2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)

Data Management Challenges for Deep Learning

Aiswarya Raj, Jan Bosch
Department of Computer Science and Engineering
Chalmers University of Technology
{aiswarya, jan.bosch}@chalmers.se

Helena Holmström Olsson
Department of Computer Science and Media Technology
Malmö University
helena.holmstrom.olsson@mau.se

Anders Arpteg, Björn Brinne
Peltarion AB
Stockholm, Sweden
{anders, bjorn}@peltarion.se

***Abstract*—Deep learning is one of the most exciting and fast-growing techniques in Artificial Intelligence. The unique capacity of deep learning models to automatically learn patterns from the data differentiates it from other machine learning techniques. Deep learning is responsible for a significant number of recent breakthroughs in AI. However, deep learning models are highly dependent on the underlying data. So, consistency, accuracy, and completeness of data is essential for a deep learning model. Thus, data management principles and practices need to be adopted throughout the development process of deep learning models. The objective of this study is to identify and categorise data management challenges faced by practitioners in different stages of end-to-end development. In this paper, a case study approach is employed to explore the data management issues faced by practitioners across various domains when they use real-world data for training and deploying deep learning models. Our case study is intended to provide valuable insights to the deep learning community as well as for data scientists to guide discussion and future research in applied deep learning with real-world data.**

***Index Terms*—Deep learning, Data Management, Machine learning, Artificial intelligence, Deep Neural Networks**

I. Introduction

Over recent years, deep learning has reached the pinnacle of popularity due to its ability to learn deep representations. The capability to learn multiple levels of representations and abstractions from data makes it unique among machine learning techniques [1]. It has been used successfully in image classification [2], object detection [3], natural language processing and information retrieval [4]. Even though the terminologies like machine learning and deep learning are used interchangeably, they do not refer to the same concepts. Machine learning requires a significant amount of work spent on feature engineering [5]. However, deep learning is a particular type of machine learning techniquemore refined Artificial Intelligence technique that can learn from unlabelled data which is an attractive feature demanded by most of the real-world applications [9]. Even though deep learning models have remarkable abstraction and generalisation capabilities, these systems are data hungry in nature. i.e a massive amount of data is required to train Deep Neural Networks. As the requirement for a large amount of data is significant, large-scale data management issues arise in collecting, processing, analysing, sharing and deploying datasets. Although deep learning models are extensively used in a variety of applications, data management for deep learning has received limited attention from researchers and practitioners.

Over the years, there has been a significant advancement in deep neural networks and algorithms. However, this advancement has not been matched with similar progress in data management. Therefore, there is a strong need for new techniques and automated tools to be designed that can assist practitioners in preparing and ensuring quality data throughout the data pipeline workflow.

In this paper, we discuss six real-world industrial applications of deep learning in different domains such as medical imaging, gaming, real-estate, manufacturing systems highlighting the key data challenges to data that can significantly impact the overall performance of DL systems. We do not aim to provide a comprehensive background on technical details and general application of deep learning (see e.g. [6], [7]) nor do we explain extensive challenges faced by real-world software-intensive systems as [8]. Instead, we focus on different data management challenges faced by DL experts while building DL application models. In this paper, we introduce a number of example applications of deep learning frameworks, we explain significant challenges and we categorise these according to the development phase in which it is encountered. The contribution of this paper is twofold. First, it presents the main data management challenges, that need to be addressed for developing high performance and operational deep learning models. Second, the paper classifies the challenges according to the phase in which they are encountered and identifies the main areas that requires attention.

The rest of this paper is organised into six sections. Section II is a description of the background and related works. In section III, we introduce the research methodology adopted for conducting the study. Section IV details the cases explored in the study. Section V focuses on findings of the case study and maps data management challenges encountered at each stage of the data pipeline with the use cases. Finally, Section VI summarises our conclusions and completes this paper.

II. Background

Deep learning [9] provides major advancements in solving the problems which were previously unbeatable by artificial intelligence and machine learning techniques. Due to this reason, it is being used in hard scientific problems like reconstruction of brain circuits [10], mutation analysis in DNA

This work is in part supported by Vinnova.

978-1-7281-3421-5/19/$31.00 ©2019 European Union
DOI 10.1109/SEAA.2019.00030

[11], structure-activity prediction of potential drug molecules [12] and online particle detection [13]. Deep neural networks are also opted to decipher many challenging tasks in speech recognition [14] and natural language processing [15].

Deep learning became the focal point after Krizhevsky et al. [16] demonstrated the remarkable performance of a Convolutional Neural Network (CNN) [17] based model on a challenging large-scale visual recognition task [18] in 2012. A substantial credit for the current reputation of deep learning can also be attributed to this influential work. Great contributions to deep learning research have been made by the computer vision community by providing solutions for the problems encountered in medical science to mobile applications since 2012. The recent breakthrough in artificial intelligence in the form of tabula-rasa learning of AlphaGo Zero [19] also owes a fair share to deep Residual Networks (ResNets) [20] that were originally proposed for the task of image recognition.

A. Data - The fuel

Data is the fuel for deep learning models. Massive datasets are used to train deep neural networks in order to mimic human intelligence. It is data which allows industries to stay on top of trends, provide answers to problems, and analyse new insights to great effect. There are numerous algorithms in deep learning tailored for various applications which deliver high performance. However, no algorithm can guarantee the same performance over all the datasets. This is a clear indication of the importance and effect of data in the performance of DL models.

B. Data Management

Data management for deep learning can be defined as a process which includes collecting, processing, analysing, validating, storing protecting, and monitoring data to ensure the consistency, accuracy, and reliability of the data. Deep learning has been successfully applied in industry products that take advantage of the large volume of digital data. However, real-world data needs to be processed and managed before feeding it as input to the deep learning models. Training a deep learning model with such massive and variegated data sets is challenging and several aspects need to be considered. E.g. data sparsity, redundancy, and missing values. In order to ensure high performance of DL models, not only good algorithms but also the management of data is required. A set of good data management practices should be followed from data collection, through data processing and analysis, dataset preparation and deployment of the model.

As DL applications are highly data-driven, it could be benefited from data management and database techniques to accelerate the speed of training. Wang et. al [21] describes how certain challenges like data dependency, memory management, concurrency, data inconsistency can be solved by combining database techniques and deep neural networks.

DL models demand large volume and variety of data which relates it to the field of Big Data. Popular companies like Apple [22], Google [23], Facebook, Microsoft are collecting a copious amount of data on a daily basis through applications like Siri, Google translator, Bing voice search [24] to provide a variety of other services such as reminders, weather reports, personalised recommendations. Although big data offers numerous opportunities, it also imposes consequential engineering challenges [25]. X. W. Chen et al. describes the big data challenges such as streaming data, high-dimensional data, scalability of models, and distributed computing [26]. However, in these papers, deep learning is considered as a solution for management of data. Data management challenges involved in implementing deep learning models are not seriously considered and our paper intends to focus on that perspective.

III. Research Method

In order to set the scope for the type of empirical studies we address in this paper, an interpretive multiple-case study approach was adopted adhering to the guidelines by [27]. Usage of multiple cases should be considered same as the duplication of a study or an experiment which means that the inferences from one case should be compared and contrasted with the results from the other case(s). The objective of this study is to identify challenges specifically related to the management of data in various real-world DL applications. The challenges identified are based on our interpretations of the experiences of experts who implement DL systems in a real-time scenario with real-world datasets. This type of case study research is appropriate as it facilitates the exploration of the real-life challenges in its context through a variety of lenses [28]. The overall research design and the major steps in the research process of the study are described below.

A. Expert Interviews

The objective of the study is to explore data management challenges encountered while implementing DL models in real-world settings. Each case in the study refers to a software-intensive system that incorporates DL components developed by an organisation. For the study, a sample pool of DL experts who works in seven different domains were selected by their expertise in the area of study. The selected seven practitioners include two authors of this paper. From the acknowledgment in the literature (and our experiences when soliciting interviewees), it can be inferred that only a few experienced practitioners are skilled in the area of intersection between DL and SE, Table 1 illustrates the vast experience of our interviewees in incorporating DL components across multiple domains.

B. Data Collection

Semi-structured interviews were used to acquire qualitative data. Based on the objective of research to explore data management challenges for deep learning systems, an interview guide with 40 questions categorised into four sections was formulated. The first and second sections focused on the background of the interviewee. The third section concentrated on

the importance of data in various projects and the last section inquired in detail about data management, the challenges faced during every phase of the data processing pipeline. The interview guide was reviewed by the authors and some additional questions were added, a few similar questions were merged together and some totally irrelevant questions were removed finally forming an interview protocol with 20 questions spread across four different categories. All interviews were face-to-face except for one which was done via video conference and each interview lasted 45 to 55 minutes. All the interviews were recorded with the permission of respondents and were transcribed later for analysis.

C. Data analysis

After the interviews, audio recordings of interview were sent for transcription and a summary of each interview was made by the first author highlighting the main focus points of the interview. The analysed points from the summary were cross-checked several times with the audio recordings and interview transcripts obtained after transcription. A theoretical thematic data analysis approach was opted for coding [30]. First, the author coded each segment of the interview transcript that was relevant to or captured something interesting about data in NVivo. In the first iteration, the aim was to identify the phases of data pipeline. After identifying the phases, a second iteration was performed to code the data management challenges encountered in each phase by setting high level themes as *(i) Data Collection, (ii) Data Exploration, (iii) Data Preprocessing, (iv) Dataset Preparation, (v) Data Testing, (vi) Deployment, (vii) Post-deployment*. The results deduced from the analysis were tabulated and sent to the authors for comments and then the final summary of the cases and mapping were sent to the interviewees for validating the inferred results.

TABLE I
DESCRIPTION OF USE CASES AND ROLES OF THE INTERVIEWEES

Case	Use case of DL components	Interviewed Experts	
		ID	Role
A	Recommending products to the users in a personalised fashion	P1	Principal Data Scientist
B	Predicting the wind power using the historical weather data	P2	Data Scientist
		P3	Head of Data Analytics team
		P4	Data Scientist
		P5	AI Research Engineer
C	Estimating and predicting the price of houses	P2	Data Scientist
		P3	Head of Data Analytics team
D	Automated classification of skin lesions into benign and malignant	P2	Data Scientist
		P3	Head of Data Analytics team
		P4	Data Scientist
		P5	AI Research Engineer
D	Detecting the credit card frauds during gaming	P2	Data Scientist
		P3	Head of Data Analytics team
E	Predicting quality of paper boards	P4	Data Scientist
		P5	AI Research Engineer

IV. CASES

This section describes different real-world DL cases that has been chosen for this research. All the cases reported here are using real-world dataset. A mapping between different data management challenges and projects is presented in a later section.

A. Recommender System

Many e-commerce and retail companies are leveraging the power of data and boosting their sales by implementing recommender systems on their websites. When a customer visits the website, the recommender system predicts users' interest and recommends electric products for them based on previous customer reviews and purchase history. Many times customers tend to look at the website for their recommendations. Personalised recommendations from the system would increase customer satisfaction and thus customer retention. In recommender systems, DL components are trained on user reviews and their purchase history.

"It's very difficult to focus on the things that aren't visible feature wise, such as tracking data. So that means that often you first develop the features the way you want them, and if you have time in the end you put tracking in, so you put the gathering of the data, that part of the code in. So that obviously means that it's not as well tested, and it's not as well tracked. It doesn't get the same love when you develop and so on. So that usually makes data quality bad."

B. Wind Power Prediction

Wind power is dependent on weather and so it is irregular and fluctuates over different time scales. Thus accurate forecasting of wind power can be considered as a major contribution for reliable large-scale wind power integration. DL model is utilised to predict accurately how much electricity, how much power are all of the wind turbines going to generate within 24 to 48 hours so that an accurate report can be submitted to the power companies for which energy is supplied. The power companies have quite strict requirements like they have to accurately say how much power they are going to deliver and if not there are penalties that need to be paid if they don't manage to deliver the reported energy. A combination of wind and weather are predicted from which the power generated by the wind turbines can be calculated. The wind power is predicted based on the meteorological data obtained from the National meteorological agency. Deep Learning is used to forecast weather and thereby predicting the wind power that can be generated in the future.

"We have gotten our data in all sorts of different ways. When we got the data for the weather prediction case we actually got them on physical tapes. Those ... Really in boxes, with physical tapes. And then we had to digitalize them ourselves. So, I do not think there is any standard or framework to accomplish this. That is why data management itself is a problem when dealing with deep learning."

C. House Price Prediction

Predicting property values are of great interest to various parties in an economy. Estimation of house price is important to prospective homeowners, developers, investors, appraisers,

tax assessors and other real-estate market participants, such as mortgage lenders and insurers. Real-estate investors and portfolio managers devise and carry out their investment decisions based on periodic evaluations of their real-estate portfolios. Individuals are interested in knowing the values of their properties before setting up their list prices. Tax authorities rely on the estimates of the properties' value as the basis for levying property taxes. Banks and mortgage providers conduct housing collateral valuation to qualify the borrowers for their mortgage applications. Initially, house price was predicted on the basis of comparison between cost and sale price and there were no accepted standards or certification process. So, the house price prediction model helps to fill up the information gap that existed before and also enhance the efficiency of the real-estate market. The house price prediction case was initially built on traditional assorted database system where SQL queries and data pipeline scripts were used whereas now it utilises deep learning technique where the model is trained with historical sales data about properties, geography, and demography in the Swedish market. The house price prediction system is a long-running DL model deployed in production and is used by many banks in Sweden. *"Someone has to make a design choice, like is this interesting to collect, on what level, and what kind of metadata do I attach to it for example. Because it's easy to just log something, but then when we come later as data scientists and look at the data and we're like "okay, that's really good, but which user was that? Ah, we didn't log it". Okay, if you didn't log that, then I can't really combine that data with my other data set where I have it on user level. So I can't really, you know, all of those missing pieces are challenges"*

D. Melanoma Detection

Melanoma is a type of skin cancer, which is not that common like basal cell and squamous carcinoma, but it has dangerous implications since it has the tendency to migrate to other parts of the body. Therefore, early detection can prevent it from spreading to other parts; otherwise, it becomes incurable. Deep learning bypasses all the complex methods of pre-processing, segmentation and low-level feature extraction. Although a lot of datasets like MED-NODE, ISIC Archive and many more are publicly available, dealing with real-world data is still challenging. Automated classification of skin lesions using images is a difficult task because of the unavailability of fine-grained varieties in the appearance of skin lesions. The skin cancer detector not only intends to detect whether a person has skin cancer or not but also what kind of cancer it is and thus how serious it is. Here, the deep learning model is used for diagnostic classification of dermoscopic images of lesions of melanocytic origin. Datasets are formed over several years by working in close collaboration with clinics. The company has restrictions on the use of the dataset and the requirement is even the data cannot leave the servers. With these restrictions, the practitioners adapt to the rules specific to the dataset and move the code and model to the server where data is stored for developing the DL model. The skin cancer detector is still not production ready.

"It is very difficult to scale the data collection, because you have to get something from a patient. Very intrusive things, like sticking electrodes into their skin and taking images, or something like that. So there, it is kind of hard to increase the amount of data you have quickly, because you need to see patients that go through the health care system, and you know. What you can do is try to use publicly available data that is similar, but then you always had issues with the data not being quite the same. Not the same distributions, different cameras, different machines etc. So that is a difficult domain."

E. Financial Fraud Detection

Frauds in finance still amount for significant amounts of money. Around the globe, hackers and crooks are trying to find new ways of committing financial frauds. Therefore, trusting financial fraud detection systems programmed based on conventional rule-based method alone will not serve the purpose. This is where Deep Learning shines as a unique solution. The DL model uses customer details like payment history and activity history and payment request data such as payment method, amount location, etc. The post-payment signs of abnormal pay are also taken into account for detecting fraud. When it comes to modelling fraud detection as a classification problem, the main challenge comes from the fact that in real-world, the majority of the transactions are not fraudulent. However, in order to train DL models, counterexamples are also required.

"If you have a company that deals with credit card fraud or something like that, and then they record all the examples of when people have had the fraud. And then if they don't have the counter examples of the normal examples, then it's again difficult. "

F. Manufacturing Systems

Paper mill industry creates paper from pulp and then dry that into carton and cardboard which is further used for making milk cartons. A DL component is incorporated in the system to predict the quality of the resulting product based on all the measurements in the machine and measurements on the pulp that goes in. And there's also images of what's happening at the beginning of the machine, and images, microscope images of the fibers in the pulp. The company manufactures large quantities of paper board each year and wanted to minimise the material cost as much as possible while maintaining high quality. Quality of the paper board is predicted based on data from process sensors and images of wood fibers taken with the PulpEye technology. The DL models serve as a stepping point for controlling the manufacturing process so that the same quality could be maintained with less input material and waste.

"I think this data engineering in the beginning or, that is supposed to be in the beginning, has always turned out to be a much bigger problem than you think. Because you usually realize after you have started modeling that you have had some

assumption that was not really correct, and then you have to go back and kind of redo the data engineering again."

V. Findings

This section presents a list of concisely described data management challenges encountered by practitioners while implementing DL components in real-world applications. Based on the study, we have identified seven stages through which data flows and the data management challenges raised in each phase. Our study is carried out with six use cases as mentioned above. Many of the challenges identified are use case specific and so a mapping between these use cases and data management challenges are shown in table 2.

A. Data Collection

The systematic process of gathering data from a range of sources relevant to the context is termed as data collection. Deep learning model should be constantly fed with data to continue improving performance while deployed in production-ready systems. Acquiring data is thus a crucial phase which needs attention.

1) Lack of metadata: Metadata is required for the practitioners as they might not be experts in the domain where they implement DL components. Practitioners mentioned that in many projects, lack of metadata creates confusion and poor understanding of the data. Due to poor organisation, the semantics of data is often obscured which in turn leads to ambiguities. When a dataset is handed in for building a stock market price prediction, the dataset may have different prices like opening price, closing price, quoted price, session price and without providing associated metadata information, it is hard for the practitioners to identify and distinguish different prices. Without metadata, it is not always possible to figure out if some pattern makes sense or not. If you know that a particular signal represents a temperature reading and it is always zero, then you know it is wrong. On the other hand, if it is an on/off switch, then maybe it is zero all the time which is fine.

2) Data Granularity: Data aggregation may remove important data points which cannot be collected again. Even after collecting a huge amount of data and then aggregating it after a certain span of time will spoil the detail in the data. Thus fine granularity in data is lost through data aggregation techniques. Like in mobile networks, counter data is collected and aggregated some value over 15 minutes, and that's what gets saved. Because saving every second's data point is not affordable. And then in that aggregation, a lot of information is lost, which could mean that even though a lot of data are in place when looking at it in detail, granularity actually needed for a use case will not be there. So even if data is collected over ten years, the problem is still kind of limited by data collection choices which are difficult to get around. In our study, the recommender system case experience this data granularity problem. When the reviews from users are all logged for a long period and handed over for building recommender system, but failed to log the user's identity, the data granularity is lost. And it is not possible to combine that data with other data set on user level.

3) Shortage of diverse samples: Upon training, the deep neural network should be given all possible instances and varieties of data so that it will not fail on inputting unseen data in production. However, during data collection, many companies collect a large number of normal samples and fail to collect the counterexamples of data. The DL model needs to be trained with counterexamples as well. From our case study, one extreme example we got is that financial fraud detection cannot be developed only with the samples of fraud cases, it also requires normal transaction instances in the dataset. Deep learning models cannot learn the normal cases by themselves when only the abnormal samples are fed during the training which leads to weird outputs after deployment.

4) Need for sharing and tracking techniques: Sharing the collected data with the practitioners is required while implementing the Deep Neural Networks. There is no defined channel or medium for sharing the collected data. According to the size of the data, different people choose different means of sharing. Some companies may opt to share the data in the form of excel files over email, FTP server or even in the form of physical tapes. Two of the experienced practitioners identifies tracking as an important measure by which data quality can be assured. However, due to the tight limit on time and resources, often data tracking is not focused much or is kept at a least priority leading to poor data quality.

5) Data Storage: Deep learning systems are powerful in memorising each and every piece of information given to it. So the amount of training data has the biggest impact on the performance of the model. General Data Protection Regulation(GDPR) is a regulation in EU law to protect online personal data. GDPR is a set of legislative rules which impose restrictions on processing and storage of information. Major companies who focused on collecting and maintaining datasets are able to build better DL models to a certain extent. The problem with small scale companies is that they do not have clear knowledge on how to collect and store data complying to the rules of GDPR and there is no framework or protocol to help them to do data collection efficiently. In such cases, a certain percentage of revenue needs to be paid as a penalty for not following the regulations of GDPR which end up in the deletion of a huge portion of data they collected over time. Even though this is a problem experienced by only one case in the entire study, it is still important as it has significant legal and financial complications involved.

B. Data Exploration

Data exploration is analysing the distribution of different datasets and data fields, checking the number of outliers and existence of missing data, examining how to connect the data together and build up basically a dataset that can be fed directly into the model.

1) Statistical Understanding: When confronted with data that needs to be analysed, the first step is to carefully identify the distribution of data. Statistical understanding is much

required for determining the distribution of data. Even with sufficient knowledge in statistics, it is challenging to identify the distribution of data. The normal distribution or Gaussian distribution is that nice, familiar bell-shaped curve. But, data comes from a range of devices out in the wild and there is no point in assuming an easy to handle normal distribution. For instance, consider an image processing application, to model the pixel values efficiently, assumption of Gaussian distribution is inaccurate as it violates the boundary properties. In such cases, models like BMM(Beta Mixture Model) is opted. Without clear knowledge of statistical distributions, it will become difficult to model the distribution.

2) Deduplication Complexity: Dataset often has a lot of duplicates, some with slight variations and some exact copies. So analysing the dataset for duplicates and deduplication is a complex task. For example, consider a song recommender system trained on a dataset of songs. If you take a random song, there can be 200 versions of the same song with slight variations in it, but it's more or less the same song. If the model is trained with such a dataset, the result may turn out horrible such that it may recommend 50 copies that sound more or less the same. In such cases, deduplication becomes complex. Because if the dataset has 100,000,000 songs, you need to compare a song with every other song in the dataset. So it's a quadratic complexity of that problem, it's impossible to do from a time point of view in a single machine and you have to run it on hundreds and hundreds of machines.

3) Heterogeneity in data: Format, size and encoding techniques varies from data to data. A single dataset itself may have data in audio, video and text formats. If a dataset with only textual data is examined, some text will be in UTF-8, some in UTF-16, some in CSV, comma-separated format, some others in tab-separated format, some having HTML code in the actual text, some having an additional weird like placeholders embedded inside the actual national language text. So it is required to invest a lot of time and effort in just transforming the text into a uniform format and coding for the data. All six cases studied here have this problem.

C. Data Preprocessing

Real-world data is often incomplete, inconsistent, and erroneous. So data preprocessing is an inevitable task before creating datasets which resolves the issues inherent with raw data. As data is coming from different sources, there can be missing data, wrong values, and ill formatted data which spoils the consistency and needs to be solved before feeding it to the DL models.

1) Dirty Data: Raw data comes up with a lot of imperfections like missing values, wrong values, and ill-formatted values. These unclean or noisy data is known as dirty data. Deep neural networks are good at deriving patterns from the given input. So, it is dangerous to feed noisy data to the DL models. Also, the DL experts might not be experts in the domain and so they are totally unaware of what needs to be filled when there are missing values and how to identify the wrong or ill-formatted values. For example, if there is a column for age and some of the values are missing. The system is supposed to make predictions based on each individual user and you do not have the age for 10% of them. That column can be filled out with the average or minus one. In order to fill the column, it is required to know what the column is meant to be and what can be filled in to replace the missing/wrong/misformatted values. All practitioners agree that they have faced this unclean data issue in all the cases they handled until now and in most of the cases, discussion with the people who collected the data was the only practical solution.

2) Managing categorical data: Categorical data are nothing but variables with label values instead of numerical values. Categorical variables can be both nominal as well as ordinal. Deep learning models cannot operate on label values as it requires all input variables in the numeric form. Even though one-hot encoding is used very frequently, it can be frustrating during implementation. When there are thousands of categories, the complexity again increases. If you have text data, for example, that needs to be cleaned up and transformed into numeric form. Then it might not be possible to do it with a laptop or even a big server. There are core systems like Hadoop, Spark or Google DataFlow where big data processing can be done. However, it's still very dependent on the person doing it, what they are comfortable with, and also the data, how big is it, how difficult is it and what needs to be done with it. There are no predefined sets or standards to handle this.

3) Managing sequences in data: Metadata management should be considered with equal importance in order to manage the sequences in data. Storing the sequencing data alongside the contextual metadata is a bit challenging especially when the data quantity is too large. For instance, for chronological data, there is a time series which needs to be divided chronologically somehow, so you do not end up predicting the past.

D. Dataset Preparation

During dataset preparation, the main large dataset is divided into three different sets namely training, validation and testing dataset. Deep understanding of domain and problem will aid in relevant structuring values in the datasets.

1) Data Leakage: Data leakage is the challenge of not splitting the training and validation/test dataset properly so that the training data for the model happens to have the data which needs to be predicted. For instance, data leakage happens when the same data instance occurs in both training and testing dataset. This hides the actual performance of the model and when it is exposed to new and unseen data, the performance will not be as expected. So proper attention should be taken while splitting the dataset. Based on the study, we could infer that checking the data distribution is not always a solution to reduce data dependency.

2) Data Quality: Quality of data is crucial as poor quality data can cause severe performance degradation and exaggerated results. Data consistency is one of the factors deciding the quality of data. However, consistency is a hard to achieve

target in many applications. For example, based on our study, the images collected from the hospitals are all taken in different conditions with different lighting. Accuracy, completeness, timeliness, validity are some other factors that ensure the quality of data. However, there is no exhaustive list of factors that should be checked to ensure the quality of data which is challenging.

E. Data Testing

Testing the data is a critical step which ensures the data quality and reduces the possible occurrence of defected data that affects the efficiency of the process. Absent, obsolete or wrong test data may prevent the practitioners from executing the test cases or give unreliable test results.

1) Expensive Testing: Data testing is highly expensive in the sense that it requires a lot of effort and time to define and automate test-cases specific to DL models. It's pretty hard to do regression testing on data, because data comes from users out in the wild where exerting control is impossible.

2) Tooling: Tooling comes as a challenge in most of the phases of the data pipeline. The major advantage of conventional software systems is that there exists a large variety of tools, especially for testing. As deep learning is a recently emerged approach, tools for testing such models are yet to be developed. All the cases included in our study experience tooling problem.

F. Deployment

Dl models need to be operationalised or put into production to measure the real performance and to generate a positive return for the investment in system development. When systems are ready for deploying in production, there are a unique set of challenges encountered which is explained as follows.

1) Data Extraction methods: Training-serving skew is a typical problem encountered when running deep learning models in production where the data seen at serving time differs in some way from the data used to train the model, leading to reduced prediction quality. For example, Google once built a system called quick access in Google drive which recommends a list of documents to open [29]. When the system was built, they first extracted data and made a training dataset, trained a model on it and did the evaluation which looked great. So, they put it in production, and it didn't work. When analysed, they realised that when they extracted the data for the training, they had a certain pipeline that it went through, but when they put it in production they had the data extracted from an API, and that API wasn't matching with the extraction they had for training. So it was some additional transformation happening in the API that caused the model to not work.

2) Overfitting: Overfitting is the situation when deep neural network memorises and fits itself so closely to the training set that it loses the capability to generalise and make predictions for new and unseen data. For instance, in a medical imaging case referred above where tabular data is used along with images, it turned out that the model was just learning the ID number of a certain hospital, and that hospital was a popular hospital to which the more severe cases were sent. So actually, the model was not learning anything from the images, rather it was just learning that the patients in that hospital are more likely to be sick, which is because they were sent there.

G. Post Deployment

Continuous monitoring is required even after deploying the DL components in production. This is because the real-world data is prone to all kind of shifts and distribution changes and the model learns constantly. The possible data management challenges after deployment are listed below.

1) Changes in data sources and distribution: When a certain problem is modelled, a distribution is postulated based on the data available at that time. However, consistency in data distribution cannot be expected all the time. Consider the house price prediction system in our study which is trained on historical real-estate data. When some sudden environmental disaster or society-wide effect takes place, the usual distribution will be disturbed and the trend in data changes. When data distribution changes, deep neural networks may not always be able to handle the new distribution. A sudden change in the source that supplies data can also lead to unexpected and undesired outcomes.

2) Data Drifts: Data drifts are also known as data shifts which happen over time. When data shifts happen, deep learning models may deliver weird and erroneous results. Consider systems, such as mobile interactions, sensor logs, and web clickstreams. Whenever the business tweaks or updates happen, the data those type of systems generate changes continuously. The sum of these changes is data drift. Other common examples of structural drift are fields being added, deleted and re-ordered, or the type of field being changed. For example, to support a growing customer base, a bank adds leading characters to its text-based account numbers. This kind of data changes causes the bank's customer service system to conflate data related to bank account 00-56789 with account 01-56789. All practitioners agree that most of the cases that they handle are subjected to this challenge.

3) Feedback loops: Feedback loops are sometimes beneficial and at times detrimental. For instance, if you implement recommendation systems, of course, there will be feedback loops. Because, the data collected will be mostly from your own customers and if you give them suggestions on what to buy, of course, they will buy more of that. Then the model sort of reinforces itself.

During the case study, all the practitioners agreed that while building any deep neural network, management of data requires more effort and time than model creation and coding as there exists a number of readily available algorithms for performing any deep learning task.

If companies are able to act quickly to embrace naive ideas and opportunities, they will gain a valuable first-mover advantage. Companies who can get their data management and DL capabilities in order now will be in prime position to benefit from the next generation of AI operations tools as soon

TABLE II
MAPPING BETWEEN DATA MANAGEMENT CHALLENGES AND USE CASES

Phase	Challenge	Use cases of DL components					
		RS[1]	WPP[2]	HPP[3]	MD[4]	FFD[5]	MS[6]
Data Collection	Lack of metadata	X	X	X	X	X	X
	Data Granularity		X	X			
	Shortage of diverse samples		X	X	X	X	
	Need for sharing and tracking techniques	X	X	X	X	X	
	Data Storage	X					
Data Exploration	Statistical Understanding		X			X	X
	Deduplication Complexity	X	X	X	X	X	X
	Heterogeneity in data	X	X	X	X	X	
Data Preprocessing	Dirty data	X	X	X	X	X	X
	Managing sequences in data					X	X
	Managing categorical data				X	X	
Dataset Preparation	Data Dependency	X	X	X	X	X	X
	Data Quality	X	X	X	X	X	X
Data Testing	Tooling	X	X	X	X	X	X
	Expensive Testing	X			X		X
Deployment	Data Extraction Methods	X	X	X	X	X	X
	Overfitting				X	X	
Post Deployment	Data sources and Distribution	X	X	X			
	Data drifts	X	X	X			
	Feedback loops	X					

[1]Recommender System [2]Wind Power Prediction
[3]House Price Prediction [4]Melanoma Detection
[5]Financial Fraud Detection [6]Manufacturing Systems

as they hit the market. This could give them an opportunity to secure a decisive edge over the competition.

VI. CONCLUSION

Deep learning has established itself as one of the most popular techniques in the area of Artificial Intelligence and data management is an integral part of deep learning models as the performance of these models largely rely on data. However, without extensive research and highly developed supporting infrastructure, companies may face significant challenges while building production-ready systems with DL components.

In this paper we identified main data management challenges while building systems with DL components. Six use cases were described to identify the challenges and also exemplify the potential for making use of the AI and specifically the DL technique. For these cases, the main problematic areas and challenges with building these systems were identified. To clarify these problem areas in more detail, a set of 20 challenges were identified and described across the phases of data pipeline. The challenges identified in this paper help practitioners to foresee the roadblocks that may encounter while managing data for deep learning systems. It also provides an overview of research challenges to be addressed by the academic community. The study helps to identify the probable blind spots for the companies wishing to implement deep neural networks as well as guide future research.

REFERENCES

[1] S. Zhang, L. Yao, A. Sun, Y. Tay, "Deep Learning based Recommender System: A Survey and New Perspectives" ACM Computing Surveys, Vol. 1, No. 1, Article 1, July 2018

[2] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In Proceedings of CVPR, pages 770–778, 2016. arxiv.org/abs/1512.03385

[3] C. Szegedy, A. Toshev, and D. Erhan. "Deep neural networks for object detection". In NIPS, 2013

[4] X. Liu, J. Gao, X. He, L. Deng, K. Duh, YY. Wang. "Representation learning using multi-task deep neural networks for semantic classification and information retrieval". In NAACL, 2015

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012

[6] R. Miotto, F. Wang, S. Wang, X. Jiang, J. T. Dudley. "Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics",2017, pages 1–11

[7] T. Wang, C. K. Wen, H. Wang, F. Gao, "Deep learning for wireless physical layer: opportunities and challenges," preprint, 2017. [Online]. Available: https://arxiv.org/abs/1710.05312

[8] A. Arpteg , B. Brinne , L. Crnkovic-Friis , J. Bosch, "Software Engineering Challenges for Deep learning". SEAA Conference, 2017.

[9] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning", Nature, vol. 521, pp. 436-444, 2015

[10] K. Pei, Y. Cao, J. Yang, S. Jana. "DeepXplore: Automated Whitebox Testing of Deep Learning Systems". ArXiv e-prints, 2017

[11] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, W. Denk, "Connectomic reconstruction of the inner plexiform layer in the mouse retina", Nature, vol. 500, no. 7461, pp. 168-174, 2013.

[12] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, "The human splicing code reveals new insights into the genetic determinants of disease", Science, vol. 347, no. 6218, pp. 1254806, 2015.

[13] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, V. Svetnik, "Deep neural nets as a method for quantitative structure–activity relationships", J. Chem. Inf. Model., vol. 55, no. 2, pp. 263-274, 2015.

[14] T. Ciodaro, D. Deva, J. M. de Seixas, D. Damazio, "Online particle detection with neural networks based on topological calorimetry information", J. Phys. Conf. Series, vol. 368, no. 1, pp. 012030, 2012.

[15] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82-97, Nov. 2012.

[16] K. Alex, S. Ilya, G. E. Hinton, "ImageNet classification with deep convolutional neural networks", Advances in neural information processing systems, pp. 1097–1105, 2012.

[17] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Proc. Adv. Neural Inf. Process. Syst., pp. 1097-1105, 2012.

[18] Y. LeCun, B. Boser, J. S. Denker, D. Henderson., "Backpropagation applied to handwritten zip code recognition", Neural Comput., vol. 1, no. 4, pp. 541-551, 1989.

[19] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang,A. Guez, Y. Chen, "Mastering the game of go without human knowledge". Nature, vol. 550, no. 7676, pp. 354-359, 2017

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016

[21] W. Wang, M. Zhang, G. Chen, H. Jagadish, B. C. Ooi, and K. L. Tan, "Database meets deep learning: Challenges and opportunities," ACM SIGMOD Record, vol. 45, no. 2, pp. 17–22, 2016.

[22] X. W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," IEEE Access, vol. 2, pp. 514–525, May 2014

[23] A. Efrati. "How 'deep learning' works at Apple, beyond". Information [Online]. Available: https://www.theinformation.com/How-Deep-Learning-Works-at-Apple-Beyond, 2013

[24] N. Jones, "Computer science: The learning machines," Nature, vol. 505, no. 7482, pp. 146–148, 2014.

[25] M. M. Najafabadi,F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, E. Muharemagic, "Deep learning applications and challenges in big data analytics", J. Big Data, vol. 2, no. 1, pp. 1-21, 2015.

[26] Y. Wang, D. Yu, Y. Ju, and A. Acero, "Voice search," in Language Understanding: Systems for Extracting Semantic Information From Speech, G. Tur and R. De Mori, Eds. New York, NY, USA: Wiley, 2011, ch. 5.

[27] P. Runeson, M. Host, "Guidelines for conducting and reporting case study research in software engineering". Empirical Software Engineering, 14(2) (2008)

[28] P. Baxter, S. Jack, "Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers ", 13(4), vol. 13, issue4, Jan 2008

[29] S. Tata, A. Popescul, M. Najork, M. Colagrosso, J. Gibbons, A. Green, A. Mah, M. Smith, D. Garg, C. Meyer, "Quick Access: Building a Smart Experience for Google Drive. In Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)". 1643–1651, 2017.

[30] M. Maguire, B. Delahunt, "Doing a Thematic Analysis: A Practical, Step-by-StepGuide for Learning and Teaching Scholars", vol. 3, 2017