GTA / UFRJ
GRUPO DE TELEINFORMÁTICA E AUTOMAÇÃO

Politécnica UFRJ

COPPE 60 anos UFRJ

Computação na borda e aprendizado de máquina: aplicações e desafios

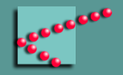
Rodrigo de Souza Couto
www.gta.ufrj.br

Realizado com recursos do CNPq, CAPES e FAPERJ

Poli – Depto. de Eng. Eletrônica e de Computação
COPPE – Programa de Engenharia Elétrica
Universidade Federal do Rio de Janeiro

Grupo de Teleinformática e Automação

GTA



GTA

Fundado em 1986

Imagem por W. Liller, retirada de Wikimedia

Qual objetivo?

Fomento de novas tecnologias e aplicações em redes de computadores

Quem faz parte?

Professores
(DEL/Poli e PEE/COPPE)

Estudantes
(Graduação, Mestrado e Doutorado)

Assuntos de interesse

Redes de computadores

- Sistemas distribuídos
- Internet das Coisas
- Inteligência artificial
- Segurança da informação
- Computação em nuvem/borda
- Redes veiculares
- Redes móveis
- ....

Equipe

Professores

Pedro Cruz

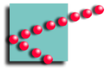
Rodrigo de
Souza Couto

Miguel Elias
M Campista

Luís Henrique
M K Costa

Resumo da minha trajetória acadêmica

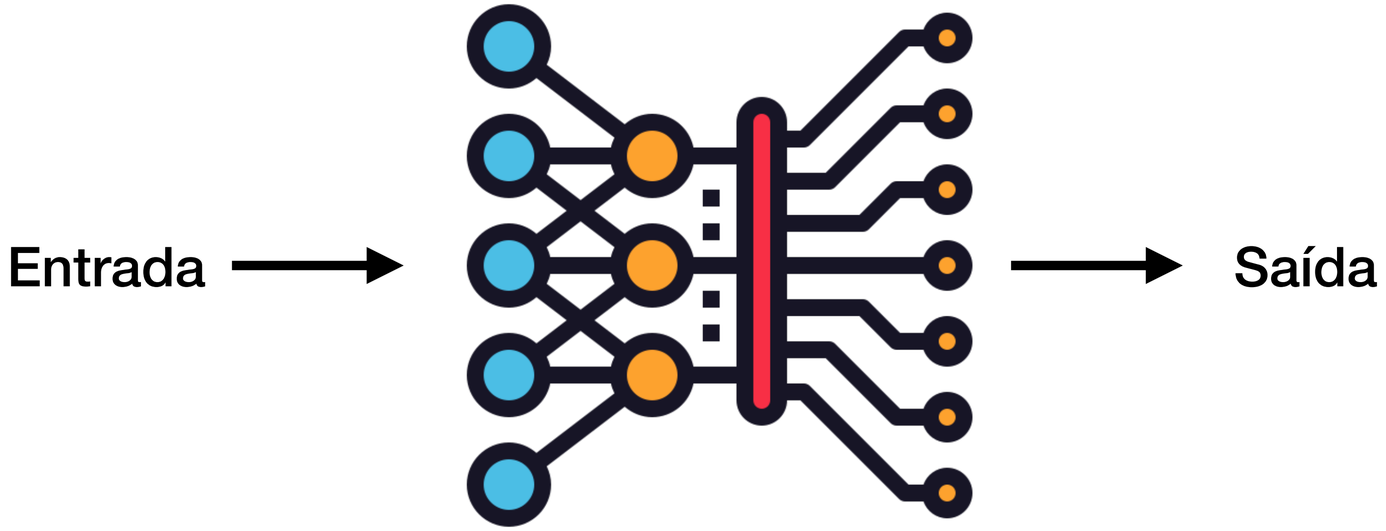
- Engenharia Eletrônica e de Computação - Poli/UFRJ - 2011
  - Iniciação científica no GTA desde 2008
- Doutorado em Engenharia Elétrica - COPPE/UFRJ - 2015
  - Período sanduíche na Sorbonne Université (Paris)
  - Orientado pelos Profs. Luís Henrique e Miguel do GTA/UFRJ
- Professor na UERJ entre 2015 e 2018
- Professor na UFRJ desde 2018



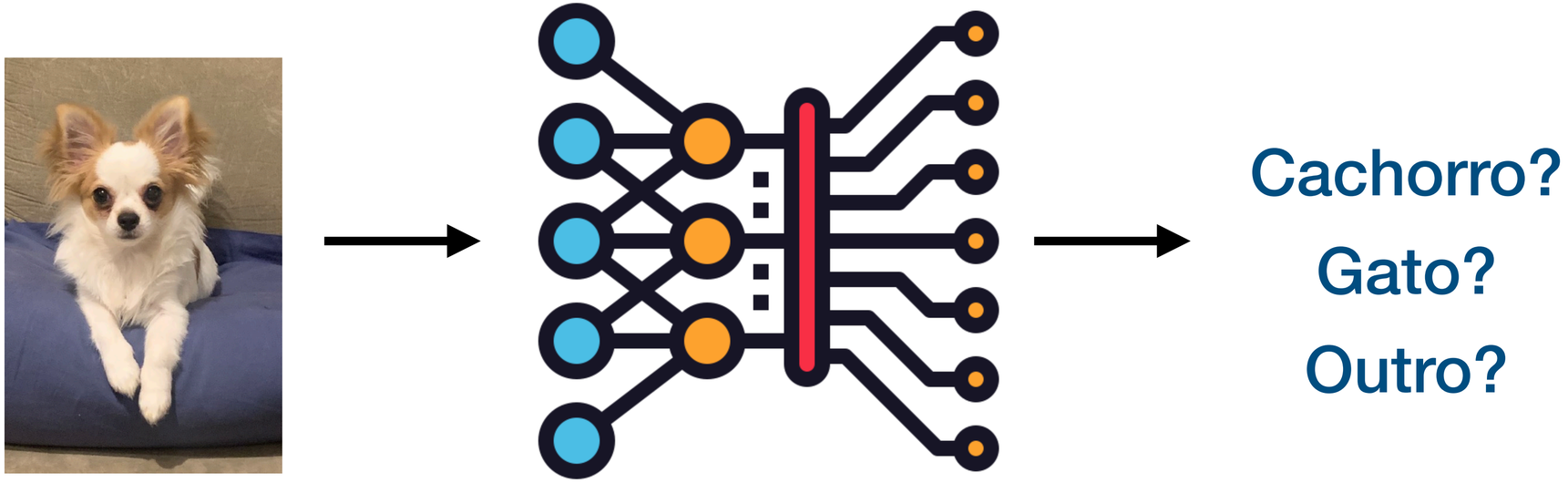
Pesquisa em dispositivos móveis e embarcados no GTA

- Redes Veiculares
- Infraestrutura para aprendizado de máquina
- Mobilidade de usuários
- Aprendizado federado
- Computação na borda

Deep Neural Network (DNN)

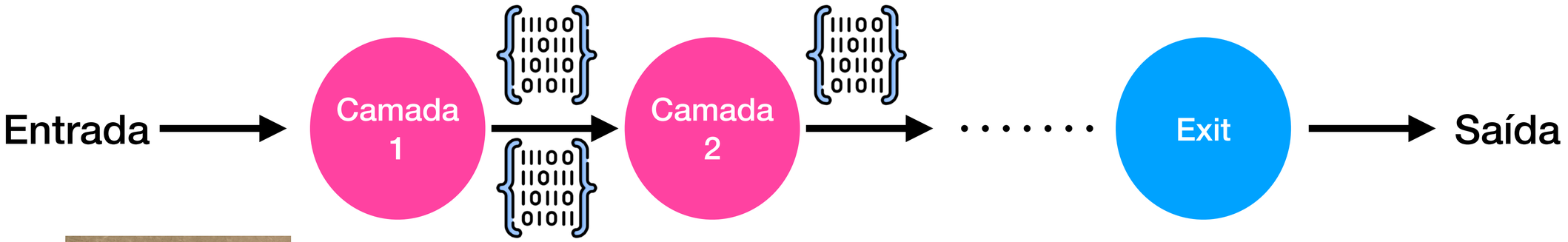


Inferência em DNNs

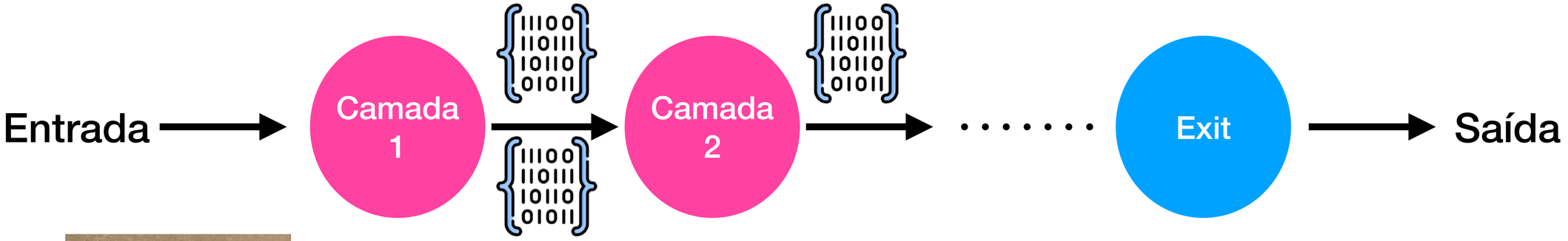


Inferência: Fornecer uma saída para classificar uma imagem de entrada

DNNs para classificação de imagens



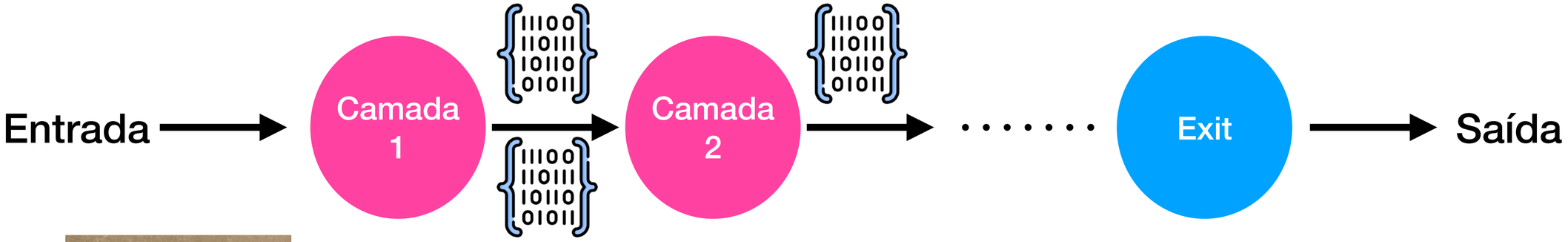
DNNs para classificação de imagens



Probabilidade

Cachorro	0.8
Gato	0.1
Outro	0.1

DNNs para classificação de imagens



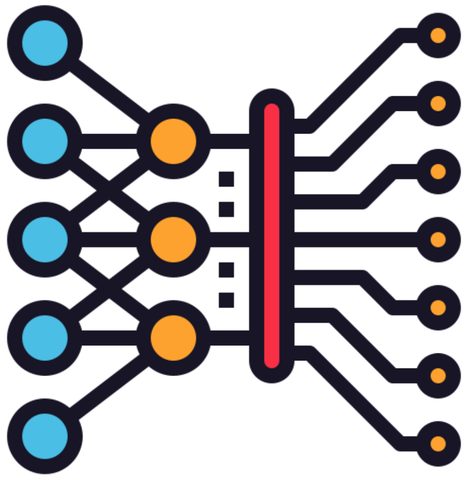
É um cachorro!

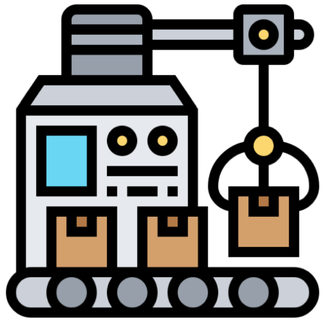
Probabilidade

Cachorro	0.8
Gato	0.1
Outro	0.1

DNNs em dispositivos móveis e embarcados

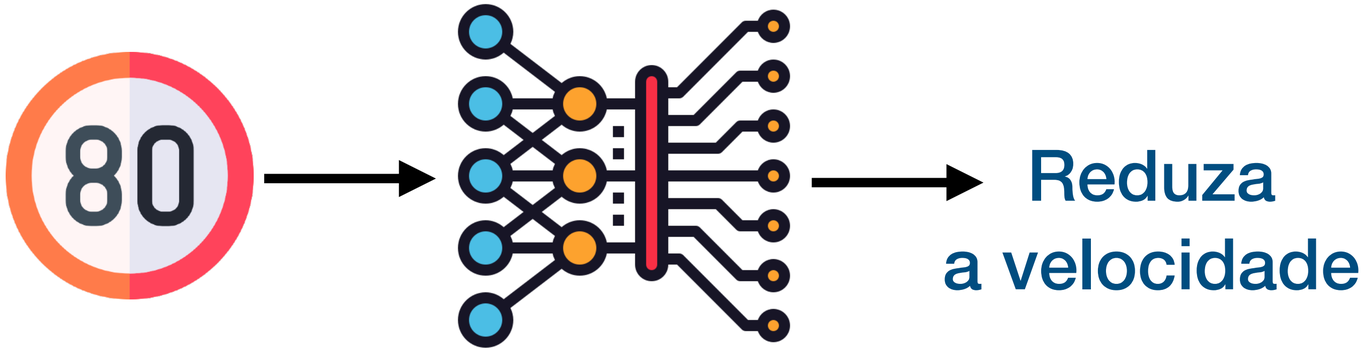
- Aplicações inteligentes em diversas áreas
  - Ubiquidade dos dispositivos





DNNs em dispositivos móveis e embarcados

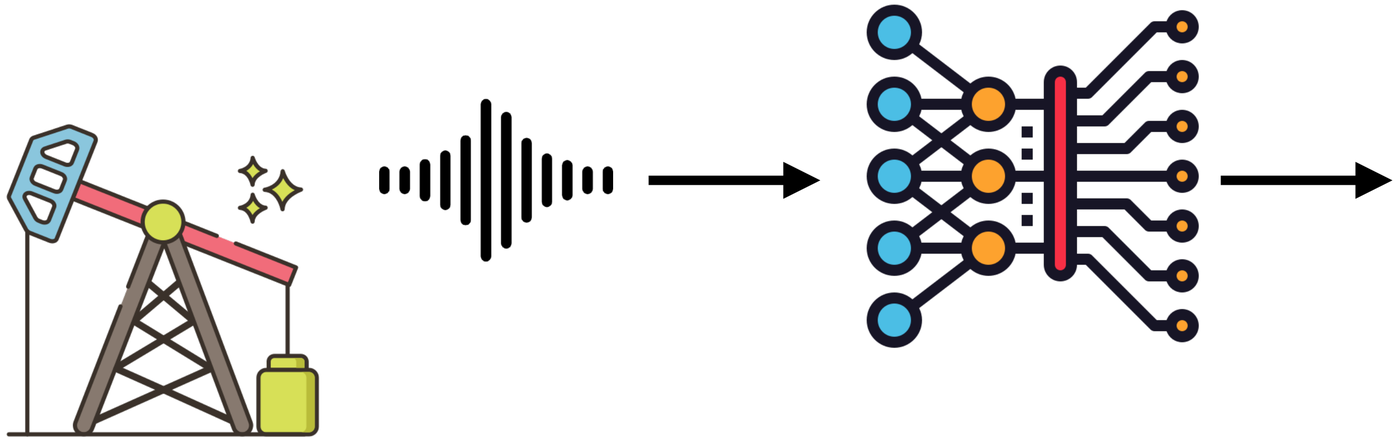
- Veículos inteligentes
  - Identificação de placas de trânsito
  - Processamento de sinais de sensores do veículo



Ortiz, F. M., Sammarco, M., Costa, L. H. M. K., & Detyniecki, M. (2022). Applications and Services Using Vehicular Exteroceptive Sensors: a Survey", em IEEE Transactions on Intelligent Vehicles

DNNs em dispositivos móveis e embarcados

- Sistemas industriais inteligentes
  - Identificação de problemas nas máquinas baseados no som

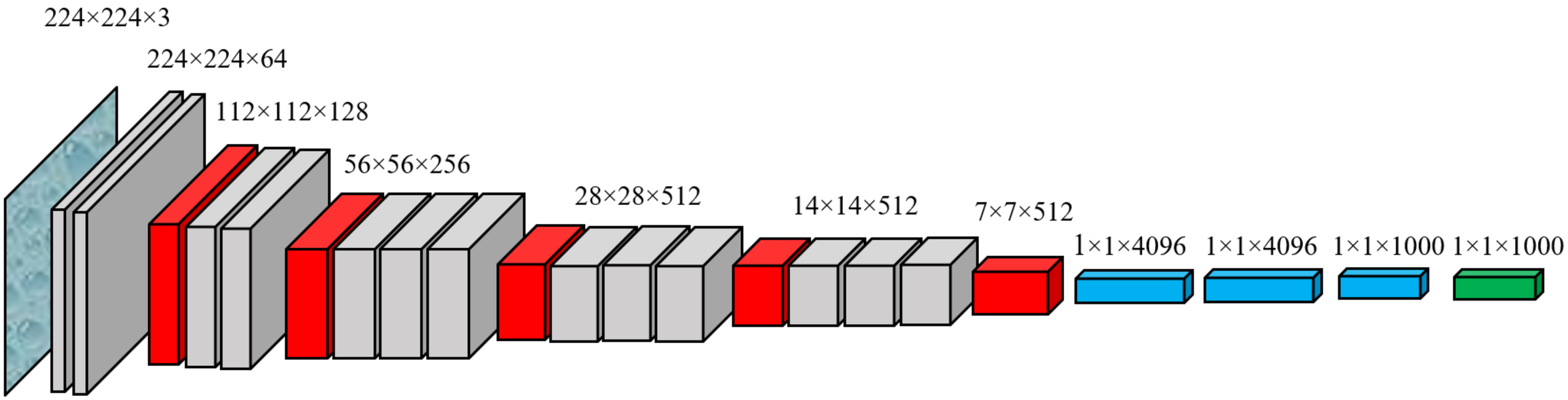


A bomba está vazando!

Gantert, L., Sammarco, M., Detyniecki, M. & Campista, M. E. M. (2022). Super Learner Ensemble for Sound Classification using Spectral Features, em IEEE Latin-American Conference on Communications (LATINCOM 2022)

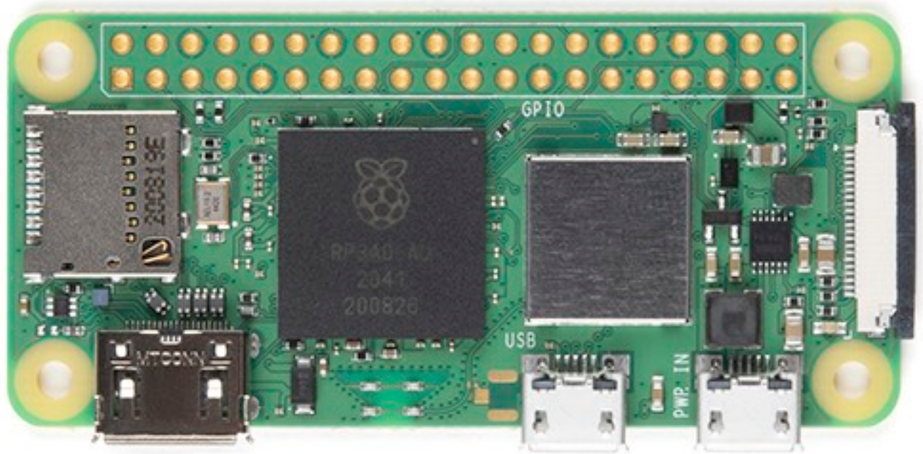
Desafios

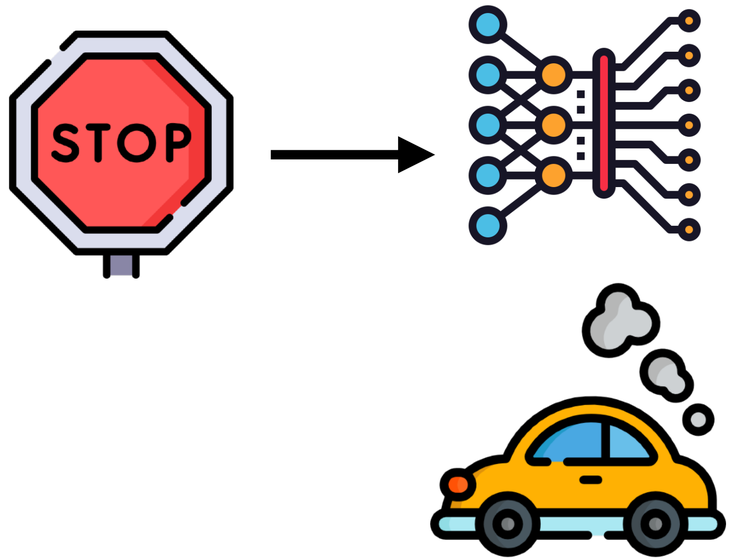
- DNNs estão cada vez mais complexas



Desafios

- Dispositivos móveis e embarcados
  - Muitas vezes possuem baixo poder computacional





Author:SparkFun Electronics

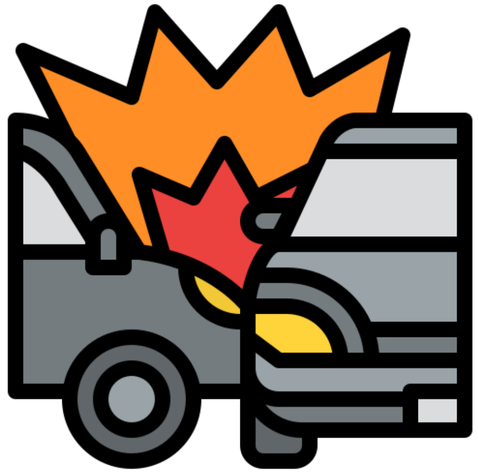
Source:https://commons.wikimedia.org/wiki/File:18713-Raspberry_Pi_Zero_2_W-04_(cropped).jpg

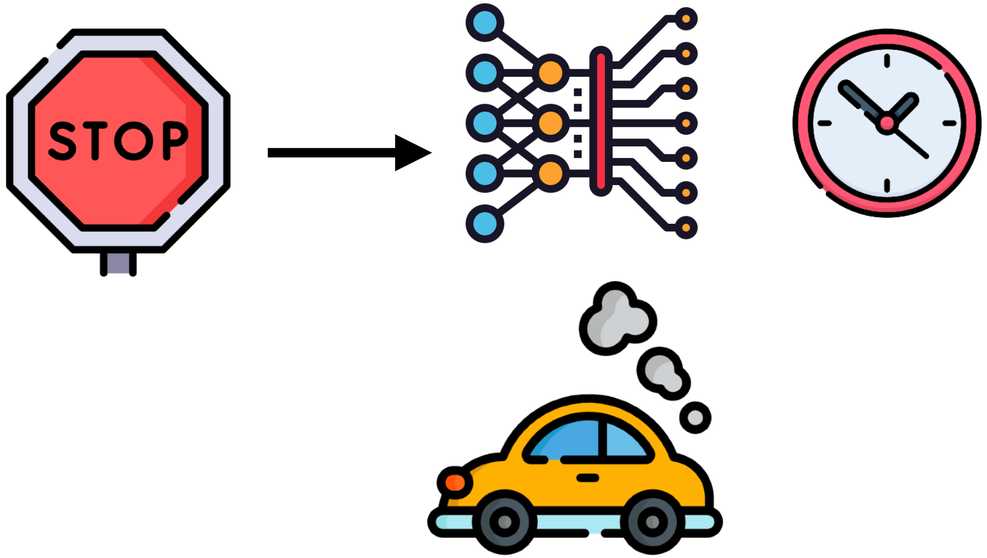
Icon made by Becris from www.flaticon.com

Icon made by Freepik from www.flaticon.com

Alto tempo de inferência

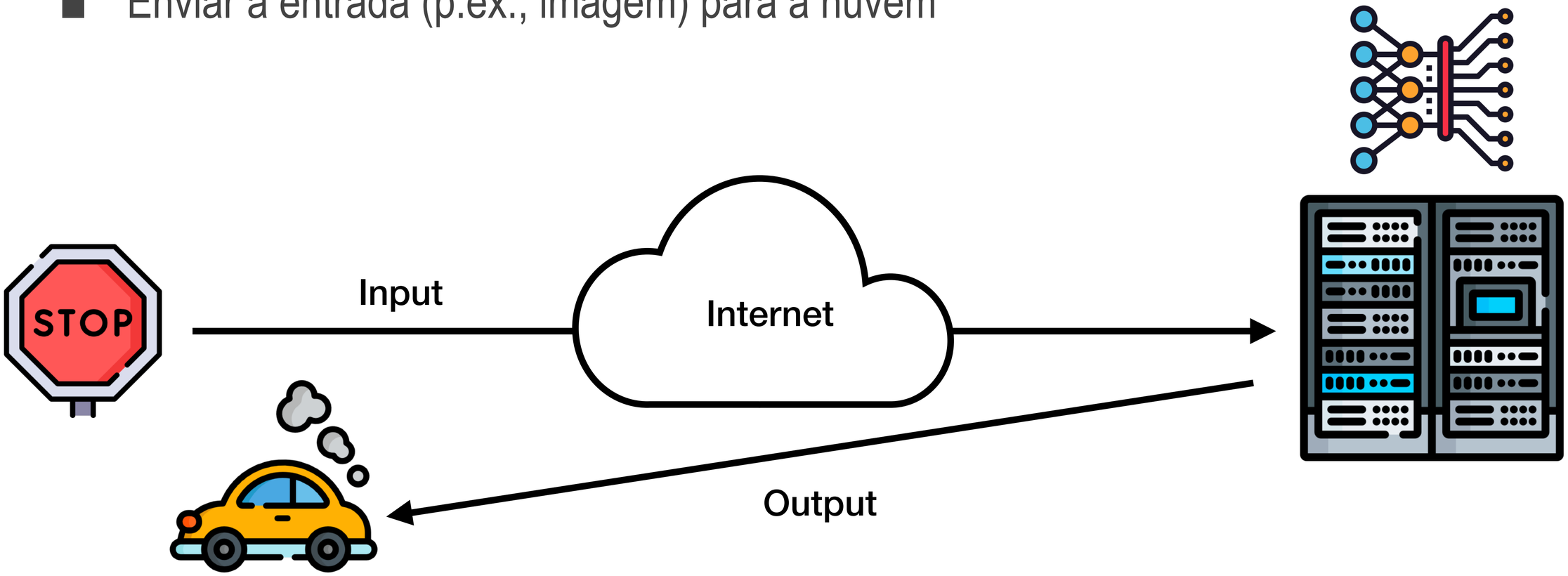
- Inviável para aplicações em tempo real





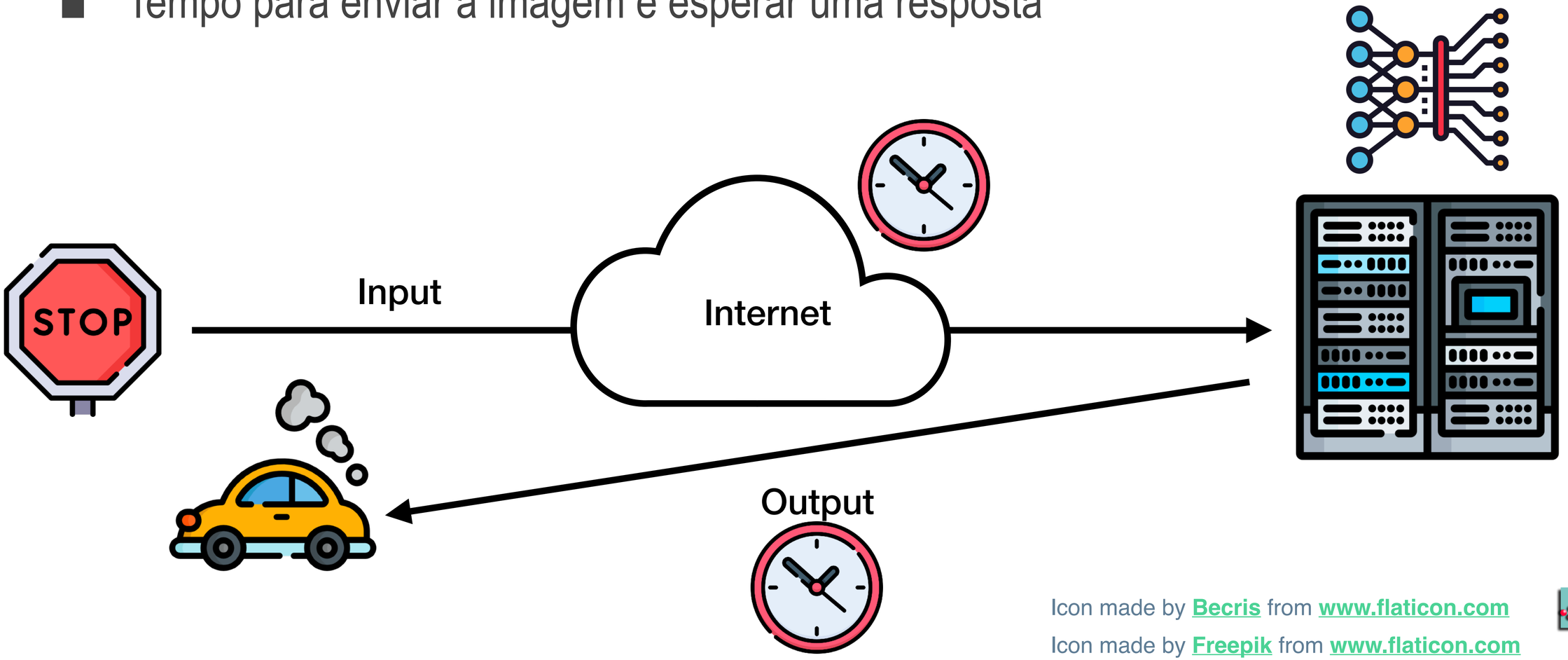
Possível solução

- Enviar a entrada (p.ex., imagem) para a nuvem



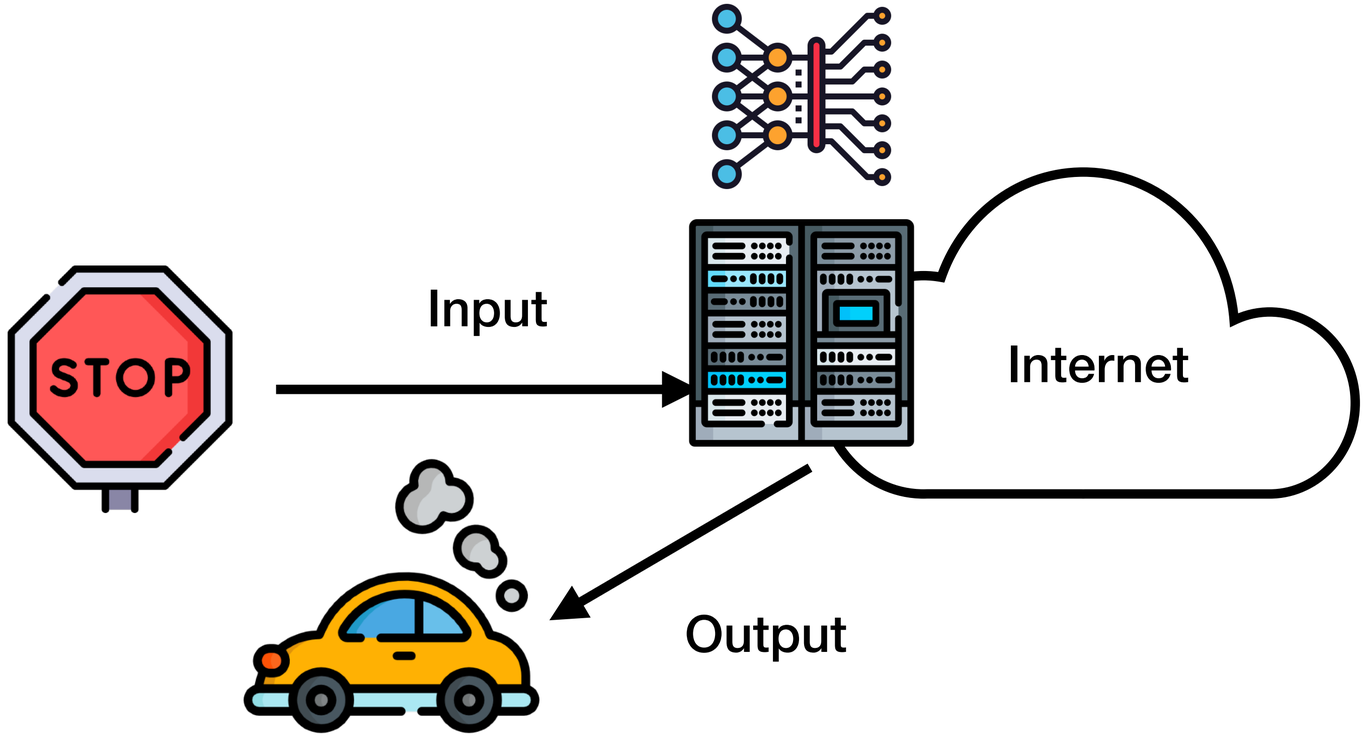
Alto tempo de inferência

- Tempo para enviar a imagem e esperar uma resposta



Computação na borda

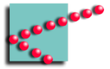
- Realizar tarefas de processamento perto do usuário



Outras funcionalidades da Borda

- Melhor responsividade
  - Menor latência da comunicação

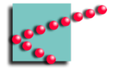
- Maior escalabilidade
  - Natureza distribuída



Outras funcionalidades da Borda

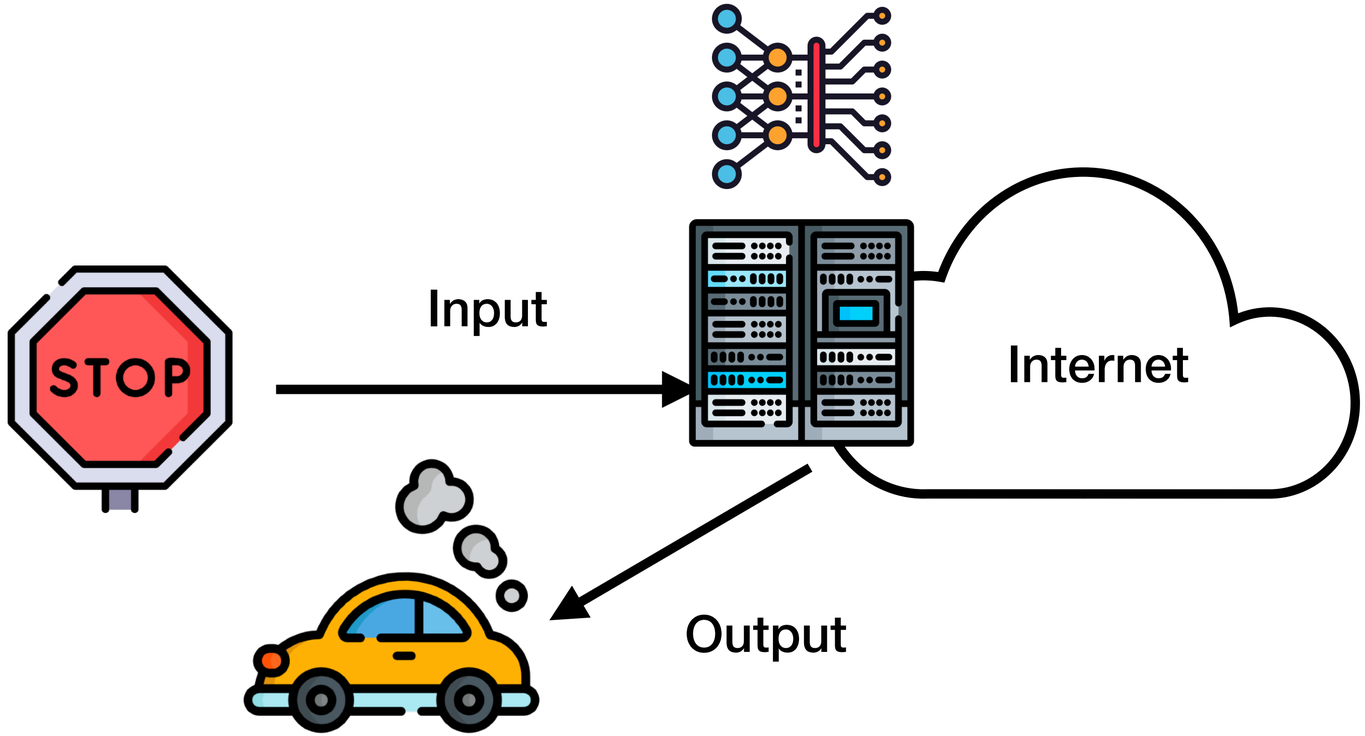
- Maior escalabilidade
  - Natureza distribuída

- Maior privacidade
  - Preservação de dados próximos ao usuário



Computação na borda

- Realizar tarefas de processamento perto do usuário



Reduz o problema,
mas ainda podemos melhorar!

Offloading adaptativo

- Algumas imagens não necessitam passar por todas as camadas da DNN

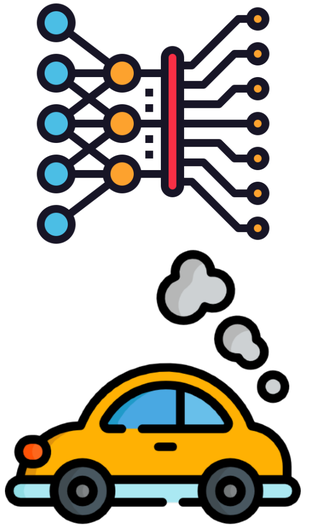
Cachorro?

Gato?

Outro?

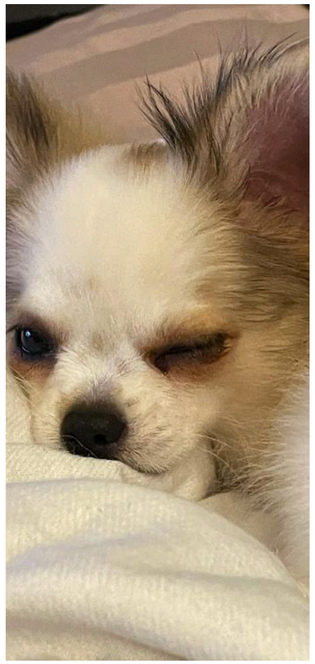
Offloading adaptativo

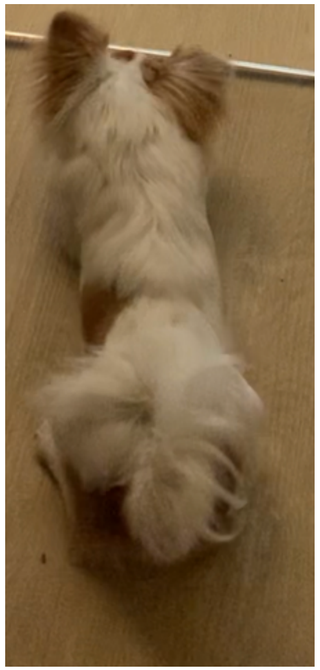
- Algumas imagens não necessitam passar por todas as camadas da DNN
  - Podem ser processadas no dispositivo

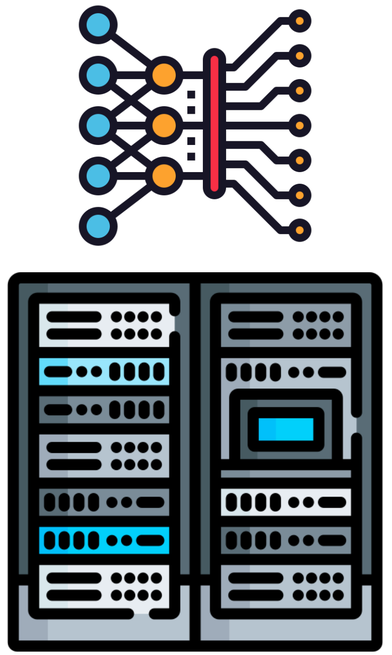


Offloading adaptativo

- Entradas mais complexas podem ser enviadas à nuvem ou à borda



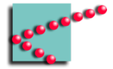




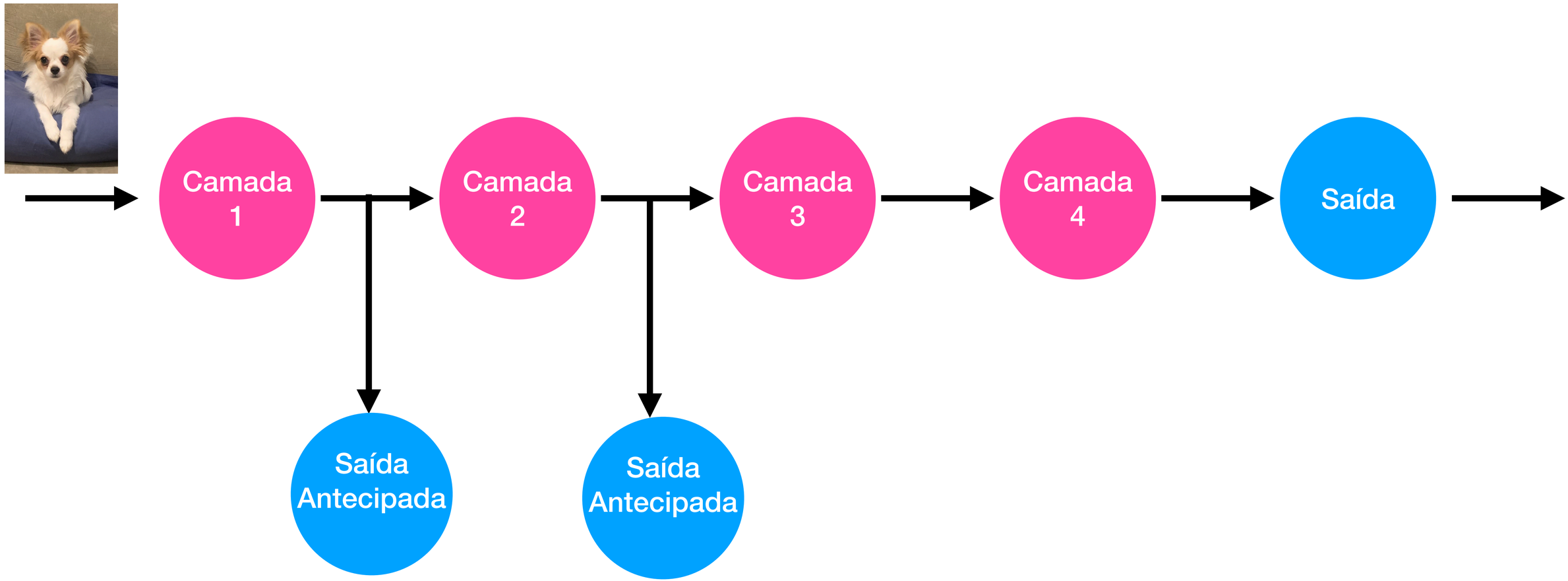
Pilares do offloading adaptativo

- DNNs com saídas antecipadas
  - Introduzidas pelo artigo da BranchyNet

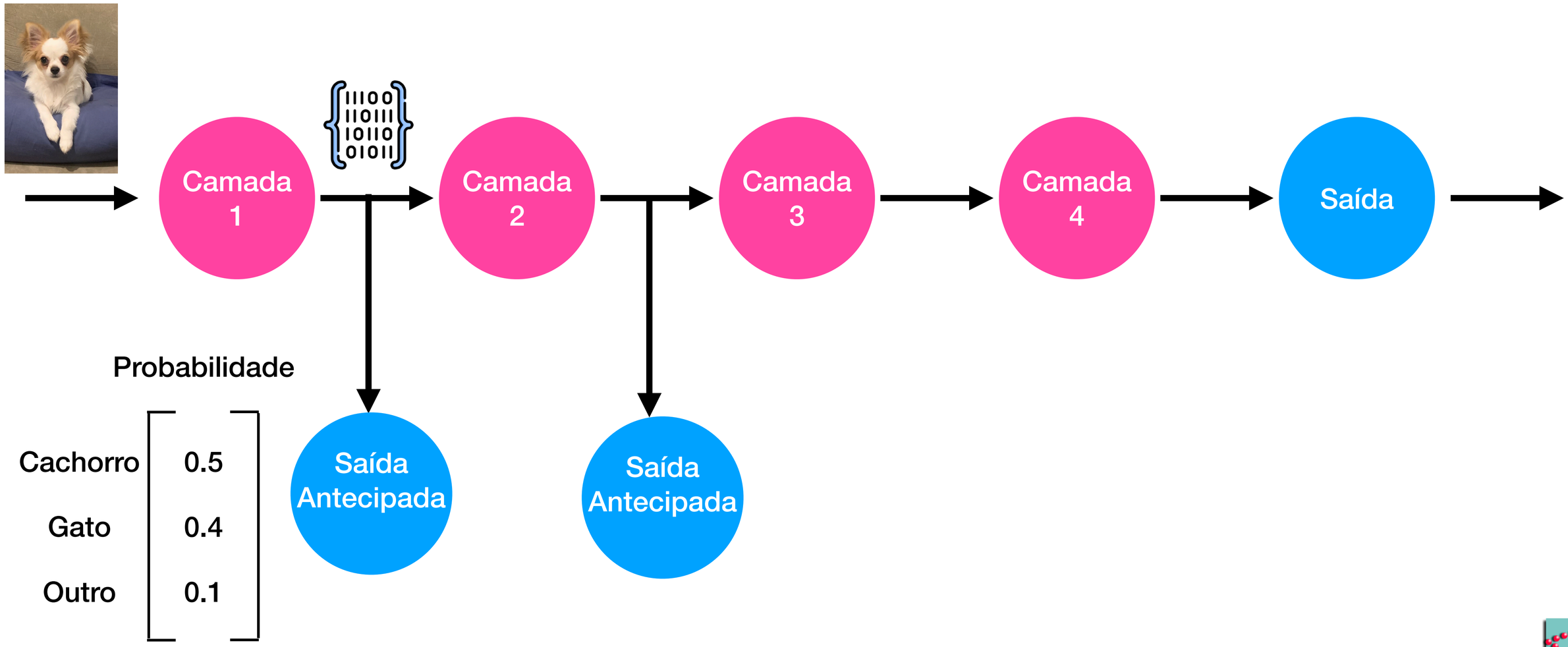
Teerapittayanon, S., McDanel, B., & Kung, H. T. (2016). Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR).*



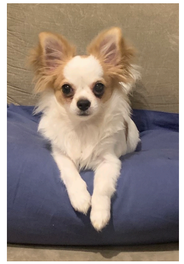
DNNs com saídas antecipadas



DNNs com saídas antecipadas



DNNs com saídas antecipadas



Confiança > 0.7 (limiar) ?

Camada 1

Camada 2

Camada 3

Camada 4

Saída

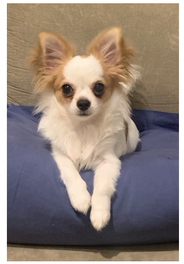
Saída Antecipada

Saída Antecipada

Probabilidade

Cachorro	0.5
Gato	0.4
Outro	0.1

DNNs com saídas antecipadas



Confiança > 0.7 (limiar) ?

Camada 1

Camada 2

Camada 3

Camada 4

Saída

Saída Antecipada

Saída Antecipada

Probabilidade

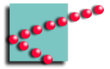
Cachorro	0.8
Gato	0.1
Outro	0.1

É um cachorro!

Pillars of adaptive offloading

- Particionamento de DNNs
  - Introduzido pelo artigo do Neurosurgeon

Kang, Y., Hauswald, J., Gao, C., Rovinski, A., Mudge, T., Mars, J., & Tang, L. (2017). Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News*



Camada 1

Camada 2

Camada 3

Camada 4

Saída

Menos dados são enviados pela rede

Author:https://commons.wikimedia.org/wiki/User:Laserlicht

Source:https://commons.wikimedia.org/wiki/File:Raspberry_Pi_4_Model_B_-_Side.jpg

Camada 1

Camada 2

Camada 3

Camada 4

Saída

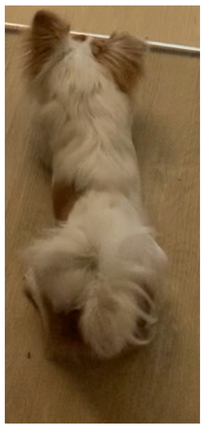
Dispositivo processa menos camadas

Author:https://commons.wikimedia.org/wiki/User:Laserlicht

Source:https://commons.wikimedia.org/wiki/File:Raspberry_Pi_4_Model_B_-_Side.jpg

Offloading adaptativo

Confiança > 0.7 (limiar) ?



Camada 1

Camada 2

Camada 3

Camada 4

Saída

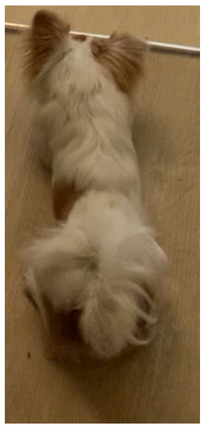
Saída Antecipada

Saída Antecipada

Probability

Dog	0.4
Cat	0.3
Other	0.3

Offloading adaptativo



Confiança > 0.7 (limiar) ?

Camada 1

Camada 2

Saída Antecipada

Saída Antecipada

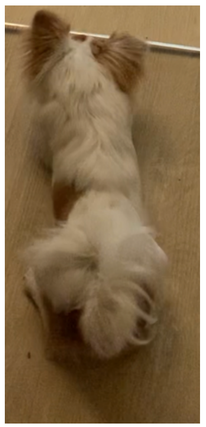
Camada 3

Camada 4

Saída

Offloading adaptativo

Confiança > 0.7 (limiar) ?



Camada 1

Camada 2

Saída Antecipada

Saída Antecipada

Camada 3

Camada 4

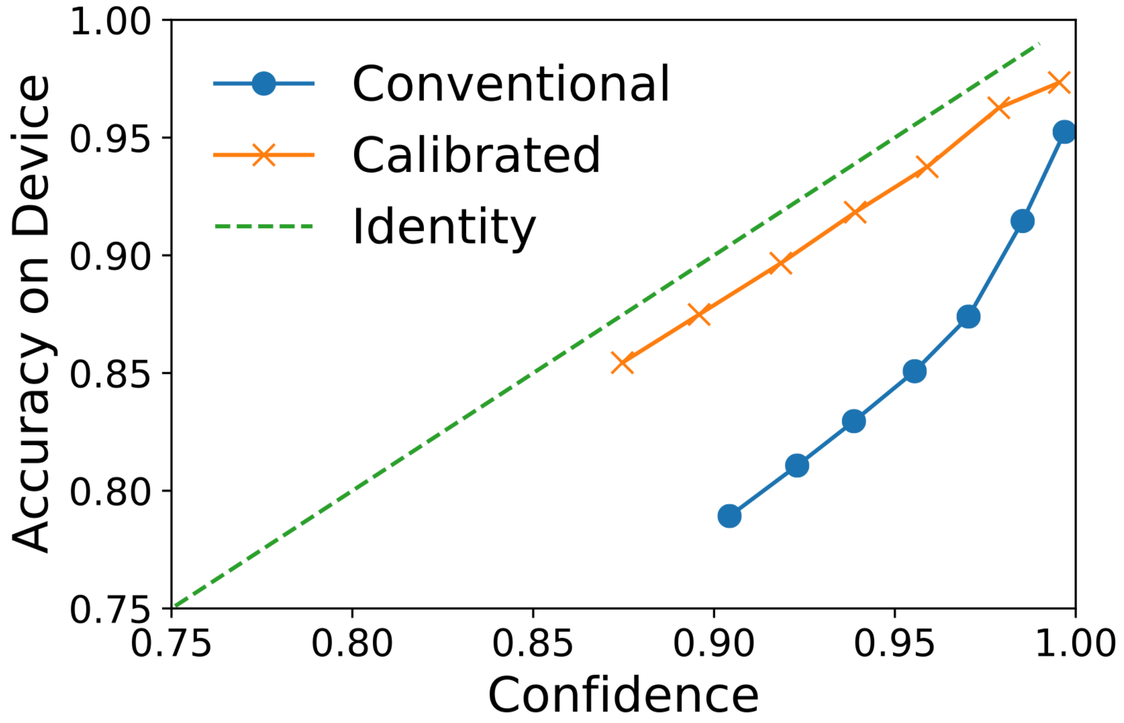
Saída

Probability

Dog	0.8
Cat	0.1
Other	0.1

Calibração de DNNs

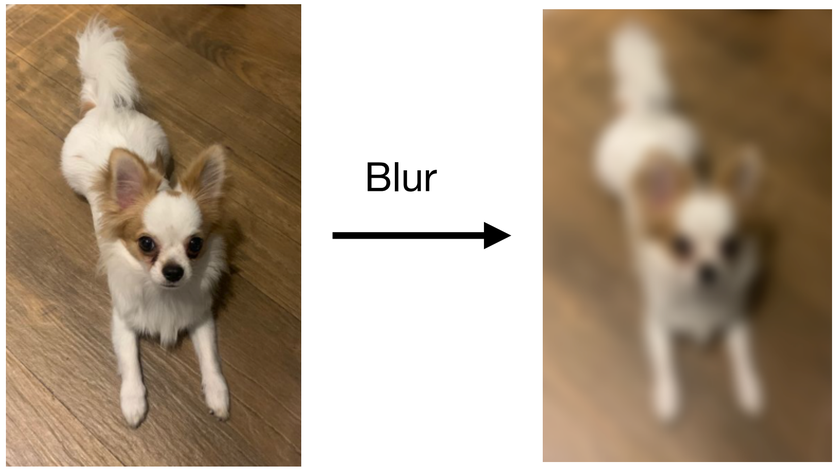
- Saídas antecipadas podem ser super confiantes
  - Leva a decisões ruins de offloading
- O trabalho mostra que uma técnica de calibração simples reduz o problema
- Menos amostras são classificadas na borda
- Trabalho em andamento tenta lidar com esse compromisso



Pacheco, R. G., Couto, R. S., & Simeone, O. (2021). Calibration-Aided Edge Inference Offloading via Adaptive Model Partitioning of Deep Neural Networks. Em *IEEE International Conference on Communications* (ICC).

Distorção de imagens

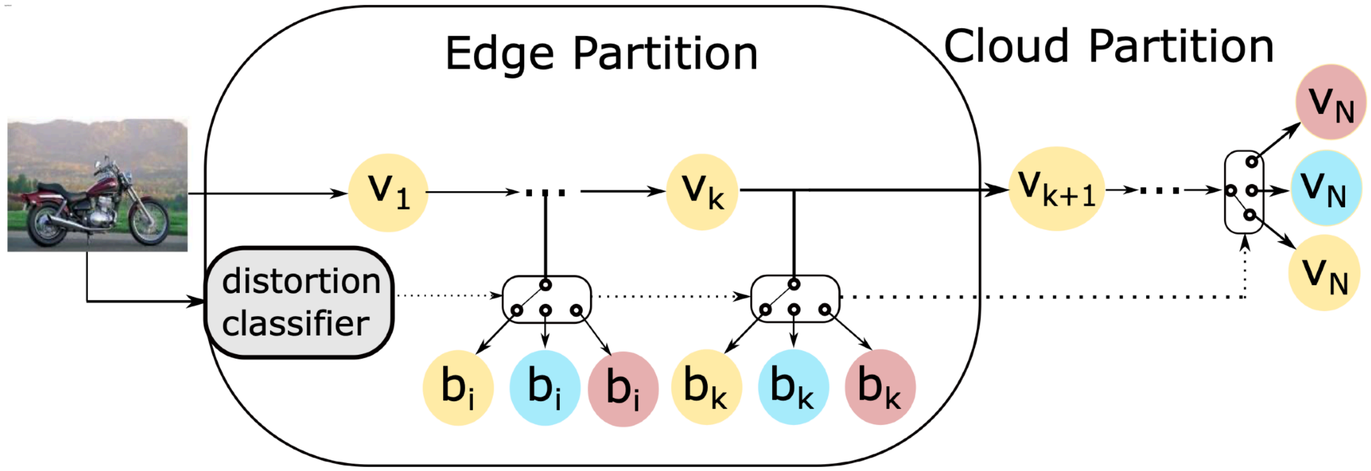
- Imagens distorcidas podem impactar o offloading
  - P.ex., motion blur

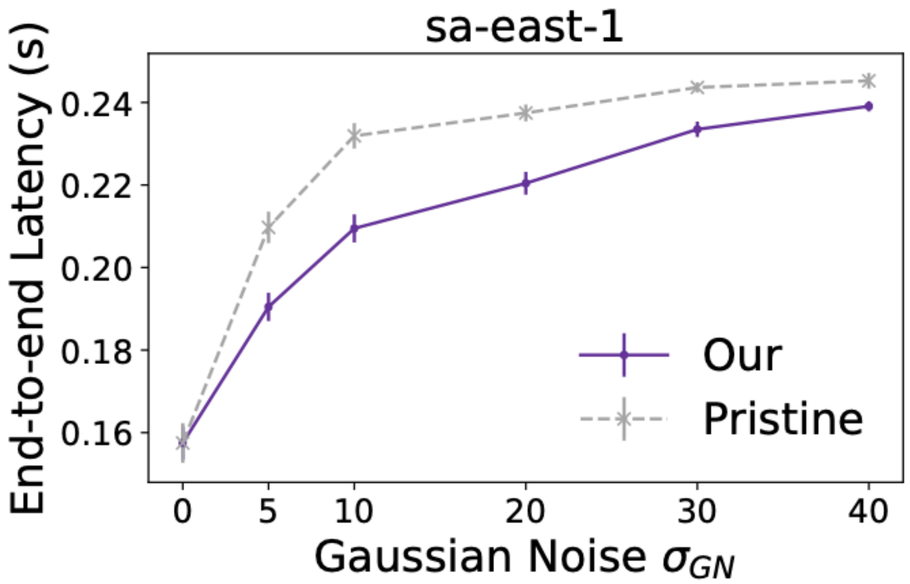


Pacheco, R. G., Oliveira, F. D., & Couto, R. S. (2021). Early-exit deep neural networks for distorted images: providing an efficient edge offloading. ", *IEEE Global Communications Conference (GLOBECOM).*

Distorção de imagens

- Saídas antecipadas especialistas

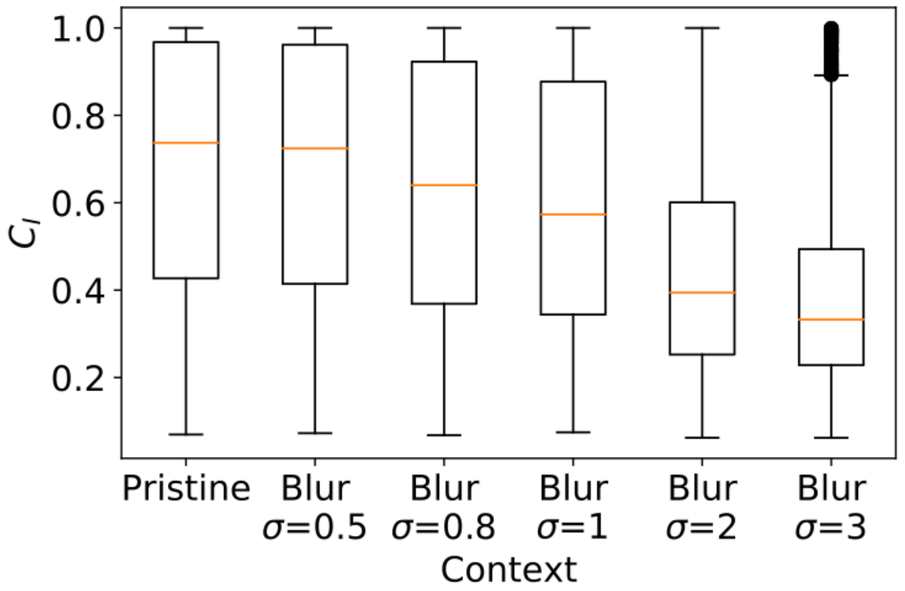




Pacheco, R. G., Oliveira, F. D., & Couto, R. S. (2021). Early-exit deep neural networks for distorted images: providing an efficient edge offloading. ", *IEEE Global Communications Conference (GLOBECOM).*

Escolha do limiar

- Limiar baixo pode levar a baixa acurácia
- Limiar muito alto pode não levar a alta acurácia
  - Imagens que já não teriam confiança alta na nuvem/borda
- Uso de aprendizado por reforço para ajuste do limiar
  - Multi-Armed Bandits
  - Adaptação ao contexto
    - P.ex., imagens distorcidas



Pacheco, R. G., Shifrin, M., Couto, R. S., Menasche, D. S., Hanawal, M. K & Campista, M. E. M. (2023) AdaEE: Adaptive Early-Exit DNN Inference Through Multi-Armed Bandits.", *IEEE International Conference on Communications* (ICC).

Exemplo de projeto de inovação no GTA

- Participação do programa de GTs (Grupo de Trabalho) da RNP (Rede Nacional de Ensino e Pesquisa)

- Objetivo do programa de GTs é fomentar embriões de startups
  - Spin-offs dos laboratórios da universidade
  - Primeiros clientes podem ser instituições do sistema RNP

- GT-CampusEdge
  - Monitoramento do patrimônio universitário por processamento de vídeo na borda

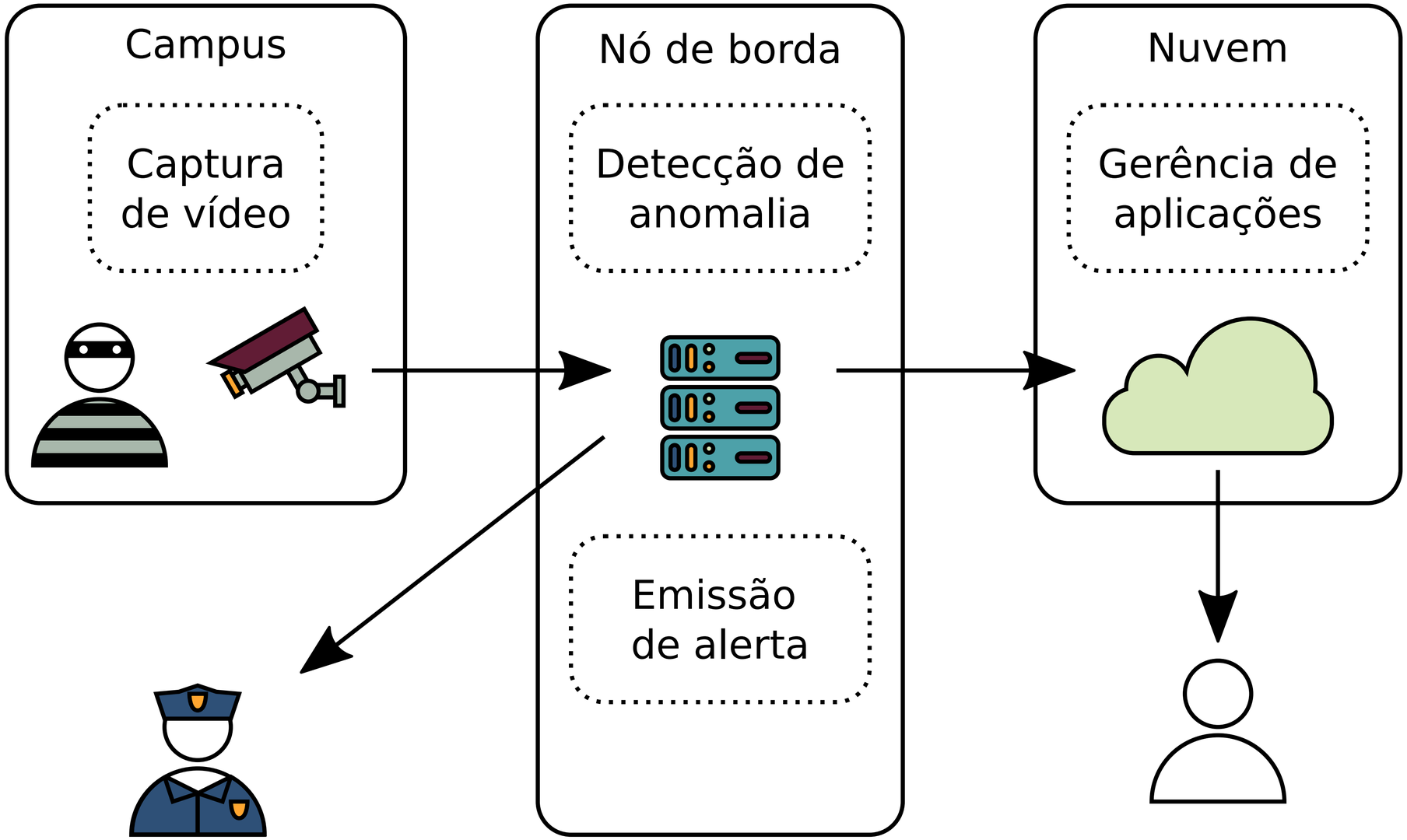
Segurança patrimonial em um campus

- Furtos desafiam gestores
  - Unidades externas de ar-condicionado split, cabos de cobre e pequenos
  - Prejuízos significativos

- Há câmeras de segurança
  - Mas equipe de vigilância é reduzida
  - Fora do horário comercial problema se agrava

- Objetivo do GT-CampusEdge
  - Desenvolver aplicações de IA que detectam eventos de segurança patrimonial

Ideia Geral do GT-CampusEdge



Estado do GT-CampusEdge

- Início em Janeiro de 2023

- Capacitação empreendedora
  - Modelagem do MVP (*Minimum Viable Product*)
  - Metodologia iCorps

- Desenvolvimento do MVP
  - Verificação de anomalias no vídeo em tempo real
  - Presença de pessoas não autorizadas nas proximidades de um equipamento

Obrigado

- Contato
  - rodrigo@gta.ufrj.br
  - www.gta.ufrj.br

Apoio

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

FAPESP 60
CIÊNCIA, CULTURA E DESENVOLVIMENTO 1962 - 2022

CAPES

RNP
ORGANIZAÇÃO SOCIAL DO MCTI

FAPERJ
Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro

GTA / UFRJ

GRUPO DE TELEINFORMÁTICA E AUTOMAÇÃO

www.gta.ufrj.br