Resource-aware Machine Learning Model Optimization for Edge Computing

Otimização de modelos de aprendizado de máquina para computação de borda

xGMobile
Centro de Competência EMBRAPII
Inatel em Redes 5G e 6G

Inatel

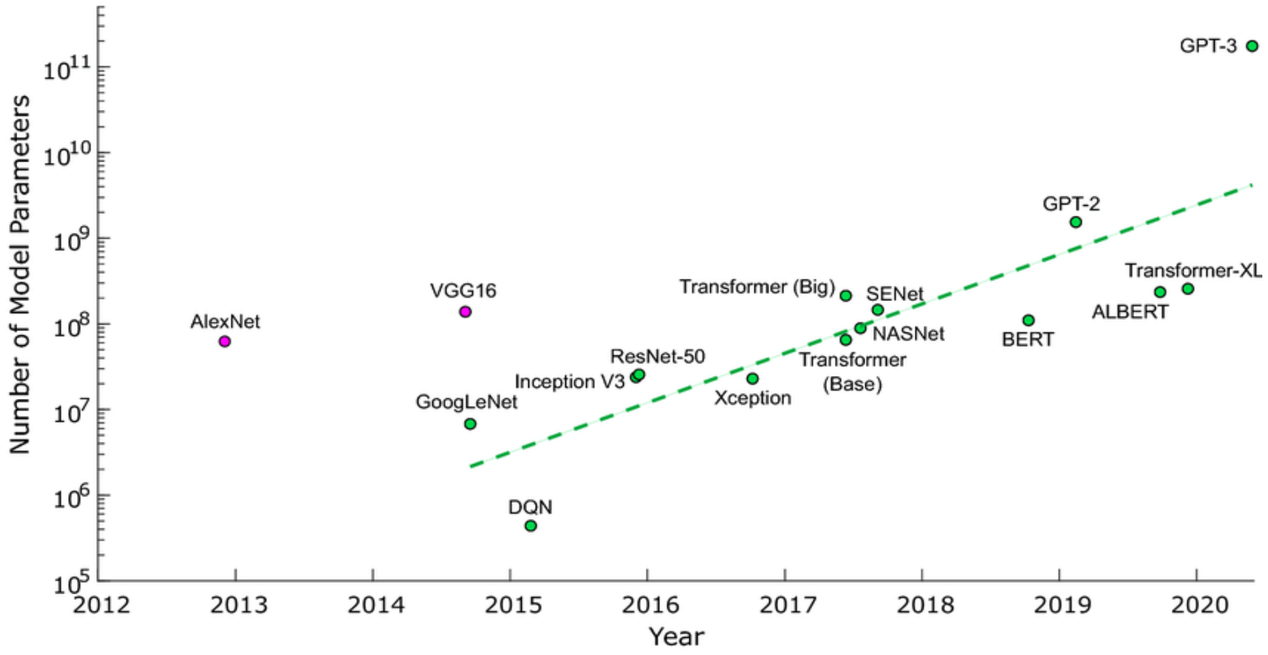
WAI
Inatel Labs

Wireless and Artificial Intelligence

Felipe Augusto Pereira de Figueiredo
felipe.figueiredo@inatel.br

Como executar IA em dispositivos com recursos limitados?

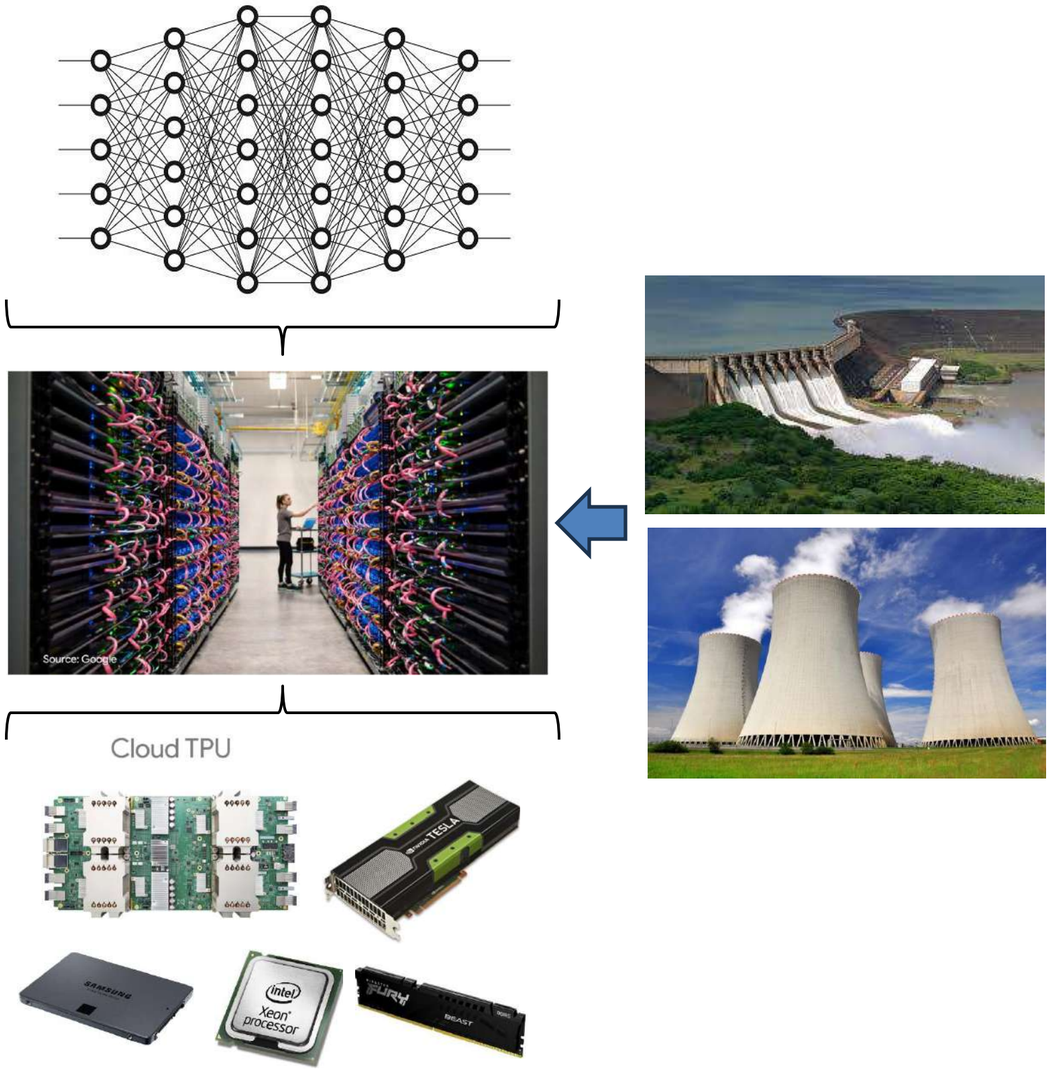
A escala dos modelos de IA atuais



Devido à **complexidade** crescente dos **problemas**, à **abundância** de **dados** e **HW** mais **poderoso** e **barato**, os modelos não param de crescer.

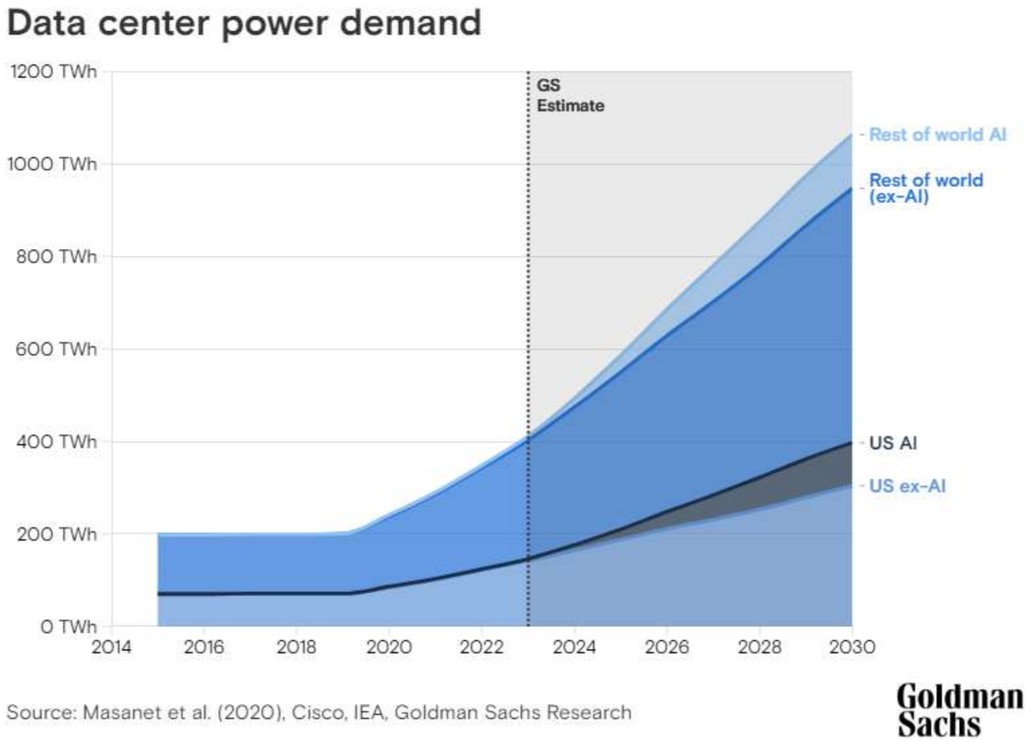
Nos últimos anos, o **poder computacional** necessário para **treinar** os modelos mais usados hoje ***teve que crescer 300.000 vezes***.

O desafio dos modelos complexos de IA

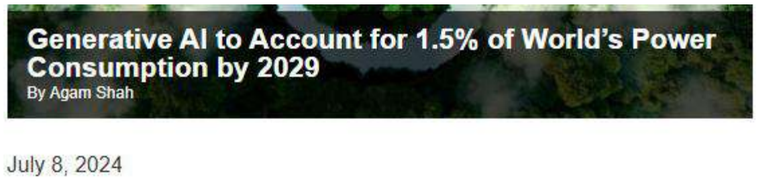


No entanto, à medida que esses **modelos** se tornam mais **complexos**, também **aumentam** os **custos computacionais** (e.g., processamento e armazenamento) e de **energia** associados, além da necessidade de **grandes espaços físicos**.

O desafio dos modelos complexos de IA



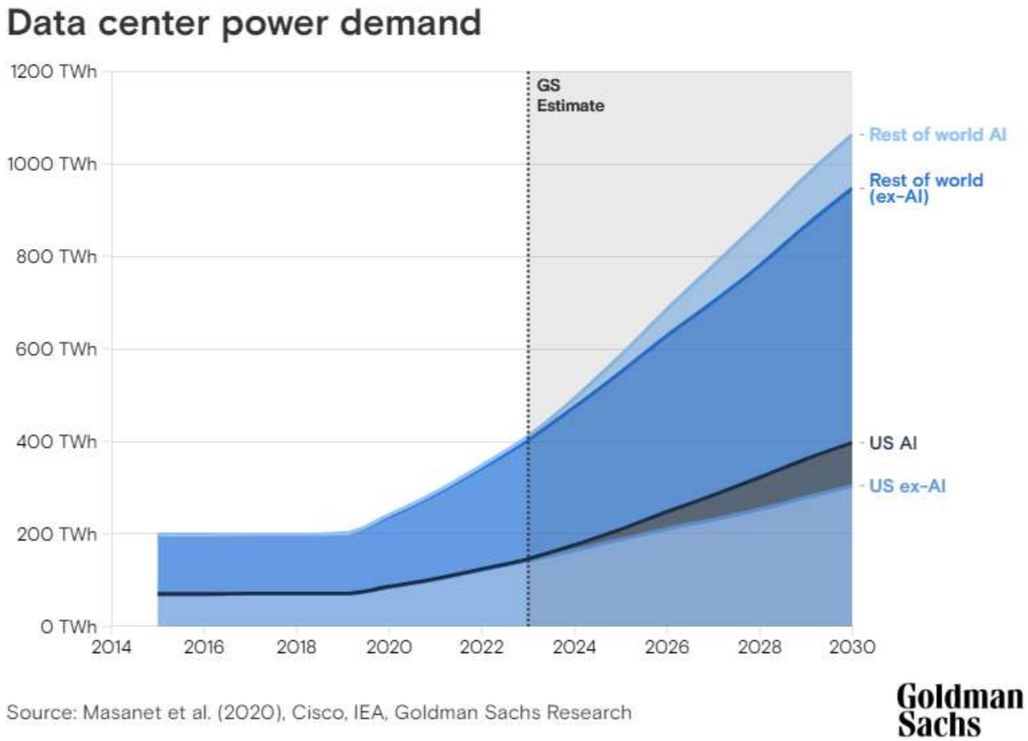
- Estima-se que a **demanda de energia** para *data centers* **crescerá 160% até 2030**.
- Até 2019, GPUs consumirão 2.318 TWh de energia, atingindo **1,5% do consumo global**.
- O **GPT-3** usa quase **1.300 MWh**, o consumo anual de **cerca de 130 residências nos EUA**.



WAI Inatel Labs | Wireless and Artificial Intelligence

xGMobile Centro de Competência EMBRAPII Inatel em Redes 5G e 6G | Inatel

O desafio dos modelos complexos de IA



Precisamos de soluções mais eficientes para a execução de modelos de IA!

BBC

Home News Sport Business Innovation Culture Arts Travel Earth Video Live

Google turns to nuclear to power AI data centres

15 October 2024

Share Save

João da Silva
Business reporter

VOA

November 09, 2024
4:46 AM

By Associated Press

As data center industry booms, English village becomes battleground

AI already uses as much energy as a small country. It's only the beginning.

The energy needed to support data storage is expected to double by 2026. You can do something to stop it.

by **Brian Calvert**
Mar 28, 2024, 9:00 AM GMT-3

Generative AI to Account for 1.5% of World's Power Consumption by 2029

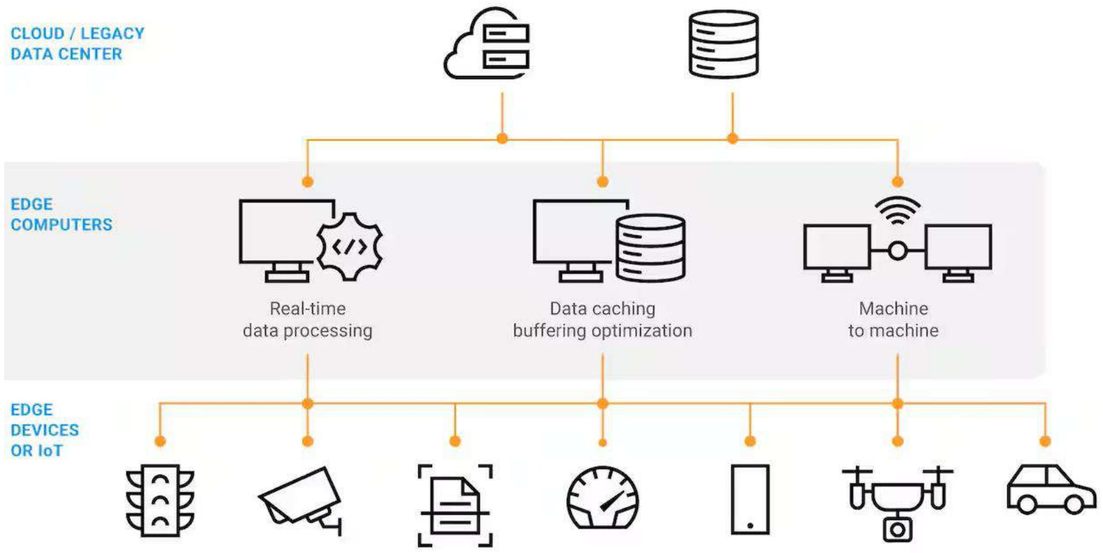
By Agam Shah

July 8, 2024

WAI Inatel Labs | *Wireless and Artificial Intelligence*

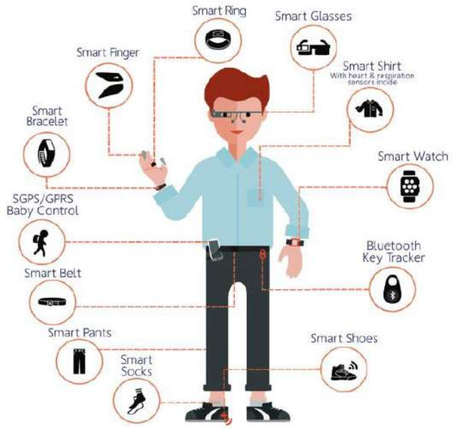
xGMobile Centro de Competência EMBRAPII Inatel em Redes 5G e 6G | Inatel

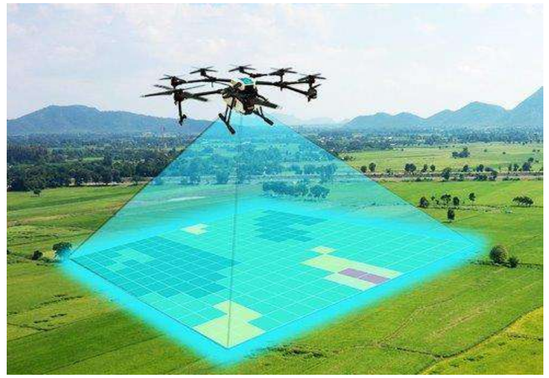
Computação de borda e IoT ao resgate!



- Computação de borda: processamento **próximo** ao local ou onde os **dados são gerados**.
  - Exemplos de **dispositivos**: sensores IoT, drones, smartphones, veículos autônomos.
- **Benefícios** em relação à **computação centralizada**:
  - menor latência e
  - maior segurança e privacidade.
- Além disso, **reduz ou elimina a transferência de grandes volumes de dados** para servidores, **reduzindo o consumo de energia e os custos**.

A oportunidade do aprendizado de máquina na borda





- O **processamento na borda** possibilita **executar modelos** de ML diretamente nos **dispositivos**, incluindo nós, próximos à fonte de dados.
- Exemplos de aplicações:
  - Monitoramento de saúde com dispositivos vestíveis.
  - Automação agrícola com drones.
  - Detecção de falhas em linhas de produção.
  - Aumento do desempenho, eficiência energética e robustez de redes móveis.

Limitações dos dispositivos de borda

- Porém, em geral, estes **dispositivos** apresentam algumas **limitações**:
  - Alimentação via bateria ou limitada.
  - Baixa capacidade de processamento.
  - Armazenamento limitado.

- Entretanto, essas **limitações são as motivações** para o desenvolvimento de **modelos mais eficientes**.

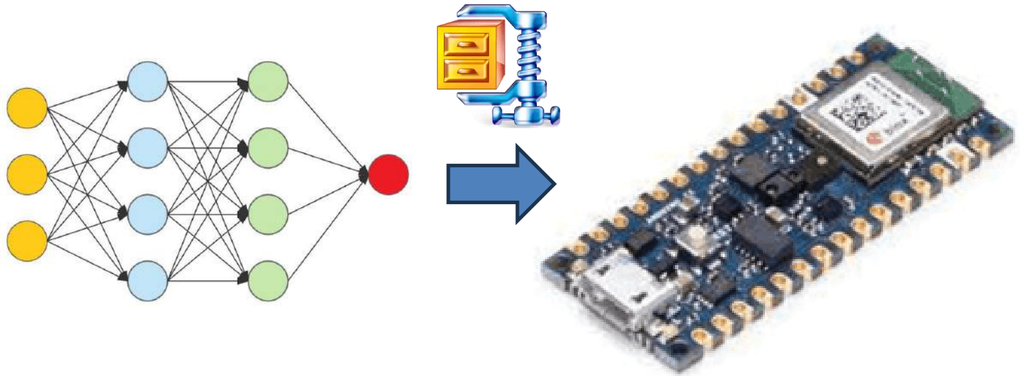
- Uma das formas de se **aumentar a eficiência** é através da **otimização dos modelos**.

Por que otimizar modelos de aprendizado de máquina?

- Porque necessitamos de **modelos menores, mais rápidos e menos exigentes.**

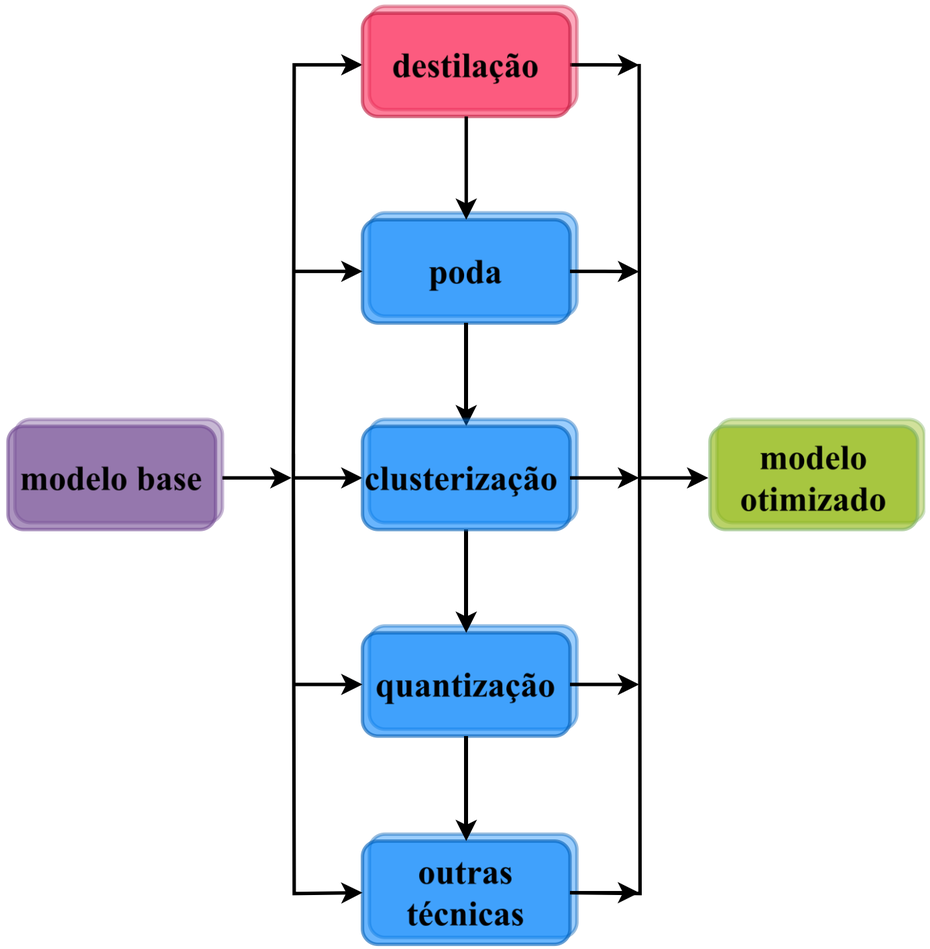
- **Benefícios** diretos do **ML na borda**:
  - Maior eficiência energética.
  - Menor tempo de inferência.
  - Suporte a aplicações móveis e *offline*.
  - Maior privacidade e segurança.

Principais técnicas de otimização



- **Poda de redes neurais**: remoção de conexões e nós irrelevantes.
- **Quantização**: redução da precisão numérica (e.g., 32 bits para 8 bits).
- **Clusterização de pesos**: substituiu pesos semelhantes pelo centroide do cluster mais próximo.
- **Distilação de modelos**: transferência de conhecimento de um modelo grande para um menor.

Combinação de técnicas de otimização



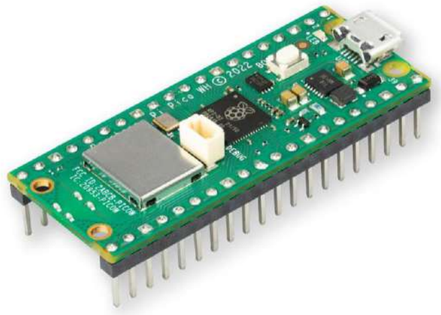
Por que combinar técnicas?

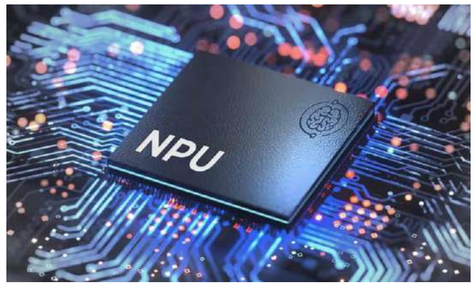
- Aplicar **técnicas** individuais **sequencialmente** (e.g., poda seguida por clusterização) **sem nenhum critério**, pode levar a **redução dos benefícios** e/ou do **desempenho**.

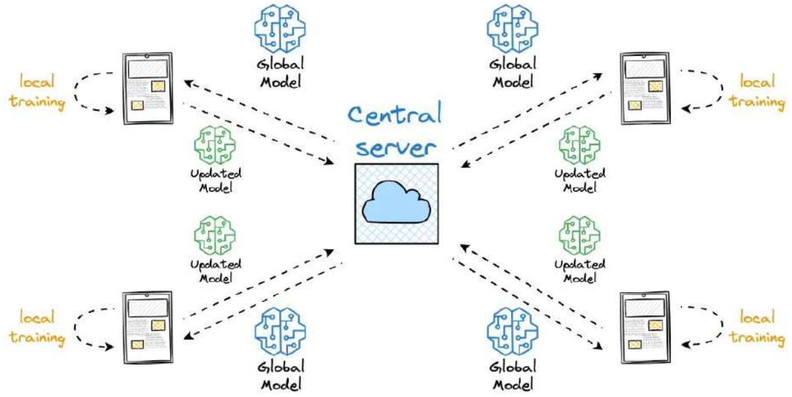
- A **otimização colaborativa** busca criar uma **sinergia** entre **diferentes técnicas**.

- Tudo isso aliado à otimização dos hiperparâmetros dos modelos.

Desafios e perspectivas futuras







Modelo	tamanho (pixéis)	mAPval 50-95	Velocidade CPU ONNX (ms)	Velocidade T4 TensorRT10 (ms)	params (M)	FLOPs (B)
YOLO11n	640	39.5	56.1 ± 0.8	1.5 ± 0.0	2.6	6.5
YOLO11s	640	47.0	90.0 ± 1.2	2.5 ± 0.0	9.4	21.5
YOLO11m	640	51.5	183.2 ± 2.0	4.7 ± 0.1	20.1	68.0
YOLO11l	640	53.4	238.6 ± 1.4	6.2 ± 0.1	25.3	86.9
YOLO11x	640	54.7	462.8 ± 6.7	11.3 ± 0.2	56.9	194.9

Desafios persistentes:

- Limitações de HW.
- Treinamento local ainda inviável em muitos casos.
- Modelos ainda muito complexos.

Perspectivas futuras:

- Desenvolvimento de HW para aceleração dos modelos.
  - e.g., CPUs com NPUs ou TPUs criadas em FPGA/ASIC.
- Aprendizado federado e/ou distribuído para treinamento local.
- Concepção de modelos altamente eficientes para execução em HW limitado.

Resultados


xGMobile
Centro de Competência EMBRAPII
Inatel em Redes 5G e 6G
Inatel
WAI
Inatel Labs
Wireless
and Artificial
Intelligence

EdgeML aplicado ao monitoramento de veículos

Hyago V. L. B. Silva, Davi Rosim, et al.

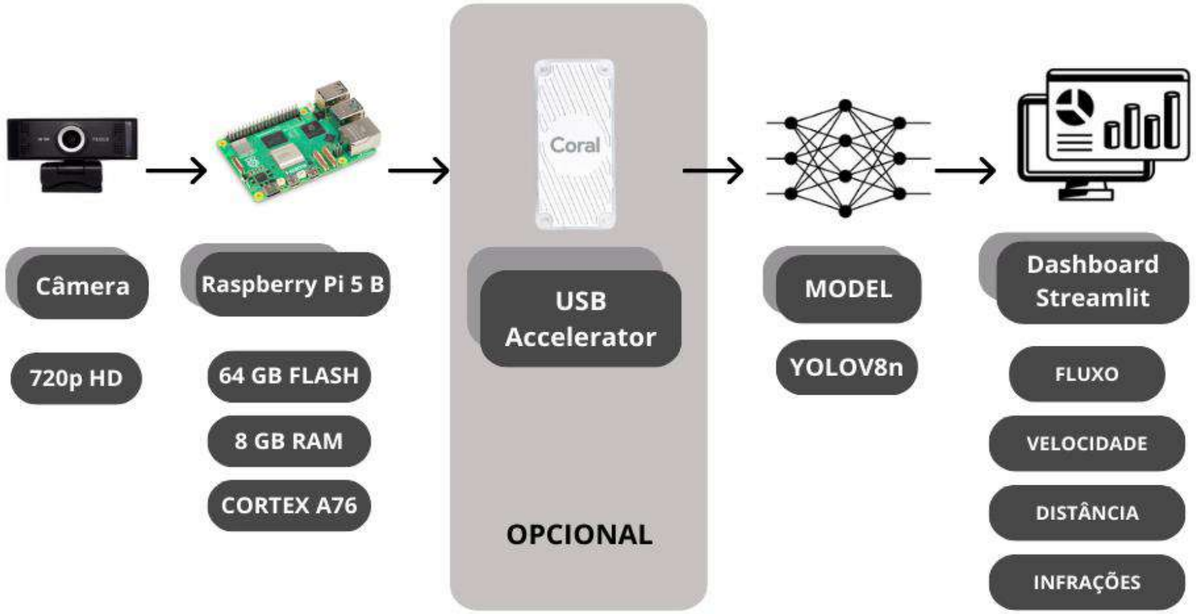
- **Problema**: o **aumento do volume e da complexidade** do tráfego torna o **monitoramento de veículos** uma t**arefa desafiadora** para **controladores** e **agentes** de trânsito.

- **Proposta**: desenvolver uma **solução EdgeML** para **monitoramento autônomo** de veículos em **tempo real**, reduzindo a **carga de operadores e agentes** de trânsito.

WAI Inatel Labs | Wireless and Artificial Intelligence | xGMobile Centro de Competência EMBRAPII Inatel em Redes 5G e 6G | Inatel

EdgeML aplicado ao monitoramento de veículos

Hyago V. L. B. Silva, Davi Rosim, et al.



- A solução usa uma câmera HD 720p, uma Raspberry Pi 5 (ARM Cortex-A76 quad-core 64-bit @2.4GHz com 8 GB de RAM) e, opcionalmente, uma edge TPU.
- O YOLOv8n (nano) é usado para rastrear objetos (i.e., veículos e pessoas).
- A solução analisa autonomamente e em tempo real o fluxo, velocidade e distância de objetos.
- Gera notificações de excesso de velocidade, incluindo as placas, e alertas de colisões eminentes.

EdgeML aplicado ao monitoramento de veículos

Hyago V. L. B. Silva, Davi Rosim, et al.

Métricas de desempenho

Versão	mAP0.5 (%)	P (%)	R (%)	F1	Size (MB)
float32	74.2	71.5	69.9	70.7	9
float16	74.0	70.0	69.0	70.0	5
int8	70.0	70.0	68.5	69.5	3

Modelo rodando na CPU

Versão	FPS	CPU (%)	RAM (GB)
float32	15	20	0.9
float16	16	15	0.7
int8	17	12	0.5

Modelo rodando na Edge TPU

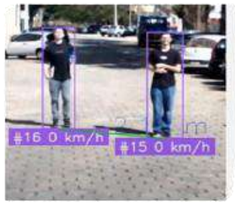
Versão	FPS	CPU (%)	RAM (GB)
float32	26	17	0.7
float16	32	15	0.6
int8	35	10	0.3

Testes de velocidade e distância

Medição: 30km/h
Resultado: 28km/h

Medição: 40km/h
Resultado: 36kh/h

Medição:1 m
Resultado: 0.91 m

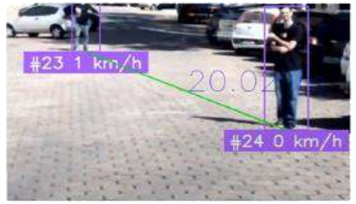


Medição:2 m
Resultado: 2.11 m

Medição: 20km/h
Resultado: 20 km/h

Medição: 10km/h
Resultado: 10km/h

Medição:5 m
Resultado: 4.95 m



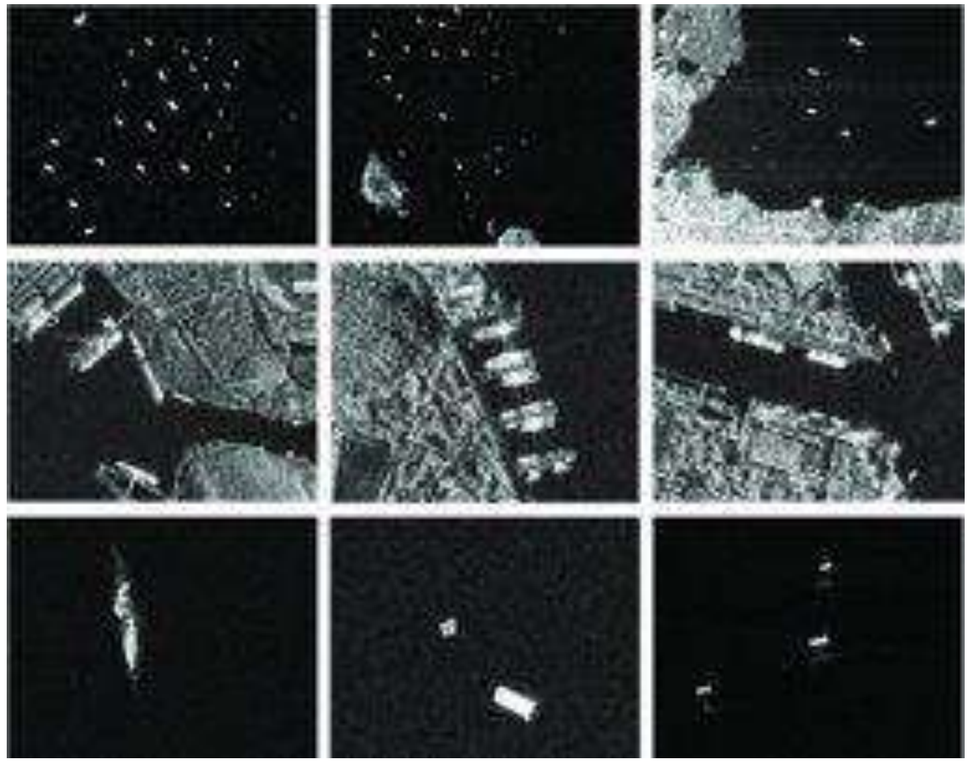
Medição:20 m
Resultado: 20.02 m

- **Resultados:** o modelo opera em **tempo real** com a edge TPU, mantém **bom desempenho mesmo quantizado** e consome **poucos recursos** (CPU, RAM e Flash).
- Estimação de velocidade e distância com erros de +/- 4 Km/h e +/- 0.5 m.
- Futuros trabalhos explorarão *knowledge distillation* e modificações na arquitetura do modelo para melhorar sua eficiência e desempenho sem usar edge TPUs.

WAI Inatel Labs
Wireless and Artificial Intelligence
xGMobile Centro de Competência EMBRAPII Inatel em Redes 5G e 6G
Inatel

YOLOv8 leve para rastreamento de navios em imagens SAR

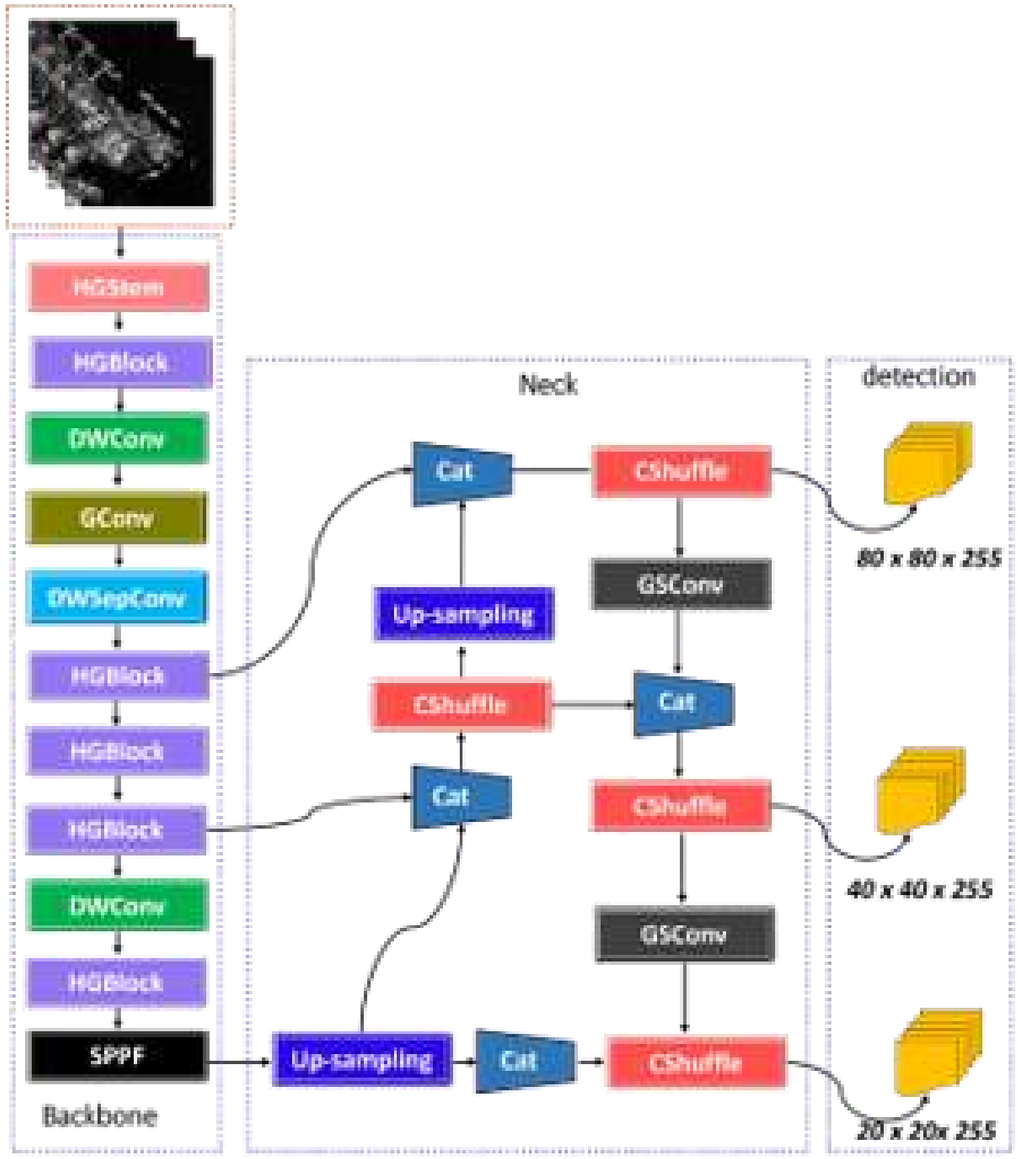
Muhammad Yasir et al.



- **Problema**: rastrear pequenas embarcações em sequências de imagens de radares de abertura sintética (SAR).
- **Proposta**: desenvolver um modelo
  - **mais preciso,**
  - **menor e**
  - **menos complexo computacionalmente.**

YOLOv8 leve para rastreamento de navios em imagens SAR

Muhammad Yasir et al.



- Por ainda ser um modelo SOTA e a menor versão, modificou-se o YOLOv8n.
- O *backbone* e os módulos do pescoço e da cabeça foram trocados por estruturas menos complexas.
- Além disso, aplicou-se *knowledge distillation* do YOLOv8x para melhorar a precisão sem aumentar a complexidade.

YOLOv8 leve para rastreamento de navios em imagens SAR

Muhammad Yasir et al.

Model	P (%)	R (%)	F1 (%)	mAP50 (%)	FLOPs (G)	Params (M)	Size (MB)	FPS
YOLOv8n	89.2	85.6	87.4	90.1	8.3	3.02	5.63	254.1
ours	**92.9**	**90.7**	**91.8**	**95.4**	**4.5**	**2.05**	**3.92**	**339.7**

Model	HOTA (%)	MOTA (%)	MOTP (%)	IDF1 (%)	IDS	FPS
DeepSort	67.4	79.1	86.0	75.3	**47**	27
StrongSort	71.9	79.2	85.4	79.1	91	19
OC Sort	69.4	79.3	85.1	75.2	103	50
ByteTrack	70.6	76.9	82.9	80.0	111	53
ours	**72.8**	**79.9**	**87.9**	**80.7**	79	**81**

- **Resultados**: o modelo superou o original em todas as métricas, sendo **mais preciso, menor e mais eficiente computacionalmente**.
- Além disso, **superou algoritmos SOTA de rastreamento** em várias métricas, exceto no ID *Switch*.
- Resultados obtidos com um computador de alto desempenho com GPU.
- Trabalhos futuros explorarão técnicas de otimização para sua execução em tempo real em dispositivos com recursos limitados.

EdgeML aplicado à detecção de incêndios florestais

Pedro M. R. Pereira et al.

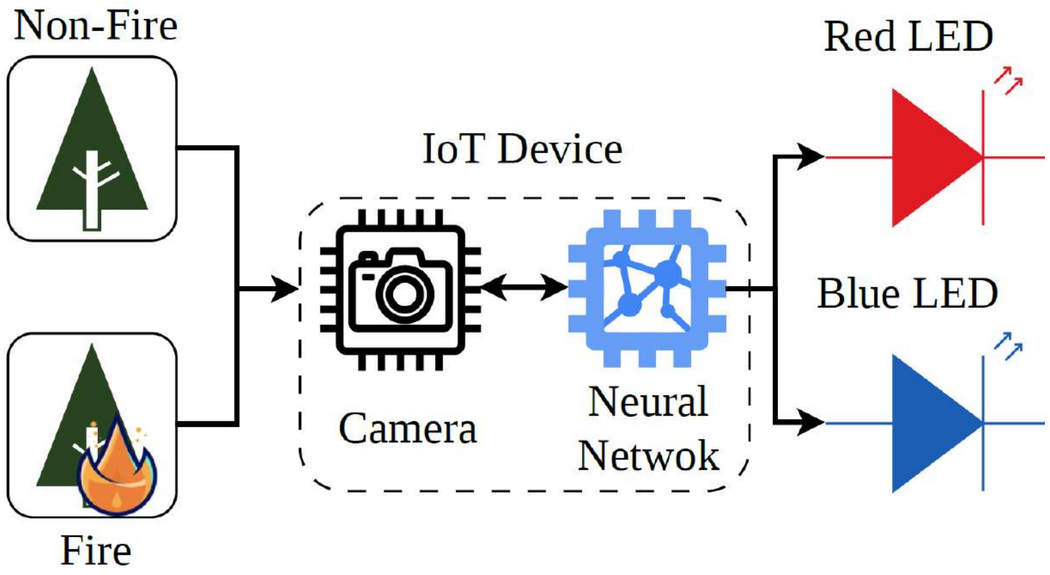
- **Problema**: métodos tradicionais de detecção de incêndios, como **vigilância humana em torres**, têm **limitações de cobertura, atraso na detecção e custos elevados**.
- **Proposta**: desenvolver um **sensor** EdgeML **preciso**, de **baixo custo, consumo e complexidade** para **detecção** de **incêndios em tempo real**.
- O sensor pode ser a base para uma rede de sensores espalhados por uma floresta.

WAI Inatel Labs | Wireless and Artificial Intelligence | xGMobile Centro de Competência EMBRAPII Inatel em Redes 5G e 6G | Inatel

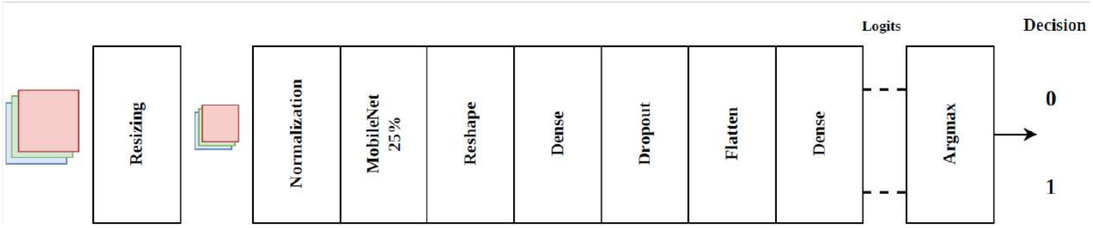
EdgeML aplicado à detecção de incêndios florestais

Pedro M. R. Pereira et al.

Arquitetura do sensor



Modelo proposto baseado no MobileNetv1



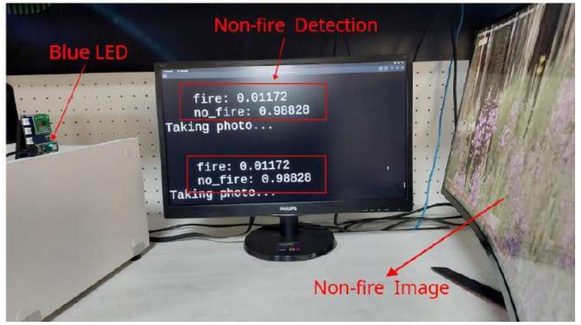
- O sensor usa câmera de 300 KPixels e Arduino Nano 33 BLE (ARM Cortex-M4, 32 bits, @64 MHz, 256 KB RAM e 1 MB Flash).
- O modelo proposto utiliza **25% das camadas do MobileNetv1** como **extrator de características** e adiciona **duas camadas densas** para a **classificação**.
- Usou-se *transfer learning* com pesos pré-treinados na ImageNet para reduzir a necessidade de grandes bases de dados.
- O modelo também é **quantizado em 8 bits** para redução de seu tamanho e complexidade.

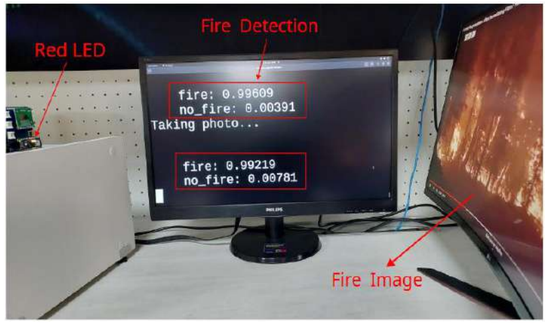
EdgeML aplicado à detecção de incêndios florestais

Pedro M. R. Pereira et al.

CLASSIFICATION METRICS FOR FLOATING POINT AND INTEGER MODELS.

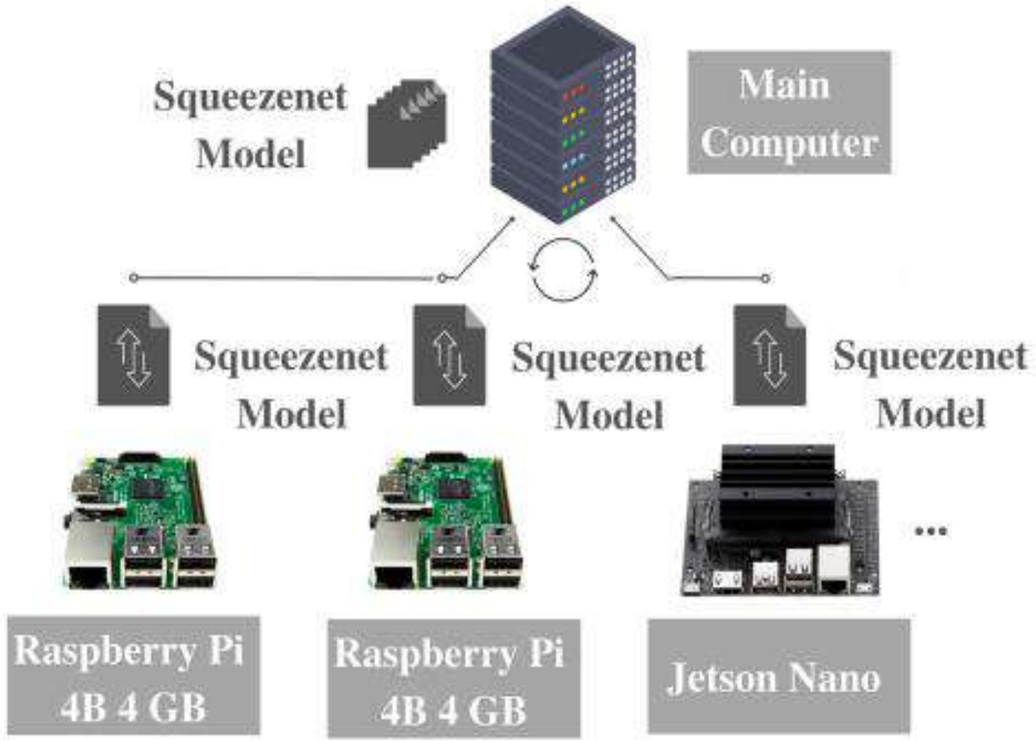
float-32 model					
	Class	Precision	Recall	F1-score	Support
	Fire	0.953	0.958	0.955	424
	Non-fire	0.957	0.952	0.954	418
Accuracy				0.955	842
Macro Average		0.955	0.955	0.955	842
Weighted Average		0.955	0.955	0.955	842
quantized model					
	Class	Precision	Recall	F1-score	Support
	Fire	0.976	0.950	0.963	424
	Non-fire	0.951	0.976	0.963	418
Accuracy				0.963	842
Macro Average		0.963	0.963	0.963	842
Weighted Average		0.964	0.963	0.963	842





- **Resultados**: o modelo **quantizado** apresenta **melhor desempenho**, **menor uso de recursos e tempo de inferência** que o original.
- Trabalhos futuros envolverão o uso de **câmeras com maior campo de visão e resolução**, **dispositivos mais baratos**, como Raspberry Pi Pico (4 USD) e a criação de uma **rede de sensores usando LoRa**.

Federated Learning, CV, and IoT for Edge Computing-based ship identification

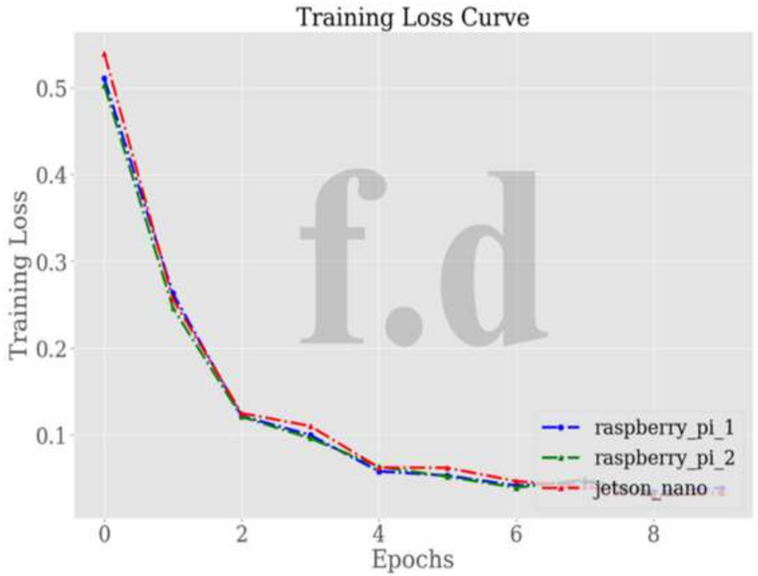


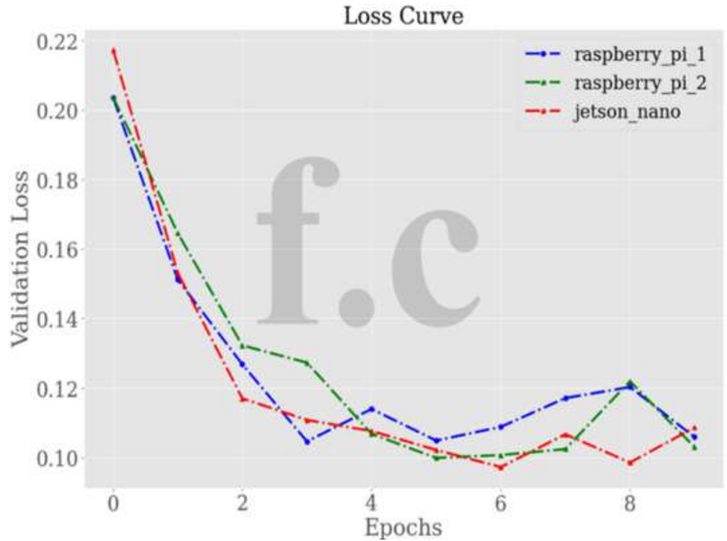
Problema central: rede de sensores visuais ao longo da costa para identificação de diferente tipos de embarcações em ambientes com ***dados privados***.

Metodologia: usamos modelos **squeezenet1.0** para classificar embarcações utilizando ***aprendizado federado*** para treiná-los de forma ***descentralizada***, sem compartilhar dados sensíveis, em dispositivos embarcados como raspberry pi 4 e Jetson Nano.

- Dataset com 6596 imagens para treinamento (aplicou-se *data aumentation*, indo para 17648) e 1536 para teste.

Federated Learning, CV, and IoT for Edge Computing-based ship identification

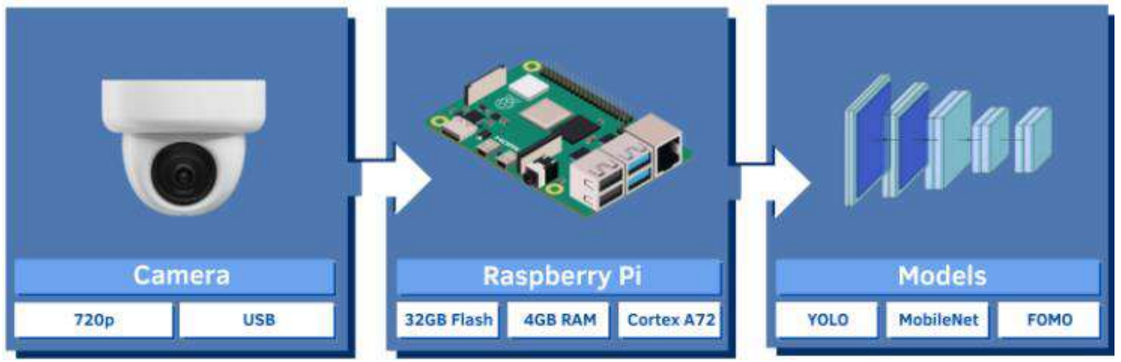




Resultados: a acurácia dos modelos atinge cerca de 98%, mas percebe-se que está possivelmente ocorrendo ***sobreajuste***, pois o erro de treinamento diminui enquanto o de validação permanece constante ao longo das épocas.

Conclusões: O uso de ***federated learning*** para o treinamento distribuído de modelos é uma boa estratégia quando os dados são confidenciais. Entretanto, precisa-se investigar o sobreajuste.

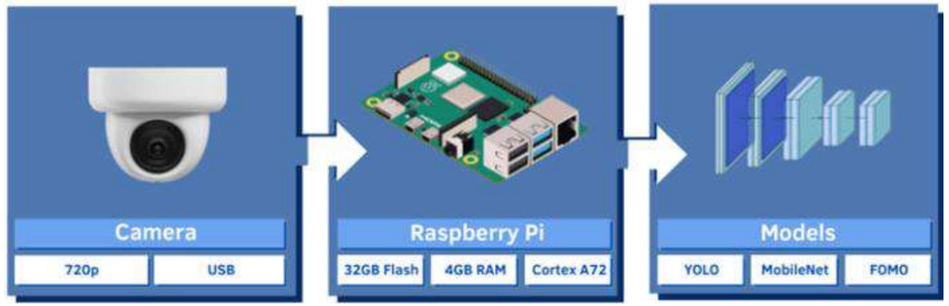
EdgeML-based Maritime Surveillance on a Raspberry Pi



Problema central: vigilância baseada em humanos tem se tornado inadequada devido ao crescente volume e complexidade do tráfego marítimo, além da sobrecarga dos operadores.

Modelos de ***detecção e rastreamento de objetos*** podem auxiliar operadores a tomarem decisões mais informadas e diminuir sua sobrecarga.

EdgeML-based Maritime Surveillance on a Raspberry Pi



Metodologia: comparar modelos de ***detecção de embarcações*** com diferentes níveis de ***quantização*** e identificar o mais adequado para vigilância marítima baseada em ***dispositivos com recursos limitados***, para criar uma ***rede de sensores*** que auxilie a tomada de decisões.

- **Dataset**: 5000 imagens de 10 classes diferentes divididas em 70%, 20%, 10%.
- **Dispositivos**: Raspberry Pi 4 com ARM Cortex A72 (Quad Core) @1.5 GHz e 4 GB de RAM e uma câmera de alta definição de 720p.
- **Modelos**: FOMO, MobileNet SSD e YOLOs v5, 8, 10.

EdgeML-based Maritime Surveillance on a Raspberry Pi

Model	Type	FPS	CPU	RAM
FOMO	float32	50	13%	150 MB
	int8	**60**	**12%**	**110 MB**
MobileNetV2 SSD FPN Lite 320x320	float32	5	32%	500 MB
	int8	7	30%	400 MB
YOLOv5n	float32	2	71%	860 MB
	float16	5	65%	800 MB
	int8	6	15%	210 MB
YOLOV8n	float32	2.3	25%	450 MB
	float16	2.9	23%	410 MB
	int8	3	20%	400 MB
YOLOV10n	float32	1	90%	500 MB

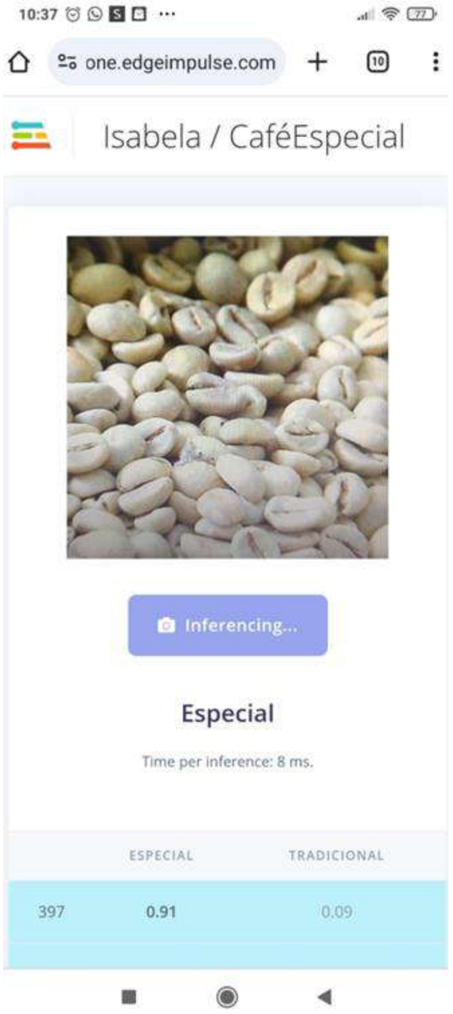
Model	Type	mAP	Precision	Recall	F1-score	Size
FOMO	float32	-	58%	74%	64%	83 kB
	int8	-	57%	72%	64%	**55 kB**
MobileNetV2 SSD FPN Lite 320x320	float32	84%	90%	82%	86%	10.9 MB
	int8	82%	85%	82%	83%	3.58 MB
YOLOv5n	float32	91%	**93%**	81%	87%	3.63 MB
	float16	91%	92%	80%	86%	3.48 MB
	int8	83%	80%	70%	75%	2 MB
YOLOV8n	float32	**93.1%**	89%	88%	**88%**	11.89 MB
	float16	92%	85%	88%	86%	5.99 MB
	int8	91%	86%	89%	87%	3.1 MB
YOLOV10n	float32	92.7%	87%	**90%**	**88%**	5.6 MB

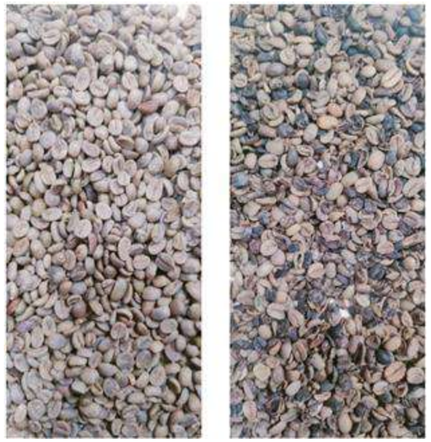
Resultados: FOMO é o modelo de menor tamanho e consumo de RAM/CPU com o maior número de FPS, mas tem o menor f1-score.

Em geral, aplicações de vigilância marítima não exigem alto FPS devido à velocidade das embarcações. Assim, os YOLOv5n e v8n int8 são os mais adequados para a detecção de navios, devido à sua boa precisão, FPS razoável e consumo moderado de recursos. Entretanto, apenas o YOLOv8n pode rastrear objetos.

O YOLOv10n apresenta bom desempenho, mas o menor FPS, além de não suportar quantização.

Tiny Machine Learning for Classifying Specialty Coffees





Problema central: a classificação de tipos de café, especialmente os especiais, é uma tarefa sujeita a interferência humana e subjetividade, podendo envolver interesses conflitantes.

Este estudo analisa o uso de VC para criar uma solução que classifique cafés especiais, eliminando subjetividade e aumentando a precisão.

Metodologia: comparação do desempenho de modelos de classificação de imagens (MobileNetv1, v2 e EfficientNet) com o intuito de usar o melhor em uma ***aplicação móvel*** para classificar cafés especiais.

Tiny Machine Learning for Classifying Specialty Coffees

Model	Accuracy		Loss
	Validation	Train	
MobileNetV2	100%	100%	0.03
MobileNetV1	78.8%	64.29%	0.46
EfficientNet	84.8%	92.86%	0.26

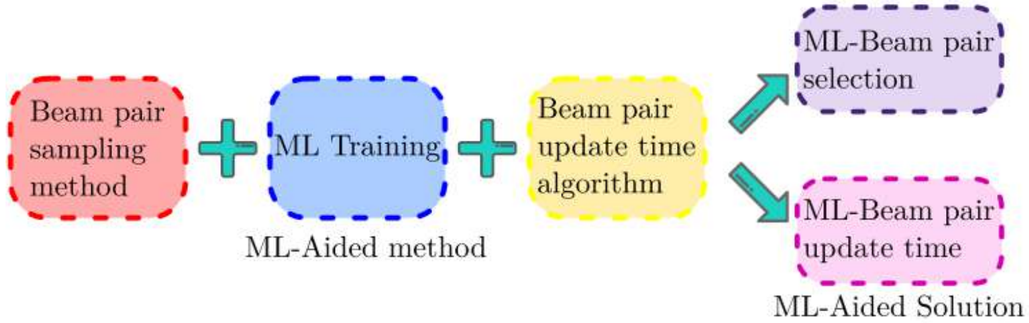
Metodologia:

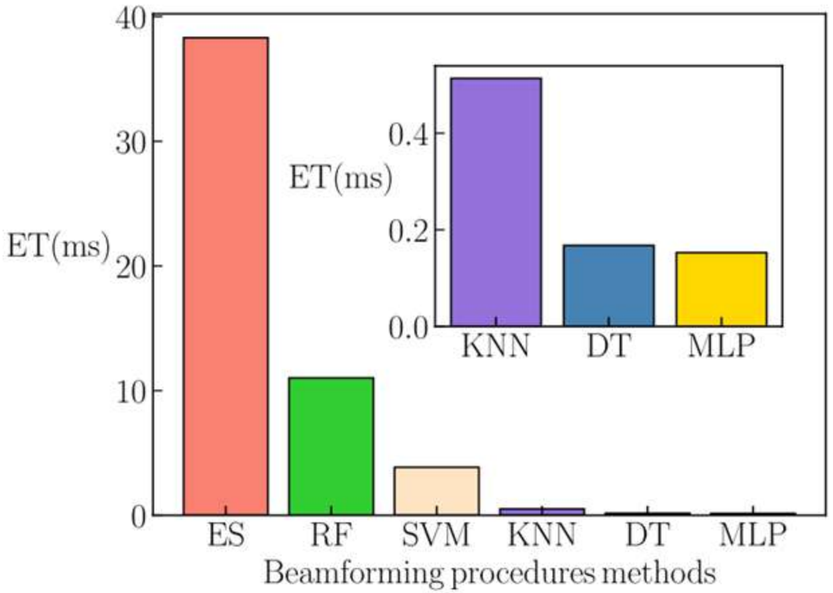
- Dataset com 176 imagens (162 train/15 test).
- Data augmentation foi aplicado às imagens de treinamento (flipping, cropping e shearing).

Resultados: o MobileNetv2 é o modelo com melhor desempenho, atingindo 100% de acurácia em ambos os conjuntos e ocupa apenas 14 MB. Porém, é necessário um estudo com uma base maior e mais classes.

Conclusões: desenvolveu-se um uma aplicação para smartphones inovadora e barata. A aplicação tem o potencial de ***reduzir a subjetividade da classificação*** de cafés especiais.

ML-aided method for optimizing beam selection and update period in 5G networks and beyond



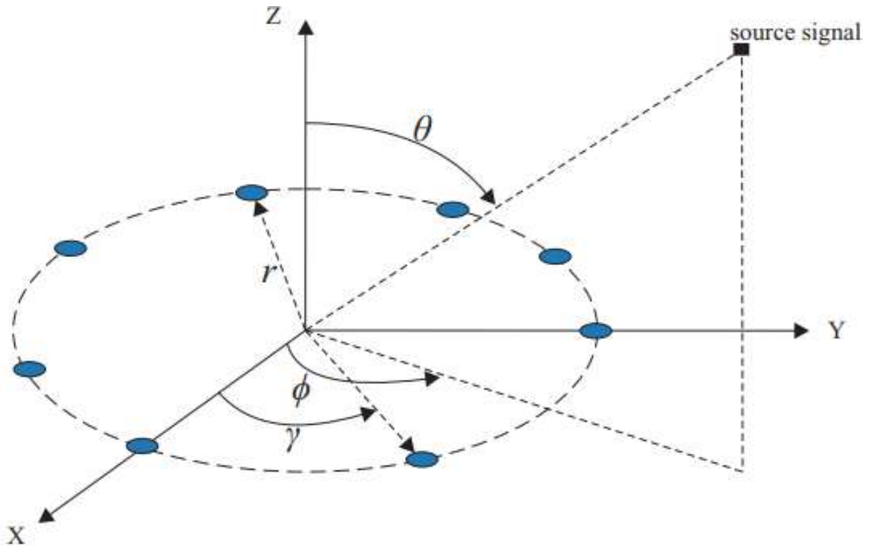


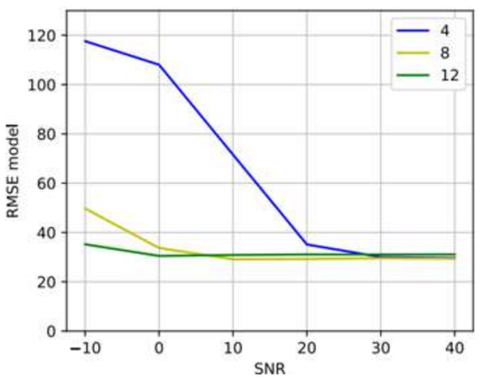
Problema: o tempo de seleção do par de feixes em sistemas 5G e além é alto e o período para uma nova seleção é constante.

Solução: Usamos as predições e as regiões de decisão de uma rede MLP para reduzir o tempo e a complexidade da ***seleção e atualização do par de feixes***, aumentando a eficiência espectral.

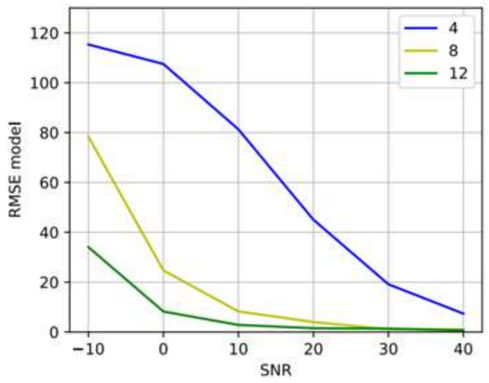
Resultados: redução de 99.5% no tempo de seleção. Redução de 85.7% na frequência de procedimentos de seleção de pares de feixes. Aumento de 4% na vazão de dados.

Simultaneous Estimation of Azimuth and Elevation Angles Using a Decision Tree-Based Method





MUSIC



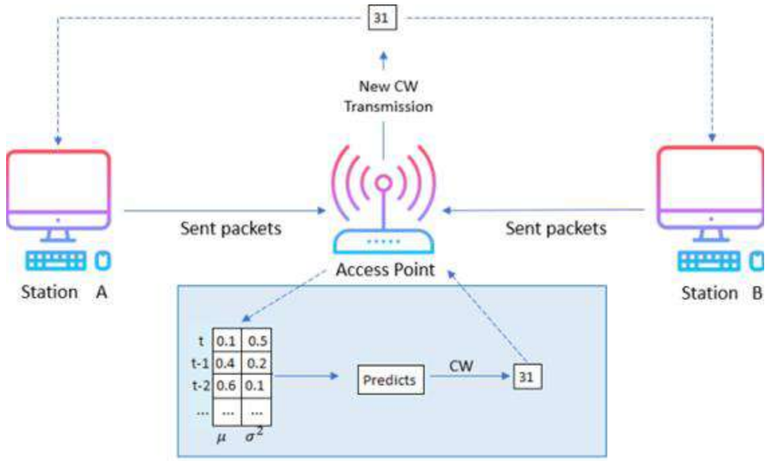
DT

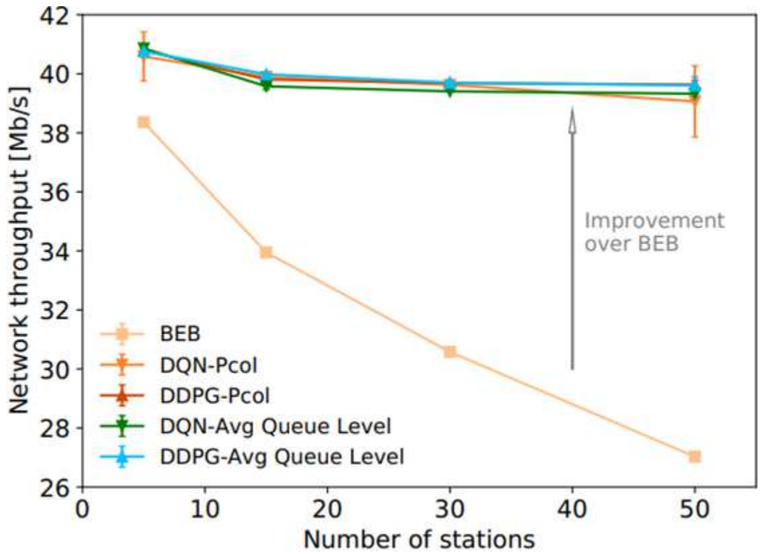
Problema: utilizar os sinais transmitidos por estações rádio base 5G e além para ***estimar a direção de dispositivos***. Aplicações em direcimento de feixes e localização e posicionamento.

Solução: usar árvores de decisão e os sinais recebidos por um sistema com múltiplas antenas para estimar os ângulos de azimute e elevação de dispositivos.

Resultado: redução de mais de 90% no erro de estimação e 50% no tempo de predição em relação ao algoritmo MUSIC.

Reinforcement Learning-based Wi-Fi Contention Window Optimization



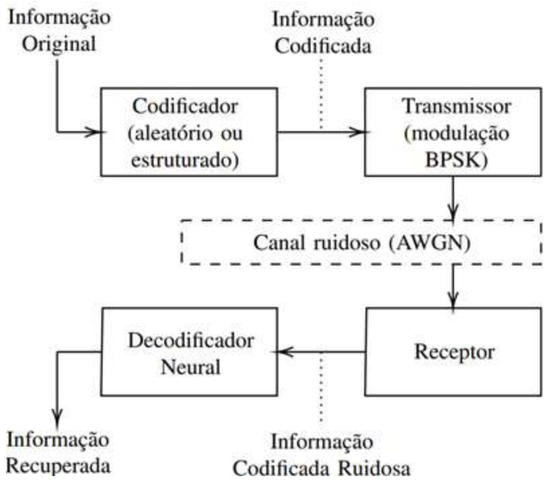


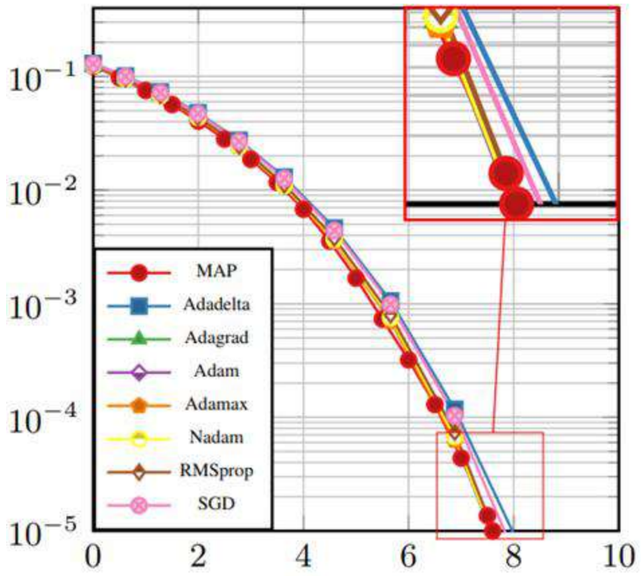
Problema: o mecanismo de prevenção de colisões do padrão WiFi não é ótimo, principalmente para caso onde a rede precisa servir muitas estações. Ele aumenta o *backoff* ao detectar uma colisão, diminuindo a utilização do espectro de rádio.

Solução: utilizar aprendizado por reforço para otimizar o uso do espectro minimizando o número de colisões.

Resultado: aumento de 45.52% na vazão de dados com 50 estações quando comparado com o mecanismo padrão.

Otimizando o Treinamento e a Topologia de um Decodificador de Canal baseado em Redes Neurais





Problema: a decodificação de canal, principalmente em SW, é muito custosa computacionalmente, impactando na vazão de redes 5G e 6G.

Solução: encontrar modelos de *deep learning* que decodifiquem códigos polares com baixa complexidade e desempenho próximo ao dos decodificadores human-engineered.

Resultado: uma rede com 4 camadas densas e 47.848 parâmetros usando otimizador Nadam atingiu desempenho similar ao decodificador human-engineered. Próximos estudos irão comparar o tempo de inferência e decodificação.

ML-based Novelty Detection and Classification of IoT Threats using Network Traffic Analysis

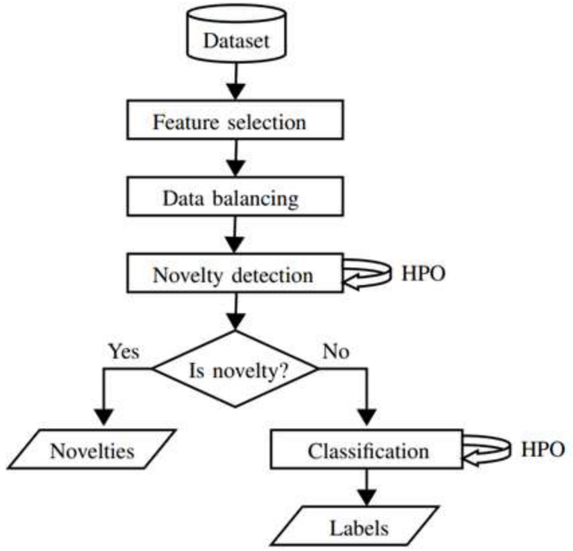


TABLE I: Novelty detection accuracy.

Scenario	Elliptic Envelope		Isolation Forest		Local Outlier Factor		SGD One-Class SVM	
	Mean ± SD	Max	Mean ± SD	Max	Mean ± SD	Max	Mean ± SD	Max
DoS	0.557 ± 0.494	**1.000**	0.594 ± 0.465	**1.000**	0.109 ± 0.105	0.263	0.620 ± 0.485	**1.000**
Mirai	0.747 ± 0.293	0.991	0.913 ± 0.087	0.993	0.414 ± 0.128	0.574	0.650 ± 0.477	**1.000**
MITM	0.391 ± 0.395	**1.000**	0.074 ± 0.093	0.474	0.208 ± 0.119	0.330	0.740 ± 0.439	**1.000**
Scan	0.537 ± 0.466	**1.000**	0.828 ± 0.238	0.949	0.601 ± 0.311	0.894	0.595 ± 0.491	**1.000**

TABLE III: Classification accuracy.

Scenario	Decision Tree		LightGBM		Random Forest		XGBoost	
	Mean ± SD	Max	Mean ± SD	Max	Mean ± SD	Max	Mean ± SD	Max
DoS	0.875 ± 0.141	**0.969**	0.765 ± 0.241	0.954	0.917 ± 0.054	0.957	0.947 ± 0.008	0.957
Mirai	0.851 ± 0.162	**0.969**	0.781 ± 0.243	0.953	0.917 ± 0.057	0.958	0.948 ± 0.008	0.957
MITM	0.864 ± 0.152	**0.967**	0.798 ± 0.213	0.953	0.916 ± 0.060	0.959	0.947 ± 0.008	0.956
Scan	0.865 ± 0.154	**0.966**	0.776 ± 0.242	0.953	0.916 ± 0.057	0.957	0.947 ± 0.008	0.955

Problema: a quantidade de dispositivos IoT e ataques a eles só tem crescido. Portanto, identificar ataques conhecidos e detectar novos tipos de ameaças se faz extremamente importante em redes IoT.

Solução: propor uma metodologia baseada em ML para detecção de novidades e classificação de ataques.

Resultado: SGD One-Class SVM teve melhor desempenho na detecção de novidades, enquanto o Decision Tree foi o melhor modelo de classificação.

Perguntas?


xGMobile
Centro de Competência EMBRAPII
Inatel em Redes 5G e 6G
Inatel
WAI
Inatel Labs
Wireless
and Artificial
Intelligence

Obrigado!


xGMobile
Centro de Competência EMBRAPII
Inatel em Redes 5G e 6G
Inatel
WAI
Inatel Labs
Wireless
and Artificial
Intelligence

Backup e Figuras

xGMobile
Centro de Competência EMBRAPII
Inatel em Redes 5G e 6G

Inatel

WAI
Inatel Labs

Wireless
and Artificial
Intelligence

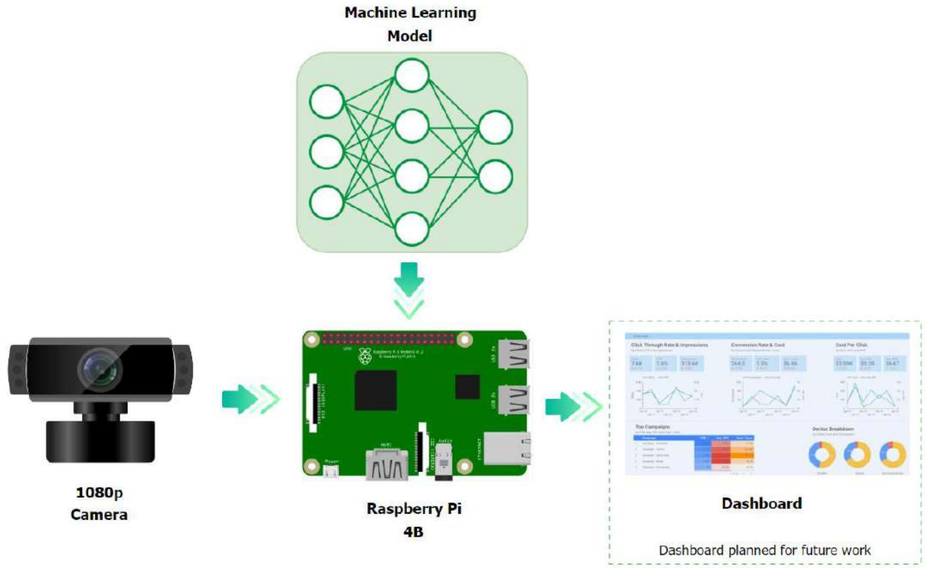
EdgeML Aplicado à Detecção de Defeitos em PCBs

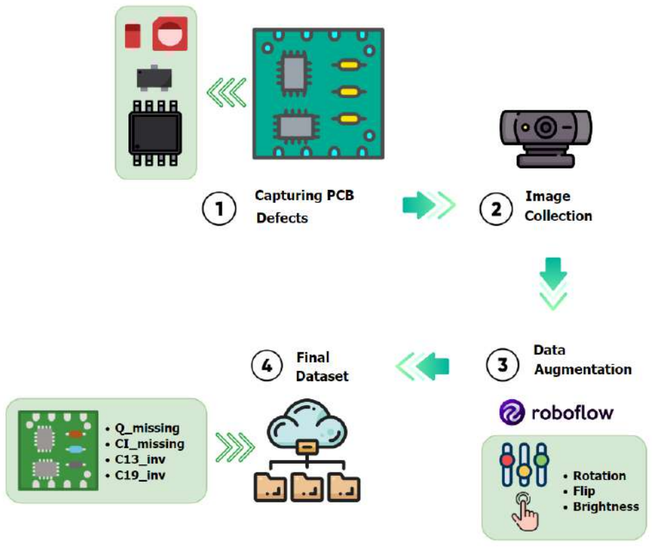
Felipe G. F. Rocha et al.

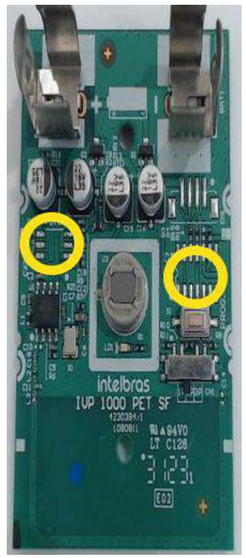
- **Problema**: a dependência de amostragem e de processos manuais para detecção de defeitos em PCBs é propensa ao erro devido à quantidade de amostras, subjetividade e limitações do foco e interpretação dos funcionários.
- **Proposta**: desenvolver uma solução EdgeML precisa, de baixo custo e complexidade e operando em tempo real para detecção de defeitos.

EdgeML Aplicado à Detecção de Defeitos em PCBs

Felipe G. F. Rocha et al.







- A solução usa uma câmera HD 1080p de 2 Mpx, uma Raspberry Pi 4 (ARM Cortex-A72 quad-core 64-bit @1.8 GHz com 4 GB de RAM).
- Criou-se uma base de dados com imagens de PCBs defeituosas e sem defeitos.
- Realizou-se um estudo comparativo entre 3 modelos de detecção de objetos: FOMO, YOLOv8 e MobileNetv2.

WAI Inatel Labs | Wireless and Artificial Intelligence | xGMobile Centro de Competência EMBRAPII Inatel em Redes 5G e 6G | Inatel

EdgeML Aplicado à Detecção de Defeitos em PCBs

Felipe G. F. Rocha et al.

TABLE I: YoloV8 results

	Size	CPU	RAM	FPS	Prec.	Recall	F1	mAP
Float 32	22 MB	30%	670 MB	0,2	96,8%	92,7%	94,6%	96,1%
Int 8	11 MB	20%	550 MB	0,3	94,3%	94,8%	94,5%	95,6%

TABLE II: FOMO results

	Size	CPU	RAM	FPS	Prec.	Recall	F1	mAP
Float 32	152 kB	25%	250 MB	10	97%	99%	98%	99%
Int 8	91 kb	20%	230 MB	20	97%	99%	98%	99%

TABLE III: MobileNetV2 results

	Size	CPU	RAM	FPS	Prec.	Recall	F1	mAP
Float 32	11 MB	30%	350 MB	3	61%	87,5%	71,8%	100%
Int 8	3 MB	25%	300 MB	6	55%	79%	65%	65%

- **Resultados**: dentre os modelos, o FOMO quantizado se destaca por apresentar melhor desempenho, menor uso de recursos computacionais e maior quantidade de FPS.
- Trabalhos futuros envolvem a criação de um dashboard para análise estatística dos defeitos com acesso remoto, uso de aprendizado federado/distribuído e testes práticos em linhas de produção.

Estimativa de canal veicular com redes neurais líquidas

Ana Flávia dos Reis et al.

- **Problema**: a comunicação veicular requer estimativas de canal precisas e eficientes, mas a dependência de muitos pilotos reduz a capacidade de dados úteis e aumenta a complexidade.
- **Proposta**: desenvolver um estimador de canal com alto desempenho e baixa complexidade computacional.

Estimativa de canal veicular com redes neurais líquidas

Ana Flávia dos Reis et al.

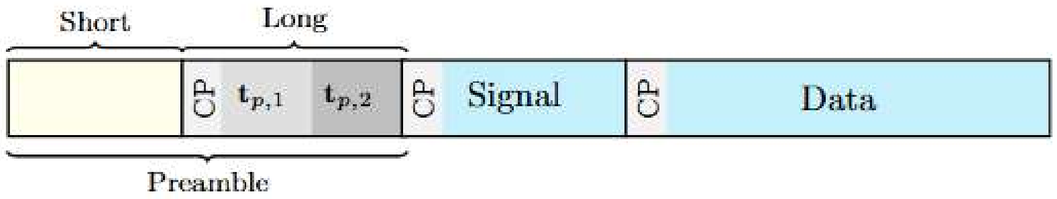
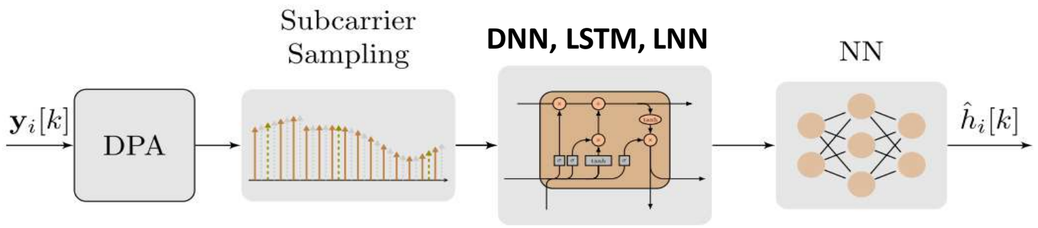


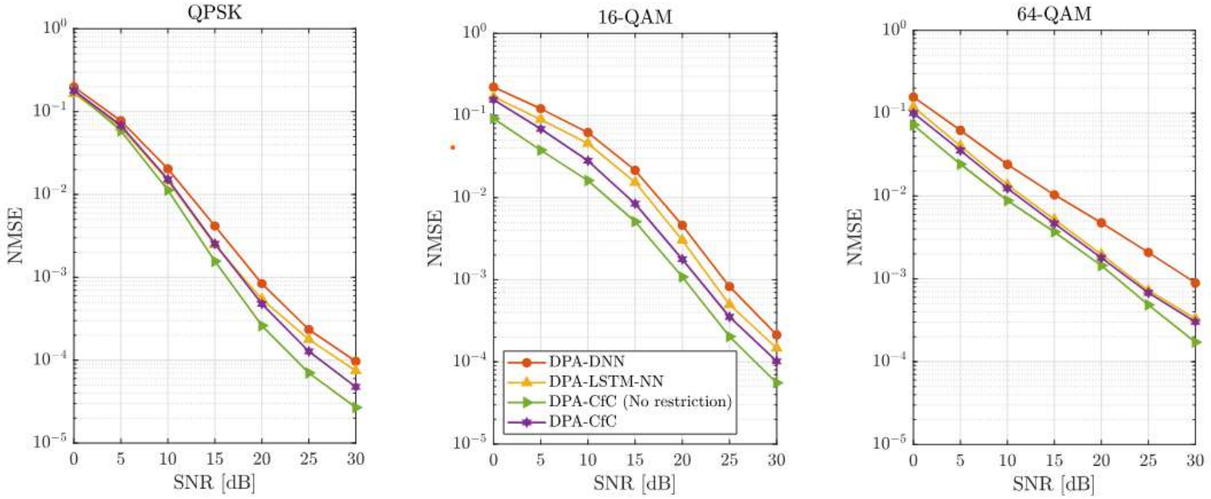
Fig. 1. IEEE 802.11p frame structure [13].



- O modelo do sistema baseia-se no padrão IEEE 802.11p.
- A estrutura do pacote consiste em um preâmbulo, um campo de sinal e um campo de dados.
- De um total de 52 subportadoras, apenas 4 são usadas como pilotos, enquanto as demais são de dados.
- Adota-se estimativa de canal auxiliada por dados e pilotos (DPA).
- A saída do DPA alimenta modelos baseados em camadas DNN, LSTM e LNN.
- Usa-se NAS para encontrar arquitetura com baixa complexidade e bom desempenho.

Estimativa de canal veicular com redes neurais líquidas

Ana Flávia dos Reis et al.

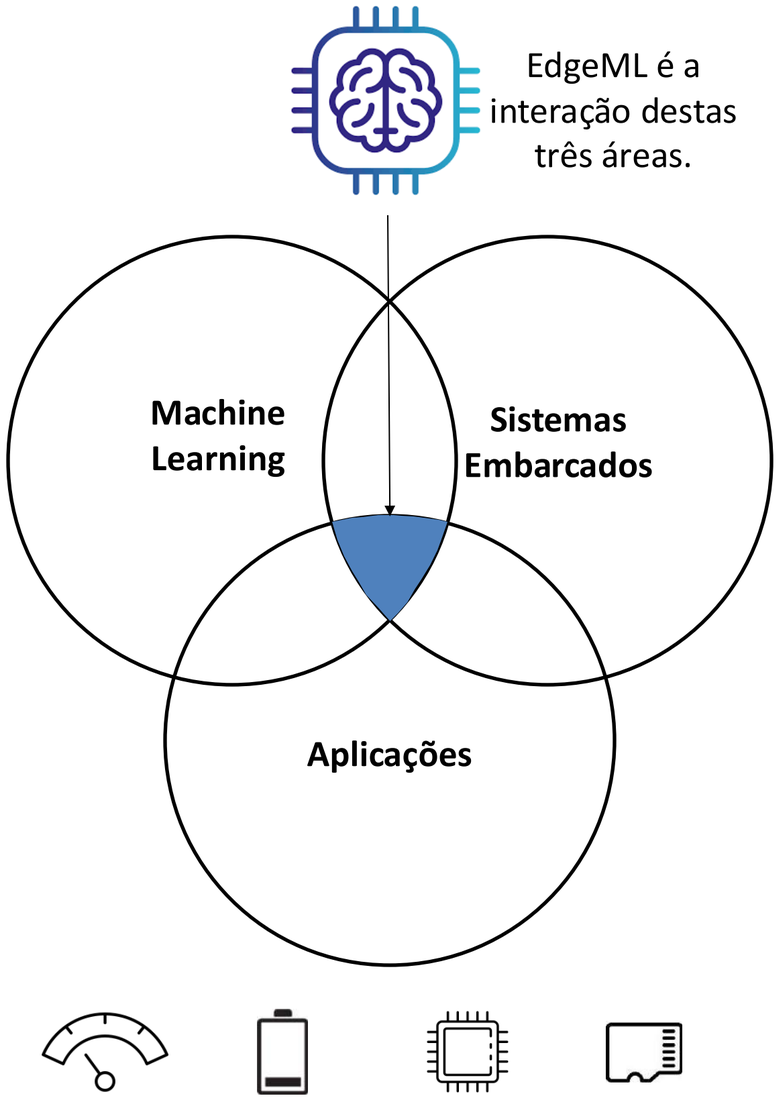


COMPUTATIONAL COMPLEXITY.

Method	Parameters	FLOPs
DPA-DNN [12]	10124	20044
DPA-LSTM-NN [6]	35115	161500
DPA-CfC (No restriction)	77186	370600
DPA-CfC	22764	126880

- **Resultados**: modelos baseados em camadas LNN superam os baseados em DNNs e LSTM.
- O uso de NAS com função de objetivo que penaliza modelos complexos encontrou um modelo de baixa complexidade e desempenho próximo ao do modelo baseado em LSTM.
- Trabalhos futuros estenderão a análise para sistemas MIMO e considerarão não linearidades de dispositivos, como amplificadores.

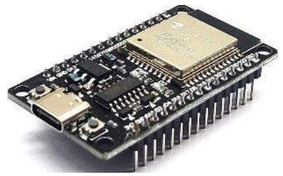
EdgeML



- A computação de borda traz o processamento e armazenamento de dados para perto de sua origem, melhorando desempenho, eficiência e segurança.
- Isso viabiliza o aprendizado de máquina local, permitindo que dispositivos façam predições e tomem decisões de forma autônoma.
- Contudo, devido a limitações de recursos, a otimização dos algoritmos é essencial para reduzir complexidade, latência e consumo de energia.

Comparação de hardwares embarcados para EdgeML

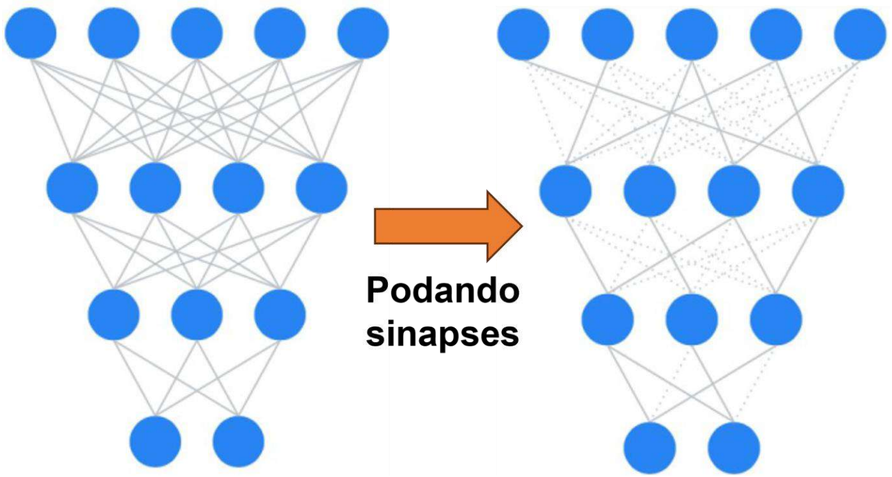


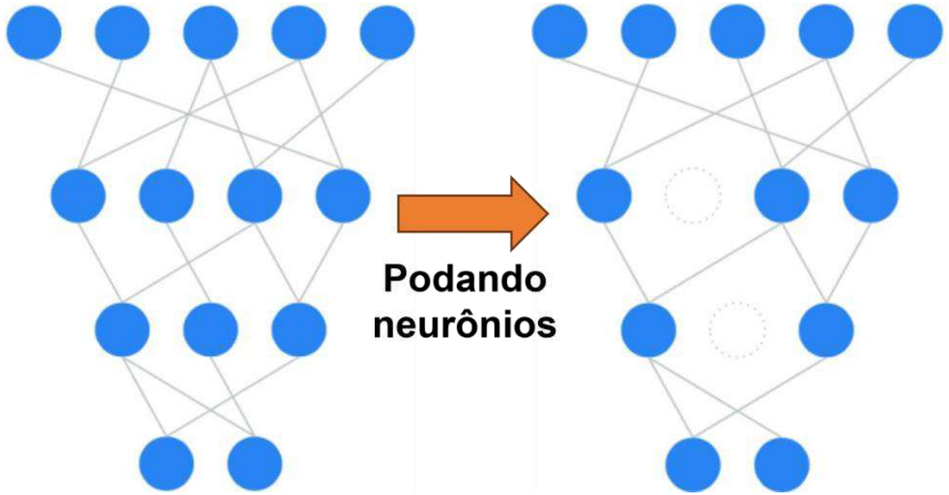


	Raspberry Pico	Arduino Nano Sense	ESP 32	Seeed Wio Terminal	Arduino Portenta
CPU*	Dual Core ARM Cortex M0+	Single Core ARM Cortex M4	Dual Core Cadence Xtensa LX6	Single Core ARM Cortex M4	Dual Core ARM Cortex M4/M7
Clock	133 MHz	64 MHz	240 MHz	120 MHz	240/480 MHz
RAM	264 KB	256 KB	520 KB	192 KB	1 MB***
Flash	2 MB	1 MB	2 MB	4 MB	2 MB
Rádio	-	BLE**	BLE/WiFi	BLE/WiFi	BLE/WiFi
Sensores	Não	Sim	Não	Sim	Não
Preço	~ $ 5,00	~ $ 40,00	~ $ 16,00	~ $ 44,00	~ $ 100,00

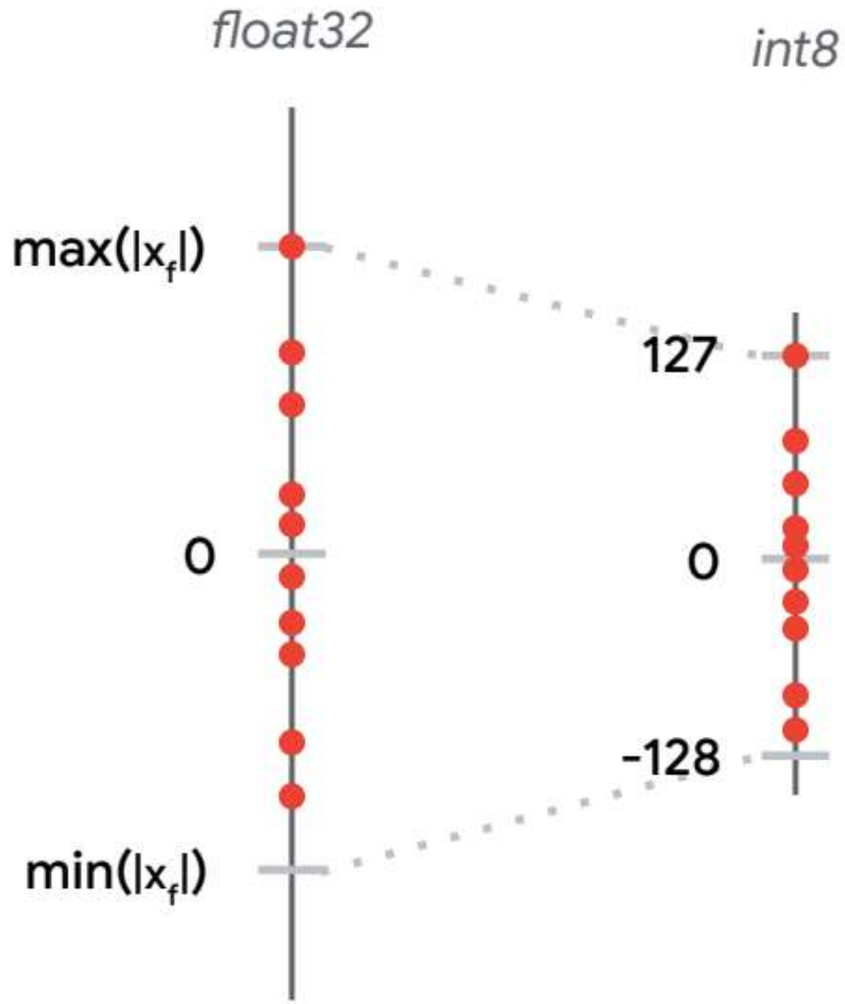
* Todas as CPUs são de 32 bits. ** Chip BLE multiprotocolo: suporta os protocolos ZigBee e Thread. *** É possível rodar aplicações de detecção de objetos.

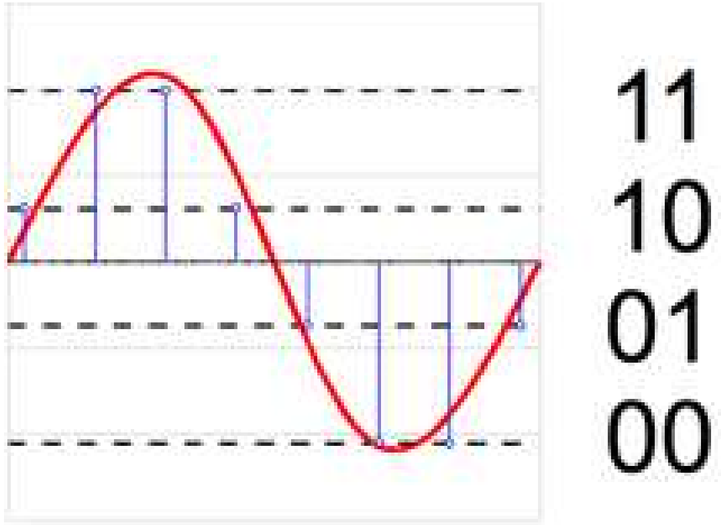
Pruning



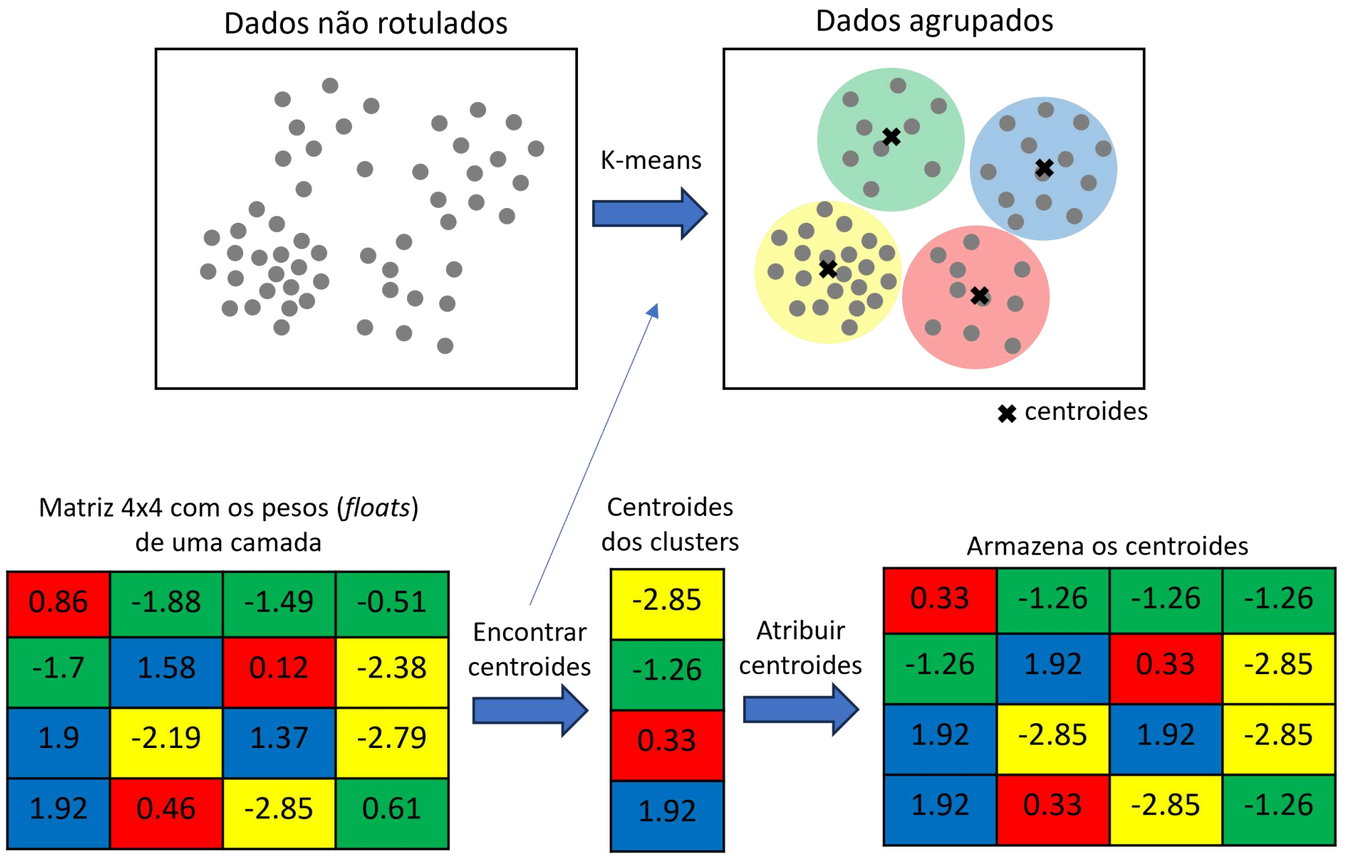


Quantização

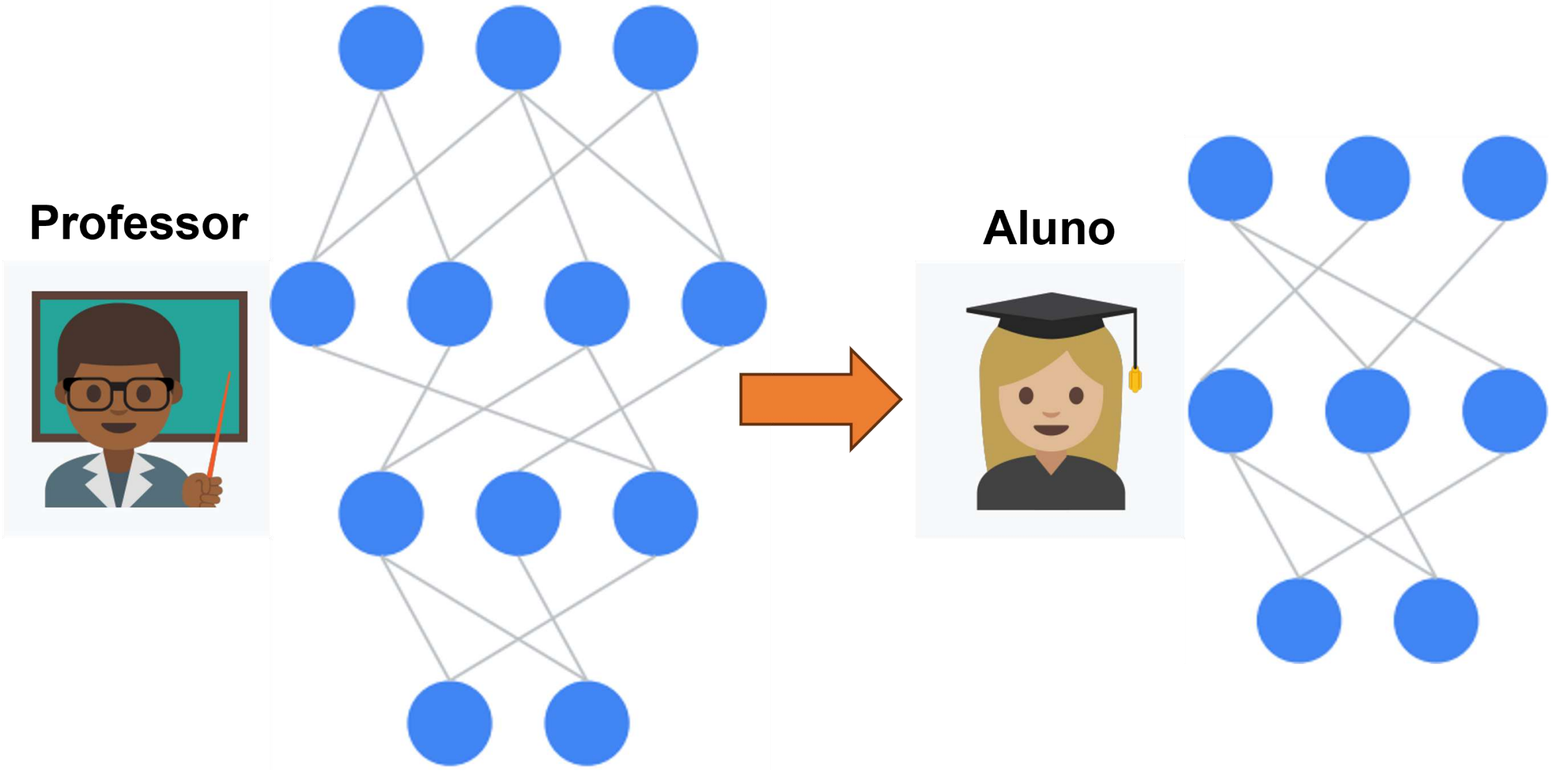


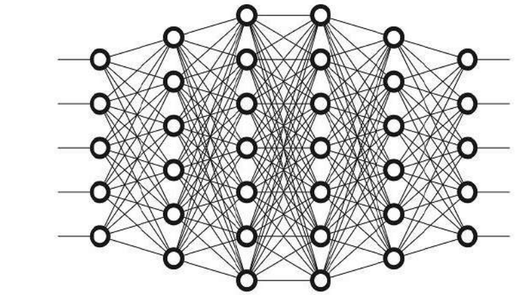


Clustering ou compartilhamento de pesos



Knowledge distillation





Cloud TPU


xGMobile
Centro de Competência EMBRAPII
Inatel em Redes 5G e 6G
Inatel


xGMobile
Centro de Competência
EMBRAPII Inatel em
Redes 5G e 6G
Inatel


xGMobile
Centro de Competência EMBRAPII
Inatel em Redes 5G e 6G
Inatel


xGMobile
Centro de Competência
EMBRAPII Inatel em
Redes 5G e 6G
Inatel