Preparation of synthetic data to be used as inputs for neural network using CAD system.

Jozef Ondriga[1], Jozef Jenis[1], Jakub Fiacan[1], Slavomír Hrcek and Michal Lukac[1]

[1] University of Žilina, Univerzitná 8215/1, 01026 Žilina, Slovakia
jozef.ondriga@fstroj.uniza.sk

Abstract. The aim of the work was to create input data for the neural network. A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. A neural network needs a large amount of data to be properly trained. Creo was used as a tool, which has a generative design tool in it that brings several significant benefits. The output of generative design is usually dozens of different design options, which the designer only must evaluate based on various criteria and choose the most suitable design. Primarily, it is used to reduce weight by approximately 30 to 40% compared to a conventional design. The lower weight not only brings material savings, but above all increased functionality of the components, which can achieve higher speeds and accelerations due to lower momentum, etc. The subject of the generation was the frame of the bicycle. A set of designs were generated that meet the input criteria and then will be used to train the neural network.

Keywords: frame, artificial intelligence, generative design, Creo

1 Introduction

1.1 Motivation

With the development of modern high accuracy and high resolution experimental techniques and measurement tools, the amount of experimental data produced is rapidly increasing. In order to process such voluminous and complex data sets, it is often necessary to use powerful computers and a variety of specialized software tools to gain the required information. This need for computer processing is even more accentuated when a large number of repeated measurements is performed. In such cases, the entire measurement procedure is often automated and controlled by software.

It is good to note, that what is described here for physical (real) experiments, also applies to mathematical models and their outputs. Although mathematical models and numerical (virtual) experiments can provide very useful and detailed insight into many real life phenomena, they still do relay on validation and calibration using physical measurements and experimentally obtained data sets.

An integral part of most experiments and measurements, is the preparation of a final report or protocol. The form and content of such protocol follows the requirements of the customer but often also has to respect the relevant legislation and technical standards. Given the complexity of the whole process, it is advantageous if the

M. Petr et al. (eds.), *62nd International Conference of Machine Design Departments (ICMD 2022)*, Atlantis Highlights in Engineering 28,
https://doi.org/10.2991/978-94-6463-423-5_39

- large, labelled datasets (digital images, social networks, etc.) that could be used to train DNNs.

The GPU allowed the creation of much larger networks with a higher number of layers ("Deep Neural Network") and this combination gave rise to the new paradigm of "Deep Learning". [2] Neural networks need large amounts of data to learn to generalize well and not look for false associations in the data. This paper discusses the possibility of using CAD systems to prepare synthetic data for neural networks.

2 Artificial intelligence and machine learning

The set containing methods and approaches that machines use to acquire useful behavior without having to explicitly pre-program this behavior is called machine learning. The term artificial intelligence is broad, and machine learning is only one subset of it. The field of AI also includes knowledge representation issues (explicit, implicit, vague/sharp), intelligent planning issues, optimization issues, and more. [3]

2.1 The distribution of machine learning

Machine learning has a wide range of applications. Based on the types of tasks, it can be divided into the following areas:

- supervised learning (or supervised learning with a teacher)
- reinforcement learning
- unsupervised learning (learning without a teacher)
- semi-supervised learning

3 Key terms in machine learning theory

3.1 Generalization

One of the key concepts in machine learning theory is the notion of generalization. The ability to generalize is a property of a learning system that determines how and to what extent the system is able - based on knowledge acquired through learning - to respond correctly even to inputs that it has not explicitly learned about. By generalization, then, we mean the ability of a system to extract knowledge from data in such a form that inferences can be made about other data, or about other previously unknown cases. [12] As an example, consider regression: a learning system is given a set of data representing a certain dependency - i.e., certain inputs and their corresponding desired outputs. The system is then tasked with identifying what the dependency is and creating a model that allows the computation of outputs for inputs other than those in the original training set. If the system generalizes correctly, it can also determine these outputs correctly (at least with some reasonable accuracy). If it does not generalize correctly, the results will be erroneous. Another example might be pattern

recognition in the input data - for example, image recognition or speech recognition. Obviously, in a practical application, the input image (or speech recording) will not be identical to the input from the training set - it will often differ significantly from it. Differences may arise due to interference, changes in conditions - such as lighting level, type of camera, in the case of speech recognition, type of microphone, nature of the speaker's voice, mood, environment, etc...

3.2 Overfitting

The phenomenon where a learning system learns to respond accurately to training data but does not generalize is called overfitting. There are certain regularities in the training data that the learning system should be able to identify. However, unless an extremely (perhaps infinitely) large training data set is available, there will be other, random regularities in the data caused by the choice of those training patterns. To generalize well, the system must correctly identify true regularities while ignoring spurious, random regularities. There are two basic challenges associated with the problem of overlearning - how to detect whether overlearning has occurred (or whether the system generalizes correctly) and how to prevent overlearning. [13, 14]

3.3 Validation by splitting

To check how well the system generalizes, we can split the original dataset into two into two parts: training data and testing data. The system learns on the training dataset. After the learning takes place, the properties of the resulting model are verified on the test dataset. The data from the test set is not used in the learning process, so it can be used to test whether the model responds correctly to data that it has not learned from - i.e., whether it generalizes correctly. This type of generalization validation is also referred to somewhere as split validation. A schematic representation of split validation is shown in **Fig. 1**. [3]

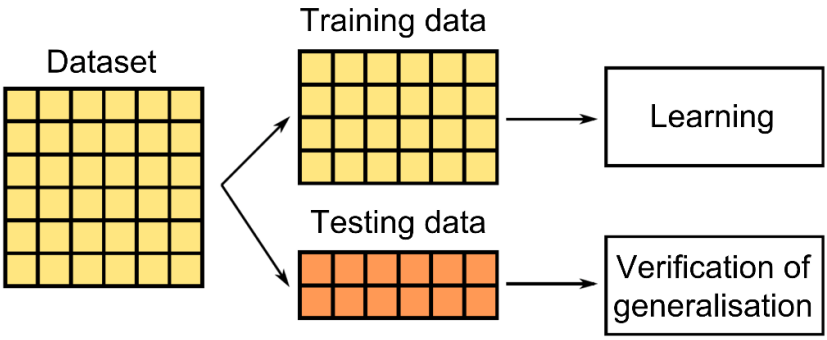

Fig. 1. Split validation scheme. [3]

3.4 How to prevent overestimation/improve generalization

Although there are many regularization methods designed specifically for a particular type of model or learning method, there are also several generic principles that apply in general. From these we can the following:
Get more data: generally, the best way to prevent
overlearning is to simply get more data. (However, this is not always possible.) As we said above, there are both real patterns and random, spurious patterns in the data. With more data, the amount of false, random regularities will understandably decrease, while the real ones will get stronger.
Limit the capacity of the model: If the model adapts too well to spurious regularities (effectively learning the data by rote - as when students memorize answers a, b, c, d from last year's tests), generalization can often be improved by reducing the number of model parameters - so that the model is no longer able to remember all the data, but is forced to identify true patterns in the data.
Use learning by committee: another way that generalization can often be improved is to train a larger number of models and determine the output of the whole system by averaging or voting among them. [3, 15]

Fig. 2 clearly demonstrates that the performance of traditional ML approaches shows better performance for lesser amounts of input data. As the amount of data increases beyond a certain number, the performance of traditional machine learning approaches becomes steady, whereas DL approaches increase with respect to the increment of the amount of data. [7] Therefore, it is necessary to have as large a training dataset as possible. This article discusses the generative design module implemented in Creo 9.0 as one of the suitable tools to create a large and high-quality dataset.

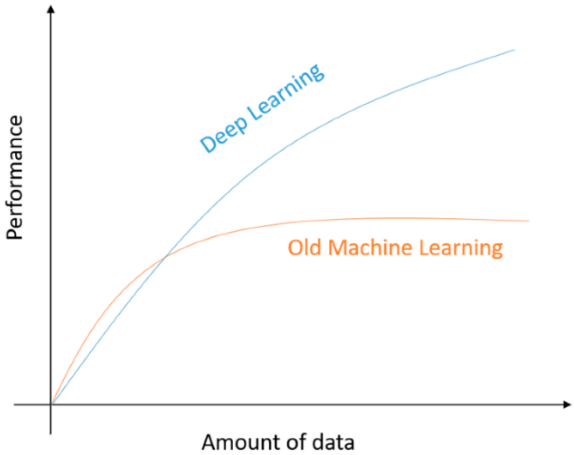

Fig. 2. The performance of deep learning with respect to the amount of data. [7]

4 Data types

In terms of origin, we can divide the data into the following categories:

Real data: Real data is data that contains real values obtained from real life. Examples include ID photos, a database of insurance claims, a file containing material properties, ...

Combined data: Sometimes a small data modification is used to get a larger data set from a smaller one, e.g., images are usually modified for example by rotation, or something is added to them. It is also common to use a basic template, especially for image data, which is then supplemented with additional data (ID card template and only personal data and photos of persons are changed).

Synthetic data: Synthetic data is data that you can create to any extent, whenever and wherever you need it. Synthetic data reflects the balance and composition of real data, making it ideal for supporting machine learning models. [4]
The uniqueness of synthetic data is that data scientists, developers, and engineers are in control. You don't have to rely on unreliable, incomplete data or struggle to find enough data for machine learning at the scale you need. Simply create it yourself. [8, 9]

Advantages of synthetic data. For data scientists, the real or synthetic nature of data is irrelevant. What really matters are the characteristics and patterns inside the data – its quality, balance, and bias. Synthetic data allows you to optimize and enrich your data, unlocking several key benefits.

Increased data quality: Real-world data isn't just hard and expensive to source. It's also prone to errors, inaccuracies and bias that can severely impact the quality of your machine learning model.
With synthetic data generation, you get increased confidence in data quality, variety, and balance. From auto-completing missing values to automated labeling, it's a way to dramatically increase the reliability and accuracy of your data and, in turn, the accuracy of your predictions. [4]

Scalability: Fueling the machine learning economy takes a huge amount of data. Few data scientists can access exactly the data they lack on the scale they need to test and train powerful predictive models. Synthetic data can close that gap. [10]
Many data scientists supplement their real-world records with synthetic data, rapidly scaling up existing data – or just the relevant subsets of this data – to create more meaningful observations and trends. [16]

Powerful simplicity: Finally, synthetic data is refreshingly easy to generate. With real-world data, developers need to:

- Ensure privacy and confidentiality
- Label data in a uniform way
- Filter out duplicate data
- Remove erroneous records
- Collate data from multiple sources, often in multiple formats

With synthetic data, you can control how the resulting data is structured, formatted, and labeled.

5 Preparation of synthetic data

Generative modules of 3D CAD systems are suitable for preparation of synthetic data in terms of quantity and quality. The most well-known CAD systems that offer generative design modules are shown in (**Table 1**). [5, 11]
In terms of connectivity, a distinction can be made between CAD systems with integrated modules for generative design (GD) and software solutions that only offer GD as a stand-alone solution. The stand-alone products offer data exchange with different CAD systems.

Table 1. Selection of CAD-Systems for Generative Design

Software	Provider	Origin	Connectivity	Storage
Fusion 360	Autodesk	USA	Integrated	Cloud
CogniCAD	ParaMatters	USA	Stand-alone	Local
Solid Edge	Siemens	Germany	Integrated	Local
Creo 9.0	PTC	USA	Integrated	Local
MSC Apex	MSC Software	USA	Stand-alone	Local
CATIA V6	Dassault Systemes	France	Integrated	Local
NX	Siemens	Germany	Integrated	Local

Creo 9.0 was selected to create the basic model of the bicycle frame. The preparation of synthetic data can be divided into several steps:

1. Preparation of the basic 3D model - in this case the frame structure of the bicycle (The frame geometry is identical to the cross-country (XC) bicycle CTM Rascal 1.0 model year 2022 size M) (see **Fig. 3**).
2. Use the created model in the generative design module in the following steps (see **Fig. 4**):
   a. Define the boundary conditions for the generative design - determine the constraints and applied loading forces on the frame.
   b. Selecting design goals - There are two options, either trying to maximize stiffness or minimize mass.

c. Specifying design criteria for the generative process - One can select the type of design constraints: build direction, parting line, linear extrusion, planar symmetry, material spreading, minimum crease radius
d. Assignment of the material with mechanical properties from which the frame is to be generated

3. Start the shape optimization process based on the specified criteria. Results can be represented either as tessellated geometry for additive manufacturing or reconstructed as full B-Rep CAD geometry for use in downstream processes.
4. After optimizing the shape, it is possible to view the simulation results (Von Mises stresses, displacement, safety factor), deformation animation (see **Fig. 5**)

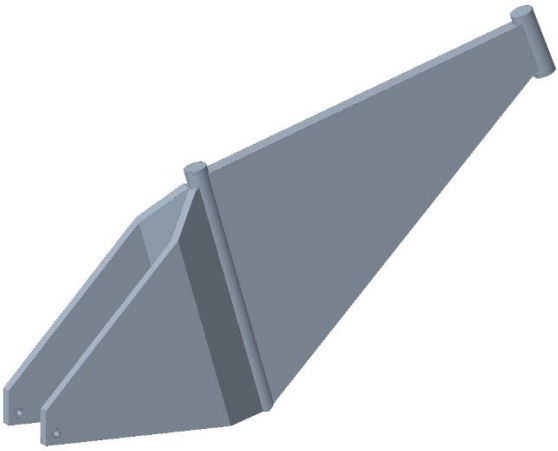

Fig. 3. Base 3D model of a bicycle frame for generative design

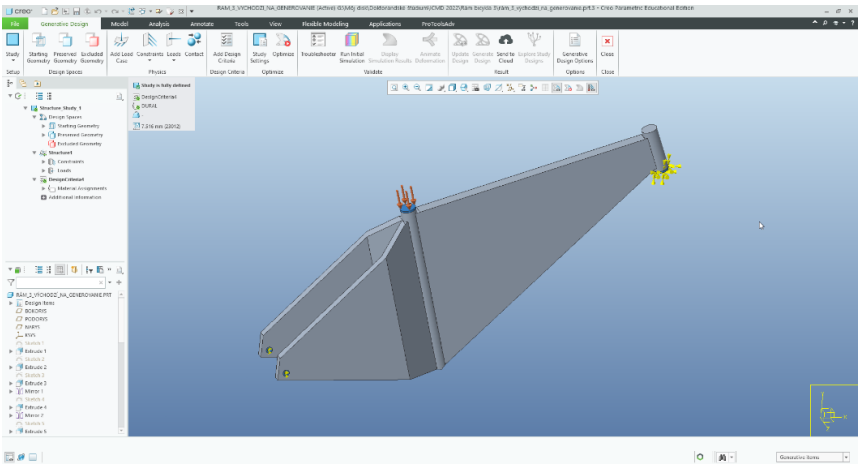

Fig. 4. Base bicycle frame with boundary conditions and load.

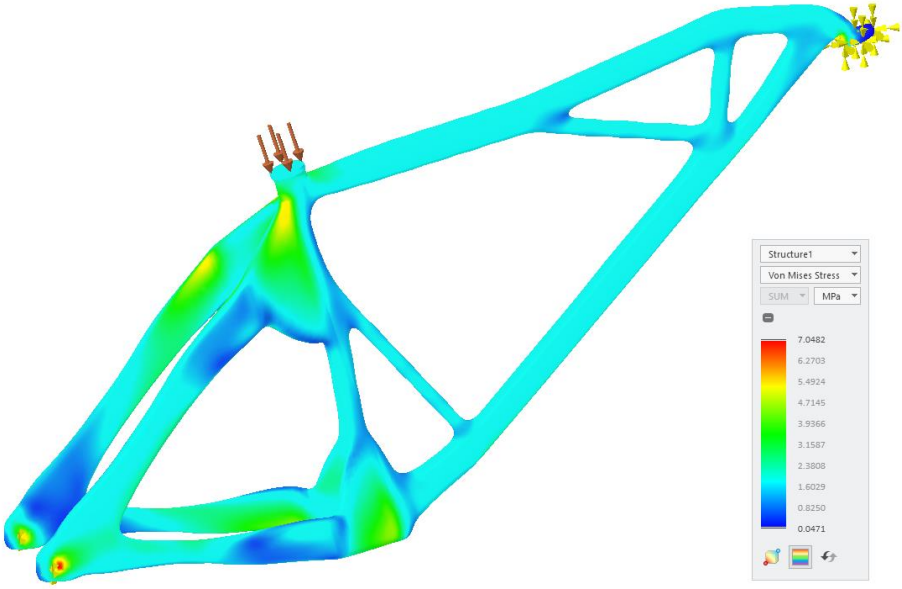

Fig. 5. One of the generated frames with Von Mises stresses shown.

6 Results and discussion

It is indeed often advisable that data repository that is used to build Machine Learning models, should contain data that run into huge numbers. While it is generally true, but a simple Machine Learning model can be built on as few as only a few hundred records. More complex systems, however, do require large datasets and models such as Image Recognition might even need records in millions of records.

The real problem with insufficient data lies in the fact that with less data, variance increases. Variance, which can easily be defined as the variability of model prediction for a given data point or a value which tells us how the data is spread. High variability in models, means that the model will fit the training data perfectly, but will stop working as soon as new data is fed into it. [6]

In this paper, a method for generating synthetic data using generative design to reduce variance has been described. Reducing the variance increases the probability that the neural network will learn to generalize correctly. By changing the boundary conditions, design goals, design criteria and by changing the assigned material, it is possible to influence the resulting shapes and thus generate many different 3D frame models. (**Fig. 6**) The set of generated frame designs can then be used to train neural networks. In the figure you can see examples of frames generated by changing some of the input parameters. [17]

With the proposed procedure, it is possible to prepare any dataset of synthetic data. It can also be applied in the case of complex high-precision systems such as harmonic

transmissions (in the design process), where they can be used as input data for deep learning.

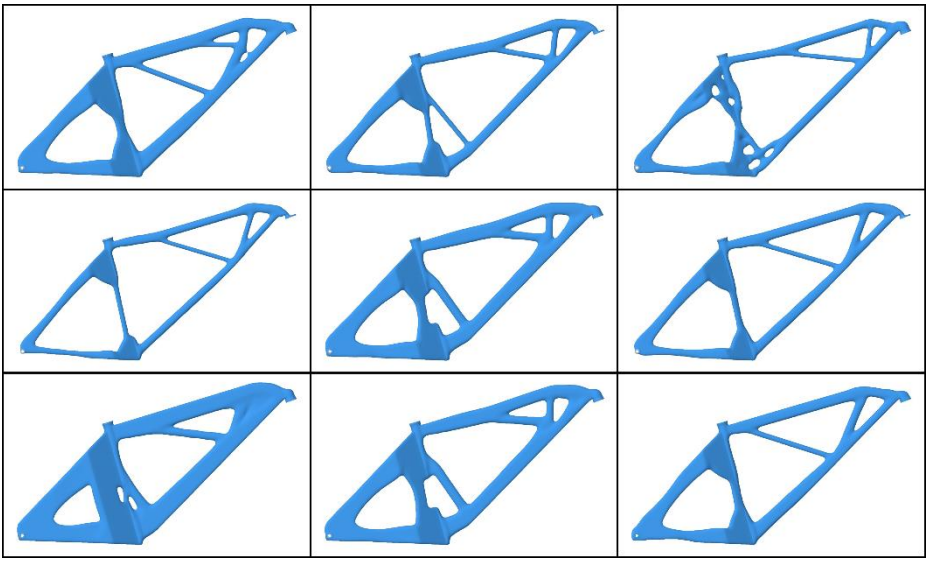

Fig. 6. Example of generated frames that serve as synthetic data.

7 Conclusion

Compared to real data, generating synthetic data is faster, more flexible, and more scalable. Parameter adjustment can also be an efficient way to model and generate data that does not exist in the real world.
Synthetic data allows data scientists to feed machine learning models with data that represents any situation. Synthetic test data can reflect "what if" scenarios, making it an ideal way to test hypotheses or model multiple outcomes.
Synthetic data is a more accurate and scalable replacement for real-world records. Synthetic data gives data scientists a way to do new, innovative things that are impossible with real-world data alone, feeding the models that will affect the way we all live in our data-driven future.
Generative design options are proving useful for the creation of synthetic data. The generated synthetic dataset will be used to train deep neural network in the next step.

8 Acknowledgements

This study was supported by the Slovak Research and Development Agency under contract no. APVV-18-0450 - Research of the influence of design parameters of special transmissions with a high gear ratio with respect to kinematic properties.

This study was supported by Kega Project no. 015ŽU-4/2020 Innovation of the educational process using new CAD technologies.

References

1. Artificial intelligence: definition and applications: https://www.europarl.europa.eu/news/sk/headlines/society/20200827STO85804/umela-inteligencia-definicia-a-vyuzitie
2. Cognition, artificial intelligence, and machine learning: https://www.efocus.sk/kategoria/znalosti-a-inovacie/clanok/poznavanie-umela-inteligencia-a-strojove-ucenie/
3. Michal, G., Dušan, N., Marián, H., Juraj, S.: Artificial intelligence 2. 1st edn. CEIT, Zilina (2018).
4. Datomize - Why Use Synthetic Data vs Real Data? https://www.datomize.com/why-use-synthetic-data-versus-real-data/
5. Junk, S., & Burkart, L. Comparison of CAD systems for generative design for use with additive manufacturing. Procedia CIRP, 100, 577-582. (2021).
6. Towards Data Science - Problems in Machine Learning Models? Check your Data First: https://towardsdatascience.com/problems-in-machine-learning-models-check-your-data-first-f6c2c88c5ec2
7. Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K. A state-of-the-art survey on deep learning theory and architectures. Electronics, 8(3), 292. (2019).
8. Bistak, M., Brumercik, F., & Lukac, M. Weighing systems in traffic. Zeszyty Naukowe. Transport/Politechnika Sląska. (2017).
9. Majchrak, M., Kohar, R., Lukac, M., & Skyba, R. The process of creating a computational 3d model of a harmonic transmission. MM Sci. J, 13, 3926-3931. (2020).
10. Kucera, Ľ., & Gajdosik, T. The vibrodiagnostics of gears. In Modern Methods of Construction Design (pp. 113-118). Springer, Cham. (2014).
11. Buczaj, A., Krzysiak, Z., Pecyna, A., Caban, J., & Brumercik, F. Safety during chemical transport of dangerous goods. Przemysl chemiczny, 98(8), 1276-1280. (2019).
12. Brumercik, F., Lukac, M., Caban, J., Krzysiak, Z., & Glowacz, A. Comparison of selected parameters of a planetary gearbox with involute and convex–concave teeth flank profiles. Applied Sciences, 10(4), 1417. (2020).
13. Mitka, M., Bastovansky, R., Brumercik, F., & Ignaciuk, P. Local resistance of heating molybdenum sheet in a test device. Advances in Science and Technology. Research Journal, 11(3). (2017).
14. Caco, M., Kohar, R., Hrcek, S., Tribula, R., & Scerba, P. Use the method of TRIZ in optimizing automated machine for ultrasonic welding. Procedia engineering, 192, 80-85. (2017).
15. Sarkan, B., Kuranc, A., & Kučera, Ľ. Calculations of exhaust emissions produced by vehicle with petrol engine in urban area. In IOP Conference Series: Materials Science and Engineering (Vol. 710, No. 1, p. 012023). IOP Publishing. (2019, December).
16. Kucera, Ľ., Patin, B., Gajdosik, T., Palencar, R., Palencar, J., & Ujlaky, M. Application of metrological approaches in the design of calibration equipment for verification of float level gauges. Measurement Science Review, 20(5), 230-235. (2020).
17. Tomasikova, M., Tropp, M., Gajdosik, T., Krzywonos, L., & Brumercik, F. Analysis of transport mechatronic system properties. Procedia engineering, 192, 881-886. (2017).