Atomic Spectroscopy
Article

An Examination Of The Use Of Deep Learning For Identifying Data Curation Activities: A Comprehensive Analysis

Deepesh, Renu
Dronacharya College of Engineering
Assistant Professor, Dronacharya College of Engineering

ABSTRACT: In the field of data science and machine learning, data plays a crucial role as the fundamental resource that drives the algorithms, models, and insights that underpin contemporary technology. Nevertheless, it is imperative to acknowledge that the quality and trustworthiness of data cannot be assumed. This is the point at which the practice of data curation becomes relevant. Data curation refers to the systematic procedure of gathering, refining, arranging, and up keeping datasets to guarantee their precision, comprehensiveness, and suitability for analysis. While it may not possess the same level of allure as the training of sophisticated deep learning models, it is unquestionably one of the most crucial stages in the data science workflow. The primary aim of this study is to provide a thorough examination of the application of deep learning techniques in the identification of data curation activities. This study utilizes a qualitative research methodology. Through this study, Data Curation is a persistent issue that requires innovative solutions to effectively manage the growing big data environment. Deep Learning is increasingly being adopted in several fields, encompassing both computer science and other domains. The convergence of these two fields will initiate a sequence of research endeavours resulting in practical solutions for numerous Data Curation challenges.

Keywords: Data Curation; Deep Learning; Data

INTRODUCTION

Data Curation (DC) refers to the systematic procedure of identifying, consolidating, and refining data in order to facilitate further analytical activities [1], [2]. This process has significant importance for organizations as it enables them to derive tangible commercial benefits from their data. The database community has intensively examined the following significant difficulties within the realm of data curation [3]:

- The process of data discovery involves the identification of pertinent data for a particular task.
- Schema matching and schema mapping involve the identification of similar columns and the understanding of the transformation between these matched columns. Entity resolution pertains to the identification of pairs of tuples that represent the same entity [4].
- Lastly, data cleaning encompasses the identification of errors in the data and the potential repair of such errors.

In addition to issues pertaining to outlier detection, data imputation, and data dependencies, other challenges arise. The Figure 1 presented below provides a comprehensive depiction of the data curation pipeline.

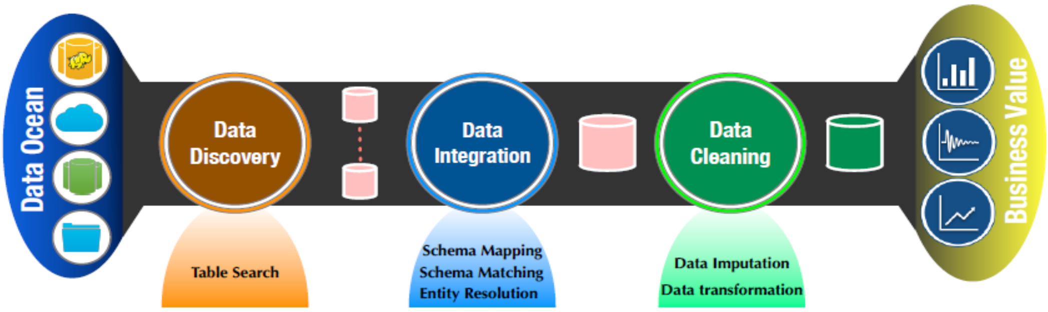

Figure 1: Data Curation pipeline [1]

The paradigm of Deep Learning (DL) has demonstrated considerable success in various domains of machine learning (ML), including but not limited to computer vision, natural language processing, speech recognition, genomics, and other related fields. DL has achieved exceptional performance in various domains because to the combination of massive data, improved algorithms, and increased computing power. As a result of these achievements, there has been a significant amount of recent research aimed at utilizing DL in various domains, both inside and beyond the field of computer science.

Currently, there is no DL architecture that takes into account the specific features of DL tasks, such as the representation of tuples or columns, integrity restrictions, and other related factors. Applying existing DL architectures without careful consideration may be effective for certain DC problems, but not for all. Alternatively, developing novel DL architectures specifically designed for DL tasks that are aware of the unique properties of these tasks would enable efficient learning in terms of both training time and the amount of training data needed.

The primary objective of this work is to analyse the application of DL in identifying operations related to data curation.

LITERATURE REVIEW:

The subsequent table provides a comprehensive overview of previous scholarly works pertaining to the function of DL in data curation applications.

Table 1: Related works

AUTHORS AND YEARS	METHODOLOGY	FINDINGS
Thirumuruganathan et al., (2018) [6]	This study examined how DL breakthroughs could improve and create new data curation solutions.	This report outlined the present DL landscape, identified promising research areas, and dispelled myths. The synthesis of these essential topics should spark a succession of research initiatives that will improve various data curation jobs.
Marciano et al., (2019) [7]	Explained the benefits of computational thinking (CT)-based digital curation.	Using a case study of computational treatments of World War II Japanese-American Incarceration Camp Records, this study showed how CT can detect personally identifiable information, develop name registries, integrate vital records, design controlled vocabularies, map events and people, and connect events and people through networks.
Zhang & Gao (2021) [8]	This research reviewed data curation approaches such data denoising, outlier identification, imputation, balancing, and semantic annotation that have successfully extracted information from noisy, incomplete, insufficient, and/or unannotated data.	Model interpretation approaches that address DL's "black-box" characteristic for model transparency are highlighted.
Cappuzzo (2022) [9]	This study created models to create distributed representations of mixed data kinds by translating relational tables into graphs. ED methods use DL to create embeddings for target applications using tabular data.	UseD attention to train GRIMP and another ML imputation algorithm on related qualities, demonstrating improved quality outcomes after considering functional dependencies.
Waskom et a., (2023) [10]	Approach uses model output confidence scores to decide whether to use manual abstraction.	Hybrid curation achieves expert-level mistake rates with minimal manual abstraction. Showed that hybrid variable research cohorts had similar demographics to conventional cohorts.

Research Gap: Based on previous research, the significance of data curation has prompted numerous commercial and academic endeavours including various facets of data curation, such as data exploration, data integration, and data cleansing. The majority of DC solutions are not completely automated, as they are frequently improvised and necessitate significant exertion to produce elements, such as characteristics and annotated data, that are utilized to create rules and train machine learning models. Practitioners require pragmatic and implementable solutions that have the potential to substantially mitigate the human cost. So, the main aim of this research is to conduct a study on the role of DL models in data curation.

METHODOLOGY:

The research methodology utilized in this study was qualitative in nature. Secondary data collecting technique was employed for the purpose of data collection. The process of collecting secondary data focused on utilizing several internet databases such as Scopus, Semantic Scholar, IEEE Xplore, Science Direct, Springer, Web of Science, and PubMed. Despite extensive efforts made by researchers and practitioners over the course of several decades, the task of data science remains one of the most labour-intensive and least pleasurable endeavours. Data curation is crucial in most businesses to properly harness the potential of big data. Regrettably, the existing solutions have failed to adapt to the constantly evolving data landscape due to their significant reliance on human resources. In recent times, DL has demonstrated significant advancements in various domains, including but not limited to image identification, natural language processing, and speech recognition. This vision paper examines the potential utilization of key advancements in DL to enhance current data curation solutions and facilitate the development of novel ones.

RESULTS AND DISCUSSIONS:

Data curation for artificial intelligence (AI) encompasses the systematic procedure of carefully choosing, refining, and structuring data in order to render it appropriate for utilization in AI and machine learning endeavours. The objective of data curation is to furnish AI models with data that is of superior quality, precise, and pertinent for the purpose of training and enhancing them. The procedure encompasses the elimination of extraneous or repetitive information, rectification of inaccuracies, completion of missing values, and assurance of data consistency. Data curation plays a crucial role in ensuring the accuracy of predictions and the delivery of relevant outcomes by supplying AI systems with high-quality data. A prevailing notion among technology professionals is that providing AI with any acquired data is adequate until they confront the actuality of tainted and prejudiced data in subsequent phases of growth. In order to surmount this obstacle, it is imperative to re-examine the initial data, implement the required modifications, retrain the model, and scrutinize the outcomes.

If data annotation is initiated without prior cleaning or curation, there exists a potential for the resultant data to lack high quality or be unsuitable for utilization in AI applications. This has the potential to yield erroneous or untrustworthy outcomes, hence impacting the efficacy and precision of the AI models constructed using the data. During the annotation process, any errors, duplication, or missing values in the data will not be rectified. Hence, the annotated data might contain mistakes, potentially resulting in biased or deceptive AI models. Likewise, if the data lacks a uniform structure, it can provide challenges when it comes to annotating and utilizing the data in AI applications. For instance, let us contemplate a hypothetical situation in which a computer vision model is being trained to identify pedestrians inside an urban setting. The performance of the model can be influenced by variations in lighting conditions, camera angles, and resolutions within the training data. The model's ability to generalize to novel photos captured under varying settings may be limited, resulting in inaccurate predictions and diminished accuracy.

In the event that the training data comprises photographs lacking appropriate annotations or labels, the model's ability to effectively detect pedestrians within these images may be compromised. Such occurrences may result in inaccurate forecasts, such as categorizing a tree or a lamppost as a pedestrian. Hence, it is crucial to perform data cleansing and curation before annotating it, to guarantee the data's superior quality and appropriateness for utilization in AI and machine learning applications.

Data curators gather data from various sources, consolidate it into a single format, then verify, oversee, store, safeguard, retrieve, and depict it. The initial step in the curation of datasets for DL commences prior to the acquisition of datasets. Data curation generally encompasses many methodologies, as depicted in the following diagram.

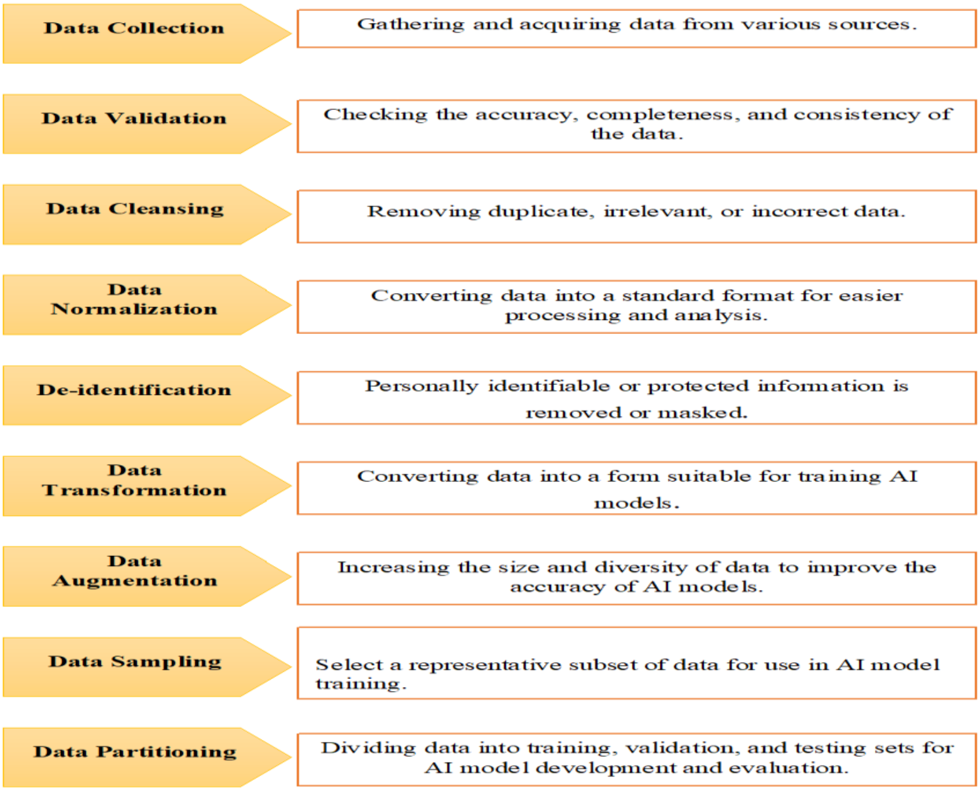
Data Collection
Gathering and acquiring data from various sources.
Data Validation
Checking the accuracy, completeness, and consistency of the data.
Data Cleansing
Removing duplicate, irrelevant, or incorrect data.
Data Normalization
Converting data into a standard format for easier processing and analysis.
De-identification
Personally identifiable or protected information is removed or masked.
Data Transformation
Converting data into a form suitable for training AI models.
Data Augmentation
Increasing the size and diversity of data to improve the accuracy of AI models.
Data Sampling
Select a representative subset of data for use in AI model training.
Data Partitioning
Dividing data into training, validation, and testing sets for AI model development and evaluation.

Figure 2: Methods in Data Curation [3]

These methodologies are employed in diverse combinations and implemented repeatedly to attain data of superior quality for the purpose of training and developing DL models. The lifespan of data involves various stages of change. For improved predictive accuracy, it is imperative that the data encompasses accuracy, diversity, and comprehensive coverage of all edge instances.

High Quality Data: The significance of data quality in AI models lies in its direct impact on the precision of their predictions. AI models rely on the patterns they acquire from the data they undergo training with. Consequently, if the data is of substandard quality or contains inaccuracies, the model will generate inaccurate predictions. In order to attain data of superior quality, companies must guarantee the accuracy, comprehensiveness, consistency, and currency of their data. The aforementioned objective can be accomplished by employing a blend of data validation, data cleansing, and data integration methodologies.

Diverse Data: Having a wide range of impartial data is crucial for training AI models as it guarantees that the model faithfully represents the real-life situation it is intended for. A model that has undergone training using biased or homogeneous data has the potential to provide outcomes that are distorted or inaccurate, hence resulting in inequitable or potentially detrimental consequences.

Edge Case Data: Ensuring comprehensive coverage of all edge cases in the acquired data is crucial for enhancing the accuracy of AI predictions. This is because AI models rely on patterns they acquire from the training data to make informed conclusions. If the dataset is constrained and fails to encompass all potential extreme scenarios, the model will lack a comprehensive comprehension of the problem it aims to address, resulting in potentially inaccurate predictions.
The Deep Learning-Driven Data Curation and Model Interpretation for Smart Manufacturing that Zhang and Gao [8] proposed is depicted in the figure that can be found below.

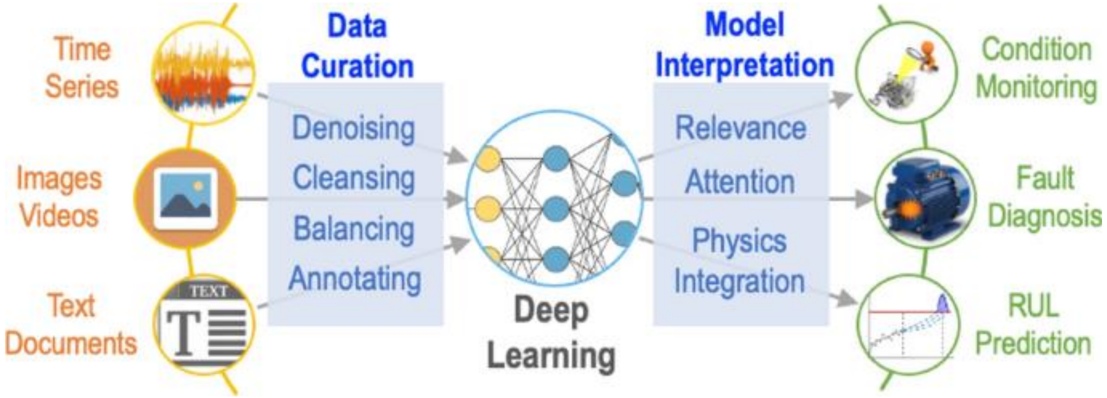

Figure 3: DL based Data Curation proposed in [8]

The following are four empirical instances of data curation in the field of DL :
ImageNet: ImageNet is widely recognized as a prominent dataset employed in the field of machine learning for the purpose of image identification jobs. This serves as a prominent illustration of comprehensive data curation. ImageNet is a repository with more than 14 million photos that have undergone manual annotation and organization in accordance with the WordNet hierarchy, a framework that classifies entities based on their semantic associations. This meticulously curated dataset has been essential in propelling the field of computer vision forward, demonstrating the transformative potential of superior, annotated data in driving advancements in machine learning. The main point to be emphasized is the considerable importance of manual annotation and the utilization of hierarchical classification in the process of data organization.

Twitter Sentiment Analysis: Sentiment analysis is a prevalent task in the field of natural language processing. For example, scholars or organizations may gather a substantial quantity of tweets and subsequently curate this dataset for the purpose of conducting sentiment analysis. The data would undergo a cleaning process, which involves removing duplicate tweets, addressing missing values, and eliminating spam. Subsequently, the data would be annotated by assigning positive, negative, or neutral labels to each tweet. Finally, the text would be converted into a numerical format suitable for comprehension by a machine learning model. This example highlights the significance of conducting thorough cleaning and annotating procedures for unstructured data, such as text.

Autonomous Vehicle Training: DL models utilized in autonomous vehicles are taught using well collected datasets. As an illustration, a dataset could comprise a vast number of photos and video frames obtained from cameras installed on cars. Subsequently, these photos undergo meticulous annotation to accurately detect pedestrians, other cars, traffic signs, and other relevant elements. The data undergoes a process of cleansing to eliminate photos that are irrelevant or of low quality. Subsequently, it is frequently converted into various formats to facilitate the training of diverse models, such as neural networks. The key point to be emphasized is the pivotal significance of data curation in intricate, safety-critical domains such as autonomous driving.

Healthcare: Data curation in the healthcare sector is frequently regarded as a complex yet necessary undertaking. The meticulous collection, cleansing, and integration of data derived from electronic health records, wearables, and medical imaging is crucial. In the context of a disease progression prediction project, data scientists are required to address various challenges such as missing values, discrepancies, measurement standardization, and patient data anonymization. The significance of thorough data curation in healthcare applications is underscored by the high stakes involved, as it is crucial to uphold accuracy and privacy. The aforementioned examples serve to demonstrate the wide range of data curation jobs that exist within various areas. In every instance, meticulous data curation plays a crucial role in generating efficient DL models.

CONCLUSION:

To summarize, data curation is a challenge that has been around for a long time and requires innovative solutions in order to function well inside the expanding big data ecosystem. The concept of DL is gaining popularity across a wide variety of fields, including many that are not related to computer science. By bringing together these two fields of study, a chain of research endeavours will be triggered, which will ultimately result in solutions that can be implemented for a variety of DC jobs. Through our research, this study was able to identify research possibilities in the areas of learning distributed representations for database aware objects, as well as creating DC-aware DL architectures for various real time applications.

REFERENCES

1. Thirumuruganathan, S., Tang, N., Ouzzani, M., & Doan, A. (2020). Data Curation with Deep Learning. In *EDBT* (pp. 277-286).
2. Do, S., Song, K. D., & Chung, J. W. (2020). Basics of deep learning: a radiologist's guide to understanding published radiology articles on deep learning. *Korean journal of radiology*, *21*(1), 33.
3. Yoon, A., Kim, J., & Donaldson, D. R. (2022). Big data curation framework: Curation actions and challenges. *Journal of Information Science*, 01655515221133528.
4. Stonebraker, M., & Ilyas, I. F. (2018). Data Integration: The Current Status and the Way Forward. *IEEE Data Eng. Bull.*, *41*(2), 3-9.
5. Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., ... & Raghavendra, V. (2018, May). Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 international conference on management of data* (pp. 19-34).
6. Thirumuruganathan, S., Tang, N., Ouzzani, M., & Doan, A. (2018). Data curation with deep learning [vision]. *arXiv preprint arXiv:1803.01384*.
7. Marciano, R., Agarrat, S., Frisch, H., Hunt, M. R., Jain, K., Kocienda, G., ... & Xu, J. (2019, December). Reframing digital curation practices through a computational thinking framework. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 3126-3135). IEEE.
8. Zhang, J., & Gao, R. X. (2021). Deep learning-driven data curation and model interpretation for smart manufacturing. *Chinese Journal of Mechanical Engineering*, *34*(1), 71.
9. Cappuzzo, R. (2022). *Deep learning models for tabular data curation* (Doctoral dissertation, Sorbonne Université).
10. Waskom, M. L., Tan, K., Wiberg, H., Cohen, A. B., Wittmershaus, B., & Shapiro, W. (2023). A hybrid approach to scalable real-world data curation by machine learning and human experts. *medRxiv*, 2023-03.