# Sequential Neural Models with Stochastic Layers

Replication of: Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, Ole Winther
Presenters: Paula Wesselmann, Yeziwei Wang, Áine Cahill

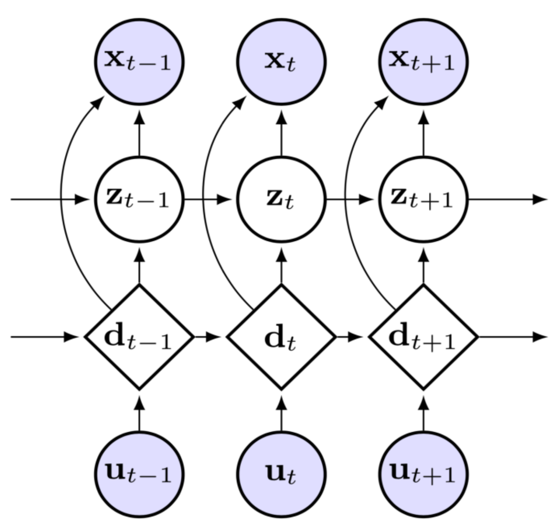
## Introduction

This paper models polyphonic music sequence data using a Stochastic Recurrent Neural Network (SRNN). SRNN combines deterministic history representation of RNNs and stochastic latent variables of SSMs. This gives the SRNN architecture the ability to model multi-modal uncertainty and long-term temporal dependency of sequence data.

## Generative Model

$$p_\theta(\boldsymbol{x},\boldsymbol{z},\boldsymbol{d}|\boldsymbol{u})$$
$$= p_{\theta_x}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{d})p_{\theta_z}(\boldsymbol{z}|\boldsymbol{d},z_0)p_{\theta_d}(\boldsymbol{d}|\boldsymbol{u},d_0)$$
$$= \prod_{t=1}^{T} p_{\theta_x}(x_t|z_t,d_t)p_{\theta_z}(z_t|z_{t-1},d_t)p_{\theta_d}(d_t|d_{t-1},u_t)$$

$p_{\theta_x}$ and $p_{\theta_z}$ are parameterised by NNs. Hidden layer $d_t = f_{\theta_d}(d_{t-1},u_t)$ is deterministically implemented using GRU with distribution $p_{\theta_d}(d_t|d_{t-1},u_t) = \delta(d_t - \tilde{d}_t)$.



} SSM layer models uncertainty in latent states

} RNN layer captures long-term dependencies

## ELBO Training

Parameter learning: max. log-likelihood of training set $\mathcal{L}(\theta) = \max_\theta [\sum_{sequences} \log p_\theta(\boldsymbol{x}|\boldsymbol{u})]$ where

$$p_\theta(\boldsymbol{x}|\boldsymbol{u}) = \iint p_\theta(\boldsymbol{x},\boldsymbol{z},\boldsymbol{d}|\boldsymbol{u}) d\boldsymbol{z}, d\boldsymbol{d}$$

$$= \iint p_{\theta_x}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{d})p_{\theta_z}(\boldsymbol{z}|\boldsymbol{d})p_{\theta_d}(\boldsymbol{d}|\boldsymbol{u}) d\boldsymbol{z}, d\boldsymbol{d}$$

Because of the intractability of maximising log-likelihood, we instead maximise the variational evidence lower bound (ELBO) of the log-likelihood:

$$\mathcal{F}(\theta,\phi)$$
$$= \mathrm{E}_{q_\phi(\boldsymbol{z},\boldsymbol{d}|\boldsymbol{x},\boldsymbol{u})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{d})]$$
$$- KL(q_\phi(\boldsymbol{z},\boldsymbol{d}|\boldsymbol{x},\boldsymbol{u})||p_\theta(\boldsymbol{z},\boldsymbol{d}|\boldsymbol{u}))$$

## Variational Inference

Approximate inference of the model is possible using a network that tracks the factorisation of the model's posterior distribution for training the SRNN,

$$p_\theta(\boldsymbol{z},\boldsymbol{d}|\boldsymbol{x},\boldsymbol{u}) = p_\theta(\boldsymbol{d}|\boldsymbol{u})p_\theta(\boldsymbol{z}|\boldsymbol{d},\boldsymbol{x}) = p_\theta(\boldsymbol{d}|\boldsymbol{u})\prod_t p_\theta(z_t|z_{t-1},d_{t:T},x_{t:T})$$

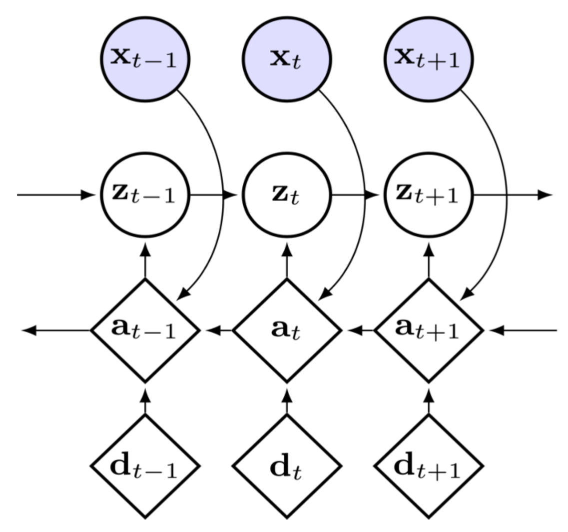
Variational approximation of the posterior is,

$$q_\phi(\boldsymbol{z},\boldsymbol{d}|\boldsymbol{x},\boldsymbol{u}) = p_\theta(\boldsymbol{d}|\boldsymbol{u})\prod_t q_\phi(z_t|z_{t-1},a_t)$$

As both the generative model and inference network factorise over time steps, the ELBO separates as a sum over time steps:

$$\mathcal{F}(\theta,\phi)$$
$$= \sum_t \mathrm{E}_{q_\phi^*(z_{t-1})}[\mathrm{E}_{q_\theta(z|z_{t-1})}[\log p_\theta(x_t|z_t,\tilde{d}_t)]$$
$$- KL(q_\phi(z_t|z_{t-1},\tilde{d}_t)||p_\theta(z_t|z_{t-1},\tilde{d}_t))]$$

Approximate dependency of $z_t$ on $d_{t:T}$ and $x_{t:T}$ by introducing auxiliary deterministic states $a_t$ from an RNN running backwards in time.



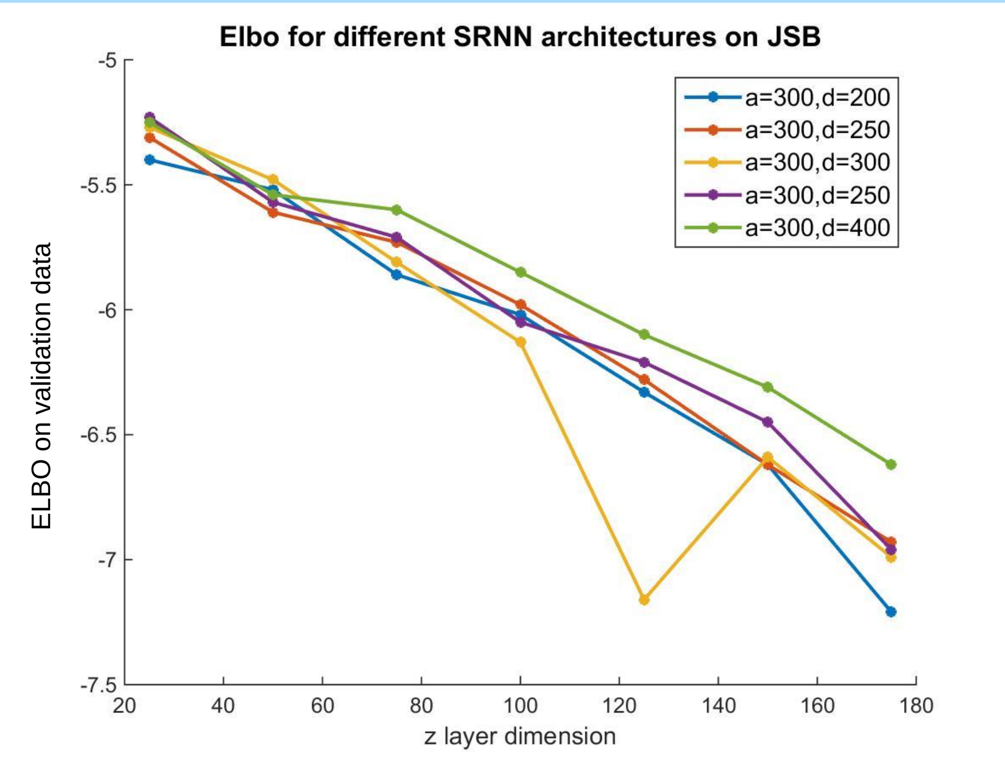
} auxiliary deterministic states

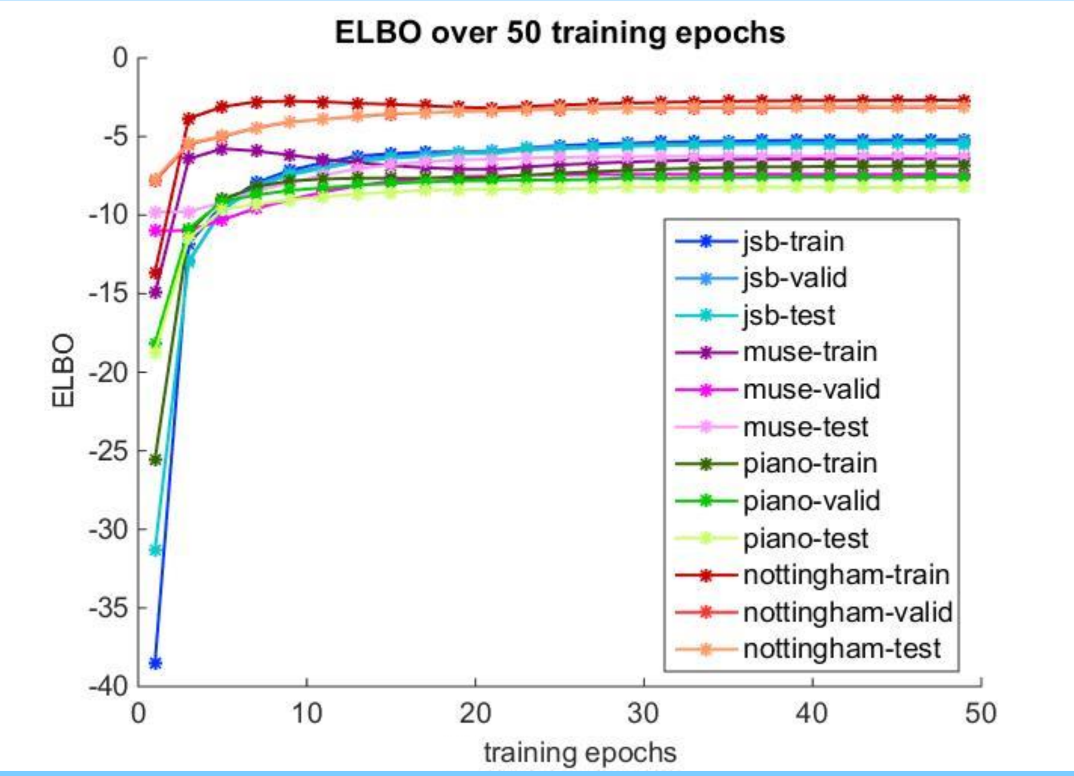
$$a_t = g_\phi(a_{t+1},[d_t,x_t])$$

## Results

| | Nottingham | JSB chorales | MuseData | piano-midi.de |
|---|---|---|---|---|
| SRNN (replication) | -3.18 | -5.37 | -7.43 | -7.57 |
| SRNN (original) | **-2.94** | **-4.74** | **-6.28** | -8.20 |
| RNN [1] | -4.46 | -8.71 | -8.13 | -8.37 |
| TSBN [2] | -3.67 | -7.48 | -6.81 | -7.98 |
| SRNN (small dim) | -3.31 | -5.04 | -7.56 | **-7.46** |
| SRNN (merged data) | -3.27 | -5.64 | -7.64 | -8.03 |

**ELBO values for replication task and references.**




## Experiment

Train SRNN on 4 polyphonic MIDI music datasets of varying tempo and complexity. Approximate log-likelihood using ELBO.

1. Paper replication experiment: 50 training epochs. SRNN: {s=100, a=300, d=300, z=100}. (s: sequence length)
2. Alternative SRNN architectures: 20 training epochs; combinations of s, a, d, z.
3. Investigate over-complexity of model: 20 training epochs on small dimension SRNN {s=100 , a=30, d=30, z=10}.
4. 50 training epochs on combined datasets.

## Discussion

- SRNNs achieve state-of-the-art for speech.
- Music data requires simpler SRNN architecture. Achieved 2% better than original Piano ELBO using small SRNN.
- Replication results for Muse, Nottingham and JSB Chorales are within 18% of ELBO for original results. Replication achieved 8% improvement for Piano data.
- Decreasing z increases ELBO. Negligible effect on ELBO by changing a, d and s.
- Small dimensional SRNN achieves similar ELBO to replication. Improved performance for Piano and JSB Chorales.
- Combining datasets for training and validation achieved 2% improvement of ELBO on Piano dataset compared to the paper's results.

## Future Work

- Use augmented music data to provide more data for training.
- Investigate the optimal architecture for music data of different complexities.
- Investigate evolution of KL-divergence during training. Ensure it does not vanish, meaning the effects of latent variables are not ignored.
- Compare to other latent variable RNN models, including the state-of-the-art on music data, RNN-NADE [1].

**References:**
[1] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv:1206.6392*, 2012.
[2] Z. Gan, C. Li, R. Henao, D. E. Carlson, and L. Carin. Deep temporal sigmoid belief networks for sequence modeling. In *NIPS*, pages 2458–2466, 2015.