Gated Complex Recurrent Neural Networks

Moritz Wolter, Angela Yao

wolter@cs.uni-bonn.de

July 13, 2018
Motivation

- RNN and neural networks in general suffer from unstable gradients.
- Distribution over a sum using gating is one fix for vanishing gradients (GRU, LSTM, ...)
- Norm preserving matrices are another way to fix this. \[ \|W h\|_2 = \|h\|_2 \]
- Orthogonal (real) and unitary (complex) matrices are norm preserving.
Motivation

- Unitary matrices are more expressive than orthogonal ones.
- Complex networks must be interoperable with real components.
- Mappings from $\mathbb{C}$ to $\mathbb{R}$ are not complex differentiable.
Wirtinger-Calculus [Wir27][MG09][KD09]

For a complex function \( f(z) = u(x, y) - iv(x, y) \) we have:

\[
\mathbb{R}\text{-derivative} \triangleq \left. \frac{\partial f}{\partial z} \right|_{\bar{z} = \text{const}} = \frac{1}{2} \left( \frac{\partial f}{\partial x} - i \frac{\partial f}{\partial y} \right),
\]

(1)

\[
\overline{\mathbb{R}}\text{-derivative} \triangleq \left. \frac{\partial f}{\partial \bar{z}} \right|_{z = \text{const}} = \frac{1}{2} \left( \frac{\partial f}{\partial x} + i \frac{\partial f}{\partial y} \right).
\]

(2)

Based on these derivatives, one can define the chain rule for a function \( g(f(z)) \) as follows:

\[
\frac{\partial g(f(z))}{\partial z} = \frac{\partial g}{\partial f} \frac{\partial f}{\partial z} + \frac{\partial g}{\partial \bar{f}} \frac{\partial \bar{f}}{\partial z}
\]

where \( \bar{f} = u(x, y) - iv(x, y) \).

(3)
\[ x_t = W_{rec} f(x_{t-1}) + W_{in} u_t + b. \]  (4)

\[ \frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \leq t \leq T} \frac{\mathcal{E}_t}{\partial \theta}, \]  (5)

\[ \frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \leq k \leq t} \left( \frac{\mathcal{E}_t}{\partial x_t} \frac{\partial x_t}{\partial x_k} \frac{\partial^+ x_k}{\partial \theta} \right), \]  (6)

\[ \frac{\partial x_t}{\partial x_k} = \prod_{t \geq i > k} \frac{\partial x_i}{\partial x_{i-1}} = \prod_{t \geq i > k} W_{rec}^T \text{diag}(f'(x_{i-1})). \]  (7)
Stiefel Manifold Weight Updates \([WPH^{+16}]\)

\[ W_{k+1} = (I + \frac{\lambda}{2} A_k)^{-1}(I - \frac{\lambda}{2} A_k)W_k, \]  

(8)

where \( A = W\nabla_w F^T - W^T \nabla_w F. \)  

(9)

**Figure:** Fix the optimized matrix eigenvalues onto the unit circle. The key idea behind stiefel-manifold optimization.
Unitary evolution network performance

\[ x_t = U_{rec} f(x_{t-1}) + W_{in} u_t + b. \]  

Figure: Current state of the art performance on memory and adding problem for \( T=250 \). Models have approximately 40k weights.
Complex equivalents of tanh and Relu

\[ f_{\text{Hirose}}(z) = \tanh \left( \frac{|z|}{m^2} \right) e^{-i \cdot \theta z} = \tanh \left( \frac{|z|}{m^2} \right) \frac{z}{|z|}, \quad (11) \]

\[ f_{\text{modReLU}}(z) = \text{ReLU}(|z| + b) e^{-i \cdot \theta z} = \text{ReLU}(|z| + b) \frac{z}{|z|}. \quad (12) \]

We will compare their performance as state-to-state non-linearities.
Complex gated Recurrent Recurrent Nets

Gate equation:

\[ g_r = f_g(z_r), \quad \text{where} \quad z_r = W_r h + V_r x_t + b_r, \quad (13) \]
\[ g_z = f_g(z_z), \quad \text{where} \quad z_z = W_z h + V_z x_t + b_z, \quad (14) \]

Update equations:

\[ \tilde{z}_t = W(g_r \odot h_{t-1}) + Vx_t + b, \quad (15) \]
\[ h_t = g_z \odot f_a(\tilde{z}_t) + (1 - g_z) \odot h_{t-1}, \quad (16) \]

\[ \mathbb{C} \rightarrow \mathbb{R}, \text{ mapping:} \]
\[ o_r = W_o[\Re(h) \Im(h)] + b_o. \quad (17) \]
Complex gate activations

\[ f_{\text{prod}}(z) = \sigma(\Re(z)) \cdot \sigma(\Im(z)), \]  \hspace{1cm} (18)

\[ f_{\text{gate hirose}} = \tanh\left(\frac{|z|}{m^2}\right)\sigma\left(a \frac{z}{|z|} + b\right), \]  \hspace{1cm} (19)

\[ f_{\text{mod sigmoid}}(z) = \sigma\left(\alpha \Re(z) + \beta \Im(z)\right). \]  \hspace{1cm} (20)

With \( \alpha \in [0, 1] \) and \( \beta = (1 - \alpha) \).
Comparison to state of the art

Figure: Comparison of our complex gated RNN (cgRNN, blue, $n_h = 80$) with the unitary RNN [ASB16] (uRNN, orange, $n_h = 140$) and standard GRU [CvMG+14] (orange, $n_h = 112$) on the memory (left) and adding (right) problem for $T = 250$. 
Stiefel optimization and activations

Figure: Comparison of non-linearities and norm preserving state transition matrices on the complex gated RNNs for the memory (a) and adding (b) problems for $T=250$. We use $n_h = 80$ for all experiments.
Figure: Motion prediction Euler angle errors for the complex gated RNN (green) versus GRU (blue), where each line indicates a separate test sequence. The final error after 20,000 iterations is shown in the adjacent table.
Gates must be able to saturate to work!

In order to further stabilize the gradients we explored normalizing the recurrent matrices in the gate equations.

**Figure:** Orthogonal recurrent gate matrices prevent the gates from functioning.
Complex gate coupling. Just one complex gate equation, \( r = \sigma(\Re(g)) \), \( z = \sigma(\Im(g)) \). Reduces complex overhead.

- Explore frequency domain networks using Hilbert or Fourier transformed input data.

- Explore dynamic mode decomposition as an alternative complex input representation.
References I


Complex Gated Recurrent Neural networks

References II


Feedback

Thanks for your attention and feedback.
Later: wolter@cs.uni-bonn.de