Qualcomm Response to NIST Request for Information

NIST seeks to understand the:

Status and plans regarding the availability, use, and development of AI technical standards and tools in support of reliable, robust, and trustworthy systems that use AI technologies;

Needs and challenges regarding the existence, availability, use, and development of AI standards and tools; and

Introduction

Qualcomm welcomes the opportunity to comment on NIST's request for information 84 FR 18490 (Docket Number: 190312229-9229-01) regarding the development and use of AI standards. With respect to standardization in AI, it is important that U.S. industry and the U.S. government understand that a consensus should be reached about the following technical topics: common AI terminology, relevant international AI standards, important use-cases for AI, industrial specialization of different types of AI techniques and a common Machine Learning neural-network workflow with associated reference stages.

Qualcomm provides information and feedback from an end-to-end (mobile devices to cloud) AI and ML platform perspective. We recommend emerging AI technical standards efforts relevant to our business, an example AI application and a recommended AI workflow to use as reference for technical and policy related discussion questions listed in the NIST RFI.

Overview

There has been a renaissance in AI capabilities due to breakthroughs in a specific AI sub-field known as Machine Learning (ML) using new reinforcement learning mathematical techniques to improve algorithms called Convolutional, Recurrent and LSTM Neural Networks (NNs). These technical improvements rapidly accelerated the adoption of AI technology in mobile-enable devices. The fluid state of AI development to prevent the development underscores the need to avoid developing standards that restrict progress.

The following list of technical standards are still at the start of their development cycle and NIST has an important role to play in their development by supporting US industry with relevant use-cases from various government departments (i.e. DoD, NOAA, Dept of Commerce, DOT, etc.), large, high-quality data sets and finally a venue to bring together industry, academic and government AI implementers to work on important common challenges.

Relevant AI Technical Standards and Open-Source Projects

Standard Name	Description and Value
ISO/IEC/SC42 – Artificial Intelligence	Broadest effort for AI Standardization. Serves as the focus and proponent for JTC 1's standardization program on Artificial Intelligence. Provide guidance to JTC 1, IEC, and ISO committees developing Artificial Intelligence applications.
SG 01 *Computational approaches and characteristics of artificial intelligence systems*	Different technologies (e.g., ML algorithms, reasoning etc.) used by the AI systems including their properties and

Standard Name	Description and Value
	characteristics. Study of existing specialized AI systems (e.g., NLP or computer vision) to understand and identify their underlying computational approaches, architectures, and characteristics. Study of industry practices, processes and methods for the application of AI systems. This is an important area for technical standardization
WG 02 AI *Big Data* Functions ISO/IEC 20546:2019 Big Data Information technology • Overview and Vocabulary • Use cases and derived requirements • Big data reference architecture • Big Data Standards roadmap • Framework and application process • Big Data reference architecture	Machine learning makes use of vast data sets and the standardization of Big Data for Artificial Intelligence is important.
WG 03 AI *Trustworthiness*	Approaches to establish trust in AI systems through transparency, verifiability, explainability, controllability, etc. Engineering pitfalls and assess typical associated threats and risks to AI systems with their mitigation techniques and methods. Approaches to achieve AI systems' robustness, resiliency, reliability, accuracy, safety, security, privacy, etc.
WG 04 AI *Use cases and applications*	Identify different AI application domains and the different context of their use. Collect representative use cases to provide best practices/guidance on domains, drive liaisons and garner insights from applications and suggest application area focus to SC 42.
IEEE P7006 - Standard for Personal Data Artificial Intelligence (AI) Agent	This standard describes the technical elements required to create and grant access to a personalized Artificial Intelligence (AI) that will comprise inputs, learning, ethics, rules and values controlled by individuals.
Open Source Tensorflow.org	TensorFlow is an end-to-end open source platform for machine learning with a flexible ecosystem of tools, libraries and resources to build and deploy ML models and applications.
Open Source ONNX.ai	ONNX is an open format to represent deep learning models to enable AI developers to use different AI development tools.
Open Source Keras.io	Keras is a high-level NN API written in Python capable of running on top of lower-level machine learning modeling libraries. It enables fast experimentation and training of models.
Open Source caffe.berkeleyvision.org	Caffe is a deep learning framework designed for expression, speed and modularity.
Open Source pytorch.org and caffe2.ai	An open source deep learning platform that provides a seamless path from research prototyping to production deployment.
Open Source mlperf.org	A broad ML benchmark suite for measuring performance of ML software frameworks, hardware accelerators, Edge and Cloud platforms.

Background and Definition of Artificial Intelligence

The NIST defines AI Technologies as follows:

AI technologies and systems are considered to be comprised of software and/or hardware that can learn to solve complex problems, make predictions or solve tasks that require human-like sensing (such as vision, speech, and touch), perception, cognition, planning, learning, communication, or physical action. Examples are wide-ranging and expanding rapidly. They include, but are not limited to, AI assistants, computer vision systems, automated vehicles, unmanned aerial systems, voicemail transcriptions, advanced game-playing software, facial recognition systems as well as application of AI in both Information Technology (IT) and Operational Technology (OT).

Qualcomm Mobile, Compute, Automotive, Audio, IoT and Data Center/Server semiconductor hardware and software products all use the AI technologies defined above.

AI technology has been in development for the last 50 years and already has accumulated a wide body of theory and mature classical algorithms (Decision Trees, Naïve Bayes Classification, Least Squares Regression, SVM, Ensemble, Clustering, PCA, SVD, ICA). Artificial Intelligence refers to the generally broad category of algorithms and techniques that solves problems and tasks typically done better by human biological intelligence. Machine Learning (ML) algorithms and techniques *learn* from past experience data with respect to some class of task and *automatically* improve performance of this task based on specific performance measures.

Unlocking a Smartphone with Facial Recognition AI Example

Users want access to their data and phone features quickly and easily with the expectation of high-security authentication. Traditionally, this was done with a password or pin-code but that proved to be much slower and less secure than new biometrics techniques made possible with AI. Phones can now use 2D and 3D cameras to scan a user's face and provide convenient, fast and secure authentication using new advanced AI/ML algorithms. The AI workflow below illustrates how this can work and highlights the need for standardization at each stage to ensure privacy, accuracy, security and integrity.

Standard Reference Machine Learning Workflow

The following Figure 1 outlines a typical modern AI workflow. The 7 stages each have specific data storage requirements which will benefit from standardized interfaces and policy protocols. All standards listed above are applicable to this workflow.

The biometric facial authentication problem is complex and can be broken down into multiple phases:

1. Capturing a reference phone owner's facial information. This is done very infrequently with most people only face training their system once or twice during the ownership period.
2. Using that reference facial information to unlock the phone. Users typically unlock their phone between 20 and 30 times **per day**.

In each of the stages above, AI algorithms can use the workflow below to create neural network models to detect the location of a face, create posing and projected faces to allow different authentication positions for maximum unlock flexibility and then measure the facial parameters to create a unique digital signature that represents a specific user. The captured facial data structure, storage, security,

privacy and integrity can be standardized to allow a consistent and secure facial recognition applications and there are many existing standards to facilitate this effort.

It is important that AI technical standards recognize the importance of the physical location of the data at each stage to ensure security, integrity and privacy. Smartphone devices storing data locally have different security and related technical considerations compared to when they transfer the data electronically from the device to cloud-based storage servers. The ISO SC42 AI Big Data and many other biometric and security standards are applicable to general AI data systems.

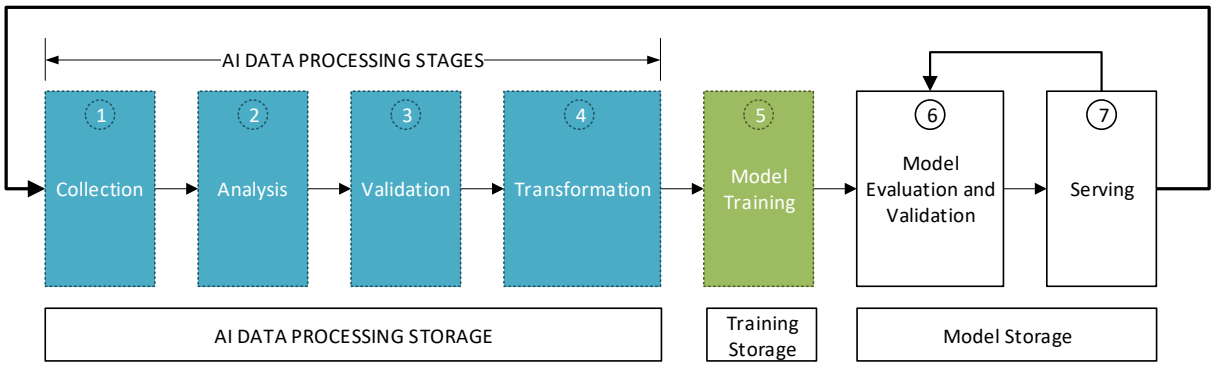

Figure 1 - AI Data Flow Processing

AI Workflow Analysis Descriptions

AI Data Processing Stages 1 through 4.

Applicable standards: ISO SC42 – WG02 AI Big Data, WG03 Trustworthiness, IEEE P7006 Personal Data Agents and all software based NN frameworks (e.g. TensorFlow, ONNX, etc.)

1. Data Collection Stage

Deep learning with neural-networks requires vast initial data sets collected from various sources such as a file, database, sensor and many other such sources. This initial data set cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data.

In our reference example, facial data gathered in this collection stage must come from the widest variety of sources to prevent skin tone, facial features, occlusions and other biases from affecting the end prediction results. However, this initial set will contain erroneous data, poor quality images and a host of other problems that must be corrected in the next stage.

2. Data Analysis Stage

The analysis stage is one of the most important steps in machine learning to ensure the final NN-model produces the most accurate predictions. This stage requires careful review and analysis of the raw data collected in Stage 1 to determine the state of the raw data collection and categorize each sample into specific areas for subsequent validation processing. This stage is purely analytical and no modifications are performed on the raw data.

In our example, real-word raw facial data may have the following problems:

2.1. **Missing Data**: Specific facial features may be missing (e.g., different racial and ethnic faces, tattoos, birthmarks, eye-patches, make-up, lighting conditions, etc.).
2.2. **Noisy Data**: There may be technical errors, distortions and other image artifacts invalidating specific raw samples from the data set. There may also be unusual outliers that confuse data training algorithms in later stages.
2.3. **Inconsistent Data**: Raw data may have been inaccurately labeled (e.g., labeled male when female, incorrect file formatting names, etc.). These types of errors affect the training algorithms by confusing proper results with strangely labeled data.

3. Data Validation Stage

The data validation stage filters out unwanted data, removes biases and finally normalizes the data format for all subsequent machine learning stages. This stage may be largely automated using filtering information from the previous stage to filter data samples to be used in the next training stages. In this stage, the data samples are rejected, sorted and filtered but not modified to correct any specific issues.

4. Data Transformation Stage

The data transformation stage may modify the raw data samples for the following reasons:

4.1. Correcting gross mistakes and errors in categorization and labeling,
4.2. Reformatting the data into a standardized scaled values, data structures and formats
4.3. Performing calculations on numeric fields and generate commonly used statistical data

In our example, the smart phone facial recognition training may rotate, scale and reposition the data to ensure a wide variety of poses are available to train the network to recognize the user from different angles and distances.

Stages 5, 6 and 7 – Applicable Technical Standards: All of those listed above including the machine learning benchmarking (mlperf.org) with special emphasis on SG-01: Computational approaches.

5. Neural Network Model Training Phase

Machine learning algorithms use the data from the previous stages to *train* a neural-network model (NN-Model). The ML algorithms are called learning algorithms because they automatically derive general purpose models from the presented data. The training data prepared in previous stages must contain correct answers called target attributes. The learning algorithm automatically finds patterns in the training data to map input data to predicted output data with a specific accuracy.

6. NN Model Evaluation and Validation (Training and Inferencing Trials)

The NN-Model generated in the last stage is evaluated for accuracy, speed, performance and a variety of other technical metrics. Once the NN-Model has completed evaluation trials, a subset of the data from the transformation stage is used to validate the final accuracy and performance of the model. This final validation trial determines whether the NN-Model can be put into production in the finalize NN-serving stage.

7. NN Serving (Production Inferencing Phase)

It is important to note the AI workflow illustrated above has two major distinct phases: the data training phases (1 through 6) and the production inferencing phase (stage 7). The NN-service stage is the final production stage for NN-Models where un-validated input data is presented to the model for it to make a prediction and provide a specific accuracy on that prediction.

The vast majority of machine learning work is performed in this final stage to recoup the investment incurred in the previous stages. Once the neural network model has completed training, the model no longer undergoes any further changes. Consideration should be given to the unique properties of the original training stage and the final production stage to ensure correct technical standards definitions account for the differences.

In our example, software applications use frameworks like Keras with TensorFlow to supply novel input data to the trained model to identify the smartphone user's facial location and then match that face to a specific individual. The facial recognition application uses a software framework because many trained models may be executed to complete a single facial recognition task.

Power consumption, performance and efficiency of the smartphone hardware determines the speed with which the phone can train to recognize a specific face and the time required to unlock the phone. These key characteristics can benefit from impartial, consistent and accurate benchmark standards as outlined in the section below.

Qualcomm is heavily focused on the inferencing stage to optimize the speed and cost of execution for consumer products and cloud services. The application of AI technology to the consumer electronics industry is relatively recent and no true world standards exist yet. Industry Alliances and ad-hoc open source groups have started to coalesce around the workflow described above.

The production and deployment of NN-Models relies on the frameworks described above. There are a variety of server-side inference frameworks: TensorFlow and PyTorch. However, the majority of inferencing takes place on millions of mobile devices using software frameworks like open source TensorFlow-Lite and Qualcomm Snapdragon Neural Processing Engine (SNPE), MACE and other optimized proprietary systems.

Machine Learning and AI Performance Benchmarking

The industry needs a standard set of benchmarks for data, training, and inference technical evaluation. There are national efforts by specific countries to fund and influence the development of these types of benchmarks and it is important these benchmarks remain accurate, impartial, consistent, open and to prevent skewed computational favoritism of specific hardware or software implementations.

Inferencing - AI Benchmarks and Performance Analysis. These benchmarking organizations are developing techniques to accurately measure the speed and performance of DNN for a variety of AI inferencing applications.

Benchmark Name	Contributing Organization	Capability
Mlperf.org	Open Industry Ad-Hoc Group	Training and Inference
AI-Matrix	Ad-Hoc Industry Group	Training and Inference

Benchmark Name	Contributing Organization	Capability
AIIA DNN Benchmark	Artificial Intelligence Industry Alliance	Training and Inference
AnTuTu	Chinese Commercial Company	Inference
Baidu Deep Bench	Chinese Commercial Company	Training
AI-Benchmark	Computer Vision Lab, ETH Zurich, Switzerland	Inference
Fathom	Harvard University	Inference
DAWNBench	Stanford University	Training and Inference

Conclusion

Artificial Intelligence and Machine Learning technology is undergoing rapid development and emerging technical standards are still largely in draft form. The NIST can best support the development of US AI by supplying government AI/ML use-cases, large high quality data sets and finally by providing a forum for continued discussion and collaboration on AI and ML standards development.