**Notes on the given relevant literatures**

**Zihao Liu**

1. “Combining Recurrent, Convolutional, and Continuous-time Models with Linear State-Space Layers”

**[Notes]:**

* The focus is on the expression of Linear State-Space Layer (LSSL) and the purpose of it is to generalizes Recurrent neural networks (RNNs), temporal convolutions, and neural differential equations (NDEs) for time-series data processing. The authors even show that RNN methods are special cases of LSSLs.
* LSSL maps the time-series u(t) → y(t) by discretizing linear continuous-time

state-space based on the expression of

The authors show that LSSLs (1) are recurrent, (2) are convolutional, and (3) are continuous-time. It highlights that LSSL has a close relation with the three algorithms above. The continuous-time feature of LSSL makes it suitable to deal with missing continuous time-series data.

* The authors show that LSSLs do not sacrifice expressivity, so it can inherit the strengths of the three algorithms above. Also, the limitations are kept by naïve LSSLs: (1) the very long dependences are difficult to be captured. (2) The state matrix A and timescale is important but difficult to be chosen.
* The authors raise two specific methods to solve the two limitations above by using: (1) a carefully chosen class of structured matrices A. (2) And a new LSSLs algorithm is raised to learn A and . After these limitations are released, the authors show that stacked LSSLs in a deep network are widely effective on the databases in that paper to capture the long dependences. If the sequence is very long (e.g., a very long speech), LSSLs are still outperform other algorithms.

1. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation

**[Notes]:**

* This Conditional Score-based Diffusion models for Imputation (CSDI) model is the most powerful competitor of SSSD (S4 or SA) model (our target). CSDI is a probabilistic imputation method by directly learning the conditional distribution with conditional score-based diffusion models.
* CSDI is exactly used for the imputation and forecast of missing values in time-series also for the probabilistic forecasting. It is based on score-based conditional diffusion models conditioned on observed data. The authors show the outstanding effectiveness of CSDI by the self-training method.
* There is a reverse process in the conditional diffusion model to convert noise into plausible time series. At each step t, the reverse process removes noise from the output of the previous step (t + 1). The reverse process can take observations as a conditional input. These observed values are the conditional information when CSDI is trained. The missing time-series information is the imputation targets.
* (\*) There is another point we should pay attention. In this paper, the authors also mention the issue of that the ground-truth missing values are unknown. Also, the training data may not contain missing data at all. Then the machine cannot be trained effectively. This issue is exactly our **question (f).** Here, the authors provide a way to validate the CSDI effectiveness: using a self-supervised training method that separates observed values into conditional information and imputation targets inspired by masked language modeling. Maybe this can give us some ideas to answer question (f).
* The authors mention that diffusion models are generally designed for data generation and do not take conditional observations as inputs. Some researchers will use a Markov chain model to approximate the conditional distribution . The authors raise that noise will be added by this way and the imputation quality can be improved without any approximation. To solve this issue, the authors raise a conditional diffusion model without approximations where the observed values () are included.

1. Mega: Moving Average Equipped Gated Attention (will read in detail later)

**[Notes]:**

* It is about attention mechanism. The authors raise a moving average equipped gated attention mechanism (Mega) model to solve the two limitations simultaneously of general attention mechanism: (1) weak inductive bias and (2) quadratic computational complexity. The key idea is to incorporate inductive biases into the attention mechanism across the timestep dimension, by leveraging the classic exponential moving average (EMA) approach.
* Also, Mega-chunk is raised by the authors by splitting split the sequences of queries.

1. Simplified State Space Layers for Sequence Modeling (will read in detail later)

**[Notes]:**

* Based on the structured state space sequence (S4) layer (our target model) which combines linear state space models (SSMs), the authors raise a S5 layer. Whereas an S4 layer uses many independent single-input, single-output SSMs, the S5 layer uses one multi-input, multi-output SSM. Also, S5 uses an efficient and widely implemented parallel scan. This removes the need for the convolutional and frequency-domain approach used by S4.
* The motivation to raise this S5 layer is to capture very long sequences.
* In this paper, the authors also establish a relationship between the dynamics of S5 and S4.
* In experiments, the authors show that the S5 layer matches the performance and efficiency of the S4 layer. Meanwhile, as S5 removes the need for the convolutional and frequency-domain approach used by S4, does not require an additional position-wise linear layer (since these features are already mixed), so the authors believe S5 is better than S4.