様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １７ 日現在

機関番号：３２６７５
研究種目：基盤研究(C)（一般）
研究期間：2018～2023
課題番号：１８Ｋ１１４４９

研究課題名（和文）深層学習における内部状態の統計的手法による表現と新しい学習手法の構築

研究課題名（英文）Statistical Representation of Internal States of Depth Neural Networks and Exploration of New Learning Methods

研究代表者

柴田　千尋（Shibata, Chihiro）

法政大学・理工学部・准教授

研究者番号：００６３３２９９

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：深層学習のモデルが，内部でどのような処理を行っているのかを明らかにすることは，説明可能なAIへ向けてのアプローチの一つである．本研究では，実際に学習させたRNNやTransformerなどの深層学習モデルに対して，構文構造の観点から分析を行った．とくに形式言語モデルを用いて，どのような統語的(=構文的)特徴を学習することができ，それらがどのように内部ベクトルに表現されているかについて，一部明らかにし，追求を行った．また，内部表現以前の問題として，構文的には間違えているが僅かな違いしか持たないような敵対的なデータセットを用いて，深層学習モデルが本当に構文的な正誤を獲得できているのかを検証した．

研究成果の学術的意義や社会的意義
RNNやTransformer などの言語モデルがどの程度構文的な知識を獲得できるのか，また，獲得できるとすれば，それらがどのように埋め込まれるのか，言語モデルの理論に照らし合わせて追求することで，未だにブラックボックスである深層学習モデルの説明可能性に対して一定の方向性を示すことができたと考える．また，今後とも，形式言語クラスやそのアルゴリズム的学習の理論的な研究と，実際の産業で使われるような深層学習の分野との架け橋としての役割を果たしていきたい．

研究成果の概要（英文）：Tracing and extracting the internal representations of deep learning models is one of the approaches towards enhancing explainablity of AI. In this research, we analyzed deep learning models such as RNNs and Transformers, which were trained from various datasets, from the perspective of syntactic structures. Particularly, we used formal language models to explore what syntactic features can be learned and how these are represented within internal vectors. Additionally, to clarify underlying issues related to internal representations in advance, we employed adversarial datasets. Adversarial datasets contain syntactic errors but only minor differences. We verified whether deep learning models truly acquire the ability to discern syntactic correctness.

研究分野：機械学習

キーワード：深層学習(ディープラーニング)

様式Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

自然言語や運転操作の履歴など、記号化された時系列データにおいて、長短期記憶リカレントニューラルネットワーク(LSTM-RNN)や、ゲートリカレントニューラルネットワーク(GRU-RNN)などの RNN が、どのように、どの程度、長期の依存関係や、自然言語における構文情報を捉えるのかについて、近年様々な研究が行われている。しかし、実際に、内部表現にまで踏み込んで議論ができている研究例は我々の知る限りごく小数であることが現状である。近年、深層ニューラルネットワーク(Deep Neural Network:DNN)を単にブラックボックスとして使うだけでなく、実際にその内部でどのような挙動をしているのかについて、人間が理解できる形で提示すること(説明可能性)が求められている。畳み込みニューラルネットワーク(CNN)については、ある程度、その内部の挙動が理解されつつある。一方で、RNN においては、内部の挙動について、未だ説明可能な形で解明がなされていない。また、リカレント構造ではなく、代わりに注意機構を採用したモデルである Transformer は、LSTM などの既存の RNN よりも、ニューラルネットワークの構造として、学習結果の精度などの観点から優れていることが、近年の研究論文により広く知られている。Transformer モデルについても、説明可能性についての研究が盛んになってきている。

２．研究の目的

本研究の目的は、学習の結果、情報が深層ニューラルネット内にどのように表現されているかを探求するものである。上述のようなモデルに対するホワイトボックス化のために、英語における句構造をあらわす構文木を線形化し文として与えて学習させたときに、RNN 内で構文情報がどのようにエンコードされるかについて詳細に分析する。また、LSTM だけではなく、Transformer を視野に入れ、どのように文の構造が表現されるのかについて、探求を行う。Transformer は、任意の二つの離れた位置にある単語に対して、その信号の関係を計算(主にアテンションと呼ばれる機構)するレイヤーを積み重ねた構造をしているため、LSTM よりも直接的に、構文木が内部で表現されていることが知られている一方、理論的な観点からは、RNN のほうが表現できる言語クラスが大きいことが知られている。実際に、ある特殊な言語クラスに属する人工言語からサンプルした文の集合を訓練データとして学習したときに、RNN と Transformer では、どのような性質上の違いがあるのかについて、実験を通して探求する。

３．研究の方法

本研究では RNN や Transformer を対象とし、構文の構造がどのように表現されるかについて探索を行う。Transformer は、任意の二つの離れた単語間の信号関係を計算するレイヤー(主にアテンション機構と呼ばれる)を積み重ねた構造を持っており、一般的に言ってより優れた言語モデルを構築可能とされているが、理論的な観点からは、RNN が表現できる言語クラスがより大きいとされている。教師あり、教師なしを問わず、深層ニューラルネットの内部において、より解釈性が高く、かつ質の高い埋め込み表現を得るための試みを行う。文書分類を行う深層ニューラルネットに対し、超球面上への分布エンコーディング、および情報ボトルネック法を用い、内部状態の可読性が向上するかどうか、および、識別率が上昇するかどうかを検証する。また、LSTM(超短期記憶)の解釈性、また、質の高い埋め込み表現の追求を行う。一方で、Transformer の学習で十分優れた精度を得るためには、実際には非常に規模の大きい訓練データを直接用意するか、または、規模の大きい別のデータにおいて事前学習を行っておく必要がある。また、人工言語からサンプルしたデータは、単語の数が少ないことや、単語の分布が異なるなど、自然言語と性質的に異なることがあるが、それらを考慮しながら実験ベースで研究を進める。

４．研究成果

まず１つ目の成果として、その具体的な手法として、まず、隠れベクトルがクラス情報だけを保持するように、クラスごとにアンカーとなるような分布の中心ベクトルを用いる。一般に、深層ニューラルネットを用いた文書分類では、出力層に近い層において、各ラベルごとに、埋め込みベクトルのクラスタが存在することが知られている。本研究で用いる、分布の埋め込み表現と情報ボトルネックの目的は、ラベルベクトル(以降アンカーとよぶ)と文書の埋め込み表現とを陽に類似させるようにするという点、もう一つは、情報ボトルネック法つまり、内部ベクトル
相互情報量を最大化することにより、よりよい表現を得ることができる点である。とくに、分布埋め込み表現を超球面上に限定することで埋め込み表現の自由度を制限することにより、高精

度化と可解釈性の向上を図る手法を提案した。実験の結果、分布エンコーディングを行わず、代わりにドロップアウト法を用いるなどしたベースラインの手法と比較して、より識別性が高く、かつ、t-sne などで 2 次元にマップした際にもより解釈性の高い内部表現が学習の結果得られることを示した。文書分類という比較的容易なタスクに限定しているものの、実際にどのように表現されているかを可視化することが可能であった点、および分布エンコーディングや情報ボトルネック法によって、表現がどのように変化するかを明らかにした。また、超球面上の分布にエンコーディングする手法を新たに提案し、それにより分類精度の高精度化および、可解釈性の向上が可能となった。

次に、LSTM(超短期記憶)の解釈性、また、質の高い埋め込み表現の追求に付いてだが、LSTM が、どのように文脈を捉えるのか、LSTM 内のどの中間ベクトルに長期的な情報がどのようなメカニズムで捉えられるのかについて、追求を行った。その結果、品詞が状態更新ベクトルの空間中に最も顕著に獲得されていることや、”that ”などの複数の意味を持つ機能語が、意味ごとにクラスタリングされていることなどが、実験を通して発見された。また、LSTM 内部の、状態更新ベクトルにおける各要素のアクティベーションの値の学習前と学習後のヒストグラムの変化から、学習によって、量子化が進むことを示した。これは、LSTM 内部の関連する重みパラメータが、学習により、より大きい分散を持つことになったことを意味している。より大きな分散を持つようになった理由としては、学習時にモデルを更新する際に、シグモイド関数や tanh 関数のもつ性質上、自然にそのようになるためで、この仕組が、学習が進んだときのロスや内部ベクトルの安定的な挙動にも寄与しているものと予想される。以上の結果は、LSTM を翻訳モデルなどに用いると、構文的な誤りが少ないことがよく知られているが、実際に、動詞や名詞などの、単語の持つ構文に関する基本的な情報が、状態更新ベクトルの一部の部分空間で、自然に獲得されているということを意味する。

最後に、複雑さの階層をもつサブレギュラー言語クラス群を対象にした適切な訓練データを用いて Transformer の性質の検証を行った。また、実際にサブレギュラー言語クラスに相当するデータセットに対して Transformer モデルの学習を行い、どの程度の大きさの Transformer で学習できるかの検証を行った。また、RNN との比較を行った。その結果、RNN と Transformer モデルでは、データセットのターゲットとなっている言語クラスに入るかどうかという binary classification の学習においては、ほとんど精度上の変化が出ないことがわかった。一方で、causal language model(ある時刻 t までの記号列(prefix)を入力し、次の時刻 t+1 の記号の生成される確率を学習するモデル)では、Transformer モデルのほうが優れていることがわかった。このことから、形式言語のクラスのような、理論的な観点から作成された人工的な時系列データにおいても、Transformer モデルのほうが有効であることがわかった。一方、Transformer モデルでは、positional embedding があるなど、RNN のような記号列の情報を固定長ベクトルに集約するタイプのモデルではないことから、例えばオートマトンの状態と RNN の内部状態の対応関係のように、埋め込み表現との直接的な比較で分析することは難しいため、Transformer の attention 機構に特化した分析の方法に関して、フューチャーワークとして今後研究を進めたい。

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　4件／うち国際共著　2件／うちオープンアクセス　5件）

１．著者名 Sam van der Poel, Dakotah Lambert, Kalina Kostyszyn, Tiantian Gao, Rahul Verma, Derek Andersen, Joanne Chau, Emily Peterson, Cody St. Clair, Paul Fodor, Chihiro Shibata, Jeffrey Heinz	４．巻 2
２．論文標題 MLRegTest: A Benchmark for the Machine Learning of Regular Languages	５．発行年 2023年
３．雑誌名 arXiv	６．最初と最後の頁 1-38
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Shibata Chihiro	４．巻 113
２．論文標題 Learning (k,l)-context-sensitive probabilistic grammars with nonparametric Bayesian approach	５．発行年 2021年
３．雑誌名 Machine Learning	６．最初と最後の頁 3267-3301
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10994-021-06034-2	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Shibata Chihiro、Uchiumi Kei、Mochihashi Daichi	４．巻 -
２．論文標題 How LSTM Encodes Syntax: Exploring Context Vectors and Semi-Quantization on Natural Text	５．発行年 2020年
３．雑誌名 Proceedings of the 28th International Conference on Computational Linguistics	６．最初と最後の頁 4033-4043
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2020.coling-main.356	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 守屋俊，　柴田千尋	４．巻 J103-D(4)
２．論文標題 文字レベル畳み込みニューラルネットに対するトピック分布を用いた事前学習	５．発行年 2020年
３．雑誌名 電子情報通信学会論文誌	６．最初と最後の頁 280-290
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2019PDP0004	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Chihiro Shibata and Jeffrey Heinz	4．巻 -
2．論文標題 Maximum Likelihood Estimation of Factored Regular Deterministic Stochastic Languages	5．発行年 2019年
3．雑誌名 In Proceedings of the 16th Meeting on the Mathematics of Language	6．最初と最後の頁 102-113
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/W19-5709	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 守屋俊，町田秀輔，柴田千尋	4．巻 -
2．論文標題 超球面上への分布エンコーディングを用いた文書分類	5．発行年 2020年
3．雑誌名 言語処理学会第26回年次大会予稿集	6．最初と最後の頁 1435-1438
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 岡本(柴田) 千尋，内海 慶，持橋 大地
2．発表標題 構文情報を陽に与えたときの LSTM-RNN による内部表現について
3．学会等名 第237回自然言語処理研究会　2018年9月26日　情報処理学会
4．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	持橋　大地 (Mochihashi Daichi) (80418508)	統計数理研究所・数理・推論研究系・准教授 (62603)	

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	吉仲　亮 (Yoshinaka Ryo) (80466424)	東北大学・情報科学研究科・准教授 (11301)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関