第2章 ChatGPTやStable Diffusionでも使われる

深層学習の最新アーキテクチャ Transformer

佐藤 聖

● CNNやRNNに代わるニューウェーブ

Transformerは，最新のニューラル・ネットワーク・アーキテクチャの1つです．従来のCNN（畳み込みニューラル・ネットワーク）やRNN（リカレント・ニューラル・ネットワーク）に取って代わる新技術です．2017年にグーグルとトロント大学の研究者によって開発されました．現在，汎用人工知能に向けた開発が続いています．

● 特徴…高精度で軽量

8ビット浮動小数点（FP8）精度の演算でも十分なモデル精度が出せます．トレーニング時間の短縮だけでなく，消費電力を抑えつつ，精度の高いモデルが開発できる環境が整いつつあります．例えば，Transformer Engine搭載のH100 Tensorコア GPU（エヌビディア）があります．大規模なモデル構築に重要なFP8精度のトレーニングを，従来の6倍以上で処理できるそうです．数週間のトレーニングを数日に短縮できるようです．

● 既に皆さんも利用している

プログラム開発をしない人でも，Transformerモデルを使っています．ウェブの検索エンジン，動画検索エンジン，Googleドキュメントのスペルや文法チェック機能，Google翻訳，スマートフォンのカメラ，iPhoneのSiri，SNSのお勧め機能などで利用しています．

従来の深層学習における課題

● CNNやRNNはデータセットの準備が大変

従来，画像分析にはCNN，文書翻訳にはRNNが用いられてきました．トレーニングにはラベル付きデータセットの準備が必要でした．この準備に多大な労力，コスト，時間が掛かり，約8割の機械学習プロジェクトはデータセットの準備で失敗すると言われていました．

▶ Transformerではラベル付けが不要に

Transformerは，データセットの要素間のパターンを数学的に発見します．また，マルチモーダル深層学習の深耕を目指しており，テキスト，画像，音声などのモダリティなタスクを統合して，機械学習の実行を目指しています．データセットにラベル付け作業が不要になるので，作業が大幅に簡略化できます．ただし，トレーニングには，膨大な画像やテキストのデータを必要とします．

また，IoTセンサ・データからリアルタイムにデータ分析したい場合にも，Transformerは向いています．複数のIoTセンサから送られてくるデータを使い，システム全体がどのような状況であるか分析するのに有効です．大量のデータからモニタリング対象の状態を総合的に判断するのに役立ちます．

● CNNから進化したRNN…言語処理に向くとするがモデルが巨大に

CNNは，人間の脳が視覚処理する方式を，漠然と模倣するアプローチを取りました．主にニューラル・ネットワークの情報の伝播や重みのチューニングに重きが置かれています．この方式の欠点として，汎用的な言語モデル処理が困難でした．

そこで，文章情報を処理する方式として，RNNが登場しました．文章だけでなく，時系列データの分析にも使われます．Transformerアーキテクチャを使ったBERTモデルが登場するまで，日本語と英語などの文章翻訳によく使われてきました．

RNNには2つの欠点があります．

1つ目は，部分的に逐次処理である点です．常に単語の順序が重要になり，単語の出現順序をニューラル・ネットワーク内に保存します．トレーニングのコストや時間がかかり，モデルが巨大になります．長文の翻訳には不向きでした．

2つ目は，処理の並列化に課題があります．コンピュータの演算リソース（特にGPU）を増やして演算規模をスケーリングしても学習の高速化に限界がありました．

RNNの欠点を克服したTransformerアーキテクチャ

Transformerアーキテクチャは，もともと翻訳処理

◆参考・引用*文献◆
(1) * Attention Is All You Need.
https://arxiv.org/pdf/1706.03762.pdf