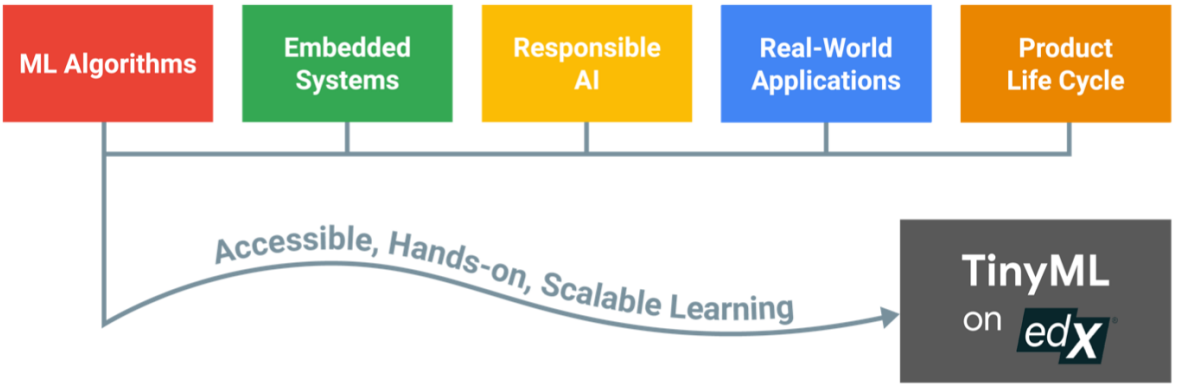
TinyML - What is it and how to make it accessible for everyone?

Summary of the paper Widening Access to Applied Machine Learning with TinyML

Author: Morsinaldo de Azevedo Medeiros - Undergraduate student in Computer Engineering at the Federal University of Rio Grande do Norte - Brazil
E-mail: morsinaldo.medeiros.075@ufrn.edu.br
LinkedIn: https://www.linkedin.com/in/morsinaldo-medeiros-288053105/



Source: Reddi (2021)

English version

In recent years, the application of Machine Learning in the most diverse areas has enabled the emergence of several solutions to the most diverse problems, besides the emergence of new technologies such as virtual assistants, autonomous vehicles, robots, etc. In particular, the area of embedded machine learning has been gaining more and more notoriety because it deals with the use of TinyML, that is, machine learning that can be embedded in devices with low computational cost and low energy cost. However, as Reddi (2021) mentions in his article, there is a deficit of trained people in the area, as well as there are people who have had the most varied types of training and who wish to work in the area, whether as a professional, hobbyist, etc. With this in mind, the author proposes in his article that there should be a collaboration between universities and industry in order to mitigate this deficit. Thus, the partnership between Harvard University and Google resulted in the creation of the massive-open-online-course (MOOP) which aims to facilitate the access of people around the world to knowledge in this area in order to train more and more people. This is of fundamental importance, as it bridges the gap between academia and the job market, since students often get stuck in the classroom and end up not having this kind of experience outside the university walls. In the course, the first part introduces the students to basic concepts of embedded systems (latency, memory, embedded operating systems, etc) and Machine Learning (gradient descent and convolution, for example), as well as showing the practical part using Python code running in the Google Colaboratory environment. The second

part of the course is focused on showing the students practical end-to-end applications of embedded systems while the third part is focused on how to actually deploy the devices, a step that already requires knowledge in electrical engineering and computer science. Thus, this third part also introduces knowledge of TensorFlow lite for microcontrollers, which allows artificial intelligence models to run on embedded devices. In the fourth part of the course, the concept of scalability, that is, how to jump from one device to a thousand, is introduced. This is done by also introducing concepts of machine learning operations (MLOps), which provide best practices for automating the workflow of the data, the model, and the project as a whole, consequently. As there are many different ways of learning, the course organization provides different types of materials such as video lectures, short texts, and exercises on Google Collaboratory in order to cover the most varied types of learning. Furthermore, the partnership between academia and the private sector is important, as it fosters the reallocation of people in the job market, since the emergence of new technologies tends to make certain types of jobs disappear while new ones emerge. One of the problems generated from this is the reallocation of people in the labor market. Because of this, the author also reinforces in his article the ethical responsibilities that this process of development and creation of technologies using Artificial Intelligence requires. There are still two very important aspects raised by the author, the first one being related to the promotion of diversity in the course, that is, the inclusion of women and black people. The second is related to their concern about access to the hardware kit required from the third part of the course onwards, since, despite the attempt to lower the cost to the students to the price of $50.00 many countries charge various taxes on imported products, not to mention the shipping fee. Because of this, they also provide a simulator so that students can deploy the AI models on an embedded device. It is not the same experience as being able to operate a real device, but it is still a great solution for those people who cannot have access to the device.

Portuguese version

Nos últimos anos, a aplicação de Machine Learning nas mais diversas áreas vêm possibilitando o surgimento de várias soluções para os mais diversos problemas, além do surgimento de novas tecnologias como assistentes virtuais, veículos autônomos, robôs *etc*. Em particular, a área de machine learning embarcado vem ganhando cada vez mais notoriedade por se tratar da utilização de TinyML, isto é, o aprendizado de máquina que pode ser embarcado em dispositivos com baixo custo computacional e com baixo custo energético. Contudo, como Reddi (2021) menciona em seu artigo, existe um déficit de pessoas capacitadas na área, bem como existem pessoas que tiveram os mais variados tipos de formação e que desejam atuar na área, seja como profissional, hobbista *etc*. Tendo isso em vista, o autor propõe em seu artigo que deve haver uma colaboração entre as universidades e a indústria com o objetivo de amenizar esse déficit. Assim, a parceria entre a universidade de Harvard e a Google resultou na criação do massive-open-online-course (MOOP) o qual possui como objetivo facilitar o acesso das pessoas ao redor do mundo ao conhecimento dessa área visando capacitar cada vez mais

pessoas. Isso é de fundamental importância, pois diminui a distância entre a academia e o mercado de trabalho, dado que muitas vezes os alunos ficam muito presos à sala de aula e acabam não tendo esse tipo de experiência fora das paredes da universidade. No curso, a primeira parte introduz para os alunos conceitos básicos de sistemas embarcados (latência, memória, sistemas operacionais embarcados *etc*) e de Machine Learning (gradiente descendente e convolução, por exemplo), bem como mostra a parte prática utilizando códigos em Python rodando no ambiente do Google Collaboratory. A segunda parte do curso é focada em mostrar para os alunos aplicações práticas fim-a-fim de sistemas embarcados ao passo que a terceira parte é focada em como fazer o deploy, de fato, nos dispositivos, passo esse que já requer conhecimentos na área de engenharia elétrica e ciência da computação. Assim, essa terceira parte introduz também os conhecimentos do TensorFlow lite para microcontroladores, a qual permite os modelos de inteligência artificial rodar em dispositivos embarcados. Na quarta parte do curso, é introduzido o conceito de escalabilidade, isto é, como saltar de um dispositivo para mil. Isso é feito introduzindo também conceitos de machine learning operations (MLOps), os quais fornecem as melhores práticas para automatizar o workflow dos dados, do modelo e do projeto como um todo, consequentemente. Como existem várias formas de aprendizado diferentes, a organização do curso provê diferentes tipos de materiais como vídeo-aulas, textos curtos e exercícios no Google Colaboratory visando abranger os mais variados tipos de aprendizado. Além disso, a parceria entre a academia e a iniciativa privada é importante, pois fomenta a realocação de pessoas no mercado de trabalho, dado que o surgimento de novas tecnologias tendem a fazer com que certos tipos de empregos desapareçam ao passo que ocorre o surgimento de novos. Um dos problemas gerados a partir disso é a realocação das pessoas no mercado de trabalho. Por causa disso, o autor reforça ainda em seu artigo as responsabilidades éticas que esse processo de desenvolvimento e criação de tecnologias utilizando Inteligência Artificial requerem. Ainda há dois aspectos muito importantes levantados pelo autor, sendo o primeiro deles relacionado ao fomento à diversidade no curso, ou seja, a inclusão de mulheres e de pessoas pretas. O segundo está relacionado à preocupação deles do acesso ao kit de hardware necessário a partir da terceira parte do curso, uma vez que, apesar da tentativa de baixar o custo para o estudantes até o preço de U$ 50.00, muitos países cobram vários impostos sobre produtos importados, fora a taxa de frete. Por causa disso, eles também disponibilizam um simulador para que os alunos possam realizar o deploy dos modelos de inteligência artificial em um dispositivo embarcado. Não é a mesma experiência de poder mexer um dispositivo real, mas ainda é uma ótima solução para aquelas pessoas que não podem ter acesso ao dispositivo.

Reference:

Reddi, Vilay Janapa et al. Widening Access to Applied Machine Learning with TinyML, 2021.