Universidad de Costa Rica
Facultad de Ingeniería
Escuela de Ingeniería Eléctrica

IE-0624 Laboratorio de Microcontroladores

EIE
Escuela de
Ingeniería Eléctrica

Introducción a ML con MCU

MSc. Marco Villalta Fallas - marco.villalta@ucr.ac.cr

II Ciclo 2022

Contexto

Que es IA?

Inteligencia artificial

- Inteligencia que muestra un máquina (HW y SW).
- Sistemas que imitan la inteligencia humana para realizar tareas y pueden mejorar iterativamente a partir de la información.
- Componentes fundamentales: Sistemas computacionales, datos y gestión de los mismos, algoritmos de IA.
- Investigación académica inicio en 1956 por Allen Newell, Herbert Simon, Marvin Minsky, Arthur Samuel y John McCarthy.

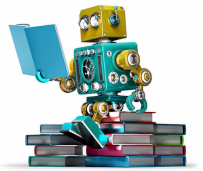
Usos de IA

- Reconocimiento facial y de voz.
- Estrategias para operaciones bursátiles.
- Conducción autónoma.
- Detección de sintomas.
- Mantenimiento predetivo.
- Reconocimiento de escritura.
- Sugerencias de compras.
- Distribución de contenido en redes sociales.

Que es ML?

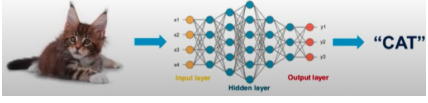
Machine Learning / Aprendizaje automático

- Subconjunto de IA que se centra en desarrollar sistemas que aprenden, o mejoran el rendimiento, en función de los datos que consumen.
- Permite que las máquinas aprendan sin ser expresamente programadas
- Programas con la capacidad de indentificar patrones complejos en millones de datos, construir modelos y generar predicciones de comportamientos futuros, basados en ejemplos de información.

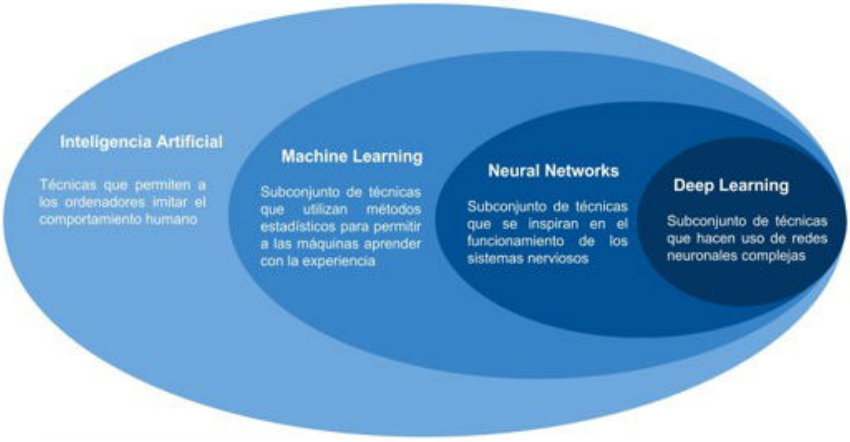


Tipos de ML

- Aprendizaje por refuerzo: se produce cuando una máquina aprende por medio de prueba y error hasta alcanzar la mejor manera de completar una tarea dada
- Aprendizaje autónomo supervisado: El algoritmo se capacita mediante un conjunto de datos que ya está etiquetado y tiene un resultado predefinido.
- Aprendizaje autónomo no supervisado: Algoritmo aprende a identificar procesos y patrones complejos sin que un ser humano proporcione una guía cercana y constante. El aprendizaje autónomo no supervisado implica la capacitación basada en datos que no tiene etiquetas o un resultado específico definido.



De IA a DL


Inteligencia Artificial
Técnicas que permiten a los ordenadores imitar el comportamiento humano
Machine Learning
Subconjunto de técnicas que utilizan métodos estadísticos para permitir a las máquinas aprender con la experiencia
Neural Networks
Subconjunto de técnicas que se inspiran en el funcionamiento de los sistemas nerviosos
Deep Learning
Subconjunto de técnicas que hacen uso de redes neuronales complejas

ML y MCUs

Que es TinyML?

Tiny Machine Learning

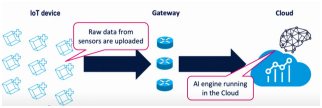
- La creación de modelos de aprendizaje automático para dispositivos de borde.
- Los modelos son adecuados para dispositivos con memoria y potencia de procesamiento limitadas, con conectividad limitada o nula.
- Aplicaciones de complejidad baja: Analisis de sensores, reconocimiento de actividad, analisis de estres
- Aplicaciones de complejidad media: Audo, reconocimiento del habla, detección de objetos
- Aplicaciones de complejidad alta: Detección de objetos/clasificación/rastreo, síntesis del habla, análisis del lenguaje natural.

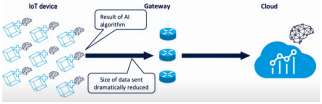
Porque TinyML?

- Aplicaciones requiere inferencia en el dispositivo.
- Aplicaciones hacen que sea atractivo comercialmente.
- Microncontroladores son baratos.
- Modelos de TinyML son posibles por tecnicas avanzadas para hacer redes mas compactas y eficientes.
- Capacidad para procsar y usar los datos sensados usando modelos de ML han estado limitados por la conectividad y acceso a servicios en la nube.

Modelos de procesamiento de ML para MCU

- ML se ha ido moviendo en los ultimos años(5) a los dispositivos
- Se ejecutar carga del trabajo en el dispositivo, cerca de la fuente de información.
- Mucho de la computación de AI se realiza en la nube, implica:
  - Dispositvos Iot envian datos
  - Alto consumo de energía
  - Problemas de seguridad y privacidad
  - Latencia





Pasos generales de TinyML

1. Capturar datos.
2. Limpiar datos, etiquetar datos, contruir topologia NN.
3. Entrenar modelo NN.
4. Convertir la NN en código optimizado para MCU.
5. Procesar y analizar datos usando NN entrenada.

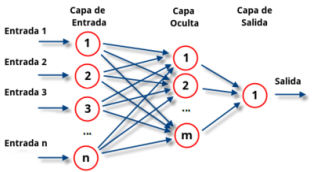
- Pasos 1-3 se realizan en la nube con plataformas como TensorFlow, TensorFlow Light, Keras, Caffe, etc.
- Fabricantes de MCUs tienen herramientas de desarrollo que simplifican estos pasos.
- No se realiza ningun entrenamiento/aprendizaje en los dispositivos de borde.

Redes neuronales en MCUs

- La mayoria de MCUs no tienen la memoria o el poder computacional para ejecutar algoritmos complejos de ML y crear redes neuronales complejas
- MCUs puede ejecutar DNNs optimizadas para los MCUs
- Mayoría requiere de MCUs de al menos 32 bits y con al menos 1 o 2MB de RAM.
- Aplicaciones simples como HAR (Human Activity Recognition) pueden ocupar 2Kb RAM y 1Mhz de procesamiento.
- MCUs simples pueden implementar arquitecturas simples como MLP.

MLP

- MLP corresponde a multilayer perceptron
- Es una red neuronal artificial del tipo *feedfoward* formada por múltiples capas
- Consiste de al menos tres capas de nodos: Una de entrada, una escondida y una de salida
- Exceptuando los nodos de entrada, cada nodo es una neurona que usa una funcion no lineal de activacion.
- Utiliza una tecnica de aprendizaje supervisado llamado *backpropagation* para el entrenamiento.



Referencias

- https://www.juanbarrios.com/inteligencia-artificial-y-machine-learning-para-todos/
- https://www.tensorflow.org/lite
- https://www.tensorflow.org/lite/microcontrollers
- http://www.moretticb.com/Neurona/