東北大学
TOHOKU UNIVERSITY

Press Release

2024 年 12 月 10 日

報道機関　各位

国立大学法人東北大学

生成 AI をスピントロニクスで省エネに
― ガウス乱数を出力する「ガウシアン確率ビット」を実現 ―

【発表のポイント】

- 生成 AI における画像や文章の生成過程を省エネ化するスピントロニクス(注 1)技術を開発しました。
- 室温で確率的に動作するスピントロニクス素子を用いて、ガウス乱数(注 2)を出力する「ガウシアン確率ビット」を実現し、動作を実証しました。
- 現行の半導体技術でガウス乱数を生成する場合と比べて回路面積は約 1/3000、消費エネルギーは約 1/150 に低減できます。

【概要】

昨今、対話式で画像や文章を生成する生成 AI が急速に普及しています。生成 AI は圧倒的な利便性を有する反面、その普及に伴って情報技術が消費する電力の増大が深刻な課題となりつつあります。生成 AI にて多大な電力を消費する過程の一つに、「拡散モデル(注 3)」を用いて行われる画像や文章の生成があり、ここに大量のガウス乱数が用いられています。このガウス乱数の AI 利用には、2024 年ノーベル物理学賞受賞者であるカナダ・トロント大学名誉教授の Geoffrey Hinton 氏らも近年注力しています。

東北大学電気通信研究所の深見俊輔教授らは、米国カリフォルニア大学サンタバーバラ校の Kerem Camsari 博士らと共同で、優れた省エネ性と小型性を兼ね備えたガウス乱数生成の新技術を開発しました。鍵となったのは、自然の熱で確率的に動作するスピントロニクス素子です。同グループは以前からこの素子を用いて二値乱数(注 2)を出力する「確率ビット(注 4)」とそれを用いた「確率論的コンピューター(注 4)」の開発を先導していました。今回同グループはこの「確率ビット」を組み合わせてガウス乱数を生成する「ガウシアン確率ビット」を実現し、連続変数を用いた組合せ最適化などの原理実証に成功しました。

今回開発された「ガウシアン確率ビット」は、消費電力の増大が重要課題となっている生成 AI の省エネ化に貢献しうるものです。今後スピントロニクス素子とそれを用いた回路、アルゴリズムの開発が進展することで、利便性と省エネ性を両立する AI 社会実現の切り札となっていくことが期待されます。

本成果は 2024 年 12 月 7-11 日（米国時間）に米サンフランシスコで開催される学術会議「International Electron Devices Meeting: IEDM」で発表されます。

研究

【詳細な説明】

研究の背景

コンピューターとの対話を通して文章や画像、動画を生成できる生成 AI が急速に普及しています。生成 AI はユーザーに圧倒的な利便性を提供する反面、その裏で従来とは桁違いの電力が消費されています。実際に、生成 AI の普及に伴って情報技術が消費する電力の増大が顕在化しつつあり、このまま行くと 2030 年までには人類が発電できる総電力の約 10%を情報機器が消費するとの予測もされています。このことから利便性を維持しつつ、省エネ動作が可能な生成 AI のためのハードウェアレベルでの革新技術の導入が重要な課題となっています。

現在の生成 AI では、「拡散モデル」に立脚したデータの学習や生成が行われており、その過程で大量のガウス乱数が用いられ、ここで多くの電力が消費されています。すなわちガウス乱数を高効率で生成できれば生成 AI の省エネ化が可能と見込まれます。なおガウス乱数は拡散モデル以外でも AI 技術全般において重要性が高く、2024 年のノーベル物理学賞が授与された AI の父とも形容される Geoffrey Hinton 氏らは、ガウス乱数を用いた機械学習法である「ガウシアン・ベルヌーイ・ボルツマンマシン」に関する研究を近年精力的に展開しています。

今回の研究を行った東北大学とカリフォルニア大学サンタバーバラ校の共同研究チームは、以前から自然の熱で確率的に状態が更新されるスピントロニクス素子を用いた効率的な乱数生成とそのコンピューティング応用に取り組んでいました。この乱数生成ユニットは「確率ビット（Probabilistic bit; p-bit）と名付けられています。これまで示されていた確率ビットは 0 または 1 の二値乱数(注 2)を出力するものでした。一方で前述のように生成 AI では連続的な値を取り、かつその統計がガウス分布に従う「ガウシアン確率ビット（Gaussian probabilistic bit; g-bit）」が必要であり、そのための新たな技術開発が求められていました。

今回の取り組み

今回、東北大学の深見俊輔教授、金井駿准教授、大野英男教授らは、米国カリフォルニア大学サンタバーバラ校の Kerem Camsari 博士らと共同で、これまで研究開発を進めてきた二値乱数を出力する確率ビット（以下、バイナリー確率ビット）を複数用いることでガウシアン確率ビットを実現できることを示しました。

図 1 にその模式図(a)、実機の写真(b)、及び出力信号の測定結果の一例(c)が示されています。今回の研究では 5 つのバイナリー確率ビットを相互作用させることで 1 つのガウシアン確率ビットを構成しています。それぞれのバイナリー確率ビットは確率動作するスピントロニクス素子を 1 つ有し、二値乱数を出力します。これが 5 つ組み合わさることで 32 (= 2^5) の状態を出力できます。バイナリー確率ビット間の相互作用はプログラマブル半導体回路 FPGA (注 5) によって付与されています。図 1(c)の測定結果から分かるように、出力されるガウス

乱数の平均値、標準偏差を、バイナリー確率ビット間の相互作用を変えることで任意に調整できています。また複数のガウシアン確率ビット間に相互作用を印加し、互いの出力信号分布に相関を持たせられる（ガウス分布のピークを近づける/遠ざける、など）ことも確認しました。

本研究ではここからさらに発展させ、2 つのガウシアン確率ビットと 2 つのバイナリー確率ビットを用いた混合整数問題の実験や、5 つのガウシアン確率ビットと 20 個のバイナリー確率ビットを用いたナップザック問題のシミュレーションなどを行いました。

加えて、研究チームは現在利用されている半導体集積回路と、今回開発したスピントロニクス素子からなるガウシアン確率ビットを用いてシステムを構築した場合に必要となるトランジスタ数と消費電力を見積もりました。その結果、今回の技術を用いることで、必要なトランジスタ数（回路の面積を決定）は約 1/3000 に、演算に要するエネルギーは約 1/150 に低減できることが分かりました。

今後の展開

本研究で得られた成果は、省エネ化が重要課題となっている生成 AI において、多くの電力を消費しながら行っている拡散モデルの省エネ処理を可能とするものです。それに加え、ノーベル賞受賞者の Hinton 氏も着目するガウシアン・ベルヌーイ・ボルツマンマシンなどへの展開も期待されます。

今回の研究はコンセプトの実証に重きが置かれていますが、研究チームは今後ある程度の規模で性能も含めてその効果を検証することを計画しています。また同時進行で、スピントロニクス素子自体のより一層の高性能化、高信頼化や、回路・アルゴリズムの研究開発にも取り組む必要があります。これらの課題に一つ一つ取り組むことで、利便性と省エネ性を両立する生成 AI が実現され、持続可能社会の担い手の一つとして社会実装されていくことが期待されます。

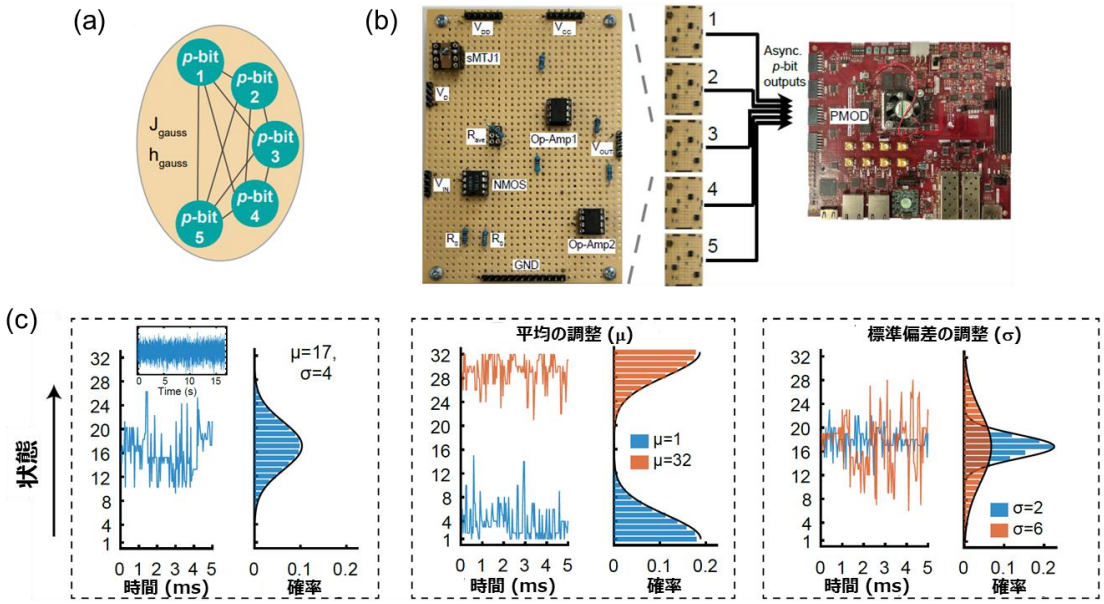

図 1. ガウシアン確率ビットの構成と測定結果。(a) 5 つのバイナリー確率ビットを用いてガウシアン確率ビットを構成する方法の模式図。バイナリー確率ビットを相互作用させ、かつそれぞれにバイアスを導入することで任意の平均値（μ）と標準偏差（σ）を持ったガウス乱数を出力するガウシアン確率ビットを構築できる。(b) 作製したガウシアン確率ビットの実機の写真。確率動作スピントロニクス素子を含むバイナリー確率ビット 5 つを FPGA で相互作用させることで実現。(c) 平均値（μ）と標準偏差（σ）の異なるガウス乱数出力の測定結果。各パネルの左側は時間領域での出力信号であり、右側は時間平均した際の出力信号の確率である。

【謝辞】

本研究は科学技術振興機構（JST）AdCORP（JPMJKB2305）、ASPIRE（JPMJAP2322）、CREST（JPMJCR19K3）、PRESTO（JPMJPR21B2）、文部科学省次世代 X-NICS 半導体創生拠点形成事業（JPJ011438）などの支援の下で行われました。

【用語説明】

注1.　スピントロニクス

物質中の電子が持つ、電気的な性質（電荷）と磁気的な性質（スピン）が協調することによって発現する現象を理解し、工学的な応用を目指す学問分野。特に、磁性体のスピンの向き（上・下）を情報（0,1）の担い手として制御する、磁気抵抗ランダムアクセスメモリ(MRAM)や磁気センサー等への応用が代表的。

注2.　ガウス乱数、二値乱数

一切の法則性を持たず、完全にランダムに並んだ数の列を乱数列と言う。二値乱数とは出力される値が 0 か 1 のいずれかである乱数であり、対してガウス乱数は出力される値が連続値であり、その分布がガウス分布（正規分布）に従う乱数。

注3.　拡散モデル

ノイズから徐々に構造化されたデータを生成する確率モデルであり、主に画像生成などに用いられる AI モデル。大きな特徴は、データにノイズを加えて崩し、それを逆に再構築するプロセスを学習する点にある。前進過程（ノイズの追加）により元のデータを完全にランダムなノイズに変換する過程と、逆拡散過程（ノイズの除去）によりノイズを取り除いて元のデータに近づける過程からなる。生成されるデータに多様性が生じ、また高品質でリアルなデータを生成することができる。

注4.　確率ビット、確率論的コンピューター

確率ビット（P ビット）とは、短時間で 0 と 1 の信号を確率的に出力し、かつ各ビットを電気的に相関させられる情報処理の基本単位。確率論的コンピューターは確率ビットを用いて演算を行うコンピューター。

確率ビットは 0 と 1 の重ね合わせ状態を持ち、かつビット間でもつれあい（相関状態）を形成できる量子ビットとは本質的に異なるが一定の類似性があり、確率論的コンピューターは量子コンピューターと並んで新概念コンピューターの一つとして注目されている。1981 年にリチャード・ファインマンが行った講演において、量子コンピューターと並んで、確率的な現象を効率的に計算する仕組みとして紹介されている。

注5.　FPGA（Field Programmable Gate Array; プログラマブル半導体回路）

ユーザーが現場（Field）で論理回路の機能をプログラムできる論理集積回路。パソコンの頭脳である CPU（Central Processing Unit）と比べると、汎用性では劣るものの、論理回路の構成自体を変えられることから、ユーザーがプログラムした計算を行う速度は速くなる。一方、ASIC（Application Specific Integrated Circuit）と比べると、速度では劣るものの、ユーザーが機能を書き換えられ汎用性が高いという利点がある。

【論文情報】

タイトル：“Beyond Ising: Mixed Continuous Optimization with Gaussian Probabilistic Bits using Stochastic MTJs”（イジングマシンのその先へ：確率動作磁気トンネル接合で構成されるガウシアン確率ビットによる連続変数の組合せ最適化）

著者： Nihal Sanjay Singh, Corentin Delacour, Shaila Niazi, Kemal Selcuk, Daniel Golenchenko, Haruna Kaneko, Shun Kanai, Hideo Ohno, Shunsuke Fukami and Kerem Y. Camsari
国際会議：70th Annual IEEE International Electron Devices Meeting (IEDM 2024)
DOI：1 **月頃に付与予定**
URL: https://iedm24.mapyourshow.com/8_0/sessions/session-details.cfm?ScheduleID=418

【問い合わせ先】
（研究に関すること）
東北大学電気通信研究所
教授　深見 俊輔
TEL: 022-217-5555
Email: s-fukami@tohoku.ac.jp
（兼）東北大学大学院工学研究科電子工学専攻
（兼）東北大学先端スピントロニクス研究開発センター (CSIS)
（兼）東北大学国際集積エレクトロニクス研究開発センター (CIES)
（兼）東北大学材料科学高等研究所 (WPI-AIMR)
（兼）公益財団法人稲盛科学研究機構 (InaRIS)

（報道に関すること）
東北大学電気通信研究所 総務係
TEL: 022-217-5420
Email: riec-somu@grp.tohoku.ac.jp