

intel
GAUDI

コスト・パフォーマンスと効率性を両立し 豊富な選択肢を提供する、高速かつ実用的なAI

Hugging Faceの測定によると、インテル® Gaudi® 2 AI アクセラレーターは生成AIやLLMなど最先端AIモデルの事前学習、実行、ファイン・チューニングでNVIDIA A100と比較した結果、トップクラスのパフォーマンスを発揮¹

顧客サポートの向上

自然言語処理

最大

1.7倍

NVIDIA A100と比較したBERT-Largeモデルのパフォーマンス²

事前処理フェーズ1での
処理文数 / 秒
BS=64、BF16/FP32



言語処理AIの
パフォーマンス向上により
業務運用を効率化

共通の問題を解決

情報に基づいた
ビジネス上の
意思決定で
収益を拡大

サポートを強化し
顧客の
維持率と
獲得率を向上

高速分析による
顧客の
フィードバックや
トレンドの特定

スピーディーな新規コンテンツ制作

テキストからの画像生成

最大

2.8倍

NVIDIA A100と比較したテキストから
画像を生成するStable Diffusionの
推論パフォーマンス³

レイテンシー
BS=8、BF16



生成AIの精度を高めて
新規コンテンツの
制作にかかる時間を短縮

共通の問題を解決

短時間で
ビジネスニーズに
合ったコンテンツを
制作

データ処理を
高速化し
リアルタイムの
テキスト生成画像分析や
意思決定に活用

大容量データを処理して
高まる需要を
AI生成コンテンツに
反映

データアクセスと顧客インタラクションの高速化

大規模言語モデル

最大

2.8倍

推論
BS=1、BF16

BLOOMZ-7Bの高速化

&

最大

1.4倍

推論
BS=1、BF16

NVIDIA A100と比較した BLOOMZ-176B
推論パフォーマンスの高速化⁴



多言語 LLM をスムーズに導入し、
データアクセスと
顧客のインタラクションを
高速化

共通の問題を解決

効率化と的確な
意思決定を促す
言語データの
高速分析

パフォーマンスを
高速化して
言語ベースの製品と
サービスの品質を向上

顧客の満足度、
ロイヤルティ、
維持率を向上

コンテンツの要約と分類を高速化

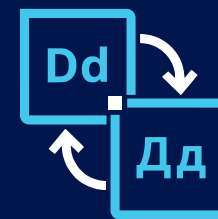
AI 処理によるテキストからテキストへの変換

最大

2.4倍

処理文数 / 秒
BS=1 (インテル® Gaudi®
アクセラレーター)
BS=16 (NVIDIA A100)
FP32

NVIDIA A100と比較した
T5-3B テキスト要約の
ファイン・チューニング性能⁵



テキスト関連
AI 言語処理の精度と
効率を向上

共通の問題を解決

テキストベースの
処理を高速化して
生産性アップ

問い合わせへの
効率的な対応と
顧客体験の
向上

手作業を削減し、要約、翻訳、
問い合わせへの対応、
分類などの業務にかかる
時間を短縮

intel.
GAUDI

詳細はお問い合わせください

インテル® Gaudi® 2 アクセラレーターは、AI Everywhere を実現し、大規模 AI の時代に求められる学習処理と推論の要件を満たす、コスト・パフォーマンス、性能、効率性の高いディープラーニング・コンピューティング・ソリューションで企業をサポートするために設計されています。

インテル® Gaudi® 2 アクセラレーターの詳細情報 : <https://habana.ai/products/gaudi2/> (英語)

¹ インテル® Gaudi® 2 アクセラレーターに対するサードパーティーの評価スコアに基づく。すべてのテストを Hugging Face が実施。BERT、T5-3B、BloomZ、Stable Diffusion のパフォーマンスを測定。詳細は、<https://habana.ai/products/gaudi2/> (英語)、<https://huggingface.co/blog/habana-gaudi-2-benchmark/> (英語)、<https://huggingface.co/blog/habana-gaudi-2-bloom/> (英語) を参照してください。

² BERT のパフォーマンスを測定。インテル® Gaudi® 2 アクセラレーターは複数の精度 (BF16/FP32) の組み合わせ、NVIDIA A100 は FP16 精度で実行。詳細は、<https://huggingface.co/blog/habana-gaudi-2-benchmark/> (英語) を参照してください。2024 年 2 月 21 日時点の更新情報。

³ Stable Diffusion のパフォーマンスを測定。インテル® Gaudi® 2 アクセラレーターは BF16、NVIDIA A100 は FP16 で実行。詳細は、<https://huggingface.co/blog/habana-gaudi-2-benchmark/> (英語) を参照してください。2024 年 2 月 21 日時点の更新情報。

⁴ BloomZ のパフォーマンスを測定。詳細は、<https://huggingface.co/blog/habana-gaudi-2-bloom/> (英語) を参照してください。2024 年 2 月 21 日時点の更新情報。詳細は、<https://habana.ai/products/gaudi2/> (英語)、<https://huggingface.co/blog/habana-gaudi-2-benchmark/> (英語)、<https://huggingface.co/blog/habana-gaudi-2-bloom/> (英語) を参照してください。

⁵ T5-3B のパフォーマンスを測定。インテル® Gaudi® 2 アクセラレーター、NVIDIA A100 とともに、勾配チェックポイントを有効にし、FP32 で実行。詳細は、<https://huggingface.co/blog/habana-gaudi-2-benchmark/> (英語) を参照してください。2024 年 2 月 21 日時点の更新情報。

性能は、使用状況、構成、その他の要因によって異なります。ワークロードと構成については、<https://habana.ai/habana-claims-validation/> (英語) を参照してください。

性能の測定結果は、構成に示されている日付時点のテストに基づいています。また、現在公開中のすべてのアップデートが適用されているとは限りません。構成の詳細については、補足資料を参照してください。絶対的なセキュリティを提供できる製品やコンポーネントはありません。実際のコストや結果は異なる場合があります。インテルは、サードパーティーのデータについて管理や監査を行っていません。ほかの情報も参考にしてデータの正確さを評価してください。

Intel、インテル、Intel ロゴ、その他のインテルの名称やロゴは、Intel Corporation またはその子会社の商標です。その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。

©2024 Intel Corporation. 無断での引用、転載を禁じます。