リザバーコンピューティングを用いた
次世代型自然言語処理モデルの開発
―リザバー計算に触発された軽量型 Transformer の提案―

1. 背景

近年、大規模言語モデル（Large Language Model: LLM）の著しい発展に伴い、機械翻訳や文章生成といった自然言語処理の分野において高い性能が実現されている。しかし、モデルの規模拡大に伴い、学習および推論に要する計算コストや消費電力が急増し、環境負荷や運用コストの増大が深刻な課題となっている。特に、エッジ環境や災害・医療などの現場では、クラウドに依存せず低電力かつ限られた計算資源下で高精度な言語処理を行う必要性が高まりつつある。一方、リザバーコンピューティングは、固定パラメタをもつ中間層の非線形ダイナミクスを活用し、低コストかつ高速に学習できる手法として期待されてきた。本プロジェクトでは、こうしたリザバーコンピューティングの特性を Transformer アーキテクチャに統合し、学習パラメタ数の削減と翻訳性能の維持を両立する新たな軽量型モデルの開発を目指した。これにより、従来の大規模モデルが抱えていた計算量と消費電力の問題を軽減し、エッジ環境を含む多様な応用分野での高精度な自然言語処理を可能とする道を拓くことを目的とする。

2. 目的

本プロジェクトの目的は、大規模化による計算負荷と消費電力の増大という問題に対処すべく、リザバーコンピューティングを取り入れた新たな Transformer モデルを構築することである。具体的には、以下の点に重点を置いた。

(ア)軽量かつ省エネルギーなアーキテクチャの確立

リザバー層（固定層）と Transformer 層（学習層）を組み合わせることで、学習パラメタ数や計算負荷を削減する。

(イ)英日翻訳タスクにおける有用性の検証

上記アーキテクチャを英日翻訳に適用し、従来の全層学習型 Transformer と同等に近い翻訳性能を維持しながら省電力化を図る。

(ウ)将来的なエッジ実装への展望

固定層をハードウェア（FPGA 等）で再利用する仕組みを見据え、分散学習や量子化など追加的な高速化・軽量化技術との親和性を検討する。

3. ソフトウェア開発内容

3.1　提案モデル

本プロジェクトでは、図 1 に示すように、Encoder 部分を「固定層（リザバー層：R 層）」と「学習層（L 層）」で交互に構成したモデルを開発した。R 層は初期化後に重みを更新せず、非線形ダイナミクスによる特徴変換を担う。一方、L 層は従来の Transformer ブロックとして勾配を通して学習し、タスクに適応する表

現を獲得する。さらに、R 層どうし、L 層どうしでパラメタを共有することで、モデル全体のパラメタ数を削減し、学習時のメモリ消費量と計算コストを抑制した。

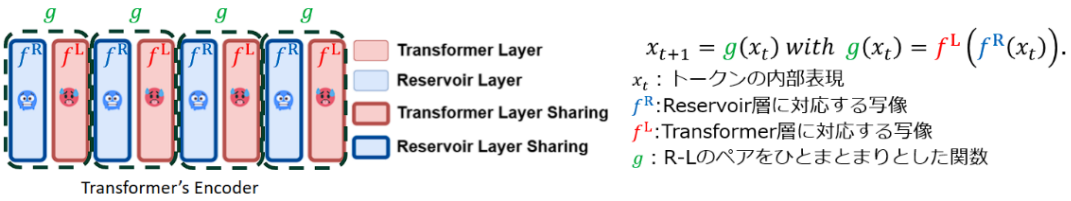


図 1. 提案モデルの Encoder 部の概念図

3.2　英日翻訳タスクへの適用

(ア)データセット: 田中コーパス、TED 字幕、青空文庫、JParaCrawl など複数の日本語-英語対訳を統合し、約 110 万文対の学習用データを作成した。

(イ)学習環境: PyTorch および fairseq を用い、NVIDIA A100 などの GPU 上で分散学習を実施した。学習時間を 30 分とし、その間に到達した BLEU スコアや学習ステップ数を比較指標とした。

3.3　実験結果の概要

本プロジェクトでは、下記 5 種類のモデルを比較した。

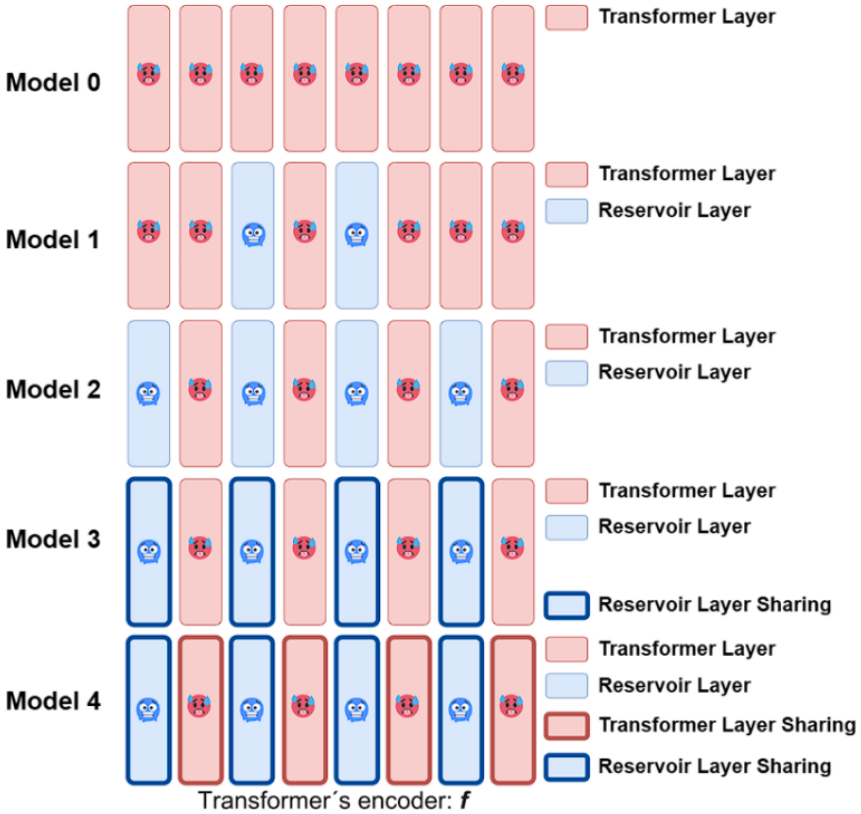
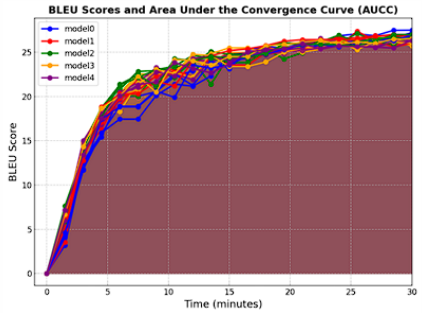


図 2　実験に用いた 5 つのモデルの概要

評価の結果を

表 1 に示す。Model 4 のように固定化の割合を増やしても、学習パラメタが大幅に減る一方、BLEU スコアはベースライン（Model 0）と同等レベルの約 27〜28 を維持できた。さらに、固定層を増やすほど 1 ステップあたりの計算量が減少し、同一学習時間内でより多くのステップを回せることが確認された。以上により、提案モデルが省エネルギー化と翻訳性能を高水準で両立し得ることが示唆された。

表 1　英日翻訳タスクにおける 5 つのモデルの BLEU スコアと学習ステップ数



Model	AUCC
Model 0	640.32 ± 7.22
Model 1	640.32 ± 7.22
Model 2	**667.64 ± 10.66**
Model 3	664.43 ± 5.00
Model 4	659.86 ± 5.00

Model	パラメタ数 (M)	Epochs @30min	Steps @30min
Model 0	57.54	19.00 ± 0.00	21040.5 ± 148.34
Model 1	57.54	20.00 ± 0.00	21808.0 ± 186.77
Model 2	57.54	20.00 ± 0.00	22324.0 ± 194.04
Model 3	53.34	20.00 ± 0.00	22449.5 ± 83.13
Model 4	**49.13**	**21.00 ± 0.00**	**23335.5 ± 307.57**

4. 新規性・優位性

(ア)固定層（リザバー層）と学習層を交互に配置するハイブリッド設計

従来の Transformer では、全ての層を学習対象とするため学習パラメタ数が膨大になりやすかった。本プロジェクトで開発したモデルは、一部の層をランダム初期化後に固定（リザバー層）し、残りの層（学習層）を通常通り勾配更新する方式を採用する。これにより、学習パラメタ総数を大幅に削減しながら、非線形変換の恩恵を維持できる点が新規である。

(イ)リザバー層・学習層間でのパラメタ共有によるさらなる軽量化

リザバー層の導入に加え、同種類の層間でパラメタを共有することで、モデル全体のパラメタ数を一層減らす手法を提案している。層間共有を組み合わせることにより、従来の Transformer と比べ、メモリ使用量や学習計算量を著しく削減できる点が優位性として挙げられる。英日翻訳タスクにおいても同等レベルの性能を保持英語と日本語のように構造が大きく異なる言語間の翻訳に適用しても、提案モデルはベースライン Transformer

と同等レベルの BLEU スコア（約 27〜28）を維持できることが確認された。これは、固定層の導入とパラメタ共有による軽量化が、高精度な自然言語処理性能を損なわずに機能することを示すものである。

(ウ)省電力・高速学習が可能なアーキテクチャ

固定層が増えるほど学習時の逆伝播計算が減少し、同一学習時間内でより多くの学習ステップを回せるため、省電力性や時間短縮の観点から有利である。将来的に FPGA などのハードウェア実装へ展開した際も、リザバー層を回路ブロックとして再利用する設計が可能なため、高い演算効率と低電力化を実現しやすい点が大きな強みとなる。

5. 期待されるユーザー価値と社会へのインパクト

本プロジェクトで開発した軽量型自然言語処理モデルは、大規模クラウドに依存せずとも高精度な翻訳を実行可能とするアーキテクチャを備えている。この特性により、以下のような価値がユーザーにもたらされ、社会全体に対して大きな影響を及ぼすと考えられる。

(ア) エッジ環境での運用

省電力・高速学習の設計により、クラウド接続が困難な場所やリソースが限られた現場においても、高精度な日本語翻訳がローカルで実行可能となる。災害時や医療・防衛などの重要インフラ下でも安定的な情報処理が期待でき、運用リスクの軽減につながる。

(イ) 運用コストの大幅削減

モデルの学習パラメタや計算負荷が削減されることで、GPU リソースや電力消費を抑制できる。これにより、大規模サーバを保有しない研究機関や企業であっても、高度な自然言語処理モデルを容易に導入・運用する道が開ける。結果として、全体的な運用コストの削減が見込まれる。

(ウ) 環境負荷の低減

消費電力を抑えられることから、データセンターを含む広範なコンピューティング環境での省エネルギー化に寄与する。膨大な電力を要する大規模言語モデルが増加する中、本プロジェクトのモデルは持続可能な AI 技術として社会的意義が高く、環境負荷低減へ向けた有効な手段となり得る。

6. 氏名（所属）

中村 仁（大阪大学 大学院情報科学研究科）