

研究終了報告書

「アーキテクチャとアルゴリズムの協調による高効率深層学習システムの創出」

研究期間： 2018 年 10 月～2022 年 3 月

研究者： 高前田 伸也

1. 研究のねらい

多種多様なセンサーを介して取得されるデータに基づいて、リアルタイムに認識・判断を行うエッジコンピューティングの需要が高まりつつある。認識・判断処理アルゴリズムの観点では、高い認識精度を持つ機械学習の技術として、深層学習は高い認識精度を達成することが可能であるが、その反面、大量の積和演算を行う必要があり、リアルタイム性および消費エネルギーの観点では負担となる。

社会システムへの深層学習の浸透に伴い、安心・安全に利活用するための対策が重要となる。例えば、学習に用いた訓練データと実際の利用環境の違いにより、導入したシステムが期待通りに認識処理ができない場合がある。これを回避するためには、運用環境それぞれに対して再学習による最適化を施すのが好ましいが、学習処理は誤差逆伝播法に基づく高負荷な計算を必要とする。限られたエネルギーでリアルタイムに推論と学習を行うには、新たなアルゴリズムと計算能力が要求される。安心・安全な認識・判断処理を行うためには、認識精度だけではなく、FATE (Fairness:公平性, Accountability:説明責任, Transparency:透明性, Ethics:倫理) を考慮したシステムが求められる。そして、その実現には、推論の途中状態を検証しながら計算することや、ベイズ推定の導入などの追加計算が必要となる。リアルタイムかつ安全な認識・判断処理を可能にするには、高い計算能力を少ないエネルギーで提供するハードウェアが更に重要となる。

本研究は、利用可能な電力およびエネルギーが制限されているエッジコンピューティング環境を主な対象として、高速かつリアルタイムに安心・安全な認識・判断処理を行うことができる機械学習システム、特に、深層学習システムを実現することを目的に、ハードウェア・アーキテクチャとアルゴリズムの協調による高精度化、高速化、高電力効率化、高信頼化技術を開発するものである。

2. 研究成果

(1)概要

高速かつリアルタイムに安心・安全な認識・判断処理を行うことができる深層学習システムの実現を目的として、アーキテクチャとアルゴリズムの両面から研究を進めた。具体的には、(1)ハードウェア効率と精度を両立する量子化ニューラルネットワークのアルゴリズムとハードウェアに関する研究、(2)ハードウェアアクセラレーションに適したベイジアンニューラルネットワークのアルゴリズムとハードウェアに関する研究、(3)高効率なニューラルネットワークのアクセラレータハードウェアの開発効率を向上するドメイン特化高位合成コンパイラに関する研究、の3つの研究を行った。

研究(1)では、まず、古典的な画像処理アルゴリズムの誤差拡散法に着想を得た、二値化ニューラルネットワークのハードウェア指向・活性化関数を中心とする Dither NN アルゴリズムとハードウェアを開発した。また、Dither NN の知見を起点として、積和演算後の値を二値化するのではなく、2つの積和値の差分を二値化することで、部分的に情報の強弱を残すこと可能な二値化ニューラルネットワークのアルゴリズムとハードウェア方式 DeltaNet を開発した。これらの二値化ニューラルネットワークのアルゴリズムとハードウェア方式により、メモリ量や回路資源の観点では優れるが、精度が低いという欠点を持つ、従来の二値化ニューラルネットワークの精度の限界を高めることができた。

研究(2)では、まず、ベイジアンニューラルネットワークに適したハードウェアアクセラレータのFPGA向けオーバーレイアーキテクチャの研究を行い、学習済みの重みの確率分布から値をサンプリングする、逆関数法とルックアップテーブルに基づく軽量アーキテクチャを開発した。これにより、従来のサンプリング方式である VIBNN と同等の精度と振る舞いを達成するサンプリング回路を、従来回路の4分の1程度のLUT数で実現できるようになった。次に、サンプリング毎の重みと活性値の変動の性質に着目した、ベイジアンニューラルネットワークの順伝播演算の近似アルゴリズムとハードウェアアーキテクチャの ASBNN を開発した。

研究(3)では、ニューラルネットワークの学習済みモデルに適したアクセラレータをFPGA上に実現する際の開発効率を高めるために、学習済みモデルからハードウェアアクセラレータのハードウェア記述を自動的に生成するドメイン特化高位合成コンパイラの NNgen を開発した。

(2) 詳細

研究(1) ハードウェア効率と精度を両立する量子化ニューラルネットワークのアルゴリズムとハードウェアに関する研究

まず、古典的な画像処理アルゴリズムの誤差拡散法に着目し、二値化ニューラルネットワークのハードウェア指向・活性化関数を中心とする Dither NN アルゴリズムとハードウェアを開発した[代表的な論文 3]。図 1 にその概要を示す。二値化ニューラルネットワークは、+1 と -1 の二値(バイナリ表現であれば 1/0)、つまり 1 ビットで活性値と重みを表現可能な方式であり、乗算が論理演算の XNOR で実現できるため、メモリ量と回路面積の観点で優れる。しかし、表現可能な数値パターンが大幅に制限されるため、浮動小数や 8 ビット程度の量子化ニューラルネットワークと比べて、認識精度が低下するという問題が存在する。そこで、二値化ニューラルネットワークに特化した計算ハードウェアにおいては、入出力はそれぞれの数値が 1 ビットで表現されるが、計算回路内部では多ビットの数値表現が用いられることに着目して、出力時に二値化により 1 ビット化される際の量子化誤差を他のニューロンに繰り越すことで、認識精度を高めることが可能になった。ハードウェアの観点では、多ビットの情報を演算器内部に留めたまま再利用するデータフローを採用することで、メモリ量と回路複雑性がほとんど増加しないことを、FPGA 実装を通じて示した。本成果は、リコンフィギュラブルハードウェアに関する難関国際会議 FPT' 18 にて発表し Best Paper Award[その他の成果 5]を受賞した。

また、Dither NN の知見を起点として、積和演算後の値を二値化するのではなく、2 つの積和値の差分を二値化することで、部分的に情報の強弱を残すこと可能な二値化ニューラルネットワークのアルゴリズムとハードウェア方式 DeltaNet を開発した[その他の成果 3, 特許出願 2]。図 2 にその概要を示す。一般の多層の畳み込みニューラルネットワークの後方では空間方向のサイズが小さくなるため、空間方向に誤差拡散する Dither NN は有効に働かないが、DeltaNet はチャンネル方向で隣接するニューロンの大小関係を 1 ビットで保存する方式であるため、多層のニューラルネットワークの前方から後方まで広く利用できる手法である。Dither NN と同様に、計算回路内部に存在する多ビットの情報表現を演算器内部に留めたまま再利用するため、回路規模の増加量は小さい。本研究の一部は、成果は難関国際会議 ASAP' 19 で発表した。また、電子情報通信学会コンピュータシステム研究会にて優秀若手講演賞を受賞した。

これらの二値化ニューラルネットワークのアルゴリズムとハードウェア方式により、メモリ量や回路資源の観点では優れるが精度が低いという欠点を持つ、従来の二値化ニューラルネットワークの精度の限界を高めることができた。

画像処理における二値化と誤差拡散法

単純二値化では色潰れによる情報損失が発生するが
誤差拡散法による二値化では空間的に中間色を表現可能



画像出典:
<https://ja.wikipedia.org/wiki/ディザ>

Dither NN: 誤差拡散法に基づく二値化NN

誤差拡散法をDNNの二値化処理に適用し高精度化

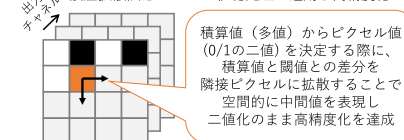


図1 低ビット精度演算に基づく
ハードウェア指向ニューラルネットワーク

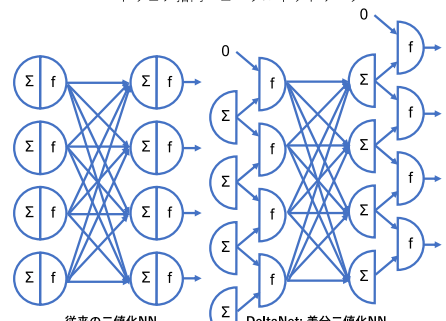


図2 差分二値化による高精度な二値化ニューラルネットワーク

研究(2)ハードウェアアクセラレーションに適したベイジアンニューラルネットワークのアルゴリズムとハードウェアに関する研究

ニューラルネットワークは一般にデータを用いて学習するものであるため、学習データとは性質の異なる入力に対しては特に誤った認識をしてしまうことがしばしばある。信頼される AI システムには、「わからないこと」がわかることが好ましい。そのための方法のひとつとして、重みや活性値に確率分布を用いるベイジアンニューラルネットワーク (Bayesian Neural Network) と呼ばれる技術がある。出力の確率分布を観察することで、その出力の確信度を知ることができるという特徴があり、確信度に基づく後続処理が可能になるという利点がある。しかし、出力の確率分布を得るためには、同一のモデルに対して、モデルの重みの確率分布から値を複数回サンプリングして、ニューラルネットワークの順伝播計算を繰り返す必要がある。そのため、一般のニューラルネットワークよりも計算コストが単純に増加するという課題が存在する。

そこで、まず、ベイジアンニューラルネットワークに適したハードウェアアクセラレータの FPGA 向けオーバーレイアーキテクチャの研究を行い、学習済みの重みの確率分布から値をサンプリングする、逆関数法とルックアップテーブルに基づく軽量アーキテクチャを開発した[代表的な論文 2]。図 5 に従来のサンプリング回路と提案サンプリング回路の構成をそれぞれ示す。ベイジアンニューラ

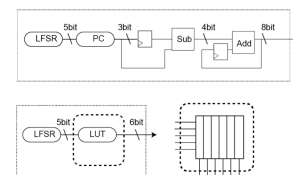


図5 逆関数法とルックアップテーブルを用いる重みサンプリング回路

ルネットワークのアクセラレータは、学習時に決定される重みの確率分布から重みの値をサンプリングする回路が必要である。そのため、積和演算と同等の並列性が求められ、回路面積および消費電力に対する影響が大きい。従来方式の VIBNN では、図 5 上部に示すように、二項分布によるガウス分布の近似を用いていたため、0/1 のカウンティング回路が必要なのに対して、図 5 下部に示す提案方式では、多ビット・並列に生成される乱数から逆関数法により、ガウス分布から直接サンプリングする方式を採用した。更に、逆関数を数ビットのルックアップテーブルで近似することで、十分高い精度の乱数をガウス分布から得ることができるようになった。提案方式を用いた場合の認識精度は従来方式の VIBNN と同等であり、重みサンプリング回路を従来の 4 分の 1 程度の LUT で実現できることを、DNN フレームワークおよび Verilog HDL による回路実装により明らかにした。本成果は国際会議 CANDAR' 19 および国際論文誌 IJNC に採択された。

次に、サンプリング毎の重みと活性値の変動の性質に着目した、ベイジアンニューラルネットワークの順伝播演算の近似アルゴリズムとハードウェアアーキテクチャの ASBNN を開発した[代表的な論文 1]。サンプリング毎に各層の出力はどの程度変化するかを図 6 のヒストグラムに示す。横軸が初回の順伝播との差分の絶対値、縦軸がその出現回数である。観察の結果、各層の出力の多くは初回の順伝播と似た値をとっており、ごく

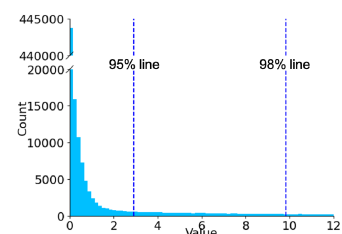


図6 ベイジアンニューラルネットワークのサンプリング毎の活性値の変化量

少数の値のみが大きく異なることがわかる。ASBNN では、この特性を活かして、各層の計算において、入力値 (= 前層の出力) が初回の順伝播と大きく異なる場所に限定して差分を計算するというアルゴリズムを採用することで、ニューラルネットワークの計算中に出現するスパース性(ゼロの割合)を増やし、計算回数を削減する。また、本アルゴリズムに対応する、差分

箇所のみを効率的に計算する、スパース構造に特化したハードウェアアーキテクチャを組み合わせることで、精度を低下することなく実際の計算コストを削減する。従来の密行列を対象とするアクセラレータおよびスパース行列に適した CNN アクセラレータ SCNN と比べて、サンプリング回数が 128 回するとき、それぞれ 3.3 倍、1.27 倍の速度向上を達成した。本成果は、難関国際会議 ASAP' 21 で発表した。

また、サンプリングフリーで高速なベイジアンニューラルネットワークの処理方式として、決定論的変分推論 (Deterministic Variational Inference, DVI) に関する研究を行い、多層の畳み込みニューラルネットワーク上で、サンプリングフリーでベイジアンニューラルネットワークを実現するためのバッチ正規化のアルゴリズムを開発した[その他の成果 2]。DVI では、それぞれの重みと活性値に適切な確率分布(主にガウス分布)を仮定し、活性値の確率分布パラメータを中心極限定理などにに基づく近似により直接求める。活性値 1 つあたりの計算は複雑化するが、1 回の計算で良いため、サンプリングによる変分推論に基づくベイジアンニューラルネットワークよりも高速化が期待される。しかし、各層では確率分布に対して計算を行うため、従来のニューラルネットワークで用いられている平均値プーリングやバッチ正規化といった演算をそのまま適用できないという課題が存在する。特にバッチ正規化は学習の発散を抑制するために、近年のニューラルネットワークで重要な演算である。

そこで、決定論的変分推論の適用可能な、確率変数に対するバッチ正規化を考案した。図 7 に確率変数を対象としたバッチ正規化の概要、図 8 に手順をそれぞれ示す。図 7 に示すように、入力ニューロン X と出力ニューロン Y がガウス分布に従うと仮定する。このとき、提案するバッチ正規化では、入力ニューロンの確率変数パラメータである平均と分散について、バッチ方向でそれぞれの統計量(つまり平均の平均と平均の分散、分散の平均、分散の分散)をもとに、確率変数パラメータの平均と分散を正規化する。これにより、確率変数パラメータ、特に分散の発散を抑制することが可能になる。実験の結果、バッチ正規化を持たない DVI、サンプリングによる変分推論に基づくベイジアンニューラルネットワーク、ドロップアウトに基づくベイジアンニューラルネットワークと比較して、提案手法を適用する DVI は高い精度を保ちながら高い確度で不確実性を表現できることを示した。本研究では、人工知能学会研究会優秀賞を受賞した。

研究(3) 高効率なニューラルネットワークのアクセラレータハードウェアの開発効率を向上するドメイン特化高位合成コンパイラに関する研究

ニューラルネットワークの学習済みモデルに適したアクセラレータを FPGA 上に実現する際、開発効率を高めるために、学習済みモデルからハードウェアアクセラレータのハードウェア記述を自動的に生成するドメイン特化高位合成コンパイラの NNgen (<https://github.com/NNgen/nngen>) を開発した[その他の成果 4, 6, 7]。図 8 にコンパイラアーキテクチャを示す。NNgen では、一般的なディープラーニングフレームワークで学習されたモデルの計算グラフを入力として、並列化や量子化等の最適化を指定するだけで、FPGA ベンダーに依存しないハードウェアアクセラレータが生成できる。モデル毎に異なる回路を生成す

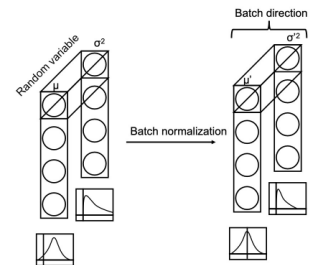


図7 決定論的変分推論に基づくベイジアンNN向けバッチ正規化

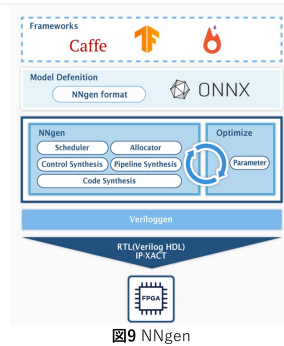
Input: $X = \{x_1 \sim N(\mu_1, \sigma_1^2), \dots, x_B \sim N(\mu_B, \sigma_B^2)\}$
Output: $Y = \{y_1 \sim N(\mu'_1, \sigma_1'^2), \dots, y_B \sim N(\mu'_B, \sigma_B'^2)\}$

```
//normalize  $\mu$ 
 $E[\mu] \leftarrow \frac{1}{B} \sum_{b=1}^B \mu_b$ 
 $V[\mu] \leftarrow \frac{1}{B} \sum_{b=1}^B (\mu_b - E[\mu])^2$ 
 $E[y'_i] \leftarrow \frac{\mu_i - E[\mu]}{\sqrt{V[\mu]}}$ 
//normalize  $\sigma^2$ 
 $E[\sigma^2] \leftarrow \frac{1}{B} \sum_{b=1}^B \sigma_b^2$ 
 $V[\sigma^2] \leftarrow \frac{1}{B} \sum_{b=1}^B (\sigma_b^2 - E[\sigma^2])^2$ 
 $V[y'_i] \leftarrow \frac{\sigma_i^2}{\sqrt{V[\sigma^2]}}$ 
//linear transform by trainable variables  $\gamma, \beta$ 
 $y_i = \gamma y'_i + \beta$ 
 $E[y_i] = \gamma E[y'_i] + \beta$ 
 $V[y_i] = \gamma^2 E[y'_i]$ 
```

図8 決定論的変分推論に基づくベイジアンNN向けバッチ正規化アルゴリズム

る点が、他の FPGA 向け DNN ハードウェアコンパイラとは異なる。NNgen のソースコードは GitHub にて Apache 2.0 ライセンスの下オープンソースソフトウェアとして公開されており、184 のスターを得ている。

上記の通り、本研究では、高速に安心・安全な認識・判断処理を行うことができる深層学習システムの実現に資する、ハードウェア・アーキテクチャとアルゴリズムの協調による高精度化、高速化、高電力効率化、高信頼化技術を開発し、当初の研究目的は達成できたと考えられる。



3. 今後の展開

二値化ニューラルネットワークに関する研究成果は、二値化ニューラルネットワークの限界を引き上げるものであるが、浮動小数や多ビット量子化ニューラルネットワークとの精度の差が依然として存在する。今後数年はハードウェア効率と認識精度を両立するための基礎的研究を進めていく必要がある。

ベイジアンニューラルネットワークに関する研究では、サンプリングに基づく方式はアルゴリズムとハードウェアアーキテクチャの両面で十分な成果が得られた。今後 2, 3 年で、信頼性が求められる用途に適用するための要件を明らかにし、実際の挙動に基づくアルゴリズムとハードウェアの拡張を行う必要がある。

ニューラルネットワーク向け高位合成コンパイラに関する研究では、NNgen は既にオープンソースソフトウェアとして公開されており、実際の製品においても利用されている。また、FPGA 入門者向けの有償セミナーの課題としても利用されている。これらのことから、十分に社会実装されていると言える。

本さがけ研究では、単一デバイス上の深層学習の推論処理にフォーカスを当て研究を行った。一方、機械学習システムは多数のクライアントやサーバーで構成される、並列分散システムであり、プライバシー保護および倫理的観点から、クライアントからサーバーへ生のデータを送信することが困難なケースが存在する。そこで、センシティブなデータをサーバーで集めることなく、大量のデータによる学習を可能にする技術として連合学習（Federated Learning）が注目されている。今後は、信頼される分散機械学習基盤・連合学習技術の創出を目指して、本さがけ研究の成果を活用して CREST 研究「D3-AI: 多様性と環境変化に寄り添う分散機械学習基盤の創出」を進める。

4. 自己評価

高速に安心・安全な認識・判断処理を行うことができる深層学習システムの実現に資する、ハードウェア・アーキテクチャとアルゴリズムの協調による高精度化、高速化、高電力効率化、高信頼化技術を開発し、当初の研究目的は十分に達成できたと考えられる。しかし、二値化ニューラルネットワークに関する成果、およびベイジアンニューラルネットワークに関する成果を、開発時間の都合により、ニューラルネットワーク向け高位合成コンパイラの成果物である NNgen に反映させることができなかった点が残念である。これらのテーマに関しては今後も研究を継続し、NNgen を通じて新規アルゴリズムとハードウェアアーキテクチャの利用促進を進

める。

本研究は、北海道大学と東京大学の修士課程・博士課程の学生と共に推進した。学生・若手研究者の待遇改善が重要であると考え、本研究費により技術補佐員やRAの形で積極的に雇用した。また、効率的な研究遂行のために、本研究費によりFPGA評価プラットフォームやGPU計算機を導入した。これらの研究費の執行は適正であったと考える。

社会への波及効果については、ニューラルネットワークのための高位合成コンパイラのNNgenを開発し、オープンソースソフトウェアとして公開の上、実際の製品や入門者向けセミナーに利用されていることから、本研究成果は社会に対して十分に影響を与えたと考えている。

5. 主な研究成果リスト

(1) 代表的な論文(原著論文)発表

研究期間累積件数:12件

1. Yoshiki Fujiwara and Shinya Takamaeda-Yamazaki: ASBNN: Acceleration of Bayesian Convolutional Neural Networks by Algorithm-hardware Co-design, The 32nd IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP 2021), July 2021.

サンプリング毎の重みと活性値の変動の性質に着目した、ベイジアンニューラルネットワークの順伝播演算の近似アルゴリズムとハードウェアアーキテクチャのASBNNを開発した。多くの活性値は重みのサンプリングによって大きく変動することなく、極一部の活性値のみが大きく変動すること、などの性質を活用する近似アルゴリズムと、そのアルゴリズムを効率的に処理するハードウェアアーキテクチャを開発した。評価の結果、従来のスパース指向アーキテクチャよりも高い性能を達成することを示した。

2. Yuki Hirayama, Tetsuya Asai, Masato Motomura, and Shinya Takamaeda-Yamazaki: A Hardware-efficient Weight Sampling Circuit for Bayesian Neural Networks, International Journal of Networking and Computing, Vol.10, No.2, pp.84-93, July 2020.

ベイジアンニューラルネットワークの計算を行うFPGA向けオーバーレイアーキテクチャを対象として、学習済みの重みの確率分布から値をサンプリングする、逆関数法とルックアップテーブルに基づく軽量アーキテクチャを開発した。多ビット・並列に生成される乱数から逆関数法により、ガウス分布から直接サンプリングする方式を採用した。更に、逆関数を数ビットのルックアップテーブルで近似することで、十分高い精度の乱数をガウス分布から得ることができるようになった。FPGA実装の結果、重みサンプリング回路を従来の4分の1程度のLUTで実現できることを示した。

3. Kota Ando, Yuka Oba, Kazutoshi Hirose, Ryota Uematsu, Takumi Kudo, Masayuki Ikebe, Tetsuya Asai, Shinya Takamaeda-Yamazaki, and Masato Motomura: Dither NN: An Accurate Neural Network with Dithering for Low Bit-Precision Hardware (Best Paper Award), The 2018 International Conference on Field-Programmable Technology (FPT'18), Tenbusu-Naha Hall, Naha, Japan, December 2018.

古典的な画像処理アルゴリズムの誤差拡散法に着想を得た、二値化ニューラルネットワークのハードウェア指向・活性化関数を中心とするDither NNアルゴリズムとハードウェアを開発

した。入出力はそれぞれの数値が 1 ビットで表現されるが、計算回路内部では多ビットの数値表現が用いられることに着目して、出力時に二値化により 1 ビット化される際の量子化誤差を他のニューロンに繰り越すことで、認識精度を高めることが可能になった。

(2) 特許出願

研究期間全出願件数: 2 件(特許公開前のものも含む)

1	発 明 者	高前田 伸也、他
	発 明 の 名 称	非公開
	出 願 人	
	出 願 日	
	出 願 番 号	
	概 要	
2	発 明 者	高前田 伸也、他
	発 明 の 名 称	非公開
	出 願 人	
	出 願 日	
	出 願 番 号	
	概 要	

(3) その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

1. Nobuho Hashimoto and Shinya Takamaeda-Yamazaki: An FPGA-Based Fully Pipelined Bilateral Grid for Real-Time Image Denoising, 31st International Conference on Field-Programmable Logic and Applications (FPL 2021), August 2021.
2. 平山 侑樹, 浅井 哲也, 本村 真人, 高前田 伸也: 人工知能学会研究会優秀賞, 決定論的変分推論に基づくベジアン CNN の検討, 人工知能学会 第 111 回人工知能基本問題研究会 (SIG-FPAI), 於 下呂市民会館, 2020 年 1 月 29 日発表, January 2020.
3. Yuka Oba, Kota Ando, Tetsuya Asai, Masato Motomura, and Shinya Takamaeda-Yamazaki: DeltaNet: Differential Binary Neural Network, IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP 2019), Cornell Tech, New York, USA, p.39, July 2019.
4. 高前田 伸也, 植松 瞭太, 藤澤 慎也, 藤崎 修一, 本村 真人: 電子情報通信学会リコンフィギャラブルシステム研究会優秀講演賞, ディープニューラルネットワーク向け拡張可能な高位合成コンパイラの開発, 電子情報通信学会研究会報告 RECONF2018-19, 於 LINE Fukuoka, 2018 年 9 月 17 日発表, September 2018 (2019 年 1 月 30 日受賞).
5. Kota Ando, Yuka Oba, Kazutoshi Hirose, Ryota Uematsu, Takumi Kudo, Masayuki Ikebe, Tetsuya Asai, Shinya Takamaeda-Yamazaki, and Masato Motomura: Best Paper Award, Dither NN: An Accurate Neural Network with Dithering for Low Bit-Precision Hardware, The 2018 International Conference on Field-Programmable Technology (FPT'18), December 2018.

6. オープンソースソフトウェア: NNgen/nngen: NNgen: A Fully-Customizable Hardware Synthesis Compiler for Deep Neural Network, <https://github.com/NNgen/nngen>
7. プレスリリース: コニカミノルタ、東京大学・高前田准教授と共同で画像 IoT、AI を実現するディープラーニング向けハードウェアコンパイラを開発、オープンソース化:
<https://www.konicaminolta.com/jp-ja/newsroom/2019/1121-02-01.html>, 2019 年 11 月 21 日