公開

研 究 終 了 報 告 書

「ランダムベクトルを用いた軽量な埋め込み表現の構築」

研究期間： 2020 年 11 月～2022 年 9 月
研 究 者： 高瀬 翔

1． 研究のねらい

画像情報処理や音声情報処理と同様に、自然言語処理においてもニューラルネットワークを用いたモデルによる性能向上が目覚ましい。これらニューラルモデルは学習コーパスの大きさやモデルのパラメータ数、学習に費やした計算時間と性能とが対数比例すると報告されており、モデルのパラメータ数や学習時間が著しく増加している。このような状況を受け、パラメータ数や学習時間のような計算コストに対し、効率の良い手法、すなわち、性能を維持しつつパラメータ数や学習時間を減らす、あるいは同程度のパラメータ数や学習時間である場合に高い性能を達成可能な手法を考えることが本研究の目的である。

自然言語処理においては 2007 年の Brants らによる Stupid Backoff や 2013 年の Mikolov らによる word2vec、Vaswani らによる 2017 年の Transformer のように、複雑な計算を捨象し、単純なモデルで置き換えることで飛躍的な性能向上を達成してきた。本研究でも効率を上げたモデルを考案し、翻訳や要約をはじめとして自然言語処理の様々な応用タスクで使用されることを目指す。本研究では特に自然言語処理のためのニューラルモデルにおいて単語を表現するために使用される埋め込みに着目し、これを省パラメータで表現可能な手法を探求する。

2． 研究成果

（1）概要

研究のねらいに記したように、本研究では自然言語処理を行うニューラルモデルについて、効率の良いモデルを探求することが目的であった。これについて、本研究では、単語をベクトル表現に変換するための単語埋め込み層、および、ニューラルモデルにおける埋め込み層以外の部分（中間層）について、パラメータ効率の良いモデルの探求を行った。具体的には、単語埋め込み層については、各単語に固有のベクトルを用意するのではなく、ベクトルを少数のランダムベクトルの組み合わせで構成することにより、通常使われている手法よりも少ないパラメータ数でありながら、同程度の性能を達成可能な手法を考案した。特に、機械翻訳の実験においては埋め込み層のパラメータ数を 1/2 に、要約生成タスクにおいては約 1/10 としても性能を維持可能なことを示した。

中間層については、従来の、パラメータ効率の良い手法として知られるパラメータ共有について、すべての層ではなく一部の層で共有すること、および、共有する層の割当の戦略を提案し、既存のパラメータ共有手法よりも計算速度、性能ともに優れた手法であることを実験を通して示した。また、現在広く使われている Transformer というニューラルモデルには正規化層の位置の違いで 2 つの構造があるが、層を積み重ねた際に学習が不安定な構造の方が学習が成功した場合に性能が良い、すなわち、パラメータ効率が良いことを示し、これの学習を追加でのパラメータや計算コストの増加なしに安定させる手法を提案した。加えて、推論時に複数の入

力からの出力を統合することで、1 つのモデルでもモデルアンサンブルのように性能が向上可能なことを示した。
パラメータ効率に加えて、計算時間の効率に関する研究にも取り組んだ。本研究では、敵対的摂動のような頑健な手法構築に用いられる正則化手法の効果を様々なタスクで検証し、モデルの頑健性の観点では単語をランダムに置き換えるような単純な手法で置き換え可能であり、学習時間に対する効率を考えると単純な手法でない場合には費やした時間に対する性能向上が見合っていないことを示した。

（2）詳細

本研究では自然言語処理を対象としたニューラルモデルの効率化に取り組んだ。具体的には単語埋め込み層、中間層のそれぞれについてパラメータ効率の良いモデルの探求を行った。また、計算時間の効率性については、学習時間に対して効率の良い正則化手法の探求を行った。

「単語埋め込み層のパラメータ効率化」

自然言語処理を行うニューラルモデルでは各単語を埋め込み表現というベクトル表現に変換し、入力を行う。この埋め込み表現は単語ごとに用意する必要があるため、語彙数×埋め込み

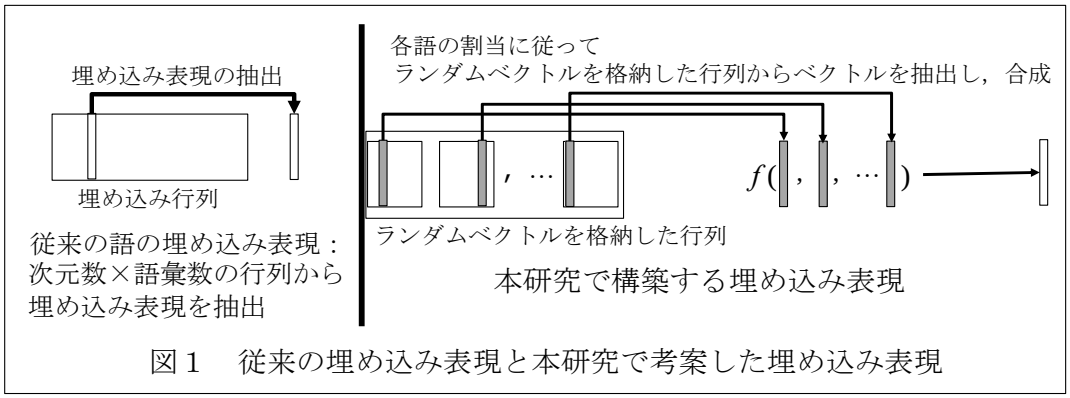

図１　従来の埋め込み表現と本研究で考案した埋め込み表現

表現の次元数の行列が必要となる。語彙数は多くの場合数万から数十万であるため、この埋め込み表現の行列はニューラルモデル内で最大の行列となることが多い。
本研究での提案手法と従来手法の概要を図 1 に示す。本研究で提案した手法では、従来のように、各単語に対応する埋め込み表現を束ねた行列を用意するのではなく、全単語で共有のランダムベクトルを複数用意し、これの組み合わせによって各単語の埋め込みを構築する。この手法において、ランダムベクトルの組み合わせが各単語によって固有であれば、各単語に固有のベクトルが構築されるため、従来の埋め込み表現の代替として利用できる。これにより、従来の、語彙数×次元数のパラメータを用意する手法に比べ、パラメータ数と使用するメモリ量を抑えることが可能となる。
従来手法との実験における比較として、表 1 に機械翻訳でのパラメータ数と性能を示した。機械翻訳での性能は BLEU という、モデルの出力と参照訳との一致率で評価する。この値は高いほど参照訳に近く、良い翻訳であるとされている。表 1 より、提案手法は BLEU の値を低下させることなく埋め込み表現のパラメータ数を減らすことに成功している事がわかる。従って、提案手法はパラメータ効率の良い埋め込み表現であると言える。本成果は機械学習に関する国際

手法	埋め込み表現のパラメータ	全パラメータ	BLEU
Transformer [Vaswani+ 17]	16.8M	60.9M	27.3
Transformer (re-run)	16.8M	60.9M	27.12
Transformer + DeFINE [Mehta+ 20]	-	68M	27.01
Transformer + 行列分解 [Lan+ 20]	8.5M	52.7M	26.56
Transformer + 提案手法	8.4M	52.5M	27.61

表1　機械翻訳での実験結果

会議である NeurIPS に採択された（主な研究成果リスト、その他の成果 1）。

「中間層のパラメータ効率化」

本研究では自然言語処理を行うニューラルモデルのうち、埋め込み層を除いたものを中間層と呼ぶ。中間層のパラメータ効率化に関する研究としては、1. パラメータ共有による効率化、2. パラメータ効率の良い構造の学習の安定化、3. 推論時に性能を引き上げる方策の 3 つに取り組んだ。以下に詳細を記す。

1. パラメータ共有による効率化：ニューラルモデルでは過剰なパラメータを抑制するために、パラメータ行列のうち形状が同一なものを共有する、という方策がしばしば採用される。例えば入力された単語列の次の単語を予測するという、言語モデルにおいては、入力、すなわち埋め込み層の次元数と中間層の出力次元数が同一な場合に、埋め込み行列と出力を計算するための行列の形状が同一となるため、共有が可能となる。

中間層のパラメータ共有については、すべての層で同一のパラメータを用いるという方策が最も多く採用されている。この手法は、実装は簡便であるが、各層の自由度を大幅に抑制するため、高い性能を達成するためには各層の次元数を大きな値に設定する必要がある。このため、パラメータ効率は高いが、計算時間に多大な悪影響がある。これを解消するため、全層でパラメータを共有するのではなく、N 層分のパラメータを用意し、これを各層に割り当てる方策を提案した。また、割り当て手法についても 3 つの手法を提示し、実験を通して比較を行った。実験を通して、提案手法は全層でパラメータを共有する手法と比べ、同一パラメータ数で同等以上の性能を達成し、また、学習や推論の速度も高速であることを示した（主な研究成果リスト、その他の成果 4）。

2. パラメータ効率の良い構造の学習の安定化：Transformer と呼ばれるニューラルモデルは機械翻訳のような系列変換タスクを効率的に処理可能なモデルとして 2017 年に発表されて以来、自然言語処理、画像処理、音声信号処理などの分野で活用されている。しかしながら、2017 年に発表された Transformer は 10 層以上に多層化して学習することが経験的に難しい構造であり、多層化にしても学習が安定する構造が主流となっていた。本研究では、最初に提案された、多層化の難しい構造の方が同パラメータ数で高い性能を達成する、すなわち、パラメータ効率の良いことを様々な実験を通して示し、また、多層化の難しい原因は勾配消失であること、層内に Residual Connection を 1 つ追加するだけで多層化が可能になることを示した。これにより、パラメータ効率の良い構造で多層化が可能になることを示し、特に、機械翻訳において、エンコーダ、デコーダのそれぞれを 100 層ずつ積み重ねたモデルの学習が可能になるこ

とを示した（主な研究成果リスト、その他の成果 2）。

3. 推論時に性能を引き上げる方策：ニューラルモデルを自然言語処理に適用する際、入力の文字列を予め単語で区切っておく必要がある。モデルの学習時に、単語の区切りを尤度に応じて変化させる、すなわち、様々な単語がモデルに入力されるようにすることでモデルの頑健性を向上させる、サブワード正則化という手法がある。サブワード正則化は訓練時にのみ使用され、推論時には最尤の区切りを採用する。本研究ではこの不一致を解消するため、推論時にも複数の区切りを尤度を元にサンプリングし、これらの入力から得られた出力の平均を最終的な出力とすることで、推論時のモデルの性能向上を実現した。本手法は、複数のモデルの出力を束ねることで単一のモデルよりも性能を向上させる、アンサンブルという手法を単一のモデルで行っているものとみなすことも可能であり、同一のパラメータ数で性能を引き上げている点からパラメータ効率の良い手法とも捉えられる。本研究は自然言語処理に関する国際会議である ACL Findings に採択された（主な研究成果リスト、その他の成果 3）。

「学習時間効率の良い正則化手法」

翻訳や要約を行うニューラルエンコーダ・デコーダモデルでは学習時は正解の単語系列のみが入力されるが、推論時には自らの予測した単語を入力として使用するため、一度誤った単語を出力してしまうと、その単語の入力をきっかけに、モデルの出力が正解から大きく逸脱していってしまう可能性がある。これを防ぐために、スケジュールドサンプリングという、学習時にモデルの出力を確率的に使用する手法が提案されている。また、モデルを頑健にする手法として、学習時にモデルの出力を毀損するようなノイズを入力に混入する、敵対的摂動という手法もある。このような手法はモデルを頑健にし、性能を向上させると主張されているが、学習時間に多大な悪影響を及ぼす点についてはほとんど議論されていない。例えば、スケジュールドサンプリングは出力系列の長さに依存した回数の予測が必要であり、敵対的摂動においては誤差逆伝播を用いて摂動を計算するため、学習時間は倍以上になってしまう。

このような状況を鑑み、これらの摂動手法は増加する学習時間に見合う利点があるのかについて、実験を通した調査を行った。上記の手法に加え、単語をランダムに別の単語に置換する手法、単語を類似度に基づいて別の単語に置換する手法、単語埋め込みをランダムにゼロベクトルに置換する手法を機械翻訳、要約、文法誤り訂正などの系列変換タスクに適用し、性能を比較したところ、図 2 に記したように、スケジュールドサンプリングと敵対的摂動は素朴に学習した際の性能に到達するまでの時間が長く、一方で、それ以外の手法は素朴に学習した際の性能に到達するまでの時間が短い、すなわち、スケジュールドサンプリングと敵対的摂動は学習時間に対して非効率であり、単語をランダムに別の単語に置換するような単純な

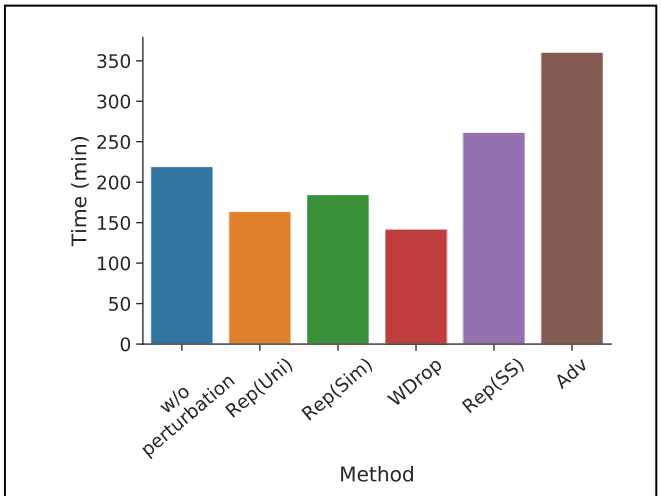

図2　ベースラインの性能に到達するまでの時間

手法のほうが効率的である、という結論が得られた。本研究は自然言語処理に関する国際会議である NAACL に採択された（主な研究成果リスト、その他の成果 2）。

3. 今後の展開

本研究では主に単語埋め込みや中間層のパラメータ効率化の研究に取り組んだ。これは近年の大規模な事前学習モデルのパラメータ数が指数的に増加していることを受けてのものであったが、本研究の成果を実際に事前学習モデルの構築や圧縮に適用できてはいない。研究機関間での GPU リソースの格差を考えると、高い性能を達成可能な事前学習モデルの圧縮は急務であると考えている。また、事前学習モデルの構築においては、既存研究で提案されている効率化手法が学習を不安定にするなどの理由で適用できなかったりするという報告がなされているが、これは経験的な知見であるため、取り組んでみなければどこに課題があるのか判然としない状況である。このため、できる限り速やかに事前学習モデルの構築に取り組みたい。

さらに、本研究ではモデルの効率に関して研究を行ったが、訓練データの効率も考慮する必要がある。訓練データをフィルタリングし、質の良い訓練データのみにした場合、1/10 のデータ数でも元の訓練データで学習した場合と同程度の性能が達成可能であるという報告がある。訓練データ内には機械翻訳や言語モデルで構築された文も多く含まれていると考えられるため、計算機の生成した文を識別し、フィルタリングする手法を探求する必要があると考えている。

このような、事前学習モデルに関する研究は、様々な企業の研究所もこぞって取り組んでいる課題であるため、成果が矢継ぎ早に出る状況となっているため、長く見積もっても 2 年以内には上記を達成できるような時間間隔で進める必要があると考えている。

4. 自己評価

本研究においての当初の目的である、パラメータ効率の良い単語埋め込み層は達成できた。また、パラメータ効率の良い中間層の研究にも取り組み、いくつかの成果を出すことができた点は評価できると考えている。一方で、本研究を大規模言語モデルの構築や、既存の大規模言語モデルの圧縮など、実応用には現状では適用できていない。本研究ではこれらを実現するための基盤技術が整理されたと考えており、今後、本研究の成果の事前学習モデルへの適用を通して、多くの研究者や末端のユーザーが利益を授与できるようになると考えている。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：0件

（2）特許出願

研究期間全出願件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. Sho Takase, Sosuke Kobayashi. All Word Embeddings from One Embedding. In Proceedings of the thirty-fourth Conference on Neural Information Processing Systems (NeurIPS), pp. 3775-3785, 2020.
2. Sho Takase, Shun Kiyono. Rethinking Perturbations in Encoder-Decoders for Fast Training. In Proceedings of 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2021), pp. 5767-5780, 2021.
3. Sho Takase, Tatsuya Hiraoka, Naoaki Okazaki. Single Model Ensemble for Subword Regularized Models in Low-Resource Machine Translation. In Findings of the Association for Computational Linguistics: ACL 2022. pp. 2536--2541, 2022.
4. Sho Takase, Shun Kiyono. Lessons on Parameter Sharing across Layers in Transformers. arXiv preprint arXiv:2104.06022. 2021.
5. Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. On layer normalizations and residual connections in transformers. arXiv preprint arXiv:2206.00330. 2022.
6. 言語処理学会第 28 回年次大会優秀賞. 2022.