-electric vehicles help reduce greenhouse gas emissions? (in review). *Submitted to Renewable and Sustainable Energy Reviews* .

Song, R. 2019. *Machine Learning for Addressing Data Deficiencies in Life Cycle Assessment*. Ph.D. thesis, UC Santa Barbara.

Sousa, I.; Wallace, D.; and Eisenhard, J. L. 2000. Approximate Life-Cycle Assessment of Product Concepts Using Learning Systems. *Journal of Industrial Ecology* 4(4): 61–81.