Topology-based Data Reduction for Green Deep Learning

Javier Perera-Lago[1], Victor Toscano-Duran[2] and Eduardo Paluzo-Hidalgo[3]

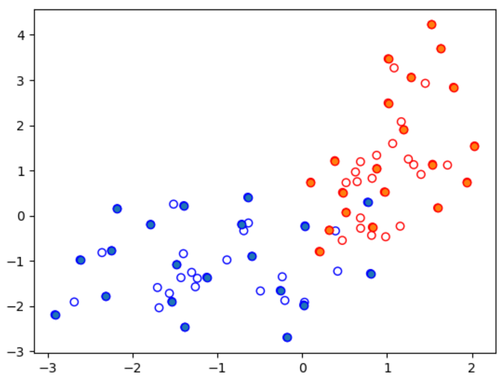
Motivation

Deep Learning has improved its ability to solve complex classification tasks, due to the development of more accurate models, the use of huge volumes of data and the advances in computer capabilities. The downside of these improvements is that the waste of memory, training time and energy resources to create and use a model has increased significantly in recent years. Data reduction is one of the many possible approaches to reduce energy consumption during the training of a deep learning model. It consists of replacing the training dataset with a smaller representative dataset, so that the model can be trained on the reduced dataset obtaining a model with similar predictive ability but less energy-consuming.
In this context, we want to test whether topology can be used to find a good representative dataset. To do this, we must find ways to reduce the dataset, find a way to measure whether the representative dataset preserves the topological properties of the full dataset, and experimentally test whether this measure is a good indicator of the quality of the reduction.

Data Reduction

There exist different methods in the literature to reduce the size of a dataset. According to the reduction strategy, we clasify the methods in the following categories:

- **Statistic-based methods,** which extract a subset either at random or using concepts from statistics and probability theory.
- **Geometry-based methods,** which use the distance matrix of the dataset to perform the reduction.
- **Ranking-based methods,** which order the training examples by some criterion and select the best ranked ones.
- **Wrapper methods,** which extract information from the training process itself and perform the data reduction at the sametime as the training.

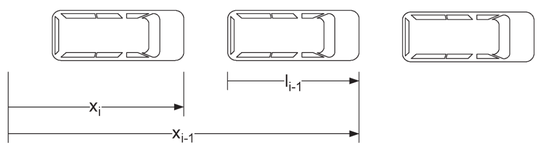


On the left, a dataset with 2 classes (blue/red), where the filled points are those selected by the MaxMin data reduction method. The QR code on the right links to a Python package that contains 8 different data reduction methods.

Collision Dataset

This is the dataset we use in our experiments. Each row of this dataset contains the parameters of a simulated vehicle platoon (number of vehicles, initial distance, initial speed, force, weight...). The task is to predict whether the vehicles will collide or not. The dataset consists of 107,210 examples with 25 numerical features and 2 classes:

- collision = 1, with 69,348 examples.
- collision = 0, with 37,862 examples.



ε-representativeness

ε-representativeness is a measure of similarity between datasets. Given a classification dataset $D = \{(x_i, y_i) | i = 1, ..., N, x_i \in X \subset R^d, y_i = 1, ..., c\}$, a reduced dataset $D_R = \{(x'_j, y'_j) | j = 1, ..., n, x'_j \in X_R \subset R^d, y'_j = 1, ..., c\}$ is ε-representative of D if for each example $(x_i, y_i) \in D$ there exists $(x'_j, y'_j) \in D_R$ with $||x_i - x'_j|| \leq \varepsilon$ and $y_i = y'_j$. The minimum ε such that D_R is ε-representative of D is:

$$\varepsilon^* = \max_{k=1,\cdots,c} \max_{y_i=k} \min_{y'_j=k} ||x_i - x'_j|| \quad (1)$$

ε-representative datasets preserve the persistent homology of the full dataset.

Theorem

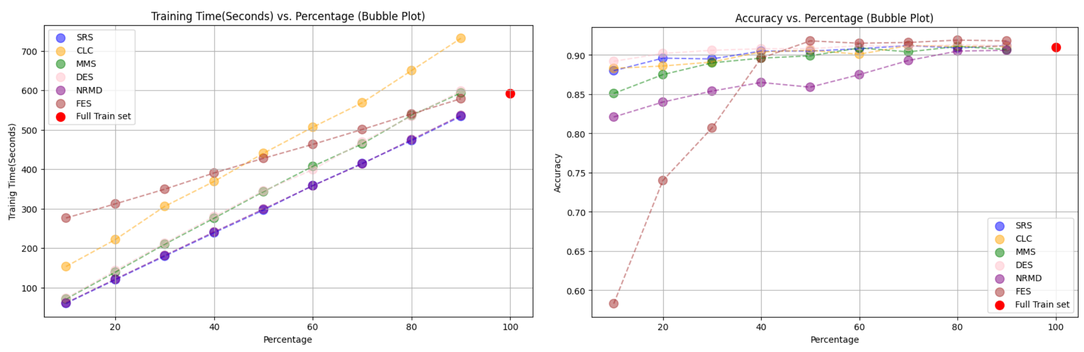
Let D_R be an ε-representative of D. Let $\mathrm{Dgm_q(X)}$ and $\mathrm{Dgm_q(X_R)}$ be the q-dimensional persistence diagrams of X and X_R respectively. Then, for $q \leq d$,

$$\frac{1}{2} d_B(\mathrm{Dgm_q(X)}, \mathrm{Dgm_q(X_R)}) \leq \varepsilon \quad (2)$$

Experiments and results

We performed some experiments to test the relationship between the data reduction method, the ε-representativeness of the reduced dataset and the performance of the trained model. The main steps of the experiments are:

1. Fix a neural architecture.
2. Split the Collision Dataset into training and test datasets.
3. Train the model with the full training dataset.
4. Evaluate the model on the test set.
5. For each data reduction method and each $p \in \{10, 20, ..., 90\}$:
   - Get a reduced dataset with $p\%$ of the training examples.
   - Train the model with the reduced dataset.
   - Evaluate the model on the test set.



The results show us that in general data reduction can help to train deep learning models equally accurate in less time, and then consuming less energy. We also used statistical inference to find a correlation between ε-representativeness and the $F1$-score of the model. The table shows a significant correlation when $p \leq 40$, meaning that reduced datasets with better ε-representativeness train better models.

	10%	20%	30%	40%	50%	60%	70%	80%	90%
Spearman's ρ	-0.38	-0.43	-0.42	-0.39	-0.22	-0.15	-0.19	-0.07	-0.14
p-value	0.0	0.0	0.0	0.0	0.1	0.24	0.14	0.58	0.3

This QR code links to a repository containing the Collision Dataset and the necessary code to reproduce these experiments and get these results.

KEY REFERENCES

[1] Toscano-Durán, V., Perera-Lago, J., Paluzo-Hidalgo, E., Gonzalez-Diaz, R., Gutierrez-Naranjo, M. Á., and Rucco, M. (2024). An In-Depth Analysis of Data Reduction Methods for Sustainable Deep Learning. arXiv preprint arXiv:2403.15150.

[2] Gonzalez-Diaz, R., Gutiérrez-Naranjo, M. A., and Paluzo-Hidalgo, E. (2022). Topology-based representative datasets to reduce neural network training resources. Neural Computing and Applications, 34(17), 14397-14413.

[3] Mongelli, M. (2021). Design of countermeasure to packet falsification in vehicle platooning by explainable artificial intelligence. Computer Communications, 179, 166-174.

ACKNOWLEDGEMENT AND PARTNERS

REXASI PRO

HORIZON-CL4-HUMAN-01 grant agreement no.101070028

Funded by the European Union

TED2021-129438B-I00

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

CONTACT INFORMATION

1. jperera@us.es
2. vtoscano@us.es
3. epaluzo@uloyola.es