

DEEPLITE, INC

AIを毎日の暮らしに。さらに速く、最適化されたAIをエッジデバイスへ。

概要

ディープラーニングのオンプレミスソフトウェアは、エッジデバイスのAIアプリケーションやディープニューラルネットワークの製品化を加速させます。

特徴

- プッシュボタン最適化:** ユーザーはDeepliteの自動ソフトウェアを使い、pre-trainedモデルとデータセットの入力、許容可能な制度の設定を効率的に実行できます。
- 比類のないモデル圧縮:** Deeplite社の既存の圧縮率は市場で比類するものがありません。圧迫されたエッジデバイスと256KBのチップメモリーリソースでAIアプリケーションを実行することが可能です。
- データプライバシーとセキュリティー:** Deepliteのソフトウェアは自社内で利用可能であり、外部で共有されたデータセットやAIモデルの関連事項は削除されます。

今後の展開

- Deepliteはスマート量産化システム、そしてARMとRISC-Vマイクロコントローラーの推論エンジンの組み合わせられた自動ソフトウェアを急速に発展させています。ディープニューラルネットワークの精度とエンドツーエンドコントロールの最適化を備えています。

対コロナへの関連

- Deepliteのソフトウェアは、マスク不着用者の探知、顔認証や複数のカメラを使つての発熱者探知など、コロナ対策AIアプリケーションの開発、販売を促進します。

問題点

ここから

クラウド/GPU インференス HW	
スループット (samples/sec)	1.0x
パワー消費	>300w
コスト (ASP)	>\$5,000

より小さく速くそしてエネルギー節約されたエッジデバイスを稼働させ、組み込まれたAIポテンシャルのロックを解除する。

ここまで

スループット

最適化されたGPU インференス

スループット (samples/sec)

1.5-5.0x Faster

圧縮

エッジ/IoT インференス

パワー消費

<10w

コスト (ASP)

~\$10

解決策- Deepliteの自動化ソフトウェア

549MB

トレンモデル

-0.5% 許容可能な制度の設定

Deeplite

一次設計KPI

Compression: モード: OPTIMIZE

✓圧縮 ✓スピード ✓ディープサーチ ターゲット

パワー消費

結果

x46

x4

x5

-0.3%

圧縮 スピードアップ パワー低減 精度変化

11.9MB

最適化モデル

Deeplite社が提供するもの。それはAIエンジニアが自動的により速く小さくそして効率的なモデルアーキテクチャを設計するためのオンプレミス最適化プログラムです。

非常に優れた結果

アプリケーション	モデル	圧縮		改善	計算量低減 (FLOPs)	精度低下 (%)	データセット
		オリジナルサイズ	最適化サイズ				
イメージ分類	VGG19	80MB	2.16MB	x37	x5	<1%	CFAR100
	Resnet50	98MB	6.71MB	x14.6	x6	<1%	CFAR100
	Resnet18	45MB	3.16MB	x14.2	x6	<1%	CFAR100
	MobileNet-v1.0	12.8MB	530KB	x22	x5	~1.5%	Visual Wake Words
行動認識	Industry use case ¹	45MB	1.8MB	x25	x4	<1%	Subset of Imagenet
行動認識	Industry use case ²	1.9MB	59KB	x32	x100	~0%	Custom dataset
物体検出	Resnet50-SSD300	54MB	18MB	x3	x3	~0%	Subset of COCO2017

Deepliteの最適化プログラムにて得られた結果 (models in FP32)。プラットフォーム対応の最適化 (INT8, 混合精度、バイナリウェイトなど) をしてインференスエンジンを使い、メモリー削減、スピードアップそしてエネルギー節約が可能です。

カスタマーサクセス-スマートマニュファクチャリングAOI

- Deepliteを利用することで、モデル最適化のための多目的アプローチにより、エンジニアは精度に焦点を当て、推論用に生産準備モデルをシームレスに作成することができます。
- 上記のステップ2に注目すると、初期のMobileNetV1モデル (約12.8 MB、精度92%のバリデーションデータ) は、ArmCortex-M4を搭載した低消費電力カメラで実行する必要があります。
- しかし、今日までAIエンジニアはTop-1精度が2%未満、256 KBのオンチップメモリーにパラメータが適合するモデルを作成する必要がありました。

①: 中国製造の生野行フェニックス作業による検査 ②: フラットカメラ: 低消費電力カメラとMCUにより、製品の品質における自動検査が可能 ③: 最終製品のAOI/パケット追跡

④: スマート製造プロセス用のパイプライン自動検査。

モデル	サイズ (バイト)	GMacc	パラメータ (100万)	精度
初期	12.8 MB	0.583	3.21	Top1:92.443%
最適化後	144 KB (89.04x)	0.112 (5.21x)	0.14 (22.26x)	Top1:90.607% (-1.84%)