Primeiros passos com IA

Como a Lenovo e a Intel impulsionam aplicações práticas entregando IA mais inteligente para todos.

Lenovo ThinkSystem SC750 V4 Neptune®, com processador Intel® Xeon® 6 – Desempenho e eficiência sem compromissos.

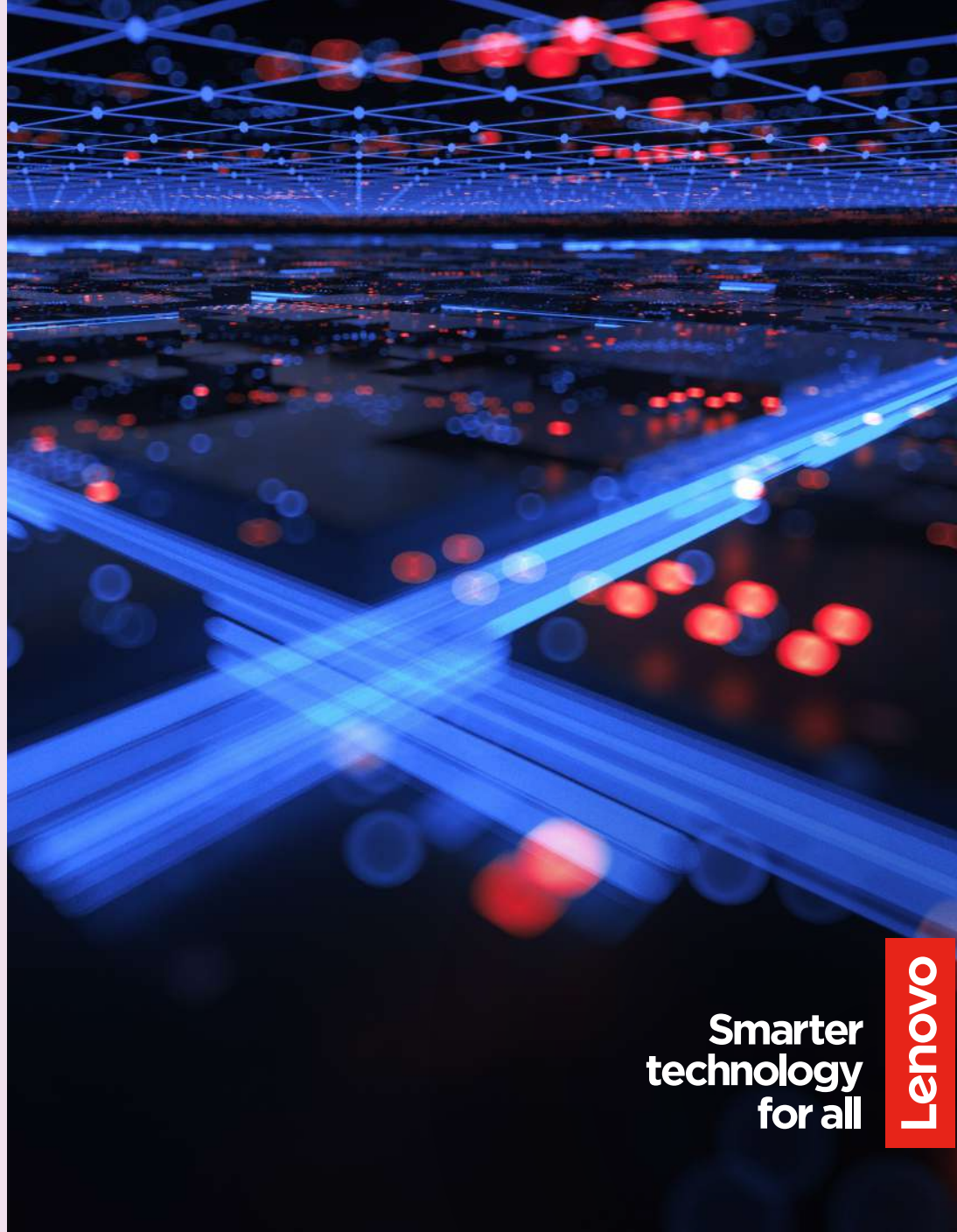


Tabela de Conteúdos

O rápido crescimento da IA

A Inteligência Artificial (IA) deu passos tremendos desde seus dias pioneiros na década de 1950. Os algoritmos estáticos predefinidos projetados para análise estatística e previsão, executados nos primeiros computadores, deram lugar aos primeiros exemplos de aprendizado de máquina na década de 1980, quando os algoritmos foram ensinados a reconhecer relacionamentos e construir modelos de sistemas complexos.

O advento de grandes redes neurais nos anos 2000 abriu caminho para expansões massivas de capacidade computacional e a introdução de IA generativa e grandes modelos de linguagem capazes de trabalhar com padrões complexos e abstratos.

De uma perspectiva de negócios, o potencial para derivar insights, reduzir cargas de trabalho e acelerar a produtividade parece quase ilimitado e as empresas estão investigando ativamente maneiras de colocar a IA para funcionar...

89%

dos líderes de TI dizem que estão pesquisando ou usando tecnologia habilitada para IA[1].

Habilitando a IA em todos os lugares

Lidar com iniciativas de IA pode ser assustador. Historicamente, a IA tem estado apenas no reino dos mecanismos de busca, instituições financeiras e pesquisa científica. Além do custo de aquisição do hardware de computação em si, em muitos casos, os data centers existentes não conseguem suportar os requisitos adicionais de energia e resfriamento, o que exige tempo e despesas de capital adicionais.

A boa notícia é que a introdução de modelos de IA de base ampla, treinados em dados públicos, diminuiu as barreiras para as organizações implementarem soluções avançadas de IA.

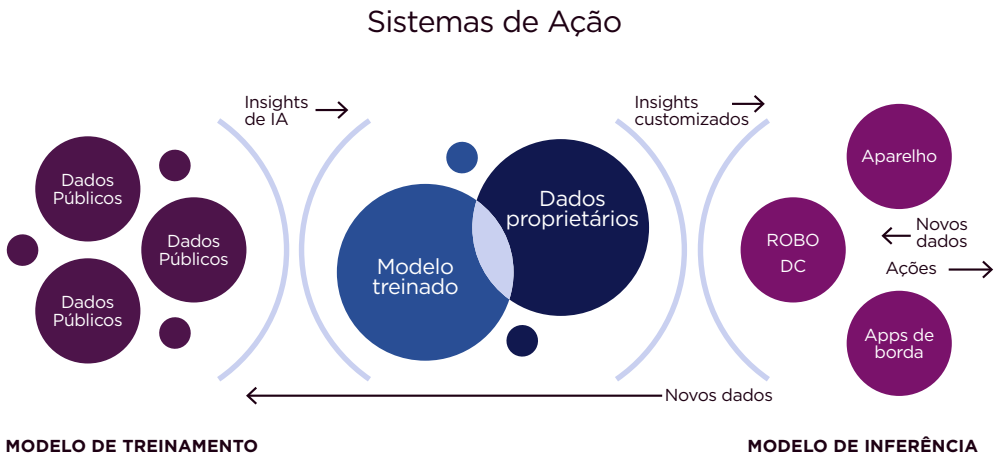
Lenovo e Intel estão colocando sua parceria de longa data para funcionar, fornecendo soluções que permitem que as empresas aproveitem todo o ótimo trabalho que foi feito até agora e apliquem a IA de maneiras práticas que entregam resultados mensuráveis.

Vamos começar com o básico

Em termos mais simples, a IA é amplamente definida como qualquer sistema de automação que simule a inteligência humana aprendendo no trabalho. A IA é implementada em duas fases:

1 **Treinamento ou desenvolvimento de modelo**
Este é o processo em que os cientistas de dados desenvolvem e otimizam modelos básicos com um conjunto de dados selecionado...

2 **Inferência**
Aplicando novos dados a um modelo treinado para obter novos insights e acelerar a automação.



Desenvolvimento de modelo de treinamento

O treinamento é alcançado por meio de um processo chamado aprendizado de máquina (ML), no qual um modelo é treinado com base em parâmetros específicos que definem a tarefa (por exemplo, cor, formas e bordas) e usa técnicas como clusterização, regressão e rede neural que processam enormes quantidades de dados para desenvolver previsões. A partir daí, o modelo continua a consumir e analisar dados enquanto melhora sua compreensão desses recursos.

Os conjuntos de dados usados para o treinamento de modelos básicos cresceram a uma escala que requer grandes quantidades de hardware de computação especializado que depende de milhares de processadores executados em paralelo para entregar os recursos necessários. É por isso que o treinamento de modelos básicos de IA tem sido tradicionalmente o domínio exclusivo de pesquisa acadêmica, financeira e governament.al

A quantidade de poder de computação necessária para treinar os maiores modelos de IA está dobrando a cada

3 a 10 meses.[2]

Apresentando a inferência de IA

A inferência de IA envolve pegar modelos treinados existentes e aplicá-los a novos conjuntos de dados proprietários para tarefas específicas de aplicação. O outcomes e os insights são então adaptados para novas aplicações que são personalizadas para entregar experiências mais precisas e relevantes.

A inferência se baseia no aprendizado já alcançado, então as demandas de processamento de gerar previsões e insights são significativamente menores do que aquelas requeridas durante o treinamento inicial.

Com os requisitos de processamento reduzidos, a inferência de IA está abrindo as portas para que mais negócios e organizações de todos os tamanhos aproveitem o poder da IA para uma ampla gama de aplicações.

Veja como a Lenovo e a Intel estão acelerando a Indústria 4.0 com inspeções visuais assistidas por IA e análises de dados mais rápidas.

Saiba mais.

O número de empresas que usam IA cresceu

300% em **5 anos.**[3]

Como as organizações não precisam desenvolver os modelos de treinamento básicos, isso acelera o desenvolvimento drasticamente e os ajuda a se mover em direção a aplicações reais mais rapidamente. Adicionalmente, como a abordagem precisa apenas de novas informações aplicadas ao modelo, os dados e demandas de processamento podem se estender além dos data centers. Isso significa que a inferência pode acontecer onde os dados são coletados, incluindo na borda.

Isso é importante porque permite o controle em tempo real de funções críticas que não incorrer em penalidades de latência indo e voltando para a nuvem, como os sistemas autônomos encontrados em carros autônomos ou fábricas automatizadas.

O modelo de inferência de IA treinado funciona apenas com os dados necessários para tomar as decisões, o que acelera o processo de decisão e reduz a necessidade de mover grandes quantidades de dados entre as redes.

Por exemplo, em um ambiente de manufatura, os servidores de borda na linha que executam modelos de inferência de IA podem usar visão computacional (baseada em modelos treinados existentes) para identificar defeitos, tomar decisões e tomar a ação apropriada (usando dados locais proprietários) para lidar com o defeito, mantendo a produtividade da linha.

Desbloqueie insights em seus dados mais rapidamente

À medida que as aplicações de IA evoluem, a tecnologia que as suporta está evoluindo para se adaptar a essas novas expectativas, enquanto acelera e capacita a implementação de IA em cada etapa, da borda à nuvem. Lenovo e Intel se uniram para entregar soluções construídas para fins específicos, projetadas especificamente para aplicações de inferência de IA.

A última geração de servidores ThinkSystem, como o **Lenovo ThinkSystem SC750 V4 Neptune,** alimentado por processadores Intel® Xeon® 6, projetados para IA. Intel® Xeon® oferece desempenho com eficiência energética e recursos de sustentabilidade para ajudar a diminuir o custo total de propriedade.
A aceleração integrada entrega um desempenho aumentado para tarefas de inferência de IA e reduz os requisitos de energia e resfriamento, o que significa que os servidores Lenovo ThinkSystem SC750 V4 Neptune podem ser implementados em data centers existentes em vez de construir novos centros.

5.5x
melhor desempenho de inferência de IA do que processadores concorrentes[4]

100%
de remoção de calor do resfriamento líquido Lenovo Neptune de 6ª Geração[5]

40%
redução no consumo de energia

Com a IA na borda, as organizações podem capitalizar informações dinâmicas em tempo real e entregar maior automação, remediação e insights onde é mais acionável - na linha de frente.

Aproveitando a Expertise em Transformação de IA

Projetar e implementar modelos de inferência de IA que entregam insights confiáveis e acionáveis exige um conjunto muito específico de habilidades e extrema atenção aos detalhes.

O **Lenovo AI Discover Center of Excellence** reúne especialistas em IA da Lenovo e da Intel para ajudar seus desenvolvedores a criar e acelerar a entrega de aplicações de IA e modelos de inferência de IA.

Nossos especialistas em IA conduzem uma ampla gama de workshops para fornecer avaliações abrangentes de negócios, avaliações de TI e planos de design documentados.

Engenheiros técnicos, parceiros e cientistas de dados otimizam seus códigos de IA usando frameworks de código aberto para executar nos servidores ThinkSystem com hardware e software Intel.

Podemos ajudá-lo a aproveitar o conjunto abrangente de recursos da Intel, como o toolkit OpenVINO™ e a oneAPI Deep Neural Network Library (oneDNN) para simplificar a implementação da inferência de aprendizado profundo para centenas de modelos pré-treinados.

A Lenovo também oferece uma ampla gama de workshops de Serviços Profissionais Lenovo para acelerar sua jornada de transformação de IA.

O Toolkit OpenVINO™

As barreiras à adoção de IA geralmente incluem a necessidade de modelos grandes, otimizados e diversos, uma ampla gama de arquiteturas xPU (muitas vezes implementadas juntas) e um ecossistema expansivo de frameworks de software de API para escolher. A implementação de IA pode ser um processo difícil e demorado, envolvendo muitas escolhas de ecossistema de fornecedores.

Com toda essa complexidade, as provas de conceito muitas vezes nunca chegam à produção, criando um "cemitério de POC".

Essas barreiras precisam ser derrubadas para criar oportunidade, e é isso que o **OpenVINO™** faz, oferecendo um toolkit de código aberto que suporta uma ampla gama de arquiteturas xPU e frameworks de software de IA.

Benefícios:

1. Ampla acessibilidade para múltiplas arquiteturas xPU através e um modelo de código aberto.

2. Uma solução de inferência de IA acessível e eficiente que reduz os custos de adoção e aplicação da tecnologia de IA da borda à nuvem para PCs locais.

3. Uma arquitetura aberta que permite colaboração em todo o ecossistema — desde cientistas de dados criando modelos para frameworks de aprendizado profundo até desenvolvedores de aplicações em uma variedade de verticais, utilizando funções de IA multimodal de visão, processamento de linguagem natural, sistemas de recomendação e IA generativa.

Emparelhado com a **Intel® Edge Platform,** soluções completas nativas da borda podem ser construídas para acelerar iniciativas de IA da borda com recursos de treinamento, otimização e desenvolvimento de aplicações para modelos de IA.

As empresas também podem integrar e gerenciar com segurança uma frota de nós de borda, aproveitando os componentes brownfield ou greenfield mais adequados e econômicos em parceria com nosso ecossistema incomparável para menor custo total de propriedade.

Veja como a Lenovo e a Intel estão facilitando a adoção de IA com o OpenVINO™

Saiba mais.

Acelere sua jornada com soluções de implementação comprovadas

Quando chega a hora de implementar sua solução de inferência de IA, o programa Lenovo AI Innovators agiliza o processo com soluções comprovadas usando o melhor software ISV da categoria na infraestrutura otimizada para IA da Lenovo e da Intel.

A Lenovo e a Intel constroem, testam e validam soluções de inferência de IA com um ecossistema de parceiros de AI Innovators comprovados para garantir implementações suaves e ideais que o mantêm dentro do cronograma e do orçamento.

- Solução de gerenciamento remoto da **Nybl**
- Solução de inspeção visual auxiliada por IA da **byteLAKE**
- Soluções de visão computacional, manutenção preditiva e detecção de anomalias da **Guise AI**
- Solução de Análise de Filas e Multidão do **WaitTime**
- Solução da **Sunlight.io** que acelera a transformação digital de restaurantes e drive-thrus
- Solução de inteligência industrial **Smartia** que conecta e transforma dados em percepções acionáveis

Continuamos a monitorar, avaliar e construir relacionamentos com parceiros ISV conforme suas soluções evoluem.

Estudo de caso: A IA está transformando as experiências do espectador

Lenovo e **WaitTime** revelam uma solução inovadora para locais para transformar a experiência do espectador da Fórmula 1® usando tecnologia de ponta. Ao combinar 18 câmeras estrategicamente instaladas em todo o autódromo do Circuit of The Americas (COTA) com a tecnologia de IA patenteada da WaitTime em servidores Lenovo ThinkEdge alimentados por Processadores escaláveis Intel® Xeon® , os operadores da COTA podem monitorar meticulosamente grupos de pessoas em filas.

“Esta plataforma de análise de dados em tempo real fornece insights inestimáveis, permitindo que os operadores entendam dinamicamente como as multidões estão crescendo, se movendo e mudando”, disse Zachary Klima, fundador e CEO da WaitTime.

“Tais informações instantâneas capacitam-nos a fazer ajustes imediatos nas operações e estratégias de receita, garantindo uma experiência ideal e contínua para os espectadores, ao mesmo tempo em que maximizam a eficiência e a receita do evento.”

Você pode ler mais sobre a solução **clicando aqui.**

Ganhe a flexibilidade para escalar perfeitamente

A implementação da inferência de IA requer muito menos em despesas iniciais em comparação com a construção e o treinamento de modelos básicos do zero, mas ainda existem custos a serem considerados para hardware, software e serviços.

Lenovo TruScale oferece a flexibilidade de um modelo escalável de pagamento conforme o uso para suas iniciativas de inferência de IA, fornecendo acesso à expertise que acelera suas iniciativas.

O modelo OpEx reduz o investimento inicial e escala com suas necessidades de negócios em constante mudança, permitindo que você leve perfeitamente projetos da prova de conceito à implementação e além.

Implementação mais rápida

Ao substituir os requisitos de aprovação de CapEx e mudar para um modelo OpEx, o TruScale pode aumentar a flexibilidade e acelerar os tempos de aquisição e implementação.

Opções escaláveis

Escolha entre um contrato fixo ou consumo medido para corresponder às necessidades da sua organização.

Expertise e serviços de IA integrados

Apoie-se nos serviços especializados da Lenovo para preencher lacunas de habilidades e recursos e ajudar a garantir o sucesso da implementação. Além disso, os gerentes de sucesso do cliente Lenovo dedicados podem ajudar a facilitar e coordenar com os recursos da Lenovo.

Essa flexibilidade não apenas torna mais fácil para uma gama mais ampla de organizações alavancar a inferência de IA, mas também torna a tecnologia à prova de futuro e elimina o risco de obsolescência à medida que a tecnologia evolui.

IA com um olho na sustentabilidade

O aumento da potência computacional necessária para treinar e operar modelos de IA significa mais eletricidade consumida e mais calor é gerado, o que continua sendo uma fonte de preocupação em todo o mundo.

E, à medida que a IA se integra a mais aspectos da vida cotidiana, o aumento resultante na capacidade de computação necessária apenas acelerará.

Por exemplo, uma pesquisa típica do Google usa menos de 0,3 watt-hora (Wh) por solicitação. Adicionar uma grande interação de modelo de linguagem à solicitação eleva esse requisito de energia para algo entre 7Wh e 9Wh por solicitação. Dado seu volume de pesquisa atual, se cada solicitação de pesquisa do Google incluísse um componente de IA, a IA do Google sozinha poderia consumir cerca de 30 terawatts-hora (TWh) por ano, ou quase o mesmo que o país da Irlanda[6].

A demanda global de energia de IA está projetada para aumentar

ao menos 10x

até 2026.[7]

A Lenovo e a Intel estão comprometidas com soluções sustentáveis, eficientes em termos de energia e ambientalmente responsáveis para a inferência de IA.

Os processadores Intel® Xeon® 6 são os processadores de data center mais sustentáveis da Intel, entregando mais de 2x melhor desempenho por watt em comparação com a geração anterior[8]. E eles podem ser implementados em data centers existentes sem requisitos adicionais de energia ou resfriamento.

No data center, a tecnologia de medição TruScale pode ajudá-lo a monitorar o consumo, a utilização e a temperatura para gerenciar o uso e os custos de forma mais eficiente. Além disso, nosso software Energy Aware Runtime (EAR) e o xClarity Energy Manager ajudam a entregar o desempenho ideal em um baixo nível de consumo de energia, otimizando os estados de energia, desligando componentes não utilizados e roteando cargas de trabalho para os recursos mais eficientes.

Otimizar seu data center com Lenovo TruScale ajuda a reduzir as emissões de CO2 e o consumo de energia em até 20%[9].

Pesquisa no Google <0.3Wh

Pesquisa no Google com tecnologia de IA: 7–9Wh

Todas as pesquisas no Google com IA: 30 TWh por ano

Uma abordagem mais inteligente para inferência de IA em todos os lugares

A inferência de IA detém uma tremenda promessa de acelerar o crescimento dos negócios, reduzir as cargas de trabalho e otimizar a eficiência para empresas em todos os setores.

Não importa onde você esteja na jornada para implementar soluções baseadas em IA em sua organização, a Lenovo e a Intel estão prontas para ajudar com soluções construídas para fins específicos, expertise líder do setor e os melhores parceiros da categoria.

Visite a página Intel AI Alliance para saber mais.

Fontes:

1 CIO, "How AI is transforming business today," setembro de 2024
2 Accenture, "Technology Vision 2023," março de 2023
3 Tidio, "10+ Essential AI Statistics You Need to Know for 2023," outubro de 2023
4 Dados da Lenovo com base em pesquisa interna da Lenovo ISG
5 Com base em testes internos da Lenovo comparando com sistemas semelhantes com resfriamento a ar em um data center típicoDe Vries, "The growing energy footprint of artificial intelligence," October 2023
6 De Vries, "The growing energy footprint of artificial intelligence," outubro de 2023See [9A2] at intel.com/processorclaims: Intel Xeon 6. Results may vary.
7 Agência Internacional de Energia, "Electricity 2024: Analysis and forecast to 2026," janeiro de 2024
8 Veja [9A2] em intel.com/processorclaims: Intel Xeon 6. Os resultados podem variar.
9 TruScale IaaS gera relatórios precisos sobre consumo de energia e emissões de CO_2, permitindo que infraestruturas gerenciadas sejam projetadas, implementadas e ajustadas não apenas para desempenho e capacidade, mas também para emissões de CO_2. O monitoramento contínuo do sistema com Lenovo XClarity Power Monitor e dados de desempenho dos sistemas é usado para otimizar o consumo de energia da infraestrutura. As emissões de CO_2 são calculadas com base na pegada de carbono local da fonte de energia utilizada.

Lenovo ThinkSystem SC750 V4 Neptune®, com processador Intel® Xeon® 6 – Desempenho e eficiência sem compromissos.