Derin Öğrenme Sistemleri Geliştirme/Konuşlandırma Platformları ve Model Optimizasyonu

Prof. Dr. Alptekin Temizel

Enformatik Enstitüsü, ODTÜ

Gerçek Hayatta Makine Öğrenmesi Sistemleri

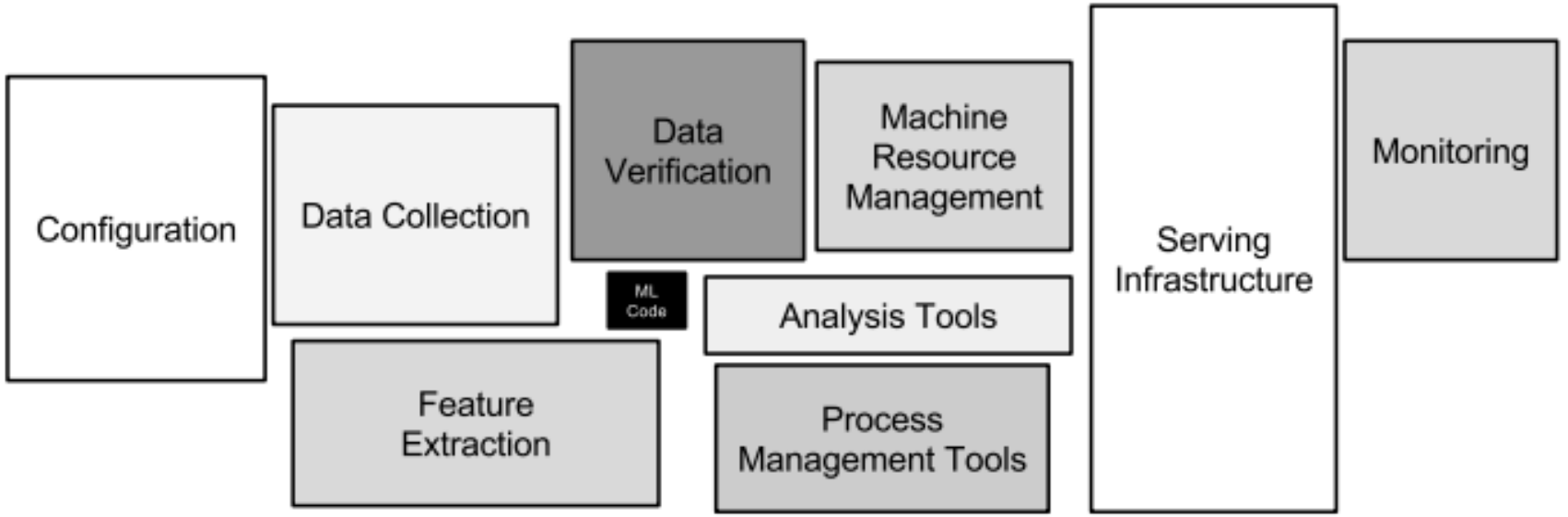


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in neural information processing systems (pp. 2503-2511).

Makine Öğrenmesi Yaşam Döngüsü

İlk önce:

- İş ihtiyaçlarını anlamak ve hedeflerin tanımlanması
- Veri Yönetimi
  - Veri Toplama
  - Ön İşleme
  - Veri Artırma (augmentation)
  - Veri Analizi

- Model Eğitimi ve En İyilenmesi
- Paketleme
- Doğrulama
- Konuşlandırma
- İzleme ve Güncelleme

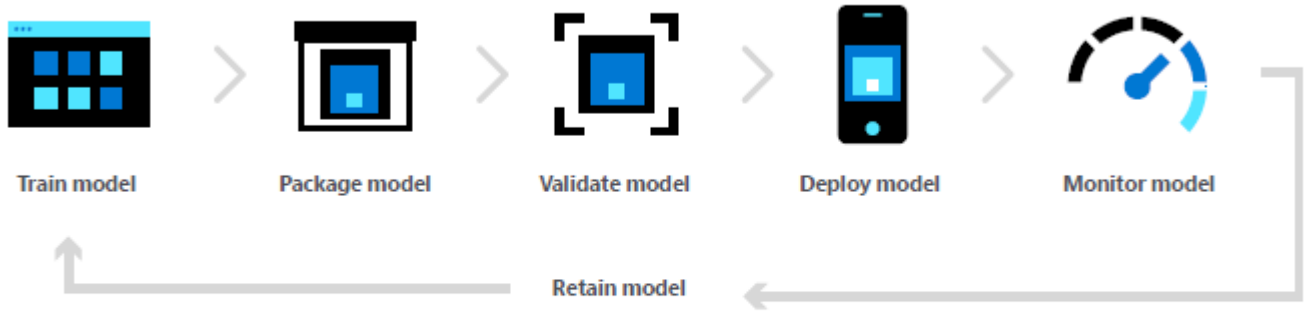


Figure 1. The end-to-end ML life cycle.

Figure from “Drive Efficiency and Productivity with Machine Learning Operations”, Microsoft Azure White Paper

Makine Öğrenmesi Yaşam Döngüsü

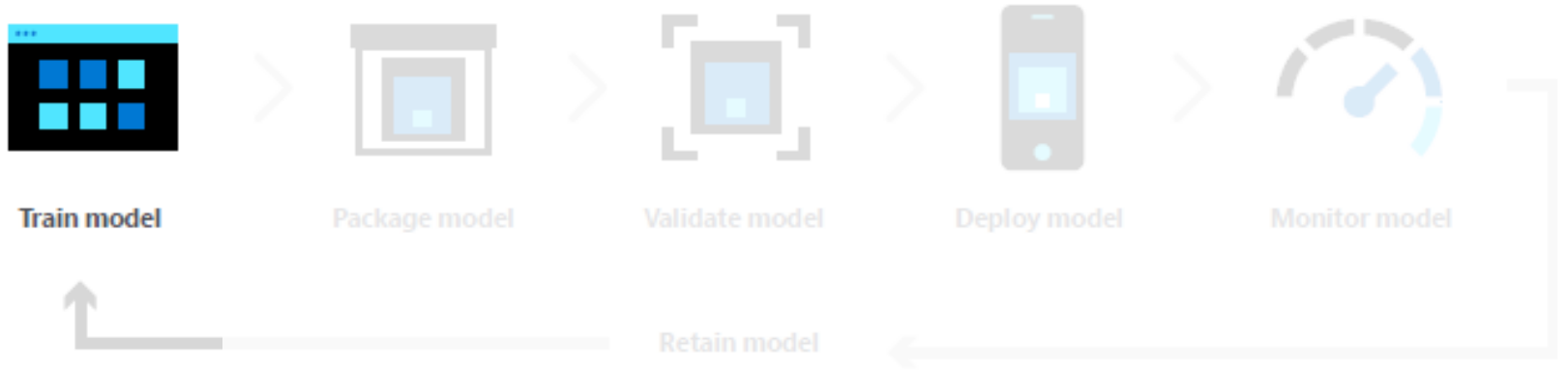


Figure 1. The end-to-end ML life cycle.

Figure from “Drive Efficiency and Productivity with Machine Learning Operations”, Microsoft Azure White Paper

Grafik İşleme Üniteleri (GPU)

- Esas geliştirilme amaçları görselleştirme (rendering) ve oyunlardır.
- PC'lerde, oyun konsollarında bulunmakla birlikte son yıllarda cep telefonları, tabletler, gömülü sistemler ve arabalarda da bulunurlar.

AiB graphics chips	Q4 / 2019	Q1 / 2020	Q2 / 2020	Q3 / 2020	Q4 / 2020
AMD	31.1%	30.8%	22%	23%	17%
nVidia	68.9%	69.2%	78%	77%	83%

Source: Jon Peddie Research

GPU’lar ile Parallel İşleme

GPUs are fast…

GTX 285 has 240 cores, 1 TFLOPS

GTX 480: 1345 GFLOPS 250W, March 2010

GTX 590: 2488 GFLOPS 244W, March 2011

GTX 680: 3090 GFLOPS 195W, March 2012

GTX 780Ti: 5046 GFLOPS 250W, November 2013 (649$)

GTX 980: 4612 GFLOPS , 165W, September 2014 (549$) (Later: 5632 GLOPS, 250W)

GTX 980 notebook: 4612 GFLOPS, 145 W, September 2015

GTX 1080: 9 TFLOPS, 180W, May 2016 (599$)

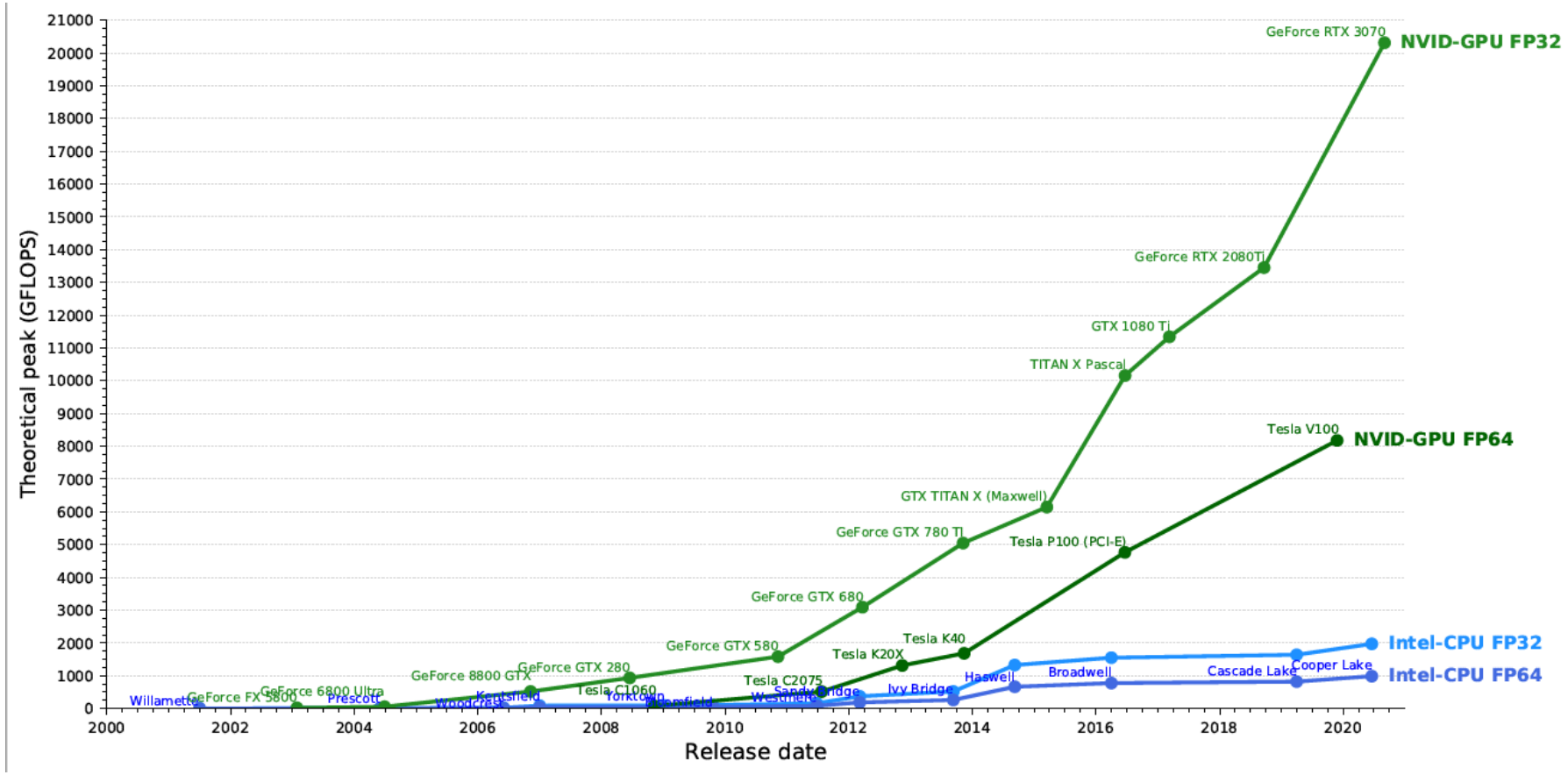
GTX 1080TI: 11.3 TFLOPS, 250W March 2017 (699$)

RTX 2080TI: 13.4 TFLOPS, 250W, Sept 2018 (999$)

RTX 3080: 29.8 TFLOPS, 320W, Sept 2020 (700$)

Note: Intel Core i7-8700K 6-core CPU has a performance of 218 GFLOPS @95W

Grafik İşleme Üniteleri (GPU)



Eğitim: Donanımlar - NVIDIA

EĞİTİM

FULLY INTERGRATED DL SUPERCOMPUTER

DGX-1 & DGX Station

DESKTOP

RTX/GTX series

DATA CENTER

Tesla A100
Tesla V100

INFERENCE

DATA CENTER

Tesla A100/V100

Tesla T4

AUTOMOTIVE

Drive PX2

EMBEDDED

Jetson TX2

Jetson Xavier

Eğitim: Donanımlar - Goggle

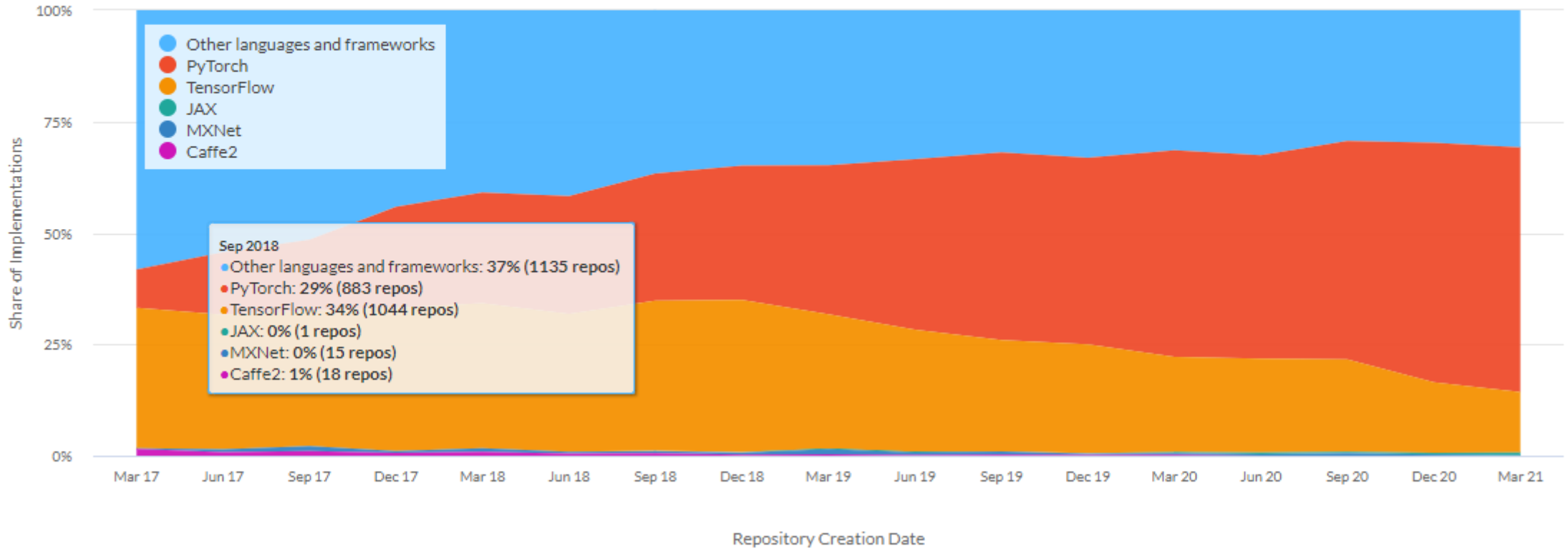
Google Cloud Tensor Processing Units (TPUs)

- TPU’lar makine öğrenmesi algoritmalarını hızlandırmak üzere özel olarak tasarlanmış uygulamaya özgül tüm devrelerdir (ASIC)
- Temel olarak lineer cebir hesaplamalarını hızlandırırlar ve modellerin Tensorflow ile eğitilmeleri amacıyla kullanılabilirler

Eğitim: Kütüphaneler

Frameworks

Paper Implementations grouped by framework



Kaynak: https://paperswithcode.com/trends

Eğitim: Hiper-Parametre En İyilenmesi

Model en iyilenmesi ve hiper parameterlerin ayarlanması: Hyper-parameter optimization (HPO)

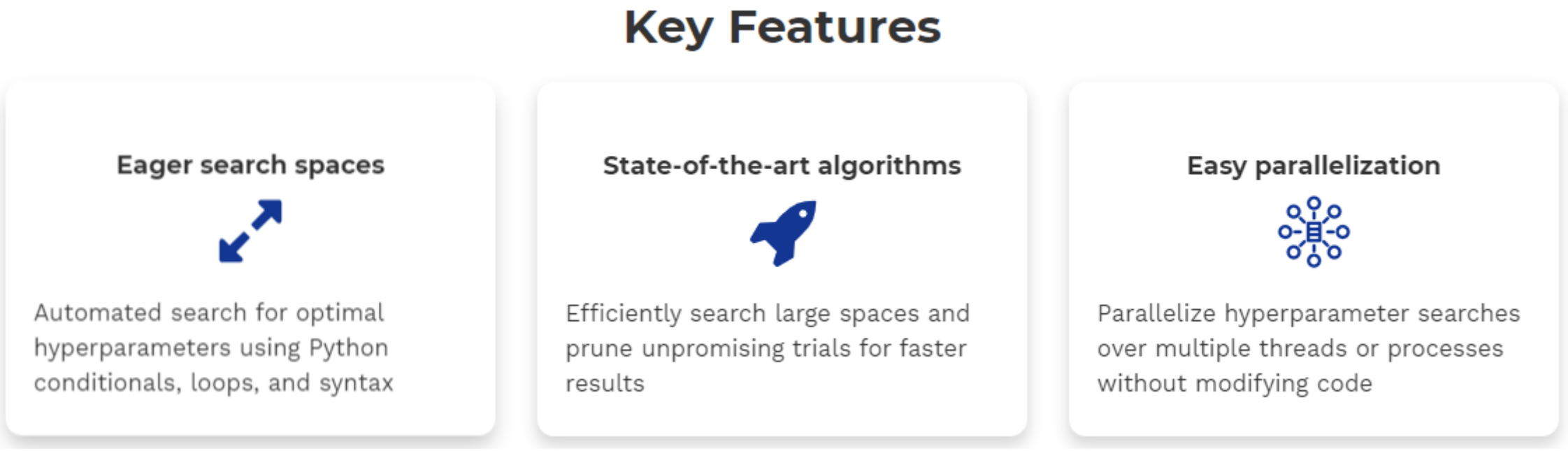
- HPO en iyi hiper-parametrelerin seçilmesi işlemidir.
- HPO yöntemleri modelin pek çok kez eğitilmesini temel alır.
- Her yeni hiper-parametre yeni bir boyut eklenmesine sebep olur ve işlem karmaşıklığını üssel olarak artırır, bu nedenle pahalı ve yoğun kaynak kullanımı gerektiren işlemlerdir
- İdeal olarak HPO işlemi modelin koşturulacağı hedef platformun belirlenerek dikkate alınmasını gerektirir. Gömülü sistemler ve mobil cihazlarda enerji tüketimi ve hafıza kısıtlarının farkında olan (hardware-aware ML) yaklaşımlarla model doğruluğu ve donanım verimliliğini bir arada en iyilenmelidir.

Paleyes, A., Urma, R.G. and Lawrence, N.D., 2020. Challenges in deploying machine learning: a survey of case studies. *arXiv preprint arXiv:2011.09926*.

Eğitim: Hiper-Parametre En İyilenmesi

Optuna: Optimize Your Optimization

Açık kaynaklı hiper-parameter en iyileme çatısıdır, hiper-parameter taramasını otomatikleştirmeyi hedefler



https://optuna.org/

Makine Öğrenmesi Yaşam Döngüsü

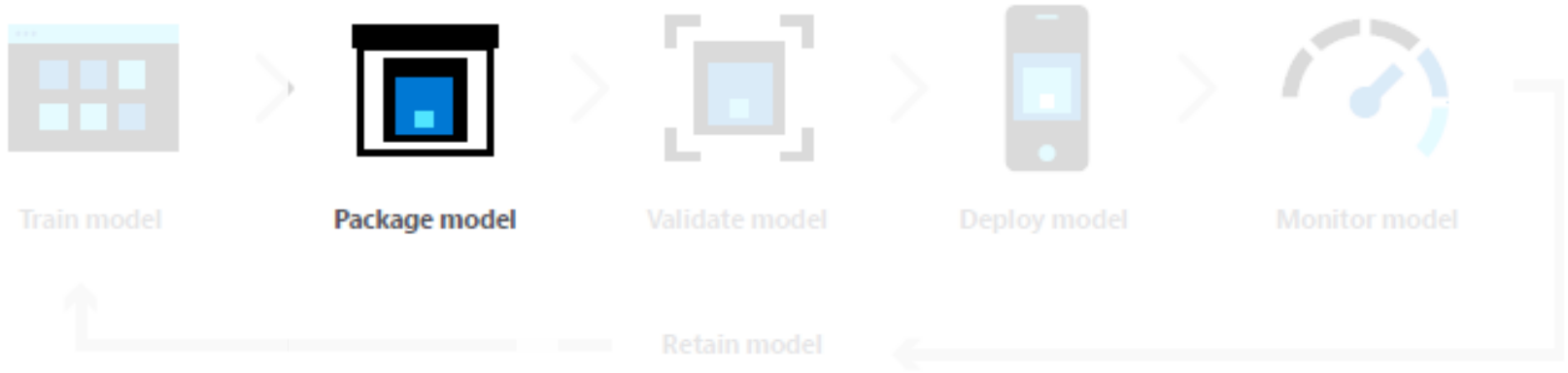


Figure 1. The end-to-end ML life cycle.

Figure from “Drive Efficiency and Productivity with Machine Learning Operations”, Microsoft Azure White Paper

Paketleme: Open Neural Network Exchange (ONNX)

- Eğitim çerçeveleri çıkarsamanın verimli olması için tasarlanmamışlardır. Ayrıca model yapılarının ileride değişmesi olasıdır.
- Eğitimin tamamlanmasının ardından modellerin dışarıya uyumlu bir yapıda aktarılarak özelleşmiş çıkarsama motorlarının kullanılması tercih sebebidir.

KEY BENEFITS

Interoperability

Develop in your preferred framework without worrying about downstream inferencing implications. ONNX enables you to use your preferred framework with your chosen inference engine.

SUPPORTED FRAMEWORKS ›

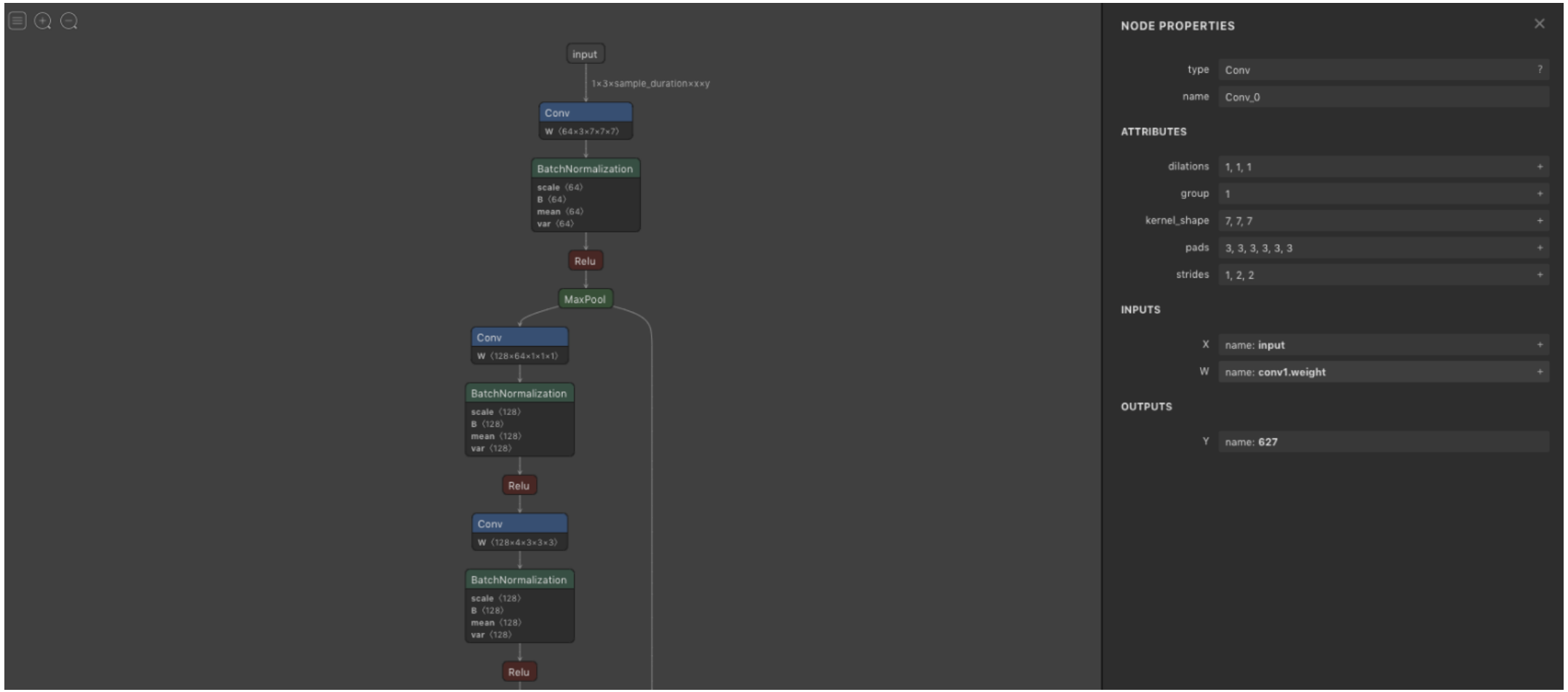
Hardware Access

ONNX makes it easier to access hardware optimizations. Use ONNX-compatible runtimes and libraries designed to maximize performance across hardware.

SUPPORTED ACCELERATORS ›

Paketleme: Open Neural Network Exchange (ONNX)

- ONNX’de model grafik yapısında biçimlendirilir.



Paketleme: Open Neural Network Exchange (ONNX)

- ONNX, veriyi Google tarafından geliştirilmiş olan Protocol Buffer (protobuf) adlı bir formatta saklar.
- Bu format Tensorflow ve Caffe kütüphaneleri tarafından da kullanılmaktadır.
- Protobuf yapısında sadece Float32 gibi veri türleri ve verinin sırası belirtilir, her verinin anlamı kullanılan yazılıma bırakılır. Kavramsal olarak, json gibidir.
- Kütüphanelerin ONNX çıktısı birbiri tekrar eden modüllerden oluşabileceği icin, ek bir basitleştirme adımına tabi tutulabilir.
    - ONNX Simplifier: https://github.com/daquexian/onnx-simplifier
- ONNX dosyalari Netron ile görselleştirilebilir.

Makine Öğrenmesi Yaşam Döngüsü

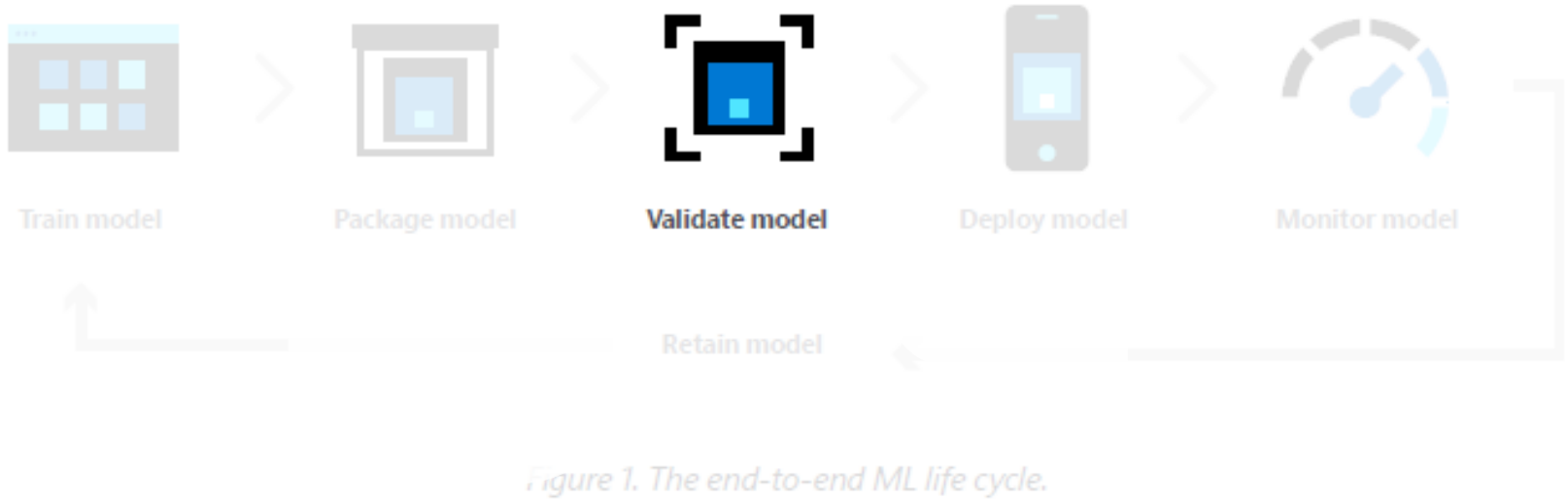


Figure 1. The end-to-end ML life cycle.

Figure from “Drive Efficiency and Productivity with Machine Learning Operations”, Microsoft Azure White Paper

Model Doğrulanması

- Gereksinim Kodlaması (Requirement Encoding)
- Resmi Doğrulama (Formal Verification)
- Teste Dayalı Doğrulama (Test-Based Verification)

Makine Öğrenmesi Yaşam Döngüsü

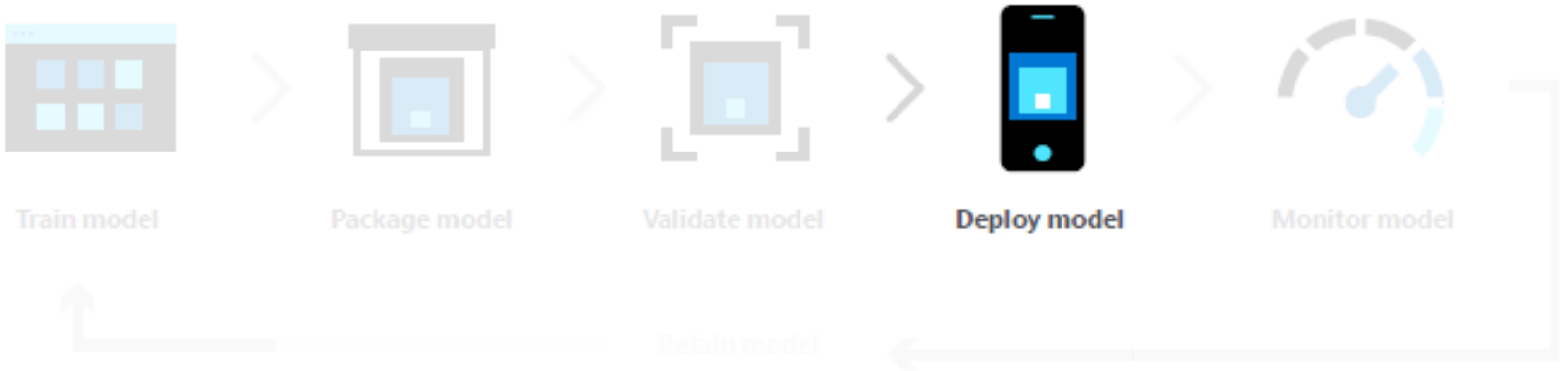


Figure 1. The end-to-end ML life cycle.

Figure from “Drive Efficiency and Productivity with Machine Learning Operations”, Microsoft Azure White Paper

Derin Öğrenme Donanımları

TRAINING

FULLY INTERGRATED DL SUPERCOMPUTER

DGX-1 & DGX Station

DESKTOP

RTX/GTX series

DATA CENTER

Tesla A100
Tesla V100

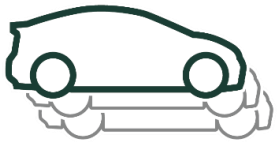
INFERENCE

DATA CENTER

Tesla A100/V100

Tesla T4

AUTOMOTIVE



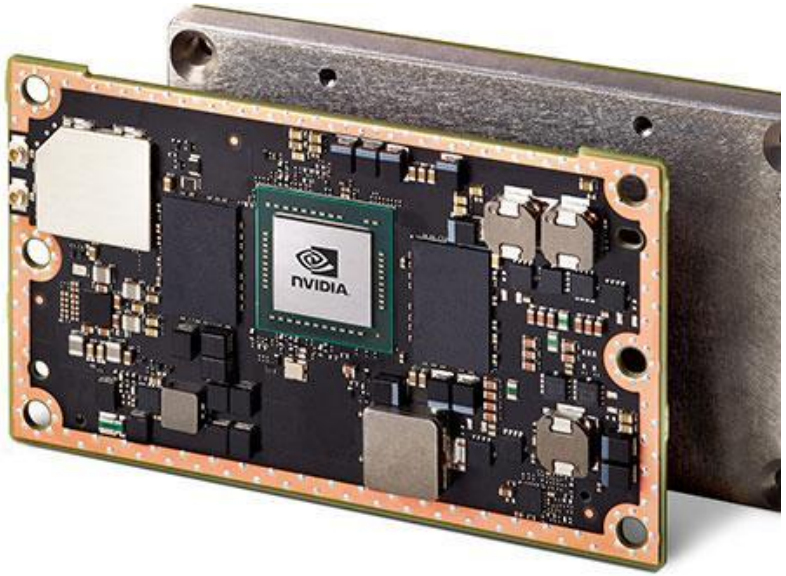
Drive PX2

EMBEDDED

Jetson TX2

Jetson Xavier

NVIDIA Jetson



	Jetson TX2	Jetson TX1
GPU	NVIDIA Pascal™, 256 CUDA cores	NVIDIA Maxwell ™, 256 CUDA cores
CPU	HMP Dual Denver 2/2 MB L2 + Quad ARM® A57/2 MB L2	Quad ARM® A57/2 MB L2
Video	4K x 2K 60 Hz Encode (HEVC) 4K x 2K 60 Hz Decode (12-Bit Support)	4K x 2K 30 Hz Encode (HEVC) 4K x 2K 60 Hz Decode (10-Bit Support)
Memory	8 GB 128 bit LPDDR4 59.7 GB/s	4 GB 64 bit LPDDR4 25.6 GB/s
Display	2x DSI, 2x DP 1.2 / HDMI 2.0 / eDP 1.4	2x DSI, 1x eDP 1.4 / DP 1.2 / HDMI
CSI	Up to 6 Cameras (2 Lane) CSI2 D-PHY 1.2 (2.5 Gbps/Lane)	Up to 6 Cameras (2 Lane) CSI2 D-PHY 1.1 (1.5 Gbps/Lane)
PCIE	Gen 2 | 1x4 + 1x1 OR 2x1 + 1x2	Gen 2 | 1x4 + 1x1
Data Storage	32 GB eMMC, SDIO, SATA	16 GB eMMC, SDIO, SATA
Other	CAN, UART, SPI, I2C, I2S, GPIOs	UART, SPI, I2C, I2S, GPIOs
USB	USB 3.0 + USB 2.0	
Connectivity	1 Gigabit Ethernet, 802.11ac WLAN, Bluetooth	
Mechanical	50 mm x 87 mm (400-Pin Compatible Board-to-Board Connector)	

NVIDIA Jetson Xavier

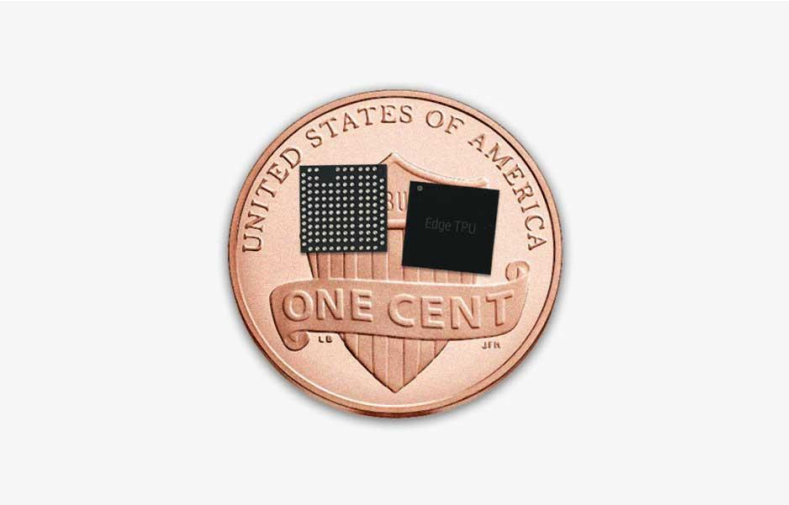
The Tech Specs

Jetson Xavier	
GPU	512-core Volta GPU with Tensor Cores
DL Accelerator	(2x) NVDLA Engines
CPU	8-core ARMv8.2 64-bit CPU, 8MB L2 + 4MB L3
Memory	16GB 256-bit LPDDR4x | 137 GB/s
Storage	32GB eMMC 5.1
Vision Accelerator	7-way VLIW Processor
Video Encode	(2x) 4Kp60 | HEVC
Video Decode	(2x) 4Kp60 | 12-bit support
Mechanical	100mm x 87mm with 16mm Z-height (699-pin board-to-board connector)

I/O	
Display	(3x) eDP/DP/HDMI at 4Kp60 | HDMI 2.0, DP HBR3
Camera Inputs	16 lanes CSI-2, 40 Gbps in D-PHY V1.2 or 109 Gbps in CPHY v1.1 8 lanes SLVS-EC Up to 16 simultaneous cameras
PCIe	5x 16GT/s gen4 controllers | 1x8, 1x4, 1x2, 2x1 • (3x) Root Port + Endpoint • (2x) Root Port
USB	(3x) USB 3.1 (10GT/s) (4x) USB 2.0 Ports
Ethernet	(1x) Gigabit Ethernet-AVB over RGMII
Other I/Os	UFS, I2S, I2C, SPI, CAN, GPIO, UART, SD

Edge TPU

- Edge TPU: Google tarafından tasarlanan ve düşük güçlü cihazlar için yüksek performanslı çıkarsama sağlayan uygulamaya küçük bir özel entegre devre (ASIC)'tir.
- MobileNet V2 gibi modelleri 100+ fps'de güç açısından verimli bir şekilde çalıştırabilir.
- Tensorflow Lite desteği de bulunmaktadır.



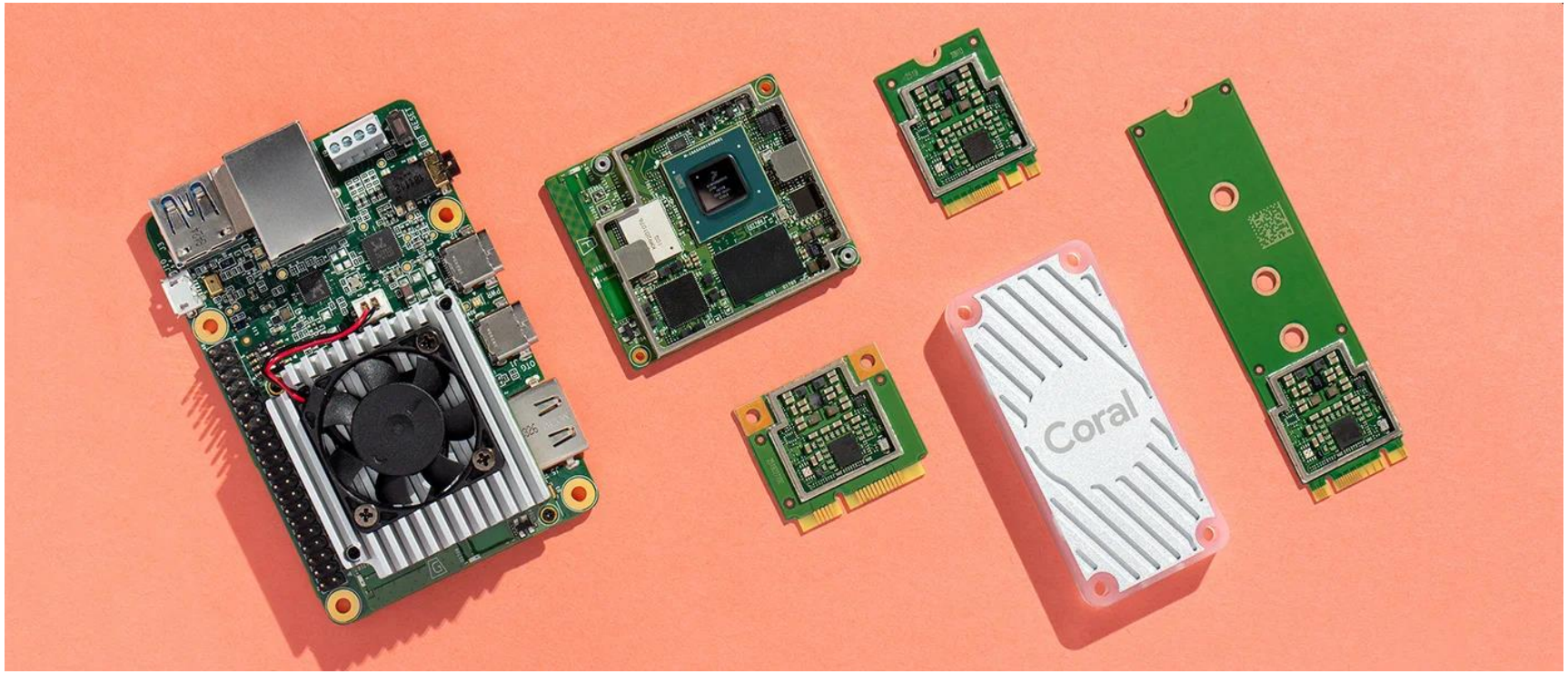
Bir Cent üzerinde iki Edge TPU yongası.

Edge TPU'lu USB Hızlandırıcı.

https://coral.withgoogle.com/tutorials/edgetpu-faq/

Edge TPU – Coral Toolkit

- Coral gömülü AI çözümleri için bütüncül bir çözüm içerir.
- Prototip oluşturmak için tek-kart bilgisayar ve USB cihazları bulunmaktadır.
- Üretime hazır çözümler sunan sistem modülü ve PCI-E modülü mevcuttur.

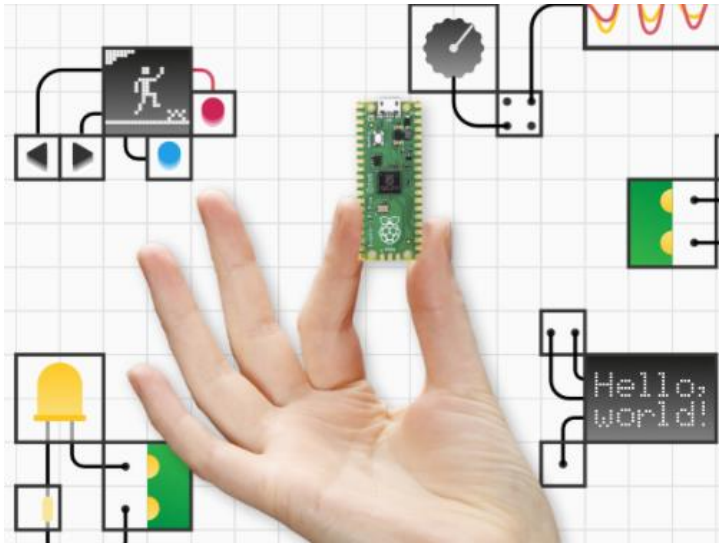


Edge TPU – Intel® Movidius™

- Çıkarsama uygulamaları için hızlandırıcı

IoT Devices – Raspberry Pi

- TensorFlow Lite desteği bulunmaktadır
- Çıktılar 4$ fiyatı bulunan Raspberry Pi Pico üzerinde dahi çalıştırılabilir
- Bu cihaz iki çekirdekli Arm Cortex-M0+ işlemci barındırır, 264KB dahili RAM'i bulunmaktadır ve 16MB'a kadar çip dışında Flash bellek desteği vardır.



MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

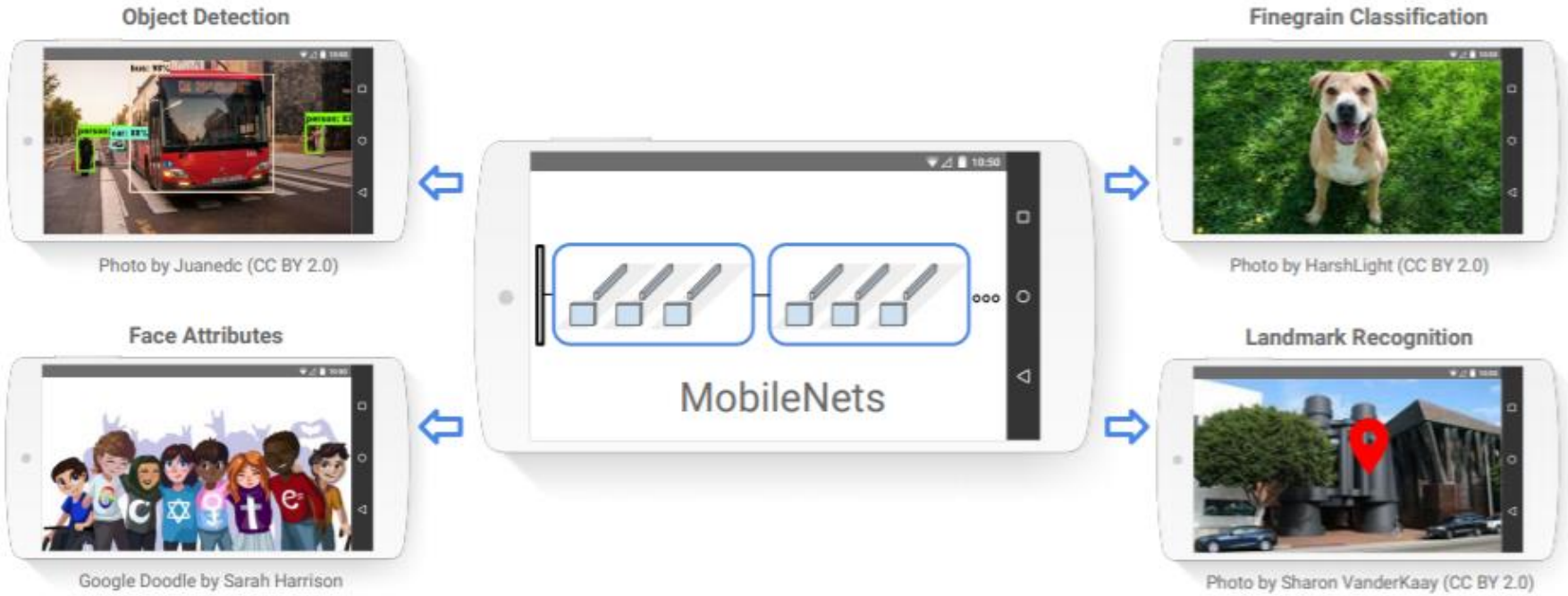
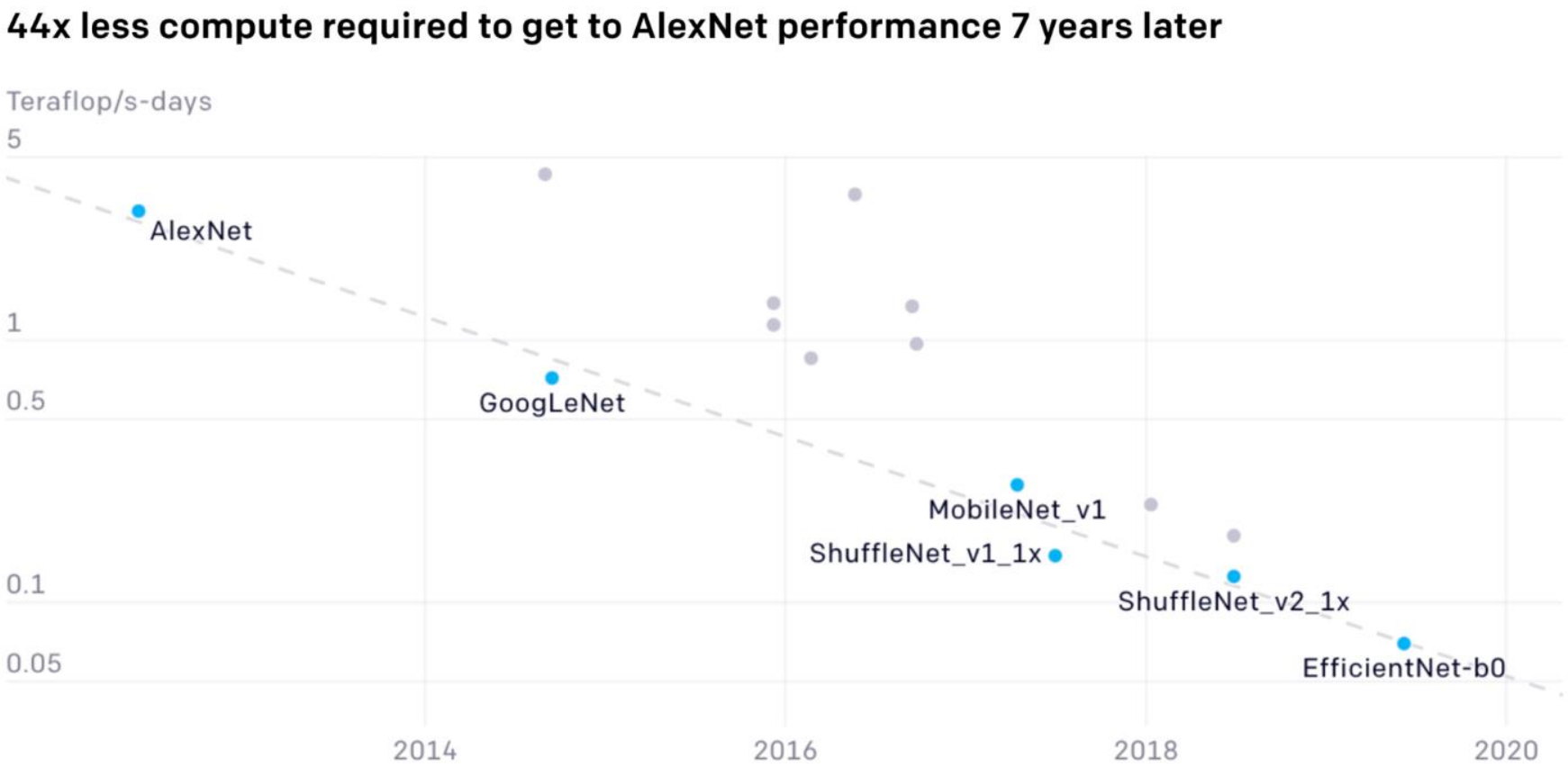


Figure 1. MobileNet models can be applied to various recognition tasks for efficient on device intelligence.

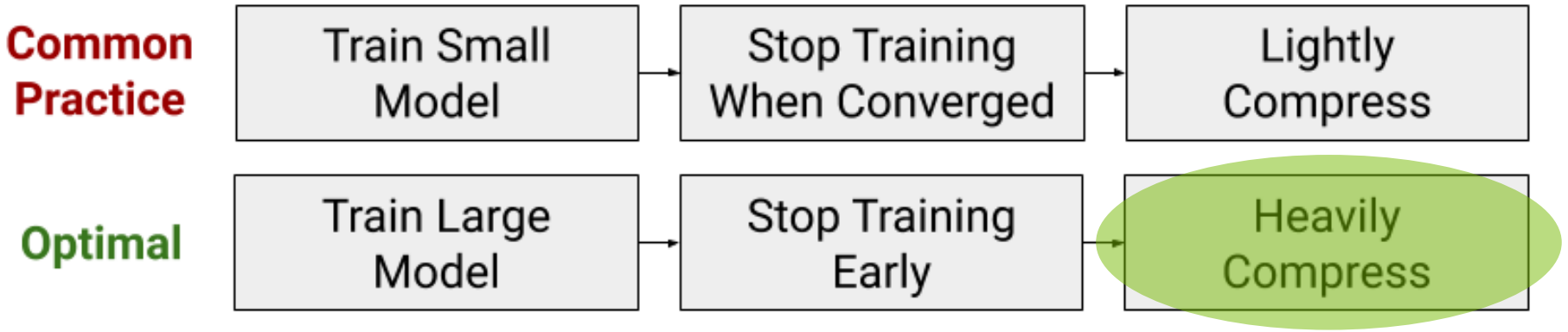
Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Efficientnet: Rethinking model scaling for convolutional neural networks



Tan, M. and Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105-6114).

Büyük Model Eğitip Daha Sonra Sıkıştırma



https://bair.berkeley.edu/blog/2020/03/05/compress/

Li, Z., Wallace, E., Shen, S., Lin, K., Keutzer, K., Klein, D., & Gonzalez, J. E. (2020). Train large, then compress: Rethinking model size for efficient training and inference of transformers. arXiv preprint arXiv:2002.11794.

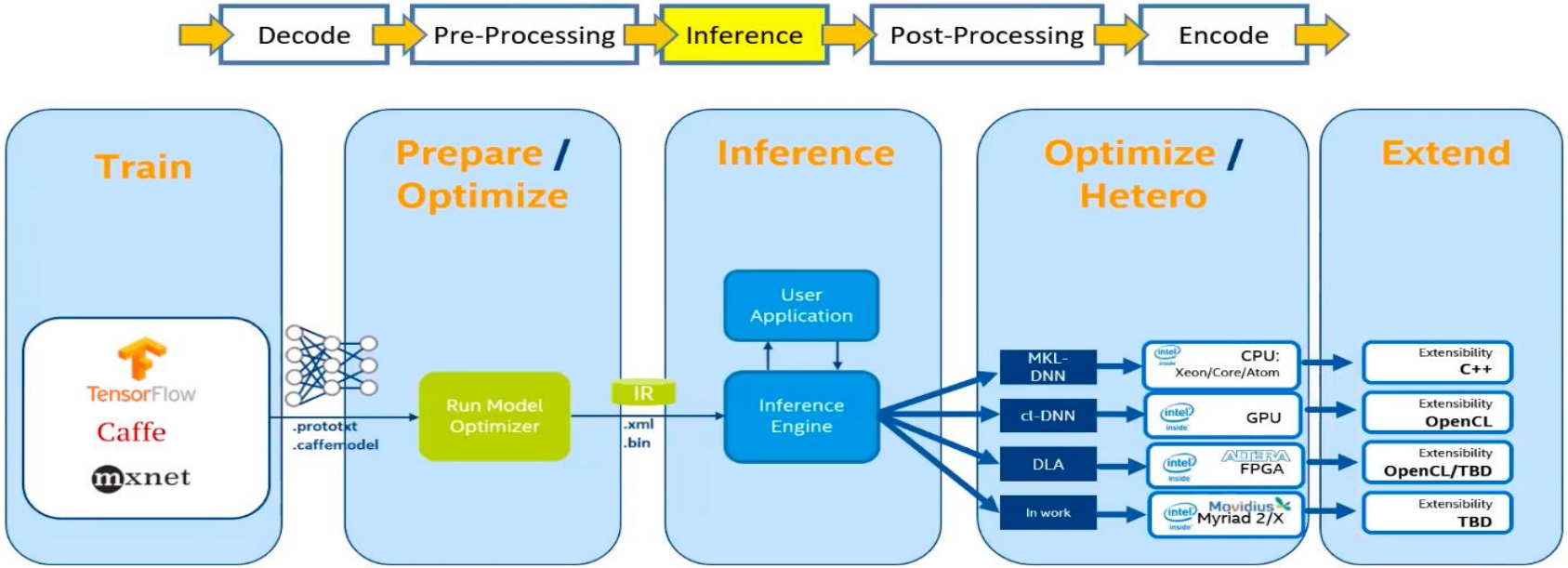
Model En İyilenmesi

- Quantization: Çıkarsamada Kullanılan veri tipinin hassasiyetini azaltarak hız kazanılabilir.
  - Örneğin ResNet50 modelinin Tesla T4 üzerinde 32-bit float yerine 16-bit kullanıldığında 4.27 kat, Int8 kullanıldığında 7.95 kat hızlanma kazandığı görülmüştür

- Pruning: Modelin kullanılmayan kısımlarının kırpılması işlemidir.

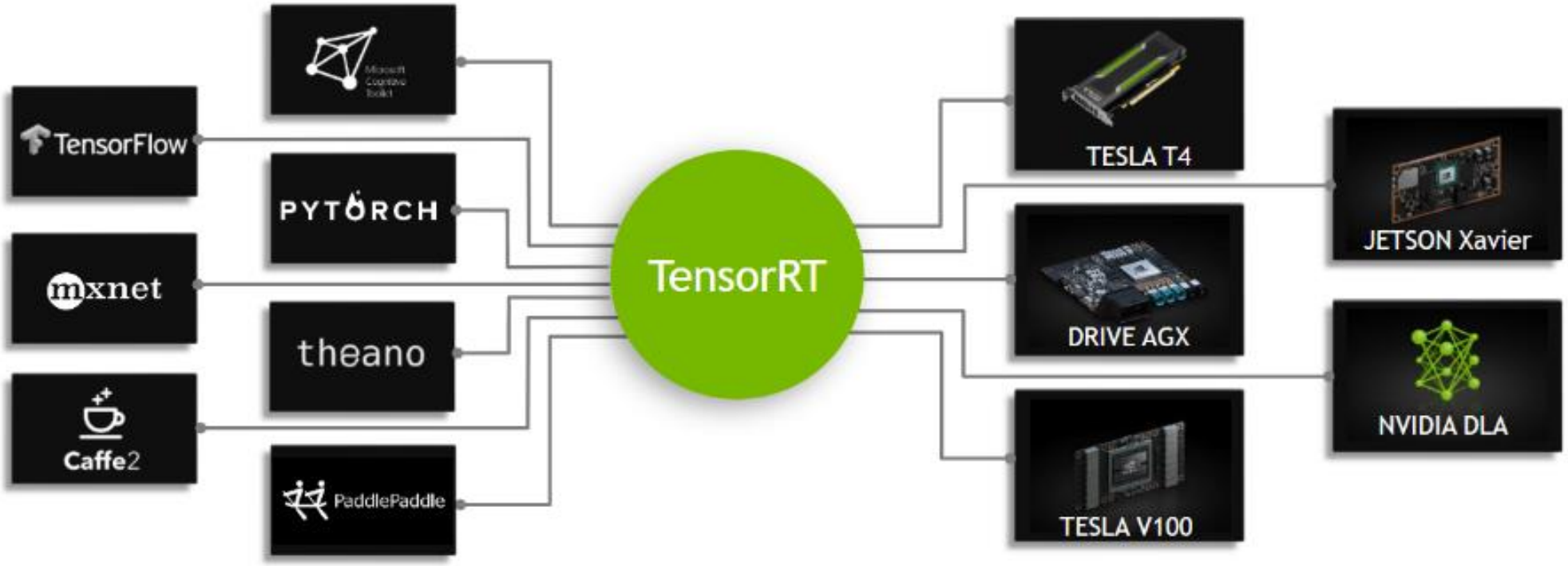
OpenVINO:Open Visual Inference and Neural Network Optimization

- Intel tarafından çıkarsama işlemlerinin hızlandırılması için sağlanan bir araç kitidir.
- CPU, GPU, Intel® Movidius™ Neural Compute Stick ve FPGA gibi farklı ortamlarda hızlandırma mümkündür.
- Farklı derin öğrenme modellerini desteklemektedir.



NVIDIA TensorRT

- Eğitilmiş ağ modeli TensorRT'ye verilerek en iyilenmesi sağlanır
- Model ONNX çıktısı olarak verilebildiğinden ONNX desteği olan tüm platformlarla birlikte çalışabilir.
- Çıktı doğrudan TensorRT Runtime tarafından farklı hedef platformlarda çalıştırılabilir ve hedef platform için eniyilenmiştir.



Google MediaPipe

- MediaPipe canlı ve akan yayınlar için açık kaynak kodlu, farklı platformları destekleyen ve özelleştirilmiş bir çözümdür

MediaPipe

End-to-end acceleration

Built-in fast ML inference and processing accelerated even on common hardware

Build once, deploy anywhere

Unified solution works across Android, iOS, desktop/cloud, web and IoT

Free and open source

Framework and solutions both under Apache 2.0, fully extensible and customizable

https://mediapipe.dev/

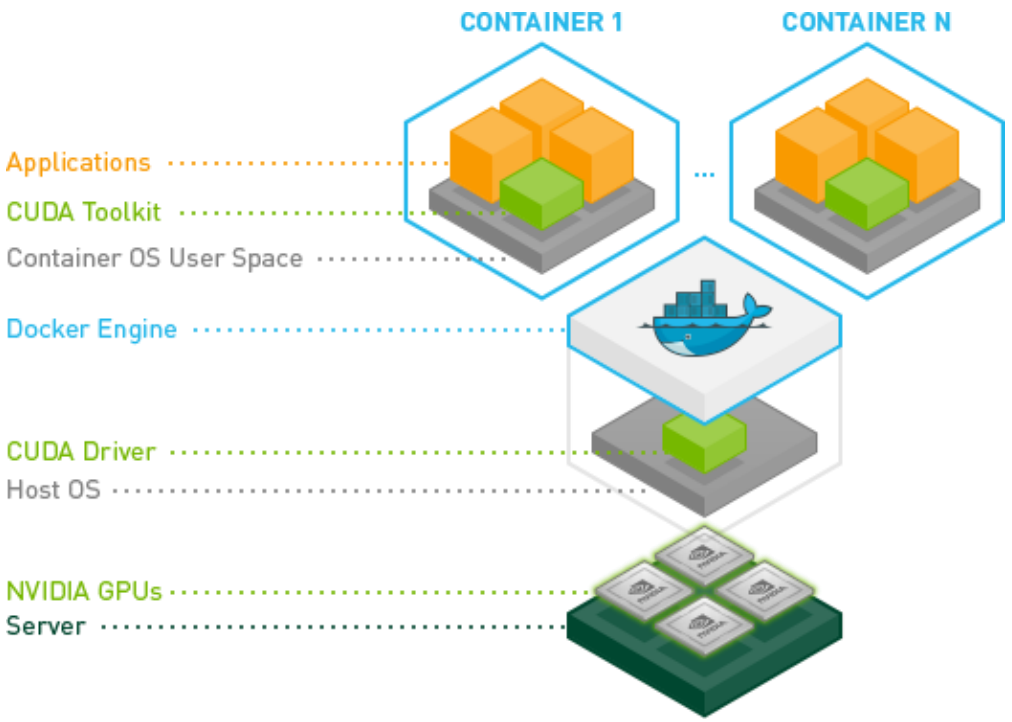
Tencent TNN

- Derin Öğrenme çıkarsama optimizasyonu çerçevesidir.
- ONNX formatina cevrilen modellerden Adreno, Mali, Apple ve Nvidia GPU'lar ve ARM x86 CPU'lar uzerinde optimize edilmis C++ derlemeleri üretmektedir.
- Openvino, TensorRT ve diğer çeşitli optimizasyonlar aynı çerçeve üzerinde yapılabilir.

https://github.com/Tencent/TNN

NVIDIA Docker

- Docker® konteynerleri CPU tabanlı uygulamaların farklı makineler üzerinde kolaylıkla kurulmasını sağlar.
- NVIDIA Docker, bu uygulamalarda GPU'ların da kullanılmasını mümkün kılmaktadır.
- Konteyner kullanımı ile:
  - Tekrarlanabilir derlemeler mümkün olur
  - Farklı makinalara konuşlandırma kolayca yapılabilir



Sonuçlar

- Model geliştirme makine öğrenmesi sistemlerinin en önemli aşamalarından bir tanesidir
- Ancak modelin geliştirilmesinin ardından gerçek hayatta istenen doğrulukta ve performansta çalışacak sistemlerin oluşturulması önemli çabalar gerektirir.
- Bu sistemlerin oluşturulması konusunda yoğun çabalar bulunmaktadır, bu sayede farklı donanım ve yazılım platformları ortaya çıkmıştır.
- Bu platformlar arasından uygun olanın seçilmesi ve modelin hedef donanıma göre en iyilenmesi önemli bir konudur.