**Clusters e modelos de IA-ML**

Introdução

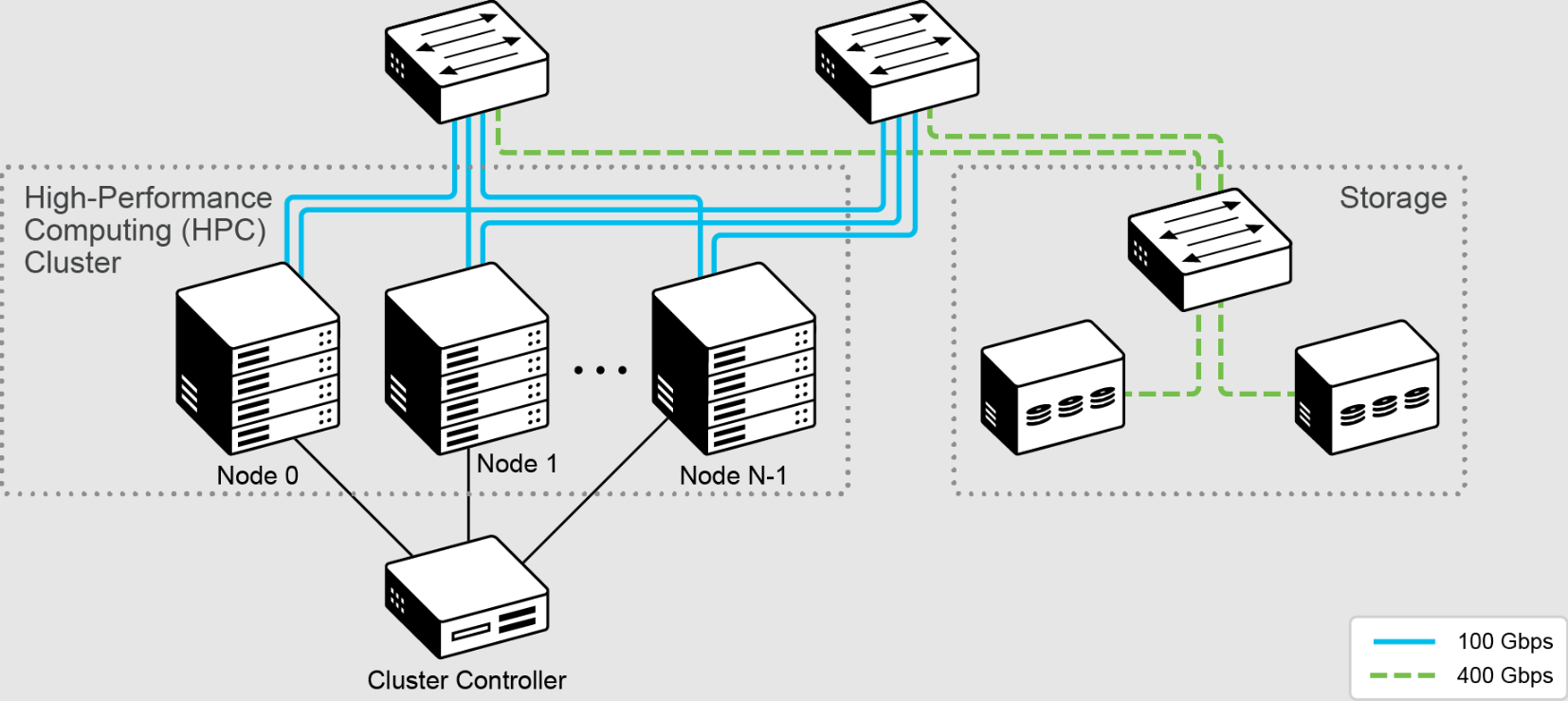
No mundo atual, impulsionado por dados, as organizações estão produzindo e analisando mais dados do que nunca. Você sabia que, até 2025, a esfera global de dados deverá atingir 175 zettabytes? Esse crescimento exponencial ressalta a necessidade crítica de infraestruturas avançadas de IA com recursos computacionais para processar e analisar vastos conjuntos de dados com eficiência.

Este treinamento examinará as complexidades dos clusters de computação de IA-ML, explorando seus componentes e casos de uso para obter insights sobre como esses sistemas são otimizados para os desafios específicos impostos pela IA e pelo aprendizado de máquina. Além disso, o treinamento explorará o desenvolvimento e a otimização de modelos de IA, abrangendo as ferramentas e técnicas que permitem a construção e a otimização eficientes de modelos adaptados a tarefas específicas, bem como como modelos de IA pré-construídos e pré-treinados podem ser ajustados às suas necessidades. Você também explorará as opções para implantar serviços de IA-ML, seja no local, na nuvem ou em sistemas distribuídos.

Clusters de computação de IA-ML

Devido aos constantes avanços tecnológicos em todos os setores, os humanos produzem, coletam, armazenam e analisam mais dados do que nunca. A análise de dados é uma das atividades computacionais mais intensivas que as organizações enfrentam atualmente. O treinamento de modelos de IA requer recursos computacionais substancialmente maiores para processar vastos conjuntos de dados e otimizar algoritmos complexos ao longo de inúmeras iterações.

Para combater a necessidade desses recursos, os sistemas habilitados para IA geralmente implantam um cluster de computação, um grupo de computadores interconectados, ou nós, que atuam como um sistema coeso. Para que um cluster seja eficaz na redução do tempo necessário para concluir tarefas de computação, seus nós devem processar dados em paralelo. Ao dividir uma carga de trabalho em vários nós que lidam com uma pequena parte da tarefa geral, o tempo de computação pode ser reduzido significativamente.



Treinar um modelo avançado de IA com clusters contendo milhares de nós ainda leva semanas ou meses. Um modelo pode ter bilhões ou trilhões de parâmetros e pontos de dados. O processo de treinamento de um modelo de IA nessa escala, sem o desempenho aprimorado dos clusters, pode levar centenas de anos.

Clusters de computação são altamente escaláveis, permitindo adicionar facilmente mais recursos de computação a eles para atender à demanda. Além disso, com múltiplos nós, os clusters promovem alta disponibilidade, garantindo que a falha de um nó não interrompa a operação de todo o sistema. Ferramentas de orquestração podem ajudar no gerenciamento de clusters de computação com funcionalidades como escalonamento automático para cima ou para baixo e substituição de nós com defeito.

Componentes do Cluster

Os nós são o núcleo de um cluster de computação, lidando com o trabalho pesado da computação. Os nós em um cluster otimizado para IA usam GPUs (Unidades de Processamento Gráfico) e TPUs (Unidades de Processamento Tensor) para computação. Ao contrário das CPUs, que possuem alguns núcleos poderosos otimizados para processamento sequencial, GPUs e TPUs são projetadas com muitos núcleos, muitas vezes milhares. Esses dispositivos são especializados para executar muitas operações simples e repetitivas simultaneamente, como calcular o valor que representa a cor que um pixel deve exibir e realizar cálculos matriciais e vetoriais envolvidos no treinamento de modelos de aprendizado profundo.

Tarefas como treinamento distribuído em larga escala envolvem uma grande quantidade de comunicação nó a nó. Portanto, os nós do cluster devem ser conectados por meio de uma estrutura de rede de altíssima velocidade, como InfiniBand ou RDMA sobre Ethernet Convergente (RoCEv2), frequentemente utilizando uma topologia de rede spine-leaf. Outras tarefas, como a execução de inferência, não exigem alta largura de banda, mas exigem alta redundância, disponibilidade e tempo de atividade. Devido aos requisitos exclusivos das diversas tarefas de engenharia ou computação relacionadas à IA, essas tarefas frequentemente serão tratadas por clusters completamente diferentes com especificações igualmente exclusivas. O mesmo se aplica ao armazenamento de dados distribuído. Em geral, a inferência com um modelo treinado requer menos recursos do que qualquer etapa do processo de treinamento.

Tabela

O conteúdo gerado por IA pode estar incorreto.

As configurações específicas de cada cluster podem ser definidas e gerenciadas por meio de um software de gerenciamento de cluster. Devido à natureza leve e portátil dos aplicativos em contêineres, eles são uma escolha comum para aplicativos executados em clusters. Docker e Kubernetes (K8s) são plataformas poderosas de orquestração de contêineres que podem ser usadas para implantar, gerenciar e escalar esses aplicativos com facilidade. O K8s também oferece diversos outros recursos úteis, como reinicialização automática de contêineres com falha, substituição de contêineres, balanceamento de carga, orquestração de armazenamento, gerenciamento de segredos e configurações, entre outros.

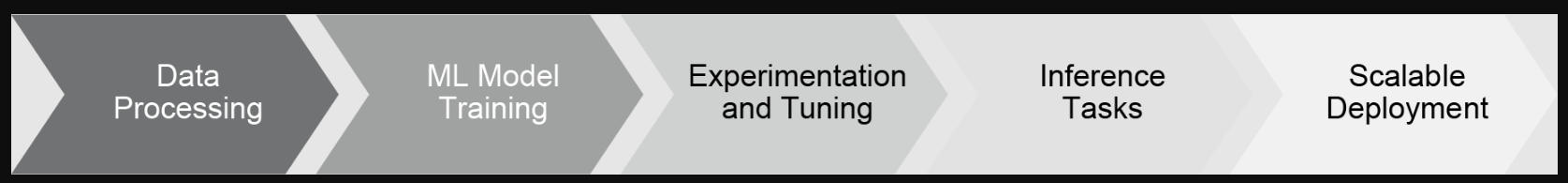
Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Casos de uso de cluster de IA-ML

Imagine que você é um cientista de dados em uma grande empresa de tecnologia encarregado de desenvolver um novo modelo de IA para melhorar o desempenho e a confiabilidade da rede para clientes corporativos. O uso de clusters de IA-ML no desenvolvimento do seu modelo de IA em cada etapa do processo de desenvolvimento aprimora o fluxo de trabalho e permite que você alcance seus objetivos com eficácia.

O cluster AI-ML oferece a capacidade de processar grandes quantidades de dados de rede, treinar modelos complexos, ajustar um modelo de ML com diferentes parâmetros, otimizar o desempenho da rede por meio de inferência e implantar um sistema escalável.



É muito importante pré-processar e limpar seu conjunto de dados de treinamento antes de começar a treinar e implantar seu modelo de IA. Essa tarefa requer a remoção de todos os dados irrelevantes, o tratamento de valores ausentes e a formatação de tudo para uso com o modelo. Clusters de IA-ML são excelentes para limpeza de dados, pois podem lidar com conjuntos de dados enormes, distribuindo cargas de trabalho por vários nós. Como resultado, mesmo os maiores conjuntos de dados são preparados rapidamente, tornando seu pipeline de dados robusto e escalável.

Assim que seu conjunto de dados estiver pronto, você poderá treinar seu modelo. Treinar em uma máquina com um conjunto de dados grande pode ser tediosamente lento e ineficiente. Usando clusters de IA-ML, o treinamento pode ser distribuído entre vários nós, cada um trabalhando em alguma parte dos dados ao mesmo tempo. Iterar pelos modelos com mais rapidez acelera o desenvolvimento, permitindo que você experimente diferentes algoritmos ou arquiteturas com facilidade.

Pode ser necessária uma quantidade significativa de experimentação com diversas arquiteturas de modelos e hiperparâmetros para encontrar o modelo de IA ideal. Esses experimentos podem ser realizados simultaneamente em um cluster de IA-ML. Se você quiser saber como o desempenho muda em função do modelo base, da taxa de aprendizado ou do número de camadas em uma rede neural, precisará executar várias instâncias desses testes em paralelo. Em um cluster de IA-ML, você pode implantar vários modelos ou versões do mesmo modelo com parâmetros diferentes simultaneamente para determinar rapidamente a melhor especificação para o seu sistema.

Após o treinamento e a otimização do seu modelo, você poderá implantá-lo para monitorar e otimizar o desempenho da rede em tempo real. Utilizando o poder de processamento coletivo dos nós, o cluster de IA-ML pode processar altos volumes de solicitações de inferência simultaneamente. À medida que o tráfego de rede flui pelo seu sistema, o modelo pode analisar padrões rapidamente, detectar anomalias e otimizar rotas sem atrasos.

Por fim, a capacidade de escalar adicionando mais nós ao seu cluster de IA-ML garante que você possa lidar com cargas de trabalho crescentes sem comprometer o desempenho. Seja para processar conjuntos de dados maiores, treinar modelos mais complexos ou lidar com mais solicitações de inferência, a escalabilidade do cluster permite que ele cresça conforme suas necessidades. Essa flexibilidade garante que sua infraestrutura possa se adaptar às demandas dinâmicas do desenvolvimento de IA, fornecendo uma solução à prova de futuro para seus projetos.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Modelos de IA personalizados – Processo

Às vezes, um modelo de IA pronto para uso não será capaz de resolver o seu problema ou lidar com as suas tarefas específicas de fluxo de trabalho. Nesses casos, você pode criar um modelo personalizado, criado do zero ou significativamente modificado a partir de modelos existentes para atender às necessidades específicas da sua aplicação. Modelos personalizados oferecem maior precisão, melhor desempenho e maior flexibilidade do que modelos pré-construídos, tornando-os muito mais adequados para tarefas especializadas.

Há várias etapas envolvidas no processo de construção de um modelo de IA personalizado. Primeiro, você precisa determinar o que o modelo pretende alcançar e como seu sucesso será medido. Esta etapa orienta o design, o treinamento e a avaliação do modelo.

Uma imagem contendo Diagrama

O conteúdo gerado por IA pode estar incorreto.

Após definir o objetivo do seu trabalho, os dados que sustentam esse objetivo devem ser coletados e processados ​​para uso no treinamento do seu modelo. Dependendo da intenção, os dados usados ​​para o treinamento podem vir de vários locais, incluindo bancos de dados existentes, web scraping, sensores e outros dispositivos de coleta de dados. Os dados também podem ser adquiridos em um marketplace de dados de IA, como Innodata, Databricks Marketplace ou Defined.ai. Os dados agregados devem ser limpos para remover ruídos e lidar com valores ausentes para garantir a qualidade. Nesse processo, técnicas de transformação de dados, como escalonamento de dados para um intervalo comum (normalização), podem ser aplicadas para garantir que os dados sejam adequados para análise e modelagem.

Após garantir um conjunto de dados de alta qualidade, desenvolva seu modelo, escolhendo, primeiramente, a arquitetura que será usada para treiná-lo. A arquitetura escolhida dependerá do tipo de problema que você está tentando resolver. Por exemplo, redes neurais são adequadas para tarefas de reconhecimento de imagens, enquanto árvores de decisão podem ser apropriadas para tarefas de classificação. Depois de decidir sobre sua arquitetura, você provavelmente precisará experimentar algumas estruturas para ver qual delas tem o melhor desempenho para sua tarefa.

Parte do processo de experimentação é o ajuste de hiperparâmetros, que otimiza os parâmetros de um modelo de aprendizado de máquina que não são aprendidos a partir dos dados, mas definidos antes do processo de treinamento. Esses parâmetros, conhecidos como hiperparâmetros, controlam vários aspectos do processo de treinamento e a estrutura do próprio modelo. O ajuste adequado desses hiperparâmetros pode melhorar significativamente o desempenho do modelo. O ajuste envolve o ajuste de parâmetros como a taxa de aprendizado, o tamanho do lote e o número de camadas para melhorar o desempenho. Ao trabalhar nesses detalhes, você deve trabalhar com uma versão menor do seu conjunto de dados se tempo e custo forem fatores importantes. Lembre-se de que o modelo e as configurações de melhor desempenho serão dimensionados com o tamanho do conjunto de dados de treinamento.

Por fim, o modelo é implantado em um ambiente de produção, onde pode fazer previsões com base em novos dados. O monitoramento contínuo é necessário para garantir que o modelo mantenha sua precisão e eficiência ao longo do tempo, bem como para aprimorá-lo no futuro.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Modelos de IA personalizados – Ferramentas

Diversas ferramentas estão disponíveis para cada etapa do processo de construção de um modelo de IA personalizado. Existem diversos fornecedores que oferecem ferramentas para pré-processamento de dados, seleção de modelos, treinamento e ajuste, avaliação de desempenho e implantação. Além das ferramentas listadas para implantar um modelo, todas as ferramentas mostradas na figura podem ser utilizadas com Python. Python é considerado o padrão da indústria para a maioria das tarefas envolvidas na construção de modelos de ML personalizados. Python oferece um amplo conjunto de bibliotecas e frameworks, facilidade de uso, forte suporte da comunidade e recursos de integração.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Para coleta e preparação de dados, ferramentas como Pandas e NumPy podem ser usadas para manipulação de dados e cálculos numéricos. TensorFlow e PyTorch são frameworks populares para aprendizado profundo, oferecendo flexibilidade e recursos poderosos para construção e treinamento de redes neurais.

Outras ferramentas, como o Optuna, oferecem otimização automática de hiperparâmetros, facilitando buscas exaustivas e aleatórias sobre valores de parâmetros. Esse processo permite que você experimente diversas configurações, conforme prescrito pelo método escolhido, para determinar a configuração ideal para o seu modelo.

Depois de decidir os valores dos hiperparâmetros e treinar seu modelo, a avaliação de seu desempenho pode ser realizada usando bibliotecas de visualização como Matplotlib e Seaborn, que ajudam a criar visualizações estáticas, animadas e interativas.

É comum que uma aplicação de modelo de ML seja conteinerizada. Plataformas como Docker e Kubernetes podem ser usadas para implantar e gerenciar qualquer aplicação conteinerizada, incluindo aquelas que utilizam uma implementação de modelo de ML. O Tensorflow é uma plataforma de código aberto para aprendizado de máquina que utiliza gráficos de fluxo de dados, permitindo que algoritmos de aprendizado de máquina sejam descritos como um gráfico de operações conectadas.

Por fim, existem ferramentas mais robustas que oferecem recursos que abrangem várias etapas da construção de um modelo de IA personalizado. Ferramentas como Scikit-learn, MLflow e Kubeflow podem ser usadas para construir, gerenciar e implantar modelos de aprendizado de máquina com eficiência, garantindo fluxos de trabalho otimizados e infraestrutura escalável. O Scikit-learn oferece funcionalidades abrangentes para preparação de dados, seleção e treinamento de modelos, ajuste de hiperparâmetros e avaliação de modelos. Ferramentas como MLflow e Kubeflow auxiliam no gerenciamento e na implantação de modelos de aprendizado de máquina, fornecendo uma plataforma para gerenciar o ciclo de vida completo do aprendizado de máquina, incluindo rastreamento de experimentos, empacotamento de modelos e implantação.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Otimização de modelo de IA pré-construído

Modelos de IA pré-construídos são modelos de aprendizado de máquina pré-treinados, desenvolvidos e treinados em grandes conjuntos de dados por especialistas na área de IA. Esses modelos são organizados em categorias como "Processamento de Linguagem Natural", "Visão Computacional", "Áudio" e assim por diante.

Para determinar qual modelo de IA atende às suas necessidades específicas, é comum executar o treinamento inicial em um conjunto muito reduzido de dados do seu conjunto de dados de treinamento com vários modelos. O resultado desse processo estabelece uma linha de base que indica qual modelo terá melhor desempenho após o ajuste fino.

Diagrama

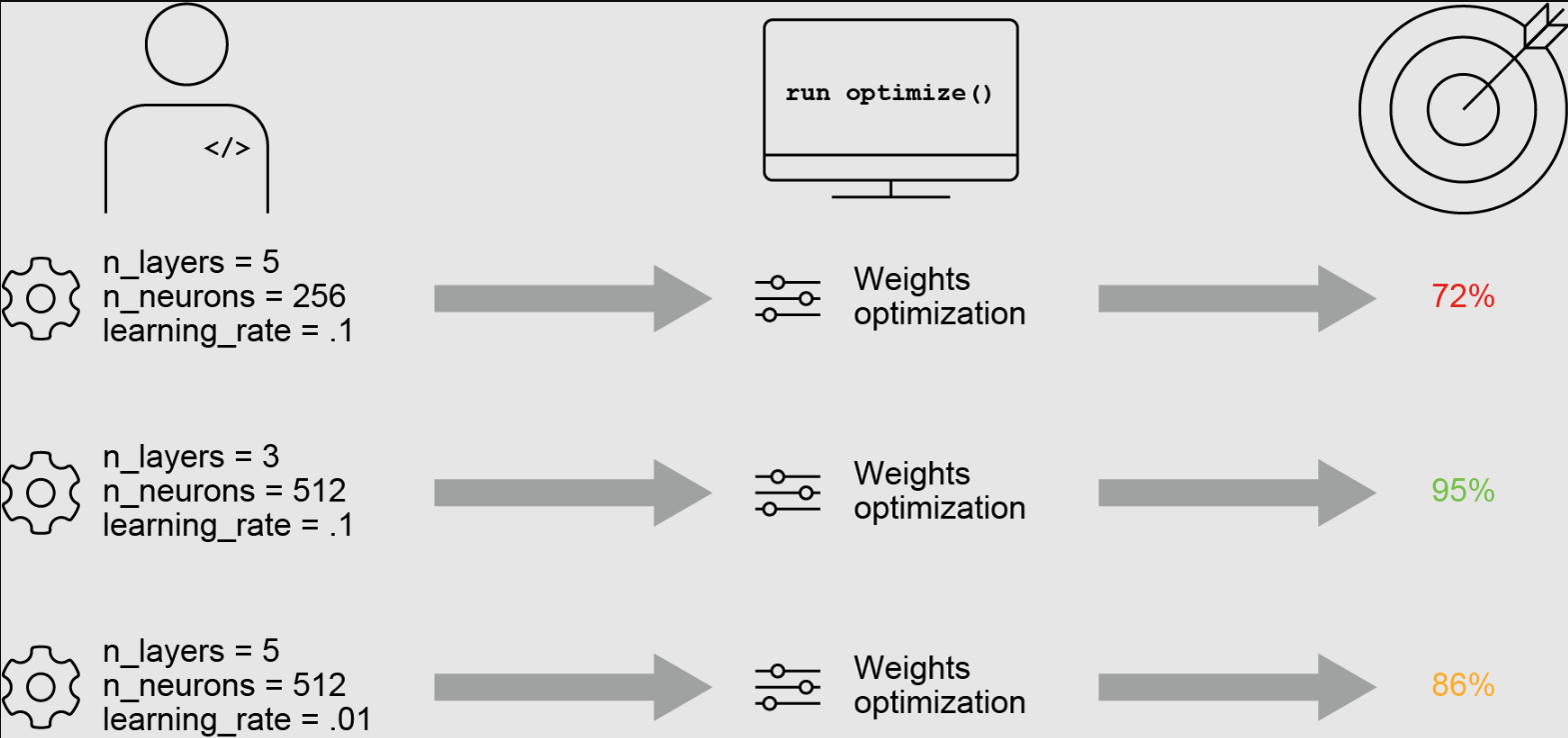
O conteúdo gerado por IA pode estar incorreto.

Otimizar modelos de aprendizado de máquina é uma etapa necessária no desenvolvimento de modelos para melhorar seu desempenho, eficiência e confiabilidade. Pular essa etapa seria como comprar o tipo certo de sapato, mas não o tamanho certo. As configurações de uma alteração no modelo dependem da qualidade, do tamanho e de outros aspectos do conjunto de dados de treinamento ou da tarefa de IA. O processo de otimização garante que os modelos forneçam previsões o mais precisas possível e utilizem os recursos computacionais de forma eficiente, alinhando o modelo à tarefa específica.

Ajuste de hiperparâmetros

O ajuste de hiperparâmetros é uma das etapas mais importantes do processo de otimização do modelo. Os hiperparâmetros mais impactantes e conhecidos incluem a taxa de aprendizado, o número de camadas em uma rede neural e o número de neurônios por camada.

A taxa de aprendizado controla o quanto os parâmetros do modelo são ajustados a cada iteração ou período durante o treinamento. Em outras palavras, a taxa de aprendizado controla quantas vezes o algoritmo de aprendizado trabalhará em todo o conjunto de dados de treinamento. O número de camadas e neurônios por camada determina a capacidade e a complexidade do modelo, influenciando sua capacidade de aprender com os dados e generalizar para novos dados. Outros hiperparâmetros comuns, como o número de períodos e o tamanho do lote (o número de exemplos de treinamento usados ​​em uma iteração), também desempenham um papel importante no processo de treinamento e no desempenho do modelo.



A combinação correta de hiperparâmetros pode afetar significativamente a capacidade de um modelo de aprender com os dados e generalizar para dados novos e inéditos. O ajuste adequado pode melhorar substancialmente o desempenho do modelo, tornando-o mais preciso e confiável em aplicações do mundo real.

Engenharia de Recursos

Características são as características associadas aos dados que auxiliam um modelo de IA a fazer previsões. Por exemplo, em uma aplicação de IA imobiliária, as características incluem o tamanho de uma casa, o número de quartos e a localização. O processo de engenharia de características inclui a seleção das características mais adequadas e, em seguida, a sua alteração para que sejam mais úteis. Esse processo inclui dimensionar as características para um intervalo comum, converter categorias em números e garantir que as características estejam no formato correto para o modelo entender.

Técnicas como poda e quantização podem ser usadas para aumentar a eficiência e reduzir o tempo de inferência em modelos. A poda remove parâmetros redundantes, como neurônios desnecessários em uma rede neural ou pesos redundantes em uma camada, simplificando o modelo sem afetar significativamente o desempenho. A quantização reduz a precisão dos parâmetros do modelo, acelerando a inferência e diminuindo o tamanho do modelo.

Técnicas de Regularização

Técnicas de regularização adicionam pequenas penalidades para evitar que o modelo se torne muito complexo, reduzindo a quantidade de dados irrelevantes. Técnicas de regularização que podem ser usadas para garantir o bom desempenho do seu modelo incluem as seguintes:

Interface gráfica do usuário, Texto, Aplicativo, Site

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Site

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Site

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Modelos de IA pré-treinados

Treinar seu próprio modelo de aprendizado de máquina leva tempo, seus próprios dados e recursos computacionais poderosos. Um modelo pré-treinado é um tipo de modelo de aprendizado de máquina que foi treinado em um grande conjunto de dados e pode ser usado como um modelo independente ou como ponto de partida para modelos personalizados e ajustados.

BERT, usado em processamento de linguagem natural, e ResNet, usado para reconhecimento de imagens, são exemplos de modelos pré-treinados populares de código aberto. Esses modelos podem ser personalizados ou ajustados com base no seu conjunto de dados específico, permitindo que você desenvolva o conhecimento já internalizado pelo modelo. Usar o treinamento extensivo pelo qual o modelo pré-treinado já passou não só economiza tempo e dinheiro, como também produz resultados quase impossíveis de serem alcançados com dados limitados.

Assim como muitos produtos de software de código aberto, modelos pré-treinados gratuitos são hospedados em repositórios. Repositórios populares incluem o TensorFlow Hub, o PyTorch Hub e o Hugging Face Model Hub, que oferecem uma ampla gama de modelos para diferentes tarefas, como classificação de imagens, processamento de texto e muito mais. Como visto na figura a seguir, o site do Hugging Face possui modelos organizados por categorias e subcategorias de tarefas de ML, facilitando a busca por um modelo que atenda às suas necessidades.

Interface gráfica do usuário, Texto, Aplicativo, Email, Site

O conteúdo gerado por IA pode estar incorreto.

Essas plataformas não só fornecem acesso a modelos pré-treinados de alta qualidade, como também oferecem ferramentas e documentação para ajudar você a integrá-los ao seu projeto. O ajuste fino adapta o modelo para que ele se destaque na sua tarefa específica, aplicando o conhecimento adquirido durante o pré-treinamento antes de você colocá-lo em prática.

Para incorporar um modelo pré-treinado ao seu conjunto de dados específico, você executaria o seguinte:

1. Instale a biblioteca Hugging Face Python, que inclui o modelo pré-treinado.
2. Carregue o modelo pré-treinado e seu tokenizador, uma ferramenta que converte dados em dados numéricos para serem usados ​​como dados de entrada para o modelo.
3. Incorpore seus próprios dados preparando seu conjunto de dados em um formato compatível com o modelo, normalmente tokenizando texto ou pré-processando imagens.
4. Treine o modelo no seu conjunto de dados, permitindo que ele ajuste seus parâmetros para melhor se adequar aos seus dados específicos. Você pode usar uma taxa de aprendizado mais baixa para fazer ajustes finos em vez de mudanças drásticas, pois o modelo já possui um bom entendimento e desempenho desde o treinamento inicial.
5. Monitore o desempenho do modelo em seus dados, ajustando hiperparâmetros se necessário, até atingir o nível desejado de precisão e generalização.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Parâmetros do modelo de IA

Os parâmetros do modelo de IA são as variáveis ​​internas que um modelo de aprendizado de máquina ajusta durante o treinamento para fazer previsões precisas. Esses parâmetros são aprendidos a partir dos próprios dados, o que significa que mudam à medida que o modelo é treinado para melhor se ajustar aos dados. Os parâmetros do modelo são diferentes dos hiperparâmetros, que são definidos antes do treinamento e guiam o processo de aprendizado sem serem diretamente influenciados pelos dados. Os parâmetros do modelo moldam a forma como um modelo interpreta os dados de entrada e gera previsões. À medida que um modelo aprende, ele ajusta continuamente esses parâmetros para minimizar erros, efetivamente "aprendendo" os melhores valores que levam aos resultados mais precisos.

Parâmetros comuns de modelo em aprendizado de máquina incluem pesos e vieses:

* **Pesos** — os fatores aplicados a cada recurso de entrada para determinar sua influência na previsão do modelo.
* **Viés** — um valor constante adicional que permite ao modelo fazer previsões precisas mesmo quando todos os recursos de entrada são zero.

Em um modelo de regressão linear, por exemplo, um modelo de IA multiplica cada característica de entrada pelo seu peso correspondente e, em seguida, soma esses produtos. O viés é então adicionado a essa soma para deslocar a saída, fornecendo uma previsão final.

Por exemplo, em um modelo que tenta determinar o preço de uma casa, há três características que representam fatores como metragem quadrada, número de quartos e localização. Por meio do treinamento, o modelo determinará os pesos apropriados e, portanto, o viés dessas características com base em como elas parecem afetar o valor final da casa em todo o conjunto de dados.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Ao contrário de uma rede neural profunda, modelos de regressão linear como o mostrado têm apenas uma camada, a camada de entrada, na qual os valores de entrada são usados ​​no algoritmo do modelo, e um valor é produzido. Embora este exemplo seja uma regressão linear bastante simples, pode haver milhares ou até milhões de recursos em uma aplicação como análise de genoma ou processamento de texto que exigiriam um algoritmo extremamente complexo e bem ajustado.

Como o treinamento é a forma como os valores dos parâmetros são determinados, para melhorar sua precisão, você precisa fazer alterações nos elementos que podem ser configurados manualmente, como hiperparâmetros, recursos e qualidade dos dados. Entender e otimizar esses elementos é fundamental para garantir que o modelo tenha um bom desempenho não apenas com dados de treinamento, mas também com dados novos e inéditos.

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

**Posicionamento de serviços – local vs. nuvem vs. distribuído**

Como em qualquer aplicação, há inúmeras considerações a serem feitas ao determinar como e onde você hospedará seu serviço de IA/ML. O posicionamento do serviço pode depender dos requisitos funcionais e não funcionais do sistema:

* **Requisitos funcionais** — definem O QUE um sistema faz, como escalabilidade, integração com infraestrutura ou serviços existentes, experiência do usuário e segurança.
* **Requisitos não funcionais** — definem COMO o sistema executa suas funções. Esta categoria inclui fatores como latência, custo, confiabilidade, disponibilidade, conformidade regulatória e distribuição geográfica.

A alocação de serviços de IA/ML no local, na nuvem e distribuídos tem suas vantagens e desvantagens. A alocação que você escolher depende de qual consideração é mais importante para sua organização.

Gráfico, Gráfico de radar

O conteúdo gerado por IA pode estar incorreto.

No local

A implantação de serviços on-premises envolve a hospedagem de clusters de IA/ML nos próprios data centers da organização. Essa estratégia traz grandes benefícios de segurança e controle, necessários para o gerenciamento de dados confidenciais, especialmente em setores com requisitos de conformidade rigorosos, como finanças e saúde. A implantação on-premises também reduz a latência, o que é vital para a tomada de decisões em tempo real em situações como o monitoramento do tráfego de rede para detectar ameaças à segurança ou otimizar o desempenho da rede. As implantações on-premises permitem uma integração próxima com a infraestrutura de rede existente, permitindo que os modelos de IA/ML funcionem perfeitamente em conjunto com outras ferramentas de gerenciamento de rede.

Infelizmente, os custos de implantação de um modelo ou aplicativo local são muito altos, pois envolvem um investimento inicial substancial em hardware e esforços contínuos de manutenção. Escalar essa infraestrutura para lidar com demandas crescentes pode ser difícil e custoso, exigindo grandes comprometimentos de recursos.

Nuvem

Os serviços de alocação em nuvem oferecem muitas vantagens para aplicações de IA/ML, principalmente escalabilidade e flexibilidade. As plataformas em nuvem permitem que as organizações aumentem ou diminuam seus recursos de IA/ML, dependendo da análise de rede e da demanda flutuante. Essa capacidade facilita o gerenciamento de grandes volumes de tráfego de rede ou economiza dinheiro quando a atividade é baixa. A nuvem facilita a implantação de ferramentas avançadas de rede sem a necessidade de construir e manter a infraestrutura internamente. A implantação em nuvem também oferece recursos de gerenciamento centralizado, benéficos para organizações que gerenciam redes complexas e geograficamente dispersas.

No entanto, as desvantagens associadas ao uso de posicionamentos em nuvem incluem um maior risco de segurança que pode existir se os dados de rede forem processados ​​na nuvem. Além disso, pode haver atrasos no processamento em tempo real, vital para tarefas como detectar e responder à atividade da rede e gerar insights. As organizações também dependem da confiabilidade e do preço do provedor de nuvem, o que pode afetar a estabilidade operacional a longo prazo.

Distribuído

A implantação de serviços distribuídos oferece uma abordagem híbrida que combina os benefícios de ambientes locais e em nuvem, tornando-a particularmente adequada para aplicações de IA/ML em redes que exigem baixa latência e capacidade de processamento escalável. Em uma implantação distribuída, os modelos de IA/ML podem ser implantados na borda da rede — mais próximos da fonte de dados — para fornecer análises e tomada de decisões em tempo real. Por outro lado, o processamento e o armazenamento de dados em larga escala podem ocorrer na nuvem. Essa abordagem aumenta a resiliência por meio da redundância geográfica, garantindo que, se uma parte da rede ou infraestrutura falhar, outras partes possam continuar operando e fornecendo serviços críticos.

No entanto, gerenciar um ambiente distribuído é complexo, exigindo ferramentas sofisticadas e expertise para sincronizar dados e cargas de trabalho em diferentes locais. Essa complexidade pode aumentar os custos, pois as organizações precisam manter recursos locais e na nuvem. Garantir a consistência e a conformidade dos dados em vários ambientes também apresenta desafios significativos de rede, especialmente ao lidar com conjuntos de dados grandes e sensíveis.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Resumo

Considere as seguintes questões ao aplicar os conceitos que aprendeu em seu próprio ambiente:

1. Como você pode usar clusters de IA-ML para melhorar a eficiência e a escalabilidade de suas tarefas atuais de processamento de dados?
   1. Considere os desafios computacionais específicos que você enfrenta e como distribuir cargas de trabalho em vários nós pode resolver esses problemas.
2. Que etapas você seguiria para criar um modelo de IA personalizado, adaptado a um problema específico no seu ambiente de trabalho?
   1. Pense nos dados que você precisaria, na arquitetura do modelo que poderia ser mais eficaz e nas ferramentas que você usaria para cada estágio do processo de desenvolvimento.
3. Qual opção de posicionamento de serviço (no local, na nuvem ou distribuído) atenderia melhor aos requisitos funcionais e não funcionais de seus aplicativos de IA/ML?
   1. Avalie suas necessidades em termos de escalabilidade, custo, segurança e latência para determinar a estratégia de implantação mais adequada.