ModelArts

Visão geral de serviço

Edição 01

Data 07-11-2022

HUAWEI TECHNOLOGIES CO., LTD.

Marcas registadas e permissões


e outras marcas registadas da Huawei são marcas registadas da Huawei Technologies Co., Ltd. Todos as outras marcas registadas e os nomes registados mencionados neste documento são propriedade dos seus respectivos detentores.

Aviso

Os produtos, serviços e funcionalidades adquiridos são estipulados pelo contrato feito entre a Huawei e o cliente. Todos ou parte dos produtos, serviços e funcionalidades descritos neste documento pode não estar dentro do âmbito de aquisição ou do âmbito de uso. Salvo especificação em contrário no contrato, todas as declarações, informações e recomendações neste documento são fornecidas "TAL COMO ESTÁ" sem garantias, ou representações de qualquer tipo, seja expressa ou implícita.

As informações contidas neste documento estão sujeitas a alterações sem aviso prévio. Foram feitos todos os esforços na preparação deste documento para assegurar a exatidão do conteúdo, mas todas as declarações, informações e recomendações contidas neste documento não constituem uma garantia de qualquer tipo, expressa ou implícita.

Huawei Technologies Co., Ltd.

Endereço: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Site: https://www.huawei.com

Email: support@huawei.com

Índice

1 Infográficos

2 O que é o ModelArts?

O ModelArts é uma plataforma de desenvolvimento de IA voltada para desenvolvedores e cientistas de dados de todos os níveis. Ele permite que você crie, treine e implante modelos rapidamente em qualquer lugar (da nuvem até a borda), e gerencie fluxos de trabalho de IA de ciclo de vida completo. O ModelArts acelera o desenvolvimento da IA e promove a inovação com recursos-chave, incluindo pré-processamento de dados e rotulagem automática, treinamento distribuído, construção automatizada de modelos e execução de fluxo de trabalho com um clique.

O ModelArts abrange todos os estágios do desenvolvimento de IA, incluindo processamento de dados e treinamento e implantação de modelos. As tecnologias subjacentes do ModelArts suportam vários recursos de computação heterogêneos, permitindo que os desenvolvedores selecionem e usem recursos de forma flexível. Além disso, a ModelArts suporta estruturas populares de desenvolvimento de IA de código aberto, como TensorFlow, MXNet, e PyTorch. O ModelArts também permite que você use estruturas de algoritmo personalizadas adaptadas às suas necessidades.

O ModelArts visa simplificar o desenvolvimento da IA.

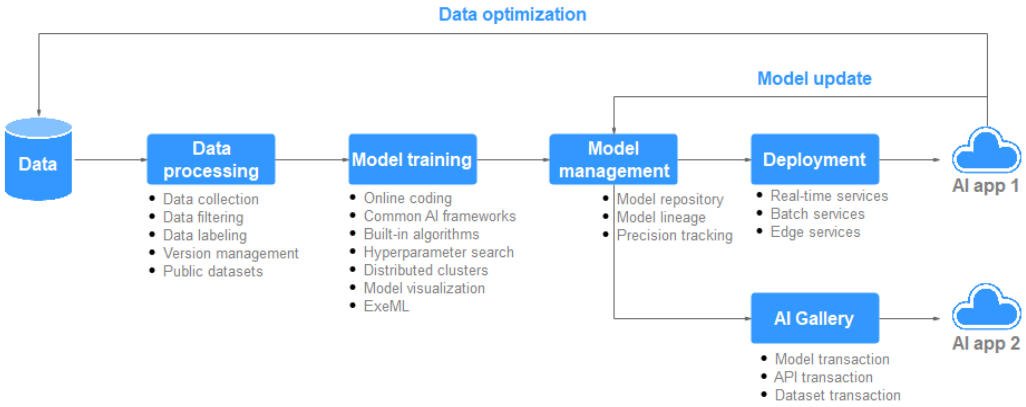
O ModelArts é adequado para desenvolvedores de IA com diferentes níveis de experiência em desenvolvimento. Os desenvolvedores de serviços podem usar o ExeML para criar rapidamente aplicativos de IA sem codificação. Os iniciantes podem usar diretamente algoritmos internos para criar aplicativos de IA. Os engenheiros de IA podem usar vários ambientes de desenvolvimento para compilar rapidamente o código para modelagem e desenvolvimento de aplicativos.

Arquitetura do produto

O ModelArts oferece suporte a todo o processo de desenvolvimento, incluindo processamento de dados e treinamento, gerenciamento e implantação de modelos.

O ModelArts oferece suporte a vários cenários de aplicativos de IA, como classificação de imagens, análise de vídeo, reconhecimento de fala, recomendação de produtos, e detecção de exceções.

Figura 2-1 Arquitetura do ModelArts



Vantagens do produto

- **Plataforma one-stop**

  A plataforma de desenvolvimento de IA pronta para uso e de ciclo de vida completo fornece serviços de one-stop de processamento de dados completo, e desenvolvimento, treinamento, gerenciamento e implantação de modelos.

- **Alto desempenho**
  - Múltiplos modelos integrados fornecidos e uso gratuito de modelos de código aberto
  - Otimização automática de hiperparâmetros
  - Desenvolvimento livre de código e operações simplificadas
  - Implantação de modelos com um clique para a nuvem, borda e dispositivos
- **High performance**
  - A estrutura de aprendizado profundo MoXing autodesenvolvida acelera o desenvolvimento e o treinamento de algoritmos.
  - A utilização otimizada da GPU acelera a inferência em tempo real.
  - Modelos executados em chips Ascend AI obtêm inferência mais eficiente.
- **Flexível**
  - Frameworks de código aberto populares disponíveis, como TensorFlow, Spark_MLlib, MXNet, Caffe, PyTorch, XGBoost-Sklearn, e MindSpore
  - Os GPU populares e chips Ascend proprietários da Huawei
  - Uso exclusivo de recursos dedicados
  - Imagens personalizadas para frameworks e operadores personalizados

Uso do ModelArts pela primeira vez

Se você é um usuário iniciante, as informações a seguir ajudarão você a se familiarizar com o ModelArts:

- **Conceitos básicos**

  4 Conhecimento básico descreve os conceitos básicos de ModelArts, incluindo o processo básico e conceitos de desenvolvimento de IA e conceitos específicos e funções de ModelArts.

- **Introdução**

  O documento **Introdução** fornece guias de operação detalhados para orientá-lo na construção do modelo no ModelArts.

- **Melhores práticas**

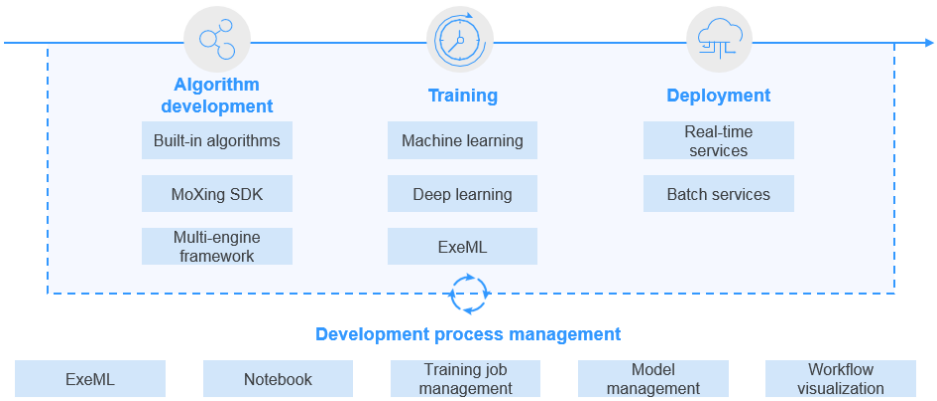
  O ModelArts oferece suporte a vários mecanismos de código aberto e fornece casos de uso extensivos com base nos mecanismos e funções. Você pode criar e implantar modelos consultando **as melhores práticas**.

- **Outras funções e guias de operação**
  - Se você é um desenvolvedor de serviços, pode usar o ExeML para criar modelos rapidamente sem codificação. Para obter detalhes, consulte **Guia de usuário (ExeML)**.
  - Se você é um engenheiro de IA, pode gerenciar o ciclo de vida de desenvolvimento de IA, incluindo gerenciamento de dados e desenvolvimento de modelos, treinamento, gerenciamento e implantação. Para obter detalhes, consulte **DevEnviron**, **preparación de datos**, **etiquetado de datos**, **desarrollo de modelos** ou **inferencia**
  - Se você for um desenvolvedor e quiser usar as API ou os SDK do ModelArts para desenvolvimento de IA, consulte **Referência de API** ou **Referência de SDK**.

3 Funções

Os engenheiros de IA enfrentam desafios na instalação e configuração de várias ferramentas de IA, preparação de dados e treinamento de modelos. Para enfrentar esses desafios, a plataforma one-stop de desenvolvimento de IA do ModelArts é fornecida. A plataforma integra preparação de dados, desenvolvimento de algoritmos, treinamento e implantação de modelos no ambiente de produção, permitindo que os engenheiros de IA realizem o desenvolvimento one-stop de IA.

Figura 3-1 Visão geral das funções



O ModelArts possui os seguintes recursos:

- **Governança de dados**

  Gerencia a preparação de dados, como filtragem e rotulagem de dados e versões de conjuntos de dados.

- **Treinamento rápido e simplificado de modelo**

  Permite treinamento distribuído de alto desempenho e simplifica a codificação com a estrutura de aprendizado profundo MoXing desenvolvida.

- **Cloud-edge-device synergy**

  Implanta modelos em vários ambientes de produção como dispositivos, borda, e nuvem, e suporta inferência em lote e em tempo real.

- **Aprendizagem automática**

Permite a criação de modelos sem codificação e suporta classificação de imagens, detecção de objetos e análise preditiva.

4 Conhecimento básico

4.1 Introdução ao ciclo de vida de desenvolvimento de IA

O que é IA

A inteligência artificial (IA) é uma tecnologia capaz de simular a cognição humana através de máquinas. A capacidade central da IA é fazer um julgamento ou previsão com base em uma determinada entrada.

Qual é o propósito do desenvolvimento de IA

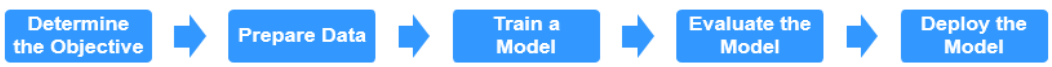
O desenvolvimento de IA visa processar e extrair centralmente informações de volumes de dados para resumir padrões internos dos objetos de estudo.

Grandes volumes de dados coletados são computados, analisados, resumidos e organizados usando estatísticas apropriadas, aprendizado de máquina e métodos de aprendizado profundo para maximizar o valor dos dados.

Processo básico de desenvolvimento de IA

O processo básico de desenvolvimento de IA inclui as seguintes etapas: determinação de um objetivo, preparação de dados e treinamento, avaliação e implantação de um modelo.

Figura 4-1 Processo de desenvolvimento de IA



Passo 1 Determinar um objetivo.

Antes de iniciar o desenvolvimento da IA, determine o que analisar. Quais problemas você quer resolver? Qual é o objetivo do negócio? Classifique a estrutura de desenvolvimento de IA e as ideias com base no entendimento do negócio. Por exemplo, classificação de imagem e detecção de objeto. Diferentes projetos têm diferentes requisitos para dados e métodos de desenvolvimento de IA.

Passo 2 Preparar os dados.

A preparação de dados refere-se à coleta e ao pré-processamento de dados.

A preparação de dados é a base do desenvolvimento de IA. Quando você coleta e integra dados relacionados com base no objetivo determinado, o mais importante é garantir a autenticidade e a confiabilidade dos dados obtidos. Normalmente, você não pode coletar todos os dados ao mesmo tempo. Na fase de rotulagem de dados, você pode achar que algumas fontes de dados estão faltando e, em seguida, pode ser necessário ajustar e otimizar os dados repetidamente.

Passo 3 Treinar um modelo.

A modelagem envolve a análise dos dados preparados para encontrar a causalidade, as relações internas e os padrões regulares, fornecendo referências para a tomada de decisões comerciais. Após o treinamento do modelo, geralmente um ou mais modelos de aprendizado de máquina ou de aprendizado profundo são gerados. Esses modelos podem ser aplicados a novos dados para obter previsões e resultados de avaliação.

Um grande número de desenvolvedores desenvolve e treina modelos exigidos por serviços relevantes baseados em mecanismos populares de IA, como TensorFlow, Spark_MLlib, MXNet, Caffe, PyTorch, XGBoost-Sklearn, e da MindSpore.

Passo 4 Avaliar o modelo.

Um modelo gerado pelo treinamento precisa ser avaliado. Normalmente, você não pode obter um modelo satisfatório após a primeira avaliação e pode precisar ajustar repetidamente os parâmetros e dados do algoritmo para otimizar ainda mais o modelo.

Algumas métricas comuns, como a precisão, a recuperação e a área sob a curva (AUC), ajudam a avaliar efetivamente e obter um modelo satisfatório.

Passo 5 Implantar o modelo.

O desenvolvimento e o treinamento do modelo são baseados em dados existentes (que podem ser dados de teste). Depois que um modelo satisfatório é obtido, o modelo precisa ser formalmente aplicado a dados reais ou dados recém-gerados para previsão, avaliação e visualização. As descobertas podem então ser relatadas aos tomadores de decisão de maneira intuitiva, ajudando-os a desenvolver as estratégias de negócios corretas.

----Fim

4.2 Conceitos básicos de desenvolvimento de IA

O aprendizado de máquina é classificado em aprendizado supervisionado, não supervisionado e por reforço.

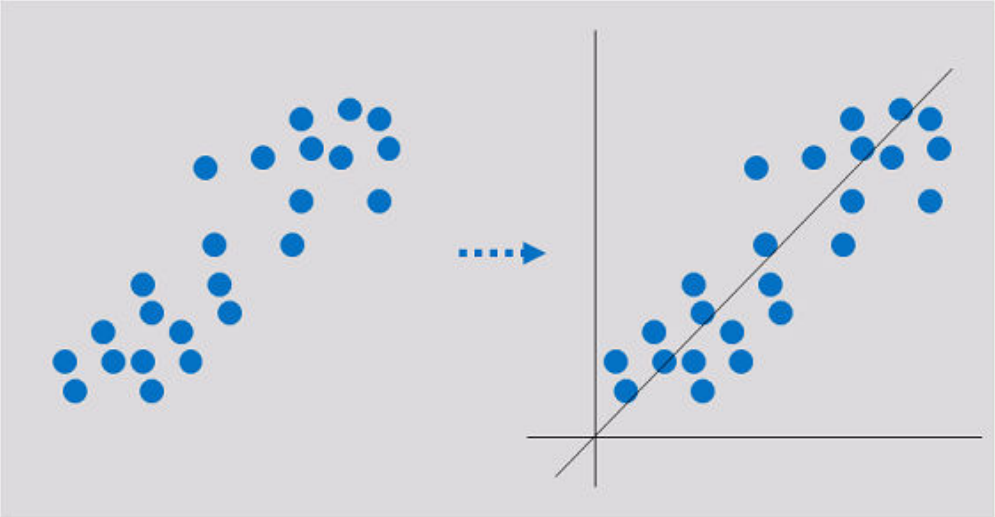
- O aprendizado supervisionado usa amostras rotuladas para ajustar os parâmetros dos classificadores para alcançar o desempenho necessário. Pode ser considerado como um

aprendizado com um professor. Aprendizagem supervisionada comum inclui regressão e classificação.

- O aprendizado não supervisionado é usado para encontrar estruturas ocultas em dados não rotulados. O clustering é uma forma de aprendizagem não supervisionada.
- Aprendizado por reforço é uma área de aprendizado de máquina preocupada com a forma como os agentes de software devem executar ações em um ambiente, de modo a maximizar alguma noção de recompensa cumulativa.

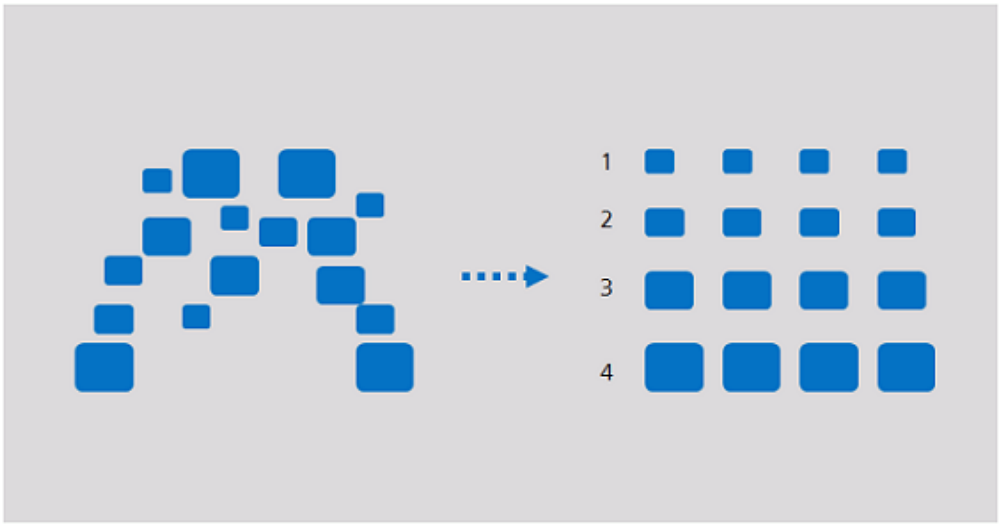
Regressão

A regressão reflete o recurso de tempo dos atributos de dados e gera uma função que mapeia um atributo de dados para uma previsão de variável real para encontrar a dependência entre a variável e o atributo. A regressão analisa principalmente dados e prevê dados e relacionamento de dados. A regressão pode ser usada para desenvolvimento de clientes, retenção, prevenção de rotatividade de clientes, análise do ciclo de vida da produção, previsão de tendências de vendas e promoção direcionada.



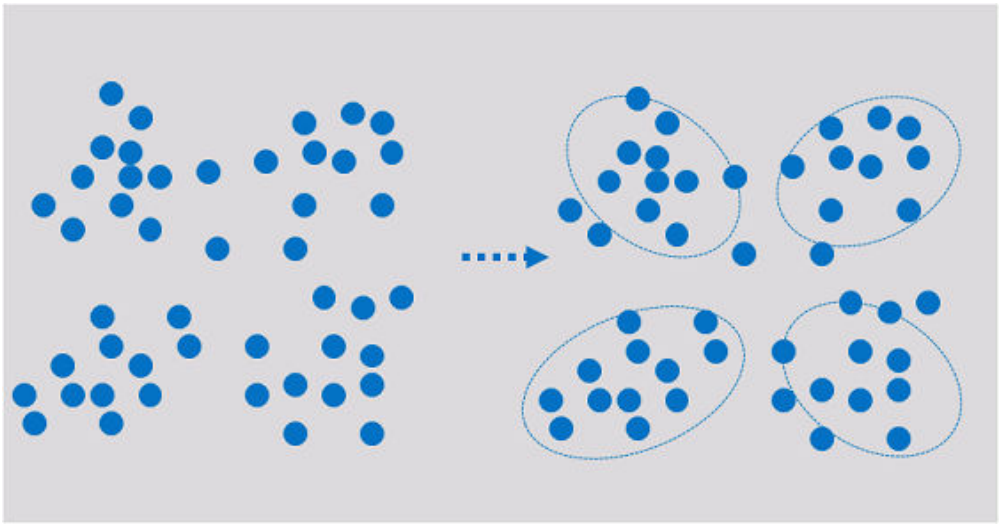
Classificação

A classificação envolve a definição de um conjunto de categorias com base nas características comuns dos objetos e a identificação de qual categoria um objeto pertence. A classificação pode ser usada para classificação de clientes, propriedades de clientes, análise de recursos, análise de satisfação do cliente e previsão de tendências de compra do cliente.



Clustering

O clustering envolve o agrupamento de um conjunto de objetos de tal forma que os objetos no mesmo grupo são mais semelhantes entre si do que aqueles em outros grupos. O clustering pode ser usado para segmentação de clientes, análise de características do cliente, previsão de tendências de compra do cliente e segmentação de mercado.



O clustering analisa objetos de dados e produz rótulos de classe. Os objetos são agrupados com base nas semelhanças maximizadas e minimizadas para formar clusters. Desta forma, os objetos no mesmo cluster são mais semelhantes entre si do que aqueles em outros clusters.

4.3 Conceitos comuns do ModelArts

ExeML

O ExeML é o processo de automatização do projeto de modelo, ajuste de parâmetros e treinamento de modelo, compactação de modelo e implantação de modelo com os dados rotulados. O processo é livre de código e não requer que os desenvolvedores tenham

experiência em desenvolvimento de modelos. Um modelo pode ser construído em três etapas: rotular dados, treinar um modelo e implantar o modelo.

Device-Edge-Cloud

Device-Edge-Cloud indica dispositivos, nós de borda inteligentes e a nuvem pública.

Inferência

Inferência é o processo de derivar um novo julgamento de um julgamento conhecido de acordo com uma determinada estratégia. Na IA, as máquinas simulam a inteligência humana e completam a inferência baseada em redes neurais.

Inferência em tempo real

A inferência em tempo real especifica um serviço web que fornece um resultado de inferência para cada solicitação de inferência.

Inferência em lote

A inferência em lote especifica uma tarefa em lote que processa dados em lote para inferência.

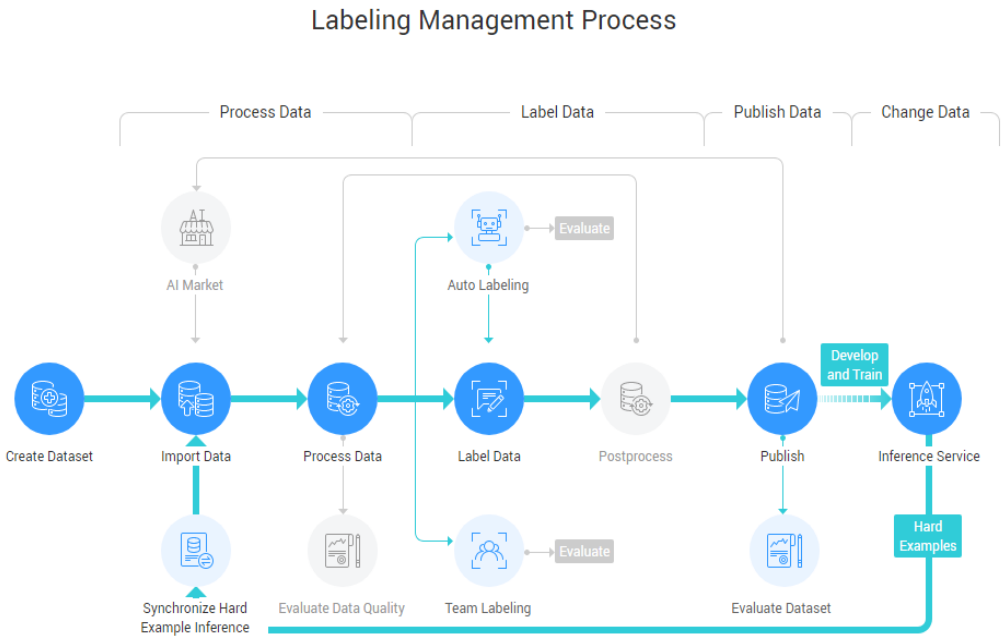
Chip Ascend

Os chips Ascend são uma série de chips de IA desenvolvidos pela Huawei com alto desempenho de computação e baixo consumo de energia.

4.4 Gerenciamento de dados

Durante o desenvolvimento da IA, grandes volumes de dados precisam ser processados, e a preparação e rotulagem de dados geralmente levam mais da metade do tempo necessário para todo o processo de desenvolvimento. O gerenciamento de dados do ModelArts fornece uma estrutura eficiente de gerenciamento e rotulagem de dados. Ele suporta tipos de dados de imagem, texto, áudio e vídeo em uma variedade de cenários de rotulagem, como classificação de imagem, detecção de objetos, rotulagem de parágrafos de fala e classificação de texto, para que o gerenciamento de dados possa ser usado em vários projetos de IA, como visão computacional, processamento de linguagem natural, e projetos de análise de áudio e vídeo. Além disso, o gerenciamento de dados ModelArts oferece funções como filtragem de dados, análise de dados, processamento de dados, rotulagem de equipe e gerenciamento de versão, permitindo gerenciar todo o processo de rotulagem de dados. **Figura 4-2** mostra o processo de rotulagem de dados.

Figura 4-2 Processo de rotulagem de dados



O gerenciamento de dados do ModelArts analisa e processa dados usando funções como análise de clustering, limpeza de dados, verificação, aumento de dados e seleção de dados, ajudando você a obter dados de alto valor que atendam aos requisitos de desenvolvimento ou projeto.

Com o gerenciamento de dados, o ModelArts permite rotular dados online para classificação de imagens, detecção de objetos, parágrafos de fala, trigêmeos de texto e vídeos. Você também pode usar a rotulagem inteligente para rotular automaticamente os dados por meio de algoritmos integrados ou personalizados, melhorando a eficiência da rotulagem.

Para apoiar a rotulagem colaborativa em larga escala, o gerenciamento de dados fornece rotulagem de equipe com gerenciamento de equipe, gerenciamento de pessoal e gerenciamento de dados para gerenciamento de projetos de processo completo, desde a criação do projeto, alocação de dados, controle de progresso, rotulagem, revisão, até a aceitação. Isso melhora a eficiência da rotulagem e minimiza os custos de gerenciamento de projetos.

O gerenciamento de dados do ModelArts garante a segurança e privacidade dos dados do usuário e permite que os dados sejam usados apenas dentro do escopo autorizado.

Na nova versão do gerenciamento de dados, os conjuntos de dados e a rotulagem de dados são desacoplados para facilitar suas operações.

4.5 Introdução às ferramentas de desenvolvimento

NOTA

Este documento descreve as funções do notebook DevEnviron da nova versão.

O desenvolvimento de software é um processo de redução dos custos do desenvolvedor e melhoria da experiência de desenvolvimento. No desenvolvimento de IA, o ModelArts

dedica-se a melhorar a experiência de desenvolvimento de IA e simplificar o processo de desenvolvimento. O DevEnviron ModelArts integra a cadeia de ferramentas de desenvolvimento para fornecer uma melhor experiência de IA na nuvem para desenvolvimento, exploração e ensino de IA.

O Notebook ModelArts para colaboração perfeita na nuvem e no local

- Plug-ins de JupyterLab na nuvem, IDE local e ModelArts para desenvolvimento e depuração remotos, adaptados às suas necessidades
- Ambiente de desenvolvimento em nuvem com recursos de computação de IA, armazenamento em nuvem e mecanismos de IA integrados
- Ambiente de tempo de execução personalizado salvo como uma imagem para treinamento e inferência

Característica 1: Desenvolvimento remoto, permitindo acesso remoto ao notebook a partir de um IDE local

O notebook da nova versão fornece desenvolvimento remoto. Depois de ativar o SSH remoto, você pode acessar remotamente o ambiente de desenvolvimento do notebook ModelArts para depurar e executar código de um IDE local.

Devido a recursos locais limitados, os desenvolvedores que usam um IDE local executam e depuram o código normalmente em um servidor de CPU ou GPU compartilhado entre os membros da equipe. Construção e manutenção do servidor de CPU ou GPU é caro.

As instâncias de notebook ModelArts estão prontas para uso com vários mecanismos e sabores internos para você selecionar. Você pode usar um ambiente de contêiner dedicado. Somente após configurações simples, você pode acessar remotamente o ambiente para executar e depurar o código do IDE local.

O notebook ModelArts pode ser considerado como uma extensão de um ambiente de desenvolvimento local. As operações como leitura de dados, treinamento e salvamento de arquivos são as mesmas realizadas em um ambiente local.

O notebook ModelArts permite que você use recursos na nuvem com os hábitos de codificação locais inalterados.

Um IDE local suporta o código do Visual Studio (VS), o PyCharm e o SSH.

Característica 2: Imagens predefinidas que são prontas para uso com configurações otimizadas e suporte a mecanismos de IA convencionais

Os motores de IA e as versões pré-configuradas em cada imagem são fixos. Ao criar uma instância de notebook, especifique um mecanismo de IA e uma versão, incluindo o tipo de chip.

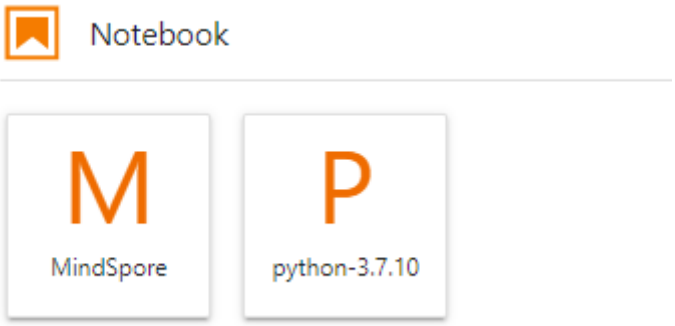
O ModelArts fornece um grupo de imagens predefinidas, incluindo imagens PyTorch, TensorFlow e MindSpore. Você pode usar uma imagem predefinida para iniciar a instância do bloco de anotações. Após o desenvolvimento na instância, envie um trabalho de treinamento sem qualquer adaptação.

As versões de imagem predefinidas no ModelArts são determinadas com base no feedback do usuário e na estabilidade da versão. Se o seu desenvolvimento pode ser feito usando as versões predefinidas no ModelArts por exemplo, o MindSpore 1.5, use imagens predefinidas.

Essas imagens foram totalmente verificadas e têm muitos pacotes de instalação comumente usados embutidos. Eles estão fora da caixa, aliviando você de configurar o ambiente.

As imagens predefinidas nos DevEnviron ModelArts são:

- Pacotes predefinidos comuns: Mecanismos comuns de IA como PyTorch e MindSpore, pacotes comuns de análise de dados como Pandas e Numpy, e ferramentas comuns como CUDA e CUDNN, atendendo aos requisitos comuns de desenvolvimento de IA.
- Ambientes predefinidos do Conda: Um ambiente Conda e um Python Conda básico (excluindo qualquer mecanismo de IA) são criados para cada imagem predefinida. A figura a seguir mostra o ambiente Conda para o MindSpore predefinido.



  Selecione um ambiente Conda com base em se o mecanismo AI é usado para depuração.
- Notebook: um aplicativo da Web que permite codificar na GUI e combinar o código, as equações matemáticas e o conteúdo visualizado em um documento.
- Plug-ins do JupyterLab: permitem mudança de variante, compartilhamento de casos na Galeria de IA para comunicação, e interrupção de instâncias para melhorar a experiência do usuário.
- Remote SSH: permite que você depure remotamente uma instância de notebook a partir de um PC local.
- Depois que as imagens predefinidas no ModelArts DevEnviron suportam o desenvolvimento, os trabalhos de treinamento podem ser executados no ModelArts.

NOTA

- Para simplificar as operações, o notebook ModelArts da nova versão não suporta alternância entre mecanismos de IA em uma instância de notebook.
- Os motores de IA variam de acordo com as regiões. Para obter detalhes sobre os mecanismos de IA disponíveis em uma região, consulte os mecanismos de IA exibidos no console de gerenciamento.

Característica 3: JupyterLab, uma ferramenta interativa de desenvolvimento e depuração online

O ModelArts integra JupyterLab de código aberto para desenvolvimento e depuração interativo online. Você pode usar o notebook no console de gerenciamento do ModelArts para compilar e depurar código e treinar modelos baseados no código, sem necessidade de instalação ou configuração de ambiente.

O JupyterLab é um ambiente de desenvolvimento interativo. É o produto da próxima geração do Jupyter Notebook. O JupyterLab permite compilar notebooks, operar terminais, editar texto Markdown, ativar interação e visualizar arquivos e imagens CSV.

5 Serviços relacionados

IAM

ModelArts usa o Identity and Access Management (IAM) para autenticação e autorização. Para obter mais informações sobre o IAM, consulte ***Guia de usuário de Identity and Access Management***.

OBS

O ModelArts usa Object Storage Service (OBS) para armazenar dados e modelos de forma segura e confiável a baixo custo. Para obter mais detalhes, consulte ***Guia de operação de console de Object Storage Service***.

Tabela 5-1 Relação entre ModelArts e OBS

Função	Sub-tarefa	Relacionamento
ExeML	Rotulagem de dados	Os dados rotulados no ModelArts são armazenados no OBS.
	Treinamento automático	Depois que um trabalho de treinamento é concluído, o modelo gerado é armazenado no OBS.
	Implementação de modelo	O ModelArts implementa modelos armazenados no OBS como serviços em tempo real.
Ciclo de vida do desenvolvimento de IA	Gerenciamento de dados	● Os conjuntos de dados são armazenados no OBS. ● As informações de rotulagem do conjunto de dados são armazenadas no OBS. ● Os dados podem ser importados do OBS.
	Ambiente de desenvolvimento	Os dados ou arquivos de código em uma instância de bloco de anotações são armazenados no OBS.

Função	Sub-tarefa	Relacionamento
	Treinamento de modelos	● Os conjuntos de dados usados pelos trabalhos de treinamento são armazenados no OBS. ● Os scripts em execução para trabalhos de treinamento são armazenados no OBS. ● Os modelos gerados pelas tarefas de treinamento são armazenados nos caminhos especificados do OBS. ● Os logs de execução dos jobs de treinamento são armazenados nos caminhos do OBS especificados.
	Gerenciamento de modelo	Depois que um trabalho de treinamento é concluído, o modelo gerado é armazenado no OBS. Você pode importar o modelo do OBS.
	Implantação de serviços	Os modelos armazenados no OBS podem ser implantados como serviços.
Configurações	-	Autoriza o ModelArts a acessar o OBS (usando uma agência ou chave de acesso) para que o ModelArts possa usar o OBS para armazenar dados e criar instâncias de notebook.

CCE

O ModelArts usa o Cloud Container Engine (CCE) para implantar modelos como serviços em tempo real. O CCE permite alta simultaneidade e proporciona dimensionamento elástico. Para obter mais informações sobre a CCE, consulte ***Guia de usuário de Cloud Container Engine***.

SWR

Para usar uma estrutura de IA que não é suportada pelo ModelArts, use o SoftWare Repository for Container (SWR) para personalizar uma imagem e importá-la para o ModelArts para treinamento ou inferência. Para obter detalhes sobre o SWR, consulte ***Guia de usuário de SoftWare Repository for Container***.

6 Como faço para acessar o ModelArts?

Você pode acessar o ModelArts pelo console de gerenciamento baseado na Web ou usando interfaces de programação de aplicativos (as API) baseadas em HTTPS.

- **Uso do console de gerenciamento**

  O ModelArts possui um console de gerenciamento simples e fácil de usar e fornece uma série de funções, incluindo ExeML, gerenciamento de dados, ambiente de desenvolvimento, treinamento de modelos, e implantação de serviços. Você pode concluir o desenvolvimento de IA de ponta a ponta no console de gerenciamento.

  Para usar o console de gerenciamento do ModelArts, você precisa se registrar na HUAWEI CLOUD primeiro. Se você se registrou na HUAWEI CLOUD, escolha **Products** > **AI**> **ModelArts** no site oficial e faça login no console de gerenciamento.

- **Uso dos SDK**

  Se quiser integrar o ModelArts a um sistema de terceiros para desenvolvimento secundário, chame os SDK para concluir o desenvolvimento. Os SDK do ModelArts encapsulam as API RESTful fornecidas pelo ModelArts para simplificar o desenvolvimento secundário. Para obter detalhes sobre os SDK e as operações, consulte ***Referência do SDK do ModelArts***.

  Além disso, você pode chamar diretamente os SDK do ModelArts ao escrever código em um notebook no console de gerenciamento.

- **Uso das API**

  Para acessar o ModelArts, use as API para integrar o ModelArts a um sistema de terceiros. Para obter detalhes sobre as API e operações, consulte ***Referência de API de ModelArts***.

7 Cobrança

O ModelArts é uma plataforma de desenvolvimento one-stop para desenvolvedores de IA. Com pré-processamento de dados, rotulagem semiautomatizada, treinamento distribuído, construção automatizada de modelos e implantação de modelos no dispositivo, borda e nuvem, o ModelArts ajuda os desenvolvedores de IA a criar modelos rapidamente e gerenciar o ciclo de vida do desenvolvimento de IA.

O ModelArts oferece dois modos de cobrança: pagamento por uso (cobrado com base na duração real) e pagamento anual/mensal (mais econômico).

Item cobrado

O ModelArts fatura você pelos recursos selecionados. **Tabela 7-1** lista os itens de cobrança. Para obter detalhes sobre o preço de cada item de cobrança, consulte **Detalhes de preços do produto**.

Tabela 7-1 Itens cobrados

Item cobrado	Descrição
Ciclo de vida do desenvolvimento de IA	Dedicado para desenvolvedores com experiência em desenvolvimento de IA. Ele suporta o desenvolvimento e implantação de algoritmos de aprendizado de máquina e aprendizado profundo, incluindo processamento de dados, desenvolvimento de modelos, treinamento e gerenciamento e implantação de serviços. Os itens de cobrança incluem ambientes de desenvolvimento de modelos (notebook), treinamento de modelos (trabalhos de treinamento e visualização) e implantação de serviços (serviços em tempo real).

Item cobrado	Descrição
ExeML	Dedicado para desenvolvedores com poucos recursos de desenvolvimento de IA. Suporta design automático, otimização e treinamento de modelos, bem como implantação de serviços, oferecendo desenvolvimento de IA sem código. Esse item de cobrança se aplica somente ao treinamento e à implantação de trabalhos do ExeML. Os itens de cobrança incluem trabalhos de treinamento, implantação de GPU e implantação de CPU no ExeML. Atualmente, apenas o modo de cobrança pagamento por uso é suportado.

Modos de cobrança

O ModelArts pode ser cobrado pelos seguintes modos:

- **Pagamento por uso**: permite que você faça uma assinatura ou cancelar a assinatura a qualquer momento. Esse modo de cobrança pode ser usado quando você seleciona recursos para criar um ambiente de desenvolvimento, criar um trabalho de treinamento ou implantar um modelo como um serviço.
- **Pacote de recursos pré-pagos**: permite que você compre um pacote com uma cota especificada. Quando você usa recursos, o sistema deduz o uso de recursos da cota e fatura os recursos do pacote em uma base de pagamento por uso. Compre um pacote na página **Dashboard** do console de gerenciamento do ModelArts.
- **Anual/Mensal**: faturado anualmente ou mensalmente. Este modo oferece um desconto maior do que o pagamento por uso.

  NOTA

  Somente pools de recursos dedicados podem ser cobrados anualmente ou mensalmente. As funções de pool de recursos dedicados e os métodos de compra variam dependendo dos sites. Para obter detalhes, consulte o console de gerenciamento.

  Para usar esse modo de cobrança, faça login no console de gerenciamento do ModelArts, clique em **Dedicated Resource Pools** no painel de navegação à esquerda e clique em **Create**. Se **Dedicated Resource Pools** não estiverem disponíveis no console de gerenciamento do ModelArts ou este modo de cobrança não estiver disponível na página para compra de um pool de recursos dedicado, a região atual não suportará o modo de cobrança anual/mensal.

Alteração do modo de cobrança

Ao usar o ModelArts você pode selecionar os recursos apropriados conforme necessário. O ModelArts fornece os seguintes métodos para você alterar os modos de cobrança após o início de uma tarefa:

- Se os recursos adquiridos não puderem atender aos requisitos de serviço, compre recursos com especificações mais altas.
- Um pool de recursos dedicado no modo de cobrança **Anual/Mensal** não oferece suporte a dimensionamento. Se você comprar um pool de recursos dedicados de pagamento por uso, poderá escalonar ou reduzir manualmente o pool de recursos dedicados. Você é cobrado com base no número de novos nós. Para obter detalhes, consulte **Escala de um pool de recursos dedicados**.

Se os métodos de alteração de configuração fornecidos pelo ModelArts não atenderem aos seus requisitos, você poderá criar um trabalho novamente e migrar dados.

Renovação

O ModelArts oferece assinatura de pacote de recursos pré-pagos e de pagamento por uso. No modo de pagamento por uso, as taxas são deduzidas a cada hora e um saldo insuficiente pode causar pagamentos em atraso. Para um pacote de recursos pré-pagos, quando você usar a cota deste pacote, o sistema cobrará automaticamente no modo de pagamento por uso. O serviço não será interrompido enquanto o saldo da sua conta for suficiente. Se sua assinatura não for renovada, seus serviços continuarão funcionando, mas entrarão em um período de retenção, durante o qual o ModelArts deixará de funcionar, mas os dados serão retidos.

- O período de retenção depende do seu nível. Para obter detalhes, consulte **Suspensão de serviço e período de release**.
- Para renovar a assinatura, acesse a página **Renewals**.

Expiração e pagamento em atraso

- Os recursos com o modo de cobrança de pacote de recursos pré-pago e pagamento por uso não expirarão. Se os recursos de um pacote de recursos pré-pagos tiverem sido usados, o uso subsequente dos recursos será cobrado em um modo de pagamento por uso. No modo pagamento por uso, as taxas são deduzidas a cada hora. Se o saldo da sua conta for insuficiente para pagar a despesa ocorrida na última hora, sua conta ficará em atraso e o ModelArts terá um **período de retenção**. Se os recursos forem renovados dentro do período de retenção, eles estarão disponíveis e você será cobrado a partir da data de expiração original.

Se a sua conta estiver em atraso, algumas operações serão restringidas. Recomendamos que você renove sua conta o mais rápido possível. **Tabela 7-2** descreve as operações restritas.

Tabela 7-2 Operações restritas devido a atrasos

Função	Operação restrita
ExeML	Treinamento e implantação de modelos
Gerenciamento de dados > Conjuntos de dados	Implantação de modelo com um clique
DevEnviron > Cadernos	Criação e início de instâncias de bloco de notas
Gestão da formação > Vagas de formação	Criação de postos de trabalho de formação
Gestão de formação > Busca automática de empregos	Criação de trabalhos de pesquisa automática
Implantação de serviços > Serviços em tempo real	Implantação de serviços em tempo real

8 Gerenciamento de permissões

Se você precisar atribuir permissões diferentes a funcionários diferentes em sua empresa para acessar recursos do ModelArts, o IAM é uma boa opção para gerenciamento de permissões refinado. O IAM fornece autenticação de identidade, gerenciamento de permissões e controle de acesso, além de fornecer acesso seguro aos recursos.

Com o IAM, você pode usar sua conta para criar usuários do IAM para seus funcionários e atribuir permissões para controlar o acesso deles a tipos de recursos específicos. Por exemplo, você tem um requisito de que certos desenvolvedores de software em sua empresa precisam usar recursos do ModelArts, mas não devem ter permissão para excluí-los ou executar operações de alto risco. Para atender a esse requisito, você pode criar usuários do IAM e conceder permissões que só permitem que eles usem recursos do ModelArts.

Se a conta tiver atendido aos seus requisitos, você não precisará criar um usuário independente do IAM para o gerenciamento de permissões. Então você pode pular esta seção. Isso não afetará outras funções do ModelArts.

O IAM pode ser usado gratuitamente. Você paga apenas pelos recursos em sua conta. Para obter mais informações sobre o IAM, consulte ***Visão geral de serviço de Identity and Access Management***.

Permissões do ModelArts

Por padrão, os novos usuários de IAM não têm nenhuma permissão atribuída. Você precisa adicionar um usuário a um ou mais grupos e atribuir políticas de permissões ou funções a esses grupos. Os usuários herdam permissões dos grupos aos quais são adicionados. Esse processo é chamado de autorização. Após a autorização, os usuários podem executar operações no ModelArts com base em permissões.

Para atribuir permissões ModelArts a um grupo de usuários, especifique o escopo como projetos específicos da região e selecione os projetos para que as permissões entrem em vigor. Se **All projects** estiver selecionado, as permissões entrarão em vigor para o grupo de usuários em todos os projetos específicos da região. Ao acessar o ModelArts os usuários precisam mudar para uma região onde foram autorizados a usar serviços em nuvem.

Você pode conceder permissões aos usuários usando funções e políticas.

- Funções: Um tipo de mecanismo de autorização de granulação grosseira que define permissões relacionadas às responsabilidades do usuário. Esse mecanismo fornece apenas um número limitado de funções de nível de serviço para autorização. Ao usar funções para conceder permissões, você também precisa atribuir outras funções das quais

as permissões dependem para entrar em vigor. No entanto, as funções não são uma escolha ideal para autorização refinada e controle de acesso seguro.

- Políticas Um tipo de mecanismo de autorização refinado que define as permissões para executar operações em recursos de nuvem específicos sob determinadas condições. Esse mecanismo permite autorização flexível baseada em políticas e atende aos requisitos de controle de acesso seguro. Por exemplo, você pode conceder permissões aos usuários do ECS que permitem que eles gerenciem apenas um determinado tipo de ECS. Para mais informações sobre as ações de API suportadas pelo ModelArts consulte **Referência de API > Políticas de permissões e ações suportadas**.

Tabela 8-1 lista todas as funções e políticas definidas pelo sistema suportadas pelo ModelArts.

Tabela 8-1 Políticas definidas pelo sistema suportadas pelo ModelArts

Nome da política	Descrição	Tipo de política
ModelArts FullAccess	Permissões de administrador para o ModelArts. Os usuários com essas permissões podem operar e usar o ModelArts.	Política definida pelo sistema
Model**Criação de uma política personalizada**Arts CommonOperations	Permissões comuns de usuário para ModelArts. Os usuários com essas permissões podem operar e usar o ModelArts mas não podem gerenciar pools de recursos dedicados.	Política definida pelo sistema

NOTA

Ao configurar permissões do ModelArts para um usuário do IAM, você precisa configurar as permissões de serviço do OBS correspondentes para que o usuário use o OBS corretamente.

- Para conceder permissões de administrador do OBS aos usuários, é necessário configurar uma política de **Tenant Administrator** que entre em vigor na região de serviço global para usuários do IAM. Para obter detalhes, consulte **Gerenciamento de permissões**.
- Para restringir as operações do usuário, você precisa configurar as permissões mínimas para usuários do ModelArts. Para obter detalhes, consulte **Criação de uma política personalizada**.

Tabela 8-2 lista as operações comuns suportadas por cada política do sistema.

Tabela 8-2 Operações comuns suportadas por cada política do sistema

Operação	FullAccess ModelArts	CommonOperations ModelArts
ExeML	Sim	Sim
Rotulagem de dados	Sim	Sim
Gerenciamento de dados	Sim	Sim

Operação	FullAccess ModelArts	CommonOperations ModelArts
Ambiente de desenvolvimento	Sim	Sim
Gerenciamento de modelo	Sim	Sim
Implementação	Sim	Sim
Galeria de IA	Sim	Sim
Pools de recursos dedicados	Sim	Não
Configurações	Sim	Sim

9 Cotas

O ModelArts utiliza os seguintes recursos de infra-estrutura:

- ECS
- EVS
- VPC

Para obter detalhes sobre como exibir e modificar a cota, consulte **Cotas**.

