フォトニック集積はどのようにAIを加速できるか

ジェフ・ヘクト

AIは自動運転車などのポテンシャルの期待に応えるため、大幅に速度を上げる必要がある。フォトニック集積はその溝を埋める可能性がある。

古参の人たちは光コンピューティングのすべてが新しいわけではないことを知っている。合成開口レーダーデータのフーリエ変換によって軍事地図を生成する光学処理は、機密扱いとして1950年代に成功した。しかし最終的に成功を収めたのは電子高速フーリエ変換だった。1980年代、故ジョン・コーフィールド氏（John Caulfield）は、新世代の光コンピューティングが、我々が想像もしなかったことをどのように実行できるかを教えてくれたが、つじつまを合わせることはできなかった。またしても、電子工学はより速く、より良く、より安価であることが証明された。半導体の機構がナノメートルスケールにまで縮小し、ムーアの法則が限界を迎えている今、次世代の集積フォトニクスは、電子工学が提供できる以上に人工知能（Artificial Intelligence：AI）の速度と処理能力を高める可能性がある。

AIの探求

AIには、実現が困難であることが証明された独自の長い見通しの歴史がある。AIのルーツは、文学上のサイエンス・フィクションに由来する。アイザック・アシモフの有名なロボットの話は、若き教授として1960年代初頭にMIT人工知能研究所を立ち上げた、マービン・ミンスキー氏（Marvin Minsky）に影響を与えた。光コンピューティングと同様に、AIは軌道に乗るのが遅く、1980年代から1990年代初頭にかけて「AIの冬」に見舞われた。

その復活は、人間の脳のようにプロセッサ間に大規模の相互接続を持つニューラルネットワークを使用し、大量の情報を収集及び分析するようコンピュータをプログラムする機械学習からもたらされた。機械学習システムは、スパムメールのフィルタリングから「Netflix（ネットフリックス）」における映画の推薦まで、さまざまなアプリケーションを見つけたが、最も有名な用途は、碁やチェスなどの複雑なゲームで人間を打ち負かすことだ。

深層学習は、より複雑なニューラルネットワークを使用して、音声認識や自動運転などのより複雑なタスクに取り組むことにより、機械学習を拡張する。ただし深層学習では、行列ベクトル操作などの複雑なプロセスを使用して大量の情報を処理する必要があり、電子コンピュータは急速に増加する需要に対応できない。開発者たちは、必要とされるスピードとパワーの増大を提供できるシリコンフォトニクスを探している。

「異質の」知能

ニューラルネットワークは機械学習の設計に影響を与えたものの、AIは人間の頭脳のように機能することはない。機械学習では、「コア計算アルゴリズムはプログラマーによって完全に提供されるわけではないが、経験を通じてコンピュータシステムにより自動的に改善または生成される」と現在、英ケンブリッジ大（University of Cambridge）に所属するチーシャン・チェン氏（Qixiang Cheng）はレビュー記事で述べている[1]。機械学習システムは、データのパターンを認識するように設計された特別なアルゴリズムを使用して入力情報を処理することにより学習する。多くの場合、画像が出発点になるが、システムは他の種類のデータも分析する。

AIは人とは異質の知性だと思うかもしれないが、異なるスキルを持っている。AIは、何度もプレイしてルールが明確に定義されたゲームで見事に機能し、我々が非常に知的であると考える人間のチャンピオンを打ち負かすことができる。ただし、AIは特定のタスク用に学習する必要があり、学習セットにない他のものを認識することはできない。AIのチェスチャンピオンはブロックの周りで車を運転することはできず、自動運転AIは、ホースが取り付けられ道路で停止してライトを点滅させている大型の赤いトラック、つまり消防車について、事前に学習していない限りどうしたらよいのかわからないだろう。人間はそれより順応性がある。

深層学習は、複数レベルの処理を備えたニューラルネットワークを使用し、より大きなデータセットを並べ替えることにより、機械学習を拡張する。例えばニューラルネットワークは鼻の形、

髪の毛の色と髪型、目や肌の色、目の間の距離、その他の顔の特徴の位置など、顔を認識するために複数の要素を使用する場合がある。行列ベクトル乗算、畳み込み、フーリエ処理などのツールを使用した多くのレベルのデータの数学的処理は、音声認識、画像分類、自動運転車の運転などのアプリケーションで成功を収めている。

成功のためには、大量のデータを非常に迅速に処理する必要がある。運転はリアルタイムで行わなければならないことから特別な課題である。自動運転車は、道路の中央で止まり、道路を横切って移動する未知の物体を特定するまで待つことはできない。そのため、大量のデータを処理するときにはレイテンシ（遅延時間）が問題になる。現代の電子コンピューティングは、処理速度だけでなく高速で急速に増加していく消費電力においても実際的な限界に直面しており、設計者は数GHz以上で動作するよう半導体プロセッサを複数のコアに分割する必要がある。

開発者は、光並列処理、フォトニック集積、及びシリコンフォトニクスの組み合わせにより、遅延時間を減らし、自動運転車などのアプリケーションにおける深層学習への厳しい制限に対処することができると期待している。

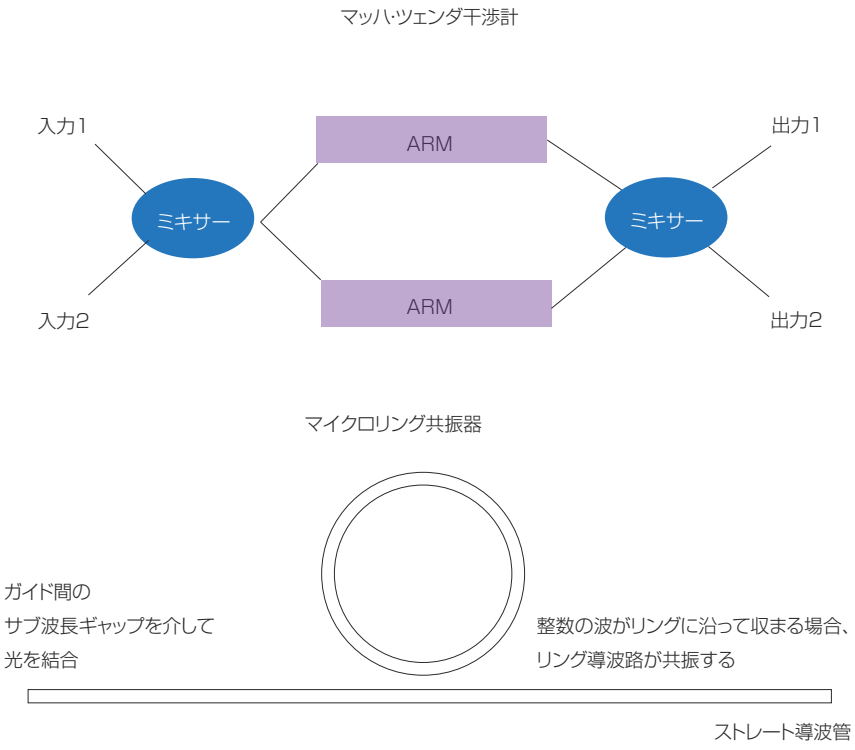

図1　集積フォトニクスの基本的な構成要素。（上）のMZIは、光がアーム内でどのように変調されているかに応じて、2つの出力間で光を分割する。（下）のMRRは、整数の波がリングの周りで適合すると共振し、サブ波長ギャップを介して隣接する導波路に光を結合する。

AIのための集積シリコンフォトニクス

深層学習において最も複雑で時間のかかる操作は、行列とベクトルの乗算である。この乗算では、M行N列の行列にN次元のベクトル（X）が乗算される。入力データは電子的に供給され、光学フォームに変換されるため、フォトニック集積回路で乗算を実行できる。これにより、Y=K・XのフォームのM次元ベクトルが生成され、電子機器に返される。その出力は、システムで使用するため、さらに処理が必要になる場合がある。

行列の乗算には複数の操作が必要であり、光学の普通の並列処理で同時に実行できる。集積フォトニクスは、古いバルク光学技術よりもはるかに効率的である。光トランシーバは電子トランシーバより消費電力が少なく、光ニューラルネットワークを完全にトレーニングできれば、マトリックスをパッシブのままにして、さらに電力を消費することなく操作を続けることができる。光行列の乗算は、一般的に数GHzの電子クロック・レートよりもはるかに速い、約100GHzまで光検出率の値を押し上げることができる。

AI集積フォトニックチップの最も一般的な2つのコンポーネントは、マッハツェンダ干渉計（Mach-Zehnder Interferometer：MZI）とマイクロリング共振器（Microring Resonator：MRR）で、どちらも**図1**に示されている。MZIの歴史は1世紀以上前にさかのぼる。左側の2つの入力が混合され、2つの並列のアームを通過する。ここで光が変調されて、右側の2つの出力間の分割が制御される。より最近に考案されたMRRは、サブ波長ギャップを介して光を他の光導波路に結合できる小さな導波路のリングである。共振は、整数の波がリングの長さに正確に適合する波長で、リングの長さに従って発生する。これらの2つの光ビルディングブロックを組み合わせると、変調器、フィルタ、マルチプレクサ、スイッチ、

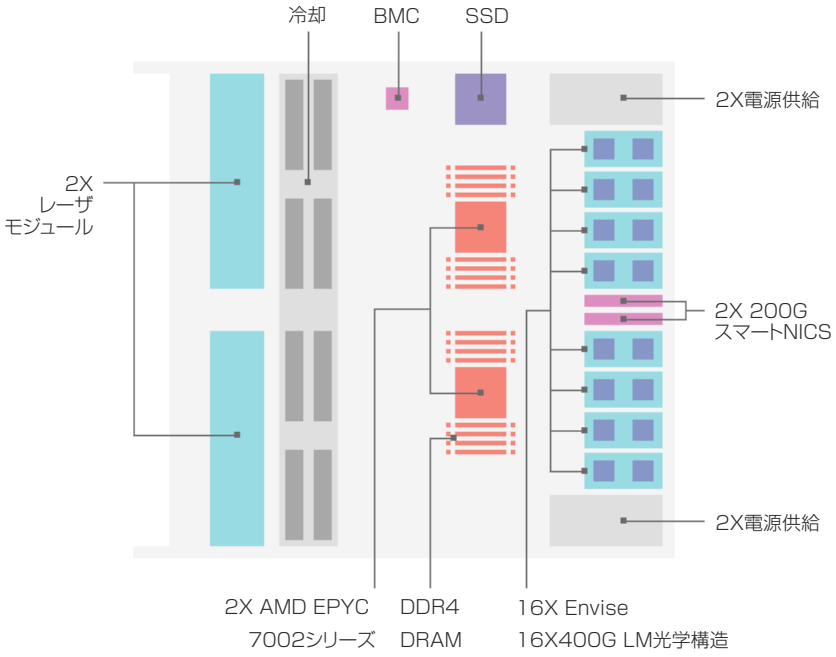

図2　Enviseフォトニックプロセッサの概略図。

及び演算器を作成できる。

マルチリング共振器とMZIは、FPGA（Field-Programmable Gate Array）と組み合わせて、光ベクトル行列の乗算を実行できる。フォトニック深層学習システムは、操作の結果を使用してオブジェクトを学習及び認識し、純粋な電子システムが現在提供できるよりも高い処理能力と速度を提供することが期待される。

研究から開発へ

開発者は、集積フォトニクスの研究、開発、製造における過去数年間の多額の投資による恩恵を受けている。その投資は、AI市場を目指して相次ぐ集積フォトニクスのスタートアップのための土壌を肥やした。

2021年3月、米マサチューセッツ工科大（Massachusetts Institute of Technology：MIT）のスピンオフである米ライトマター社（Lightmatter）は、同社が「世界初の汎用フォトニックAIアクセラレータ」と呼ぶものを発表した。「Envise」という名前のこれは、AIやニューラルネットワークの主要なモデルにプラグインしてパフォーマンスを強化できるフォトニックモジュールであり、AIの世界では1秒あたりの「推論（inference）」で測定される。Enviseには、ライトマター社の16個のフォトニックチップアレイが含まれており、4Uサーバーブレードフォーマットに組み込まれている。1TBのDRAM、3TBのソリッドステートメモリを搭載し、最大6.4Tbit/sの光相互接続を処理する。**図2**の回路図に示されているように、Enviseには、フォトニックプロセッサを駆動する2つのレーザも含まれている。基本的な構成要素は、行列の乗算用に組み立てられた同一のMZIの配列である。

図3にライトマター社のEnviseフォトニックプロセッサの内部を示す。中央の大きな光沢の箱は集積フォトニクスを収める。ベンチマークテストでは、Enviseを標準の電子システムに追加すると、1秒あたりの推論数が3倍から10倍に増加し、1秒あたりの推論の合計が57万6千から2400万近くになることが示されている。ライトマター社は、潜在的なアプリケーションとして自動車、ロボット工学、自然言語翻訳、癌検出、デジタルアシスタント、チャットボットなどを挙げている。同社はまた、ニューラルネットワークモジュールをコンパイル及び実行するためのソフトウエアと、100Tbit/sを送信できるPassageと呼ばれるプログラム可能なフォトニックテストプラットフォームを導入した。

他のいくつかのスタートアップは、集積フォトニクスを使用してさまざまなアプリケーションのAI処理を高速化するための同様の計画を持っているようだが、これまでのところ、彼らの公開するサイトはほとんど詳細を提供していない。米ルミナス・コンピューティング社（Luminous Computing）はビル・ゲイツ氏（Bill Gates）から資金を集めており、米プリンストン大（Princeton University）で学んでいた時にCTOのミッチェル・ナミアス氏（Mitchell Nahmias）が開発したコンセプトに焦点を当てていると述べている[2]。ナミアス氏の初期の研究は、集積フォトニクスを使用してアナログフォトニック操作を実装するニューロモルフィックフォトニクスに焦点を当てていた[3]。

米ライトインテリジェンス社（Light

図3　Enviseシステムの内部。中央の金属ケースにフォトニックプロセッサが収まっている。

Intelligence）は、2019年の「MIT Start up Exchange」に参加した、光ニューラルネットワークに焦点を当てたスタートアップである。そのCEO兼創設者のイチェン・シェン氏（Yichen Shen）は、MITを卒業し、コヒーレントナノフォトニック回路を使用した深層学習に関する「Nature」の論文の筆頭著者だった[4]。2020年7月26日現在、「世界最大の集積フォトニック回路」を構築しようとしていた。

仏ライトオン社（LightOn）は2015年に設立され、AIアプリケーション向けの初の大規模ハイブリッドデジタル／アナログコンピューティングプラットフォームを開発したと述べている[5]。同社は、顧客が専用のオンラインクラウドインフラを介して同社の光学処理ユニットを使用できるようにするサービスを運営している。

米ファゾム・レディアント社（Fathom Radiant）は、100mまでの距離にわたって、ワットレベルの電力で毎秒ペタビットの長距離帯域幅を実現できる光相互接続ファブリックを開発していると述べている。同社は、投資家の中にジェフ・ベゾス氏（Jeff Bezos）が含まれるとしている。

英オプタリシス社（Optalysys）は少々異なる。英ケンブリッジ大からの2001年のスピンオフであり、2013年にビッグデータに再び焦点を合わせ、画像認識のためのチップスケールのフォトニクスとフーリエ光学に焦点を合わせてきた。オプタリシス社は、2020年に初のチップレベルのフーリエエンジンを実装した。

超並列処理のための光周波数コム

光周波数コムによって生成された数千または数百万の狭いスペクトル線の波長分割多重化に基づいて、新しいレベルの、より大規模な光並列処理が実験室から出現している。当初は分光法と計測学で使用されていた周波数コムは、テオドール・ヘンシュ氏（Theodore Hänsch）とジョン・ホール氏（John Hall）により2005年のノーベル物理学賞を受賞した。Natureの2021年1月7日号には、周波数コムを組み込んだ集積フォトニクスが光ニューラルネットワークの速度と情報処理能力をどのように向上できるかについての2つの論証が含まれていた。2つのアプローチは詳細が異なるが、基本的に互換性があるように思われる。

豪スウィンバーン工科大（Swinburne University of Technology）のデビッド・J・モス氏（David J. Moss）が率いるチームは、視覚野に触発されたタイプの畳み込みニューラルネットワークで周波数コムシステムを実証した。これにより、コンピュータビジョン、音声認識や医療診断に役立つ方法でデータを分析できる。集積フォトニック周波数コムの出力は、時間、空間、及び波長の次元で同時にインターリーブされ、目的の行列ベクトルの乗算を実行する。モス氏によると、研究者の試験は速度に焦点を当てており、超高速の畳み込みを直接実行することで、1秒あ

たり11兆の行列ベクトル演算に到達した。次に、その出力を使用して、顔画像認識のために毎秒3.8兆プロセスの速度で25万ピクセルの画像を処理した。独創的なトリックの1つは、波長分散を使用して、さまざまな波長の信号に対してさまざまな時間遅延を生成し、波長に応じてそれらを組み合わせるというものだった[6]。

このグループは、結果がハイエンドの電子ニューラルネットワークの1秒あたり200兆回の操作に匹敵するものではないことを認めた。ただし、「規模と速度の両方でパフォーマンスを向上させるための簡単なアプローチがある」と研究者はNatureに書いている。「このアプローチは、自動運転車やリアルタイムのビデオ認識といった要求の厳しいアプリケーション向けの、はるかに複雑なネットワークに対して、スケーラブルでありトレーニングが可能である」。

光周波数コムの代替

同じ問題について報告している2つ目のグループは、1秒あたり数兆回の積和演算を実行できる「テンソルコア」と呼ばれる集積フォトニックハードウエアアクセラレータと呼ばれるものについて説明した。テンソルコアは、特定用途向け集積回路（Application-Specific Integrated Circuit：ASIC）の光学版である。操作を高速化するための重要なトリックの1つは、データを別々のプロセッサに移動するのではなく、メモリ内で操作を実行することであった。これは、スカラー行列の乗算に相変化材料を使用して2019年にグループが実証した手法である[7]。

2021年1月7日のNatureの新しい論文で説明されているように、そのグループはチップベースの周波数コムを追加し、技術能力を拡張して、シングル光学時間ステップで畳み込み演算を実行したと、スイスのIBMチューリッヒ社（IBM Zurich）のスタッフメンバーでグループのリーダーであるアブ・セバスチャン氏（Abu Sebastian）は言う。その畳み込みは、2つの関数に対して実行される数学的な演算であり、一方の形式の変更がもう一方の形式によってどのように変更されるかを示す3番目の関数を出力する。ニューラルネットワークは、単一の画像または他のデータに対して数十億の操作を必要とする可能性があるため、研究者は、フォトニックチップが1秒あたり数兆の操作を実行できることを証明したいと考えていた[8]。

彼らの実験では、最大9×4の要素を持ち、時間ステップごとに最大4つの入力ベクトルを持つ行列を使用し、周波数コムチップから多波長信号を送った。マトリックスを14GHzで変調すると、1秒あたり2兆の積和演算を処理できる。秘訣は、畳み込み演算をいくつかの行列ベクトル乗算演算に変換することである。これは、電子機器では不可能な、波長分割多重化を介して並列に実行することによって実現される。

「これはほんの始まりに過ぎない」とセバスチャン氏は言う。チップ上で「妥当なスケーリングの仮定により、前代未聞の一平方ミリメートルあたりの操作がペタMAC（積和演算[Multiply-Accumulate]）、つまり1000兆を達成できると期待している」。その密度は、最先端の電子AIプロセッサよりも1000倍高くなっている。

「どちらのアプローチも非常にスケーラブルだ」とモス氏は言う。原則として、それらは彼のグループの高速直接光畳み込みをその相変化チップメモリに追加することによって組み合わせることができる。

見通し

Natureの同じ号にある2本の周波数コムの論文の解説で、フアチャン・ウー氏（Huaqiang Wu）とチョンハイ・ダイ氏（Qionghai Dai）は、AIコンピューティングの要求の厳しいタスクで電子工学を成功させるフォトニクスの可能性について大きな期待感を分かち合っている。それでも彼らは、現在の光プロセッサは小型であり、実用的な光コンピュータを構築するには「集中して学際的に取り組むことが必要」であり、多くの分野にまたがった協力が必要であると述べている[9]。

集積フォトニクスコンピュータ革命は容易なものではないが、新しい進歩は、AIのさらなる進化には必須の重要な新たな可能性をはっきりと示している。最先端の電子チップは現在、公称サイズが5nm、幅が10シリコン原子である。エレクトロニクスの形状が原子レベルに近づくほど、それらの振る舞いは量子力学的になり、フォトニックチップが優位になる。

参考文献

(1) Q. Cheng et al., Proc. IEEE, 108, 8, 1261-1282 (Feb. 10, 2020); https://ieeexplore.ieee.org/document/8988228.
(2) http://tcrn.ch/30SW6Pd.
(3) A. N. Tait et al., Sci. Rep.,7, 7430 (2017); https://doi.org/10.1038/s41598-017-07754-z.
(4) Y. Shen et al., Nat. Photonics (Jun. 12, 2017); doi:10.1038:/nphoton.2017.93.
(5) http://bit.ly/LightOnRef.
(6) X. Xu et al., Nature (Jan. 7, 2021); https://doi.org/10.1038/s41586-020-03063-0.
(7) C. Ríos et al., Sci. Adv., 5, eaau5759 (Feb. 15, 2019).
(8) J. Feldmann et al., Nature, https://doi.org/10.1038/s41586-020-03070-1.
(9) H. Wu and Q. Dai, Nature, 589, 25-26 (Jan. 7, 2021).

LFWJ