|  |
| --- |
|  |

|  |  |
| --- | --- |
| Input Contribution | |
| Meeting ID\* | RDM#51 |
| Title:\* | Data management for AI/ML |
| Source:\* | JaeSeung Song, KETI, jssong@sejong.ac.kr  Minbyeong Lee, Hyundai Motors, [minbyeong.lee@hyundai.com](mailto:minbyeong.lee@hyundai.com) |
| Date:\* | 2021-09-02 |
| Input related to\* | WI-0105 oneM2M System Enhancement to AI capabilities  TR-0068 V 0.1.0 |
| Intended purpose of  document:\* | Decision  Discussion  Information  Other <specify> |
| Impacted other TS/TR(s) |  |
| Decision requested or recommendation:\* | Agree for inclusion in TR-0068 |
| Template Version: January 2017 (Do not modify) | |

**oneM2M Notice**

The document to which this cover statement is attached is submitted to oneM2M. Participation in, or attendance at, any activity of oneM2M, constitutes acceptance of and agreement to be bound by terms of the Working Procedures and the Partnership Agreement, including the Intellectual Property Rights (IPR) Principles Governing oneM2M Work found in Annex 1 of the Partnership Agreement.

# Introduction

This contribution provides input to section 6.2 AI/ML and IoT.

In particular, the contribution describes how AI/ML manages data to build a model and perform prediction.

### -----------------------Start of change 1-------------------------------------------

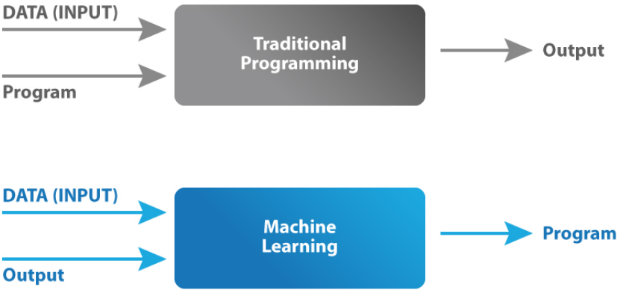
## 6.2 AI/ML and IoT

*Editor’s Note: The section introduces how AI/ML technologies can coexist with IoT technologies.*

### 6.2.1 Steps for AI/ML

Machine Learning gives devices the ability to learn from their experiences and improve their self without doing any coding. Machine Learning is the study of making machines more human-like in their behaviour and decisions by giving them the ability to learn and develop their own programs. This is done with minimum human intervention, i.e., no explicit programming. The learning process is automated and improved based on the experiences of the machines throughout the process. Good quality data is fed to the machines, and different algorithms are used to build ML models to train the machines on this data. The choice of algorithm depends on the type of data at hand, and the type of activity that needs to be automated.

In traditional programming, the input data and a program are feeded into a machine to generate output. When it comes to machine learning, input data along with the output is fed into the machine during the learning phase, and it works out a program for itself. To understand this better, refer to the illustration below:



**Several common terminologies of ML are follows:**

* **Model:** A machine learning model is the mathematical representation of a real-world process. A machine learning algorithm along with the training data builds a machine learning model.
* **Feature:** A feature is a measurable property or parameter of the data-set.
* **Feature Vector:** It is a set of multiple numeric features. We use it as an input to the machine learning model for training and prediction purposes.
* **Training:** An algorithm takes a set of data known as “training data” as input. The learning algorithm finds patterns in the input data and trains the model for expected results (target). The output of the training process is the machine learning model.
* **Prediction:** Once the machine learning model is ready, it can be fed with input data to provide a predicted output.
* **Target (Label):** The value that the machine learning model has to predict is called the target or label.
* **Overfitting:** When a massive amount of data trains a machine learning model, it tends to learn from the noise and inaccurate data entries. Here the model fails to characterise the data correctly.
* **Underfitting:** It is the scenario when the model fails to decipher the underlying trend in the input data. It destroys the accuracy of the machine learning model. In simple terms, the model or the algorithm does not fit the data well enough.

In order to perform ML, there are Seven Steps to take as follows:

**Step 1: Gathering data**

For the purpose of developing our machine learning model, the first step would be to gather relevant data. This step is very crucial as the quality and quantity of gathered data will have a direct impact to a model for prediction. Mistakes such as choosing the incorrect features or focusing on limited types of entries for the data set may render the model completely ineffective.

**Step 2: Data preparation**

The next step is data preparation where the data is located in a place and then prepared for the use in the ML training. The data preparation step split the data sets into 2 parts. The larger part (~80%) is used for training the model while the smaller part (~20%) is used for evaluation purposes. This is important because using the same data sets for both training and evaluation would not give a fair assessment of the model’s performance in real world scenarios. Apart from the data split, additional steps are taken to refine the data sets. This could include removing duplicate entries, discarding incorrect readings etc.

**Step 3: Choosing a model**

The next step is to select a model among the many that researchers and data scientists have created over the year. There are various existing models developed by data scientists which can be used for different purposes. These models are designed with different goals in mind. For instance, some models are more suited to dealing with texts while another model may be better equipped to handle images.

**Step 4: Training**

At the heart of the machine learning process is the training of the model. Bulk of the “learning” is done at this stage. The training step requires patience and experimentation. It is also useful to have knowledge of the field where the model would be implemented. For instance, if a machine learning model is to be used for identifying high risk clients for an insurance company, the knowledge of how the insurance industry operates would expedite the process of training as more educated guesses can be made during the iterations. Training can prove to be highly rewarding if the model starts to succeed in its role. This process then repeats and each cycle of updating is called one training step.

**Step 5: Evaluation**

With the model trained, it needs to be tested to see if it would operate well in real world situations. That is why the part of the data set created for evaluation is used to check the model’s proficiency. This puts the model in a scenario where it encounters situations that were not a part of its training. Evaluation becomes highly important when it comes to commercial applications. Evaluation allows data scientists to check whether the goals they set out to achieve were met or not. If the results are not satisfactory then the prior steps need to be revisited so that root cause behind the model’s underperformance can be identified and, subsequently, rectified. If the evaluation is not done properly then the model may not excel at fulfilling its desired commercial purpose.

**Step 6: Parameter tuning**

If the evaluation is successful, the next step is to perform parameter. This step tries to improve upon the positive results achieved during the evaluation step. There were a few parameters that were implicitly assumed when the training was done. Another parameter included is the learning rate that defines how far the line is shifted during each step, based on the information from the previous training step. These values all play a role in the accuracy of the training model, and how long the training will take.

**Step 7: Prediction**

ML basically answers questions using data. Therefore, the final step of the machine learning process is answer a question using prediction based on a developed model. This is the stage where the model is to be ready for practical applications.

### 6.2.2 Data Augmentation

AI/DL algorithms have become a key standard for most vision and machine learning issues. Despite its general use and high performance for many applications, they have some disadvantages. A big problem with AI/DL methods is the size of the dataset to be used for training. Appropriate training methods require a large dataset. However, a large dataset may not be available for all problems. In this case, appropriate method refer as data augmentation is use to obtain a larger dataset from original dataset.

Data augmentation techniques generate different versions of a real dataset artificially by adding slightly modified copies of already existing data or newly created artificial data from existing data to increase its size. Data augmentation is useful to improve performance and outcomes of machine learning models by forming new and different examples to train datasets. If dataset in a machine learning model is rich and sufficient, the model performs better and more accurate. This techniques enable machine learning models to be more robust by creating variations that the model may see in the real world. So, AI applications take the small dataset of images and transform the objects to different sizes by zooming in or zooming out, flipping them vertically or horizontally or changing the brightness or rotating whatever makes sense for the object as shown in Figure 1. Popular open source python packages for data augmentation in computer vision are Keras ImageDataGenerator, Skimage and OpeCV.

텍스트, 묶음, 다른, 여러개이(가) 표시된 사진

자동 생성된 설명

Figure x: Data Augmentation

Benefits of data augmentation include:

* Improving model prediction accuracy
  + adding more training data into the models
  + preventing data scarcity for better models
  + reducing data overfitting and creating variability in data
  + increasing generalization ability of the models
  + helping resolve class imbalance issues in classification
  + more
* Reducing costs of collecting and labeling data

There exist several basic but powerful data augmentation technicques that are widely used to increase the amount of data set. Followings are several well known data augmentation techniques:

1. **Adding noise**: For blurry images, adding noise on the image can be useful. By “salt and pepper noise”, the image looks like consisting of white and black dots
2. **Cropping**: A section of the image is selected, cropped and then resized to the original image size.
3. **Flipping**: The image is flipped horizontally and vertically. In flipping, the pixels are rearrange while protecting features of image. Vertical flipping is not meaningful for some photos, but it can be useful for example cosmology or microscopic photos.
4. **Rotation**: The image is rotated by a degree between 0 and 360 degree. Every rotated image will be unique in the model. For example, the selected image can be rotated 45 degree. The value can from 359 ~ -359. In the random rotate case, two values, for example, 30, and 90 can be entered. Then a random degree between the two input parameters can be selected to rotate the image. Total number of images can be decided. In this case, the target image can be rotated to generate the given number of images. For example, input parameters A degree, B degree, 100 means that generates 100 images through rotating the source image between A and B degrees.
5. **Scaling**: The image is scaled outward and inward. An object in new image can be smaller or bigger than in the original image by scaling. For example, the image can be scaled below to 100% to 50% of the image height/width. In this case, a random value between 100% ~ 50% will be selected to resize the image.
6. **Translation**: The image is shifted into various areas along the x-axis or y-axis, so neural network looks everywhere in the image to capture it.
7. **Brightness**: The brightness of the image is changed and new image will be darker or lighter. This technique allows the model to recognize image in different lighting levels. The brightness of the image can be changed using different contrast. Depending on the selected contrast various input parameters should be selected, for example, 1. Contrast options, 2. Gamma contrast, 3. Sigmoid contrast, 4. Linear contrast, 5. Various contrast filter can be used.
8. **Contrast**: The contrast of the image is changed and new image will be different from luminance and colour aspects. The following image’s contrast is changed randomly.

텍스트, 슬롯머신이(가) 표시된 사진

자동 생성된 설명

Figure x: Geometric transformations

1. **Color space transformations** – change RGB color channels, intensify any color. The color of image is changed by new pixel values.

텍스트, 화면, 디스플레이, 여러개이(가) 표시된 사진

자동 생성된 설명

1. **Kernel filters** (sharpen or blur an image): These are a very popular techniques in image processing to sharpen and blur images. These filters work by sliding an n × n matrix across an image with either a Gaussian blur filter, which will result in a blurrier image, or a high contrast vertical or horizontal edge filter which will result in a sharper image along edges.

사람, 아이, 젊은, 소년이(가) 표시된 사진

자동 생성된 설명

1. **Random Erasing** (delete a part of the initial image): It retains the overall structure of the object, only occluding some parts of object. Areas are re-assigned with random values, which can be viewed as adding noise to the image.

프로젝터이(가) 표시된 사진

자동 생성된 설명

1. **Mixing images**: Basically, this is a technique to mix images with one another. Might be counterintuitive but it work.

대지, 포유류, 실외, 집고양이이(가) 표시된 사진

자동 생성된 설명

### -----------------------End of change 1-------------------------------------------