Tunable Efficient Unitary Neural Networks (EUNN) and their application to RNNs

Massachusetts Institute of Technology

Li Jing*, Yichen Shen*, Tena Dubček, John Peurifoy, Scott Skirlo, Yann LeCun, Max Tegmark, Marin Soljačić

* equal contribution

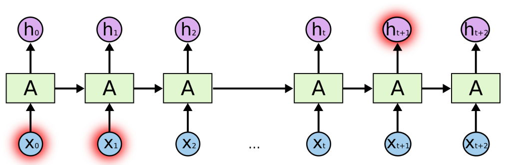
facebook research

Contribution

We present a new architecture for implementing an Efficient Unitary Neural Network (EUNNs)
1) The representation capacity of the unitary space in an EUNN is fully tunable, ranging from a subspace of SU(N) to the entire unitary space.
2) The computational complexity for training an EUNN is merely O(1) per parameter.
3) We find that our architecture significantly outperforms both other state-of-the-art unitary RNNs and the LSTM architecture, in terms of the final performance and/or the wall-clock training speed.

Background

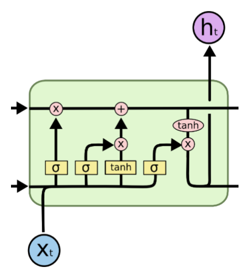
Gradient Vanishing/Explosion Problem



$$\frac{\partial C}{\partial \mathbf{h}^{(t)}} = \frac{\partial C}{\partial \mathbf{h}^{(T)}} \frac{\partial \mathbf{h}^{(T)}}{\partial \mathbf{h}^{(t)}} = \frac{\partial C}{\partial \mathbf{h}^{(T)}} \prod_{k=t}^{T-1} \frac{\partial \mathbf{h}^{(k+1)}}{\partial \mathbf{h}^{(k)}} = \frac{\partial C}{\partial \mathbf{h}^{(T)}} \prod_{k=t}^{T-1} \mathbf{D}^{(k)} \mathbf{W},$$

Conventional Solution: LSTM

require gradient clipping



New Solution: Unitary RNN

In mathematics, a complex square matrix U is unitary if its conjugate transpose U* is also its inverse

$$U^*U = UU^* = I$$

Keep the norm of vectors:

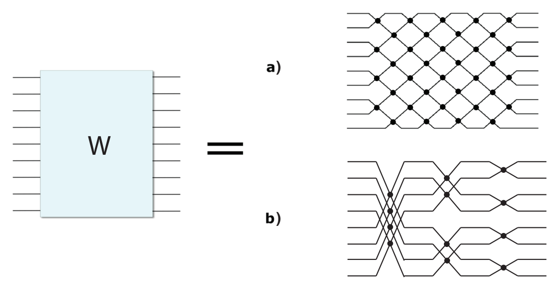
$$||Ux|| = ||x||$$

Related works

- Restricted space Unitary Matrix Paramtrization [1]
- Full-capacity Unitary Matrix by projection [2]
- Orthogonal parametrization by reflection[3]
- Orthogonal matrix by regularization[4]

Model

Tunable Efficient Unitary Parametrization

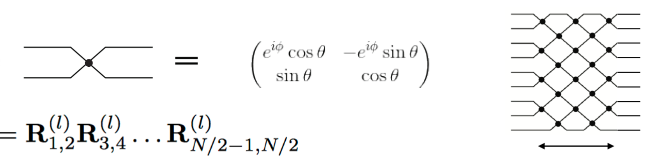


Tunable

$$\mathbf{W} = \mathbf{D}\mathbf{F}_A^{(1)}\mathbf{F}_B^{(2)}\mathbf{F}_A^{(3)}\mathbf{F}_B^{(4)}\cdots\mathbf{F}_B^{(L)}$$

FFT-style

$$\mathbf{W} = \mathbf{D}\mathbf{F}_1\mathbf{F}_2\mathbf{F}_3\mathbf{F}_4\cdots\mathbf{F}_{\log(N)}$$

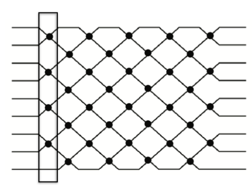


c) $$\mathrm{R}(\theta,\phi) = \begin{pmatrix} e^{i\phi}\cos\theta & -e^{i\phi}\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

$$\mathbf{F}_A^{(l)} = \mathbf{R}_{1,2}^{(l)}\mathbf{R}_{3,4}^{(l)}\cdots\mathbf{R}_{N/2-1,N/2}^{(l)}$$

capacity

Implementation Algorithm



sparse block diagonal matrix

Algorithm 1 Efficient implementation for $\mathbf{F}$ with parameter θ_i and ϕ_i.

Input: input $\mathbf{x}$, size N; parameters θ and ϕ, size $N/2$; constant permuatation index list $\mathbf{ind_1}$ and $\mathbf{ind_2}$.
Output: output $\mathbf{y}$, size N.
$\mathbf{v_1} \leftarrow$ concatenate($\cos\theta$, $\cos\theta * \exp(i\phi)$)
$\mathbf{v_2} \leftarrow$ concatenate($\sin\theta$, - $\sin\theta * \exp(i\phi)$)
$\mathbf{v_1} \leftarrow$ permute($\mathbf{v_1}, \mathbf{ind_1}$)
$\mathbf{v_2} \leftarrow$ permute($\mathbf{v_2}, \mathbf{ind_1}$)
$\mathbf{y} \leftarrow \mathbf{v_1} * \mathbf{x} + \mathbf{v_2} *$ permute($\mathbf{x}, \mathbf{ind_2}$)

no need to implement back propagation

Complexity

Model	Time complexity of one online gradient step	number of parameters in the hidden matrix	Transition matrix search space
URNN	$\mathcal{O}(TN\log N)$	$\mathcal{O}(N)$	subspace of $\mathbf{U}(N)$
PURNN	$\mathcal{O}(TN^2+N^3)$	$\mathcal{O}(N^2)$	full space of $\mathbf{U}(N)$
EURNN (tunable style)	$\mathcal{O}(TNL)$	$\mathcal{O}(NL)$	tunable space of $\mathbf{U}(N)$
EURNN (FFT style)	$\mathcal{O}(TN\log N)$	$\mathcal{O}(N\log N)$	subspace of $\mathbf{U}(N)$

Advantages

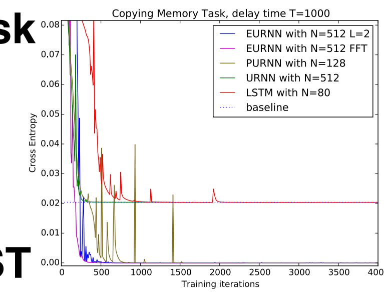
1) **Efficient**: O(1) operation per parameter
2) **Tunable**: span from small subspace to full unitary space
3) **Easy implementation**: element-wise functions
4) **FFT approximation** provides further speed-up

Experiment

We compare our model to LSTM and other unitary RNN with same number of parameters

Copying Memory Task

EURNN outperforms all other models in long memory task.

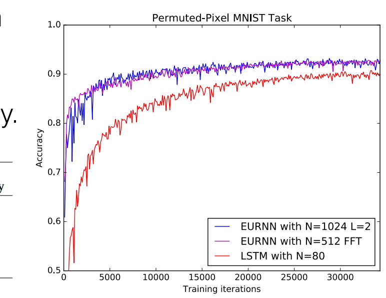


Pixel Permuted MNIST

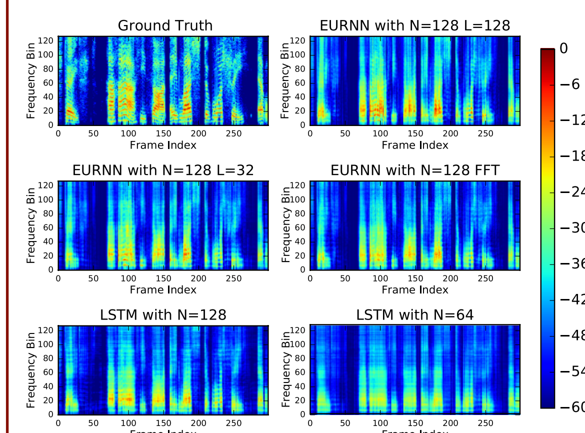
EURNN outperforms LSTM in both final performance and speed.

EURNN achieves highest accuracy.

Model	hidden size (capacity)	number of parameters	validation accuracy	test accuracy
LSTM	80	16k	0.908	0.902
URNN	512	16k	**0.942**	0.933
PURNN	116	16k	0.922	0.921
EURNN (tunable style)	1024 (2)	13.3k	0.940	**0.937**
EURNN (FFT style)	512 (FFT)	9.0k	0.928	0.925



TIMIT Speech Spectrum Prediction



TIMIT spectrum sampled in 8 GHz, normalized, Fourier transformed.

Each RNN model is required to predict next frame given previous frames.

EURNN outperforms LSTM in real task.

Reference & Code

[1] M. Arjovsky, A. Shah, and Y. Bengio, "Unitary Evolution Recurrent Neural Networks," ICML 2016
[2] S. Wisdom, T. Powers, et al, "Full-Capacity Unitary Recurrent Neural Networks," NIPS 2016.
[3] Z. Mhammedi, A. Hellicar, A. Rahman, J. Bailey, "Efficient Orthogonal Parametrisation of Recurrent Neural Networks Using Householder Reflections", ICML 2017
[4] E. Vorontsov, C. Trabelsi, S. Kadoury, and C. Pal, "On orthogonality and learning recurrent networks with long term dependencies", ICML 2017

Tensorflow: https://github.com/jingli9111/EUNN-tensorflow
PyTorch: https://github.com/jingli9111/URNN-PyTorch
Theano: https://github.com/iguanaus/EUNN-theano

ICML @ Sydney, August 8 2017