プレスリリース

2021 年 10 月 20 日

ディープラーニング推論高速化に関する研究論文の公開
～無料トライアル中『SoftNeuro』の高速化手法について技術詳細を解説～

【概要】

株式会社モルフォ（所在地：東京都千代田区、代表取締役社長：平賀督基、以下モルフォ）は、ディープラーニング推論高速化手法に関する研究論文（プレプリント版）を、論文公開サイト「arXiv」に公開しました。

本論文における研究内容は、モルフォが開発し商用化に成功した世界最速級ディープラーニング推論エンジン『SoftNeuro®』に活用されている技術です。

『SoftNeuro』は、非商用での利用 (*1) に限り 2021 年 12 月 31 日までモルフォウェブサイトにて無料で公開されており、多くの方にご利用をいただいています。ご利用と同時に『SoftNeuro』が用いている革新的な技術の詳細についてもご理解をいただくことで、より多くの方により便利にご利用いただきたいとの思いから今般の論文公開に至りました。

なお、「『SoftNeuro』無料トライアル」をお試し後、「ご利用者様アンケート」にご回答いただいた全ての方に、Amazon ギフト券 1000 円分を進呈しています。ぜひご利用ください。

- 無料トライアルサービスページ：https://softneuro.morphoinc.com

【論文公開内容について】

- 論文名称：SoftNeuro: Fast Deep Inference using Multi-platform Optimization
- 著者名：Masaki Hilaga, Yasuhiro Kuroda, Hitoshi Matsuo, Tatsuya Kawaguchi, Gabriel Ogawa, Hiroshi Miyake and Yusuke Kozawa
- 公開 URL：https://arxiv.org/abs/2110.06037

【論文のポイント】

1. DNN(ディープニューラルネットワーク)においてレイヤーとルーチンを分離
2. 動的計画法により最適なルーチンの組み合わせを探索することで、推論を高速化
3. 実験により、既存の推論エンジンよりも高速な推論が可能であることを示した

【背景および研究内容】

DNN は様々な分野で応用されています。

DNN モデルの学習は GPU などのリッチな計算資源上で行えますが、推論時にはスマートフォンや AR/VR デバイス、産業用機械など、限られた計算資源の中で「高速に」推論しなければなりません。これまでも多様な推論エンジンが公開されてきましたが、実機環境に最適な計算が行えるわけではありませんでした。

私たちが開発した『SoftNeuro』では、DNN モデル全体の最適化を行うことにより、CPU、GPU、DSP など様々な環境で高速な推論が実現できます。様々な型（float32、float16、qint8 など）やデータ形状（channel-first や channel-last）、アルゴリズム（Winograd やナイーブな手法）の中から最適なものを動的計画法により選びます。

実験により、既存の推論エンジンよりも高速な推論が可能で、tuning 速度も上回ることが示されました。

【SoftNeuro の特徴】

1. 様々な DNN フレームワークで学習されたモデルを取り込み、『SoftNeuro』用に変換可能です。具体的には TensorFlow と ONNX（これは Caffe2・Chainer・Microsoft Cognitive Toolkit・MXNet・PyTorch などから変換可能な形式）からモデルを変換できます。その際、例えば ReLU や Batch Normalization レイヤーを前のレイヤーに統合することで、高速化が図れます。

2. 変換したモデルが推論環境で高速に動くよう tuning します。その際、モデルを抽象的な計算概念であるレイヤーと、レイヤーの具体的な実装であるルーチンに分離して取り扱います。レイヤーは、計算方法を示すレイヤーパラメータと学習された値である weight から成ります。Convolution レイヤーで例えると、レイヤーパラメータは stride や padding などを指し、weight は kernel と bias を指します。一つのレイヤーに対し、具体的実装であるルーチンは複数存在し得ます。CPU・GPU・DSP などの環境、float32・qint8 などのデータ型、そして計算アルゴリズムなどの違いによって様々なルーチンが作成出来るためです。

3. tuning では、まず各ルーチンの実行時間を計測します。これを profiling と呼びます。profiling 結果を基にレイヤーの適切なルーチンを選ぶことで推論高速化が可能です。後述するアルゴリズムでルーチン選択を実行し tuning が完了します。

【推論最適化アルゴリズム】

DNN モデルは、レイヤーをノードとする有向非巡回グラフからなります。取り得るグラフ構造の中には多様な分岐パターンが含まれます（図１）。この複雑な構造の中で、レイヤーのとるルーチンの最適な組み合わせを見つけなければなりません。特にルーチン間でデータ型が変更されたり CPU から GPU へデータが転送されるような場合には、型変換やデータ転送を行う処理（adapt レイヤー）を挟まなければならず、その点も考慮する必要があります。

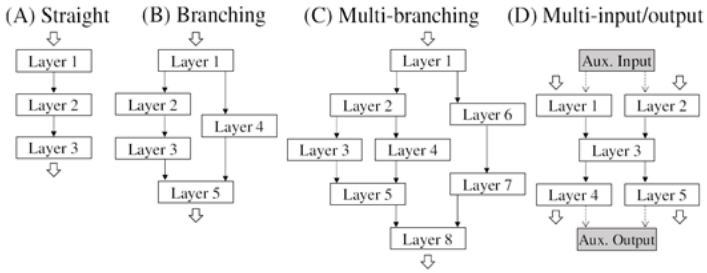


図 1

『SoftNeuro』では、レイヤーごとの取り得るルーチンと profiling 結果を基に、動的計画法のアルゴリズムを用いることで、最適なルーチン組み合わせを高速に探索します。

【実験と結果】

1. VGG16・ResNet50・MobileNetV2・MobileNetV3 それぞれのモデルについて、float32 と qint8 両方の型を利用可能な設定で tuning を行ったところ、単に float32 または qint8 のルーチンを使用した場合に比べ、推論速度が向上しました。

2. 既存の推論エンジンである Tensorflow Lite・PyTorch Mobile との比較実験を行ったところ、上述した全てのモデルに対して、『SoftNeuro』が最高速であることが示されました。（図２）（図中の単位は ms、スマートフォン(Snapdragon835)上で計測）

Network	TensorFlow Lite	PyTorch Mobile	Ours
VGG16	403.062 (±12.053)	N/A	198.453 (±3.758)
ResNet50	129.698 (±0.871)	152.172 (±0.766)	102.544 (±1.373)
MobileNetV2	25.730 (±0.639)	55.653 (±0.031)	14.931 (±0.019)
MobileNetV3	7.479 (±0.036)	N/A	5.695 (±0.066)

図 2

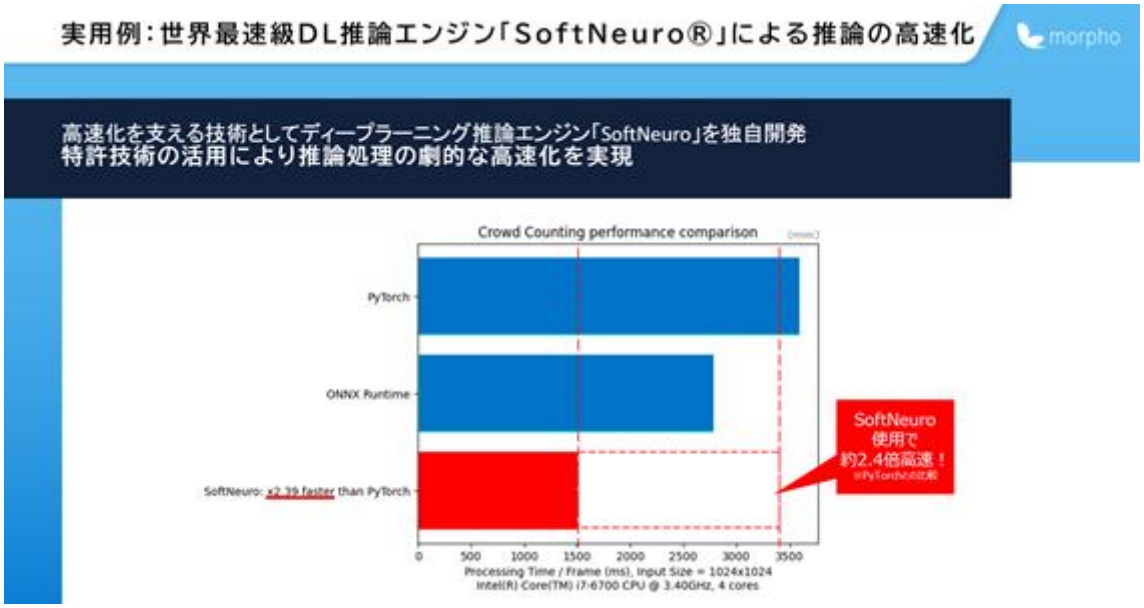
3. SoftNeuro と異なる tuning 方法を採る推論エンジン TVM との比較を行ったところ、tuning 速度および推論速度の両方で『SoftNeuro』が上回ることが示されました。

【モルフォ製品への応用例】

「群衆・混雑カウントソリューション」は、モルフォと株式会社セキュアが共同開発した、人物の密集度や混雑度を見える化する映像解析 AI です。これらを可視化することで、密集・密接を避けた形での安全なショッピングや食事の実現に活用いただくことを想定しています。

商業施設や飲食店など様々なシーンで実用化が進んでおり、混雑度合いの可視化に求められる高速で高精度な推論を、本論文でご紹介した技術が支えています。

- 群衆・混雑カウントソリューション：https://www.morphoinc.com/news/20201102-jpr-morpho_secure_ccs



*1：
非商用利用とは以下を指します。
・いかなる形でも対価の支払いを受けない。
・営利目的用途の成果物作成を目的としない。
・商業的サービスの提供を目的としない。
詳しくは利用規約（https://softneuro.morphoinc.com/terms.html）をご確認ください。

【株式会社モルフォについて】

モルフォは「画像処理／AI（人工知能）」の研究開発型企業です。高度な画像処理技術を組み込みソフトウェアとして、国内外のスマートフォン、半導体メーカを中心にグローバルに展開しています。また、カメラで捉えた画像情報をエッジデバイスやクラウドで解析する、AI を駆使した画像認識技術を車載や産業 IoT 分野へ

提供し、様々なイノベーションを先進のイメージング・テクノロジーで実現しています。

所 在 地：東京都千代田区西神田 3 丁目 8 番 1 号　千代田ファーストビル東館 12 階
代 表 者：代表取締役社長　平賀　督基（まさき）、【博士（理学）】
設　　立：2004 年 5 月 26 日
資 本 金：1,782,977 千円（2021 年 2 月 28 日現在）
事業内容：画像処理および AI（人工知能）技術の研究・製品開発。スマートフォン・半導体・車載・産業 IoT 向けソフトウェア事業をグローバルに展開。
ホームページ：https://www.morphoinc.com/
Facebook：https://www.facebook.com/morphoinc

【お問合せ先】

株式会社モルフォ　広報担当　宮崎、大野
TEL：080-8433-3415
お問い合わせフォーム：　http://www.morphoinc.com/contact