ML4SCI Hackathon 2021
NMR Challenge

Tun Sheng Tan

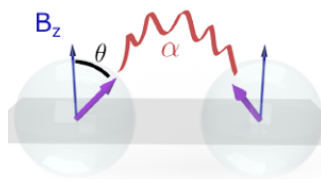
Physics Department
University of Florida

January 17 2022

Problem and Approach

Using spin-echo magnetizations $M_x(t)$ and $M_y(t)$ to predict

- the planar effective scattering strength α_x
- perpendicular effective scattering strength α_z
- the flip angle of nuclear spins θ



Approach

- Small dataset ($n = 3000$) with sequence length $L = 856$
- Uncorrelated targets $\rightarrow$ independent regressions
- Prioritize feature engineering

What does not work

1. Dimensionality reduction using PCA
2. Feature selection using F-statistic and p-values
3. Area under curve, $|M^{xy}(t)|$
4. Time-delay embedding
5. Higher order gradient of time-series[1]
6. Power spectral density & Cross power spectral density[1]
7. Regressor Chain

[1]Not as good as Fourier transform. $MSE_\theta^{PSD}, MSE_\theta^{TS} \sim 10^{\circ^2}$ vs $MSE_\theta^{FFT} \sim 1^{\circ^2}$.

Features

1. Fourier transform $|\tilde{M}^{xy}(\omega)|$[1]
2. $\frac{d}{d\omega}|\tilde{M}^{xy}(\omega)|$
3. Constant-Q transform
4. Trigonometry transform: $\log[V\cos(\psi)]$ and $\log[|\psi|]$ where

$$
\begin{aligned}
V &= \sin^{-1}\left(1.1\max\{|M^{xy}(t)|\}\right) \\
\phi(t) &= \tan^{-1}\left(\frac{M^x(t)}{M^y(t)}\right) \\
\psi(t) &= \sin^{-1}\left(\frac{M^x(t)}{V\cos(\phi(t))}\right)
\end{aligned}
$$

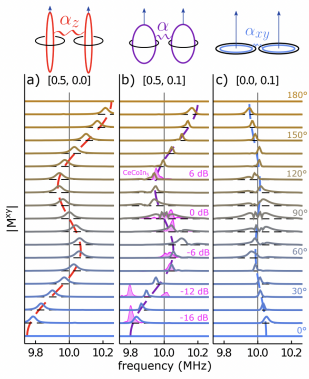


Figure: [1]

[1] arXiv:2110.06811 . See section 3 of supplementary material for details.

Base Models

Three base models without hyperparameter tuning

1. LightGBM: default settings
   - $\alpha_{x,z}$: Features 1, 2, 3, 4
   - θ: Features 1, 2, 4
2. Transformer: SGD optimizer, cyclic learning rate, SiLU activation, smooth L1 loss, 256 batch size, 1000 epochs
   - Model 1: Features 1, 2, 3, 4
   - Model 2: Features 1, 3, 4

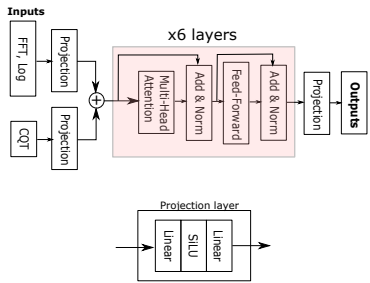


Figure: Transformer

Final Model

1. Data preprocessing:
   - standardize features (removing the mean and scaling to unit variance)
   - log transform targets and standardize
2. Perform out-of-fold prediction (k=10) using transformer 1, transformer 2 and LightGBM
3. Take weighted average of the models for each target
4. Clip $\alpha_x \in [0, 150] kHz$, $\alpha_z \in [0, 300] kH$, $\theta \in [10, 90]^\circ$

Code available at https://github.com/tunsheng/ML4SCI-NMR-2021-Solution

Backup

Ablation Study (I)

Features	MSE_{α_x}	MSE_{α_z}	MSE_{θ}
TS	231.6800	6625.0227	141.0835
FFT	94.0395	2634.2513	1.4023
CQT	41.4168	5825.2558	6.7801
FFT + GRAD FFT	91.1387	2661.7525	1.3735
FFT + CQT	38.1213	2365.2629	2.5243
TRIG + CQT	33.6687	5434.9461	6.6327
FFT + TRIG	43.5086	2515.2337	**1.2402**
FFT + GRAD FFT + TRIG	44.9065	**2280.9371**	1.3222[1]
FFT + TRIG + CQT	33.2738	2377.5268	2.5472
FFT + GRAD FFT + TRIG + CQT	**32.0861**[1]	2427.6995[1]	2.4594

Table: Mean squared errors for LightGBM

[1] Used this for training

Ablation Study (II)

Features	MSE_{α_x}	MSE_{α_z}	MSE_{θ}
TS	231.6800	6625.0227	141.0835
GRAD 6th	203.1794	3974.3285	37.2417
TS + GRAD 4th	144.8954	4465.4705	18.1563
TS + GRAD 6th	142.9788	4154.9162	19.8565
TS + GRAD 6th + TRIG	61.1777	3558.1060	37.0968
TS + TRIG	84.6014	4941.4302	80.1348
TS + FFT	103.9186	2940.1749	1.5453
TS + CQT	40.1949	5673.7429	6.9262

Table: Mean squared errors for LightGBM

Stacking

To minimize variance and improve accuracy

Model	MSE_{α_x}	MSE_{α_z}	MSE_{θ}
LightGBM	36.22663	2299.48227	1.53249
Transformer 1	40.11050	2297.86204	4.95406
Transformer 2	46.58499	2319.71042	5.06916
Ensemble	25.64227	1749.83395	1.36223

Final submission score: 237.43