

Attention ゲート付き xLSTM

日大生産工 ○石原 嵩大 日大生産工 山内 ゆかり

1. まえがき

ロング・ショートターム・メモリ (LSTM) [1]のアイデアは、リカレントニューラルネットワークの勾配消失問題を克服するために導入された。定常誤差カルーセルは、シグモイドゲートによって制御されるセル入力によるセル状態 c_{t-1} の加法更新である。入力ゲートと忘却ゲートはこの更新を制御し、出力ゲートは隠れ状態を制御する。セル状態は正規化または圧縮されることによって出力ゲートを通して隠れ状態が得られる。LSTMは様々な分野で成功を収め、2017年のTransformerの登場まで、テキスト生成などのシーケンス関連タスクで主流であった。LSTMの効果は、テキスト生成、手書き生成、シーケンス間翻訳、プログラム評価、画像キャプション生成、ソースコード生成、降雨流出モデル、洪水警報のための水文学モデルなどで実証されている。LSTMは抽象化を学ぶのに優れており、意味情報を巧みに抽出しメモリセルに保存することができる。これは、数値や構文のニューロン、言語ニューロン、感情ニューロンなどによって明らかになっている。

2. 従来研究

LSTMの限界を克服するために、xLSTM[2]はLSTMのアイデアに2つの主要な修正を行う。スカラーメモリ、スカラー更新、メモリミキシングを持つ新しいsLSTMと、行列メモリと共分散更新ルールを持ち、完全に並列化可能な新しいmLSTMである。sLSTMとmLSTMはともに指数関数ゲーティングによってLSTMを強化する。並列化を可能にするために、mLSTMはメモリ混合を放棄している。mLSTMもsLSTMも複数のメモリセルに拡張することができ、sLSTMはセル間のメモリ混合を特徴とする。さらに、sLSMTは複数のヘッドを持つことができ、ヘッド間のメモリ混合はない。このようにsLSTMにヘッドを導入することで、エクスポネンシャル・ゲーティングとともに新しいメモリ混合の方法が確立される。mLSTMでは複数のヘッドと複数のセルは等価である。

これらの新しいLSTM変種を残余ブロックモジュールに統合すると、xLSTMブロックができる。これらのxLSTMブロックをアーキテクチャ

残留的に積み重ねることで、xLSTMアーキテクチャが得られる。

2.1. sLSTM

sLSTMに記憶決定を修正する能力を持たせるために、正規化と安定化とともに指数ゲートを導入する。特に、入力ゲートと忘却ゲートは指数活性化関数を持つことができる。正規化のために、入力ゲートと将来のすべての忘却ゲートの積を合計する正規化状態を導入する。sLSTMの順伝播経路は以下の通りである。

$$c_t = f_t c_{t-1} + i_t z_t \quad (1)$$

$$n_t = f_t n_{t-1} + i_t \quad (2)$$

$$h_t = o_t h'_t, h'_t = c_t / n_t \quad (3)$$

$$z_t = \varphi(z'_t), z'_t = w^T_z x_t + r_z h_{t-1} + b_z \quad (4)$$

$$i_t = \exp(i'_t), i'_t = w^T_i x_t + r_i h_{t-1} + b_i \quad (5)$$

$$f_t = \sigma(f'_t) \text{ OR } \exp(f'_t), \\ f'_t = w^T_f x_t + r_f h_{t-1} + b_f \quad (6)$$

$$o_t = \sigma(o'_t), o'_t = w^T_o x_t + r_o h_{t-1} + b_o \quad (7)$$

指数活性化関数は、大きな値を取り、オーバーフローする。そこで、追加状態 m_t を用いてゲートを安定化させる。

$$m_t = \max(\log(f_t) + m_{t-1}, \log(i_t)) \quad (8)$$

$$i'_t = \exp(\log(i_t) - m_t) = \exp(i'_t - m_t) \quad (9)$$

$$f'_t = \exp(\log(f_t) + m_{t-1} - m_t) \quad (10)$$

Attention-Gated xLSTM

Takao ISHIHARA and Yukari YAMAUCHI

sLSTMは元のLSTMのように複数のメモリセルを持つことができる。複数のメモリセルは、隠れ状態ベクトル h からメモリセル入力 z とゲート i, f, o へのリカレント接続 R_z, R_i, R_f, R_o を介したメモリ混合を可能にする。新しいメモリミキシングにおける側面は、指数関数的ゲーティングの効果である。新しいsLSTMは複数の各ヘッド内ではメモリが混合するが、ヘッド間では混合しない。sLSTMにヘッドを導入し、指数関数的ゲーティングを行うことで、新しいメモリミキシングの方法が確立される。また、以下に各値の勾配を示す。

$$\delta_{h_t} = \delta_{h_t}^{ext} + \delta_{h_t}^R \quad (11)$$

$$\delta_{c_{t-1}} = f_t \circ \delta_{c_t} + o_{t-1} \circ n_{t-1}^{-1} \circ \delta_{h_{t-1}} \quad (12)$$

$$\delta_{n_{t-1}} = f_t \circ \delta_{n_t} - o_t \circ c_{t-1} \circ n_{t-1}^{-2} \circ \delta_{h_{t-1}} \quad (13)$$

$$\delta_{f_t} = f'_t \circ c_{t-1} \circ \delta_{c_t} + f'_t \circ n_{t-1} \circ \delta_{n_t} \quad (14)$$

$$\delta_{i_t} = i'_t \circ z_t \circ \delta_{c_t} + i'_t \circ \delta_{n_t} \quad (15)$$

$$\delta_{\tilde{z}_t} = i_t \circ \varphi'(\tilde{z}_t) \circ \delta_{c_t} \quad (16)$$

$$\delta_{o_t} = o'_t \circ c_t \circ n_t^{-1} \circ \delta_{h_t} \quad (17)$$

$$\delta_{\tilde{x}_t} = \sum_{g \in \{f, i, z, o\}} W_g^\top \delta_{\tilde{g}_t} \quad (18)$$

$$\delta_{h_{t-1}}^R = \sum_{g \in \{f, i, z, o\}} R_g^\top \delta_{\tilde{g}_t} \quad (19)$$

$$\delta_{R_g}^\top = \sum_t h_{t-1} \delta_{\tilde{g}_t}^\top \quad (20)$$

$$\delta_{W_g}^\top = \sum_t x_t \delta_{\tilde{g}_t}^\top \quad (21)$$

2.2. mLSTM

LSTMの記憶容量を高めるために、LSTMのメモリセルをスカラー $c \in \mathbb{R}$ から行列 $c \in \mathbb{R}^{d \times d}$ に増やす。時刻 t において、一対のベクトル $k_t \in \mathbb{R}^d$ と値 $v_t \in \mathbb{R}^d$ を格納したい。その後、時刻 $t + \tau$ において、問い合わせベクトル $q_{t+\tau} \in \mathbb{R}^d$ によって値 v_t を検索する必要がある。これは双方向連想記憶の設定である。キーと値のペアを格納するための共分散更新規則は以下の通りである。

$$C_t = C_{t-1} + v_t k_t^\top \quad (22)$$

入力をキーと値に射影する前にレイヤー・ノーマを仮定する。共分散更新規則は、検索された2値ベクトルの分離性が最大になるように最適化される。検索をペアワイズ相互作用に限定し、二次的複雑性を認めると、より高い分離可能性が可能になる。共分散更新ルールは $C_{t-1}, v_t k_t$ に乗じた一定の減衰率と一定の学習率を備えている。この精神に基づき、共分散更新ルールをLSTMに統合する。忘却ゲートは減衰率、入力ゲートは学習率に対応する。一方、出力ゲートは検索されたベクトルをスケールする。このマトリックス・メモリーでは、正規化状態はキー・ベクトルの加重和であり、各キー・ベクトルは入力ゲートと将来のすべての忘却ゲートによって加重される。ここでも、正規化状態はゲートの強さの記録をしている。クエリと正規化状態間のドット積はゼロに近くなる可能性があるため、このドット積の絶対値を使用し、以前と同様に閾値で下界を設定する。mLSTMの順伝播経路は以下の通りになる。

$$C_t = f_t C_{t-1} + i_t v_t k_t^\top \quad (23)$$

$$n_t = f_t n_{t-1} + i_t k_t \quad (24)$$

$$h_t = o_t \odot h'_t, \quad h'_t = C_t q_t / \max\{|n_t^\top q_t|, 1\} \quad (25)$$

$$q_t = W_q x_t + b_q \quad (26)$$

$$k_t = \frac{1}{\sqrt{d}} W_k x_t + b_k \quad (27)$$

$$v_t = W_v x_t + b_v \quad (28)$$

$$i_t = \exp(i'_t), \quad i'_t = w^\top_i x_t + b_i \quad (29)$$

$$f_t = \sigma(f'_t) \text{OR} \exp(f'_t), \quad f'_t = w^\top_f x_t + b_f \quad (30)$$

$$f_t = \sigma(f'_t) \text{OR} \exp(f'_t), \quad f'_t = w^\top_f x_t + b_f \quad (31)$$

mLSTMは元のLSTMのように複数のメモリセルを持つことができる。mLSTMの場合、メモリが混在しないので、複数のヘッドと複数のセルは等価である。mLSTMの指数のゲートを安

定化させるために、sLSTMと同じ安定化技術を使用する。mLSTMにはメモリ混合がないため、この再帰は並列バージョンで再定式化できる。以下にmLSTMの各値の勾配を示す。

$$\delta_{\tilde{h}_t}^\top = o_t \circ \delta_{h_t}^{ext} \quad (32)$$

$$\delta_{c_{t-1}} = f_t \delta_{c_t} + \frac{q_{t-1} \delta_{\tilde{h}_{t-1}}^\top}{\max\{|n_{t-1}^\top q_{t-1}|, 1\}} \quad (33)$$

$$\begin{aligned} \delta_{n_{t-1}} &= f_t \delta_{n_t} \\ &- \frac{q_{t-1}^\top C_{t-1}^\top \delta_{\tilde{h}_{t-1}}}{\max\{|n_{t-1}^\top q_{t-1}|, 1\}^2} \Omega(n_{t-1}^\top q_{t-1}) q_{t-1} \end{aligned} \quad (34)$$

$$\delta_{v_t}^\top = i_t k_t^\top \delta_{c_t}^\top \quad (35)$$

$$\delta_{k_t}^\top = i_t (v_t^\top \delta_{c_t} + \delta_{n_t}^\top) \quad (36)$$

$$\begin{aligned} \delta_{q_t}^\top &= \frac{C_t^\top \delta_{\tilde{h}_t}^\top}{\max\{|n_t^\top q_t|, 1\}} \\ &- \frac{q_t^\top C_t^\top \delta_{\tilde{h}_t}^\top}{\max\{|n_t^\top q_t|, 1\}^2} \Omega(n_t^\top q_t) n_t \end{aligned} \quad (37)$$

$$\delta_{x_t} = \sum_{g \in \{q, k, v\}} W_g^\top \delta_{g_t}^\top \quad (38)$$

$$\delta_{W_g}^\top = \sum_t x_t \delta_{g_t}^\top \quad (39)$$

$$\delta_{b_g} = \sum_t \delta_{g_t} \quad (40)$$

$$\begin{aligned} \delta_{\tilde{f}_t} &= (1^\top (C_{t-1} \circ \delta_{c_{t-1}}) 1 \\ &+ 1^\top (n_{t-1} \circ \delta_{n_t})) \gamma(\tilde{f}_t) \end{aligned} \quad (41)$$

$$\begin{aligned} \delta_{\tilde{i}_t} &= (1^\top ((v_t k_t^\top) \circ \delta_{c_{t-1}}) 1 + 1^\top (n_{t-1} \\ &\circ \delta_{n_t})) \exp(\tilde{f}_t) \end{aligned} \quad (42)$$

$$\delta_{\tilde{o}_t} = \tilde{h}_t \circ \sigma'(\tilde{o}_t) \circ \delta_{h_t} \quad (43)$$

2.3. xLSTMアーキテクチャ

xLSTMブロックは高次元空間で過去を非線形に要約し、異なる履歴やコンテキストをよりよく分離する必要がある。履歴を分離することは、次のトークンなど次のシーケンス要素を正しく予測するための前提条件である。高次元空間では、非線形に埋め込まれたパターンは元の空間よりも線形に分離できる可能性が高いというCoverの定理に頼る。ここで、二つの残差ブロック・アーキテクチャを考える。(1)逆投影後の残差ブロックで、元の空間で過去を非線形に要約し、次に高次元空間に線形マップし、非線形活性化関数を適用し、元の空間に線形マップする。(2)アッププロジェクション前の残差ブロックは、高次元空間に線形マップし、高次元空間の過去を非線形に要約し、元の空間に線形にマップする。sLSTMを含むxLSTMブロックでは、ほとんどの場合ポストアッププロジェクションブロックを使用する。xLSTMアーキテクチャは残差ブロックを積み重ねることで構築される。

3. 提案手法

sLSTM と mLSTM のゲートに Attention 機構[3]を組み込み、ゲートがどのタイムステップの情報に注目すべきかを動的に調整する仕組みを導入する。これにより、さらに長期的な依存関係をとらえることができるようになると考えられる。そのためにまず、クエリ、キー、バリューを計算する。次にクエリとキーを用いてスコアを計算し、どのゲートアクティベーションが重要かを判断する。そのスコアをソフトマックス関数で正規化することで各ゲートアクティベーションに対する重み付けを決定する。この重み付けされたゲートアクティベーションの和を計算し、最終的なコンテキストベクトルを計算する。最後にこのコンテキストベクトルを用いて各ゲートの出力を調整する。以下の式では、 $g \in \{i, f, o\}$ とする。

$$q_g = W_q a_g \quad (44)$$

$$k_g = W_k a_g \quad (45)$$

$$v_g = W_v a_g \quad (46)$$

$$s_t = \frac{q_g \cdot k_g^\top}{\sqrt{d_k}} \quad (47)$$

$$\alpha_g = \frac{\exp(s_t)}{\sum_j \exp(s_j)} \quad (48)$$

$$c_g = \sum_g \alpha_g v_g \quad (49)$$

$$i_t^{att} = i_t \cdot c_g \quad (50)$$

$$f_t^{att} = f_t \cdot c_g \quad (51)$$

$$o_t^{att} = o_t \cdot c_g \quad (52)$$

以下が逆伝播に用いる各値の勾配となっている。

$$\delta_{w_v} = \delta g^{att} \cdot \alpha_g \cdot h_t^\top \quad (53)$$

$$\delta_{s_t} = \delta \alpha_g \cdot \alpha(1 - \alpha) \quad (54)$$

$$\delta_{w_q} = \delta_{s_t} \cdot \frac{k_g^\top}{\sqrt{d_k}} \cdot h_t^\top \quad (55)$$

$$\delta_{w_k} = \delta_{s_t} \cdot \frac{q_g^\top}{\sqrt{d_k}} \cdot h_t^\top \quad (56)$$

4. 実験および検討

データセットを用いて、いくつかの手法で二乗和誤差を比較し、評価する。実験に用いるデータセットである株式会社レーザーテックの株価を以下のFigure 1に示す。

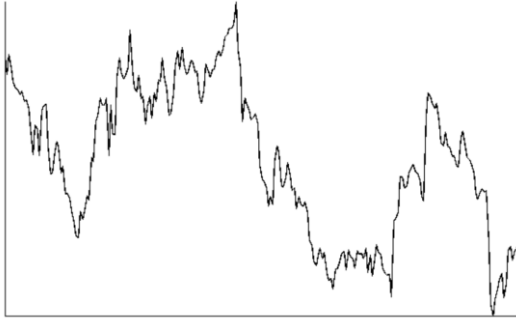


Figure 1 株式会社レーザーテックの株価

5. まとめ

xLSTMの長所である長期依存関係を学習することができるという点に焦点をあて、それをさらに引き延ばすため、Attention機構を組み込むことを提案した。これにより、従来のxLSTMのゲート機構と組み合わせると相乗効果が得られると考えられる。

参考文献

- [1] Hochreiter.S, and Schmidhuber.J, “Long Short-term Memory” *Neural Computation* , 9(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735 . (1997).
- [2] Beck.M, and Poppel.K, and Spanring.M, and Auer.A, and Prudnikova.O, and Kopp.M, and Klambauer.G, and Brandstetter.J, and

Hochreiter.S, “xLSTM: Extended Long Short-Term Memory” arXiv:2405.04517v1 [cs.LG] (2024).

[3] Vaswani.A, and Shazeer.N, and Parmar.N, and Uszkoreit.J, and Jones.L, and Gomez.A.N, and Kaiser.Ł and Polosukhin.I, “Attention Is All You Need” arXiv:1706.03762v7 (2023).