FactSheets for Hardware-Aware AI Models: A Case Study of Analog AI Models

Brandon Dominique
dominique.b@northeastern.edu
Northeastern University
Boston, MA, USA

Kaoutar El Maghraoui
kelmaghr@us.ibm.com
IBM Research
Yorktown Heights, NY, USA

David Piorkowski
djp@ibm.com
IBM Research
Yorktown Heights, NY, USA

Lorraine M. Herger
herger@us.ibm.com
IBM Research
Yorktown Heights, NY, USA

ABSTRACT

In the last few years, documenting and tracking the lineage of AI models has emerged as a important research area that can help to improve the transparency, traceability and overall effectiveness of a model when it is used or deployed by an entity that did not create it. Multiple documentation methods have been proposed and their adoption has slowly begun, but these methods tend to focus on the data science aspects of the model creation, such as the datasets used to design and train the model, the neural network structure of the model, the F1 score, the modal bias, etc. When adapted to the emerging AI hardware field of analog in-memory computing AI which aims to reduce the energy use of models, additional documentation requirements need to be considered. We use IBM's AI FactSheets (FS) 360 documentation methodology to understand and evaluate the documentation needs in this emerging domain. To do so, we interviewed 12 participants who represent various roles throughout the lifecycle of designing, training, evaluating, deploying and consuming an analog-aware AI model. We draw from the experiences of the participants in IBM's AI hardware center. From these interviews we capture these roles' documentation and collaborative needs, develop FactSheets to meet those needs, and evaluate the quality of completed FactSheets. We show that the FactSheets methodology can be applied to Analog AI models to successfully create meaningful documentation that is suitable across multiple roles.

ACM Reference Format:
Brandon Dominique, Kaoutar El Maghraoui, David Piorkowski, and Lorraine M. Herger. 2022. FactSheets for Hardware-Aware AI Models: A Case Study of Analog AI Models. In *Proceedings of The 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2022).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ESEC/FSE 2022, 14 - 18 November, 2022, Singapore

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent papers related to documentation for AI have provided a few different methods of recording important aspects of a model and its training data. While they differ in terms of how to obtain this information, they all converge to the same goal of capturing and storing key facts and details about how the model was created and tested to create more responsible and transparent AI systems. One aspect that has been overlooked by existing AI documentation efforts is to consider the hardware efficiency of AI models. This is becoming increasingly important with the exponential increase in AI models' size and complexity and the explosion of a variety of AI hardware deployment platforms ranging from AI supercomputers, cloud data centers, to personal computers and edge devices. Deep learning (DL) models are becoming more dense, with millions or even billions of parameters. More recently, models have been published that break the trillion parameter threshold for the first time, such as the Switch Transformer [7]. The increased size and complexity of DL models translates into high computational demands and energy demands. In fact, compute requirements for large AI training jobs have been doubling every 3.5 months [1]. This also translates into a high carbon footprint [11]. Additionally, most of these models are too expensive computationally to run on devices such as cell phones, embedded devices, or edge devices. As more AI models enter production in enterprises worldwide, AI practitioners, desire systems that are performing, efficiently, and sustainably in addition to being ethical, explainable and accountable. Therefore capturing the details of how to tune, optimize, compress, and scale AI models are key aspects that should be captured in AI documentation.

As the future of AI is moving to the edge and AI becomes increasingly important at the edge, we need to consider more seriously a shift from state-of-the-art model accuracy to state of the art model efficiency. Hardware-software co-design approaches are becoming increasingly needed to specialize and optimize AI models for these constrained devices and enable on-device AI. This is forcing hardware designers and AI model builders to work more closely together–two communities that historically have not collaborated closely. AI hardware designers need to understand more deeply how neural networks work at the algorithmic level to guide their AI chip design. Approaches to bridge the gap between hardware designers and AI models builders are needed to allow for collaborations and efficient communication between these two groups in order to advance the state of the art. We are experiencing these

challenges first hand in IBM's AI Hardware Center [20]. The center is focused on creating and enabling next-generation chips and systems that support the tremendous power and processing demands for emerging AI workloads. The AI Hardware Center is taking a holistic approach to building AI systems from the ground up – from materials, chips, devices, architecture, simulation, emulation and the entire software stack. This requires multiple teams across the stack to work and collaborate closely.

In our study, we have focused on ***analog-aware hardware models***. Analog AI or in-memory computing AI [22] is an emerging AI hardware acceleration technology that recently has received considerable attention. Analog AI relies on non-Volatile Memory (NVM) in memory computing technology which eliminates the Von Neumann bottleneck and allows preforming computations directly in memory. This is accomplished by mapping deep neural networks (DNNs) to analog hardware and storing the neural network's weights as a property of the NVM material itself. By eliminating data movement altogether, this AI approach can offer speedup and energy efficiency for AI workloads.

To the best of our knowledge, this is the first work to address the gap of documenting AI hardware efficiency metrics and details during the training and inference of an AI model. We leverage the FactSheets [2] framework which provides a methodology [21] that can be adapted to multiple consumers and use cases. The basic premise is that a team identifies the relevant documentation for consumers and producers, and through an iterative approach, identifies and creates specific documentation needs for each role in the lifecycle of the AI model. Once a candidate FactSheet is created, it is evaluated with consumers to identify gaps. Those gaps are addressed and the process repeats until all documentation needs are met.

In this paper, we conducted a study centered on the FactSheets methodology to address these gaps and to identify the documentation needs for Analog AI models. Specifically, we answer following research questions:

- RQ1 What are the Analog AI documentation needs specific to different roles?
- RQ2 What are the new Analog AI facts that need to be documented in this space?
- RQ3 Did the created Analog AI documentation sufficiently meet the needs of the various roles?

The remainder of the paper is organized as follows. In Section 2 we present related AI documentation work. In Section 3 we provide details on Analog AI and FactSheets. Section 4 describes how we designed our study and the various phases of our methodology. In Section 5, we present the results of the study. Concluding remarks are provided in Section 6.

2 RELATED WORK

Several research efforts have emphasized the need to capture the clarity and transparency of machine learning models. Drawing on practices from other fields, the authors of [13] explore building a dataset in a way that provides greater transparency about the reasoning behind the addition of each feature. The authors argue for a framework for transparency and accountability of datasets, similar to what is done with models. Improving dataset development documentation, the authors state, will address the concerns of transparency, accountability and visibility. In [10] the authors' focus is also on datasets as in [13]. The authors propose 'datasheets for datasets', similar to the detailed datasheets used in the electronic industry, with the intention of providing transparency and accountability. Similarly, the 2018 paper on Factsheets [2] proposes a FactSheet for AI Services with details on use, performance, safety, and security, risk, etc. The authors reference supplier's 'declarations of conformity' (SDoCs) used in several industries as a model for FactSheets. In [12], the authors introduce the 'Dataset Nutrition Label' as a diagnostic tool that provides information, such as a framework to address and mitigate some of the challenges to provide critical information (provenance, statistics, metadata, variables, and other information contained in Table 1[1] to specialists at the point of data analysis.

The goal of the research in [15] is to define standard documentation practices across the field of natural language processing (NLP) tools. The authors state that developing templates that are easy to use is challenging due to the fact that the people involved in NLP come from a wide variety of of backgrounds, skills, incentives, etc. This is similar to what we see in our work on AI FactSheets for analog hardware. We shate the same view of the the authors that best practices for documentation have seen no widespread adoption. This have motivated them to focus on common, easy to use templates which they hope can drive wide adoption. The focus in this paper is the Hugging Face (HF) dataset and the GEM benchmark. The authors distinguish between a 'data card' for datasets and a 'model card' for model documentation. In a second paper from the NLP domain [4], though specifically made for NLP researchers, 'Data Statements' are meant to "address critical scientific and ethical issues that result from the use of data from certain populations in the development of technology for other populations.

The research in [3] is focused on developing a method for a quantifiable metric that can rank overall, the transparency of the process pipelines used to generate AI systems. The authors use the literature on supply chain visibility, from the basis of a focus organization, as their starting point. The authors do a review in Section 2 on documentation for both datasets and ML models, discussing Model Cards, Datasheets for Datasets, as well as standards from other industries, such as Bill of Materials. The authors use the VIS (Visibility Quality Index) as a measure for determining the overall visibility of the model. They demonstrate the the use of VIS on models from the PyTorch Hub and ModelHub. The method demonstrates some gaps in the chosen models, as well as good visibility due to high quality documentation around the training and architectures.

Nurrahmi et al. focuses on customer product reviews, and techniques to automatically extract features, and classify them [17]. The authors use two methods: CSR (Class Sequential Rule) for feature product extraction and Opinion Lexicon for feature opinion classification. Using the CSR method, the authors describe the process to extract features, and save these into a database. Opinion Lexicon classifies opinions into positive or negative, typically based on adjectives. The authors demonstrate their techniques on five amazon.com reviews and demonstrate their performance based on

[1] Holland et al. 2019, Table 1

precision, recall and F-score. By applying CSR, Opinion Lexicon, and post processing analysis, they were able to improve the scoring. The FactSheet methodology, at this point in time, uses simpler techniques to classify the results of the interviews. As the body of work grows, it will be useful to develop a method for classifying opinions of the interviewees, and in general developing automated techniques for the overall process.

In our related work search, we have not found any specific papers related to AI Hardware documentation. There are two papers that have proposed methods of device evaluation [6]. These authors take the approach of outlining the set of requirements they believe necessary for sensor technologies in medicine, and propose a framework for putting them into practice. This approach is similar to our AI hardware methodology. In [9] the authors propose a code framework to manage documentation of IT projects. The framework is divided into five views to accommodate the roles typically found in an IT organization (development, sizing, application, security, architecture). The authors also provide a comprehensive related work section on existing IT tools and frameworks. As we are doing with AI hardware documentation, the approach is to tailor the sections to each stakeholders interest and role in the work, so they can focus on the area that is relevant to their particular work items.

3 BACKGROUND

In this section we provide introductory background on the concept of Analog AI, the participant roles associated with implementing the AI model in Analog AI hardware, and finally, the concept and purpose of the FactSheet methodology.

3.1 Analog AI

Analog compute provides three main benefits over digital computing:

(1) **High Efficiency:** Analog AI removes the requirement of moving data, by using the analog memory elements to store and compute neural network weights, thus eliminating the von Neumann bottleneck.
(2) **Speed of Operations:** Low latency due to computation in place eliminates data movements and brings extreme high performance, making it suitable to compute the hundreds of thousands of operations occurring in parallel during AI training and Inference, which provides great advantage in particular to edge-AI computing, where constraints on model's size, memeory, and speed exist.
(3) **Low Power:** Since computation is accomplished without moving large amounts of data across multiple devices and buses, and only modifying small electrical currents, power consumption is low compared to digital devices.

An in-memory computing chip typically consists of multiple crossbar arrays of memory devices that communicate with each other. A neural network layer can be implemented on (at least) one crossbar, in which the weights of that layer are stored in the charge or conductance state of the memory devices at the cross points. Figure 1 illustrates a crossbar array and how the weights are mapped to it. Usually, at least two devices per weight are used: one encoding the positive part of the synaptic weight and the other encoding the negative part. Figure 1 shows a DNN layer and how to map computations from input x from the upstream neurons to a downstream neuron y. A vector-matrix multiplication operation first multiplies each element in an x vector with the corresponding weights, then the multiplication results are summed over all elements. f is a nonlinear activation function that is required to ensure that different layers in a DNN do not effectively become one linear layer. We first map each element in x_i to the word-line rows M of a crossbar array. Each weight $w_i j$ is then mapped into two devices with analog conductance, $G+$ and $G-$, to be able to represent both positive and negative weights. The matrix multiplication and summation operations are performed using the laws of physics (Ohm's law and Kirchhoff's law). We provide inputs on the left and in parallel several neurons are doing all the necessary math with natural summations of current. The propagation of data through that layer is performed in a single step by inputting the data to the crossbar rows and deciphering the results at the columns. The results are then passed through the neuron nonlinear function and input to the next layer. The neuron nonlinear function is typically implemented at the crossbar periphery, using analog or digital circuits.

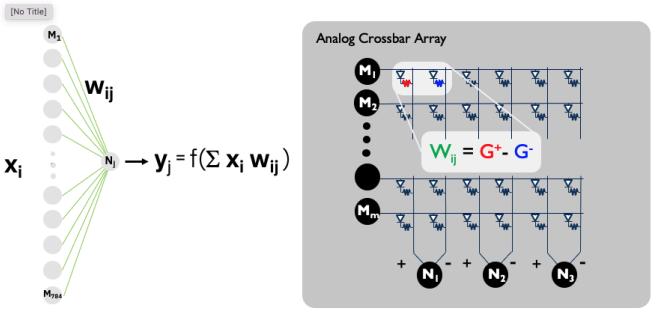


Figure 1: Mapping a Neural Network Layer to an Analog Crossbar Array

3.2 AI Hardware Lifecycle: Roles of the Participants

Building a hardware-aware model or Analog AI model in particular is a cross-functional endeavor and multiple roles are engaged in Analog AI model development. An example lifecycle from this team is provided in Figure 2. The goal of the figure is to provide the flow of the hardware-aware AI development lifecycle including Analog AI, and the participant roles in the flow. Note how different roles drop in and out of the development process. This situation makes it difficult for any one person to track all the information and assumptions related to building the model. Furthermore, each role has their own information needs which is often dependent on the work of others. This complex creation process justifies the use of the FactSheet as a complete information repository for all aspects of the model lifecycle.

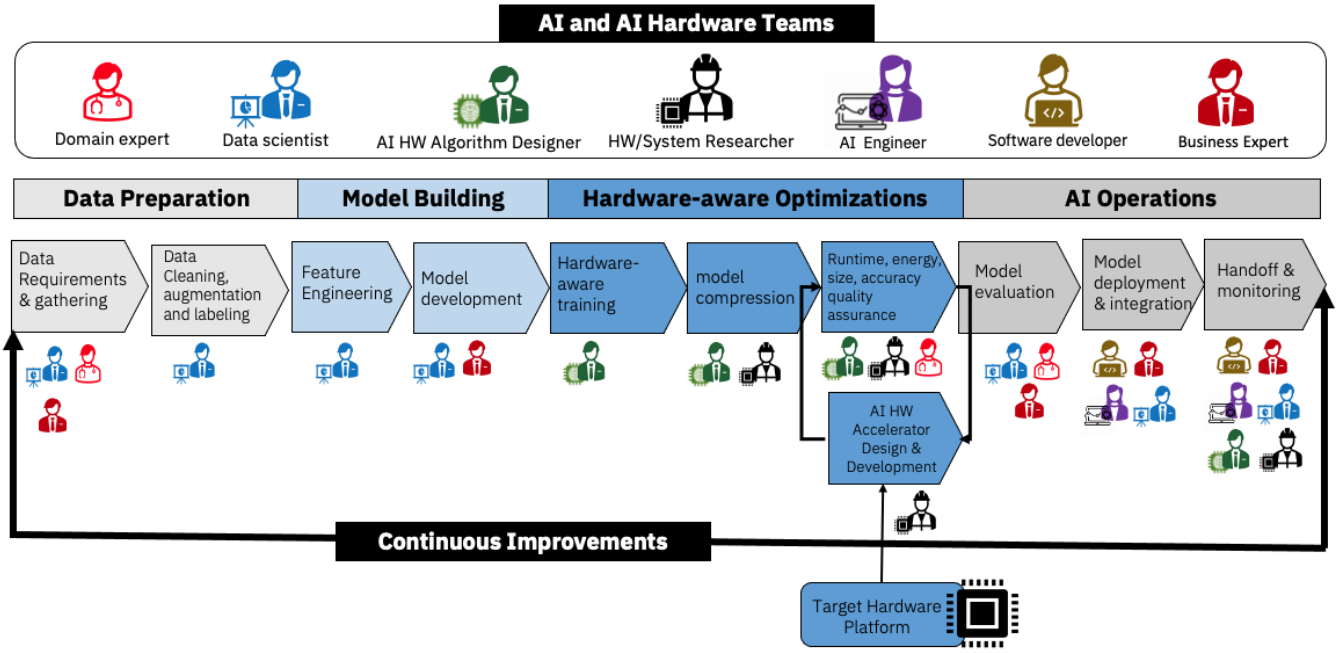


Figure 2: Lifecyle of a Hardware-aware Model Building Process

3.3 FactSheets

A FactSheet aims to document all relevant information (*facts*) about an AI model or service [2]. Provided with FactSheets is a human-centered methodology to guide the creation of a FactSheet [21]. At the methodology's core, there are two groups involved: consumers and producers. Consumers represent persons who have one or more use cases for AI documentation. Consumers can exist within the teams that build models or outside of them. In contrast, producers are the people who know the information about the model. There are of course instances of producers who are also consumers, especially within a model development team. Since AI model development involves multiple roles entering and leaving at different stages of the development lifecycle, the information that needs to be captured tends to exist across multiple roles [19]. To address this concern, the FactSheets methodology provides guidance on how to identify what is the relevant information to capture for a consumer's use case and how to engage with producers to get that information.

One of the reasons we chose to use FactSheets is that they are inherently adaptive to the domain and needs of interest. Instead of recommending a specific set of information to capture for use across all models, the FactSheets methodology is designed to elicit the documentation needs for a particular model. Given that the documentation needs for Analog AI are not well known, this approach is preferable. The methodology provides a seven-step process summarized in the steps below. Complete details can be found in [21].

(1) **Know Your FactSheet Consumers**. Understanding the information needs of FactSheet consumers is the first and most important task. In this step, the needs of consumers are identified through via discussions with them and recorded for use later.
(2) **Know Your FactSheet Producers**. Some facts can be automatically generated by tooling. Others require a human to provide them. Understanding which producers have the information needed is the key to this step. This steps identifies which information is available and which is not (and thus, may need to be found and captured).
(3) **Create a FactSheet Template**. The information gathered in the first two steps leads directly to the creation of a FactSheet template. A FactSheet template will contain what can be thought of as questions or, alternatively, the fields in a form. Each individual FactSheet will contain the answers to these questions.
(4) **Fill In FactSheet Template**. This step is where the creators of the FactSheet template attempt to fill it in for the first time. This step allows creators to informally assess the quality of the template itself by reflecting on what was learned about both consumers and producers, their skills, and information needs in the first two steps. This assessment serves as a checkpoint to highlight where improvements are needed and what information remains to be collected.
(5) **Have Actual Producers Create a FactSheet**. In this step, producers fill in the template for their parts of the lifecycle. For example, if there is a question in the template about model purpose, creators should find someone who would actually be providing that information and

have them answer the question. For example, ask a data scientist to answer the questions related to the development and testing of the model. Ask an AI engineer about deployment and so on. If the lifecycle is not that structured, the person responsible for most of the work could fill in this template,

(6) **Evaluate Actual FactSheet with Consumers**. In this step, an assessment is conducted to determine the current quality and completeness of the FactSheet produced in the prior step. If there are multiple consumer roles (not uncommon), it should be evaluated separately for each consumer role. To ground this assessment properly, each consumer should be asked to reflect and comment on how this FactSheet would actually help them perform their work or provide information to address their concerns. The results of this assessment feed into additional iterations until the quality is sufficient for all consumers involved.

(7) **Devise Other Templates for Other Audiences and Purposes**. This step returns to the beginning and is the only *defined* iteration in the methodology. There may be other consumers that need to be supported. Or, if the work so far has focused on the process of creating and deploying AI models, it would be worthwhile to consider consumers beyond this such as internal or external review boards or regulators, sales personnel, and end users or others affected by the product or service.

4 STUDY DESIGN

Following the recommendations of the FactSheet methodology, three researchers took on the role of the FactSheet team whose responsibility it is to identify the consumers' needs, create the documentation, and assess its usefulness. Concretely, the study consisted of three phases: *Need Identification* (matching steps 1 and 2 of the methodology), *FactSheet Creation* (steps 3-5), and *FactSheet Evaluation* (step 6).

Prior to the study, the researchers were given a head start since the Analog AI project team had already identified the six key consumer and producer roles to serve as participants for their documentation needs and for the study. A summary of those roles is presented in Table 1.

During the Need Identification phase, we interviewed 10 participants representing the AI Hardware or Systems Researcher, AI Hardware Algorithm Designer, Data Scientist, and Software Developer roles. We interviewed participants for up to one hour. During the interview, we first introduced participants to the concept of FactSheets and showed multiple examples. We then asked each participant which consumer role they most closely identified with. Following that, a semi-structured interview protocol was used to elicit potential use cases for that role (or roles), and their documentation needs for those use cases. After collecting documentation needs from consumers and additional discussion with producers, the researchers developed a template based on their findings.

Then, in the FactSheet Creation phase, the researchers collaborated with producers to fill out the template and create a FactSheet based on the consumer needs identified in the Need Identification phase. This process was iterative and shaped by what information producers were able to give. If a need was not able to be met, that Fact was omitted. Creating the FactSheet draft took the researchers approximately 3 weeks.

Finally, in the FactSheet Evaluation phase, one of the researchers led interviews to evaluate the created FactSheet. We interviewed 4 people representing 3 roles: HW/Systems Researcher, Data Scientist, and Software Developer. The other two roles were not available at the time of the study. In this phase, we first asked participants to score the FactSheet from the perspective of their use case on a three-point Likert scale along the following four dimensions Understandability, Relevance Presentability, and Completeness. These metrics are derived from prior work on qualitative metrics for assessing AI documentation quality [18]. More details about the metrics is provided in Section 5.3. We then asked participants to improve the FactSheet draft created in the prior phase by commenting on and modifying each of the facts one by one to better meet their needs. For example, participants were encouraged to point out any errors or confusing pieces of information, and suggest any additional text or links that could be included to provide more information to readers. This phase lasted approximately one hour per participant.

Since a stated goal of the study was to identify new documentation requirements, not to provide complete documentation for the model used in the study, we made the choice to only do one iteration of the FactSheet methodology. One iteration is sufficient to identify new needs, the corresponding documentation for those needs, and get a sense if those needs are indeed useful or not. Normally, as mentioned above, step 6 would be iterated upon until the FactSheet fully met the needs of its consumers.

4.1 BERT, the Analog AI Model Documented

The model that served as the basis for documentation was the Bidirectional Encoder Representations from Transformers (BERT) [24], used in the NLP domain. The BERT model is a transformer based model, that uses the technique of self-attention, weighing the significance of each part of an input similar to how a human may do this in cognitive attention, rather than sequentially evaluating each input, as was done previously in deep learning. Due to the size of BERT, in terms of the number of layers, parameters, and the large datasets used to train BERT, it can take several days and a large amount of compute power to train the model. The model can then be fine-tuned for the particular task.

In our case, we trained BERT on the AiMOS Supercomputer, located at Rensselear Polytechnic Institute, Troy, New York [8], in a hardware-aware environment. 'Hardware-Aware' means that the training understands that the BERT model is being trained with analog devices, and must take into consideration the hardware's design and material specifications [14]. When performing inference on Analog hardware, direct transfer of the weights trained on a digital chip to analog hardware would results in reduced network accuracy due to the non-idealities of the non-volatile memory technologies used in the analog chip. For example, a common non-ideality of Phase Change Memory is the resistance drift. This drift is due to structural relaxation of the physical material composing the memory and causes an evolution of the memory resistance over time,

Role	Description	Tasks
Hardware (HW) or Systems' Researcher	Pioneers the research on hardware and systems (defines 'what' is the technology)	- Compare and simulate with existing solutions on different dimensions - Design and build AI hardware accelerators - Integrate the accelerators in Systems on Chips (SoC) for various form factors (cloud, edge, embedded, etc.)
AI Hardware (AIHW) Algorithm Designer	AI practitioner/researcher with a solid interest in the impact of existing or emerging HW on the performance of the models	- Research and experiment applications of the hardware - Integrate and interact with different domains - Interested in the intersection of neural network architecture, computer architecture, systems design, materials, and hardware - Develop algorithms and techniques that take advantage of the hardware characteristics and specialize models to perform well on the hardware - Understand the concepts and new possibilities of the technology and the underlying code
Data Scientist	Develop neural networks and machine learning models that take advantage of the technology	- Have little knowledge of the underlying hardware. Rely on high level AI frameworks such as PyTorch and TensorFlow for training and inference - Develop/adapt networks and models that take advantage of the hardware characteristics at a macro level - Benchmark and validate models
AI Engineer	Integrates the technology into production environments	- Integrate the new AI/ML model solutions on existing pipelines - Explore robustness and interoperability of the technology for scalability
Software Developer	Develops or adapts software for producing results in a specific business context	- Compose existing software applications with the AI/ML models - Assess the performance benefits and advantages of the new hardware
Business Expert	Collaborate and partner with AI and Machine Learning solutions to analyze, measure, and refine business and maximize the ROI	- Assess the technology, from a high-level view and how it affects the business - Explore technology and its potential applications in a specific business context - Evaluate the benefits and competitive advantage of the technology in a business context - Influence and follow trends on software and technology

Table 1: A list of the roles in the AI Hardware Lifecycle that were identified before the study began.

which translates in a change in the weight stored in the memory array. This would eventually result in a decrease network accuracy over time. What one can do is to train the network on digital accelerator but in a way that is aware of the hardware characteristics that will be used in the inference pass, we refer to this as Hardware Aware Training (HWA). For instance, using the non-volatile memory (NVM) crossbar, one must consider parasitic resistance, source resistance, interconnect parasitics, process variations, noise, etc. These non-idealities introduce errors into the computations, and must be taken into account during computation. nce the model is trained, the inference compute is then accomplished on Analog AI hardware, as a hardware-aware model.

4.2 Study Participants

Over the two interview phases of the study, we interviewed 12 participants total. 11 were IBM employees working in Analog AI (2 HW/Systems Researchers, 2 AIHW Algorithm Designers, 4 Data Scientists, 2 Software Developers/Architects and 1 AI Engineer). We also interviewed an additional HW/Systems Researcher who is not an employee of IBM. Two of the participants participated in both the Need Identification and FactSheet Evaluation phases. Unfortunately, no Business Expert participants were available at the time of the study.

Details of the study participants are presented in Table 2. We tag each participant with a short code to make recalling roles easier. For example, 'P1-hw' indicates that participant 1's role is AI Hardware or Systems Researcher. We also identify which phase of the study that participants were involved in, 1 representing the FactSheet Creation phase and 3 representing the FactSheet Evaluation phase.

5 RESULTS

In this section, we discuss the interview results and the insights that were derived from applying the FactSheet methodoloy to the domain of Analog AI models.

5.1 RQ1: Roles' Analog AI Documentation Needs

Different roles identified different use cases for AI documentation. Recall that in the Need Identification phase of the study, we worked with participants to identify their use cases and what documentation needs would help them address their use case. These results are presented in Table 3. (Note that only participants who participated in that phase will have information here.) Looking across the rows of the table suggests that some information is nearly universally needed across roles such as the Training and Testing Methodology and Model Parameters, but how and in what context they wanted to see those Facts differed. We highlight these differences by role next.

For AI Hardware or System Researchers, their documentation needs centered around how the model was built and its hardware performance characteristics. For example, P2-hw discussed the need to understand *"[the model's] energy consumption in various states,"* and both P1-hw and P5-hw wanted to understand how analog hardware choices affected performance. This was reflected in their desire for information about the model's architecture, training/testing process, and hardware performance. AI Hardware or System Researchers were the only role to request information about Inference Methodology which explains the post-training process where the model can be used to make predictions against new, previously unseen data.

P#	Phase	Job Role	Years of experience	Gender	Role
P1-hw	1 and 3	Manager, Analog AI for Deep Learning	10	F	HW or Systems Researcher
P2-hw	1	Phase Change Memory Researcher	10	M	HW or Systems Researcher
P3-alg	1	AI Hardware Algorithm Designer	11	M	AI Hardware Algorithm Designer
P4-alg	1	Analog AI Algorithms Researcher	12	M	AI Hardware Algorithm Designer
P5-hw	1 and 3	Associate Professor and Material Scientist	9	M	HW or Systems Researcher
P6-data	1	Manager, Multilingual Natural Language Processing	19	M	Data Scientist
P7-data	3	Research Scientist, Multilingual Question Answering	11	F	Data Scientist
P8-data	1	Senior Technical Staff Member (STSM)	10	M	Data Scientist
P9-data	1	STSM, Chief Architect, Watson Core Language	13	F	Data Scientist
P10-dev	1	Manager, Emerging Solutions Development	19	M	Software Developer
P11-eng	3	STSM, NLP, Distributed Deep learning	24	M	AI Engineer
P12-dev	1	Senior Software Engineer	16	M	Software Developer

Table 2: Survey Participants. 'Phase' refers to which phase of the study the participant was involved in (Need Identification = 1, FactSheet Evaluation = 3). No participant was involved in phase 2 of the study.

Facts Needed	P1-hw	P2-hw	P5-hw	P3-alg	P4-alg	P6-data	P8-data	P9-data	P10-dev	P12-dev
Overview	✓		✓		✓			✓		
Purpose	✓	✓	✓		✓	✓	✓	✓		✓
Setup Instructions	✓	✓							✓	✓
Model Architecture	✓	✓		✓	✓		✓		✓	✓
Model Parameters	✓	✓	✓	✓	✓	✓	✓	✓		✓
Training/Testing Methodology	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Inference Methodology	✓	✓								
Accuracy			✓		✓	✓		✓		
Prediction Time	✓		✓		✓				✓	✓
Hardware Performance	✓	✓	✓		✓	✓		✓		
Robustness		✓					✓	✓		
limitations/challenges with the system			✓				✓			

Table 3: The facts that participants stated were useful for their particular roles.

AI Hardware Algorithm Designers instead wanted information to help them understand comparisons across models. In P3-alg's case, he wanted to make comparisons between Analog AI models and their GPU-based equivalents. He said his job is to, *"take a model that has already been trained and see if [it] can do equally as good as [Analog AI]."* P4-alg wanted to *"use existing models ... we take them and see if they are robust against analog noise."* Although their specific needs overlapped somewhat with the AI Hardware or System Researchers, how they wanted to use that information differed.

Similar to the AI Hardware Researcher role, the Data Scientists were interested in documentation related to the model's creation, evaluation, and improvement. However, they were forward-thinking in their use case identifying how this information is easily forgotten or how it may be used by others in the future. P6-data said, *"In [my organization] we have 100+ models and I always forget what domain they're trained on... the reason I'd be looking is to remind myself what they were trained and tested on."* P9-data explicitly called out how something like FactSheets is meant for others to consume. She said, *"it wouldn't be for yourself, it'd be for others to use once you give them a model."*

Finally, a Systems Architect or Developer's needs are focused on deployment or integration with other tools, such as an API or inference hardware. Because of this, the Systems Architect role is more interested in the higher-level overview of the model as opposed to any specific facts related to Analog AI hardware. P12 said: *"[The] most important decisions about parameters and stuff have already been made... [I'm] interested in the infrastructure supporting a model."* Correspondingly, they have some of the most limited needs from the FactSheet, relative to the other roles.

5.2 RQ2: How Documentation Differs in Analog AI

We found that the Analog AI Hardware domain shared many of the same documentation needs reported in prior work on AI documentation, such as details about the model's architecture, details about training and test data, and model performance measures [2, 10, 16, 19]. However, we found that new specific documentation needs for Analog AI. Examples from the FactSheet can be found in Table 4. As an example of some of the details specific to this domain, note that while the *Training Methodology* section includes information about the Learning Rate of the BERT Model (a detail that's usually shown in documentation methods), there's also information that is unique to Analog AI models - Weight/Read Noise, ADC/Output Noise, Input and Output bounds, etc. Additionally, entirely new facts that are specific to Analog AI were added - *Inference Methodology* and *Parameters and Inference Attributes*.

Parameters and Inference Attributes

Total Number of weights/ops/Multiply Accumulate (MAC):

- 110 million parameters/weights
- Total number of operations will depend on the length of the input sequence (SL) as well as the batch size
- Analog AI accelerator prototypes are being designed for optimum performance for SL up to 128 and low batch sizes (1-4)
- For a SL of 32 (Batch Size = 4) total number of ops is approximately. 21.84B (out of which 21.74B are operation performed using Analog Non-Volatile Memory (NVM) arrays or tiles). Total execution time is approximately, 0.19msec

Reuse Factor for each node/layer:

- The same weights are not re-used in this dataflow.

NVM Array Tile Size:

- Each array tile has 512x512 synaptic weights. Synaptic weights are represented as a differential pair of conductances.

Power and Chip Area Estimates:

- Average Power (Chip level): 3.13W
- Effective area: 780.4 sq.mm

Training Methodology

The model was trained using IBM's AiMOS supercomputer in a Hardware-aware (HWA) fashion before being deployed on Analog Accelerator hardware for Inference [2]. Hardware-aware training makes the model resilient to non-idealities of NVM devices as well as peripheral circuitry and leads to improved inference accuracy with Analog compute. Specifically, this is done by starting with a floating point (FP) trained baseline model and re-training it further with the addition of device circuit non-idealities in the forward pass only. The backward pass and weight update are similar to traditional FP training. These can include inputs such as programming (or weight) noise, read noise, conductance drift, ADC (or output) noise, input output bounds. The following parameters were used for HWA training of BERT-base model:

- Weight Noise (Std-dev): 1.0
- Read Noise: 0.0175
- ADC/Output noise: 0.04
- Drift model from PCM devices
- Input and Output bounds: 3.0/10.0
- Learning rate: 5E-5 to 1.5E-4

This hardware-aware training produced a model with a validation accuracy of 80.6 pct (GLUE average) (compared to state-of-the-art at 81.2 pct from BERT paper).

[2] 2020 Nature Communications. Accurate deep neural network inference using computational phase-change memory

Inference Methodology

- Number of non-idealities can impact the weight programming and stability such as programming noise, read noise, conductance drift and temperature dependent variations. These parameters can be dependent on the weight (and conductance) distributions and are thus model and network dependent.
- Some of these non-idealities are introduced during the hardware-aware training methodology described above to make the model more resilient and to improve inference accuracy.
- Conductance variations post programming such as due to drift and temperature are mitigated by algorithmic corrections to minimize the impact on inference accuracy.

Table 4: A Section of the New FactSheet. The 3 facts selected to be shown here (Parameters and Inference Attributes, Training Methodology, Inference Methodology) contain information unique to the Analog AI domain.

Table 4 shows a section of the FactSheet that had a high proportion of new Analog AI fact content. Table 5 summarizes the FactSheet template and describes the other new facts we identified. The fully completed FactSheet can be viewed in the Appendix.

As discussed in Section 3.2, The novel contribution of Analog AI isn't necessarily a more accurate model, but rather the same models with better hardware performance (speed, memory usage, etc.). The facts listed here are important because they address factors that show how well or poorly a model runs on the Analog AI hardware. A majority of these facts can be thought of as an answer to two questions: (1) "For a given model, what are the parameters of the model, and the hardware it was trained on?" and (2) "For a given model, what was its overall performance in comparison to models trained on other hardware setups?" Note that these are very similar to the types of questions that a researcher only concerned with model performance would ask of an AI model, but a correct answer here must also include information about the Analog AI hardware. FactSheets in this new domain reflect this; there's a strong need to first know what the performance of a model is across different hardware and software metrics, and secondly how this performance was achieved. Several of these documentation methods include information about the model's bias/robustness, and any other ethical considerations that must be made when using the model. While our study did not place a great deal of importance on these considerations, it is certainly plausible that there will be more of an emphasis on them as more Analog AI FactSheets are created.

To verify that these new pieces of information in Analog AI FactSheets are indeed important, as well as that the differing use cases discussed in 5.1 are sufficiently met, we continue to the evaluation of our FactSheet.

5.3 RQ3: Fact Quality

Given that the use cases varied from role to role, one representation of the FactSheet may be insufficient for their needs. To address where gaps existed, we invited participants from each role to evaluate the quality of the FactSheet from the use cases identified from the Need Identification phase of the study.

Recall that we used four dimensions to evaluate the FactSheet's quality. *Understandability* measures if the fact is written in a manner that is comprehensible, unambiguous, and at the correct level of vocabulary. *Relevance* measures if the fact's information is on-topic. *Presentability* measures if the visual and data representation of the fact is as expected. Finally, *completeness* measures if the fact has all the information necessary. Together, these quality dimensions shed light on where a fact may be deficient and provide a structured way to have participants assess a fact's quality. Participants rated each fact using the four quality dimensions. Each quality dimension used three-point scale. 'Low' indicates that the fact *did not* meet the criteria for the dimension. 'Medium' indicates the fact *partially* met the criteria, and 'High' indicates that the fact *completely* met the criteria for the dimension.

Table 6 summarizes all the low and medium scores from the evaluation. (Blank cells indicate high scores, and are omitted for clarity.) Looking at the summary of the results, two key findings emerge. First, scores for the dimensions of understandability, relevance, and presentability earned high marks whereas all the participants found the FactSheet to be incomplete. Second, P11-eng's was dissatisfied with the FactSheet along all four dimensions, and we explore why later in this section. Taken together, this suggests that the first draft FactSheet created for this study is on the right track, but in need of additional iteration.

Fact	Description
Overview	A summary of the model and its accompanying FactSheet.
Purpose	The tasks that the model was trained for.
Model Architecture	In addition to information about the structure of the Neural Network, Model Architecture is also meant to capture information about how this Neural Network is replicated on the cross bar arrays of in-memory computing chips. Some additional considerations when collecting information for this fact: - In terms of layers and nodes, how large is the model? - Is the model customizable? - What is the exact classification method being used?
Model Parameters	In addition to the parameters most commonly associated with Neural Networks, some Hardware-specific parameters such as Reuse Factor for each node/layer and Non-Volatile Memory (NVM) Array Tile Size are expected to be included.
Training/Testing Methodology	In the training process, data is processed by the DNN. After each pass, the error in the prediction (or inference) of the DNN is evaluated, and updates are made to update the weight (strengths) of the neuron connections. This process is repeated until the desired accuracy is reached and the model can be said to 'trained'. Training requires large amounts of compute power, and can take hours or days to complete.
Inference Methodology	Inference is the post-training process, in which the model can be used to make predictions against new, previously unseen data. Trained models are essential in edge application, for example, where power and compute are limited. Examples of this are traffic cameras, self-driving cars, chat bots, etc. [5, 23].
Training Data	The data used to train the model, and its basic characteristics. Any additional notes or consideration should be listed here.
Characteristics of the Training Hardware and Inference Hardware	The properties of the hardware used during the Training and Inference processes.
Test Data	The data used to test the model, and its basic characteristics. Any additional notes or consideration should be listed here.
Performance	The goal of recreating these models on Analog Hardware is to achieve the same accuracy while improving the performance of the hardware being used by the model. Here, different hardware metrics such as Prediction Time, Latency, Energy Consumption per prediction and Throughput are usually expected. Some additional considerations when collecting information for this fact: - The Size of the Model on Disk on disk - The Size of the Model on an In-Memory Computing Chip - Throughput on either a CPU or CPU + GPU in kHz/second

Table 5: The FactSheet template for the BERT model.

Fact	Understandability				Relevance				Presentability				Completeness			
	P1-hw	P5-hw	P7-data	P11-eng	P1-hw	P5-hw	P7-data	P11-eng	P1-hw	P5-hw	P7-data	P11-eng	P1-hw	P5-hw	P7-data	P11-eng
Overview		med				med						med		med	med	med
Purpose		med								med						
Model Architecture				med				med					med			med
Parameters and Inference Attributes				med			low			med		med	med			med
Training Methodology				med			low							med	low	med
Inference Methodolgy			low	med			low	med			low	med	med	med	low	med
Training Data				low				med				low		med	low	low
Hardware Characteristics				low				low				low	med		med	low
Test Data				low				med				low		med	low	low
Performance Metrics				low						med		low	med		med	low

Table 6: Participants' responses from FactSheet evaluation. Blank cells indicate a high score and have been removed for clarity.

Further analysis of the data revealed that when participants noted some missing information (low completeness score), they tended to score the other values lower as well. If a FactSheet consumer does not see all the information they need for their use case, it follows that the perceived understandability, relevance, and presentability may decrease as well. Although at first glance, this may seem like an unfavorable result, it is in fact a signal that the FactSheet methodology is working as intended. Recall that we only did one iteration of the FactSheet creation and evaluation process. Ideally, this process would repeat, as indicated by step 6 of the methodology, and future versions would incorporate feedback from the users to gather or present the information in the necessary way. However, given that the goal of this paper was to use the FactSheets methodology to understand new documentation needs in this domain and have some indication that documentation based on those needs is useful, we consider these findings to be sufficient.

The relatively lower scores of P11-eng come from his perspective that his documentation needs in a FactSheet had not been met.

This participant that we interviewed was already very familiar with Analog AI and its functionality, as well as FactSheets for regular AI models; he felt that there was more Analog AI-specific information that could have been included for some facts but wasn't, which left him confused on some of the technical details of the model. During the evaluation he said *"[Including] a link to a research paper, or to the Analog AI website, would be good [for this FactSheet]."* One likely cause for this was that due to scheduling issues, we were unable to include any AI Engineering participants during the Need Identification part of the study. So when asked to evaluate the FactSheet for his use case, the FactSheet ended up missing critical information, as indicated by the lower scores in P11-eng's Completeness column in Table 6. Had we been able to keep iterating on the FactSheet, we believe that his needs could also be met.

6 CONCLUSION

This study was the first to investigate the documentation needs of models built on Analog AI hardware. In particular, we made the following contributions:

(1) Identification of role-specific use cases for AI documentation in this domain.
(2) Documentation needs and examples for an Analog AI model.
(3) An evaluation of the usefulness of the created AI documentation across three roles.
(4) Techniques to improve the transparency of a complex AI model domain.

The FactSheets methodology proved successful in helping us shed light on the needs of this untapped and understudied domain and provided the early foundation from which other documentation and transparency issues can begin to be addressed in the emerging area of hardware-friendly AI models.

ACKNOWLEDGMENTS

We would like to thank the teams from both the IBM Research AI Hardware Center and Rensselaer Polytechnic Institute for making the AiMOS computing platform available to us. We conducted most of our experiments on the AiMOS computing platform. We are also deeply grateful for all the researchers who have participated in the interviews from the AI hardware team and the AI NLP team. We would also like to thank the National GEM Consortium for supporting this research.

REFERENCES

[1] Dario Amodei, Danny Hernandez, Girish Sastry, Jack Clark, Greg Brockman, and Ilya Sutskever. 2018. *AI and Compute.* https://openai.com/blog/ai-and-compute/
[2] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.
[3] Iain Barclay, Harrison Taylor, Alun Preece, Ian Taylor, Dinesh Verma, and Geeth Mel. 2020. A framework for fostering transparency in shared artificial intelligence models by increasing visibility of contributions. *Concurrency and Computation: Practice and Experience* 33, 19 (Dec 2020). https://doi.org/10.1002/cpe.6129
[4] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041
[5] Michael Copeland. 2016. What's the difference between Deep Learning Training and inference? https://blogs.nvidia.com/blog/2016/08/22/difference-deep-learning-training-inference-ai/
[6] Andrea Coravos, Megan Doerr, Jennifer Goldsack, Christine Manta, Mark Shervey, Beau Woods, and William A Wood. 2020. Modernizing and designing evaluation frameworks for connected sensor technologies in medicine. https://doi.org/10.1038/s41746-020-0237-3
[7] William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *CoRR* abs/2101.03961 (2021). arXiv:2101.03961 https://arxiv.org/abs/2101.03961
[8] Center for Computational Innovations at Rensselaer Polytechnic Institute. 2022. *Artificial Intelligence Multiprocessing Optimized System (AiMOS).* https://cci.rpi.edu/aimos
[9] Christophe Gaie, Bertrand Florat, and Steven Morvan. 2021. An architecture as a code framework to manage documentation of IT projects. *Applied Computing and Informatics* (2021).
[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2021. Datasheets for Datasets. arXiv:1803.09010 [cs.DB]
[11] Karen Hao. 2019. *Training a single AI model can emit as much carbon as five cars in their lifetimes.* https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/
[12] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677 [cs.DB]
[13] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. arXiv:2010.13561 [cs.LG]
[14] Vinay Joshi, Manuel Le Gallo, Simon Haefeli, Irem Boybat, S. R. Nandakumar, Christophe Piveteau, Martino Dazzi, Bipin Rajendran, Abu Sebastian, and Evangelos Eleftheriou. 2020. Accurate deep neural network inference using computational phase-change memory. *Nat Commun* 11, 1 (1 Dec. 2020). https://doi.org/10.1038/s41467-020-16108-9
[15] Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards. *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)* (2021). https://doi.org/10.18653/v1/2021.gem-1.11
[16] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency.* 220–229.
[17] Hani Nurrahmi, Warih Maharani, and Siti Saadah. 2016. Feature extraction and opinion classification using class sequential rule on customer product review. In *2016 4th International Conference on Information and Communication Technology (ICoICT).* 1–5. https://doi.org/10.1109/ICoICT.2016.7571891
[18] David Piorkowski, Daniel González, John Richards, and Stephanie Houde. 2020. Towards evaluating and eliciting high-quality documentation for intelligent systems. *arXiv preprint arXiv:2011.08774* (2020).
[19] David Piorkowski, John Richards, and Michael Hind. 2022. Evaluating a Methodology for Increasing AI Transparency: A Case Study. *arXiv preprint arXiv:2201.13224* (2022).
[20] IBM Research. 2022. *AI Hardware Center.* https://research.ibm.com/collaborate/ai-hardware-center/
[21] John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. 2020. A Methodology for Creating AI FactSheets. *arXiv preprint arXiv:2006.13796* (2020).
[22] Abu Sebastian, Manuel Le Gallo, Riduan Khaddam-Aljameh, and Evangelos Eleftheriou. 2020. Memory devices and applications for in-memory computing. *Nature nanotechnology* 15, 7 (July 2020), 529—544. https://doi.org/10.1038/s41565-020-0655-z
[23] Mary T. 2020. The difference between Artificial Intelligence, Machine Learning and deep learning. https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/The-Difference-Between-Artificial-Intelligence-Machine-Learning/post/1335666
[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

APPENDIX

BERT FactSheet

Overview

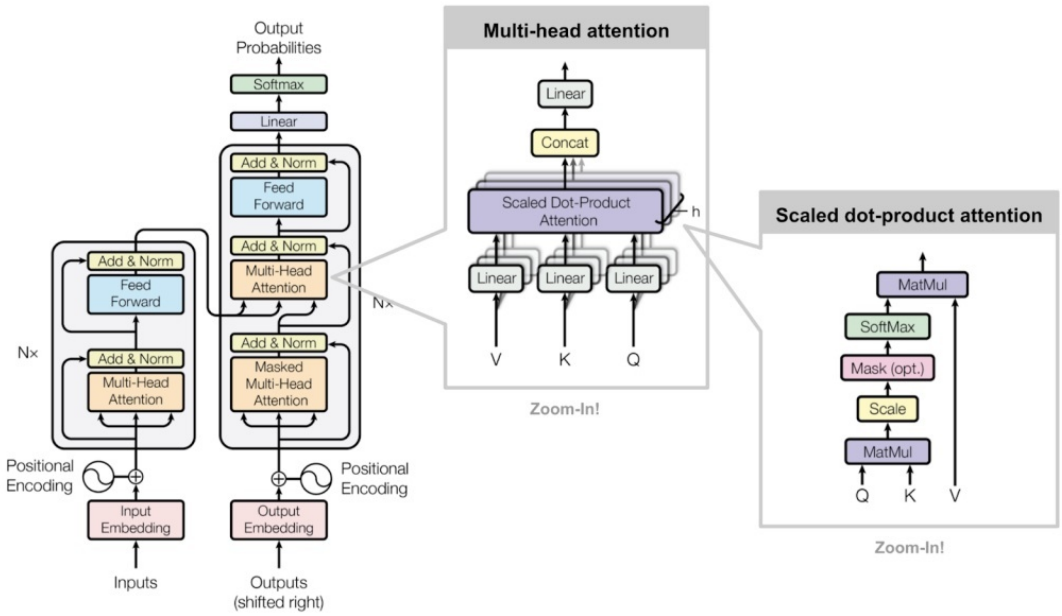
This document is a FactSheet accompanying the BERT model used in the IBM AI Hardware Center – with Analog AI Hardware. FactSheets aim at increasing trust in AI services through supplier's declarations of conformity and this FactSheet documents the process of training the Audio Classifier model as well as its expected results and appropriate use.

Purpose

BERT stands for Bidirectional Encoder Representations from Transformers. As the name suggests, this is based on the Transformer architecture. It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context. Pre-training is performed for two NLP tasks, namely, (a) Masked Language Modeling and (b) Next Sentence Prediction. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks. These can include sentiment classification, abstract summarization, question answering and conversational response generation to give a few examples.

Model Architecture

The figure below shows the nominal architecture of BERT models – these consist of many layers of Encoder layers. BERT Base (Cased or uncased) model uses: 12 layers (transformer encoder blocks) with 12 attention heads, and a total of 110 million parameters (neural network weights). We have used the hugging face implementation of BERT (https://huggingface.co/docs/transformers/model_doc/bert) which is based on the "Attention is All You need" Google paper [1].



[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc.
https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Parameters and Inference Attributes

Total Number of weights/ops/Multiply Accumulate (MAC)

- 110 million parameters/weights
- Total number of operations will depend on the length of the input sequence (SL) as well as the batch size
- Analog AI accelerator prototypes are being designed for optimum performance for SL up to 128 and low batch sizes (1-4)
- For a SL of 32 (Batch Size = 4) total number of ops is approximately. 21.84B (out of which 21.74B are operation performed using Analog Non-Volatile Memory (NVM) arrays or tiles). Total execution time is approximately, 0.19msec

Reuse Factor for each node/layer

- The same weights are not re-used in this dataflow.

NVM Array Tile Size

- Each array tile has 512x512 synaptic weights. Synaptic weights are represented as a differential pair of conductances.

Power & Chip Area Estimates

- Average Power (Chip level): 3.13W

- Effective area: 780.4 sq.mm

Training Methodology

The model was trained using IBM's AiMOS supercomputer in a Hardware-aware (HWA) fashion before being deployed on Analog Accelerator hardware for Inference [2]. Hardware-aware training makes the model resilient to non-idealities of NVM devices as well as peripheral circuitry and leads to improved inference accuracy with Analog compute. Specifically, this is done by starting with a floating point (FP) trained baseline model and re-training it further with the addition of device & circuit non-idealities in the forward pass only. The backward pass and weight update are similar to traditional FP training. These can include inputs such as programming (or weight) noise, read noise, conductance drift, ADC (or output) noise, input & output bounds.

The following parameters were used for HWA training of BERT-base model:

- Weight Noise (Std-dev): 1.0
- Read Noise: 0.0175
- Analog-Digital Converter (ADC)/Output noise: 0.04
- Drift model from Phase Change Memory (PCM) devices
- Input & Output bounds: 3.0/10.0
- Learning rate: 5E-5 to 1.5E-4

This hardware-aware training produced a model with a validation accuracy of 80.6% (GLUE average) (compared to state-of-the-art at 81.2% from BERT paper).

[2] 2020 Nature Communications. Accurate deep neural network inference using computational phase-change memory

Inference Methodology

- Number of non-idealities can impact the weight programming & stability such as programming noise, read noise, conductance drift and temperature dependent variations. These parameters can be dependent on the weight (& conductance) distributions and are thus model & network dependent.
- Some of these non-idealities are introduced during the hardware-aware training methodology described above to make the model more resilient and to improve inference accuracy.
- Conductance variations post programming such as due to drift and temperature are mitigated by algorithmic corrections to minimize the impact on inference accuracy.

Training Data

The model was trained on the GLUE benchmark dataset (8 tasks: RTE, MRPC, STS-B, CoLA, SST-2, QNLI, QQP, MNLI). It has a size from 2.5k (RTE) to 393k (MNLI). All tasks are classification, except STS-B, which is a regression task. MNLI has 3 classes and all others have 2 classes. Sequence length goes up to 128 for our purposes.

Characteristics of the training hardware and characteristics of the inference hardware

AiMOS supercomputer and Cognitive Computing Cluster at IBM Research (CCC) was used. Batch Size is optimized for GPU Memory.
For both training of the model and inference accuracy evaluations, a computing node with a single NVIDIA V100 GPU was used on the CCC cluster.

Test Data

The model was tested on the GLUE benchmark dataset (8 tasks: RTE, MRPC, STS-B, CoLA, SST-2, QNLI, QQP, MNLI). It has a size from 2.5k (RTE) to 393k (MNLI). All tasks are classification, except STS-B, which is a regression task. MNLI has 3 classes and all others have 2 classes. Sequence length goes up to 128.

Performance Metrics

- *Accuracy*: Most tasks use accuracy; MRPC and QQP use both accuracy and F1 score; CoLA uses Matthews corr.; STS-B uses Person/Spearman corr.

Task names	RTE	MRPC	STS-B	CoLA	SST-2	QNLI	QQP	MNLI	Avg.
Data set size	2.5k	3.7k	5.7k	8.5k	67k	105k	364k	393k	
Metrics *		(f1)	(Spearma)	(Matthew')					
BERT paper (digital)	66.40%	88.90%	85.80%	52.10%	93.50%	90.50%	89.20%	83.40%	81.23%
Analog AI	64.00%	88.62%	84.52%	55.65%	91.03%	89.01%	89.63%	82.12%	80.57%
* Metrics is accuracy, unless specified, e.g., f1, Spearma Co, or Matthew's Co.									

- ***Classification Time***: 0.19msec assumes a Sequence Length of 32
- ***Energy Efficiency***: 6589 (Inferences/s/W)
- ***How often do you have to reprogram the weights/refresh the model:*** Due to non-idealities such as conductance drift and retention loss, it is expected that models would need to be refreshed (re-programmed) on Analog HW every 4 weeks.