(Srivastava et al., 2015).

$$
\begin{aligned}
\mathbf{c}_t &= \mathrm{s}\left(\mathbf{W}^{\mathbf{ch}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{cx}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{c}}\right) && (2)\\
\mathbf{g}_t &= \sigma\left(\mathbf{W}^{\mathbf{gh}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{gx}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{g}} + b^{fg}\right) && (3)\\
\mathbf{h}_t &= \mathbf{g}_t \cdot \mathbf{h}_{t-1} + (\mathbf{1} - \mathbf{g}_t) \cdot \mathbf{c}_t && (4)
\end{aligned}
$$

### GRU - Gated Recurrent Unit (Cho et al., 2014)

$$
\begin{aligned}
\mathbf{r}_t &= \sigma\left(\mathbf{W}^{\mathbf{rh}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{rx}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{r}}\right) && (5)\\
\mathbf{u}_t &= \sigma\left(\mathbf{W}^{\mathbf{uh}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{ux}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{u}} + b^{fg}\right) && (6)\\
\mathbf{c}_t &= \mathrm{s}\left(\mathbf{W}^{\mathbf{ch}}\,(\mathbf{r}_t \cdot \mathbf{h}_{t-1}) + \mathbf{W}^{\mathbf{cx}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{c}}\right) && (7)\\
\mathbf{h}_t &= \mathbf{u}_t \cdot \mathbf{h}_{t-1} + (\mathbf{1} - \mathbf{u}_t) \cdot \mathbf{c}_t && (8)
\end{aligned}
$$

### LSTM - Long Short Term Memory(Hochreiter & Schmidhuber, 1997)

$$
\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W}^{\mathbf{ih}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{ix}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{i}}\right) && (9)\\
\mathbf{f}_t &= \sigma\left(\mathbf{W}^{\mathbf{fh}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{fx}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{f}} + b^{fg}\right) && (10)\\
\mathbf{c}_t^{in} &= \mathrm{s}\left(\mathbf{W}^{\mathbf{ch}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{cx}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{c}}\right) && (11)\\
\mathbf{c}_t &= \mathbf{f}_t \cdot \mathbf{c}_{t-1} + \mathbf{i}_t \cdot \mathbf{c}_t^{in} && (12)\\
\mathbf{o}_t &= \sigma\left(\mathbf{W}^{\mathbf{oh}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{ox}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{o}}\right) && (13)\\
\mathbf{h}_t &= \mathbf{o}_t \cdot \tanh(\mathbf{c}_t) && (14)
\end{aligned}
$$

### +RNN - Intersection RNN

Due to the success of the UGRNN for shallower architectures in this study (see later figures on trainability), as well as some of the observed trainability problems for both the LSTM and GRU for deeper architectures (e.g. Figure 4h) we developed the Intersection RNN (denoted with a '+') architecture with a coupled depth gate in addition to a coupled recurrent gate. Additional influences for this architecture were the recurrent gating of the LSTM and GRU, and the depth gating from the highway network (Srivastava et al., 2015). This architecture has recurrent input, $\mathbf{h}_{t-1}$, and depth input, $\mathbf{x}_t$. It also has recurrent output, $\mathbf{h}_t$, and depth output, $\mathbf{y}_t$. Note that this architecture only applies between layers where $\mathbf{x}_t$ and $\mathbf{y}_t$ have the same dimension, and is not appropriate for networks with a depth of 1 (we exclude depth one +RNNs in our experiments).

$$
\begin{aligned}
\mathbf{y}_t^{in} &= \mathrm{s1}\left(\mathbf{W}^{\mathbf{yh}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{yx}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{y}}\right) && (15)\\
\mathbf{h}_t^{in} &= \mathrm{s2}\left(\mathbf{W}^{\mathbf{hh}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{hx}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{h}}\right) && (16)\\
\mathbf{g}_t^{y} &= \sigma\left(\mathbf{W}^{\mathbf{g^y h}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{g^y x}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{g}^y} + b^{fg,y}\right) && (17)\\
\mathbf{g}_t^{h} &= \sigma\left(\mathbf{W}^{\mathbf{g^h h}}\,\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{g^h x}}\,\mathbf{x}_t + \mathbf{b}^{\mathbf{g}^h} + b^{fg,h}\right) && (18)\\
\mathbf{y}_t &= \mathbf{g}_t^{y} \cdot \mathbf{x}_t + (1 - \mathbf{g}_t^{y}) \cdot \mathbf{y}_t^{in} && (19)\\
\mathbf{h}_t &= \mathbf{g}_t^{h} \cdot \mathbf{h}_{t-1} + (1 - \mathbf{g}_t^{h}) \cdot \mathbf{h}_t^{in} && (20)
\end{aligned}
$$

In practice we used ReLU for s1 and $\tanh$ for s2.

## 2 Capacity Experiments

### 2.1 Per-parameter capacity

A foundational result in machine learning is that a single-layer perceptron with $N^2$ parameters can store at least 2 bits of information per parameter (Cover, 1965; Gardner, 1988; Baldi & Venkatesh, 1987). More precisely, a perceptron can implement a mapping from $2N$, $N$-dimensional, input vectors to arbitrary $N$-dimensional binary output vectors, subject only to the extremely weak restriction that the input vectors be in general position. RNNs provide a far more complex input-output mapping, with hidden units, recurrent dynamics, and a diversity of nonlinearities. Nonetheless, we wondered if there were analogous capacity results for RNNs that we might be able to observe empirically.