インテル® Xeon® スケーラブル・プロセッサーでの顔認識推論の最適化

この記事は、インテル® デベロッパー・ゾーンに掲載されている「Optimizing Face Recognition Inference on Intel® Xeon® Scalable Processors」の日本語参考訳です。

目次

はじめに

今日、新しいテクノロジー、特に人工知能 (AI) が現代社会に新たな生産性をもたらすことは誰も否定できません。中でも、ディープラーニングは非常に需要な役割を果たし、とりわけ顔認識において強力な改善をもたらします。顔認識は、ユニット保護、交通、金融、小売、犯罪鑑識など、さまざまなシナリオで広く使用されています。

このケーススタディーでは、インテル® Xeon® スケーラブル・プロセッサー上で顔認識推論のテストを行い、顔認識アプリケーションで AI がどのように活用されているかを示します。

通常、ディープラーニングのエンドツーエンドのパイプラインを使用する顔認識は、次のステップに従って実行されます。

1. 顔画像のデータセットを準備 (ソースは公開されているオープン・データセットまたはオンプレミス)
2. ディープラーニング・トポロジー ResNet50 を使用して顔検出モデルをトレーニング
3. トレーニング済みの検出モデルで推論
4. 推論された顔とローカル・データベースに格納されている人間の顔と特定済みの数百万の顔の特徴を比較し、顔照合アルゴリズムで類似性のスコアを付けてソート
5. 最も一致率が高い顔データを選択して類似性を表示

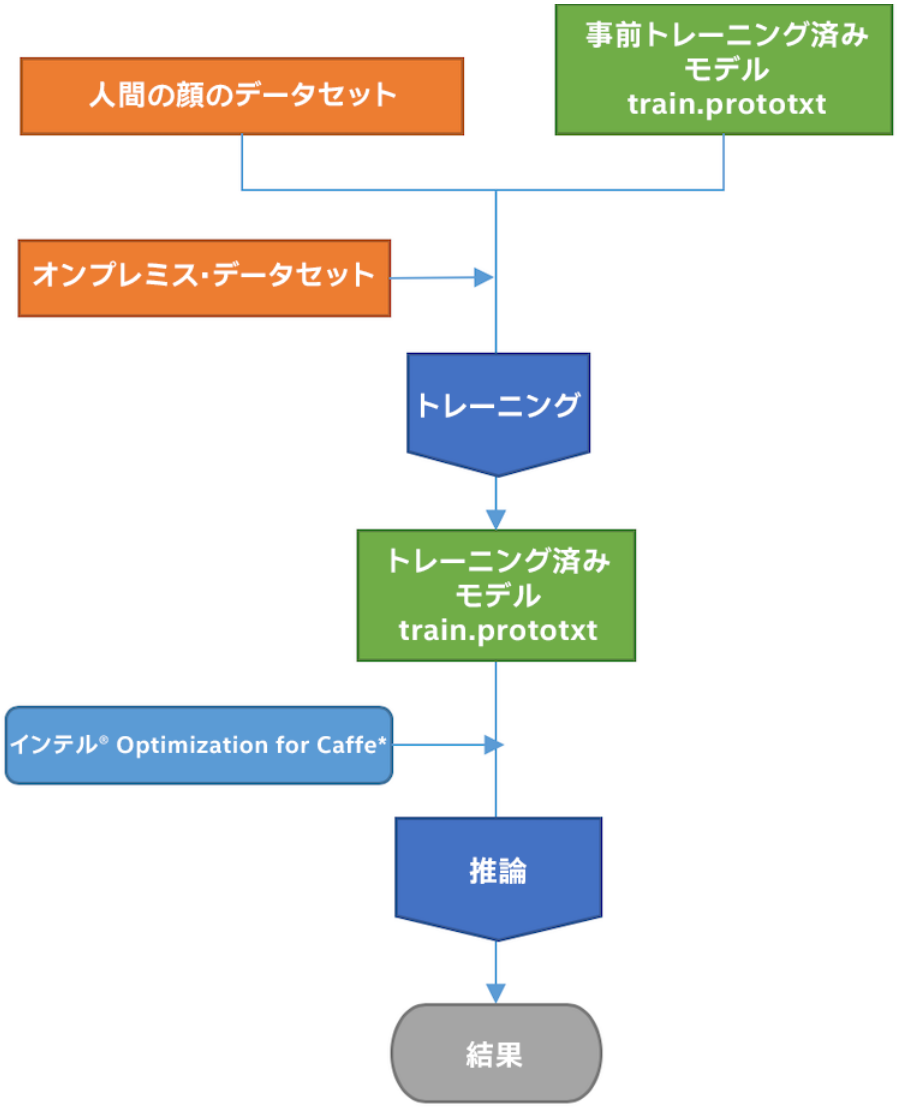
この記事では、ステップ 3 (トレーニング済み検出モデルで推論) のみ取り上げます。これは、インテル® Xeon® スケーラブル・プロセッサーで顔認識推論を最適化する上で重要なステップの 1 つです。

ソリューションのアーキテクチャーと設計

フレームワークとして、インテル® Optimization for Caffe* を使用します。このフレームワークには次の利点があります。

1. インテル® マス・カーネル・ライブラリー (インテル® MKL) やディープ・ニューラル・ネットワーク向けインテル® マス・カーネル・ライブラリー (インテル® MKL-DNN) などのライブラリーを活用して、行列の乗算と加算を高速化できます。
2. トポロジーの最適化により、同様の層は融合され、存続期間中に一度だけ計算されます。つまり、ほかの層は計算結果を非常に迅速に照会および取得できます。
3. さらに、モデルとエンドツーエンドのパイプラインの一貫性を維持するため、インテル® Optimization for Caffe* によって顔の特徴を検出するためのトレーニングがすでに行われています。

以下は、ここで使用するワークフローです。



アーキテクチャー・ワークフロー

トポロジー

メインの推論トポロジーとして ResNet50 を使用します。このトポロジーには次の利点があります。

- ここで使用する典型的な顔認識のユースケースでは、ドライバーは運転席に座っており、カメラはフロントウィンドウ越しにドライバーを監視します。SSD や Yolo などのほかのトポロジーと比較して、ResNet50 は優れた精度と安定したパフォーマンスを提供するため、このシナリオに適しています。
- ResNet50 は、「ソリューションのアーキテクチャーと設計」で述べたとおり、パフォーマンス・ブースト・モデル・トポロジーにより最適化されています。

ハードウェア構成

	構成 1	構成 2
プラットフォーム	x86_64	x86_64
ノード数	1	1
ソケット数	2	2
CPU	インテル® Xeon® プロセッサー E5-2699 v4 (55M キャッシュ、2.20GHz)	インテル® Xeon® Platinum 8180 プロセッサー(38.5M キャッシュ、2.50GHz)
ソケットごとのコア数とスレッド数	22, 44	28, 56
ucode	N/A	N/A
インテル® ハイパースレッディング・テクノロジー	無効	無効
インテル® ターボ・ブースト・テクノロジー	有効	有効
BIOS バージョン (dmidecode -s bios-version)	SE5C620.86B.00.01.0015.110720180833	SE5C620.86B.00.01.0015.110720180833
システム DDR メモリー構成	8 スロット/16GB/2400MHz	12 スロット/16GB/2666MHz
ノードごとの合計メモリー (DDR+DCPMM)	128GB	192GB
ストレージ - ブート	480GB	300GB
ストレージ - アプリケーション・ドライブ	480GB (ブートと共有)	300GB (ブートと共有)
NIC	N/A	N/A
PCH	N/A	N/A
OS	CentOS* 7.2	CentOS* 7.2
カーネル	N/A	N/A
緩和策バリアント (英語)	N/A	N/A
コンパイラー	GCC 4.8.5	GCC 4.8.5
ライブラリー	OpenCV* 3.4	OpenCV* 3.4
フレームワークのバージョン	インテル® Optimization for Caffe* 1.1.0	インテル® Optimization for Caffe* 1.1.0
インテル® MKL/インテル® MKL-DNN バージョン	インテル® MKL インテル® MKL-DNN 2018 Update2	インテル® MKL インテル® MKL-DNN 2018 Update2
データセット	オンプレミス・データセット	オンプレミス・データセット

	構成 1	構成 2
トポロジー	ResNet50	ResNet50
バッチサイズ	1,2,4,8,16,32,50,64,128	1,2,4,8,16,32,50,64,128

使用ソフトウェア

- ツール
  - インテル® Optimization for Caffe* v1.1.0
  - BVLC Caffe* v1.0
- 言語
  - Open Python* v2.7
- トポロジー
  - ResNet50 - インテル® Optimization for Caffe* に統合

必要なソフトウェアのインストール

次のリンクから必要なソフトウェアをインストールします。

インテル® Optimization for Caffe* のインストール (英語)
BVLC Caffe* のインストール (英語)

評価内容

前述の 2 つのハードウェア構成でインテル® Optimization for Caffe* と BVLC Caffe* のパフォーマンスを比較します。最初に、BVLC Caffe* (Public Caffe) をベースラインとして、インテル® Optimization for Caffe* のパフォーマンスを比較します。(図 1 と 4 を参照)

次に、2 つのハードウェア構成でインテル® Optimization for Caffe* のパフォーマンスを評価します。(図 2 と 3 を参照)

アーキテクチャーの説明で述べたとおり、インテル® MKL-DNN はディープラーニングの計算を高速化するように設計されており、一方、インテル® MKL は行列計算の基礎を成します。そのため、インテル® MKL-DNN またはインテル® MKL を使用して、異なるトポロジーのパフォーマンスを評価します。(図 5 を参照)

テストコマンド

```
caffe time --forward_only --phase TEST --iterations 100 --model <model.caffemodel> --engine <mkl/mkldnn>
```

これは推論のスクリプトです。ここで、**model.caffemodel** はトレーニング済みの Caffe モデル、**mkl/mkldnn** はインテル® MKL またはインテル® MKL-DNN を使用するエンジンです。

テスト結果

以下のグラフで、インテル® MKL-DNN エンジンを使用するインテル® Optimization for Caffe* は、BVLC Caffe* と比較して、インテル® Xeon® プロセッサー E5-2699 v4 では 5.96 倍、インテル® Xeon® Platinum 8180 プロセッサーでは 6.37 倍のパフォーマンス・ゲインを達成しています。インテル® Xeon® プロセッサー E5-2699 v4 とインテル® Xeon® Platinum 8180 プロセッサー上でのインテル® Optimization for Caffe* のパフォーマンスの比較では、後者は 1.67 倍のスピードアップを示しています。

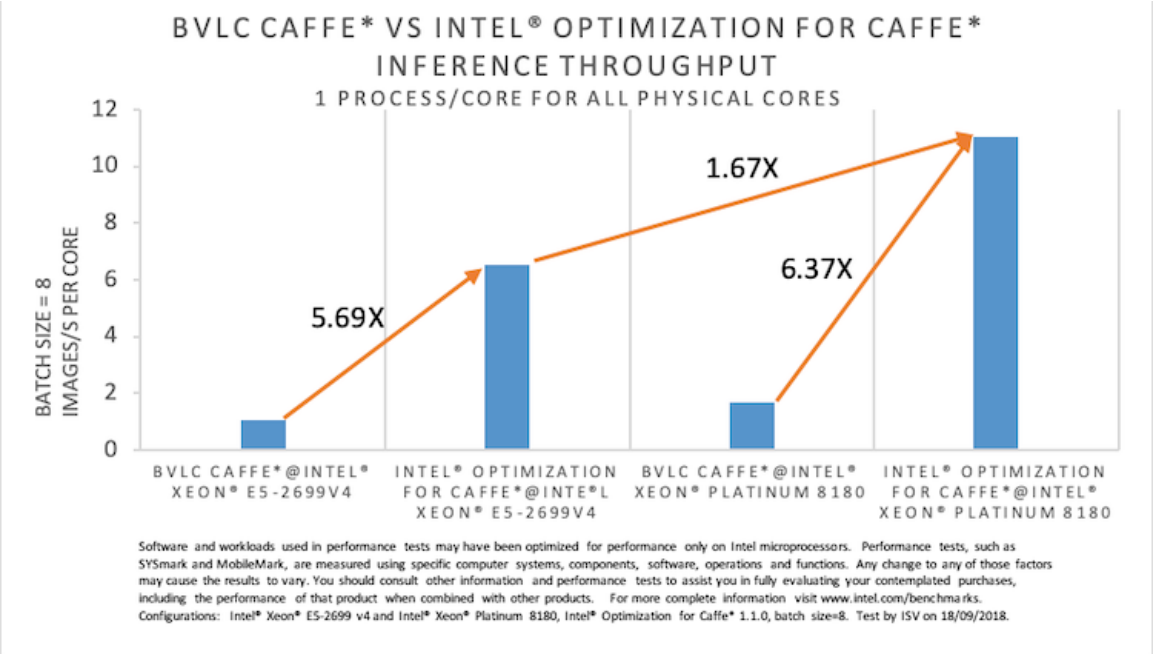


図 1. インテル® Xeon® プロセッサー E5-2699 v4 とインテル® Xeon® Platinum 8180 プロセッサー上での BVLC Caffe* とインテル® Optimization for Caffe* の推論スループットの比較

次のグラフは、インテル® Xeon® プロセッサー E5-2699 v4 とインテル® Xeon® Platinum 8180 プロセッサー上ですべてのコアを使用した場合のインテル® Optimization for Caffe* の推論スループットを示しています。インテル® Xeon® Platinum 8180 プロセッサー上のパフォーマンスは、インテル® Xeon® プロセッサー E5-2699 v4 と比較して 2.13 倍高速です。

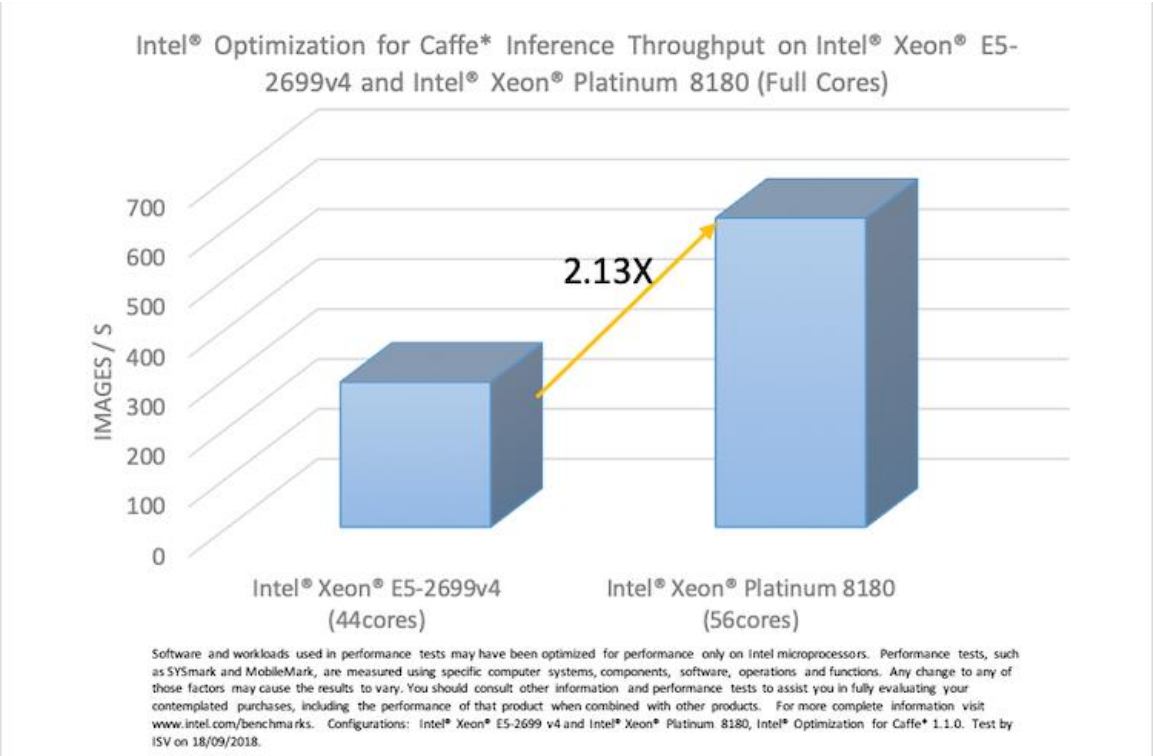


図 2. インテル® Xeon® E5-2699 v4 とインテル® Xeon® Platinum 8180 上のインテル® Optimization for Caffe* の推論スループットの比較 (すべてのコアを使用した場合)

図 2 と図 3 の違いは、それぞれのシステムにおいて、図 2 ではすべてのコアを使用していますが、図 3 では 1 コアを使用しています。その結果、インテル® Xeon® Platinum 8180 プロセッサーでは、1.63 倍の向上を達成しています。

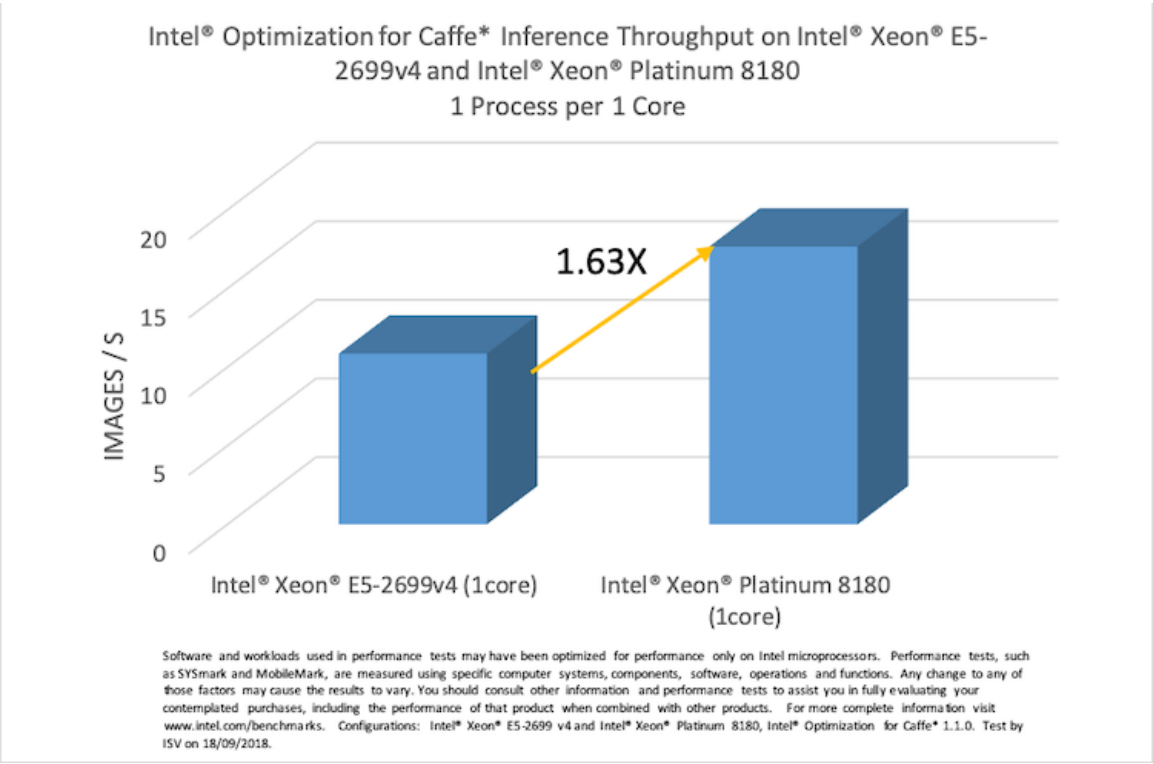


図 3. インテル® Xeon® E5-2699 v4 とインテル® Xeon® Platinum 8180 上のインテル® Optimization for Caffe* の推論スループットの比較 (1 コアを使用した場合)

図 4 は、さまざまなバッチサイズの推論スループットを比較している点で図 1 と異なります。

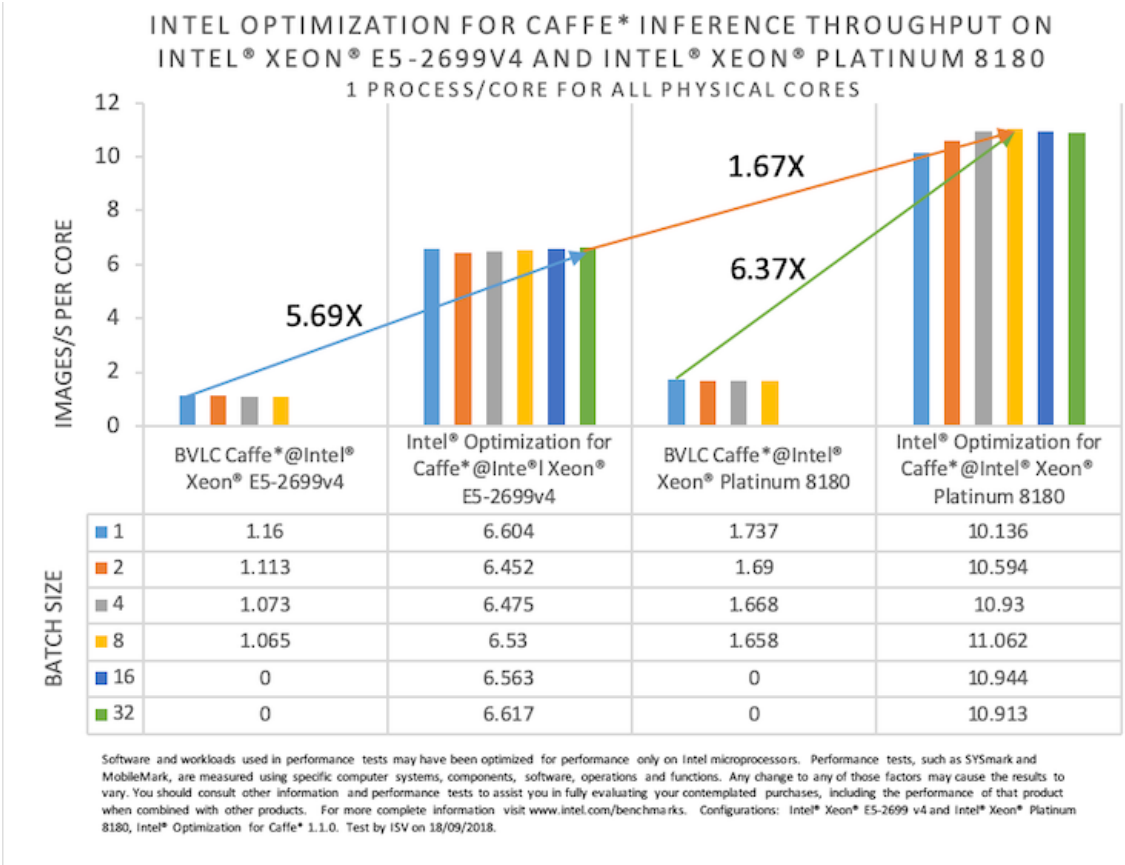


図 4. インテル® Xeon® プロセッサー E5-2699 v4 とインテル® Xeon® Platinum 8180 プロセッサー上での BVLC Caffe* とインテル® Optimization for Caffe* の推論スループットの比較 (さまざまなバッチサイズを使用)

図 5 は、インテル® Xeon® E5-2699 v4 とインテル® Xeon® Platinum 8180 上でインテル® MKL とインテル® MKL-DNN エンジンを使用した場合のインテル® Optimization for Caffe* の推論パフォーマンスの比較です。このグラフから、インテル® MKL-DNN エンジンを使用した場合のほうが、インテル® MKL エンジンを使用した場合よりも優れており、インテル® Xeon® Platinum 8180 プロセッサー上のほうがインテル® Xeon® プロセッサー E5-2699 v4 上よりもはるかに優れたパフォーマンスを達成できることが分かります。

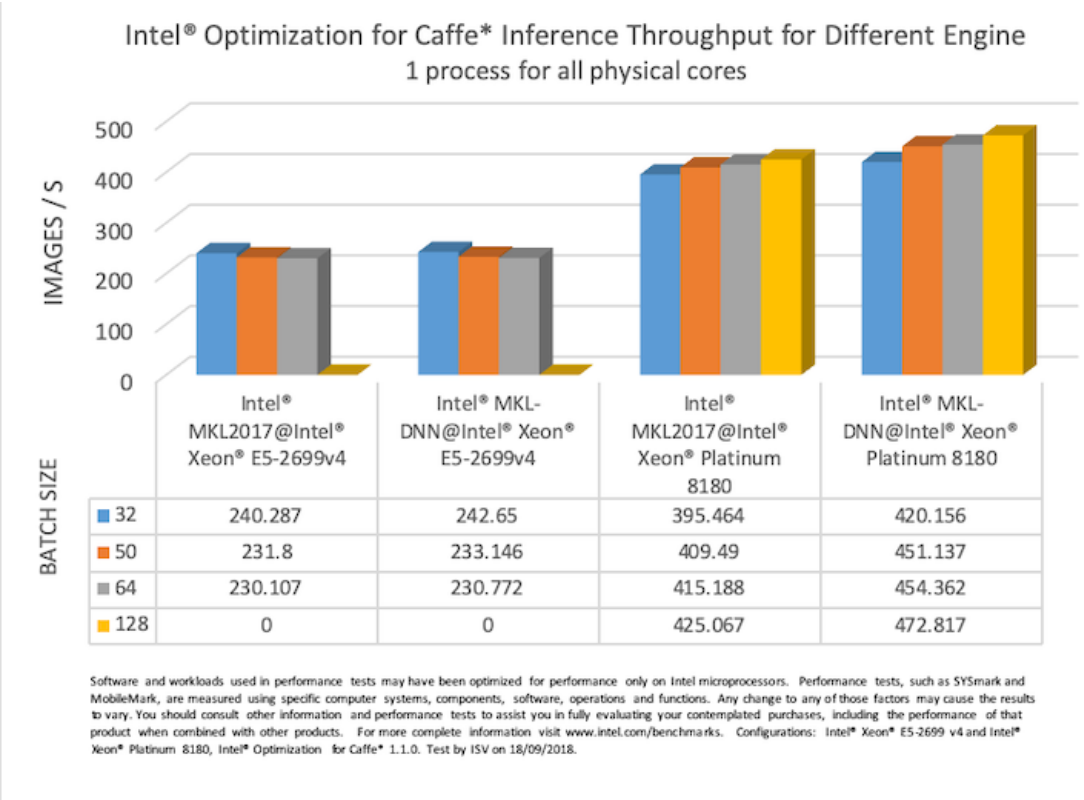


図 5. インテル® Xeon® E5-2699 v4 とインテル® Xeon® Platinum 8180 上でインテル® MKL とインテル® MKL-DNN エンジンを使用した場合のインテル® Optimization for Caffe* の推論スループットの比較 (さまざまなバッチサイズ)

まとめ

インテル® Xeon® Platinum 8180 プロセッサー上の顔認識は、すべてのコアを使用する場合インテル® Xeon® プロセッサー E5-2699 v4 よりも優れており、1 コアを使用する場合も 1.67 倍のパフォーマンス・ゲインが得られます。BVLC Caffe* と比較して、インテル® MKL-DNN を使用するインテル® Optimization for Caffe* は 6.37 倍 のパフォーマンスを達成しました。インテル® MKL-DNN エンジンの使用は、インテル® MKL エンジンを使用する場合と比較して、約 10% のパフォーマンス向上をもたらします。

関連情報

インテル® Optimization for Caffe* のインストール (英語)
BVLC Caffe* のインストール (英語)
インテル® MKL
インテル® MKL-DNN (英語)

法務上の注意書き:
性能に関するテストに使用されるソフトウェアとワークロードは、性能がインテル® マイクロプロセッサー用に最適化されていることがあります。SYSmark* や MobileMark* などの性能テストは、特定のコンピューター・システム、コンポーネント、ソフトウェア、操作、機能に基づいて行ったものです。結果はこれらの要因によって異なります。製品の購入を検討される場合は、他の製品と組み合わせた場合の本製品の性能など、ほかの情報や性能テストも参考にして、パフォーマンスを総合的に評価することをお勧めします。詳細については、www.intel.com/benchmarks (英語) を参照してください。システム構成: インテル® Xeon® プロセッサー E5-2699 v4 およびインテル® Xeon® Platinum 8180 プロセッサー、インテル® Optimization for Caffe* 1.1.0。2018 年 9 月 18 日現在の ISV による測定値。

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。