Data Filtering for a Sustainable Model Training

Francesco Scala[1,2,*], Sergio Flesca[1] and Luigi Pontieri[2]

[1]Dept. Computer Engineering, Modeling, Electronics, and Systems Engineering (DIMES), University of Calabria, 87036 Rende (CS), Italy
[2]Institute of High Performance Computing and Networking (ICAR-CNR), Via P. Bucci, 87036 Rende (CS), Italy

Abstract

The remarkable capabilities of deep neural networks (DNNs) in addressing intricate problems are accompanied by a notable environmental toll. Training these networks demands immense energy consumption, owing to the vast volumes of data needed, the sizeable models employed, and the prolonged training durations. Compounded by the principles of Green-AI, which emphasize reducing the ecological footprint of AI technologies, this poses a pressing concern. In response, we introduce DFSMT, an approach tailored to selecting a subset of labeled data for training, thereby aligning with Green-AI objectives. Our methodology leverages Active Learning (AL) techniques, which systematically identify and select batches of the most informative instances of the data for model training. Through an iterative application of diverse AL strategies, we curate a labeled data subset that preserves adequate information to maintain model quality standards. Empirical results underscore the effectiveness of our approach, demonstrating substantial reductions in labeled data requirements without significantly compromising model performance. This achievement carries particular significance in the context of Green-AI, providing a pathway to mitigate the environmental impact of AI training processes.

Keywords

Active Learning, Green-AI, Data Selection, Energy Efficiency, Sustainability

1. Introduction

Artificial Intelligence (AI) has undergone significant growth in recent years, bringing about transformative changes in various industries and offering innovative solutions to intricate problems. Its impact spans sectors ranging from healthcare and finance to manufacturing and retail, reshaping both our lifestyles and professional environments. Nevertheless, this expansive development has introduced challenges, particularly in terms of increased energy consumption and, consequently, carbon emissions. Moreover, this issue is projected to escalate significantly, as highlighted in [1]. The training phase of AI models, with its substantial demands for data and computing power, is a primary contributor to this energy-intensive process [2]. Effectively training high-performing AI models necessitates vast amounts of data and considerable computing power, resulting in a notable increase in energy consumption. The carbon emissions linked to AI predominantly stem from the electricity utilized during the training phase of these models. Since electricity predominantly originates from non-renewable energy sources, such as coal and natural gas, training AI models significantly contribute to

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy
*Corresponding author.
francesco.scala@icar.cnr.it (F. Scala); sergio.flesca@unical.it (S. Flesca); luigi.pontieri@icar.cnr.it (L. Pontieri)
0009-0007-5224-0910 (F. Scala); 0000-0002-4164-940X (S. Flesca); 0000-0003-4513-0362 (L. Pontieri)

global warming. Indeed, despite advancements, non-renewable energy sources still dominate the majority of the energy production landscape [3].

The aim of reducing the effect on global warming pushed the research community to work on the topic of Green-AI, whose aim is to reduce the environmental impact of AI by promoting the development of efficient and sustainable models and algorithms. Green-AI focuses on several key areas:

- **Reducing energy consumption**: Developing models and algorithms that require less energy for training and use;
- **Using renewable energy**: Powering AI training and use with renewable energy, such as solar and wind power;
- **Developing efficient hardware**: Designing hardware specifically for AI that is more energy efficient;
- **Recycling and reuse**: Promoting the recycling and reuse of hardware components used for AI.

In this paper, we investigate the issue of diminishing energy consumption during the training phase of AI models. Various methodologies have been introduced to tackle this challenge, including MdBR [4] for regression on static data, n-gram counting [5] for machine translation and the enhanced OPF method by Chouvatut et al. [6] which minimizes training set size for classifiers with minimal accuracy loss. Furthermore, clustering techniques have been employed to eliminate irrelevant training samples.

In this work, we investigate the possibility of leveraging Active Learning (AL) [7, 8, 9, 10, 11, 12] to reduce the volume of data required for training AI models, by meticulously selecting the most informative data points within the dataset, and consequently reduce the energy demands of their training phase.

AL techniques are designed to find the most informative data for model training, born out of the recognition that data labeling is one of the most resource-intensive and time-consuming processes in AI model training. AL selects data points for labeling, typically by a human expert annotator, to maximize learning efficiency and minimize the overall data labeling cost. Various approaches have been defined for this purpose. For instance, Least Confidence Sampling (LCS) [8] prioritizes items with the lowest confidence for their predicted label, while LAL-IGrad and its enhancements [10, 11] exploit gradient variation within artificial neural networks to estimate instance relevance. Additionally, Ash et al. [12] proposed BAIT that is a technique for selecting batches of samples by optimizing a bound on the Maximum Likelihood Estimators (MLE) error in terms of the Fisher information.

In this paper we propose DFSMT, a versatile technique that combines various AL methodologies, to actively explore the data space within a pool-based framework, thus identifying the most informative data for the model. AL techniques iteratively select the most informative subset of labeled data to achieve acceptable model quality. To retain efficiency, the emphasis is on computationally lightweight techniques; otherwise, the selection process could become more resource-intensive than training the neural network itself. Experimental results demonstrate that the proposed technique can significantly reduce the amount of labeled data required for

training AI models, while preserving high model quality. This outcome holds particular significance within the perspective of Green-AI, as our technique offers a notable reduction in the environmental impact of AI. It achieves this by significantly lowering the computational cost associated with training AI models. Rather than relying on resource-intensive backpropagation across neural networks, this technique selectively trains on a smaller, optimized dataset obtained by exploiting AL techniques. This drastic reduction in energy and computational power usage aligns with a more environmentally friendly approach to AI model training.

2. Related Work

In recent years, the field of machine learning has witnessed a growing interest in data reduction techniques. This interest is motivated by various needs, including the optimization of computational resources, the reduction of the environmental impact of artificial intelligence (Green-AI), and the improvement of model generalization. In this context, our work falls within the research line that aims to reduce the amount of data required for training machine learning models while maintaining high model quality. Several studies have explored data reduction approaches in different contexts.

For example, the MdBR [4] (Multidimensional binned reduction) method focuses on regression tasks and uses discretization and non-parametric reduction techniques to achieve significant data reduction (over 99%) while maintaining or even improving model performance. However, MdBR is limited to static data and cannot handle time series. In the field of machine translation, Lewis et al. [5] proposed an n-gram counting approach that reduces the size of datasets by up to 90%, without a significant loss of quality (measured by the BLEU score [13]). This method is scalable to large datasets and offers advantages beyond data reduction, such as faster training times and smaller model sizes.

Koggalage et al. [14] proposed a strategy that uses clustering techniques to identify and remove irrelevant training samples that do not affect the decision boundary, this approach allows to reducing the training set size without compromising classification accuracy, but it is specific for SVM. Chouvatut et al. proposed the improved OPF (Optimum-Path Forest) [6] method was developed to reduce the training set size for classifiers. This method is based on a graph-based algorithm and a segmented linear regression approach to achieve a 7-21% reduction in the training set size while maintaining similar accuracy (with a 0.2-0.5% decrease). In some cases, the improved OPF even achieves the exact same accuracy as the original OPF algorithm.

Yang et al. [15] proposed a method called incremental adaptive deep model (IADM) that addresses the challenges of training deep models on streaming data with evolving distributions. It employs an adaptive attention mechanism to adjust model depth and utilizes an attention-based Fisher information matrix to prevent catastrophic forgetting, enabling efficient and accurate learning on incremental data.

Our work differs from previous ones in the following aspects:

- **Combination of different active learning (AL) strategies**: DFSMT uses a combination of AL techniques, potentially offering greater flexibility and adaptability compared to single-strategy approaches;

- **Focus on Green-AI**: Our work explicitly emphasizes environmental impact reduction as a key aspect of data reduction, a unique focus in the current landscape;
- **Potentially broader applicability**: Our approach aims for broader applicability, not limited to a specific task or data type.

By highlighting these strengths and comparing our work to related studies, we can effectively position our research within the current landscape of data reduction techniques and emphasize its potential contributions to Green-AI and other research fields. Our proposal contributes to this line of research by combining different active learning techniques to identify the most informative data points iteratively. This approach has the potential to further reduce the amount of labeled data required for training high-quality AI models, contributing to more efficient and environmentally friendly AI development.

3. Proposed Approach

A classification problem consists in associating every instance taken from a predefined domain $\mathcal{D}$ with a label selected from a fixed domain of labels $\mathcal{L}$. We assume the presence of a set of instance-label pairs $LS \subseteq \mathcal{D} \times \mathcal{L}$, where for each pair $\langle x, y \rangle \in LS$, x is an instance in $\mathcal{D}$ and y is the label associated with x. Algorithm 1 shows the general schema of the proposed approach, named ***D**ata **F**iltering for a **S**ustainable **M**odel **T**raining* and algorithm 2 shows how the selection is performed. DFSMT receives in input the dataset LS, a neural network model NN, the number $epch$ of the training, the number $steps$ of the selection process, p_s the number of relevant instances at start, p_n the number of relevant instances to select at each step and AS the a set of AL techniques. SelectionAlgorithm receives in input LS the instances not already selected in the dataset, k the number of relevant instances to select and AS the a set of AL techniques.

The DFSMT algorithm starts by selecting a number of instances and placing them in the TS for initial training. The model iteratively learns: at each step, p_n additional instances are added to the TS using SelectionAlgorithm that receives as input LS, AS, a set of statistics about the samples needed for AL techniques (which may differ from the techniques themselves), and p_n. During each iteration, the model is updated/-trained with both the new and existing instances. Finally, the trained model is returned.

Algorithm 1: DFSMT

Data: LS: dataset, NN: neural network model, $epch$: number of epochs, $steps$: number of steps, p_s: number of relevant instances at start, p_n: number of relevant instances to select at each step, AS: a set of AL techniques

1 $TS \leftarrow$ SelectionAlgorithm (LS, p_s, AS)
2 Train NN on TS for $epch$ epochs
3 **for** $i = 1 \dots steps$ **do**
4 *stats* $\leftarrow$ *getStats*(LS, *NN*, AS)
5 $TS \leftarrow TS \cup$ SelectionAlgorithm(LS,AS, *stats*,p_n)
6 Train NN on TS for $epch$ epochs
7 **return** NN

The core of the proposed approach is the SelectionAlgorithm, which is responsible for selecting the instances to be used for training. This algorithm combines the active learning techniques present in the AS set. For each instance in LS, the algorithm calculates a relevance score and then combines them. Finally, the $topk$ instances with the highest scores are selected and returned. It is obvious that more techniques in AS, more accurate the selection should be, but at the expense of energy consumption and computation time.

Algorithm 2: SelectionAlgorithm

Data: LS: not selected instances in the dataset, AS: a set of AL techniques, *stats*: A set of data statistics necessary for AS, k: number of relevant instances to select.

$S \leftarrow []$
for $instance \in LS$ **do**
    $score \leftarrow 0$
    for $technique \in AS$ **do**
        $score \leftarrow score + f_{technique}(instance,\ stats_{instance})$
    $S \leftarrow S \cup score$
$topK \leftarrow Selecttop{-}kinstancesfromLSbasedonthescoresinS$
$LS \leftarrow LS \backslash topK$
return $topK$

3.1. Computational reduction

Active learning (AL) offers a pathway to streamline AI model development while aligning with the principles of Green-AI. The core concept lies in the strategic selection of the most informative data samples from a larger labeled dataset. By training on this optimized subset, AL techniques can reduce the overall computational costs associated with reaching a target accuracy level. The potential for energy reduction is directly linked to the following factors:

- **Energy Cost per Data Point**: The hardware used (CPUs, GPUs or TPUs) and the complexity of the neural network architecture dictate the energy expenditure on processing each data point during training. Optimizing algorithms for specific hardware can further reduce this cost;
- **Data Reduction Effectiveness**: A core measure of AL effectiveness is its ability to drastically reduce the training set size while preserving model performance. The greater the reduction achievable, the higher the potential energy savings;
- **AL Complexity**: Active learning techniques range in computational overhead. Simpler methods like uncertainty sampling may have minimal cost, while more sophisticated approaches can introduce higher computation, Indeed using some computationally intensive AL technique may render ineffective the proposed method, because the selection process can become more burdensome wrt the neural network's training;
- **Impact on Training Convergence**: The interaction between data reduction and the model's convergence behavior cannot be ignored. In some cases, a highly informative dataset might lead to fewer training iterations, amplifying savings. However, it's also possible that more iterations might be required to converge, partially offsetting the energy gains.

The significance of energy conservation has long been recognized [16, 17, 18], leading to ongoing advancements in power consumption estimation methodologies. Alongside these theoretical developments, practical tools for building energy consumption modeling have emerged. For the purpose of calculating energy savings, we employed the following formula, established in the work of Lannelongue et al. (2020) [19]:

$$E = t \times (n_c \times P_c \times u_c + n_m \times P_m) \times PUE \times 0.001 \quad (1)$$

Where:

- t: is the running time (hours);
- n_c: the number of cores;
- n_m: the size of memory available (gigabytes);
- u_c: the core usage factor (between 0 and 1);
- P_c: the power draw of a computing core;
- P_m: the power draw of the memory (Watt);
- PUE: is the efficiency coefficient of the data centre.

4. Experimental Evaluation

Data. We used the following dataset to execute the experimental evaluation:

- **MNIST** [20]: which consists of 60000 instances representing 28x28 gray scale images, labeled using 10 mutually exclusive classes, with 6000 images per class. The dataset is organized into 60000 instances as the training set and 10000 instances as the test set. The latter contains exactly 1000 randomly-selected images from each class, while the training set is comprised of five training batches, which contain 6000 images from each class;
- **Fashion-MNIST** [21]: which consists of 60000 instances representing 28x28 gray scale images, labeled using 10 mutually exclusive classes, with 6000 images per class. The dataset is organized into 60000 instances as the training set and 10000 instances as the test set. The author intends Fashion-MNIST to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms. It shares the same image size and structure of training and testing splits.

Baseline methods. We compared the performance of DFSMT with a classical training approach that uses all the data available in the dataset. This allowed us to evaluate how our technique reduces the amount of data required to achieve comparable performance to classical training, measured in terms of model accuracy. As AL technique we utilized the *LCS* technique due to its light weight capabilities. However, this does not preclude the use of other techniques or their combination. More precisely, given an instance x and a classification model θ, the *LCS* method measures the uncertainty of x w.r.t. θ ($\phi(x)$) as $\phi(x) = (1 - P_{\theta(y^*|x)}) \times \frac{m}{m-1}$, where $P_{\theta(y^*|x)}$ denotes the probability that the model θ assigns to the label y^* for the instance x, y^* is the label for which θ yields the maximum probability on x (i.e., $y^* = \arg\max_y P_{\theta(y|x)}$), and m is the

cardinality of the set of labels. Note that the uncertainty function ranges between $[0, 1]$, where 1 is the most uncertain score.

Settings and assessment criteria. To evaluate the effectiveness of DFSMT, we conducted experiments on two standard image datasets just described. For each dataset, we used the following neural networks:

- **MNIST**: A CNN with two convolutional layers (10 and 20 filters, respectively), followed by a dropout layer and two fully connected layers (50 and 10 neurons);
- **Fashion-MNIST**: This CNN architecture starts with two convolutional layers, each using 3x3 filters for local pattern extraction. Batch normalization speeds up training, and ReLU activations provide non-linearity. Max pooling reduces dimensionality. Fully connected layers then interpret the features, with dropout preventing overfitting. The final 10-output layer likely corresponds to a 10-class classification task.

The stochastic gradient descent (SGD) [22] optimization algorithm was used to optimize the model parameters of the neural network for MNIST, chosen due to its efficiency and reliability in a variety of machine learning problems. For Fashion-MNIST, however, the Adam [23] optimization algorithm was selected, potentially due to its faster convergence and adaptability to complex datasets.

For MNIST the negative log-likelihood (*nll_loss*) loss function was used. This function is specific to the multi-class classification. It measures how closely the model predictions align with the ground truth labels. For Fashion-MNIST, which is a multi-class classification problem, as the previous ones, the cross-entropy loss (*CrossEntropyLoss*) function was used. This function measures the distance between two probability distributions and has been shown to be effective for classification problems with a high number of classes.

Classical training involves using the entire dataset to train the model in a single phase, doing 100 training epochs. This approach can be computationally expensive and require significant training time, especially for large datasets and models. Incremental training, on the other hand, adopts an iterative approach. Initially, a small subset of the dataset is used to train the model (1000 samples), subsequently, the model is updated incrementally with new data acquired iteratively (1000 samples) per 10 incremental steps, in which are performed 10 training epochs. This approach can significantly reduce the training time, energy consumption and the amount of data required, while maintaining high model accuracy.

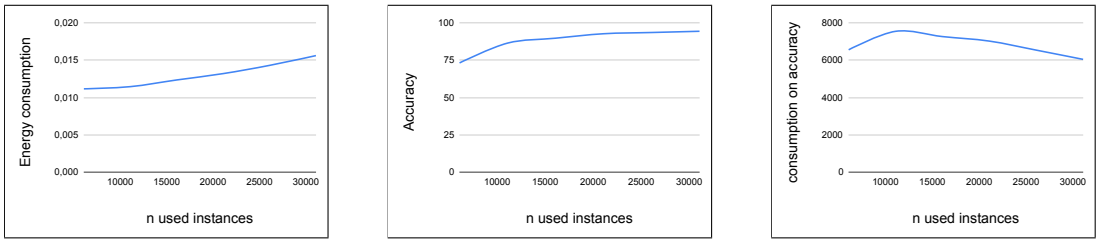
We analyzed how the behavior of DFSMT changes when varying the amount of data selected at each training step with the MNIST dataset [1]. Table 1 summarizes our analysis and figure 1 shows them. It includes the amount of data selected at each step of the process, the final amount of data used at the end of training, the model's accuracy, average CPU utilization (note that values exceeding 100% indicate multi-core usage), processing time in milliseconds, energy consumption (expressed in kWh) calculated using equation 1, and a metric relating accuracy to energy efficiency (efficiency ratio) calculated as $accuracy/energy$.

Then we analyzed the accuracy and loss curves during both classical and incremental training. This allowed us to monitor the model's learning in both cases, comparing its evolution with

[1]Experiments were carried out on an Intel Core i5 CPU @2.30GHz 8259U, 8GB RAM, with Intel Iris Plus Graphics 655 GPU

instances per step	tot. instances	accuracy	CPU	time	energy	efficiency ratio
500	6000	73.18%	108%	512463	0.011	6558.32
1000	11000	86.47%	112%	519593	0.011	7550.01
1500	16000	89.76%	112%	559421	0.012	7268.01
2000	21000	92.75%	114%	594989	0.013	7025.61
2500	26000	93.58%	114%	644746	0.014	6541.62
3000	31000	94.41%	112%	708402	0.016	6045.24

Table 1
This table shows the relationship between the number of samples selected per training step and corresponding changes in computer usage parameters.



(a) energy vs instances (b) accuracy vs instances (c) effectiveness

Figure 1: Graph (a) demonstrates how energy requirements increase alongside the number of instances, graph (b) similarly illustrates the rise in accuracy with instance quantity, graph (c) visualizes the relationship between these two trends.

full and reduced data sets. Accuracy is the primary metric for evaluating a model's ability to correctly classify images. The loss measures the model's error in predicting labels. By monitoring the loss during training, we can evaluate the model's ability to learn from the data and improve its predictions.

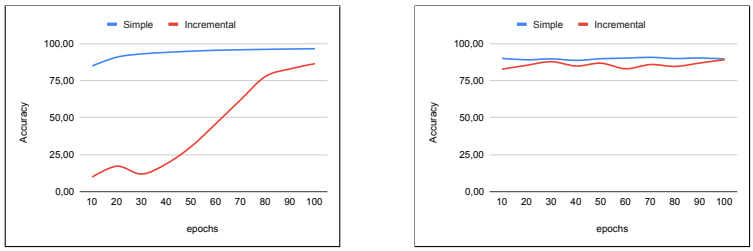
Results. The analysis focuses on three key aspects: computational savings, accuracy and loss, comparing the performance of DFSMT with classical training on two datasets of varying complexity: MNIST and Fashion-MNIST. As observed in Table 1, increasing the number of training instances naturally leads to higher accuracy and energy consumption. Our experiments aimed to identify the optimal parameters for maximizing the accuracy-energy consumption relationship. We determined that the *"n instances per step"* parameter is the primary influencing factor, with 1000 instances yielding the best results. Consequently, we used this parameter for our comparative analysis against classical training. While classical training achieved slightly higher accuracy (96.58% vs. 94.41%), its energy consumption was significantly greater (0.027 kWh vs. 0.011 kWh). This translates to a superior efficiency ratio of DFSMT of 7550.01 compared to 3642.04 with classical training. Figure 1 clearly demonstrates the differing growth patterns of accuracy and energy consumption. While accuracy increases logarithmically, energy consumption follows a different trajectory. This highlights the inherent trade-off between these two metrics, emphasizing the need to carefully select parameters for the most efficient model training.

DFSMT demonstrated remarkable potential on the Fashion-MNIST dataset. It achieved a

significantly higher efficiency ratio (3737.75 vs. 663.19 with classical training) and drastically reduced energy consumption (0.024 kWh vs. 0.134 kWh) while maintaining comparable accuracy (89.21% vs. 89.62%). These results, obtained under identical MNIST settings, underscore DFSMT's advantages. By comparing the accuracy trends during classical and incremental training, we observed:

- **Classical Training**: Accuracy increased gradually with the number of epochs, reaching a plateau towards the end of training;
- **Incremental Training**: DFSMT exhibits a faster learning rate (i.e., steeper upward trajectory) than classical training on MNIST as the number of training examples increases. On Fashion-MNIST, this difference is less pronounced.

Our analysis of accuracy and loss validates DFSMT's ability to reduce energy consumption in machine learning training. Even with less data, incremental training achieved comparable accuracy to classical training, demonstrating its potential as a more efficient and sustainable approach.



(a) mnist (b) fashion-mnist

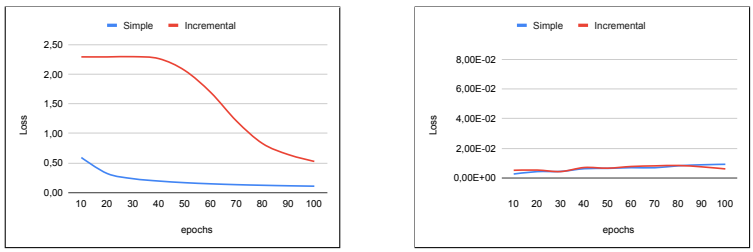
Figure 2: These two graphs (one for each dataset used) compare the accuracy between classical training and our proposed incremental approach at the varying of the training epochs.

Our analysis reveals that classical training converges to the optimum faster than DFSMT, as evidenced by both loss curve and accuracy trends. While DFSMT's loss curve initially shows slightly less stability due to less training data, it eventually stabilizes as the number of training instances increases.

DFSMT stands out for its significant computational savings compared to classical training. The advantage becomes more pronounced with increasing dataset's instances size.

5. Conclusion

Based on the conducted analysis, we can confidently state that DFSMT represents an efficient and performant machine learning method for handling large datasets. The algorithm offers significant computational savings compared to classical training, without notable sacrificing model accuracy. The computational efficiency of DFSMT makes it a promising solution for machine learning on resource-constrained devices, and also in the context of Green AI, which is becoming increasingly important due to the climate crisis. Moreover, its ability to handle



(a) mnist (b) fashion-mnist

Figure 3: These two graphs (one for each dataset used) compare the loss between classical training and our proposed incremental approach at the varying of the training epochs.

large datasets opens up new possibilities for the use of machine learning models in a variety of applications, with a positive impact on the efficiency and sustainability of such systems. At the led of these results we continue the research in this direction making some improvements to DFSMT exploiting for example the information supplied from the dataset as the label (in contrast of a simple AL setting) and applying some optimizations to the selected data in order to keep the dataset balanced. Building upon these findings, our future research endeavors will focus on refining DFSMT by leveraging dataset-specific information such as the label of the instances, diverging from simple active learning settings, and implementing optimizations to maintain dataset balanced. These enhancements aim to further elevate the performance and versatility of DFSMT, fostering its broader adoption across diverse domains and reinforcing its role in advancing both efficiency and sustainability in machine learning practices.

Acknowledgement

This work was partly supported by project FAIR - Future AI Research - Spoke 9 (Directorial Decree no. 1243, August 2nd, 2022; PE 0000013; CUP B53C22003630006), under the NRRP (National Recovery and Resilience Plan) MUR program (Mission 4, Component 2 Investment 1.3) funded by the European Union – NextGenerationEU.

References

[1] A. de Vries, The growing energy footprint of artificial intelligence, Joule 7 (2023) 2191–2194. URL: https://www.sciencedirect.com/science/article/pii/S2542435123003653. doi:https://doi.org/10.1016/j.joule.2023.09.004.

[2] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 3645–3650. URL: https://doi.org/10.18653/v1/p19-1355. doi:10.18653/V1/P19-1355.

[3] S. Flesca, F. Scala, E. Vocaturo, F. Zumpano, On forecasting non-renewable energy pro-

duction with uncertainty quantification: a case study of the italian energy market, Expert Systems with Applications 200 (2022). URL: http://www.sciencedirect.com/science/article/pii/S0957417422003670. doi:http://doi.org/10.1016/j.eswa.2022.116936.

[4] J. Wibbeke, P. Teimourzadeh Baboli, S. Rohjans, Optimal data reduction of training data in machine learning-based modelling: A multidimensional bin packing approach, Energies 15 (2022). URL: https://www.mdpi.com/1996-1073/15/9/3092. doi:10.3390/en15093092.

[5] W. Lewis, S. Eetemadi, Dramatically reducing training data size through vocabulary saturation, in: Proceedings of the Eighth Workshop on Statistical Machine Translation, WMT@ACL 2013, August 8-9, 2013, Sofia, Bulgaria, The Association for Computer Linguistics, 2013, pp. 281–291. URL: https://aclanthology.org/W13-2235/.

[6] V. Chouvatut, W. Jindaluang, E. Boonchieng, Training set size reduction in large dataset problems, in: 2015 International Computer Science and Engineering Conference (ICSEC), 2015, pp. 1–5. doi:10.1109/ICSEC.2015.7401435.

[7] B. Settles, Active Learning Literature Survey, Technical Report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[8] B. Settles, M. Craven, An analysis of active learning strategies for sequence labeling tasks, in: Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 1070–1079.

[9] S. Kee, E. del Castillo, G. Runger, Query-by-committee improvement with diversity and density in batch active learning, Information Sciences 454-455 (2018) 401–418. URL: https://www.sciencedirect.com/science/article/pii/S0020025518303700. doi:https://doi.org/10.1016/j.ins.2018.05.014.

[10] S. Flesca, D. Mandaglio, F. Scala, A. Tagarelli, A meta-active learning approach exploiting instance importance, Expert Systems with Applications 247 (2024) 123320. URL: https://www.sciencedirect.com/science/article/pii/S0957417424001854. doi:https://doi.org/10.1016/j.eswa.2024.123320.

[11] S. Flesca, D. Mandaglio, F. Scala, A. Tagarelli, Learning to active learn by gradient variation based on instance importance, in: 2022 26th International Conference on Pattern Recognition (ICPR), 2022, pp. 2224–2230. doi:10.1109/ICPR56361.2022.9956039.

[12] J. T. Ash, S. Goel, A. Krishnamurthy, S. M. Kakade, Gone fishing: Neural active learning with fisher embeddings, in: M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 8927–8939. URL: https://proceedings.neurips.cc/paper/2021/hash/4afe044911ed2c247005912512ace23b-Abstract.html.

[13] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[14] R. Koggalage, S. K. Halgamuge, Reducing the number of training samples for fast support vector machine classification, 2004. URL: https://api.semanticscholar.org/CorpusID:6688904.

[15] Y. Yang, D.-W. Zhou, D.-C. Zhan, H. Xiong, Y. Jiang, Adaptive deep models for incremental learning: Considering capacity scalability and sustainability, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD

'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 74–82. URL: https://doi.org/10.1145/3292500.3330865. doi:10.1145/3292500.3330865.

[16] E. García-Martín, C. F. Rodrigues, G. Riley, H. Grahn, Estimation of energy consumption in machine learning, Journal of Parallel and Distributed Computing 134 (2019) 75–88. URL: https://www.sciencedirect.com/science/article/pii/S0743731518308773. doi:https://doi.org/10.1016/j.jpdc.2019.07.007.

[17] D. A. Patterson, J. Gonzalez, Q. V. Le, C. Liang, L.-M. Munguía, D. Rothchild, D. R. So, M. Texier, J. Dean, Carbon emissions and large neural network training, ArXiv abs/2104.10350 (2021). URL: https://api.semanticscholar.org/CorpusID:233324338.

[18] J. Xu, W. Zhou, Z. Fu, H. Zhou, L. Li, A survey on green deep learning, ArXiv abs/2111.05193 (2021). URL: https://api.semanticscholar.org/CorpusID:243861089.

[19] L. Lannelongue, J. Grealey, M. Inouye, Green algorithms: Quantifying the carbon footprint of computation, Advanced Science 8 (2021) 2100707.

[20] L. Deng, The mnist database of handwritten digit images for machine learning research, IEEE Signal Processing Magazine 29 (2012) 141–142.

[21] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, CoRR abs/1708.07747 (2017). URL: http://arxiv.org/abs/1708.07747. arXiv:1708.07747.

[22] S. Ruder, An overview of gradient descent optimization algorithms, 2017. arXiv:1609.04747.

[23] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. arXiv:1412.6980.