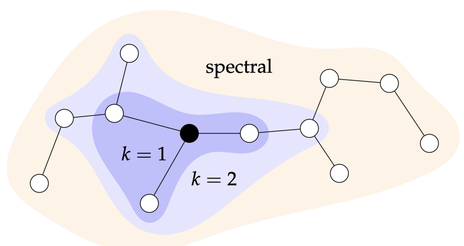
Spatio-Spectral Sequence Processing

Nikita Kostin, Simon Geisler, Arthur Kosmala, Stephan Günnemann

TLDR: We augment sequence models with a learnable spectral branch (S2Seq) to capture global structure. On the LRA benchmark, S2Seq shows consistent gains on certain datasets, with just $K = 32$ frequencies needed for near-optimal performance. We further propose linear-time extensions for autoregressive prediction.

Background: Spatio-Spectral GNNs

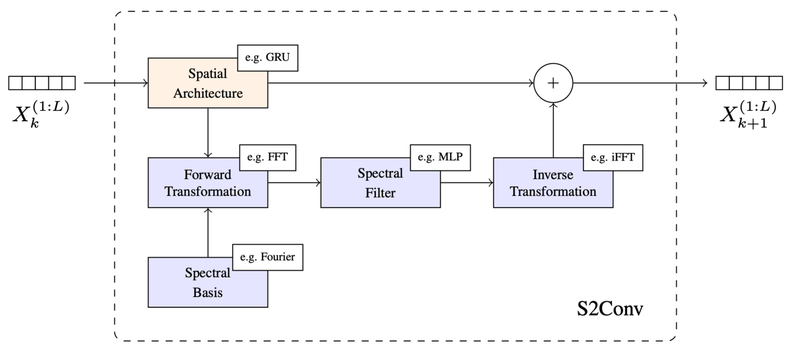


- S2GNNs provide a framework for combining **local** and **global** geometric information on graphs
- Local information is modelled with a regular GNN architecture
- Global information is captured via learnable spectral filtering on the graph **Laplacian eigenbasis**
- Sequences are **chain graphs** ⇒ graph Laplacian corresponds to DCT

Motivation

- Transformers are **quadratic** ⇒ prohibitive for long sequence
- S2GNNs are efficient at capturing global information - sequences are **chain graphs**!
- Goal: a lightweight and **model-agnostic** spectral augmentation for sequence architectures

Architecture



- **Modular** add-on on top of a sequence architecture (e.g. GRU, Conv1D, Mamba)
- Forward transform on the spatial output → spectral filter → inverse transform
- Choice of **spectral basis** (DFT/DST/DCT) does not influence the result
- Small parameter overhead

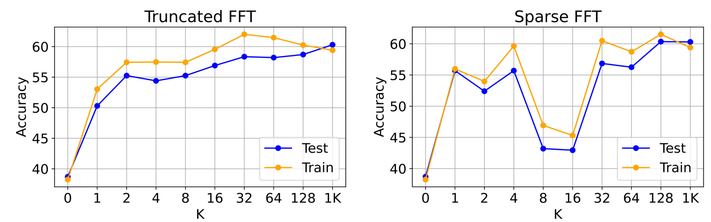
Long-Range Arena Benchmark

- 5 long-sequence (L between $1k$ and $4k$) classification datasets
- Spectral augmentation provides consistent gains on certain datasets
- S2Seq on GRUs approaches **SOTA performance** on ListOps

Architecture	S2Seq	ListOps	Text	Retrieval	Image	Path	Avg.
Transformer	✗	36.37	64.27	57.46	42.44	71.40	54.39
BigBird	✗	36.05	64.02	59.29	40.83	74.87	55.01
S4	✗	59.60	86.62	90.90	**88.65**	94.20	83.99
S5	✗	**62.15**	**89.31**	**91.40**	88.00	**95.33**	**85.24**
Conv1D	✗	33.70	84.59	74.58	54.29	✗	61.79
GRU	✗	38.70	82.10	82.89	65.28	73.85	68.56
Conv1D	✓	46.44 ↑	85.03	81.17 ↑	55.09	✗	66.93
GRU	✓	60.30 ↑	83.40	84.53 ↑	65.33	73.20	73.35

Partial Spectra

- Low number of frequencies K is enough to give a **performance boost**
- At $K = 32$ the performance is **almost optimal**
- This enables partial spectrum computation, e.g. via truncated or sparse FFT - can be $o(L)$ where L is the length



Autoregressive Prediction

Challenges

- Autoregressive spectrum must be **causal**
- Spectrum update must be **efficient** - naive implementation is $O(L^2 \log L)$

Proposed approaches

- Sliding window
- Sliding window with doubling at exponential points
- Exponential window

$$X_n(\omega) = \sum_{m=-\infty}^{0} e^{m/\tau} x[n+m] e^{-j\omega m}$$

Sliding and exponential windows remain in $O(L)$!