Transformer

品川政太朗（奈良先端科学技術大学院大学）
sei.shinagawa@is.naist.jp

1. Transformer とは

Transformer[1] とは，系列情報を処理する，注意機構を主体とした深層学習モデルである．最初に提案された当初は英語をドイツ語に翻訳するといった自然言語処理の系列変換タスクで有効性を示したが，今日では様々なタスクやモダリティに応用されている．Transformer の大きな貢献は系列データの学習の並列化にあり，当時主流だった再帰ニューラルネットワークベースの方法に比べて大規模なテキストデータを処理できるようになった．これが現在の大規模言語モデル時代につながっている．Transformer 内部の基本処理単位である Transformer block は，近年の深層学習モデルに共通して用いられており，Transformer や Transformer から派生した方法論は深層学習の基礎としての地位を確立している．深層学習の初学者が学習初期に Transformer を学ぶことも当たり前になってきた．本稿は，Transformer とは何なのか，何をもたらしたのかを簡潔にまとめ，Transformer とその周辺について読者の理解の助けとなることを目指す．

2. Transformer の仕組み

Transformer はエンコーダ・デコーダと呼ばれる典型的な系列変換モデルであり，一つのエンコーダモデルと一つのデコーダモデルが接続される形を成している．英日翻訳のようにある言語（原言語）の単語列を別の言語（目的言語）に変換するタスクを想定すると，エンコーダは原言語の特徴を抽出し，デコーダはエンコーダから得られた特徴を参照しつつ，目的言語の単語を一つずつ生成する．

エンコーダとデコーダの内部の構造は図 1 のようにほぼ共通しており，Transformer Block（TB）と呼ばれるモジュールが基本単位になる．TB は多頭注意（Multi-head Attention），層正規化（Layer Normalization），残差接続（Residual Connection），2 層 MLP（Multi-layer Perceptron）で構成される．ここで，層正規化の位置は多頭注意や 2 層 MLP の前に配置する PreNorm の方式に従っており，原論文とは異なることに注意されたい．PreNorm を採用した理由は，現在の多くのモデルが学習の安定化の観点からこの PreNorm を採用しているのに加え，Transformer の著名な解説である The Annotated Transformer[2] も PreNorm で解説しており，本稿もこれに倣った方が読者の学習が円滑に進むと考えたためである．

デコーダはエンコーダと異なり，エンコーダの特徴を参照しながら単語を一つずつ予測するために相互注意（Cross-attention）と注意マスク（Attention Mask または Causal Mask）と呼ばれる仕組みを持つ．相互注意はエンコーダの特徴をデコーダに取り込む仕組みであり，注意マスクは単語を一つずつ生成する際に未来の情報の参照を防ぐ仕組みである．これによって，Transformer は系列処理において複数の時刻を並列に学習することができる．

3. Transformer の構成要素

注意機構（Attention Mechanism）は元々 2014 年に発明されたアイデアであり，これまでに様々な形態が存在していたが，現在では Transformer が採用している Query，Key，Value による自己注意，相互注意という形が標準的に用いられている．

Query，Key，Value はそれぞれベクトルのトークン列として与えられる．それぞれが入力トークンの線形変換で与えられ，Query が同じ情報源に由来するときは自己注意，異なる情報源に由来するときは相互注意となる．Transformer では基本的には自己注意が用いられ，エンコーダの出力をデコーダで参照する際に相互注意が用いられる．

Key と Value は同じベクトルに別々の線形変換をかけるが，トークンの対応関係を仮定する．注意機構の役割は必要な情報の選択である．これは実際には，Value の各トークンを重み付けして足し合わせる内挿操作により実現される．この重み付けに用いる重みがいわゆる注意あるいは注意重み（Attention Weights）と呼ばれているものである．

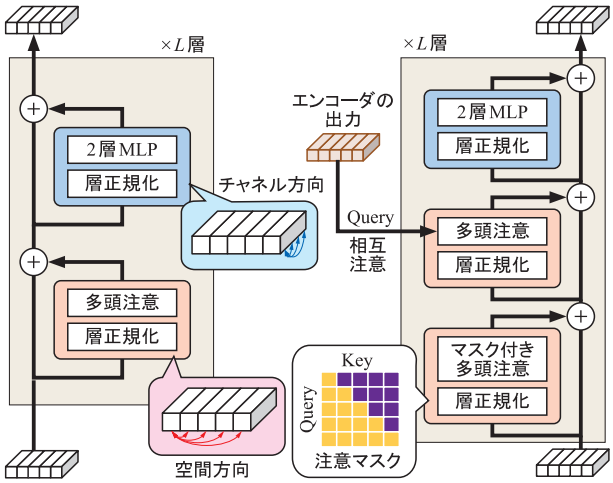

図 1　エンコーダ（左図）とデコーダ（右図）内部の Transformer Block

本会ハンドブック「知識の森」
https://www.ieice-hbkb.org/portal/doc_index.html

この注意重みはQueryとKeyの内積（ただし，正規化処理を含む）によって計算される．注意機構は計算上，Queryを入力としてQueryと同じサイズのベクトルを出力する変換器として解釈できるが，QueryとValueが間接的に計算されることで強力な非線形変換を実現できる．

多頭注意は注意機構の計算をより高度化する工夫である．多頭注意は，Query，Key，Valueの各ベクトルをチャネル方向に同じサイズの小さなベクトルに分割してから計算を行い，その後ベクトルを結合して元のサイズのベクトルに戻す．次元の大きなベクトル同士では内積計算を行う際に小さな相関を持つ成分が無視される問題があるが，事前にベクトルを小さく分割することでこれらの相関を捉えることが可能になる利点がある．次に，他の構成要素について述べる．

- 層正規化：層正規化は学習の安定化に用いられる．トークンごとに正規化を行うため，系列の長さが変化しても安定して正規化を行うことができる．
- 残差接続：学習の安定化と多層化に貢献する．入力を二つに分岐させ，一方はモジュールに入力して出力を得たあと，もう一方の分岐を統合し，元の入力を足し合わせる仕組みである．層正規化との位置関係は学習に重要であり，Transformerに用いられていたPostNorm（残差接続の後に層正規化）は現在ではあまり使われておらず，より安定的なPreNorm（非線形変換モジュールの前で層正規化）が好まれる．
- 2層MLP：非線形変換を行う機構として用いられる．多頭注意がトークン間の関係性を捉える役割であるのに対して，2層MLPはトークンごとに処理を行い，系列によらない，モデル内部に学習された情報を抽出する役割があると言われている．
- 位置埋め込み：TBは各トークンの順序を考慮しないため，系列の位置情報を考慮する場合には別途位置情報を挿入する必要がある．位置情報は位置埋め込みと呼ばれるベクトルとして入力トークンに挿入するか注意機構の注意重み計算時のバイアスとして加算することが多い．Transformerの位置埋め込みはSinusoidal Positional Encoding（SPE）と呼ばれる方法で作成される．SPEはsin関数とcos関数によって位置埋め込みを表現する方法で，相対位置がcos関数の位相として表現される．また，学習不要で用いることができ，任意の系列長に適用できる．

4. Transformerがもたらしたもの

Transformerの最大の貢献は，現在の大規模事前学習の世界を切り開いたことにある．2017年当時，自然言語処理において，系列情報を扱うモデルとして主流だったのは再帰ニューラルネットワークだったが，このネットワークは系列情報を時系列順に順番に入力して隠れ層を更新する必要があり，学習に時間がかかっていた．一方，Transformerは注意マスクを利用することで，並列に訓練を行うことができた．これが自然言語処理における大規模事前学習という発想につながったと考えられる．

Transformerによる大規模事前学習の威力を示したのはBERT[3]である．BERTはエンコーダ型のTransformerで，大規模データによるTransformerモデルの事前学習（Pre-training）と下流タスクでの微調整（Fine-tuning）という考え方を普及させた．

昨今よく耳にする「大規模言語モデル（LLM : Large Language Models）」という言葉が登場し始めたのも，BERTの登場以降である．

BERTでは，単語をランダムにマスクし，その単語を予測するMasked Language Modelingと，入力した2文が連続した2文か否かを当てるNext Sentence Predictionの2種類の自己教師あり学習（Self-supervised Learning）により，データ自身が持っている情報や対応関係を予測することで，人手によるラベル付けに頼ることなく事前学習を行う．この事前学習によって訓練された学習済みのBERTは，様々な下流タスクにおいて高い転移学習性能（学習済みモデルを下流タスクのデータセットで微調整したときのモデルの予測性能）を示した．

2018年には，大規模言語モデルの代名詞となっているChatGPTやGPT-4の原点であるGPT-1[4]も登場した．GPT-1は，Transformerを利用したデコーダモデル，つまり言語モデルである．GPT-1は次のトークンを予測する最ゆう推定で学習され，目的の下流タスクに合わせて微調整して分類タスクを解く．GPT-2[5]，GPT-3[6]では，学習データとモデル規模が増大した．少数の事例やタスクを文章で与えることで，その後に生成される文章を回答にするように学習するInstruction Tuningというアイデアも登場し，これが昨今の大規模言語モデル時代につながっている．近年の大規模言語モデルは，画像や音声といったあらゆるモダリティも組み合わせてマルチモーダル化する方向に進んでおり，自然言語処理に限らない多くの分野でも威力を発揮している．

5. おわりに

本稿ではTransformerについて解説した．より本格的に学びたい読者は，まずThe Annotated Transformer[2]を参照されるのがよいだろう．Transformerの仕組みを論文に沿ってコード付きで解説しており，学びに最適な資料の一つとして評判が高い．2022年に刷新された最新版[7]が存在するので，初学者にはこちらの方がより適していると思われる．その後は興味に応じて著名な自然言語処理や大規模言語モデルの入門書[8,9]を参照されるとよいだろう．

文　献

（1） A. Vaswani, et al.,“Attention is all you need,” Proc. NeurIPS, vol. 30, 2017.
（2） A. Rush,“The annotated transformer,” Proc. NLP-OSS workshop, pp. 52-60, 2018.
（3） J. Devlin, et al.,“BERT : Pre-training of deep bidirectional transformers for language understanding,” Proc. NAACL, pp. 4171-4186, 2019.
（4） A. Radford, et al.,“Improving language understanding by generative pre-training,” OpenAI Blog, 2018.
（5） A. Radford, et al.,“Language models are unsupervised multitask learners,” OpenAI blog, 2019.
（6） T. Brown, et al.,“Language models are few-shot learners,” Proc. NeurIPS, vol. 33 1877-1901, 2020.
（7） https://nlp.seas.harvard.edu/annotated-transformer/（2024年3月確認）
（8） 岡崎直観，鶴岡慶雅，荒瀬由紀，宮尾祐介，鈴木　潤，IT Text 自然言語処理の基礎，オーム社，2022.
（9） 山田育矢，鈴木正敏，山田康輔，李　凌寒，大規模言語モデル入門，技術評論社，2023.

（2024年4月16日受付）