Improving Generalization of Encoder-Decoder Neural Network with Stochastic Latent Space

R·I·T
B. Thomas Golisano College of Computing and Information Sciences

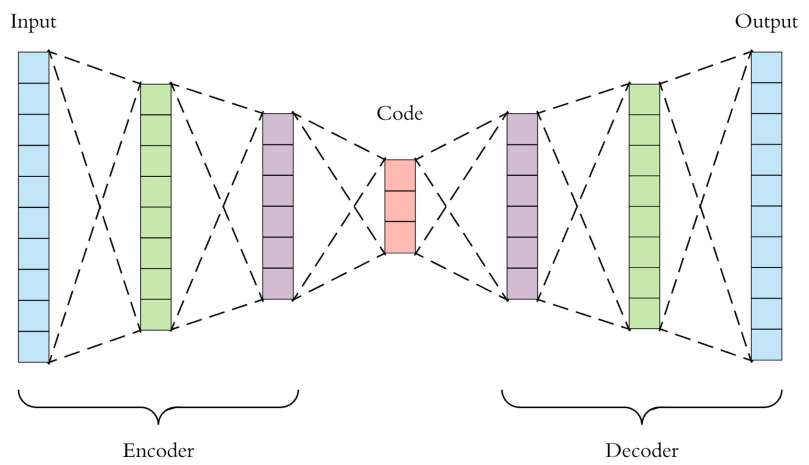
Sandesh Ghimire and Linwei Wang
sandesh.ghimire@mail.rit.edu ; www.sandeshgh.com

Computational Biomedicine Lab

Introduction

Encoder-Decoder Neural Network

- Autoencoder
- Encoder-Decoder



Key Ideas

- How to improve generalization ability of an encoder decoder network?

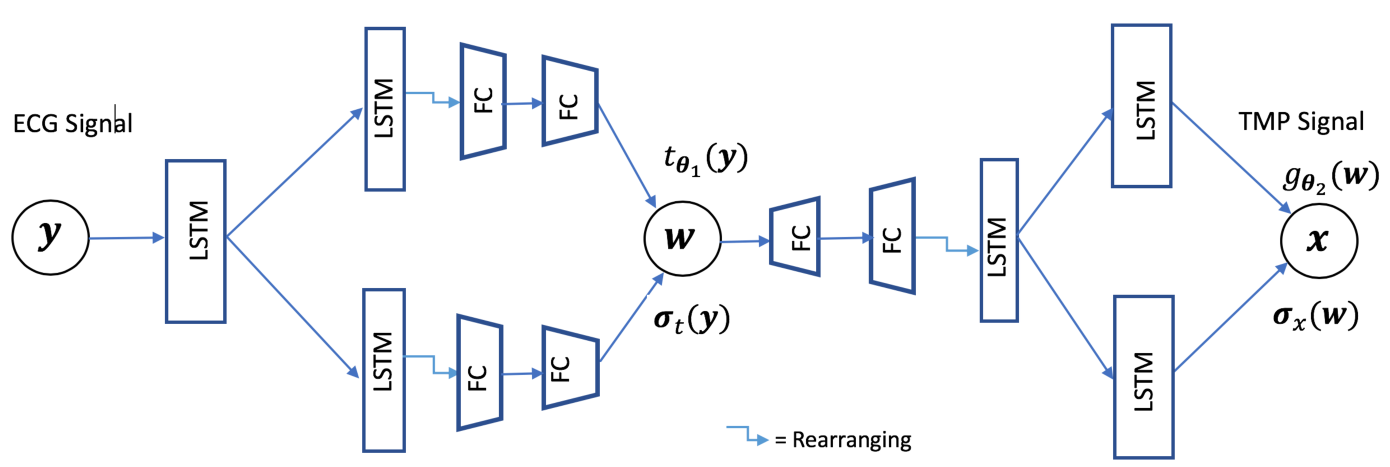
→ With constrained stochasticity of latent space

- Why does constrained stochasticity help?

→ Because it renders decoder less sensitive to variation in latent space

Application to ECGi

- Inverse reconstruction of TMP from ECG signal
- Ill posed inverse problem
- Estimating sequence from sequence



- LSTM and FC layer in both encoder and decoder
- Conditional distributions as Gaussian

Contribution

- Theoretically show that stochastic latent space helps in improving generalization
- Experimentally support the theory with application to ECGi

Theory

Generalization Gap

$\mathcal{T}$ = Encoder

f = loss composite decoder

Theorem 1 ((Kawaguchi and Bengio 2018)). *For any ℓ, let $(\mathcal{T}, f)$ be a pair such that $\mathcal{T} : (\mathcal{Z}, \mathcal{S}) \to ([0,1]^d, \mathcal{B}([0,1]^d))$ is a measurable function, $f : ([0,1]^d, \mathcal{B}([0,1]^d)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is of bounded variation as $V[f] < \infty$, and $\ell(\boldsymbol{x}, h(\boldsymbol{y})) = (f \circ \mathcal{T})(\boldsymbol{z}) \forall \boldsymbol{z} \in \mathcal{Z}$, where $\mathcal{B}(A)$ indicates the Borel σ- algebra on A. Then for any dataset pair (D_m, Z_m) and any $\ell(\boldsymbol{x}, h_{\mathcal{A}(D_n)}(\boldsymbol{y}))$,*

$$E_\mu \ell(\boldsymbol{x}, h_{\mathcal{A}(D_m)}(\boldsymbol{y})) - E_{Z_m} \ell(\boldsymbol{x}, h_{\mathcal{A}(D_n)}(\boldsymbol{y})) \leq V[f]\mathcal{D}^*[\mathcal{T}_*\mu, \mathcal{T}(Z_m)]$$

where $\mathcal{T}_*\mu$ is pushforward measure of μ under the map $\mathcal{T}$.

Implication:
Generalization gap can be reduced by decreasing either variation or discrepancy

Proposition 1 ((Kawaguchi and Bengio 2018)). *Suppose that $f_{j_1,..j_k}$ is a function for which $\partial_{1,\dots k} f_{j_1,..j_k}$ exists on $[0,1]^k$. Then,*

$$V^k[f_{j_1 \dots j_k}] \leq \sup_{\boldsymbol{t}_{j_1},..,\boldsymbol{t}_{j_k} \in [0,1]^k} |\partial_{1,\dots k} f_{j_1,..j_k}(\boldsymbol{t}_{j_1},..,\boldsymbol{t}_{j_k})|.$$

If $\partial_{1,\dots k} f_{j_1,..j_k}$ is also continuous on $[0,1]^k$,

$$V^k[f_{j_1 \dots j_k}] = \int_{[0,1]^k} |\partial_{1,\dots k} f_{j_1,..j_k}(\boldsymbol{t}_{j_1},..,\boldsymbol{t}_{j_k})| dt_{j_1}..dt_{j_k}.$$

$$\partial_{1,\dots k} f_{j_1,..j_k}(\boldsymbol{t}_{j_1},..,\boldsymbol{t}_{j_k}) = \frac{\partial^k \ell}{\partial \boldsymbol{t}_{j_1},..,\partial \boldsymbol{t}_{j_k}}$$

Learning with Stochastic Latent Space

$$p_{\boldsymbol{\theta}_1}(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{t}_{\theta_1}(y), \boldsymbol{\sigma_t}^2(\boldsymbol{y}))$$
$$p_{\boldsymbol{\theta}_2}(\boldsymbol{x}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{g}_{\theta_2}(\boldsymbol{w}), \boldsymbol{\sigma_x}^2(\boldsymbol{w}))$$

$$\text{minimize} - E_{P(x,y)} \log \int p_{\boldsymbol{\theta}_2}(\boldsymbol{x}|\boldsymbol{w}) p_{\boldsymbol{\theta}_1}(\boldsymbol{w}|\boldsymbol{y}) d\boldsymbol{w}$$
$$\text{such that } D_{KL}(p_{\boldsymbol{\theta}_1}(\boldsymbol{w}|\boldsymbol{y}) || \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{I})) < \delta$$

Loss Function

$$\mathcal{L} \leq E_{P(\boldsymbol{x},\boldsymbol{y})}\Big[E_{\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\Big(\sum_i \overbrace{\frac{1}{\boldsymbol{\sigma_{x}}_i^2}(x_i - g_i(\boldsymbol{t} + \boldsymbol{\sigma_t} \odot \boldsymbol{\epsilon}))^2}^{T_1} + \log \boldsymbol{\sigma_{x}}_i^2\Big) + \lambda . D_{KL}(p_{\boldsymbol{\theta}_1}(\boldsymbol{w}|\boldsymbol{y}) || \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{I}))\Big] + constant$$

Result I
(Using Taylor series)

Deterministic loss

$$T_1 = \sum_i \frac{1}{\boldsymbol{\sigma}_{xi}^2}\Big[\ell_i(\boldsymbol{x}_i, \boldsymbol{t}) + \langle \boldsymbol{\sigma}_t \odot E_\epsilon[\boldsymbol{\epsilon}], \frac{\partial}{\partial t}\ell_i(\boldsymbol{x}_i, \boldsymbol{t})\rangle$$
$$+ \frac{1}{2}\langle [\boldsymbol{\sigma}_t \otimes \boldsymbol{\sigma}_t] \odot E_\epsilon[\boldsymbol{\epsilon} \otimes \boldsymbol{\epsilon}], \Big[\frac{\partial^2}{\partial \boldsymbol{t}_{j_1}, \partial \boldsymbol{t}_{j_2}}\ell_i(\boldsymbol{x}_i, \boldsymbol{t})\Big]\rangle + ..$$
$$+ \frac{1}{k!}\langle [\boldsymbol{\sigma}_t \otimes^k \boldsymbol{\sigma}_t] \odot E_\epsilon[\boldsymbol{\epsilon} \otimes^k \boldsymbol{\epsilon}], \Big[\frac{\partial^k}{\partial \boldsymbol{t}_{j_1}, .., \partial \boldsymbol{t}_{j_k}}\ell_i(\boldsymbol{x}_i, \boldsymbol{t})\Big]\rangle + ..\Big]$$

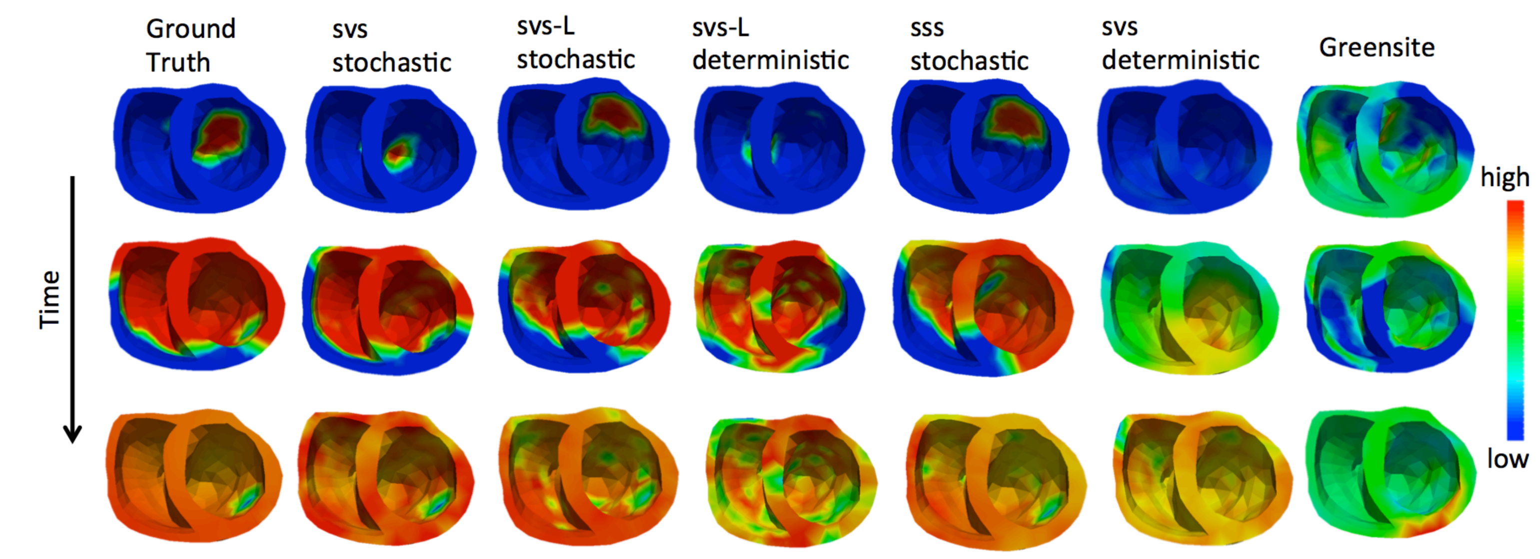
Weight for partial derivative
Tensor modulating noise
Noise tensor
Partial derivative tensor

Experiment

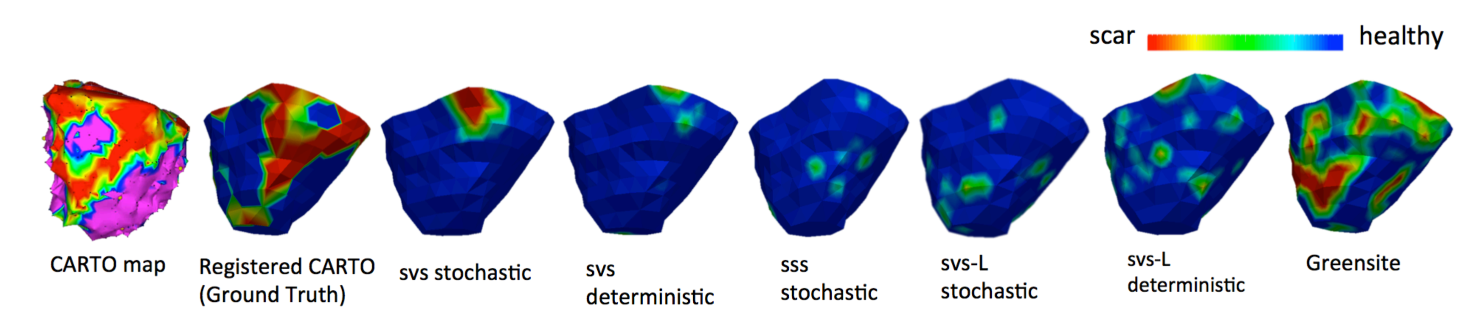
Settings:

- TMP simulated using Aliev Panfilov model on 3 human-torso geometry models.
- Simulations varying origin of excitation and tissue property
- Test data simulated with new origin of excitation and tissue property

Experiment on Simulated data



Experiment on Real data



Stochastic latent space generalizes well in both svs and svs-L architectures qualitatively and quantitatively

Metric	MSE	TMP Corr.	AT Corr.	Dice Coeff.
svs stochastic	**0.037± 0.021**	**0.885± 0.061**	**0.885± 0.072**	**0.645± 0.181**
svs deterministic	0.075 ± 0.013	0.77 ± 0.038	0.12 ± 0.13	0.01 ± 0.006
svs-L stochastic	0.068 ± 0.023	0.838 ± 0.053	0.601 ± 0.074	0.28 ± 0.154
svs-L deterministic	0.067 ± 0.02	0.84 ± 0.053	0.57 ± 0.052	0.165 ± 0.092
Greensite	–	–	0.514 ± 0.006	0.138 ± 0.005

Conclusion

By drawing from analytical learning theory, we have shown that constrained stochasticity of latent variable improves generalization ability and then supported the theory with experiments on electrophysiological imaging.

Acknowledgements

This work is supported in part by the National Institutes of Health [grant no: R21HL125998] and the National Science Foundation [grant no:ACI-1350374]