目次

まえがき

お気付きの点がありましたらこの原稿のリポジトリの Issues までお願いいたします。
https://github.com/CookieBox26/notes/issues

オーバービュー

NeurIPS 2021 で発表された (時) 系列データを扱うモデルに関連する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 !?

あ、あまりに多いので機械的に絞り込みましょう。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……。画像認識, GAN, 強化学習系のタイトルは断腸の思いでとばしていきましょう……。

――こうしてみると、**「セルフアテンションの計算量に対処する」**は引き続き人気 (?) なテーマである一方、**「RNN を理論的に理解する」**という研究も割りにみられるように感じられます。以下は私見による整理です。

- **RNN を理論的に理解する** [5] [18] [21] [22]。
- 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]。
- トランスフォーマーを理論的に理解する [2] [16]。
- トランスフォーマーを工夫する。
  - 機械的にプレ処理 (トレンド-季節性分解) をする [23]。
  - グリッド分割をさらにする (ビジョントランスフォーマー) [17]。
  - **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]。
  - ヘッド間で Q, K の分布を一致させる正則化をする [15]。
  - トランスフォーマーのアーキテクチャ自体を再考する。
    * 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]。
- 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]。
- 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。
  - トランスフォーマー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]。

グループ「RNN を理論的に理解する」

RNN の理論的な理解に関する論文が複数みられました。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？

- RNN がある再生核ヒルベルト空間におけるカーネル法と捉えられることを示す [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]。
- これまでの理論保証の制約を緩和する [21] [22]。

グループ「セルフアテンションの計算量に対処する」

トランスフォーマー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。

- $QK^\top$ 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]。
- $QK^\top$ の計算箇所でだけ入力系列を短い系列に射影する [4]。
  - 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]。
- 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]。

各論 (未執筆)

カーネル法の計算量削減のアプローチを応用する [3]

原題: Skyformer: Remodel Self-Attention with Gaussian Kernel and Nyström Method (Chen et al.)

トランスフォーマーはセルフアテンション層が計算量のボトルネックとなっていますが、カーネルマシンもまた内積計算がボトルネックになっている、と。そうですね、カーネル法のグラム行列のサイズもデータサイズに応じて $N \times N$ になりますものね。そこで、カーネル法で用いられる Nyström 近似を適用できるようにして適用したトランスフォーマーがスカイフォーマーであると。なぜスカイフォーマーなのか少し気になったので論文を覗いてみると Symmetrization of Kernelized attention for NYström method なのですね…。

以下未執筆。

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 neural 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.