AI 特許紹介(86)
AI 特許を学ぶ！究める！
～Infini-attention 特許～

2026 年 3 月 10 日
河野特許事務所
所長弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要
特許出願人　Google
出願日　2025 年 4 月 10 日
公開日　2025 年 11 月 25 日
公開番号 WO2025230701
発明の名称　効率的なロングコンテキストトランスフォーマーのための圧縮メモリ

701 特許は、圧縮メモリをトランスフォーマベースの生成モデルのアテンション機構に組み込むことで、限られたメモリと計算量で、潜在的に無限長の入力へのスケーリングを可能にする Infini-attention 技術に関する。

2.特許内容の説明
トランスフォーマベースの大規模言語モデル等の生成モデルアーキテクチャは、アテンション機構固有の性質により、コンテキスト依存のメモリに関連する課題に直面している。アテンション機構の計算要求とメモリフットプリントは、シーケンス長の 2 乗で増加する。たとえば、アテンションのキーバリュー(KV) 状態は、大規模モデルと長いコンテキスト長に対して大量のメモリリソースを必要とする。100 万トークン以上等の

長いシーケンスを処理できるように生成モデルをスケーリングすると、トランスフォーマに基づくものも含め、さまざまな生成モデルアーキテクチャで問題が生じる。さらに、コンテキスト長が増加する生成モデルを提供した場合、計算コストが増大する。

本発明は、圧縮メモリをトランスフォーマベースの生成モデルのアテンション機構に組み込むことで、限られたメモリと計算量で、潜在的に無限長の入力へのスケーリングを可能とする。

下記図は知識システム 100 のネットワーク環境を示すブロック図である。

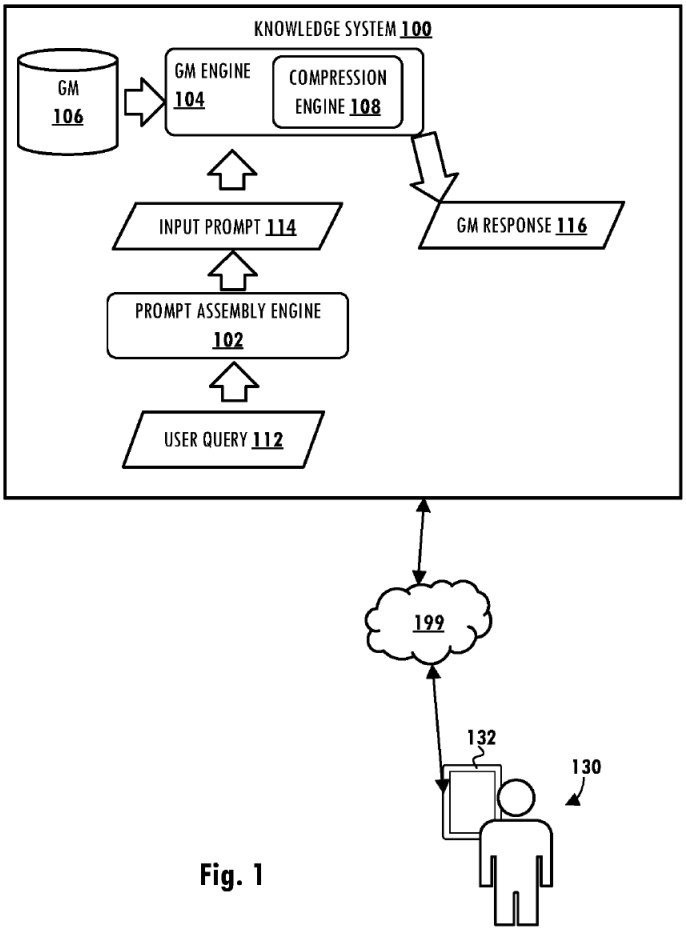



Fig. 1

知識システム 100 は、プロンプトアセンブリエンジン 102、生成モデル（GM；generative model）エンジン 104、GM106、および圧縮エンジン 108 を含む。プロンプトアセンブリエンジン 102 は、クライアントデバイス 132 からユーザクエリ 112 を受信する。次に、プロンプトアセンブリエンジン 102 は、ユーザクエリ 112 から入力プロンプト 114 を組み立て、入力プロンプト 114 を GM エンジン 104 に提供する。GM エンジン 104 は、入力プロンプト 114 を処理して、GM106 を使用して生成される GM

応答 116 を生成する。GM エンジン 104 は、GM 応答 116 をクライアントデバイス 132 に提供する。

Transformer ベースのアーキテクチャは、アテンション機構の性質上、従来、コンテキスト依存メモリの制約を受ける。場合によっては、アテンション機構はメモリフットプリントと計算時間の両方で 2 乗の複雑性を示す。例えば、バッチサイズが 512、コンテキスト長が 2048 の 5000 億パラメータモデルの場合、アテンションのキーバリュー（KV）状態は最大 3TB のメモリフットプリントを持つ。そのため、LLM などの生成モデルをより長いシーケンス（100 万トークンなど）にスケーリングすることは Transformer アーキテクチャでは困難になり、ますます長いコンテキストモデルの提供には計算コストと費用を要する。

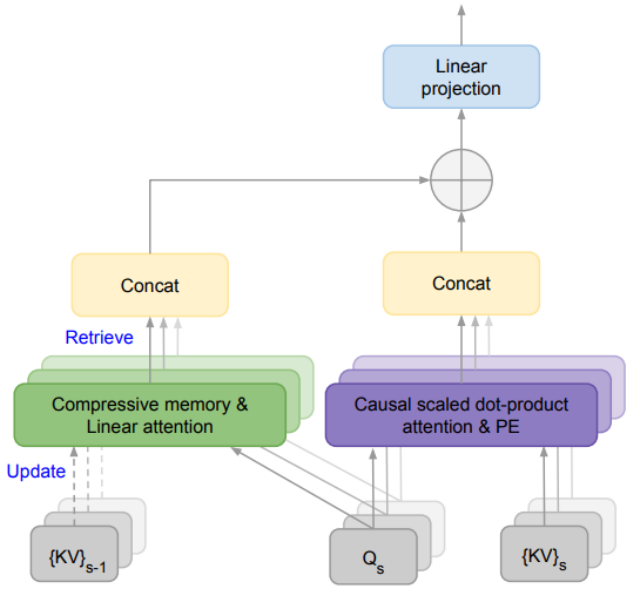
圧縮メモリシステムは、極めて長いシーケンスに対して、アテンション機構よりもスケーラブルで効率的であることが期待される。入力シーケンスの長さに応じて増加する配列を使用する代わりに、圧縮メモリは主に固定数のパラメータを維持し、限られたストレージコストと計算コストで情報を保存および呼び出す。圧縮メモリでは、パラメータを変更することで新しい情報がメモリに追加され、後でその情報を復元できる。

このため、GM エンジン 104 は、圧縮メモリを活用して生成モデルによるより長いコンテキスト長のより効率的な処理を可能にするように構成された圧縮エンジン 108 を含む。圧縮エンジンは、GM エンジン 104 のための情報を統合する。例えば、圧縮エンジン 108 は、ローカルアテンション計算からキーとバリューの状態を受け取る。これらの古い状態を破棄する代わりに、圧縮エンジン 108 は、キーとバリューの状態のバインディングを圧縮メモリに格納する。後続のシーケンスを処理する際、圧縮エンジン 108 は、アテンションクエリ状態を使用して圧縮メモリからバリューを取得することができる。その後、長期メモリから取得されたバリューとローカルアテンションコンテキストを集約して、最終的なコンテキスト出力を計算する。

圧縮エンジン 108 は、アテンション層のスタックに供給される最初のトークンセグメントから開始する。各アテンション層では、トークンはアテンション処理を受け、最終的にその層に固有のローカルアテンション状態が生成される。このアテンション処理の副産物として、そのセグメントの隠れ状態が導出され、圧縮メモリに格納される。この処理は、アテンション層のスタックを通過する 2 番目のトークンセグメントに対しても繰り返される。各層において、トークンはアテンション処理され、再びローカルアテンション状態が生成される。次に、1 番目のセグメントから以前に格納された隠れ状態が、2 番目のセグメントから新たに生成されたローカルアテンション状態と統合される。

圧縮メモリは、2 番目のセグメントへのアテンション処理の副産物として導出された隠れ状態を格納することで再び更新される。この反復処理により、知識システム 100 は異なるトークンセグメントにまたがって情報を保持し、過去の知識と現在の入力を統合することで、コンテキストアウェアネスを向上させることができる。

下記図は、Infini-attention 機構を示す説明図である。




Infini-attention は、無限長のコンテキストを処理するための線形アテンションを備えた追加の圧縮メモリを備えている。{KV}s−1 と{KV}s は、それぞれ現在の入力セグメントと前の入力セグメントのアテンションキーとバリューであり、Qs はアテンションクエリである。PE は位置埋め込みを表す。

図に示すように、本アーキテクチャは、従来のアテンション機構に圧縮メモリを組み込み、マスクされたローカルアテンション機構と長期線形アテンション機構の両方を単一のトランスフォーマーブロックに組み込んでいる。図のアーキテクチャは、標準的なアテンション計算のキー、バリュー、およびクエリ状態をすべて、長期メモリの統合および検索に再利用する。アテンション機構の古い KV 状態は、従来のアテンション機構のように破棄されるのではなく、圧縮メモリに保存する。その後、後続のシーケンスを処理する際に、アテンションクエリ状態を用いて圧縮メモリからバリューを検索する。最終的なコンテキスト出力を計算するために、長期メモリから検索されたバリューは、ローカルアテンションコンテキストと集約される。

3.クレーム

701 特許のクレーム 1 は以下の通りである。

1.1 つまたは複数のプロセッサを使用して実装される方法において、

複数のアテンション層のスタックを使用してトークンの第 1 セグメントを処理し、トークンの第 1 セグメントの処理は、各アテンション層において、以下のステップを含み、

各アテンション層の第 1 セグメントローカルアテンション状態を生成するために、第 1 セグメントから導出されたトークンにアテンションし、

アテンションの副産物を、各アテンション層の第 1 セグメント隠れ状態として圧縮メモリに格納し、

複数のアテンション層のスタックを使用してトークンの第 2 セグメントを処理し、トークンの第 2 セグメントの処理は、各アテンション層において、以下のステップを含み、

各アテンション層の第 2 セグメントローカルアテンション状態を生成するために、第 2 セグメントから導出されたトークンにアテンションし、

各アテンション層の第 1 セグメント隠れ状態を、各アテンション層の第 2 セグメントローカルアテンション状態と集約し、

第 2 のセグメントから導出されたトークン全体にアテンションすることによる副産物を、各アテンション層に関連付けられた第 2 セグメント隠れ状態として、圧縮メモリに格納する。

4. 本特許に関連する論文

本特許に関する論文“Leave No Context Behind:Efficient Infinite Context Transformers with Infini-attention”[1]が、Google の Tsendsuren Munkhdalai 氏らにより公表されている。下記表には、50 万語の書籍要約（BookSum）の結果が示されている。

| Model | Rouge-1 | Rouge-2 | Rouge-L | Overall |
|---|---|---|---|---|
| BART | 36.4 | 7.6 | 15.3 | 16.2 |
| BART + Unlimiformer | 36.8 | 8.3 | 15.7 | 16.9 |
| PRIMERA | 38.6 | 7.2 | 15.6 | 16.3 |
| PRIMERA + Unlimiformer | 37.9 | 8.2 | 16.3 | 17.2 |
| Infini-Transformers (Linear) | 37.9 | 8.7 | 17.6 | 18.0 |
| Infini-Transformers (Linear + Delta) | **40.0** | **8.8** | **17.9** | **18.5** |

要約タスク向けに特別に構築されたエンコーダー・デコーダーモデル（Lewis et al.,

[1] Tsendsuren Munkhdalai, et al. “Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention” arXiv:2404.07143v2 [cs.CL] 9 Aug 2024

2019; Xiao et al., 2021）およびそれらの検索ベースの長文脈拡張（Bertsch et al., 2024）と、本モデルを比較した。本モデルは、書籍の全テキストを処理することで、これまでの最高結果を上回り、BookSum で新たな SOTA を達成している。

著者紹介

河野英仁

河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース、生成 AI ビジネスコース修了。

AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 3」、「ブロックチェーン 3.0(共著)」がある。