東京大学
THE UNIVERSITY OF TOKYO

記者会見　開催のお知らせ
「あなたのスマホへ最新の人工知能をお届けします
～アプリ不要の高速ディープニューラルネットワーク実行システムを開発 ～」

１．会見日時：　２０１７年１０月１７日（火）１４：００～１５：００

２．会見場所：　東京大学 本郷キャンパス 工学部２号館 ８階 81C1 号室　（別紙参照）

３．出席者：

原田 達也　（東京大学 大学院情報理工学系研究科 知能機械情報学専攻 教授）
牛久 祥孝　（東京大学 大学院情報理工学系研究科 知能機械情報学専攻 講師）
日髙 雅俊　（東京大学 大学院情報理工学系研究科 知能機械情報学専攻 博士課程３年）
木倉 悠一郎（東京大学 大学院情報理工学系研究科 創造情報学専攻 修士課程２年）

４．発表のポイント

- 市販のパソコンやスマートフォンに標準搭載されている Web ブラウザ上で、ディープニューラルネットワークを高速実行（世界最速）できるシステムを開発しました。
- 計算方法の最適化および Web の最新技術の活用により、画像認識処理において従来の約 50 倍の実行速度を達成しました。
- 画像生成等の人工知能研究の最新の成果を誰でも容易に試せるようにする基盤として有用であり、さまざまなデモンストレーションやアプリケーションへの応用が期待されます。

５．発表概要：

東京大学 大学院情報理工学系研究科 知能機械情報学専攻の原田達也教授、日髙雅俊大学院生、木倉悠一郎大学院生、牛久祥孝講師は、市販のパソコンやスマートフォンに標準搭載されている Web ブラウザ上で、ディープニューラルネットワーク(DNN)を高速（世界最速）に実行できるソフトウェアフレームワーク「WebDNN」を開発しました。

DNN は近年急速に発展を遂げている人工知能の一形態で、画像や音声の認識や生成に有効な手法です。しかしながら計算負荷が高く、Web サービスに組み込むにはサービス提供者側がユーザ数に応じた大量の計算機を用意するか、ユーザの端末に専用アプリケーションをインストールして処理を行う必要がありました。本研究では、DNN 内の冗長な計算を除去する技術および Web の最新規格を活用して端末の能力を最大限引き出す技術を開発し、Web サービスに容易に組み込めるシステムとして提供しました。WebDNN を搭載した Web サービスでは、ユーザがこれを Web ブラウザで開くだけで、端末上で DNN を高速に実行できます。これにより、人工知能の最新の研究成果を用いたデモンストレーションやアプリケーションを低コストかつ利便性高く提供することが可能となりました。

会見当日は、本研究のご紹介と共に、実演も合わせて行います。つきましては、本件について記事掲載の取材を是非お願いいたしたく、ご案内申し上げます。

６．発表内容：

東京大学 大学院情報理工学系研究科 知能機械情報学専攻の原田達也教授、日髙雅俊大学院生、木倉悠一郎大学院生、牛久祥孝講師は、市販のパソコンやスマートフォンに標準搭載されている Web ブラウザ上で、ディープニューラルネットワーク(DNN)を高速に実行できるソフトウェアフレームワーク「WebDNN」を開発しました。

本研究は、科学技術振興機構（JST）「戦略的創造研究推進事業 CREST」および文部科学省「ポスト「京」プロジェクト」の支援を受けて開発しました。

科学技術振興機構（JST）　戦略的創造研究推進事業（CREST）
研 究 領 域 名：ビッグデータ統合利活用のための次世代基盤技術の創出・体系化
（研究総括：喜連川 優）
研 究 課 題 名：膨大なマルチメディアデータの理解・要約・検索基盤の構築
研 究 代 表 者：東京大学 大学院情報理工学系研究科 教授 原田達也

文部科学省ポスト「京」開発事業
（フラッグシップ 2020 プロジェクト：FLAGSHIP2020_Project）
萌芽的課題名：萌芽的課題 4　思考を実現する神経回路機構の解明と人工知能への応用
課　題　名：脳のビッグデータ解析、全脳シミュレーションと脳型人工知能アーキテクチャ
研 究 代 表 者：沖縄科学技術大学院大学 神経計算ユニット 教授 銅谷賢治

研究の背景・目的

DNN は近年急速に発展を遂げている人工知能の一形態で、画像や音声の認識や生成に有効な手法です。しかしながら計算負荷が高く、Web サービスに組み込むにはサービス提供者側がユーザ数に応じた大量の計算機を用意することが必要となり、コストが高くなります。別の手段として、DNN の処理を行う機能を組み込んだ専用アプリケーションをユーザに配布し、端末（パソコンやスマートフォン等）上で計算を行うことも考えられます。この場合、サービス提供者のコストは下がるもののユーザは専用アプリケーションのインストールという手間がかかり、気軽に試すことが難しくなります。これらの問題を解決する手段として、Web ページの中に DNN の処理を行うソフトウェアを組み込み、Web ブラウザでこれを開いたユーザの端末上で計算処理を行わせるというアイデアが以前より提案されています。しかし、このアイデアに基づいて作られた既存のシステムは処理速度が遅く、実用的なサービスを提供することが難しいという課題がありました。

本研究ではパソコンやスマートフォンの Web ブラウザ上で、DNN を大幅に高速処理することのできるソフトウェアフレームワーク「WebDNN」を開発しました。

WebDNN を用いることにより、DNN を用いた Web サービスを低コストに提供できるようになります。ユーザはただ Web サービスの URL を開くだけで済み、手間が増えることはありません。また処理が端末内で完結することから、処理対象の写真等をサービス提供者側に送信する必要がなくなり、プライバシー保護の観点での安全性も高まります。

手法の概要

WebDNN では、DNN を Web ブラウザ上で高速に実行することを目的に、(1)実行結果が変わらない範囲での計算量の削減、(2)端末の性能を最大限引き出すための Web 最新規格の活用の 2 点の技術を開発しました。

計算量の削減の面では、例えば「2×3×変数」という計算が必要な場合、「2×3」の部分は事前に計算しておき、「6×変数」という処理に変換（最適化）できます。この例では、計算時間が短縮されるだけでなく、処理に必要なデータのダウンロード時間の短縮にもつながります。WebDNN にはこのような変換ルールが 10 以上組み込まれています。

端末の性能を引き出す面では、Apple 社のスマートフォン iPhone に標準搭載されている Web ブラウザ Safari の現バージョン（2017 年 9 月 20 日リリース）に搭載の新規格「WebGPU」の活用が挙げられます。この規格はすでに普及している「WebGL」規格に対してコンピュータグラフィックスの高速化を目指すものですが、DNN の処理でも大幅な高速化をもたらすことがわかりました。WebDNN では WebGPU 規格にいち早く対応することで、Web ブラウザ上で DNN の処理を行う他のソフトウェアと比べ大幅な高速化を達成しました。他社の Web ブラウザを利用している場合でも、「WebAssembly」と呼ばれる最新規格への対応により、同じ端末でも性能をより有効活用することが可能となりました。

なお、DNN は利用前にそのパラメータを深層学習（大量のデータを用いて調整を行う）により決定する必要がありますが、これを行うさまざまなソフトウェア（Tensorflow、Chainer 等）に対応しています。

実験の結果

Web ブラウザ上での DNN 実行ソフトウェアとして、本研究で開発した WebDNN および既存ソフトウェアの Keras.js (Chen ら、2016)の速度を比較しました。画像中の物体を認識する DNN として著名な VGG16 (Simonyan ら、2014)などを実行し、その処理時間を計測しました。結果を図 1 に示します。各ソフトウェアにおいて最速の設定をした場合で、WebDNN は他のソフトウェアと比較し約 50 倍の速度で計算を行うことができました。

今後の展望と課題

- Web ブラウザ上で DNN を高速実行可能にしたことにより、人工知能研究のデモンストレーションや、ユーザがカメラで撮影した画像を処理するサービスなどの開発が促進されることが期待されます。
- 将来的に、さらなる高速化やデータダウンロード時間の削減、利用可能な DNN の種類の増加を行いたいと考えています。

７．発表学会：

本研究の詳細は、米国加州マウンテンビューで開催されるマルチメディア系、特にマルチメディアコンテンツ解析に関する下記の国際会議にて現地時間（米国太平洋標準時）2017 年 10 月 25 日に発表します。

学会名：ACM Multimedia 2017 (http://www.acmmm.org/2017/)
論文タイトル：WebDNN: Fastest DNN Execution Framework on Web Browser
著者：Masatoshi Hidaka*, Yuichiro Kikura*, Yoshitaka Ushiku, Tatsuya Harada
*equally contributed
DOI 番号： 10.1145/3123266.3129394

８．注意事項：

日本時間１０月１７日（火）午後３時以前の公表は禁じられています。

９． 添付資料：

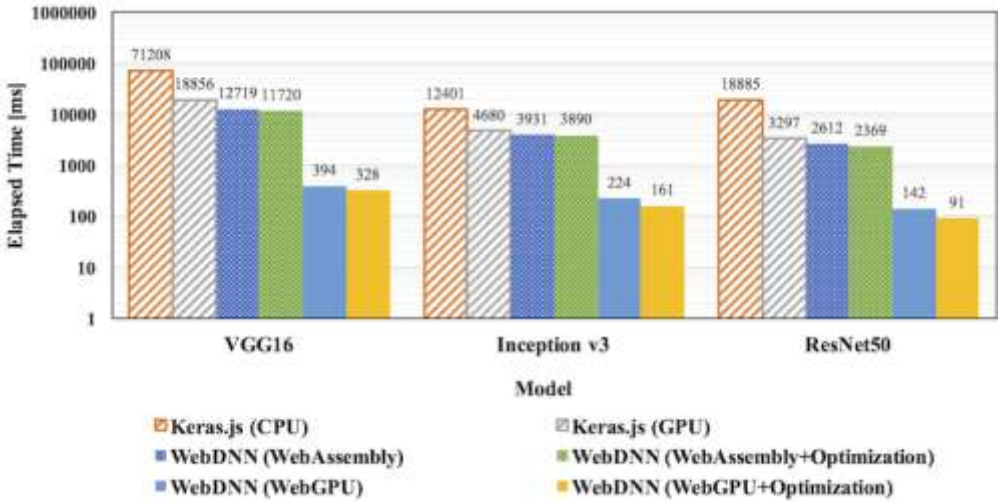

図 1 WebDNN および Keras.js における画像認識 DNN の処理時間の比較。
縦軸は処理時間であり、短い方が良い。

（別紙）

記者会見開催場所

東京大学 本郷キャンパス 工学部 2 号館 8 階 81C1 号室（〒113-8656 文京区本郷 7-3-1）