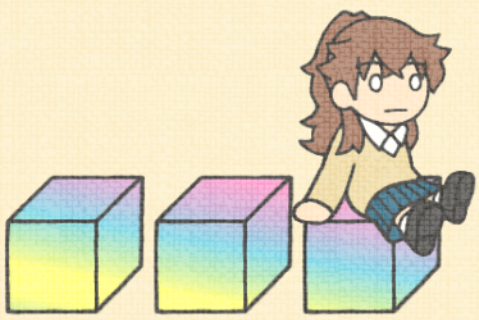
DRAFT 2022-01-15

https://github.com/CookieBox26/notes/tree/main/20211223_sequence_models

NeurIPS 2021 にみる

最近のニューラル系列モデルへの発見・工夫・理解



目次

まえがき

本書は機械学習の国際会議 NeurIPS 2021 で発表された論文から、(時) 系列データを処理するためのニューラルネットワークモデル――長いのでニューラル系列モデルとよびます――に関連する研究を私見で眺めたものです。網羅的でも排他的でもありません。

本書の内容は著者の理解であることにご留意ください。著者の誤りは著者に帰属します。

本書の内容についてお気付きの点がありましたら、大変お手数ですが、この原稿があるリポジトリの Issues、または著者ブログのコメント欄までお知らせください。著者ブログへのコメントはただちには公開されません。非公開希望の方はその旨をお知らせください。非公開希望であって返信が必要な場合はご連絡先の明記をお願いいたします。

リポジトリ https://github.com/CookieBox26/notes/
著者ブログ https://cookie-box.hatenablog.com/

本書に関連している記事は以下です。

メモ前編 https://cookie-box.hatenablog.com/entry/2021/11/28/191332
メモ後編 https://cookie-box.hatenablog.com/entry/2021/12/23/124713

登場人物紹介

この人はベイズ統計部の部長です。1 年生です。とある目的のためにベイズ統計部を立ち上げ、統計や機械学習を勉強しています。ベイズ統計部には部長と副部長しかいません。姉が 2 人います。

この人はベイズ統計部の副部長です。2 年生です。海外からの編入生でラクロス部に入部しようとしていましたが、部長に勧誘されてベイズ統計部に入部しました。数学が得意ですがなぜか著者を超える数学力が出せません。

NeurIPS 2021 のニューラル系列モデルを眺める

(時) 系列データを処理するためのニューラルネットワークモデルの動向を知るために、NeurIPS 2021 で発表された研究をみていきましょう。NeurIPS 2021 で発表された論文の総数は……2334 本[a] !?

[a] 2021 年 11 月 28 日時の https://proceedings.neurips.cc/paper/2021 のリンク数に基づく。

あ、あまりに多いので機械的に絞り込みましょう。さしあたりタイトルに time series, sequential, rnn, recurrent, transformer, attention, state space のいずれかを含む論文は……それでも 155 本……。ただこの絞り込みだと画像認識, GAN, 強化学習の研究も多そうですね。それらも興味深いですが、(時) 系列に関する研究を優先するべく断腸の思いでとばしていきましょう……。

とばしても 39 本……多いですね……こう多いと何が何だかわかりません。モデルの切り口でグループ分けして、アブストラクトからの理解で整理してみましょう。

まず最大勢力は Transformer ですね、検索語に含めたのでヒットするのは当然ですが、多くを占めます……**「セルフアテンションの計算量に対処する」**は昨年以前から引き続き人気 (?) なテーマであるようです。このサブグループは後で改めてメモしましょう。

グループ「Transformer」

- Transformer の性質を理解する。
  - 行列分解でパラメータを 10 倍削減しても性能が出ると示す [2]。
  - セルフアテンションを生物学的な記憶モデルと解釈する [16]。
- Transformer の使用方法を工夫する。
  - グリッド分割をさらにグリッド分割する (Vision Transformer) [17]。
  - 機械的にプレ処理 (トレンド-季節性分解) をする [23]。
  - 状態空間モデルと組み合わせて時系列の長期予測等をする [27]。
  - データ間でセルフ (?) アテンションする [36]。
- Transformer の訓練方法を工夫する。
  - ヘッド間で Q, K の分布を一致させる正則化をする [15]。
- Transformer のアーキテクチャを工夫する。
  - **セルフアテンションの計算量に対処する** [3] [4] [7] [11] [19] [20] [24] [25] [37]。
  - その他 Transformer のアーキテクチャを再考する。
    * 言語処理に適した構造を探索する [8]
    * セルフアテンションの代わりにゲート付 MLP にする [10]。
    * 時系列予測のために位相的アテンションを導入する [29]。
    * セルフアテンションを Softmax しない Transformer で偏微分方程式を解く [30]。

次は RNN ですね。**「RNN を理論的に理解する」**という研究が割にみられるように感じられます。理論解析が進めばどのような系列データにどのようなニューラルアーキテクチャを用いるべきかにつながるのでしょうか……??

グループ「RNN」

- **RNN を理論的に理解する。**
  - RNN がある再生核ヒルベルト空間におけるカーネル法であると示す [5]。

  - スイッチング線形動的システムで RNN をリバースエンジニアリングする [18]。
  - RNN が学習できると保証される関数の制約を撤廃する [21] [22]。
  - タスクの解空間を構造化して RNN の性質を調べる [31]。
  - 勾配消失/爆発しない RNN のサブセットを突き止める [32]。
  - 成長するメモリ付き固定精度 RNN がチューリング完全であると示す [35]。
- RNN を工夫する。
  - 訓練時に隠れ状態にノイズ添加してロバストにする [6]。
  - RNN 自体が時間変化できるようにする [9]。
  - ドロップアウトを活用して LSTM の計算量を削減する [28]。

後はその他とでもしましょうか。

その他

- 機械的に汎用的なプレ処理 (成分クラスタリング) をする [14]。
- 時系列を生成する混合分布を推定するためのコアセットを構築する [26]。
- 微分方程式で記述されるシステムをニューラルネットで実現する。
  - 線形時不変連続時間システムをニューラルネットで実現する [1]。
  - 連立微分方程式システムをベイズフィルタで解く [13]。
- 系列モデルを新しい用途に活用する。
  - Transformer を活用してガウス過程モデル適用時のカーネルを同定する [12]。
- 時系列のオンライン異常検知の偽陽性率を制御する [33]。
- Transformer と CNN のロバスト性を比較する [34]。
- ニューラル時系列モデルにおいて誤差の自己相関を調整する [38]。
- 継続期間も明示的に利用するレジームスイッチングモデルを実現する [39]。

最後に**「セルフアテンションの計算量に対処する」**を改めてみましょう。Transformer の計算量が取り沙汰されるのは $\mathrm{Softmax}\left(QK^\top/\sqrt{d}\right) \in \mathbb{R}^{N\times N}$ を求めるのに系列長 N に対して $\mathcal{O}(N^2)$ の計算量がかかるためですが、$\mathcal{O}(N^2)$ を回避するために、以下のようなアプローチが取られているようです。スパース化、低ランク近似自体はこれまでも計算量削減の基本路線であったと思いますが、新たな切り口を導入しているのと、その他の独自路線アプローチもみられるのではないでしょうか。

グループ「Transformer」のサブグループ「セルフアテンションの計算量に対処する」

- $QK^\top$ の成分を間引く (スパースにする)。
  - どの成分か不要なのか自体を学習する [11]。

- $QK^\top$ を低ランク近似する (行列分解する)。
  - カーネル法の計算量削減のアプローチを応用する [3]。
- スパース化と低ランク近似を統合する [19]。
- 長距離依存性の計算量を削減する。
  - 短距離依存性はそのまま計算し、長距離依存性は短い系列に射影する [20]。
  - 長距離依存性については重み付き期待値に対してアテンションする [24]。
  - FMM(高速多重極法) を応用して長距離依存性を低ランク近似する [37]。
- $QK^\top$ の計算箇所でだけ入力系列を短い系列に射影する [4]。
- 最初のセルフアテンション層では $QK^\top$ を計算するが、2 番目以降ではそれを時間発展させる [7]。
- アテンションの計算に高速フーリエ変換を応用する [25]。

それぞれのお話

[1] RNN、CNN、連続時間モデルを結合したい

ほげ。

[2] ほげ

ほげ。

[3] Skyformer——セルフアテンションの Nyström 近似

これまでのセルフアテンションの計算量削減には往々にして近似誤差の理論保証がないというようにいっていますね。だから手法間の比較もできなくなっているし、ハイパーパラメータによる計算量削減度合いの調整もできなくなっていると——

Nyström 近似?

[4] ほげ

ほげ。

[5] ほげ

ほげ。

[6] ほげ

ほげ。

[7] ほげ

ほげ。

[8] ほげ

ほげ。

[9] ほげ

ほげ。

[10] ほげ

ほげ。

[11] ほげ

ほげ。

[12] ほげ

ほげ。

[13] ほげ

ほげ。

[14] ほげ

ほげ。

[15] ほげ

ほげ。

[16] ほげ

ほげ。

[17] ほげ

ほげ。

[18] ほげ

ほげ。

[19] ほげ

ほげ。

[20] ほげ

ほげ。

[21] ほげ

ほげ。

[22] ほげ

ほげ。

[23] ほげ

ほげ。

[24] ほげ

ほげ。

[25] ほげ

アブストラクトは、「セルフアテンションの計算量を削減する既存研究の多くは『内積をとってからソフトマックスする』方式にしか対応できない」と主張しているようにみえます。そして、それだと「相対位置エンコーディング (RPE) に対応できない」と……これまでに提案されているセルフアテンション計算量削減ってそんなに制約があったんですか??

[26] ほげ

ほげ。

[27] ほげ

ほげ。

[28] ほげ

ほげ。

[29] 時系列予測のための位相的アテンション

ほげ。

[30] セルフアテンションを Softmax しない Transformer で偏微分方程式を解く

ほげ。

[31] タスクの解空間を構造化して RNN の性質を調べる

ほげ。

[32] SBO-RNN――勾配消失/爆発しない RNN のサブセット

こちらの SBO-RNN は、RNN の中でも勾配消失/爆発せず安定的に学習できる構造を突き止め、そのサブセットに SBO-RNN と名付けたということなのでしょうか?

[33] 時系列のオンライン異常検知の偽陽性率を制御する

時系列データのオンライン異常検知の話ですが、「FDRC ルール」とは読んで字のごとく偽陽性率を抑えるためのルール、なのでしょうか……?

[34] Transformer と CNN のロバスト性を比較する

「Transformer が CNN よりロバストとされているがそんなことはない」といったアブストラクトですが、そもそも Transformer が CNN よりロバストとされているんですか?

[35] 成長するメモリ付き固定精度 RNN がチューリング完全であると示す

チューリング完全って何ですか?

[36] データ間でセルフ (?) アテンションする Non-Parametric Transformer

検証したのがテーブルデータや CIFAR-10 であって系列データといった向きのデータではなさそうですが、タイトルが気になりました。Self-Attention Between Datapoints というのは、言語データに喩えるなら、単語から文章内の他の単語へアテンションするのではなく、文章から他の文章へアテンションするということなのでしょうか。それって Self なんでしょうか……? それはさておき、本当に「データセット全体を入力とする」のであれば訓練や推論のコストが膨大になりそうですが……?

[37] FMMformer——FMM(高速多重極法) を応用して長距離依存性を低ランク近似する

FMM(高速多重極法) というのは粒子間の相互作用を近距離成分と遠距離成分に分けて計算量を削減する電磁気学分野の手法なのでしょうか? こういわれると、Transformer の相互作用にも応用できる気配がしますが、具体的にどのような手法なのでしょうか?

[38] ニューラル時系列モデルにおける誤差の自己相関の調整

通常ニューラルネットで時系列データを学習するときにステップ間で誤差に自己相関はないとしていますが、現実には自己相関するので誤差の自己相関係数も学習するといっていますね?

[39] RED-SDS――継続期間も明示的に利用するレジームスイッチングモデル

時系列のレジームの切り替わりを捉えたいといっていますね……レジームというのはこの時点を境に好景気から不景気になったというような環境の変化のようなものですよね、適当な訳語がわかりませんが……。それで、RED-SDS: Recurrent Explicit Duration Switching Dynamical System なる提案モデルでは状態にも時間にも依存してレジームをスイッチングできるようにしたんですね? うーん、いまいちどう価値があることをしたのかわからないのですが……。

おそらくレジームスイッチングモデルは元々は何らかの変数 (観測不可能なら状態といった方がいいかな) に依存してスイッチングするモデルとして考案されたんだよね。ある変数がこうなってきたらここから不景気レジームだな、みたいに。でも、レジームの継続期間にもパターンがあるならそれを積極的に利用した方がいいよね。わからないけど、この病気の流行は 1 ヶ月で落ち着く、みたいな知識があったりしたらさ。それが Explicit Duration Switching の意味かなと思うんだけど、この発想自体は前からあって、この論文の新規性はそれを状態スイッチングモデルと組み合わせてディープで実現したところにあるのかな?

結び――NeurIPS 2021 にみる最近のニューラル系列モデルへの発見・工夫・理解

Appendix

公開コードを動かす――[3] Skyformer

Skyformer [3] を動かしてみましょう。

```
from models.model_LRA import ModelForSC, ModelForSCDual
from config import Config

model_config = Config["lra-text"]["model"]
model_config["mixed_precision"] = True
model_config["attn_type"] = "softmax"
model = ModelForSC(model_config)
print(model)
```

```
ModelForSC(
  (model): Model(
    (embeddings): Embeddings(
      (word_embeddings): Embedding(512, 64)
      (position_embeddings): Embedding(4000, 64)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (transformer_0): TransformerLayer(
      (norm1): LayerNorm((64,), eps=1e-05, elementwise_affine=True)
      (mha): Attention(
        (W_q): Linear(in_features=64, out_features=64, bias=True)
        (W_k): Linear(in_features=64, out_features=64, bias=True)
        (W_v): Linear(in_features=64, out_features=64, bias=True)
        (attn): SoftmaxAttention(
          (drop_attn): Dropout(p=0.1, inplace=False)
        )
# 以下省略
```

これに Long Range Arena のデータを渡せばよいですね。

参考文献

[1] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, Christopher Ré. Combining Recurrent, Convolutional, and Continuous-time Models with Linear State Space Layers. In NeurIPS 2021.
[2] Aliakbar Panahi, Seyran Saeedi, Tom Arodz. Shapeshifter: a Parameter-efficient Transformer using Factorized Reshaped Matrices. In NeurIPS 2021.
[3] Yifan Chen, Qi Zeng, Heng Ji, Yun Yang. Skyformer: Remodel Self-Attention with Gaussian Kernel and Nyström method. In NeurIPS 2021.
[4] Xuezhe Ma, Xiang Kong, Sinong Wang, Chunting Zhou, Jonathan May, Hao Ma, Luke Zettlemoyer. Luna: Linear Unified Nested Attention. In NeurIPS 2021.
[5] Adeline Fermanian, Pierre Marion, Jean-Philippe Vert, Gérard Biau. Framing RNN as a kernel method: A neural ODE approach. In NeurIPS 2021.
[6] Soon Hoe Lim, N. Benjamin Erichson, Liam Hodgkinson, Michael W. Mahoney. Noisy Recurrent Neural Networks. In NeurIPS 2021.
[7] Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, Tanmoy Chakraborty. Redesigning the Transformer Architecture with Insights from Multi-particle Dynamical Systems. In NeurIPS 2021.
[8] David So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, Quoc Le. Searching for Efficient Transformers for Language Modeling. In NeurIPS 2021.
[9] Aston Zhang, Yi Tay, Yikang Shen, Alvin Chan Guo Wei, SHUAI ZHANG. Self-Instantiated Recurrent Units with Dynamic Soft Recursion. In NeurIPS 2021.
[10] Hanxiao Liu, Zihang Dai, David So, Quoc Le. Pay Attention to MLPs. In NeurIPS 2021.
[11] Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, Łukasz Kaiser, Wojciech Gajewski, Henryk Michalewski, Jonni Kanerva. Sparse is Enough in Scaling Transformers. In NeurIPS 2021.
[12] Fergus Simpson, Ian Davies, Vidhi Lalchand, Alessandro Vullo, Nicolas Durrande, Carl Edward Rasmussen. Kernel Identification Through Transformers. In NeurIPS 2021.
[13] Jonathan Schmidt, Nicholas Krämer, Philipp Hennig. A Probabilistic State Space Model for Joint Inference from Differential Equations and Data. In NeurIPS 2021.
[14] Zhibo Zhu, Ziqi Liu, Ge Jin, Zhiqiang Zhang, Lei Chen, Jun Zhou, Jianyong Zhou. MixSeq: Connecting Macroscopic Time Series Forecasting with Microscopic Time Series Data. In NeurIPS 2021.
[15] Shujian Zhang, Xinjie Fan, Huangjie Zheng, Korawat Tanwisuth, Mingyuan Zhou. Alignment Attention by Matching Key and Query Distributions. In NeurIPS 2021.
[16] Trenton Bricken, Cengiz Pehlevan. Attention Approximates Sparse Distributed Memory. In NeurIPS 2021.
[17] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing XU, Yunhe Wang. Transformer in Transformer. In NeurIPS 2021.
[18] Jimmy Smith, Scott Linderman, David Sussillo. Reverse engineering recurrent neu-

ral networks with Jacobian switching linear dynamical systems. In NeurIPS 2021.
[19] Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, Christopher Ré. Scatterbrain: Unifying Sparse and Low-rank Attention. In NeurIPS 2021.
[20] Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, Bryan Catanzaro. Long-Short Transformer: Efficient Transformers for Language and Vision. In NeurIPS 2021.
[21] Lifu Wang, Bo Shen, Bo Hu, Xing Cao. On the Provable Generalization of Recurrent Neural Networks. In NeurIPS 2021.
[22] Abhishek Panigrahi, Navin Goyal. Learning and Generalization in RNNs. In NeurIPS 2021.
[23] Haixu Wu, Jiehui Xu, Jianmin Wang, Mingsheng Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In NeurIPS 2021.
[24] Hongyu Ren, Hanjun Dai, Zihang Dai, Mengjiao Yang, Jure Leskovec, Dale Schuurmans, Bo Dai. Combiner: Full Attention Transformer with Sparse Computation Cost. In NeurIPS 2021.
[25] Shengjie Luo, Shanda Li, Tianle Cai, Di He, Dinglan Peng, Shuxin Zheng, Guolin Ke, Liwei Wang, Tie-Yan Liu. Stable, Fast and Accurate: Kernelized Attention with Relative Positional Encoding. In NeurIPS 2021.
[26] Lingxiao Huang, K Sudhir, Nisheeth Vishnoi. Coresets for Time Series Clustering. In NeurIPS 2021.
[27] Binh Tang, David Matteson. Probabilistic Transformer For Time Series Analysis. In NeurIPS 2021.
[28] Anup Sarma, Sonali Singh, Huaipan Jiang, Rui Zhang, Mahmut Kandemir, Chita Das. Structured in Space, Randomized in Time: Leveraging Dropout in RNNs for Efficient Training. In NeurIPS 2021.
[29] Sebastian Zeng, Florian Graf, Christoph Hofer, Roland Kwitt. Topological Attention for Time Series Forecasting. In NeurIPS 2021.
[30] Shuhao Cao. Choose a Transformer: Fourier or Galerkin. In NeurIPS 2021.
[31] Elia Turner, Kabir Dabholkar, Omri Barak. Charting and Navigating the Space of Solutions for Recurrent Neural Networks. In NeurIPS 2021.
[32] Ziming Zhang, Yun Yue, Guojun Wu, Yanhua Li, Haichong Zhang. SBO-RNN: Reformulating Recurrent Neural Networks via Stochastic Bilevel Optimization. In NeurIPS 2021.
[33] Quentin Rebjock, Baris Kurt, Tim Januschowski, Laurent Callot. Online false discovery rate control for anomaly detection in time series. In NeurIPS 2021.
[34] Yutong Bai, Jieru Mei, Alan L. Yuille, Cihang Xie. Are Transformers more robust than CNNs? . In NeurIPS 2021.
[35] Stephen Chung, Hava Siegelmann. Turing Completeness of Bounded-Precision Recurrent Neural Networks. In NeurIPS 2021.
[36] Jannik Kossen, Neil Band, Clare Lyle, Aidan N. Gomez, Thomas Rainforth, Yarin Gal. Self-Attention Between Datapoints: Going Beyond Individual Input-Output Pairs

in Deep Learning. In NeurIPS 2021.

[37] Tan Nguyen, Vai Suliafu, Stanley Osher, Long Chen, Bao Wang. FMMformer: Efficient and Flexible Transformer via Decomposed Near-field and Far-field Attention. In NeurIPS 2021.

[38] Fan-Keng Sun, Chris Lang, Duane Boning. Adjusting for Autocorrelated Errors in Neural Networks for Time Series. In NeurIPS 2021.

[39] Abdul Fatir Ansari, Konstantinos Benidis, Richard Kurle, Ali Caner Turkmen, Harold Soh, Alexander J. Smola, Bernie Wang, Tim Januschowski. Deep Explicit Duration Switching Models for Time Series. In NeurIPS 2021.

NeurIPS 2021 にみる **最近のニューラル系列モデルへの発見・工夫・理解**

YYYY 年 MM 月 DD 日　初版発行
YYYY 年 MM 月 DD 日　第 2 版発行

著　者　クッキー
発行者　クッキーの日記
https://cookie-box.hatenablog.com/