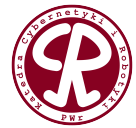
Selected topics in Artificial Intelligence
Data augmentation

Wojciech Domski

Chair of Cybernetics and Robotics,
Wrocław University of Science and Technology

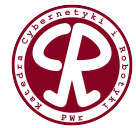
Presentation compiled for taking notes during lecture

Wrocław University
of Science and Technology



Outline

Wrocław University of Science and Technology



Dataset

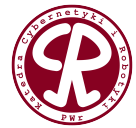
Wrocław University
of Science and Technology

Dataset (1/3)

Learning process is as good as the data it is provide hence the dataset should be:

- representative,
- unbiased,
- unique,
- accurate,
- without noise,
- Big!

Wrocław University of Science and Technology



Dataset (2/3)

Representative

Collected data should represent the problem at hand. It should only contain relevant data while the number of samples contaminating the dataset should be reduced to the minimum.

Unbiased

Collected data should evenly represent the problem space. In case of classification dataset it means that each class should have equal or similar number of samples.

Unique

Samples of the collection should be unique. Repetition of identical samples should be reduced.

Dataset (3/3)

Accurate

Samples should reflect described phenomenon as accurately as possible. Ideally, the samples should not have any noise in the prime dataset.

Without noise

Ideally, the samples should not have any noise in the prime dataset.

BIG

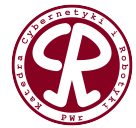
In comparison to other machine learning methods deep neural networks can operate well on vast datasets. The number of the samples should be usually at least 10$\times$ larger than dataset used for conventional methods.

Dataset (1/3)

Learning process is as good as the data it is provide hence the dataset should be:

- representative,
- unbiased,
- unique,
- accurate,
- without noise,
- Big!

Wrocław University of Science and Technology



Dataset (2/3)

Representative

Collected data should represent the problem at hand. It should only contain relevant data while the number of samples contaminating the dataset should be reduced to the minimum.

Unbiased

Collected data should evenly represent the problem space. In case of classification dataset it means that each class should have equal or similar number of samples.

Unique

Samples of the collection should be unique. Repetition of identical samples should be reduced.

Dataset (3/3)

Accurate

Samples should reflect described phenomenon as accurately as possible. Ideally, the samples should not have any noise in the prime dataset.

Without noise

Ideally, the samples should not have any noise in the prime dataset.

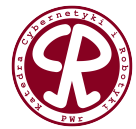
BIG

In comparison to other machine learning methods deep neural networks can operate well on vast datasets. The number of the samples should be usually at least 10$\times$ larger than dataset used for conventional methods.

Ethics and data anonymization

Depending on the process it might happen that the data set is consisting of sensitive data. Therefore, it has to be carefully processed.
Sometimes processing of the data or data collecting itself can rise ethical concerns by e.g. interfering directly or indirectly into private space.
Furthermore, some applications require special preprocessing i.e. data anonymization. Original data has to be reduced in terms of striping of information through which e.g. a person could be identified.

Wrocław University of Science and Technology

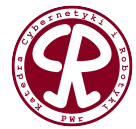


Data acquisition

Depending on the nature of the problem sometimes it is possible to obtain more data than the initial dataset. However, it is important to maintain representativeness of the data set.
In cases it is impossible to collect more data one of two possibilities are available:

- artificially generate the data, e.g. via the means of simulation.
- dataset augmentation – producing new samples based on the already existing dataset.

Wrocław University of Science and Technology



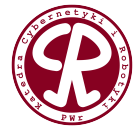
Data processing (1/3)

Each dataset has to be processed. Depending on the type of data some parts of the process can be automated to a great extend.

Gaps

It is not uncommon that the dataset has samples which are corrupted. It can be resolved by data removal or by partial data replacement. Although, depending on with what type of data we are working with neither might be possible depending on the use case.

Wrocław University of Science and Technology



Data processing (2/3)

Quality

Low quality samples are usually discarded since they carry less information or might bias the training process. However, depending on the process itself this type of data might be in demand.

Repetitions

Records which are identical should be removed. This artificially increases the dataset but makes it inaccurate.

Normalization

Normalization of the original data could be introduced to equalize the samples in the dataset. Normalized representation allows to fully use entire representation space of the digital values.

Data processing (3/3)

Saturation

Anomalies in dataset are common. However, if the anomaly is not a part of the process it has to be reduced. This can be either done by clipping values or removing a sample.

Anonymization

Certain parts of dataset ought to be reduced in order to prevent later identification.

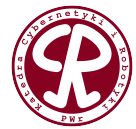
Legal consideration

The source of the original data set should be legal or a consent should be given for the data usage.

Wrocław University of Science and Technology

Data augmentation

Wrocław University
of Science and Technology

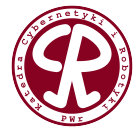


Data augmentation (1/2)

Data augmentation

It is a process of artificially enlarging the initial dataset in order to increase the number of relevant samples while not influencing bias to the dataset.

Wrocław University of Science and Technology

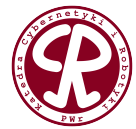


Data augmentation (2/2)

There are multiple number of techniques which allow to enlarge the initial dataset. However, not each method can be utilized for every case. Therefore, it is important to analyse if a certain strategy can be applied to the given problem.
Furthermore, some techniques are dedicated to given class of problems and can not be applied to the other class.
Data augmentation can be used to enhance the data set of:

- numerical data,
- images,
- sound,
- time series,
- ...

Wrocław University of Science and Technology



Original image

Figure: Original image

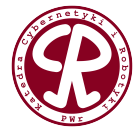
Rotation (1/2)

Figure: Rotation

Rotation (2/2)

Slight rotation, usually within $\pm$ 5° is allowed.

Wrocław University
of Science and Technology



Mirroring (1/2)

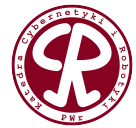
(a) Horizontal

(b) Vertical

Mirroring (2/2)

Horizontal and vertical flipping could be used with object detection, e.g. detecting a ball.

Wrocław University of Science and Technology



Noise (1/2)

(c) Gaussian noise

(d) Shot noise

Noise (2/2)

Random noise allows to create a number of different noised samples based on a original image.

Wrocław University of Science and Technology

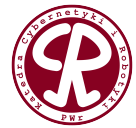


Image transformation (1/2)

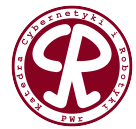
(e) Brightness

(f) Contrast

Image transformation (2/2)

Transforming images in terms of brightness, contrast or saturation (in case of colour images) can create new samples.

Wrocław University of Science and Technology



Contamination (1/2)

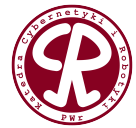
(g) Image placed on image

(h) Image with some artefacts

Contamination (2/2)

Original image can be contaminated with embedding unrelated image. However, this requires additional knowledge to not introduce occlusions.

Wrocław University
of Science and Technology



Zoom in/zoom out (1/2)

(i) Zoom in

(j) Zoom out

Zoom in/zoom out (2/2)

Zooming in or out. While zooming in we need to care about the features we are detecting. In turn, if the image is zoomed out we need to be confident that details of the picture are still recognizable.

Image shift (1/2)

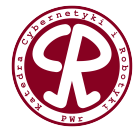
(k) Shifting along one axis

(l) Shifting along two axis

Image shift (2/2)

Shifting an image is similar to zooming in. It may be applied only when features are still visible.

Wrocław University of Science and Technology



Stacking operations (1/2)

(m) Shot noise, rotation and zoom in

(n) Rotation, zoom out and adjusted brightness and contrast

Stacking operations (2/2)

Finally, all different operations can be stacked together. This highly increases number of generated samples.

Wrocław University
of Science and Technology