Vol.2021-AVM-112 No.2
2021/2/26

電子透かしを用いた RNN 学習モデルの保護

松本幸大 [1,a)]　酒澤茂之 [2]

概要：ディープラーニングにおいてコストをかけて生成されるモデルは資産として捉えることができる．電子透かしをモデルに対して埋め込み，権利保護を行う技術が注目されている．本研究では再帰型ニューラルネットワーク（RNN）を対象として，モデル学習中に電子透かしを埋め込む．学習中に生じるタスクロスに加えて，透かしを埋め込むためのロスを定義し，電子透かしを埋め込むものとする．実験では，LSTM ネットワークで生成するモデルに対して電子透かしの埋め込みを行い，検出を行った結果，LTSM モデルに対する電子透かしの埋め込みが可能であることが示された．今後の課題として，電子透かし埋め込み時のパラメータが本来のタスクに及ぼす影響を分析し，それを踏まえてモデル精度への影響を軽減しながら電子透かしを埋め込む手法の確立が挙げられる．

キーワード：RNN，　LSTM，　電子透かし

1. はじめに

研究やビジネスにおいてデータというものは重要な資産の一つとして挙げられる．近年では膨大な時間やコストを投入してディープラーニングの学習モデル構築が行われている．そのように時間やコストをかけて構築される学習モデルも重要な資産だと捉えることができる．ディープラーニングでは転移学習やファインチューニングと呼ばれる手法を用いることで，大規模学習モデルをベースに，少しのデータかつ短時間で高精度なモデルを構築可能になる．コストを掛け，構築したモデルから簡単に高精度なモデルを生成できてしまうため，学習モデルの不正利用が考えられる．そのため本研究では学習モデルに対して電子透 かしを埋め込み，保護を行うことを目的とする．既存研究では CNN モデルに対しての電子透かしの埋め込みが研究されているが，構造の異なる RNN モデルに対しての電子透かしの埋め込みに関する検証が少ない．本論文では構造の異なる RNN モデルに対して電子透かしを施した際にどのような影響がみられるのか検証，考察を行う．

2. 関連研究

現在までにディープラーニングの学習モデルに対して電子透かしを埋め込む研究が行われている．その代表例は，内田らによる CNN モデルに対して電子透かしを埋め込む[1]という研究である．この方式では，学習モデルの重み係数に対して電子透かしを埋め込む手法が用いられている．

1 大阪工業大学大学院
Osaka Institute of Technology., Hirakata, Osaka 573-0196, Japan
2 大阪工業大学
Osaka Institute of Technology., Hirakata, Osaka 573-0196, Japan
a) m1m20a41@oit.ac.jp

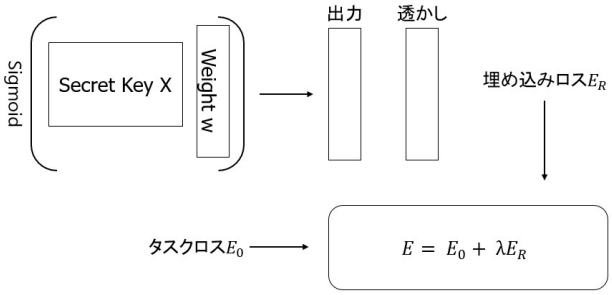

図 1 電子透かしの埋め込み

図 1 の Weight w に対して電子透かしを埋め込むために電子透かし行列 Secret Key X と掛け合わせ，計算処理を行っている．その計算の結果得られる値を電子透かしビットとし，その値が特定の値になるよう本来の学習タスクロスに埋め込み用のタスクロスとして加える．このように学習中に埋め込みを行うという方式が用いられている．また，RNN の学習モデルの保護に関する研究[2]も行われているがモデルに対して電子透かし等を埋め込んでいるのではなくトリガーとなる特定のワードそのものを学習させることで，モデルの所有がどこにあるのかを証明するものとなっている．この方式では，RNN のモデルに対して保護を行うことができるが，ファインチューニングに対する保護残存性に対しての対策は行われていない．

3. 実験

3.1 実験環境

本研究の実験環境について述べる．使用した言語は Python（v3.6）である．また主なライブラリとして Tensorflow（v2.1.0）,Keras（v2.3.1）,CUDA（v8.0）,cuDNN（v5.1.10）を用いる．また，モデルは RNN の一種である LSTM を用いて構築する．

ⓒ2021 Information Processing Society of Japan

3.2 実験概要

実験ではまず電子透かしの埋め込みが可能か検証を行う．そのために単純なタスクで学習モデルを構築し，電子透かし埋め込み可能性および埋め込んだ際の影響を確認する．また，より複雑なタスクの学習モデルに対して電子透かしの埋め込みが可能であるか，またその際の影響についても検証を行う．

3.3 実験方法

ディープラーニングでは大量のデータを用いて学習を行うが，学習データに対して過度に適合しすぎてしまう過学習という状態に陥る場合がある．与えられた学習データに対しての誤差は小さくなるが，学習データに対して過度な適合しているため未知のデータに対しては適切な値を出せないようなモデルとなってしまい，汎化能力が低くなる．また，学習データが十分に足りない場合にも過学習を引き起こしてしまう可能性もある．その過学習を防ぐために正則化という手法が存在する．正則化は学習を行う際にモデル中の重みの絶対値が大きくなりすぎないようにペナルティを設ける．そのペナルティをタスクのロスに加える，その時の値が最小になるように 学習を進めることで過学習を防ぎつつ汎化能力を高めることが可能となる．本研究において電子透かしの埋め込みは正則化の部分で埋め込みを行う．

図 1 が埋め込みを行う際の流れとなる．電子透かしを埋め込むためのロスについて説明する．まず，電子透かしの秘密鍵である行列 Secret Key X をニューラルネットワーク内の重み係数ベクトルと掛け合わせて計算を行い，その結果得られる値を電子透かしビットと定義する．その値が埋め込みたい電子透かし（0 or 1）の特定の値になるように，バイナリ交差エントロピーにより求められる値を埋め込みロス E_R とし，本来の学習タスクロス E_0 に加算している．

実験では，埋め込む電子透かしの値を 1 とし，図 1 にもある通り掛け合わせる Secret Key を乱数で生成し実験を行う．RNN では再帰的な構造を持っているため学習の入力に加えて，その直前の状態に依存し学習を行う．そのため，ネットワークには現在の入力と直前の状態の入力の 2 つが入力として与えられる．実験ではそれら 2 つの入力に対しての重みを用いて電子透かしの埋め込みを行い，現在の入力に対する重みを kernel weight，直前の状態の入力を recurrent weight として本論文では論じる．

埋め込む対象の学習モデルは電子透かしの埋め込みが可能かを検証するために，sin 波を予測するタスクの単純なモデルを用いる．sin 波の予測を行う学習モデルでは，20 個の LSTM セルを持ち，kernel weight 80 個，recurrent weight 160 個に対して電子透かしの埋め込みを行う．また，電子透かしの埋め込みを確認後，sin 波予測とはタスクの異なる文章生成を行う学習モデルに対しても電子透かしの埋め込みを行い，どのような影響があるのかを検証する．文章生成を行うモデルでは学習用のデータとして，老人と海[3]を用いる．こちらの学習モデルでは 128 個の LSTM セルを持ち，kernel weight 781,312 個，recurrent weight 65,536 個に対して電子透かしの埋め込みを行う．また，電子透かしの埋め込みビット数を変化させた場合にどの程度差が生まれるのかの検証も行う．埋め込みビット数は 1 ビット，8 ビット，256 ビットの 3 パターンでの実験を行う．

3.4 実験結果と考察

sin 波の予測を行うシンプルな学習モデルに対しての実験結果を述べる．予測用に sin 波の入力データを引き渡した後，モデルの予測結果を出力する．次の表 1 に結果を示す．

表 1 sin 波予測モデルに対する実験結果

対象の重み	埋め込み可能性
kernel weight	可能
recurrent weight	可能
上記両方	可能

sin 波予測のモデルに対して電子透かしが全て 1 の場合，乱数生成の Secret Key を用いての埋め込みが可能であることが示された．次に電子透かし埋め込みの有無で本来のタスクの精度に影響が見られるか比較を行う．

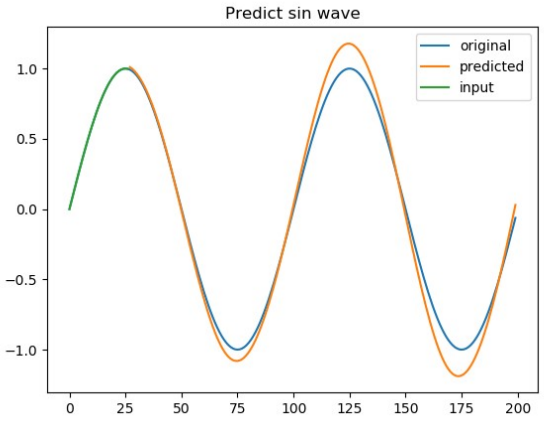

図 2 sin 波の予測(電子透かしなし)

ⓒ2021 Information Processing Society of Japan

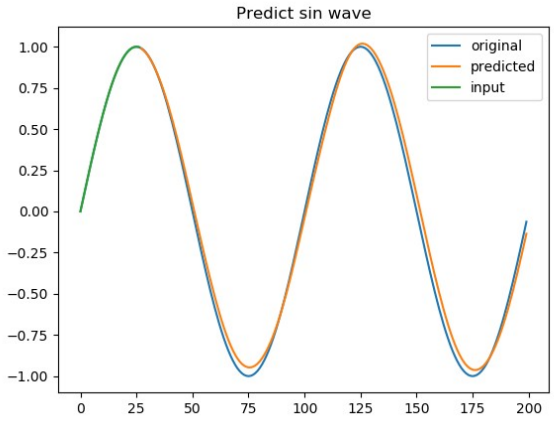

図 3 sin 波の予測(kernel weight から透かし生成)

表 2 sin 波の予測における最終的な training loss の値

正則化部分	標準の正則化関数	本研究の正則化関数
kernel weight	0.0042	6.4396e-05
recurrent weight	0.0046	7.0320e-05
上記両方	0.0102	8.8131e-05

図 2 は電子透かしを埋め込まず，L2 正則化という手法を用いて学習を行ったモデルの予測結果である．過学習を抑制し，一定の予測精度を保てていると考えられる．一方で図 3 は kernel の重みを用いて電子透かしを埋め込んだ学習モデルである．電子透かし埋め込みは確認できたが，図の sin 波予測には差がないと考えられるが，表 2 から L2 正則化を用いた場合よりも損失の値が極端に小さくなっていることが確認できる．電子透かしを埋め込む場合に用いた正則化手法において過学習を引き起こしてしまうことが判明した．

学習タスクが単純であること，学習に用いるデータの数が少ないことが過学習を引き起こしている原因になっていると考えられる．続いて文章生成を行う学習モデルに対して埋め込んだ場合について述べる．

彼は老いていた．小さな船で老人の手を見ていくれる，老人は，少年はずつ浮かり出して，ロープが両手を左手と言ってると，見えないのを何のこともないだ」「その時，魚が，イスまままでになく，俺のことは何う切れそうというしれないよ」「あいが，いあできたので，左手でたなシイラの大日にはずがぶら，その時まあ，水を打ちのめされた．それかオやりの漁師たちはと，いるの俺が人間をて『くのらいく，もからその．またわんだ水らからないと，じゃに出なかった．

図 4 生成された文章（一部）

例として，生成された文章を図 4 に示す．「彼は老いていた．」という文章から自動で文章を生成していくような学習モデルとなっている．

表 3 文章生成モデルに対する実験結果

対象の重み	埋め込み可能性
kernel weight	可能
recurrent weight	可能
上記両方	可能

表 4 文章生成モデルでの最終的な training loss の値

正則化部分	標準の正則化関数	本研究の正則化関数
kernel weight	4.2154	1.0497
recurrent weight	2.1235	1.0497
上記両方	5.02	1.1173
正則化なし	0.7489	

表 4 の training loss の値から，正則化をしない場合よりも過学習を抑制しながら電子透かしの埋め込みが可能であると考えられる．また，極端なタスクロスの低下も見られないため，正則化を行わない場合よりも過学習を抑制しながらモデル構築が可能だと考えることができる．

続いて学習モデルの文章生成精度について述べる．文章生成の分野でのモデルの精度というものは，Perplexity という評価指標[4]によって評価可能である．Perplexity は分岐の数または選択肢の数を表し，確率の逆数で定義される．この値が大きい場合は単語の特定などが難しくなり複雑なものになるが，逆にこの値が小さいと単語の特定や候補数が少なくなるため予測性能が高いと考えることができる．実験では Perplexity の変化を精度の指標として用いる．

表 5 埋め込みビット数別 Perplexity

対象の重み	1bit	8bit	256bit
kernel weight	2.9086	3.1555	3.1012
recurrent weight	3.3321	2.9338	3.4150
上記両方	3.1232	2.9145	3.1977

表 5 は電子透かしを埋め込んだ場合の Perplexity の値および電子透かしの埋め込みを行わず正則化もしなかった場合の Perplexity の値である．微小な差異はあるものの，電子透かしを埋め込んだ場合でも精度に大きな影響を及ぼしていないことが検証された．

ⓒ2021 Information Processing Society of Japan

4. むすび

本論文では，RNN 学習モデルに対して電子透かしを埋め込みその有用性および本来のタスクに対してどのような影響を及ぼすのか実験を行い，結果とその考察を論じた．実験では，電子透かしの埋め込みが可能であるかを検証するためにシンプルな sin 波の予測モデルを用い，さらに複雑なモデルに対して電子透かしの埋め込みを行った場合の影響を検証するために文章生成モデルを用いた．実験の結果, RNN 学習モデルがシンプルな場合でも複雑な場合でも電子透かし埋め込みが可能であると検証された．また，電子透かしの埋め込みビット数を変更した場合でも，電子透かしの埋め込み自体に大きな差は出ないこと，また学習モデルの本来のタスクに対する影響が少ないことも検証された．特に文章生成を行う学習モデルの kernel 部と recurrent 部の重み係数群の個数は大きく異なるが，いずれの場合でも電子透かしの埋め込みを行うことが可能だと検証された．

既存研究では CNN モデルに対して埋め込みを行っているが，本研究ではネットワーク構造の異なる RNN のモデルに対しての電子透かしの埋め込みを行った．再帰的構造を持っているが，その再帰部の重み係数を利用した電子透かしの埋め込みが可能であることが実験の結果として得られた．

モデルのタスク精度を保ちながら電子透かしの埋め込みを行うためには，正則化を行う部分のネットワーク構造特性を明らかにすることが必要となるだろう．また，電子透かしの埋め込みビット数を増やした際の影響や電子透かしに対する攻撃耐性の検証も行う必要がある．

謝辞

本研究は JSPS 科研費 JP18K11309 の助成を受けたものである．

参考文献

[1] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin’ichi Satoh. Embedding Watermarks into Deep Neural Networks. *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pp. 269–277, June 2017.

[2] Jiazhu Dai and Chuanshuai Chen. A backdoor attack against LSTM-based text classification systems. *arXiv:1905.12457 [cs]*, June 2019.

[3] Hemingway, Ernest Miller. 老人と海. Charles Scribner’s Sons, 1952.

[4] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L. Mercer. An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics*, Vol. 18, No. 1, pp. 31–40, 1992.

ⓒ2021 Information Processing Society of Japan