A Statistical Recurrent Model on the Manifold of Symmetric Positive Definite Matrices

RUDRASIS CHAKRABORTY[a], CHUN-HAO YANG[a], XINGJIAN ZHEN[b], MONAMI BANERJEE[a],
DEREK ARCHER[a], DAVID VAILLANCOURT[a], VIKAS SINGH[b] AND BABA C. VEMURI[a]

[a] CISE, UNIVERSITY OF FLORIDA, USA   [b] UNIVERSITY OF WISCONSIN MADISON, USA

NeurIPS | 2018
Thirty-second Conference on Neural Information Processing Systems
Year (2018)

ABSTRACT

In this paper, we develop a novel statistical recurrent network for data that are ordered, temporal in nature and live on a Riemannian manifold. An efficient algorithm and rigorous analysis of its statistical properties are then presented. Extensive numerical experiments demonstrating competitive performance with state-of-the-art methods but with significantly less number of parameters are presented. We also present applications to a statistical analysis task in brain imaging, a regime where deep neural network models have only been utilized in limited ways.

MOTIVATION

- ✓ Statistical recurrent unit (SRU) model on Euclidean space.
- ✓ This un-gated architecture gives competitive results compared to more complex alternatives like LSTM and GRU.
- ✓ SRU is very similar to taking the average of stochastic processes and looking at the "average process".

$$\mathbf{r}_t = \text{ReLU}\left(W^{(r)}\boldsymbol{\mu}_{t-1} + b^{(r)}\right)$$

$$\boldsymbol{\varphi}_t = \text{ReLU}\left(W^{(\phi)}\mathbf{r}_t + W^{(x)}\mathbf{x}_t + b^{(\phi)}\right)$$

$$\forall \alpha \in J\,, \quad \boldsymbol{\mu}_t^{(\alpha)} = \alpha\boldsymbol{\mu}_{t-1}^{(\alpha)} + (1-\alpha)\boldsymbol{\varphi}_t$$

$$\mathbf{o}_t = \text{ReLU}\left(W^{(o)}\boldsymbol{\mu}_t + b^{(o)}\right)$$

KEY INGREDIENTS IN SRU

(a) weighted sum **(b)** addition of bias **(c)** moving average **(d)** ReLU

SPD-SRU INGREDIENTS

- ✓ weighted sum and moving average → Weighted Fréchet mean (wFM)
- ✓ addition of bias → action by the group $\mathsf{O}(n)$
- ✓ ReLU → tangent ReLU

SPD-SRU

- ✧ Weighted Sum and Moving average can be replaced by wFM if we include a convexity constraint on the weights.
- ✧ Since $\mathsf{O}(n)$ is a subgroup of the group of isometries of $\mathsf{SPD}(n)$, naturally it preserves the metric and hence the corresponding group action can be viewed as 'translations'.
- ✧ Use ReLU on the parameter space of $\mathsf{SPD}(n)$ and then map it back on to $\mathsf{SPD}(n)$.
- ✧ Let, $X_1, X_2, \cdots X_T$ be an input temporal or ordered sequence of points on $\mathsf{SPD}(n)$. The update rules for a layer of SPD-SRU are,

$$Y_t = \mathsf{FM}\left(\left\{M_{t-1}^{(\alpha)}\right\}, \left\{w^{(y,\alpha)}\right\}\right), \quad R_t = \mathsf{T}\left(Y_t, g^{(r)}\right)$$

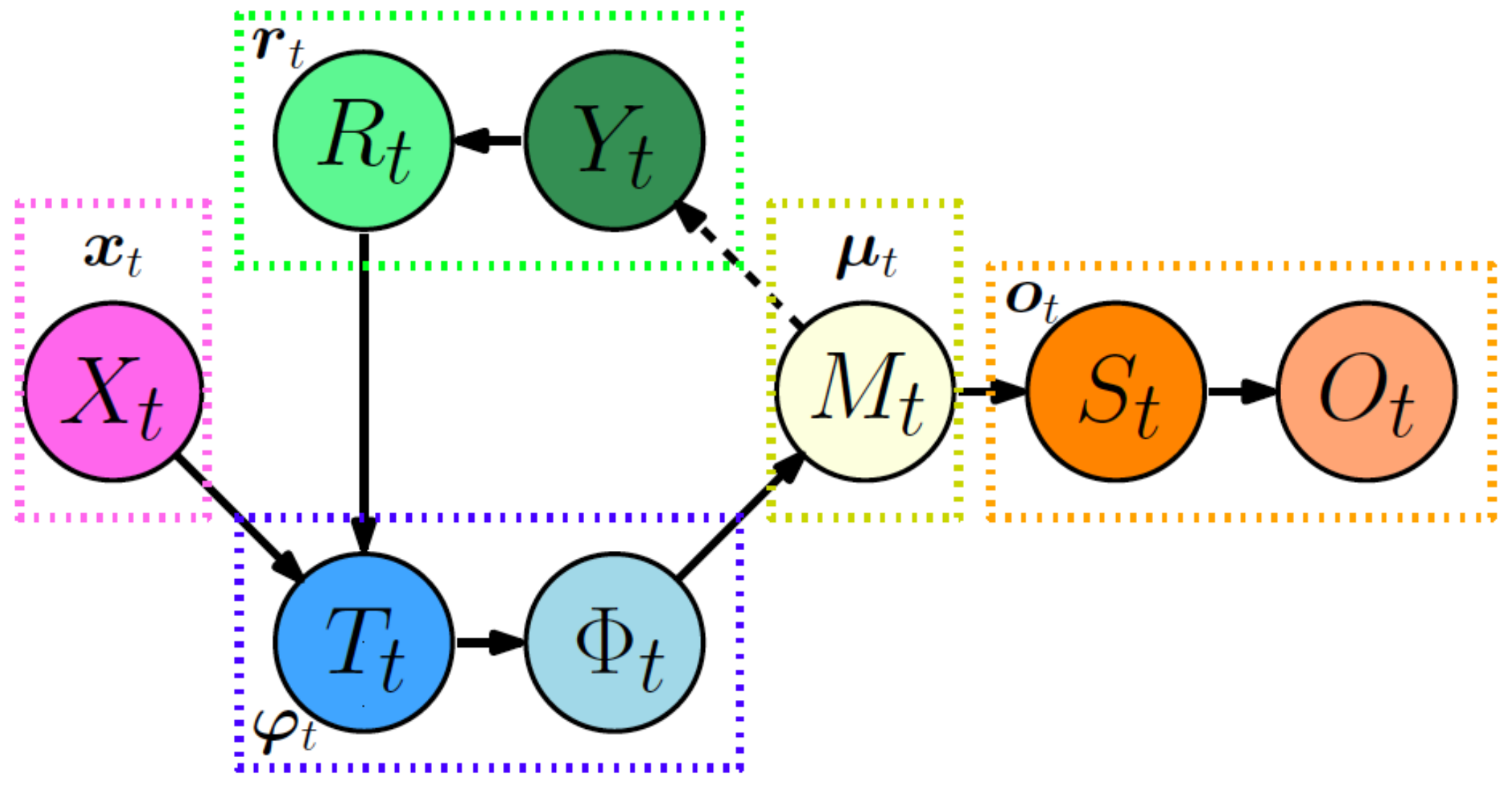
$$T_t = \mathsf{FM}\left(\left\{R_t, X_t\right\}, w^{(t)}\right), \quad \Phi_t = \mathsf{T}\left(T_t, g^{(p)}\right)$$

$$\forall \alpha \in J\,, \quad M_t^{(\alpha)} = \mathsf{FM}\left(\left\{M_{t-1}^{(\alpha)}, \Phi_t\right\}, \alpha\right)$$

$$S_t = \mathsf{FM}\left(\left\{M_t^{(\alpha)}\right\}, \left\{w^{(s,\alpha)}\right\}\right)$$

$$O_t = \mathsf{Chol}\left(\text{ReLU}\left(\mathsf{Chol}\left(\mathsf{T}\left(S_t, g^{(y)}\right)\right)\right)\right)$$

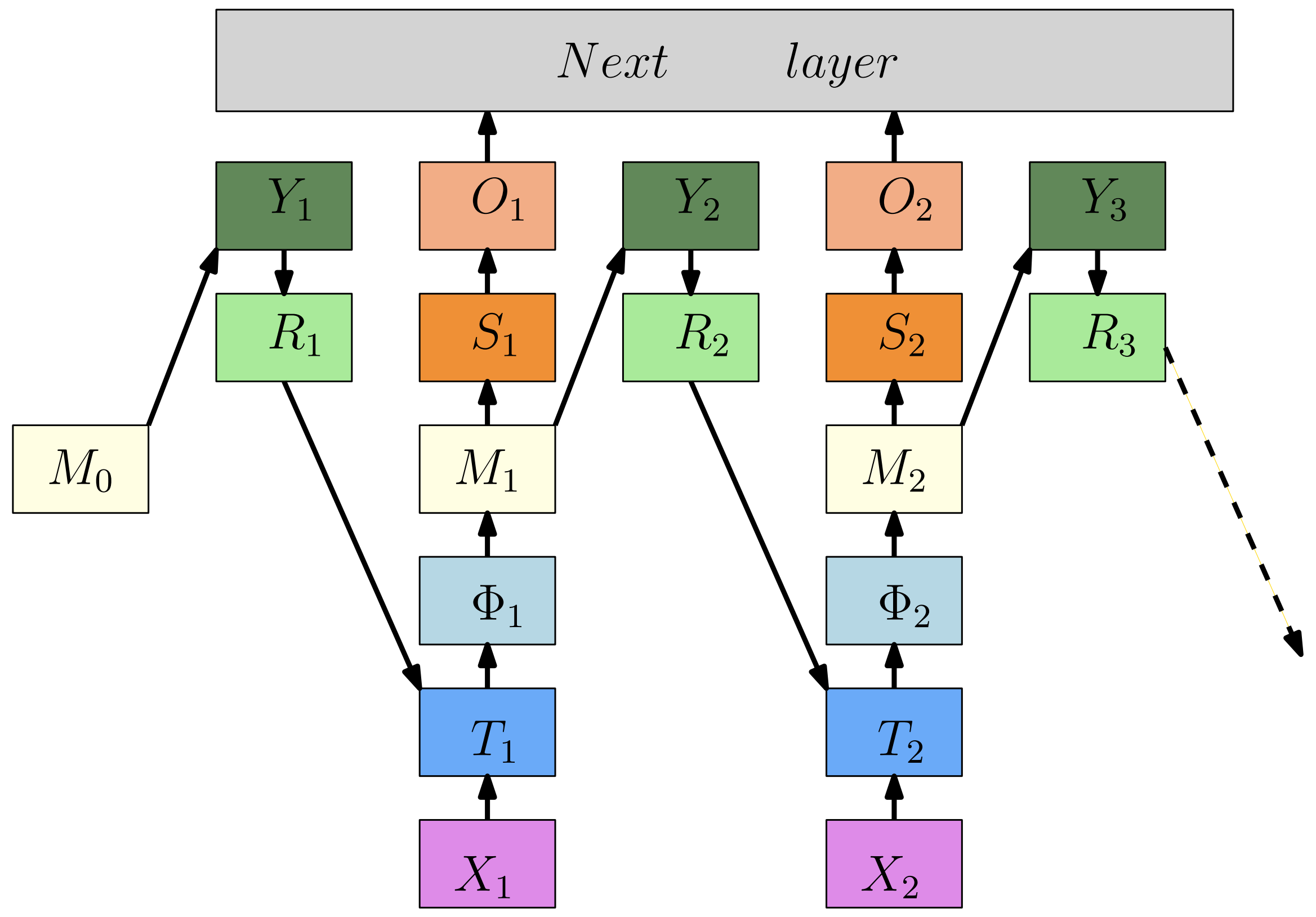
- ✧ $M_0^{(\alpha)}$ is initialized to be a diagonal $n \times n$ matrix with small positive values.
- ✧ The set J consists of positive real numbers from the unit interval.
- ✧ Now, computing the FM at the different elements of J will give a wFM at different "scales", exactly as desired.
- ✧ Sketch of an SPD-SRU and SRU layer:



- ✧ Using the parametrization of $\mathsf{O}(n)$, we learn the "bias" term on the parameter space, which is a vector space.
- ✧ In order to ensure the convexity property on the weights, we learn the square root of the weights which is unconstrained, i.e., the entire real line.
- ✧ All the trainable parameters lie in the Euclidean space and the optimization of these parameters is unconstrained, hence standard techniques are sufficient.
- ✧ The Riemannian gradient descent to compute wFM has a runtime complexity of $\mathcal{O}(iN)$, where N is the number of samples and i is the number of iterations for convergence – this runtime makes training incredibly slow.
- ✧ Our strategy: (a) use a recursive wFM computing algorithm (b) proof weak consistency of the estimator.
- ✧ Recursive wFM algorithm

$$M_k = M_{k-1}\left[\sqrt{T_k + \frac{(2w_k-1)^2}{4}\left(I - T_k\right)^2} - \frac{2w_k-1}{2}\left(I - T_k\right)\right] \text{ where, } T_k = M_{k-1}^{-1}X_k$$

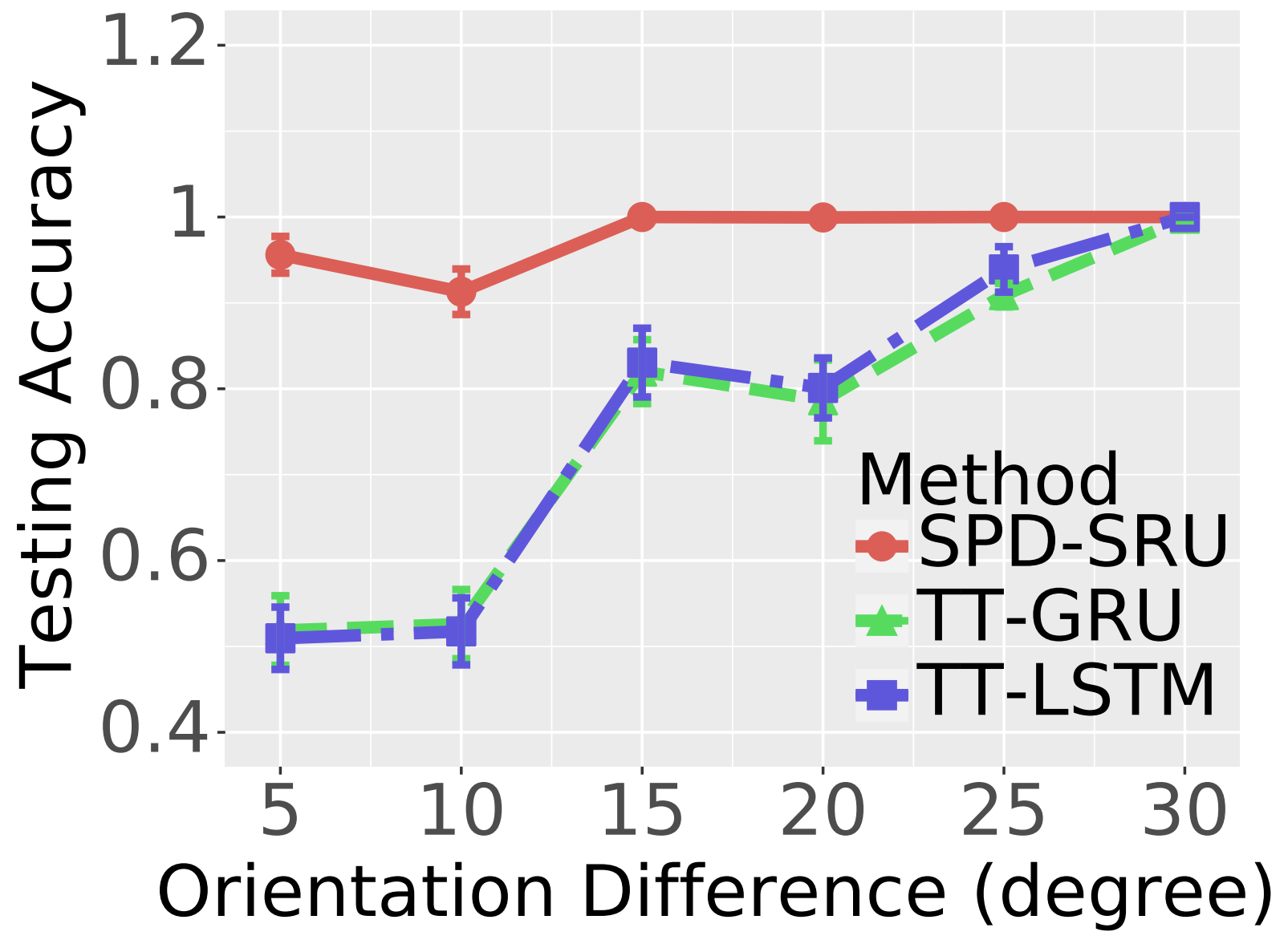
- ✧ Weak Consistency **(a)** $\text{Var}\left(M_k\right) \to 0$ as $k \to \infty$. **(b)** The rate of convergence of the proposed recursive FM estimator is super linear.
- ✧ Schematic of a SPDSRU network



EXPERIMENTAL RESULTS

- ✧ **Moving MNIST:** 11×11 SPD matrices, 1 SPD-SRU layer.

Mode	# params.	time (s) / epoch	orientation (°)		
			30-60	10-15	10-15-20
SPD-SRU	**1559**	~ 6.2	**1.00 ± 0.00**	**0.96 ± 0.02**	**0.94 ± 0.02**
TT-GRU	2240	~ **2.0**	**1.00 ± 0.00**	0.52 ± 0.04	0.47 ± 0.03
TT-LSTM	2304	~ **2.0**	**1.00 ± 0.00**	0.51 ± 0.04	0.37 ± 0.02
SRU	159862	~ 3.5	**1.00 ± 000**	0.75 ± 0.19	0.73 ± 0.14
LSTM	252342	~ 4.5	0.97 ± 0.01	0.71 ± 0.07	0.57 ± 0.13



- ✧ **UCF-11:** 8×8 SPD matrices, 5 SPD-SRU layer.

Model	# params.	time/ epoch	Test acc.
SPD-SRU	**3337**	~ 76	**0.78**
TT-GRU	6048	~ 42	**0.78**
TT-LSTM	6176	~ **33**	**0.78**
SRU	2535630	~ 50	0.75
LSTM	14626425	~ 57	0.70

- ✧ **Group differences:** (a) The data pool consists of dMRI (human) brain scans acquired from 50 'PD' patients and 44 controls ('CON'). (b) We used FSL to extract M1 fiber tracts (denoted by 'LM1' and 'RM1') which consists of 33 and 34 points respectively. (c) We fit diffusion tensors and extract 3×3 SPD matrices. (d) Now, for each of these two classes, we use 3 layers of SPD-SRU to learn the tract patterns to get two models for 'PD' and 'CON' respectively.(e) We use permutation testing based on a "distance" between learned models to get p-values to be 0.01 (for 'LM1') and 0.032 (for 'RM1').

ACKNOWLEDGMENTS

This research was funded in part by the NSF grant IIS-1525431 and IIS-1724174 to BCV and NSF CAREER award 1252725 and R01 EB022883 to VS. XZ and VS were also supported by UW CPCP (U54 AI117924).

CONCLUSION

- ✣ We presented a generalization of the RNN to the SPD manifold and analyzed its theoretical properties.
- ✣ Our proposed framework is fast and needs far fewer parameters than the state-of-the-art.