# インテル® HD グラフィックス向けにエッジベースの AI を最適化する

この記事は、インテル® デベロッパー・ゾーンに公開されている「Optimizing Edge-Based Intelligence for Intel® HD Graphics」の日本語参考訳です。

---

## AI に関する背景情報とエッジへの移行

私たちの日々の生活は、人工知能 (AI) ベースのアルゴリズムと関わりがあります。過去 60 年間、AI の研究は断続的に行われてきました。マシンラーニングあるいは多層のディープラーニングは、現代生活のあらゆる分野において AI の利用を推進しています。コンピューター・ビジョンによる識別と分類から自然言語処理や予測まで、さまざまな分野で AI が利用されています。これらの基本レベルのタスクは、意思決定などのより高いレベルのタスクにつながります。

ディープラーニングは通常、サーバー、クラウド、ハイパフォーマンス・コンピューティングと関連があります。クラウドにおける AI の使用は拡大していますが、一方でエッジ上 (PC、IoT デバイスなど) での AI 推論エンジンの利用も増えています。低レイテンシー、高可用性、コスト軽減 (例えば、サーバー上で推論アルゴリズムを実行するコスト)、プライバシー保護の必要性から、デバイスでマシンラーニングをローカルに実行するか、あるいはクラウドのみで行う傾向にあります。図 1 に、ディープラーニングのフェーズを示します。

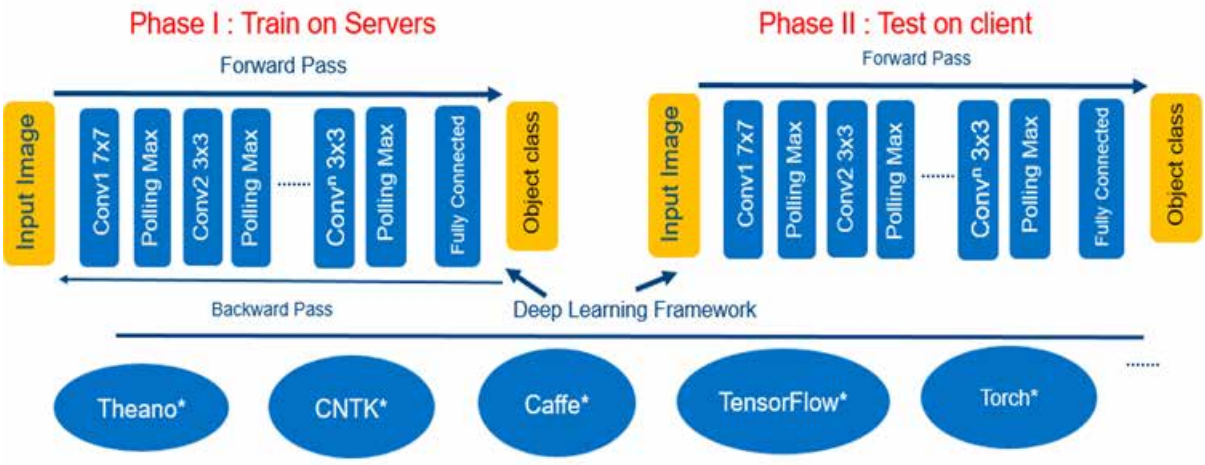



**図 1.** ディープラーニングのフェーズ

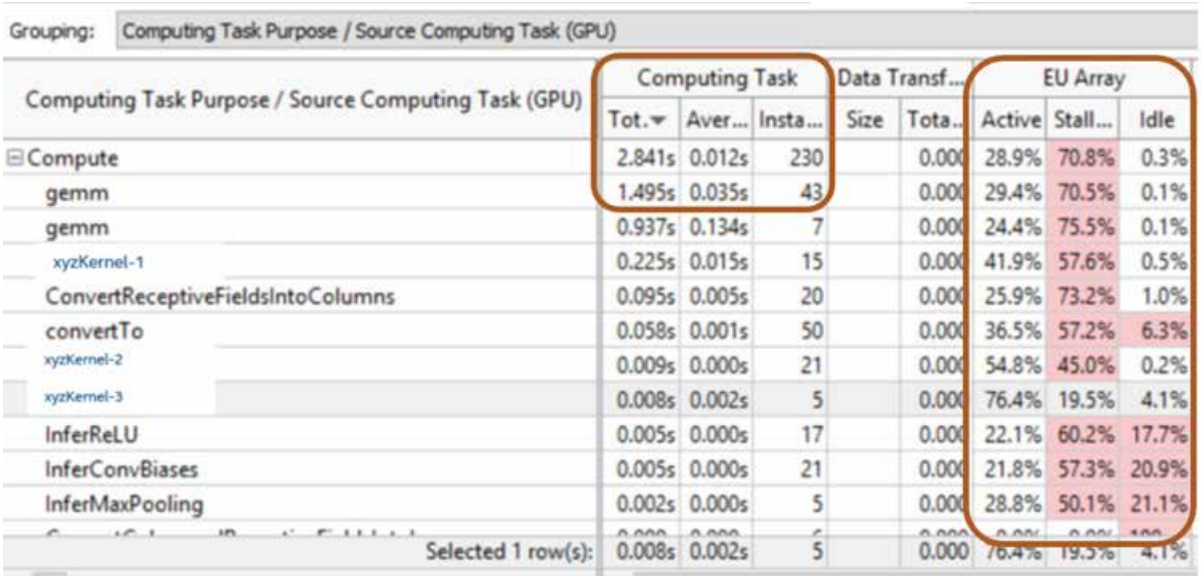
**AI の大衆化** – Personify は、リアルタイムのイメージ・セグメンテーションを行うエキスパート・システムです。2015 年に Personify は、インテル® RealSense™ カメラ内でリアルタイムのセグメンテーションを可能にしました。2016 年には、Microsoft* Skype*、Cisco Webex*、Zoom などの主要なビデオ・チャット・アプリケーションと OBS や XSplit などのストリーミング・アプリケーションにおいて、ユーザーの背景を削除/置換/ぼかすことができるアプリケーション、ChromaCam をリリースしました。ChromaCam は、リアルタイムに動的な背景置換を行うためディープラーニングを使用しており、ラップトップに搭載されている標準の 2D Web カメラで動作します。

**図 2.** Personify のセグメンテーション

Personify の要件の 1 つは、エッジ上でディープラーニング・アルゴリズムの推論エンジンの処理をできるだけ高速に実行することです。優れたセグメンテーション品質を実現するには、Personify はクラウドのレイテンシーを回避するため、推論アルゴリズムをエッジで実行する必要がありました。Personify のソフトウェア・スタックは、CPU とグラフィックス・プロセシング・ユニット (GPU) で実行しますが、元々ディスクリート・グラフィックス向けに最適化されていました。しかし、ディスクリート GPU を必要とする最適化されたディープラーニングの推論エンジンでは、一部の限られた PC 環境でしかアプリケーションを実行できません。さらに、セグメンテーションは、通常ゲームなどの高負荷のアプリケーションとともに使用され、多くのラップトップでは計算処理能力とシステムオンチップ (SoC) の熱要件の制約から、セグメンテーションの効果は非常に効率的であることが理想的です。インテルと Personify は、これらの課題に対応し、このテクノロジーをメインストリームのラップトップで利用できるようにするため、インテル® HD グラフィックス上で推論エンジンの最適化に取り組みました。

インテル® VTune™ Amplifier を利用して、インテル® HD グラフィックス 530 搭載の第 6 世代インテル® Core™ i7 プロセッサー上でディープラーニングの推論の GPU パフォーマンスのプロファイルと最適化を行いました。

図 3 は、クライアント PC における推論ワークロードのベースライン実行プロファイルです。アプリケーションは、ディープラーニング・アルゴリズムに GPU を使用していますが、パフォーマンス要件を満たしていません。インテル® HD グラフィックス上での最適化されていない推論アルゴリズムの合計実行時間は約 3 秒で、そのうち 70% は GPU がストールしています。



Grouping: Computing Task Purpose / Source Computing Task (GPU)

| Computing Task Purpose / Source Computing Task (GPU) | Computing Task | | | Data Transf... | | EU Array | | |
|---|---|---|---|---|---|---|---|---|
| | Tot. | Aver... | Insta... | Size | Tota... | Active | Stall... | Idle |
| Compute | 2.841s | 0.012s | 230 | | 0.000 | 28.9% | 70.8% | 0.3% |
| gemm | 1.495s | 0.035s | 43 | | 0.000 | 29.4% | 70.5% | 0.1% |
| gemm | 0.937s | 0.134s | 7 | | 0.000 | 24.4% | 75.5% | 0.1% |
| xyzKernel-1 | 0.225s | 0.015s | 15 | | 0.000 | 41.9% | 57.6% | 0.5% |
| ConvertReceptiveFieldsIntoColumns | 0.095s | 0.005s | 20 | | 0.000 | 25.9% | 73.2% | 1.0% |
| convertTo | 0.058s | 0.001s | 50 | | 0.000 | 36.5% | 57.2% | 6.3% |
| xyzKernel-2 | 0.009s | 0.000s | 21 | | 0.000 | 54.8% | 45.0% | 0.2% |
| xyzKernel-3 | 0.008s | 0.002s | 5 | | 0.000 | 76.4% | 19.5% | 4.1% |
| InferReLU | 0.005s | 0.000s | 17 | | 0.000 | 22.1% | 60.2% | 17.7% |
| InferConvBiases | 0.005s | 0.000s | 21 | | 0.000 | 21.8% | 57.3% | 20.9% |
| InferMaxPooling | 0.002s | 0.000s | 5 | | 0.000 | 28.8% | 50.1% | 21.1% |
| Selected 1 row(s): | 0.008s | 0.002s | 5 | | 0.000 | 76.4% | 19.5% | 4.1% |

**図 3.** インテル® VTune™ Amplifier によるセグメンテーションの GPU プロファイル (ベースライン)

ここで注目すべき重要なカラムは、一般的な行列-行列乗算 (GEMM) の *total time* (合計時間) と *execution unit (EU) stalls* (実行ユニットのストール) の 2 つです。ディープラーニングの推論エンジンを最適化しない場合、インテル® HD グラフィックス上のテレビ会議では、イメージ・セグメンテーションが非常に遅くなります。我々の目標は、メインストリームのコンシューマー向けラップトップにおいて、インテル® GPU の性能を最大限に引き出すことでした。

**最適化:** 行列-行列乗算カーネルを最適化し、EU のアクティブ時間を増やすことが最優先課題です。

図 4 は、畳み込みニューラル・ネットワークのデフォルトのパイプラインです。

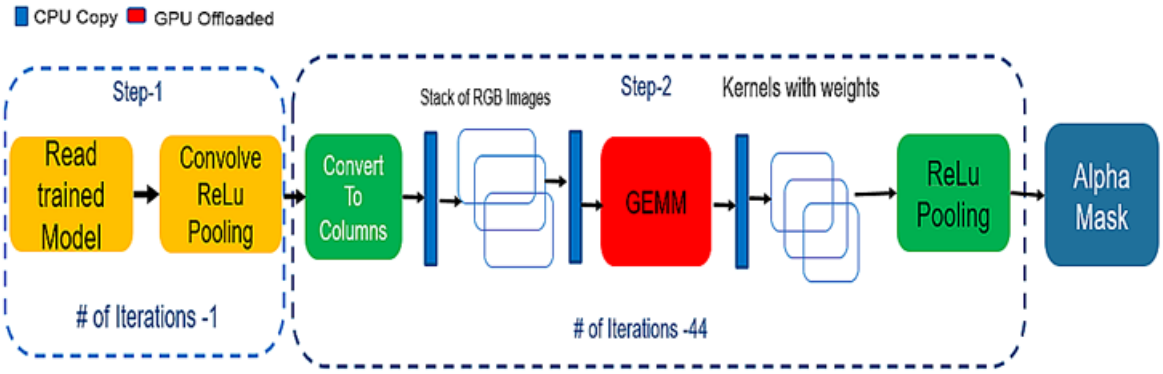



**図 4.** デフォルトのディープラーニング・パイプライン

インテル® HD グラフィックス上のディープラーニングの推論において、我々はいくつかのことに気付きました (図 5)。

- CPU コピー - アルゴリズムは、ディープラーニングの各層で処理のため CPU から GPU へデータをコピーするのに CPU を使用しています。
- GEMM 畳み込みアルゴリズム - OpenCV* および OpenCL* ベース。
- カラムへの変換 - 追加のステップとメモリーが必要です。

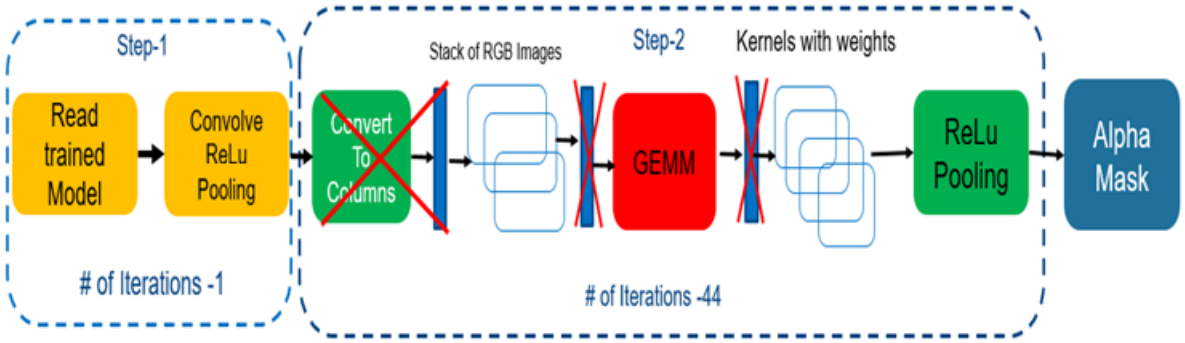



**図 5.** 余分なコピーの排除 (空間畳み込みを使用)

GEMM 畳み込みを空間畳み込みに置き換えることで、余分なメモリーコピーを回避し、速度が最適化されたコードを生成できました。また、メカニズムを自動チューニングすることで、アーキテクチャーの個々のカーネルの読み取り依存関係を解決できました (図 6)。

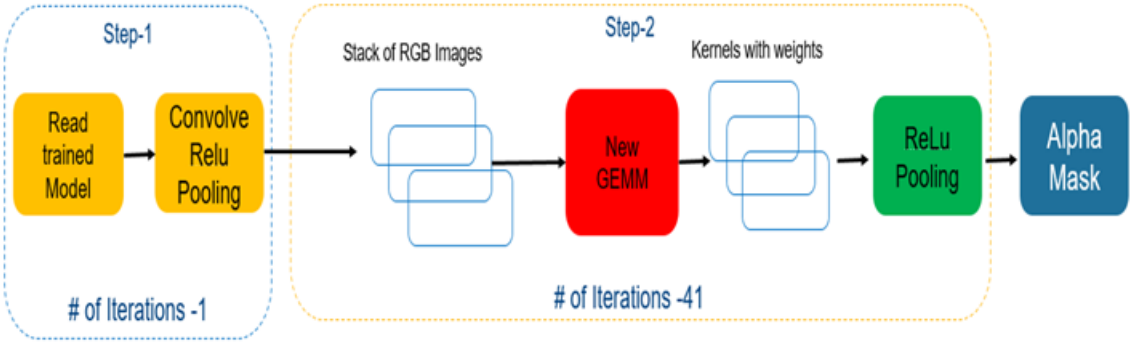



**図 6.** 新しい簡素化したアーキテクチャー

**結果:** インテル® HD グラフィックス搭載の第 6 世代インテル® Core™ i7 プロセッサー上でのテストでは、合計時間が 13 倍向上し (2.8 秒から 0.209 秒)、GPU 使用率が約 69.6%に改善され、GEMM カーネルのパフォーマンスが約 8.6 倍向上し (1.495 秒から 0.173 秒、または図 3 から 図 7)、フレームレートが向上したことでリアルタイム・セグメンテーションの品質も向上しました。

| Computing Task Purpose / Source Computing Task (GPU) | Computing Task | | | Data Transferred | | EU Array | | |
|---|---|---|---|---|---|---|---|---|
| | Total Time ▼ | Average Time | Instance Count | Size | Total, GB/sec | Active | Stalled | Idle |
| Compute | 209.398ms | 1.598ms | 131 | | 0.000 | 69.6% | 19.8% | 10.6% |
| Conv_Interleaved_32_2 | 170.884ms | 4.021ms | 41 | | 0.000 | 73.8% | 16.7% | 9.5% |
| DecodeApproxConvWeights | 25.462ms | 0.621ms | 41 | | 0.000 | 33.6% | 48.1% | 18.3% |
| SimpleGEMM | 4.456ms | 1.114ms | 4 | | 0.000 | 66.7% | 21.0% | 12.3% |
| InferConvBiases | 2.631ms | 0.146ms | 18 | | 0.000 | 24.4% | 45.7% | 29.8% |
| InferReLU | 2.376ms | 0.158ms | 15 | | 0.000 | 25.9% | 41.5% | 32.6% |
| InferMaxPooling | 0.943ms | 0.189ms | 5 | | 0.000 | 25.1% | 22.8% | 52.1% |

**図 7.** インテル® VTune™ Amplifier によるセグメンテーションの GPU プロファイル (最適化後)

モバイル環境での使用を可能にするため、ラップトップではバッテリー寿命も重要です。クライアントでは、消費電力の高いハイパフォーマンスなディープラーニング・アルゴリズムはユーザー体験に影響します。インテル® Power Gadget ツールを使用して、120 秒間のテレビ会議ワークロードの予測消費電力を解析しました。

**Personify アプリケーションの電力使用量**

- **GPU 電力使用量** 5.5W
- **SOC 電力** 11.28W

**まとめ:** クラウド・データセンターとクライアント間における計算の分担は変わりつつあり、ディープラーニング・モデルのアプリケーションはクライアントに移行しつつあります。ローカルモデルには、レイテンシー・オーバーヘッドに加えて、個人データをローカルに保持することでプライバシー保護の利点があります。インテル® プロセッサー・ベースのプラットフォームは、大きなコンシューマー・ベースのエコシステムをカバーするクライアントにおいて、ハイエンドの CPU と GPU が推論エンジン・アプリケーションを提供できるようにします。

## 参考文献


- Chromacam - https://www.chromacam.me (英語)
- インテル® GPU 向け GEMM カーネル - https://github.com/opencv/opencv/pull/8104 (英語)
- インテル® Power Gadget - https://software.intel.com/en-us/articles/intel-power-gadget-20 (英語)
- Personify - https://www.personify.com (英語)


コンパイラーの最適化に関する詳細は、最適化に関する注意事項を参照してください。