Projeto ABCIA

MÓDULO 06 - FRAMEWORKS DE DESENVOLVIMENTO PARA IA

Aula II

Prof. Dr. Leandro Balico

UFRR | DCC DEPTO DE CIÊNCIA DA COMPUTAÇÃO | Softex | ABCIA ABC DA INTELIGÊNCIA ARTIFICIAL

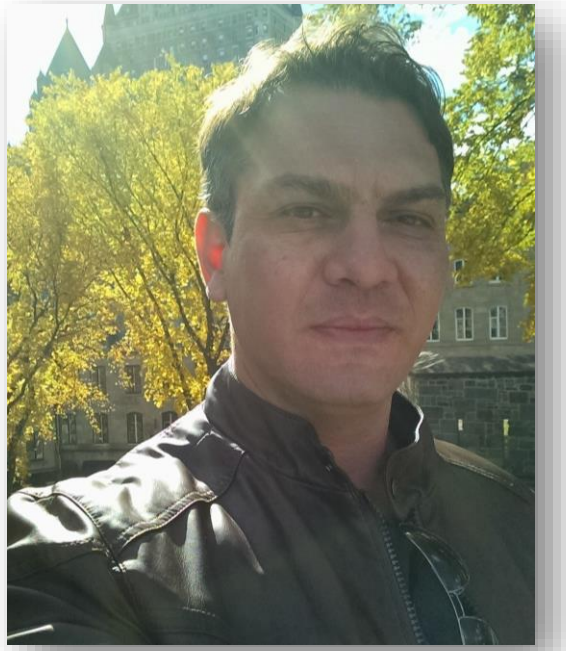
Apresentação do Professor

Formação

- Bacharel em Ciência da Computação – UFPR
- Mestre em Informática – UFAM
- Doutor em Informática – UFAM
- Certificação HCIA-IA (Huawei)

Atuação

- Atualmente é Professor Adjunto da Universidade Federal de Roraima. Tem experiência na área de Ciência da Computação, com ênfase em Redes de Sensores Sem Fio, Redes Ad Hoc, Redes Veiculares (VANets), Aprendizado de Máquina e Computação Móvel e Ubíqua, atuando principalmente nos seguintes temas: algoritmos distribuídos, localização, roteamento, consumo de energia, fusão de dados, e outros.



Prof. Dr. Leandro Balico

Objetivos

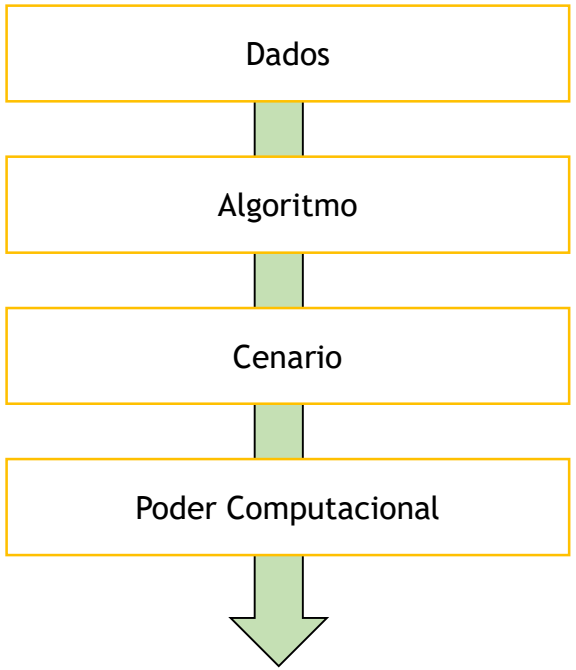
Objetivos

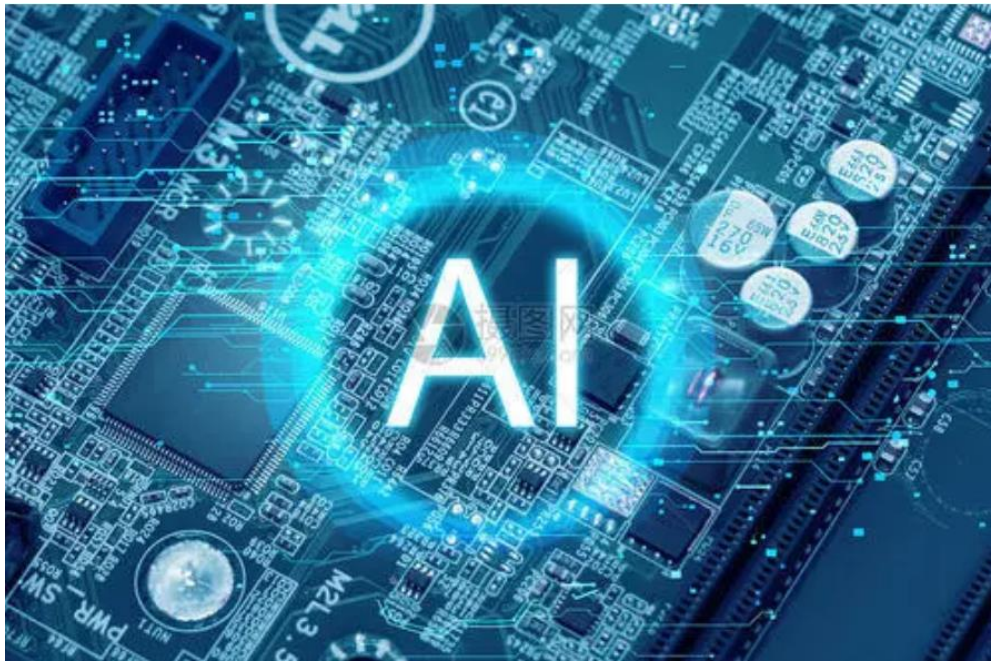
1. Chips de Inteligência Artificial
2. Plataforma de Computação de IA Atlas
3. Plataforma de Desenvolvimento de IA para Dispositivos Inteligentes
4. Plataforma de Aplicações de Inteligência Empresarial
5. Exercícios

1. Chips de Inteligência Artificial

1.1 Introdução

- **Quatro elementos da IA:** dados, algoritmo, cenário e poder de computação.
- **Os chips de IA:** também conhecidos como aceleradores de IA, são módulos de função que processam tarefas de computação massivas em aplicativos de IA.





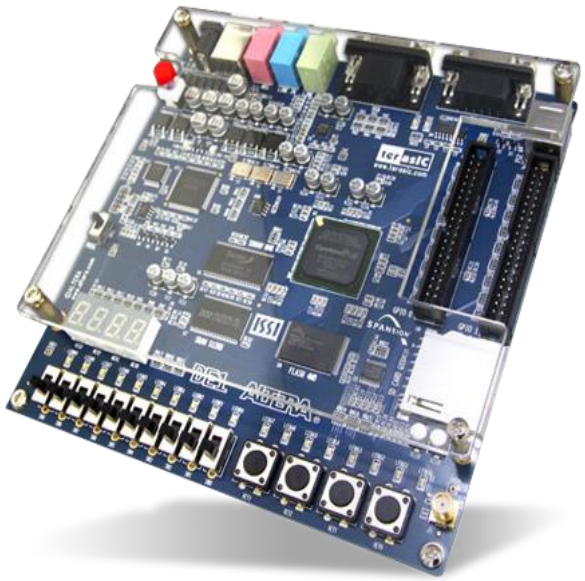
1.2 Classificação dos Chips de IA

- Os Chips de podem ser divididos em quatro tipos pela sua arquitetura:

Unidade de Processamento Central (*Central Processing Unit* – CPU)

Unidade de Processamento Gráfico (*Graphics Processing Unit* – GPU)

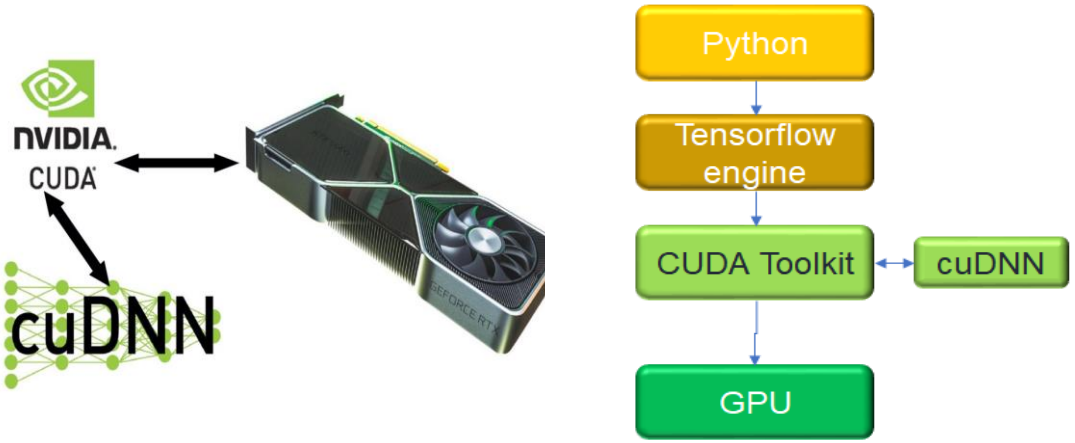
Circuito integrado de aplicação específica (*Application Specific Integrated Circuit* – ASIC)

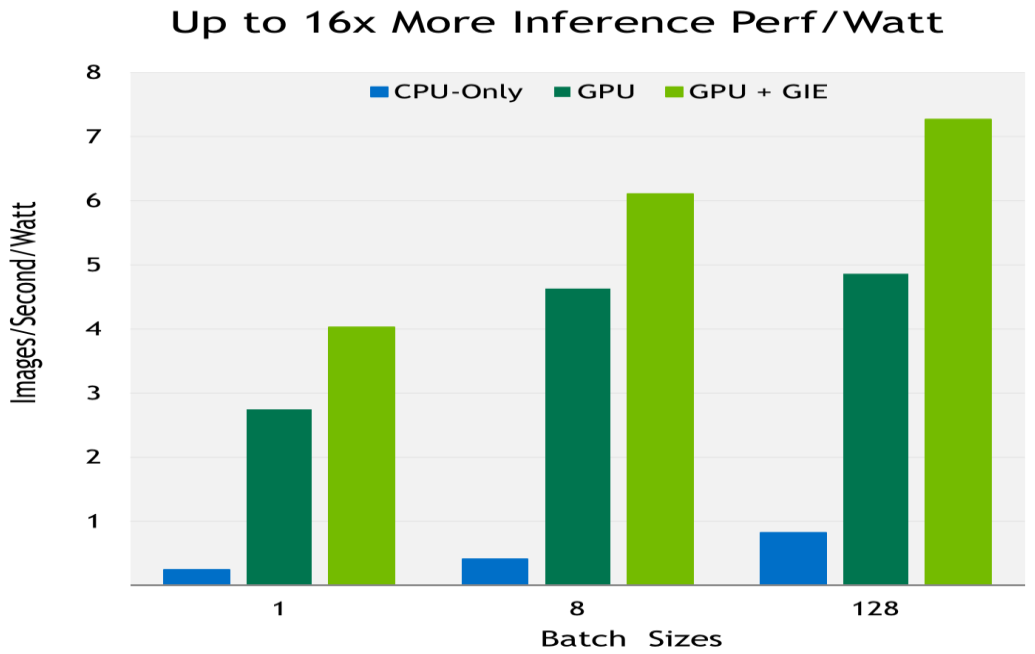


Matriz de Portas Programáveis em Campo (*Field Programmable Gate Array* - FPGA)

1.2 Classificação dos Chips de IA

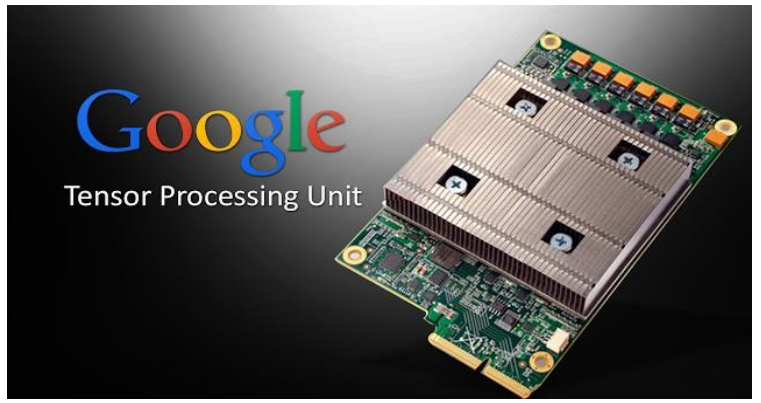
- **Unidade de processamento gráfico (GPU)**
  - Desempenho notável em computação matricial e paralela atuando como um chip de aceleração para aprendizado profundo
  - Usando a arquitetura de GPU, a NVIDIA se concentra em:
    - Diversificar o ecossistema com a biblioteca de otimização cuDNN (CUDA Deep Neural Network)
    - Melhorar a personalização suportando vários tipos de dados e módulos dedicados e otimizados para calculo de Tensores
- Problemas:
  - altos custos e consumo energético

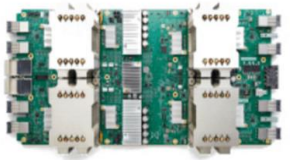




1.2 Classificação dos Chips de IA

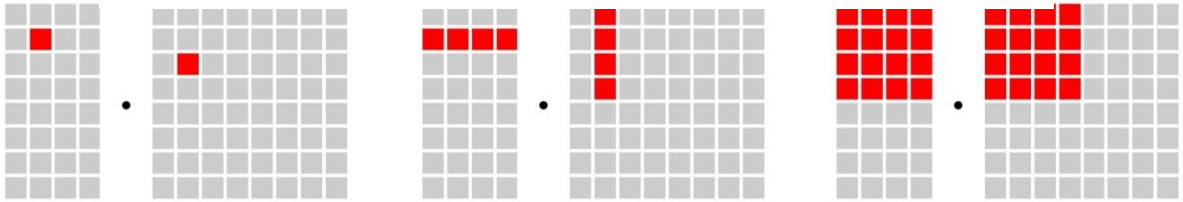
- **Unidade de processamento de Tensores (Tensor Processing Unit - TPU)**
  - São chips ASICs especializados projetados especificamente para tarefas de aprendizado profundo.
    - Projetados e otimizados para executar cálculos de matriz (Tensores).
    - Treinamento mais rápidos e maior precisão.
    - Alta eficiência computacional e são energeticamente mais eficientes.
  - Não estão amplamente disponíveis, mas estão se tornando cada vez mais populares.





CPU GPU TPU

Calculo de Primitivas



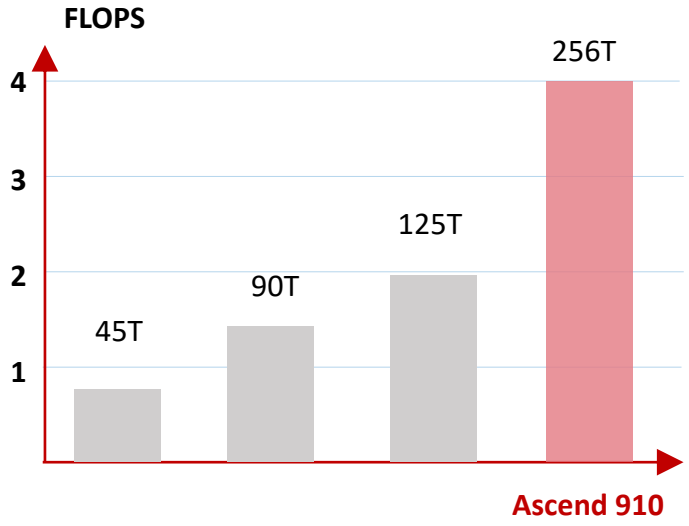
Escalares Vetores Tesores

1.3 Processadores Ascend

- São Unidades de Processamento de Rede Neurais (Neural-network Processing Unit NPU).
- Projetados e desenvolvidos pela Huawei para uso em aplicativos de IA.
- São baseados na arquitetura Da Vinci proprietária da Huawei.
- Principais vantagens: Alto desempenho, Baixa latência, Eficiência energética e versatilidade.

- Ascend-Mini
- Architecture: Da Vinci
- Half precision (FP16): 8 Tera-FLOPS
- Integer precision (INT8): 16 Tera-OPS
- 16-channel full-HD video decoder: H.264/H.265
- 1-channel full-HD video decoder: H.264/H.265
- **Max. power: 8W**

- Ascend-Max
- Architecture: Da Vinci
- Half precision (FP16): 256 Tera-FLOPS
- Integer precision (INT8): 512 Tera-OPS
- 128-channel full-HD video decoder: H.264/H.265
- Max. power: 350W

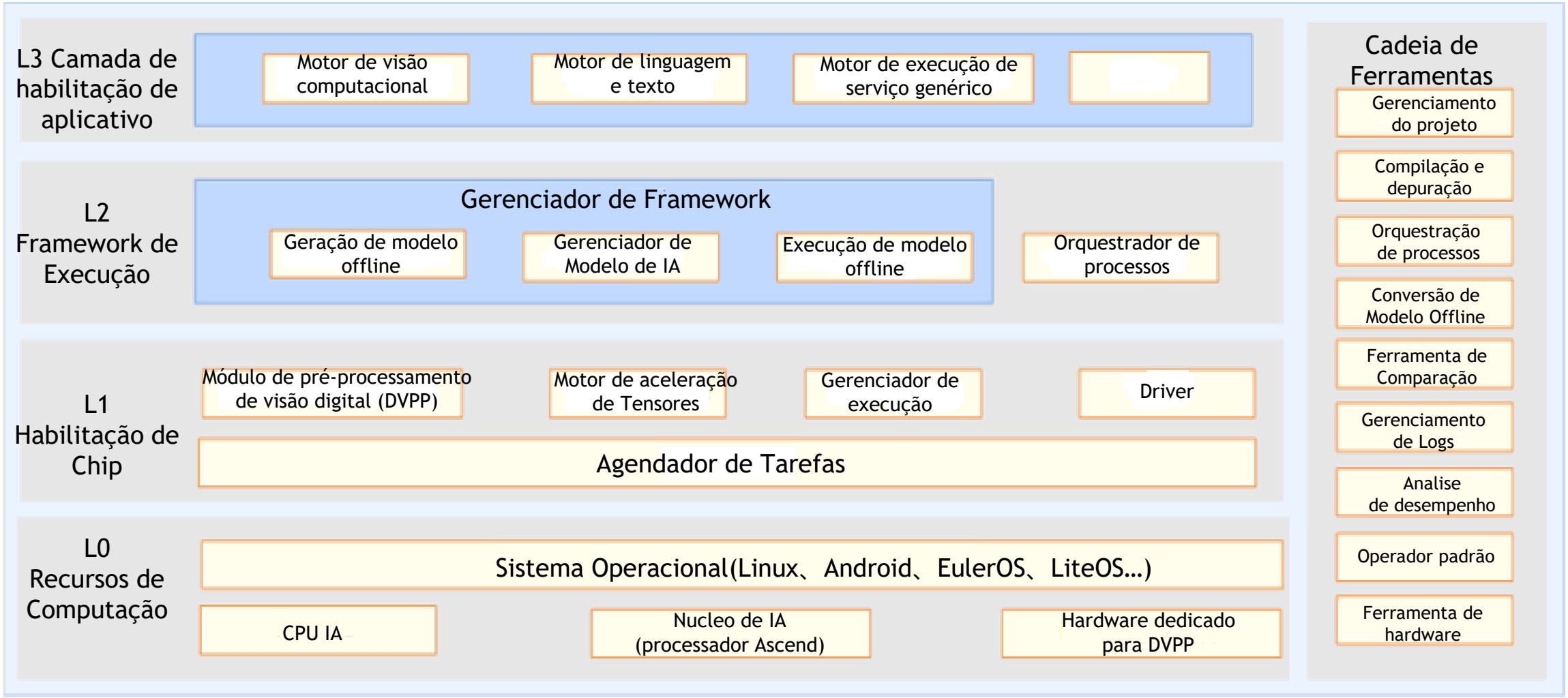


2. Plataforma de Computação de IA Atlas

2.1 A Plataforma de Computação de IA Atlas (Huawei Atlas AI Computing Platform)

- É uma plataforma de IA abrangente que fornece componentes de hardware e software para aplicativos de IA.
- Oferece uma infraestrutura de computação incluindo:
  - Algoritmos de aprendizado profundo e aprendizado de máquina.
  - Tecnologias de hardware de ponta: GPUs, TPUs e FPGAs.
  - Facilidade de uso.
  - Alto desempenho e escalabilidade.

2.2 Arquitetura Lógica do Processador Ascend AI



2.3 Portfolio da Plataforma de Computação de IA Atlas

Internet, segurança, finanças, transportes, energia, etc.

Habilitação de aplicações

SDK industrial
Motor de container
Serviço básico de repositório

Gerenciamento de Clusters
Gerenciamento de Modelos
Pré-processamento de dados

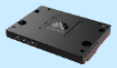
Framework IA

MindSpore
TensorFlow/PyTorch/Caffe/MxNet
Adaptadores de framework

Arquitetura de computação para redes neurais (CANN)

AscendCL
Motor de otimização de grafos
Bibliotecas de Operadores/Aceleração/Ccomunicação (BLAS, FFT, DNN, Rand, Solver, Sparse, HCCL)
Runtime
Driver

Chips & Hardware



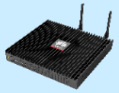
Atlas 200
16 TOPS INT8

Atlas 200 kit de desenvolvimento

Atlas 300 inference accelerator card
64 TOPS INT8

Atlas 300 placa de treinamento
256 TFLOPS FP16

Ascend 310



Atlas 500
16 TOPS INT8

Arquitetura da Vinci

Atlas 800 AI servidor de Inferencia
512 TOPS INT8

Atlas 800 AI servidor de treinamento
2 PFLOPS FP16

Ascend 910

Atlas 900
256-1024 PFLOPS FP16

2.3 Portfolio da Plataforma de Computação de IA Atlas

- Plataforma de Inferência.

Ascend 310 AI

Atlas (DK)
Kit de desenvolvimento IA

Atlas 200 AI modulo acelerador
Modelo: 3000

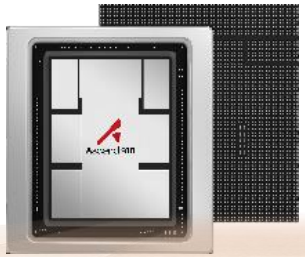
Atlas 300 AI placa aceleradora
Modelo: 3000

Atlas 500 AI estacao de borda
Modelo: 3000

Atlas 800 AI servidor
Modelo: 3000/3010

2.3 Portfolio da Plataforma de Computação de IA Atlas

- Plataforma de Treinamento.



Ascend 910
AI

Atlas 300 AI placa aceleradora
Modelo: 9000

Atlas 800 AI servidor
Modelo: 9000/9010

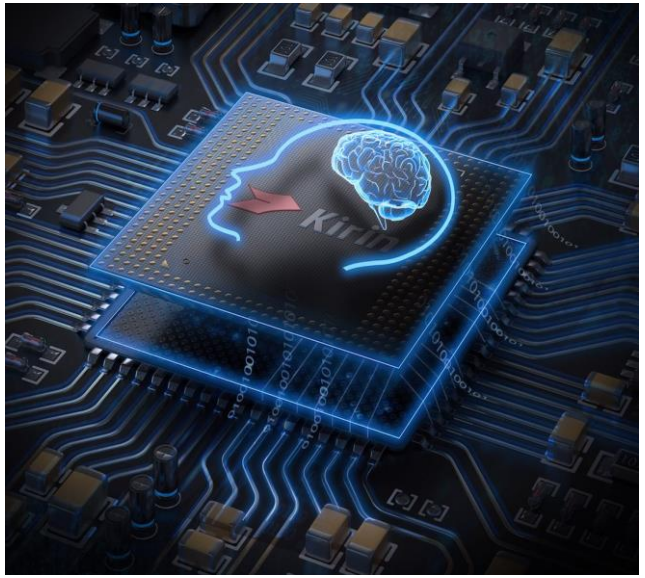
Atlas 900 AI cluster

3. Plataforma de Desenvolvimento de IA para Dispositivos Inteligentes

3.1 Plataforma de Desenvolvimento de IA para Dispositivos Inteligentes HiAI

Nuvem

Dispositivo



Chip

3.2 Ecossistema de IA de Três Camadas HiAI

Nuvem	HiAI Service	Fornece aos desenvolvedores uma API fácil de usar para criar aplicativos de IA
Dispositivo	HiAI Engine	Responsável por executar modelos de IA e lidar com tarefas como treinamento e inferência de modelos
Chip	HiAI Foundation	Fornece serviços básicos de aceleração e suporte de plataforma de hardware para IA

3.3 Beneficios da plataforma Huawei HiAI

Aceleração de IA

Compatibilidade de entre dispositivos

Desenvolvimento fácil

Segurança aprimorada

Baixo custo

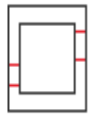
4. Plataforma de Aplicações de Inteligência Empresarial

4. Plataforma de Aplicações de Inteligência Empresarial

4.1 Plataforma de Aplicações de Inteligência Empresarial (*Huawei Cloud Enterprise Intelligence*)

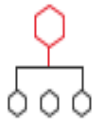
- ModelArts : Plataforma de desenvolvimento de IA unificada para desenvolvedores e cientistas de dados
- Huawei HiLens (HiLens) : Plataforma multimodal de desenvolvimento de IA
- Graph Engine Service (GES) : Consulta e análise de dados de estrutura gráfica
- Question Answering Bot (QABot): Crie e gerencie bots inteligentes de perguntas e respostas
- Image Recognition: Identifique automaticamente o conteúdo de imagens
- Content Moderation: Realize a varredura de textos e imagens em busca de conteúdo impróprio
- Optical Character Recognition (OCR): Detecta e extrai texto de imagens e o converte em um formato JSON editável
- Image Search: Pesquisa de imagens exatas e difusas
- Short Sentence Recognition: Transcrição automática de gravações de áudios curtos
- EIHealth : Acelere pesquisas e aplicações de IA em genômica, descoberta de medicamentos e exames de imagem

4.2 ModelArts



Gerenciamento de Dados

Suporta processamento de dados, como filtragem e rotulagem, e fornece gerenciamento de versão do conjunto de dados



Treinamento de modelo rápido e simplificado

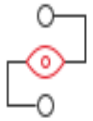
O framework de aprendizado profundo MoXing desenvolvido pela Huawei é eficiente e fácil de usar, acelerando muito o treinamento

Implantação do modelo

Pode implantar modelos em vários ambientes de produção, como implantação em nuvem para inferência online e em lote, dispositivos e borda.

ExeML

ModelArts suporta treinamento de modelos baseados em aprendizado automático para que os usuários concluam a modelagem automática e a implantação com um clique sem compilar o código.



Fluxo de trabalho visualizado

Graph Engine Service (GES) gerencia os metadados do pipeline de desenvolvimento de forma unificada, e visualiza automaticamente a evolução

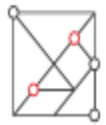
Mercado AI

ModelArts oferece suporte a modelos e conjuntos de dados comuns. e compartilhamento interno ou público de modelos empresariais no mercado

4.2 ModelArts

Plataforma Completa

Suporta processamento de dados, como filtragem e rotulagem, e fornece gerenciamento de versão do conjunto de dados



Fácil de Usar

Suporta processamento de dados, como filtragem e rotulagem, e fornece gerenciamento de versão do conjunto de dados

Excelente Desempenho

Suporta processamento de dados, como filtragem e rotulagem, e fornece gerenciamento de versão do conjunto de dados

Alta Flexibilidade

Suporta processamento de dados, como filtragem e rotulagem, e fornece gerenciamento de versão do conjunto de dados

4.3 Huawei HiLens

5. Exercícios

5.1 Hardware que fornece computação paralela.

a) CPU
b) Dimm
c) GPU
d) Placa principal

5.2 Qual das seguintes opções é a arquitetura adotada pelos chips da série Huawei Atlas?

a) Von Neumann
b) Gauss
c) Ascend
d) Da Vinci

5.3 A arquitetura da plataforma de computação móvel HIAI suporta várias estruturas de front-end principal, como Tensorflow e Caffe.

a) Verdadeiro
b) Falso

5.4 Na Huawei Cloud EI Enterprise Intelligence, quais serviços básicos de plataforma estão incluídos?

a) Aprendizado de máquina
b) Aprendizado profundo
c) Graph Engine Service
d) Processamento em lote

UFRR

DEPTO DE CIÊNCIA DA COMPUTAÇÃO

Softex

ABCIA
ABC DA INTELIGÊNCIA ARTIFICIAL

Até a próxima aula!

Continue praticando :)