

AI アクセラレータに向けたニューラルネットワーク

Neural network for AI accelerator

○相田息吹, 上村公紀, 水品圭汰, 竹内健 (中大理工)

◦Ibuki Aida, Koki Kamimura, Keita Mizushima, and Ken Takeuchi (Chuo Univ.)

E-mail: aida@takeuchi-lab.org

1. はじめに

近年、IoT(Internet of Things)の発展に伴い、ビッグデータを高速に解析する必要がある。しかし、半導体性能の向上を支えてきたムーアの法則が終焉を迎えているため従来のノイマン型コンピューティングでは処理速度に限界がある。そこで AI を用いた解析が注目されており、AI 専用のアクセラレータの開発が各所で行われている。その中でも特に不揮発性メモリを用いた AI アクセラレータ[1]への関心が高まっている。不揮発性メモリに演算機能を持たせることでより高速な処理が期待される。本論文では不揮発性メモリを用いた AI アクセラレータへの実装に向けたニューラルネットワークを提案する。

2. ニューラルネットワーク

ニューラルネットワーク[2]とは脳の神経細胞(ニューロン)による信号伝達をモデル化したものである。Fig.1 に推論時のニューラルネットワークの概要を示す。はじめに中間層のニューロンに対して複数の入力が存在し、それぞれの入力に対し重みが設定されておりその積をとった値をすべて足し合わせる。その総和が閾値に達した場合、出力ニューロンに信号を伝達する。閾値に達したかの判断には活性化関数を用いる。出力ニューロンでも同様に中間ニューロンと対応している重みの積がすべて足されるが、出力には恒等関数か Softmax 関数が使われる。出力で使われる関数はニューラルネットワークのアプリケーションによって異なり回帰問題の場合は恒等関数、分類問題の場合は Softmax 関数がそれぞれ使われる。本研究では画像認識を想定しているため分類問題つまり Softmax 関数を用いる。

3. ニューラルネットワークの学習

ニューラルネットワークの特徴として、重みパラメータの値をデータから自動で決定できるということがあげられる。多くの学習データを用いて順方向、逆方向の計算をし、計算結果から重みパラメータを更新する。これを繰り返す、分類問題を解くのに最適な重みに近づけていく。その際は重みパラメータの勾配を利用して勾配方向に重みの値を更新する作業を行う。数値微分による計算で勾配を求めるのは計算に時間がかかるため勾配の計算を効率よく計算を行える逆誤差伝播法が広く使われている。一般的な逆誤差伝播法の構成を Fig.2 に示す。逆誤差伝播法では現在の重みパラメータの計算結果が正解データとどれだけ一致していないかを示す損失関数というものを導入し、それ

を指標として損失関数の値が小さくなるように重みパラメータを更新していく。

4. AI アクセラレータ向けニューラルネットワーク

本研究では、不揮発性メモリを用いた AI アクセラレータへの実装に向けたニューラルネットワークを提案する。また、画像認識を用いた評価を行う。

参考文献

- [1] M. Prezioso, et al., "Modeling and Implementation of Firing-Rate Neuromorphic-Network classifiers with Bilayer PT/Al₂O₃/TiO₂-x/Pt Memristors", in *IEEE International Electron Devices Meeting (IEDM)*, Dec. 2015, pp. 455-458.
- [2] Warren S. McCulloch and Walter Pitts "Neurocomputing: Foundations of Research", chapter Logical Calculus of the Ideas Immanent in Nervous Activity, pp. 15-27, MIT Press, Cambridge, MA, USA, 1998.

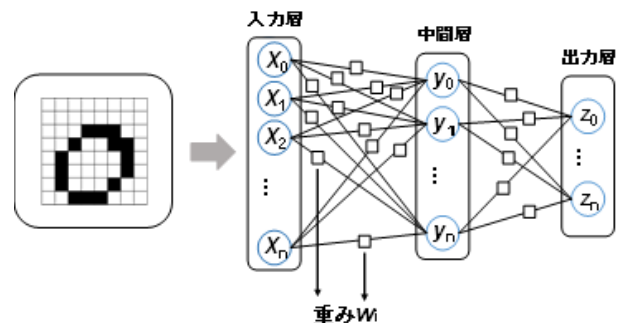


Fig.1 Neural network

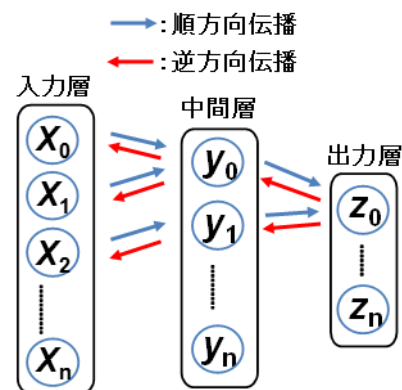


Fig.2 Back propagation