Ground_Truth_Data for AI

Towards Sustainable AI

Dossier

Romero and Salmeron
June 24, 2024

EXECUTIVE SUMMARY

1. EXECUTIVE SUMMARY

Lorem ipsum **dolor sit amet**, consectetur adipiscing elit. Aliquam eu nibh non tortor maximus tincidunt. Sed pellentesque lacus a metus pellentesque, vel ultrices leo pulvinar. Ut finibus ipsum ornare, vehicula mauris id, pulvinar enim. Proin in neque elit. Duis eget tempus turpis. Ut consectetur lacinia augue pulvinar bibendum. Nulla sed nulla ut mauris consequat egestas ac et orci.

Integer venenatis, *lectus a dapibus vehicula*, diam odio rutrum ante, at dictum erat lorem et sapien.

Suspendisse potenti :

- pellentesque habitant morbi tristique
- senectus et netus et malesuada fames ac turpis egestas
- ut sit amet ultricies nisl
- etiam laoreet condimentum lacus et elementum

Sed vehicula quis ligula non fermentum. Pellentesque semper ligula gravida "C_i", bibendum dui ut, semper nibh. Quisque a risus porta, fermentum velit vitae, pellentesque felis.

Suspendisse potenti. Aliquam id metus in purus imperdiet aliquam eu ut dolor :

$$
\begin{aligned}
\alpha_{i+1} &= \frac{A_{i+1}}{S_{i+1}} && (1.1)\\
&= \frac{A_i + \Delta A_i}{S_i + \Delta S_i} && (1.2)\\
&= \frac{A_i + \Delta A_i}{S_i + \frac{\Delta A_i}{\alpha_i}} && (1.3)\\
&= \alpha_i * \frac{A_i + \Delta A_i}{\alpha_i * S_i + \Delta A_i} && (1.4)\\
&= \alpha_i && (1.5)
\end{aligned}
$$

Vestibulum luctus consectetur vestibulum. Duis ullamcorper ligula nec mauris viverra rhoncus. Aliquam auctor est accumsan odio vehicula ornare.

> Duis et metus auctor, fringilla nulla non, dapibus felis **??**.

Morbi ipsum dui, rutrum et cursus a, porta eu ligula. Nulla lobortis leo eget odio bibendum rhoncus. Maecenas posuere nulla nec felis viverra blandit.

Sit amet elementum tortor malesuada quis. Quisque pellentesque nisl quis augue interdum vehicula. Nunc nec dictum sem. Nunc venenatis elementum hendrerit.

> Pellentesque ut justo arcu. Fusce cursus luctus tortor eu venenatis.

Maecenas viverra dui vitae eros tristique, porta elementum mi aliquam. Cras sodales erat et molestie auctor. Phasellus mauris eros, bibendum sit amet rhoncus id, bibendum ut velit.

MARKET

2. MARKET

2.1. SUSTAINABILITY

In recent years, we have witnessed an extraordinary acceleration in the growth of artificial intelligence (AI), transforming the way we live, work, and interact with technology. AI algorithms impact sectors such as healthcare, finance, manufacturing, transportation, and entertainment. These advances are driving new chip and server technologies, resulting in extreme rack power densities and presenting new challenges in the design and operation of data centers to meet the massive demand for AI. As part of their Environmental, Social, and Governance (ESG) programs, data center operators are making commitments to environmental sustainability.

Data center operators should use a standard set of metrics. The Green Grid (TGG) proposed power usage effectiveness (PUE) in 2007, which was widely adopted and helped drive efficiency improvements across the industry. A global survey conducted by Uptime Institute in 2023 showed that the average annual PUE of large data centers improved from 2.5 to 1.58 since 2007.

Schneider Electric's growth predictions estimate that AI accounts for 8% of the total power consumption of data centers in 2023, rising to 15-20% by 2028. Given that data centers worldwide consumed over 500 TWh of electricity in 2023 and are projected to consume 815 TWh in 2028 (the equivalent of 16 New York Cities), AI's power consumption equates to 40 TWh in 2023 and is expected to increase to between 122 and 163 TWh by 2028.

2.2. GROUND TRUTH DATA SOLUTION

2.2.1. Approach

At Ground Truth Data, we're tackling energy consumption in data centers with our efficient data reduction technology. This approach not only cuts down on storage space for AI data but also reduces computing times and network loads, boosting overall system efficiency.

Our solution brings a range of benefits: better use of storage resources, cost savings, longer hardware lifespans, and optimized data processing. For AI applications, our techniques improve performance and scalability, allowing for faster training and deployment of AI models, especially deep learning.

Additionally, our technology speeds up backup and recovery processes, which is crucial for the finance and healthcare sectors, and enhances the efficiency of IoT devices in edge computing. For cloud service providers, it helps meet sustainability goals and lowers operational costs.

Using our data reduction strategies leads to significant energy savings, lower operational costs, and greater environmental sustainability, all while maintaining high performance and reliability. Our solution also supports AI development, making AI technologies more accessible and efficient across various industries.

2.2.2. Competitors

bla bla

2.3. RESULTS

Ground Truth Data's solution achieves data reductions of up to 86%, leaving only 14% of the original data. This reduction translates to a 36% savings in computing for AI algorithms on IoT devices and an 86% savings in data centers.

Considering these figures and the optimistic consumption forecasts for data centers in the coming years, we can illustrate the potential energy and cost savings for the industry in the following table:

Year	AI (TWh)	Elect. price ($)	AI Cost (B$)	AI E. saved (TWh)	Cost saved (B$)	CO_2 saved (Mt)
2023	40	0.17	7	34	6	7
2024	54	0.175	9	46	8	10
2025	72	0.18	13	62	11	13
2026	92	0.185	17	79	15	16
2027	115	0.19	22	99	19	20
2028	143	0.195	28	123	24	25
2029	175	0.2	35	151	30	31
2030	212	0.206	44	182	38	38
				776 (TWh)	151 (B$)	160 Mt CO_2

MARKET

TECHNOLOGY

3. DATA REDUCTION

3.1. THE GOAL

The aim of data reduction is to simplify a dataset while preserving its key information. This can typically be accomplished by either reducing the number of features or the number of samples. Our approach will concentrate on the more challenging task of reducing the sample size.

3.2. NEED FOR DATA SIZE REDUCTION

With data being collected at an unprecedented pace, data reduction plays a critical role in boosting training efficiency. By reducing the number of samples, we create a simpler yet representative dataset, which can alleviate memory and computation constraints. This not only enhances sustainability by lowering energy consumption but also contributes to significant energy savings.

3.3. REGULAR METHODS FOR REDUCING SAMPLE SIZE

Sample reduction is typically achieved through instance selection, which involves choosing a representative subset of data samples that retain the original dataset's properties. Existing methods can be categorized into wrapper and filter methods. Filter methods select instances based on scoring functions, such as selecting border instances that often shape the decision boundary. Wrapper methods, on the other hand, select instances based on model performance, considering their interaction with the model. Additionally, instance selection techniques can address data imbalance issues by undersampling the majority class, such as with random undersampling. Recent advancements have incorporated reinforcement learning to optimize undersampling strategies. However, these regular methods neither achieve the data size reduction nor the accuracy level that our data size reduction approach can deliver.

3.4. CHALLENGES

The challenges of data reduction are twofold. First, selecting the most representative data or projecting data into a low-dimensional space with minimal information loss is complex. While learning-based methods can partially address these challenges, they often require substantial computational resources, particularly with very large datasets. Consequently, achieving both high accuracy and efficiency is difficult. Second, data reduction can potentially amplify data bias, raising fairness concerns.

4. VALIDATION TESTS

4.1. INTRODUCTION

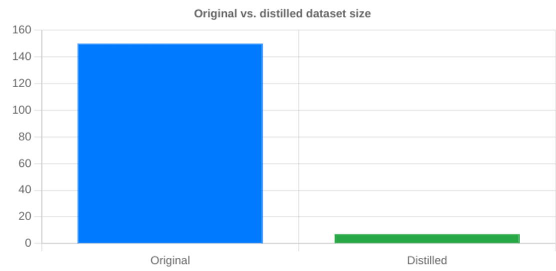
In the realm of data science and machine learning, the efficiency and effectiveness of data processing are often contingent upon the size and manageability of the datasets in use. The burgeoning volume of data necessitates innovative methods for reducing dataset size without compromising the integrity and utility of the data. This chapter focuses on validation tests for a revolutionary data size reduction method, applied to four well-known datasets: the Iris dataset, the Wine dataset, the Breast Cancer dataset, and the MNIST dataset.

4.2. TESTS DONE

4.2.1. Iris Dataset

The Iris dataset, introduced by Ronald A. Fisher in 1936, is a staple in the field of machine learning and statistics. It consists of 150 instances of iris flowers, each described by four features: sepal length, sepal width, petal length, and petal width. These features are used to classify the flowers into three species: Iris-setosa, Iris-versicolor, and Iris-virginica. The simplicity and clarity of the Iris dataset make it an ideal candidate for demonstrating basic principles of data size reduction.

Dataset	Accuracy
Original	97.4%
Distilled	94.7%



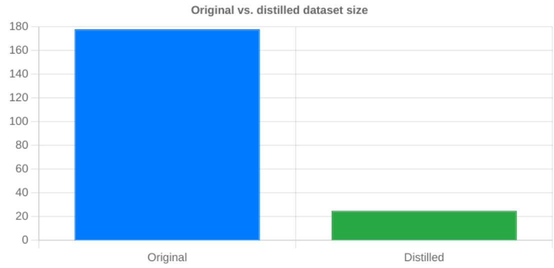
4.2.2. Wine Dataset

The Wine dataset is derived from the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. It contains 178 instances with 13 attributes including alcohol content, malic acid, ash, and others. This dataset is often used for classification problems and serves as an excellent test bed for validating data reduction techniques due to its moderate size and complexity.

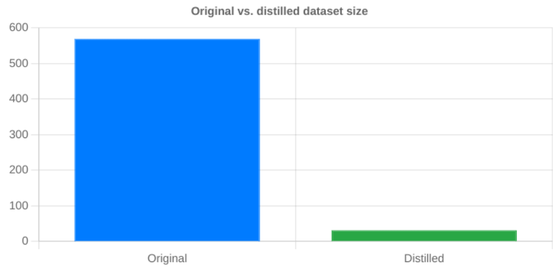
4.2.3. Breast Cancer Dataset

The Breast Cancer Wisconsin dataset, created by Dr. William H. Wolberg, is used for binary classification tasks in predicting the malignancy of breast cancer samples. It comprises 569 instances, each with 30 numeric features representing characteristics of cell nuclei present in a digitized image of a fine needle aspirate of a breast mass. This dataset is critically important for medical research and diagnostics, making the preservation of data quality essential during size reduction.

Dataset	Accuracy
Original	97.8%
Distilled	95.6%



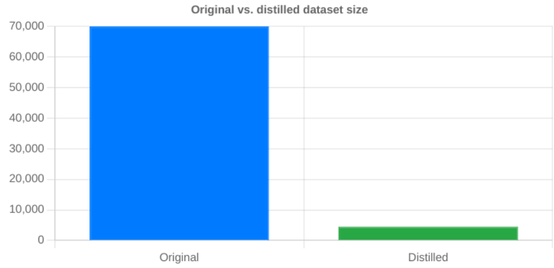
Dataset	Accuracy
Original	96.5%
Distilled	95.1%



4.2.4. MNIST Dataset

The MNIST dataset is a large collection of handwritten digits, commonly used for training various image processing systems. It includes 60,000 training examples and 10,000 testing examples, each represented by a 28x28 grayscale image of a digit (0-9). The high dimensionality and substantial size of the MNIST dataset present significant challenges for data size reduction, thus providing a rigorous test for our proposed method.

Dataset	Accuracy
Original	97.7%
Distilled	93.7%



BUSINESS MODEL

5. HENDRERIT SAPIEN

5.1. SED ULTRICES

5.1.1. Donec pellentesque

Tempus ipsum, vitae condimentum nisi efficitur id. In velit mauris, auctor eget sapien nec, viverra mattis neque. Nunc vel commodo nunc, eget cursus ex.

1. Nunc maximus consequat tristique.
2. Praesent luctus ex aliquam rhoncus consectetur.
3. Suspendisse mattis velit ante, vitae mollis est pharetra a.
4. Duis tincidunt, urna id auctor imperdiet, odio dui ullamcorper nisi.
5. Integer tincidunt enim vitae nulla iaculis, in varius metus blandit.
6. Nunc quam arcu, fermentum non dapibus condimentum, condimentum eget nunc.

5.1.2. Euismod sodales

Removed

5.1.3. Pellentesque a nulla

Removed

5.2. SED ULTRICES

5.2.1. Donec pellentesque

Tempus ipsum, vitae condimentum nisi efficitur id. In velit mauris, auctor eget sapien nec, viverra mattis neque. Nunc vel commodo nunc, eget cursus ex.

1. Nunc maximus consequat tristique.
2. Praesent luctus ex aliquam rhoncus consectetur.
3. Suspendisse mattis velit ante, vitae mollis est pharetra a.
4. Duis tincidunt, urna id auctor imperdiet, odio dui ullamcorper nisi.
5. Integer tincidunt enim vitae nulla iaculis, in varius metus blandit.
6. Nunc quam arcu, fermentum non dapibus condimentum, condimentum eget nunc.

5.2.2. Euismod sodales

Removed

5.2.3. Pellentesque a nulla

Removed

TEAM

6. TECHNOLOGY CREATORS

6.1. MEMBERS

6.1.1. Ivan Romero, PhD

Tempus ipsum, vitae condimentum nisi efficitur id. In velit mauris, auctor eget sapien nec, viverra mattis neque. Nunc vel commodo nunc, eget cursus ex.

Contact: ✉ ivanromeroruiz@gmail.com

6.1.2. Jose L. Salmeron, PhD

Prof. Jose L. Salmeron, Ph.D., Eng., ACM Senior Lifetime member has served as Principal AI & Quantum Scientist at the Hybrid Intelligence division of Capgemini and holds the position of (Catedratico) Professor of Computer Science and Artificial Intelligence at CUNEF University. He has almost 30 years of experience in technology and research, including positions at several universities, consulting in the AI/IT industry, and a wide range of projects with private and public organizations such as Intel, Cisco, Vodafone, Gilead, Microsoft, Airbus, BBVA, and others. He has also been awarded ACM Senior Membership. Prof. Salmeron has consistently been featured in Stanford list of the world's top 2% scientists since the initial version. He has authored or co-authored over 200 scientific articles, conference papers, and book chapters indexed in ISI Web of Science. Furthermore, his work has garnered over 5,900 citations from researchers in 20 different countries, resulting in an h-index of 37 on Google Scholar. His primary field of expertise encompass Distributed Artificial Intelligence, eXplainable Artificial Intelligence, Reservoir Computing, Quantum Computing, Causal Machine Learning, and Federated Learning.

Contact: ✉ joseluis.salmeron@gmail.com

FUNDING

7. FUNDS REQUIREMENTS AND APPLICATIONS

7.1. SED ULTRICES

7.1.1. Donec pellentesque

Tempus ipsum, vitae condimentum nisi efficitur id. In velit mauris, auctor eget sapien nec, viverra mattis neque. Nunc vel commodo nunc, eget cursus ex.

1. Nunc maximus consequat tristique.
2. Praesent luctus ex aliquam rhoncus consectetur.
3. Suspendisse mattis velit ante, vitae mollis est pharetra a.
4. Duis tincidunt, urna id auctor imperdiet, odio dui ullamcorper nisi.
5. Integer tincidunt enim vitae nulla iaculis, in varius metus blandit.
6. Nunc quam arcu, fermentum non dapibus condimentum, condimentum eget nunc.

7.1.2. Euismod sodales

Removed

7.1.3. Pellentesque a nulla

Removed

7.2. SED ULTRICES

7.2.1. Donec pellentesque

Tempus ipsum, vitae condimentum nisi efficitur id. In velit mauris, auctor eget sapien nec, viverra mattis neque. Nunc vel commodo nunc, eget cursus ex.

1. Nunc maximus consequat tristique.
2. Praesent luctus ex aliquam rhoncus consectetur.
3. Suspendisse mattis velit ante, vitae mollis est pharetra a.
4. Duis tincidunt, urna id auctor imperdiet, odio dui ullamcorper nisi.
5. Integer tincidunt enim vitae nulla iaculis, in varius metus blandit.
6. Nunc quam arcu, fermentum non dapibus condimentum, condimentum eget nunc.

7.2.2. Euismod sodales

Removed

7.2.3. Pellentesque a nulla

Removed

APPENDICES

8. VESTIBULUM COMMODO

INTEGER SEMPER DICTUM TELLUS

Neque eu cursus faucibus, ipsum magna tincidunt dui, vel scelerisque urna nibh ut nisl:

Duis sit amet mattis magna.
Curabitur sit amet fermentum mi.
Morbi convallis purus eu fermentum accumsan, mauris felis consequat ipsum.
Ut sollicitudin sit amet tellus et mollis.

A PORTTITOR ORCI FAUCIBUS SIT AMET

Vivamus et sapien vitae lacus ornare suscipit vitae id mauris:

Vivamus in dui arcu.
Morbi vitae dolor libero.
Tellus est, pellentesque at ultrices non, consectetur sit amet augue.
Suspendisse elementum mollis nisl at aliquam.

9. ETIAM FACILISIS

9.1. IPSUM PRIMIS

In faucibus orci luctus et ultrices posuere cubilia curae; Aliquam nunc ipsum, sollicitudin ut tempus consectetur, fermentum at lectus.

Nullam molestie viverra ?? augue sit amet gravida.

9.1.1. Mauris pellentesque

Massa sagittis malesuada

Symbol	**Unit**	**Description**
g	m/s^2	nisi in mollis gravida

9.1.2. Nulla massa

Quis imperdiet

Symbol	**Unit**	**Description**
Q_p	t/h	consectetur tortor

Magna lectus

Symbol	**Unit**	**Description**
τ_a	$°C$	integer mollis bibendum gravida
f_s	$/jour$	vestibulum porta urna at ex dignissim tincidunt

9.1.3. Nam vitae

Nibh a lacus tincidunt efficitur

Symbol	**Unit**	**Description**
X_n	m	venenatis nisi
Y_n	m	integer malesuada eu ligula nec pharetra
Z_n	m	fusce non elit lectus

LIST OF FIGURES

LIST OF TABLES