# Approximating Stacked and Bidirectional Recurrent Architectures with the Delayed Recurrent Neural Network: Supplementary Material

## A. Theorem 1 Proof

Let us recall the notation introduced in the main paper. We use superscript $(i)$ to refer to a weight matrix or vector related to layer $i$ in a stacked network, e.g., $\mathbf{W}_{\mathbf{h}}^{(i)}$, or $\mathbf{h}_t^{(i)}$. For a single-layer d-RNN, we refer to weight matrices and related vectors with "hat", e.g., $\hat{\mathbf{W}}_{\mathbf{h}}$ or $\hat{\mathbf{h}}_t$. Additionally, we define the block notation as subvector $\hat{\mathbf{v}}_t^{\{i\}}$ refers to the $i$-th block of vector $\hat{\mathbf{v}}_t$ composed of $k$ blocks. The blocks follow the definition in Equations (3)-(5).

*Proof of Theorem 1.* We prove Theorem 1 by induction on the sequence length $t$. First, we show that for $t = 1$ the stacked RNN and the d-RNN with the constrained weights are equivalent. Namely, for $t = 1$ we show that the outputs and the hidden states are the same, i.e. $\hat{\mathbf{y}}_k = \mathbf{y}_1$ and $\hat{\mathbf{h}}_i^{\{i\}} = \mathbf{h}_1^{(i)}$, respectively. Without loss of generality, we have for any $i$ in $1 \dots k$ the following:

$$
\begin{aligned}
\hat{\mathbf{h}}_i^{\{i\}} &= f^{\{i\}}\left(\hat{\mathbf{W}}_{\mathbf{x}}\mathbf{x}_i + \hat{\mathbf{W}}_{\mathbf{h}}\hat{\mathbf{h}}_{i-1} + \hat{\mathbf{b}}_{\mathbf{h}}\right) \\
&= f\left(\hat{\mathbf{W}}_{\mathbf{x}}^{\{i\}}\mathbf{x}_i + \mathbf{W}_{\mathbf{x}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{h}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i\}} + \mathbf{b}_{\mathbf{h}}^{(i)}\right) \\
&= f\left(\mathbf{0} + \mathbf{W}_{\mathbf{x}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{h}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{h}}^{(i)}\right) \\
&= f\Big(\mathbf{W}_{\mathbf{x}}^{(i)}\cdot \\
&\quad f\left(\mathbf{W}_{\mathbf{x}}^{(i-1)}\hat{\mathbf{h}}_{i-2}^{\{i-2\}} + \mathbf{W}_{\mathbf{h}}^{(i-1)}\hat{\mathbf{h}}_{i-2}^{\{i-1\}} + \mathbf{b}_{\mathbf{h}}^{(i-1)}\right) \\
&\quad + \mathbf{W}_{\mathbf{h}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{h}}^{(i)}\Big) \\
&= f\Big(\mathbf{W}_{\mathbf{x}}^{(i)}\cdot \\
&\quad f\left(\mathbf{W}_{\mathbf{x}}^{(i-1)}\hat{\mathbf{h}}_{i-2}^{\{i-2\}} + \mathbf{W}_{\mathbf{h}}^{(i-1)}\mathbf{h}_0^{(i-1)} + \mathbf{b}_{\mathbf{h}}^{(i-1)}\right) \\
&\quad + \mathbf{W}_{\mathbf{h}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{h}}^{(i)}\Big) \\
&= \dots \\
&= f\Big(\mathbf{W}_{\mathbf{x}}^{(i)} \dots f\Big(\mathbf{W}_{\mathbf{x}}^{(j)} \dots \\
&\quad f\Big(\mathbf{W}_{\mathbf{x}}^{(2)} \cdot f\left(\mathbf{W}_{\mathbf{x}}^{(1)}\mathbf{x}_1 + \mathbf{W}_{\mathbf{h}}^{(1)}\mathbf{h}_0^{(1)} + \mathbf{b}_{\mathbf{h}}^{(1)}\right) \\
&\quad + \mathbf{W}_{\mathbf{h}}^{(2)}\mathbf{h}_0^{(2)} + \mathbf{b}_{\mathbf{h}}^{(2)}\Big) \\
&\quad \dots + \mathbf{W}_{\mathbf{h}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{h}}^{(j)}\Big) \dots + \mathbf{W}_{\mathbf{h}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{h}}^{(i)}\Big) \\
&= \dots \\
&= f\Big(\mathbf{W}_{\mathbf{x}}^{(i)} \dots \\
&\quad f\left(\mathbf{W}_{\mathbf{x}}^{(j)}\mathbf{h}_1^{(j-1)} + \mathbf{W}_{\mathbf{h}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{h}}^{(j)}\right) \\
&\quad \dots + \mathbf{W}_{\mathbf{h}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{h}}^{(i)}\Big) \\
&= \dots \\
&= f\left(\mathbf{W}_{\mathbf{x}}^{(i)}\mathbf{h}_1^{(i-1)} + \mathbf{W}_{\mathbf{h}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{h}}^{(i)}\right) \\
&= \mathbf{h}_1^{(i)},
\end{aligned}
$$

where we used the initialization assumption $\hat{\mathbf{h}}_{i-1}^{\{i\}} = \mathbf{h}_0^{(i)}$ for all $i = 1 \dots k$, and the definition of the hidden state in Equations (3)-(4) for $j - 1$ blocks, in the previous steps. In particular, we have for $j = k$,

$$\hat{\mathbf{h}}_k^{\{k\}} = \mathbf{h}_1^{(k)}.$$

Plugging this result and the definition of the output weights and biases in Equation (8) into Equation (2) for computing the output, we obtain

$$
\begin{aligned}
\hat{\mathbf{y}}_k &= g\left(\hat{\mathbf{W}}_{\mathbf{o}}\hat{\mathbf{h}}_k + \hat{\mathbf{b}}_{\mathbf{o}}\right) \\
&= g\left(\mathbf{W}_{\mathbf{o}}\hat{\mathbf{h}}_k^{\{k\}} + \mathbf{b}_{\mathbf{o}}\right) \\
&= g\left(\mathbf{W}_{\mathbf{o}}\mathbf{h}_1^{(k)} + \mathbf{b}_{\mathbf{o}}\right) \\
&= \mathbf{y}_1. \qquad \text{(A.1)}
\end{aligned}
$$

Which concludes the basis of the induction.

Next, we assume that $\hat{\mathbf{h}}_{t+i-1}^{\{i\}} = \mathbf{h}_t^{(i)}$ for all $1 \le i \le k$ and $t \le T-1$, and prove that it holds for the hidden states for all layers when $t = T$: $\hat{\mathbf{h}}_{T+i-1}^{\{i\}} = \mathbf{h}_T^{(i)}, \forall 1 \le i \le k$. Without loss of generality, we have for the hidden state $\hat{\mathbf{h}}_{T+i-1}^{\{i\}}$ in

constrained weights single-layer d-RNN that,

$$
\begin{aligned}
\hat{\mathbf{h}}^{\{i\}}_{T+i-1} &= f^{\{i\}}\left(\hat{\mathbf{W}}_{\mathbf{x}}\mathbf{x}_{T+i-1} + \hat{\mathbf{W}}_{\mathbf{h}}\hat{\mathbf{h}}_{T+i-2} + \hat{\mathbf{b}}_{\mathbf{h}}\right) \\
&= f\left(\hat{\mathbf{W}}^{\{i\}}_{\mathbf{x}}\mathbf{x}_{T+i-1} + \mathbf{W}^{(i)}_{\mathbf{x}}\hat{\mathbf{h}}^{\{i-1\}}_{T+i-2}\right. \\
&\quad \left. + \mathbf{W}^{(i)}_{\mathbf{h}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{h}}\right) \\
&= f\left(\mathbf{0} + \mathbf{W}^{(i)}_{\mathbf{x}}\hat{\mathbf{h}}^{\{i-1\}}_{T+i-2} + \mathbf{W}^{(i)}_{\mathbf{h}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{h}}\right) \\
&= f\left(\mathbf{W}^{(i)}_{\mathbf{x}} \cdot f\left(\mathbf{W}^{(i-1)}_{\mathbf{x}}\hat{\mathbf{h}}^{\{i-2\}}_{T+i-3} + \right.\right. \\
&\quad \left.\left. \mathbf{W}^{(i-1)}_{\mathbf{h}}\hat{\mathbf{h}}^{\{i-1\}}_{T+i-3} + \mathbf{b}^{(i-1)}_{\mathbf{h}}\right)\right. \\
&\quad \left. + \mathbf{W}^{(i)}_{\mathbf{h}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{h}}\right) \\
&= \dots \\
&= f\left(\mathbf{W}^{(i)}_{\mathbf{x}} \dots f\left(\mathbf{W}^{(j)}_{\mathbf{x}} \dots \right.\right. \\
&\quad f\left(\mathbf{W}^{(2)}_{\mathbf{x}} f\left(\mathbf{W}^{(1)}_{\mathbf{x}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{h}}\hat{\mathbf{h}}^{\{1\}}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{h}}\right)\right. \\
&\quad \left. + \mathbf{W}^{(2)}_{\mathbf{h}}\hat{\mathbf{h}}^{\{2\}}_{T} + \mathbf{b}^{(2)}_{\mathbf{h}}\right) \\
&\quad \left. \dots + \mathbf{W}^{(j)}_{\mathbf{h}}\hat{\mathbf{h}}^{\{j\}}_{T+j-2} + \mathbf{b}^{(j)}_{\mathbf{h}}\right) \\
&\quad \left. \dots + \mathbf{W}^{(i)}_{\mathbf{h}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{h}}\right)
\end{aligned}
$$

From the inductive assumption we have $\hat{\mathbf{h}}^{\{j\}}_{T+j-2} = \mathbf{h}^{(j)}_{T-1}$ for all $1 \leq j \leq k$, then it follows

$$
\begin{aligned}
\hat{\mathbf{h}}^{\{i\}}_{T+i-1} &= f\left(\mathbf{W}^{(i)}_{\mathbf{x}} \dots f\left(\mathbf{W}^{(j)}_{\mathbf{x}} \dots \right.\right. \\
&\quad f\left(\mathbf{W}^{(2)}_{\mathbf{x}} f\left(\mathbf{W}^{(1)}_{\mathbf{x}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{h}}\mathbf{h}^{(1)}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{h}}\right)\right. \\
&\quad \left. + \mathbf{W}^{(2)}_{\mathbf{h}}\mathbf{h}^{(2)}_{T-1} + \mathbf{b}^{(2)}_{\mathbf{h}}\right) \\
&\quad \left. \dots + \mathbf{W}^{(j)}_{\mathbf{h}}\mathbf{h}^{(j)}_{T-1} + \mathbf{b}^{(j)}_{\mathbf{h}}\right) \\
&\quad \left. \dots + \mathbf{W}^{(i)}_{\mathbf{h}}\mathbf{h}^{(i)}_{T-1} + \mathbf{b}^{(i)}_{\mathbf{h}}\right) \\
&= f\left(\mathbf{W}^{(i)}_{\mathbf{x}} \dots f\left(\mathbf{W}^{(j)}_{\mathbf{x}} \dots \right.\right. \\
&\quad f\left(\mathbf{W}^{(2)}_{\mathbf{x}}\mathbf{h}^{(1)}_{T} + \mathbf{W}^{(2)}_{\mathbf{h}}\mathbf{h}^{(2)}_{T-1} + \mathbf{b}^{(2)}_{\mathbf{h}}\right) \\
&\quad \left. \dots + \mathbf{W}^{(j)}_{\mathbf{h}}\mathbf{h}^{(j)}_{T-1} + \mathbf{b}^{(j)}_{\mathbf{h}}\right) \\
&\quad \left. \dots + \mathbf{W}^{(i)}_{\mathbf{h}}\mathbf{h}^{(i)}_{T-1} + \mathbf{b}^{(i)}_{\mathbf{h}}\right) \\
&= \dots \\
&= f\left(\mathbf{W}^{(i)}_{\mathbf{x}} \dots \right. \\
&\quad f\left(\mathbf{W}^{(j)}_{\mathbf{x}}\mathbf{h}^{(j-1)}_{T} + \mathbf{W}^{(j)}_{\mathbf{h}}\mathbf{h}^{(j)}_{T-1} + \mathbf{b}^{(j)}_{\mathbf{h}}\right) \\
&\quad \left. \dots + \mathbf{W}^{(i)}_{\mathbf{h}}\mathbf{h}^{(i)}_{0} + \mathbf{b}^{(i)}_{\mathbf{h}}\right) \\
&= \dots \\
&= f\left(\mathbf{W}^{(i)}_{\mathbf{x}}\mathbf{h}^{(i-1)}_{T} + \mathbf{W}^{(i)}_{\mathbf{h}}\mathbf{h}^{(i)}_{T-1} + \mathbf{b}^{(i)}_{\mathbf{h}}\right) \\
&= \mathbf{h}^{(i)}_{T},
\end{aligned}
$$

where we used the definition of the hidden states in Equations (3)-(4). In particular, we have for $i = k$ that $\hat{\mathbf{h}}^{\{k\}}_{T+k-1} = \mathbf{h}^{(k)}_{T}$.

Now, we show that $\hat{\mathbf{y}}_{T+k-1} = \mathbf{y}_T$. By the definition of the output weights and biases in Equation (8). and by the fact that $\hat{\mathbf{h}}^{\{k\}}_{T+k-1} = \mathbf{h}^{(k)}_{T}$, we obtain

$$
\begin{aligned}
\hat{\mathbf{y}}_{T+k-1} &= g\left(\hat{\mathbf{W}}_{\mathbf{o}}\hat{\mathbf{h}}_{T+k-1} + \hat{\mathbf{b}}_{\mathbf{o}}\right) \\
&= g\left(\mathbf{W}_{\mathbf{o}}\hat{\mathbf{h}}^{\{k\}}_{T+k-1} + \mathbf{b}_{\mathbf{o}}\right) \\
&= g\left(\mathbf{W}_{\mathbf{o}}\mathbf{h}^{(k)}_{T} + \mathbf{b}_{\mathbf{o}}\right) \\
&= \mathbf{y}_T, \qquad \text{(A.2)}
\end{aligned}
$$

which completes the proof. ■

## B. Extension to d-LSTMs

A Long Short-Term Memory recurrent cell (Hochreiter & Schmidhuber, 1997) is given by the introduction of a cell state and a series of gates that control the updates of the states. The cell state together with the gates aim to solve the vanishing gradients problems in the RNN. The LSTM cell is highly popular and we refer to the following implementation:

$$\hat{\mathbf{e}}_t = \sigma\left(\hat{\mathbf{W}}_{\mathbf{xe}}\mathbf{x}_t + \hat{\mathbf{W}}_{\mathbf{he}}\hat{\mathbf{h}}_{t-1} + \hat{\mathbf{b}}_{\mathbf{e}}\right), \qquad \text{(B.3)}$$

$$\hat{\mathbf{f}}_t = \sigma\left(\hat{\mathbf{W}}_{\mathbf{xf}}\mathbf{x}_t + \hat{\mathbf{W}}_{\mathbf{hf}}\hat{\mathbf{h}}_{t-1} + \hat{\mathbf{b}}_{\mathbf{f}}\right), \qquad \text{(B.4)}$$

$$\hat{\mathbf{o}}_t = \sigma\left(\hat{\mathbf{W}}_{\mathbf{xo}}\mathbf{x}_t + \hat{\mathbf{W}}_{\mathbf{ho}}\hat{\mathbf{h}}_{t-1} + \hat{\mathbf{b}}_{\mathbf{o}}\right), \qquad \text{(B.5)}$$

$$\hat{\mathbf{g}}_t = \tanh\left(\hat{\mathbf{W}}_{\mathbf{xc}}\mathbf{x}_t + \hat{\mathbf{W}}_{\mathbf{hc}}\hat{\mathbf{h}}_{t-1} + \hat{\mathbf{b}}_{\mathbf{c}}\right), \qquad \text{(B.6)}$$

$$\hat{\mathbf{c}}_t = \hat{\mathbf{f}}_t \odot \hat{\mathbf{c}}_{t-1} + \hat{\mathbf{e}}_t \odot \hat{\mathbf{g}}_t, \qquad \text{(B.7)}$$

$$\hat{\mathbf{h}}_t = \hat{\mathbf{o}}_t \odot \tanh\left(\hat{\mathbf{c}}_t\right), \qquad \text{(B.8)}$$

where $\hat{\mathbf{e}}_t$ is the input gate, $\hat{\mathbf{f}}_t$ the forget gate, $\hat{\mathbf{o}}_t$ the output gate, $\hat{\mathbf{g}}_t$ the cell gate, $\hat{\mathbf{c}}_t$ the cell state, and $\hat{\mathbf{h}}_t$ the hidden state. The weight matrices are symbolized $\hat{\mathbf{W}}_{\mathbf{xa}}$ and $\hat{\mathbf{W}}_{\mathbf{ha}}$ as well as the bias $\hat{\mathbf{b}}_{\mathbf{a}}$, with $\mathbf{a} \in \{\mathbf{e}, \mathbf{c}, \mathbf{f}, \mathbf{o}\}$ being the respective gate. The symbol $\odot$ represents an element-wise product and $\sigma(\cdot)$ is the sigmoid function.

First, we note that the set of Equations (B.3)-(B.8) can be

expanded into the following two equations:

$$\begin{aligned}\hat{\mathbf{c}}_t =& \sigma\left(\hat{\mathbf{W}}_{\mathbf{xf}}\mathbf{x}_t + \hat{\mathbf{W}}_{\mathbf{hf}}\hat{\mathbf{h}}_{t-1} + \hat{\mathbf{b}}_{\mathbf{f}}\right) \odot \hat{\mathbf{c}}_{t-1} \\ &+ \sigma\left(\hat{\mathbf{W}}_{\mathbf{xe}}\mathbf{x}_t + \hat{\mathbf{W}}_{\mathbf{he}}\hat{\mathbf{h}}_{t-1} + \hat{\mathbf{b}}_{\mathbf{e}}\right) \\ &\odot \tanh\left(\hat{\mathbf{W}}_{\mathbf{xc}}\mathbf{x}_t + \hat{\mathbf{W}}_{\mathbf{hc}}\hat{\mathbf{h}}_{t-1} + \hat{\mathbf{b}}_{\mathbf{c}}\right), \qquad \text{(B.9)} \\ \hat{\mathbf{h}}_t =& \sigma\left(\hat{\mathbf{W}}_{\mathbf{xo}}\mathbf{x}_t + \hat{\mathbf{W}}_{\mathbf{ho}}\hat{\mathbf{h}}_{t-1} + \hat{\mathbf{b}}_{\mathbf{o}}\right) \odot \tanh\left(\hat{\mathbf{c}}_t\right). \qquad \text{(B.10)}\end{aligned}$$

Rewriting the LSTM Equations (B.3)-(B.8) in this form, allows to remain with the recurrent equations where both $\hat{\mathbf{h}}_t$ and $\hat{\mathbf{c}}_t$ depend on the previous hidden and cell states, $\hat{\mathbf{h}}_{t-1}$ and $\hat{\mathbf{c}}_{t-1}$ , and the current input $\mathbf{x}_t$.

Next, we describe the weight matrices for the single-layer d-LSTM that matches a stacked-LSTM with $k$ layers. The matrices and biases follow the exact same pattern as the RNN proof, being the same for all gates.

$$\hat{\mathbf{W}}_{\mathbf{ha}} = \begin{bmatrix} \mathbf{W}_{\mathbf{ha}}^{(1)} & \mathbf{0} & & \cdots & & \mathbf{0} \\ \mathbf{W}_{\mathbf{xa}}^{(2)} & \mathbf{W}_{\mathbf{ha}}^{(2)} & & & & \\ \mathbf{0} & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \mathbf{W}_{\mathbf{xa}}^{(i)} & \mathbf{W}_{\mathbf{ha}}^{(i)} & \ddots & \\ & & & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & & \mathbf{0} & \mathbf{W}_{\mathbf{xa}}^{(k)} & \mathbf{W}_{\mathbf{ha}}^{(k)} \end{bmatrix} \qquad \text{(B.11)}$$

$$\hat{\mathbf{b}}_{\mathbf{ha}} = \begin{bmatrix} \mathbf{b}_{\mathbf{ha}}^{(1)} \\ \vdots \\ \mathbf{b}_{\mathbf{ha}}^{(k)} \end{bmatrix}, \qquad \hat{\mathbf{W}}_{\mathbf{xa}} = \begin{bmatrix} \mathbf{W}_{\mathbf{xa}}^{(1)} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \qquad \text{(B.12)}$$

where $\hat{\mathbf{W}}_{\mathbf{xa}} \in \mathbb{R}^{kn \times q}$ are the input weights, $\hat{\mathbf{W}}_{\mathbf{ha}} \in \mathbb{R}^{kn \times kn}$ the recurrent weights, $\hat{\mathbf{b}}_{\mathbf{ha}} \in \mathbb{R}^{kn}$ the biases, for gate $\mathbf{a} \in \{\mathbf{e}, \mathbf{c}, \mathbf{o}, \mathbf{f}\}$. We follow the same notation for blocks and layers introduced with Theorem 1. We omit the equations for the output element $\hat{\mathbf{y}}_t$ as they are exactly the same as the RNN in Theorem 1, and thus require the same steps for proving that outputs are equal, i.e., $\hat{\mathbf{y}}_{T+k-1} = \mathbf{y}_T$. Therefore, for the LSTM theorem we will focus on the hidden and cell states.

**Theorem 2.** *Given an input sequence $\{\mathbf{x}_t\}_{t=1\ldots T}$ and a stacked LSTM with $k$ layers, and initial states $\{\mathbf{h}_0^{(i)}, \mathbf{c}_0^{(i)}\}_{i=1\ldots k}$, the d-LSTM with delay $d = k-1$, defined by Equations (B.11)-(B.12) and initialized with $\hat{\mathbf{h}}_0$ such that $\hat{\mathbf{h}}_{i-1}^{\{i\}} = \mathbf{h}_0^{(i)}$, $\forall i = 1\ldots k$ and $\hat{\mathbf{c}}_0$ such that $\hat{\mathbf{c}}_{i-1}^{\{i\}} = \mathbf{c}_0^{(i)}$, $\forall i = 1\ldots k$, produces the same output sequence but delayed by $k-1$ timesteps, i.e., $\hat{\mathbf{y}}_{t+k-1} = \mathbf{y}_t$ for all $t = 1\ldots T$. Further, the sequence of hidden and cell states at each layer $i$ are equivalent with delay $i-1$, i.e., $\hat{\mathbf{h}}_{t+i-1}^{\{i\}} = \mathbf{h}_t^{(i)}$ and $\hat{\mathbf{c}}_{t+i-1}^{\{i\}} = \mathbf{c}_t^{(i)}$ for all $1 \le i \le k$ and $t \ge 1$.*

*Proof.* We prove Theorem 2 by induction on the sequence length $t$. First, we show that for $t = 1$ the stacked LSTM and the d-LSTM with the constrained weights are equivalent. Namely, for $t = 1$ we show that the outputs, hidden states and cell states are the same, i.e. $\hat{\mathbf{y}}_k = \mathbf{y}_1$, $\hat{\mathbf{h}}_i^{\{i\}} = \mathbf{h}_1^{(i)}$, and $\hat{\mathbf{c}}_i^{\{i\}} = \mathbf{c}_1^{(i)}$, respectively. Without loss of generality, we have for any $j$ in $1\ldots k$ the following:

$$\begin{aligned}\hat{\mathbf{h}}_i^{\{i\}} &= \sigma\left(\hat{\mathbf{W}}_{\mathbf{xo}}^{\{i\}}\mathbf{x}_i + \hat{\mathbf{W}}_{\mathbf{ho}}^{\{i\}}\hat{\mathbf{h}}_{i-1}^{\{i\}} + \hat{\mathbf{b}}_{\mathbf{o}}^{\{i\}}\right) \odot \tanh\left(\hat{\mathbf{c}}_i^{\{i\}}\right) \\ &= \sigma\left(\mathbf{W}_{\mathbf{xo}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{ho}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i\}} + \mathbf{b}_{\mathbf{o}}^{(i)}\right) \\ &\quad \odot \tanh\left(\hat{\mathbf{c}}_i^{\{i\}}\right) \\ &= \sigma\left(\mathbf{W}_{\mathbf{xo}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{ho}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{o}}^{(i)}\right) \\ &\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{hf}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{f}}^{(i)}\right)\right. \\ &\quad \odot \hat{\mathbf{c}}_{i-1}^{\{i\}} \\ &\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{he}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\ &\quad \left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{hc}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{c}}^{(i)}\right)\right) \\ &= \ldots \\ &= \sigma\left(\mathbf{W}_{\mathbf{xo}}^{(i)}\ldots\left\{\sigma\left(\mathbf{W}_{\mathbf{xo}}^{(j)}[\ldots\right.\right.\right. \\ &\quad \sigma\left(\mathbf{W}_{\mathbf{xo}}^{(2)}\sigma\left(\mathbf{W}_{\mathbf{xo}}^{(1)}\mathbf{x}_1 + \mathbf{W}_{\mathbf{ho}}^{(1)}\mathbf{h}_0^{(1)} + \mathbf{b}_{\mathbf{o}}^{(1)}\right)\right. \\ &\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(1)}\mathbf{x}_1 + \mathbf{W}_{\mathbf{hf}}^{(1)}\mathbf{h}_0^{(1)} + \mathbf{b}_{\mathbf{f}}^{(1)}\right)\right. \\ &\quad \odot \mathbf{c}_0^{(1)} \\ &\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(1)}\mathbf{x}_1 + \mathbf{W}_{\mathbf{he}}^{(1)}\mathbf{h}_0^{(1)} + \mathbf{b}_{\mathbf{e}}^{(1)}\right) \\ &\quad \left.\left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(1)}\mathbf{x}_1 + \mathbf{W}_{\mathbf{hc}}^{(1)}\mathbf{h}_0^{(1)} + \mathbf{b}_{\mathbf{c}}^{(1)}\right)\right) \right. \\ &\quad + \mathbf{W}_{\mathbf{ho}}^{(2)}\mathbf{h}_0^{(2)} + \mathbf{b}_{\mathbf{o}}^{(2)}\Big) \\ &\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(2)}(\ldots) + \mathbf{W}_{\mathbf{hf}}^{(2)}\mathbf{h}_0^{(2)} + \mathbf{b}_{\mathbf{f}}^{(2)}\right)\right. \\ &\quad \odot \hat{\mathbf{c}}_1^{\{2\}} \\ &\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(2)}(\ldots) + \mathbf{W}_{\mathbf{he}}^{(2)}\mathbf{h}_0^{(2)} + \mathbf{b}_{\mathbf{e}}^{(2)}\right) \\ &\quad \left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(2)}(\ldots) + \mathbf{W}_{\mathbf{hc}}^{(2)}\mathbf{h}_0^{(2)} + \mathbf{b}_{\mathbf{c}}^{(2)}\right)\right) \\ &\quad \ldots] + \mathbf{W}_{\mathbf{ho}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{o}}^{(j)}\Big) \\ &\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(j)}[\ldots] + \mathbf{W}_{\mathbf{hf}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{f}}^{(j)}\right)\right. \\ &\quad \odot \hat{\mathbf{c}}_{j-1}^{\{j\}} \\ &\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(j)}[\ldots] + \mathbf{W}_{\mathbf{he}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{e}}^{(j)}\right)\end{aligned}$$

$$
\begin{aligned}
&\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(j)}[\dots] + \mathbf{W}_{\mathbf{hc}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{c}}^{(j)}\right)\Big)\Big\} \\
&\cdots + \mathbf{W}_{\mathbf{ho}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{o}}^{(i)} \\
&\odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{hf}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{f}}^{(i)}\right)\right. \\
&\odot \hat{\mathbf{c}}_{i-1}^{\{i\}} \\
&+ \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{he}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{hc}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{c}}^{(i)}\right)\right) \\
=\ &\sigma\left(\mathbf{W}_{\mathbf{xo}}^{(i)}\dots\Big\{\sigma\left(\mathbf{W}_{\mathbf{xo}}^{(j)}[\dots\right. \\
&\quad \sigma\left(\mathbf{W}_{\mathbf{xo}}^{(2)}\mathbf{h}_1^{(1)} + \mathbf{W}_{\mathbf{ho}}^{(2)}\mathbf{h}_0^{(2)} + \mathbf{b}_{\mathbf{o}}^{(2)}\right) \\
&\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(2)}\mathbf{h}_1^{(1)} + \mathbf{W}_{\mathbf{hf}}^{(2)}\mathbf{h}_0^{(2)} + \mathbf{b}_{\mathbf{f}}^{(2)}\right) \odot \mathbf{c}_0^{(2)}\right. \\
&\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(2)}\mathbf{h}_1^{(1)} + \mathbf{W}_{\mathbf{he}}^{(2)}\mathbf{h}_0^{(2)} + \mathbf{b}_{\mathbf{e}}^{(2)}\right) \\
&\quad \left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(2)}\mathbf{h}_1^{(1)} + \mathbf{W}_{\mathbf{hc}}^{(2)}\mathbf{h}_0^{(2)} + \mathbf{b}_{\mathbf{c}}^{(2)}\right)\right) \\
&\quad \left.\dots] + \mathbf{W}_{\mathbf{ho}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{o}}^{(j)}\right) \\
&\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(j)}[\dots] + \mathbf{W}_{\mathbf{hf}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{f}}^{(j)}\right) \odot \mathbf{c}_0^{(j)}\right. \\
&\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(j)}[\dots] + \mathbf{W}_{\mathbf{he}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{e}}^{(j)}\right) \\
&\quad \left.\left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(j)}[\dots] + \mathbf{W}_{\mathbf{hc}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{c}}^{(j)}\right)\right)\Big\}\right. \\
&\cdots + \mathbf{W}_{\mathbf{ho}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{o}}^{(i)} \\
&\odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{hf}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{f}}^{(i)}\right) \odot \mathbf{c}_0^{(i)}\right. \\
&+ \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{he}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{hc}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{c}}^{(i)}\right)\right) \\
=\ &\dots \\
=\ &\sigma\left(\mathbf{W}_{\mathbf{xo}}^{(i)}\dots\Big\{\sigma\left(\mathbf{W}_{\mathbf{xo}}^{(j)}\mathbf{h}_1^{(j-1)} + \mathbf{W}_{\mathbf{ho}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{o}}^{(j)}\right)\right. \\
&\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(j)}\mathbf{h}_1^{(j-1)} + \mathbf{W}_{\mathbf{hf}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{f}}^{(j)}\right) \odot \mathbf{c}_0^{(j)}\right. \\
&\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(j)}\mathbf{h}_1^{(j-1)} + \mathbf{W}_{\mathbf{he}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{e}}^{(j)}\right) \\
&\quad \left.\left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(j)}\mathbf{h}_1^{(j-1)} + \mathbf{W}_{\mathbf{hc}}^{(j)}\mathbf{h}_0^{(j)} + \mathbf{b}_{\mathbf{c}}^{(j)}\right)\right)\Big\}\right. \\
&\cdots + \mathbf{W}_{\mathbf{ho}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{o}}^{(i)} \\
&\odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{hf}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{f}}^{(i)}\right) \odot \mathbf{c}_0^{(i)}\right. \\
&+ \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{he}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{hc}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{c}}^{(i)}\right)\right) \\
=\ &\dots
\end{aligned}
$$

$$
\begin{aligned}
&= \sigma\left(\mathbf{W}_{\mathbf{xo}}^{(i)}\mathbf{h}_1^{(i-1)} + \mathbf{W}_{\mathbf{ho}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{o}}^{(i)}\right) \\
&\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}\mathbf{h}_1^{(i-1)} + \mathbf{W}_{\mathbf{hf}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{f}}^{(i)}\right) \odot \mathbf{c}_0^{(i)}\right. \\
&\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}\mathbf{h}_1^{(i-1)} + \mathbf{W}_{\mathbf{he}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\quad \left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}\mathbf{h}_1^{(i-1)} + \mathbf{W}_{\mathbf{hc}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{c}}^{(i)}\right)\right) \\
&= \mathbf{h}_1^{(i)},
\end{aligned}
$$

where we used the initialization assumptions $\hat{\mathbf{h}}_{i-1}^{\{i\}} = \mathbf{h}_0^{(i)}$ and $\hat{\mathbf{c}}_{i-1}^{\{i\}} = \mathbf{c}_0^{(i)}$ for all $i = 1 \dots k$, and the definition of the hidden and cell state in Equations (B.9) and (B.10) for $j-1$ blocks, in the previous steps. In particular, we have for layer $k$ that $\hat{\mathbf{h}}_i^{\{k\}} = \mathbf{h}_1^{(k)}$, and using the same transformations as in (A.1) with RNNs, we obtain $\hat{\mathbf{y}}_k = \mathbf{y}_1$. Furthermore, we obtained that:

$$
\begin{aligned}
\hat{\mathbf{c}}_i^{\{i\}} &= \sigma\left(\hat{\mathbf{W}}_{\mathbf{xf}}^{\{i\}}\mathbf{x}_i + \hat{\mathbf{W}}_{\mathbf{hf}}^{\{i\}}\hat{\mathbf{h}}_{i-1} + \hat{\mathbf{b}}_{\mathbf{f}}^{\{i\}}\right) \odot \hat{\mathbf{c}}_{i-1}^{\{i\}} \\
&\quad + \sigma\left(\hat{\mathbf{W}}_{\mathbf{xe}}^{\{i\}}\mathbf{x}_i + \hat{\mathbf{W}}_{\mathbf{he}}^{\{i\}}\hat{\mathbf{h}}_{i-1} + \hat{\mathbf{b}}_{\mathbf{e}}^{\{i\}}\right) \\
&\quad \odot \tanh\left(\hat{\mathbf{W}}_{\mathbf{xc}}^{\{i\}}\mathbf{x}_i + \hat{\mathbf{W}}_{\mathbf{hc}}^{\{i\}}\hat{\mathbf{h}}_{i-1} + \hat{\mathbf{b}}_{\mathbf{c}}^{\{i\}}\right) \\
&= \sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{hf}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i\}} + \mathbf{b}_{\mathbf{f}}^{(i)}\right) \odot \mathbf{c}_0^{(i)} \\
&\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{he}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i\}} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\quad \odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i-1\}} + \mathbf{W}_{\mathbf{hc}}^{(i)}\hat{\mathbf{h}}_{i-1}^{\{i\}} + \mathbf{b}_{\mathbf{c}}^{(i)}\right) \\
&= \sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}\mathbf{h}_1^{(i-1)} + \mathbf{W}_{\mathbf{hf}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{f}}^{(i)}\right) \odot \mathbf{c}_0^{(i)} \\
&\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}\mathbf{h}_1^{(i-1)} + \mathbf{W}_{\mathbf{he}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\quad \odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}\mathbf{h}_1^{(i-1)} + \mathbf{W}_{\mathbf{hc}}^{(i)}\mathbf{h}_0^{(i)} + \mathbf{b}_{\mathbf{c}}^{(i)}\right) \\
&= \mathbf{c}_1^{(i)}
\end{aligned}
$$

Which concludes the basis of the induction.

Next, we assume that $\hat{\mathbf{h}}_{t+i-1}^{\{i\}} = \mathbf{h}_t^{(i)}$ and $\hat{\mathbf{c}}_{t+i-1}^{\{i\}} = \mathbf{c}_t^{(i)}$ for all $1 \leq i \leq k$ and $t \leq T-1$, and prove that it holds for the hidden and cell states for all layers when $t = T$: $\hat{\mathbf{h}}_{T+i-1}^{\{i\}} = \mathbf{h}_T^{(i)}$, $\forall 1 \leq i \leq k$. Without loss of generality, we have for the hidden state $\hat{\mathbf{h}}_{T+i-1}^{\{i\}}$ in constrained weights

single-layer d-LSTM that,

$$\begin{aligned}
\hat{\mathbf{h}}^{\{i\}}_{T+i-1} &= \sigma\left(\hat{\mathbf{W}}^{\{i\}}_{\mathbf{xo}}\mathbf{x}_{T+i-1} + \hat{\mathbf{W}}^{\{i\}}_{\mathbf{ho}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \hat{\mathbf{b}}^{\{i\}}_{\mathbf{o}}\right) \\
&\quad \odot \tanh\left(\hat{\mathbf{c}}^{\{i\}}_{T+i-1}\right) \\
&= \sigma\left(\mathbf{W}^{(i)}_{\mathbf{xo}}\hat{\mathbf{h}}^{\{i-1\}}_{T+i-2} + \mathbf{W}^{(i)}_{\mathbf{ho}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{o}}\right) \\
&\quad \odot \tanh\left(\hat{\mathbf{c}}^{\{i\}}_{T+i-1}\right) \\
&= \sigma\left(\mathbf{W}^{(i)}_{\mathbf{xo}}\hat{\mathbf{h}}^{\{i-1\}}_{T+i-2} + \mathbf{W}^{(i)}_{\mathbf{ho}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{o}}\right) \\
&\quad \odot \tanh\left(\sigma\left(\mathbf{W}^{(i)}_{\mathbf{xf}}\hat{\mathbf{h}}^{\{i-1\}}_{T+i-2} + \mathbf{W}^{(i)}_{\mathbf{hf}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{f}}\right)\right. \\
&\quad \odot \hat{\mathbf{c}}^{\{i\}}_{T+i-2} \\
&\quad + \sigma\left(\mathbf{W}^{(i)}_{\mathbf{xe}}\hat{\mathbf{h}}^{\{i-1\}}_{T+i-2} + \mathbf{W}^{(i)}_{\mathbf{he}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{e}}\right) \\
&\quad \left.\odot \tanh\left(\mathbf{W}^{(i)}_{\mathbf{xc}}\hat{\mathbf{h}}^{\{i-1\}}_{T+i-2} + \mathbf{W}^{(i)}_{\mathbf{hc}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{c}}\right)\right) \\
&= \dots
\end{aligned}$$

$$\begin{aligned}
&= \sigma\Big(\mathbf{W}^{(i)}_{\mathbf{xo}} \dots \Big\{\sigma\Big(\mathbf{W}^{(j)}_{\mathbf{xo}}[\dots \\
&\quad \sigma\Big(\mathbf{W}^{(2)}_{\mathbf{xo}}\sigma\Big(\mathbf{W}^{(1)}_{\mathbf{xo}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{ho}}\hat{\mathbf{h}}^{\{1\}}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{o}}\Big) \\
&\quad \odot \tanh\Big(\sigma\Big(\mathbf{W}^{(1)}_{\mathbf{xf}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{hf}}\hat{\mathbf{h}}^{\{1\}}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{f}}\Big) \\
&\quad \odot \hat{\mathbf{c}}^{\{1\}}_{T-1} + \sigma\Big(\mathbf{W}^{(1)}_{\mathbf{xe}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{he}}\hat{\mathbf{h}}^{\{1\}}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{e}}\Big) \\
&\quad \odot \tanh\Big(\mathbf{W}^{(1)}_{\mathbf{xc}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{hc}}\hat{\mathbf{h}}^{\{1\}}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{c}}\Big)\Big) \\
&\quad + \mathbf{W}^{(2)}_{\mathbf{ho}}\hat{\mathbf{h}}^{\{2\}}_{T} + \mathbf{b}^{(2)}_{\mathbf{o}}\Big) \\
&\quad \odot \tanh\Big(\sigma\Big(\mathbf{W}^{(2)}_{\mathbf{xf}}(\dots) + \mathbf{W}^{(2)}_{\mathbf{hf}}\hat{\mathbf{h}}^{\{2\}}_{T} + \mathbf{b}^{(2)}_{\mathbf{f}}\Big) \\
&\quad \odot \hat{\mathbf{c}}^{\{2\}}_{T} + \sigma\Big(\mathbf{W}^{(2)}_{\mathbf{xe}}(\dots) + \mathbf{W}^{(2)}_{\mathbf{he}}\hat{\mathbf{h}}^{\{2\}}_{T} + \mathbf{b}^{(2)}_{\mathbf{e}}\Big) \\
&\quad \odot \tanh\Big(\mathbf{W}^{(2)}_{\mathbf{xc}}(\dots) + \mathbf{W}^{(2)}_{\mathbf{hc}}\hat{\mathbf{h}}^{\{2\}}_{T} + \mathbf{b}^{(2)}_{\mathbf{c}}\Big)\Big) \\
&\quad \dots] + \mathbf{W}^{(j)}_{\mathbf{ho}}\hat{\mathbf{h}}^{\{j\}}_{T+j-2} + \mathbf{b}^{(j)}_{\mathbf{o}}\Big) \\
&\quad \odot \tanh\Big(\sigma\Big(\mathbf{W}^{(j)}_{\mathbf{xf}}[\dots] + \mathbf{W}^{(j)}_{\mathbf{hf}}\hat{\mathbf{h}}^{\{j\}}_{T+j-2} + \mathbf{b}^{(j)}_{\mathbf{f}}\Big) \\
&\quad \odot \hat{\mathbf{c}}^{\{j\}}_{T+j-2} + \sigma\Big(\mathbf{W}^{(j)}_{\mathbf{xe}}[\dots] + \mathbf{W}^{(j)}_{\mathbf{he}}\hat{\mathbf{h}}^{\{j\}}_{T+j-2} + \mathbf{b}^{(j)}_{\mathbf{e}}\Big) \\
&\quad \odot \tanh\Big(\mathbf{W}^{(j)}_{\mathbf{xc}}[\dots] + \mathbf{W}^{(j)}_{\mathbf{hc}}\hat{\mathbf{h}}^{\{j\}}_{T+j-2} + \mathbf{b}^{(j)}_{\mathbf{c}}\Big)\Big)\Big\} \\
&\quad \dots + \mathbf{W}^{(i)}_{\mathbf{ho}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{o}} \\
&\quad \odot \tanh\Big(\sigma\Big(\mathbf{W}^{(i)}_{\mathbf{xf}}(\dots) + \mathbf{W}^{(i)}_{\mathbf{hf}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{f}}\Big) \\
&\quad \odot \hat{\mathbf{c}}^{\{i\}}_{T+i-2} + \sigma\Big(\mathbf{W}^{(i)}_{\mathbf{xe}}(\dots) + \mathbf{W}^{(i)}_{\mathbf{he}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{e}}\Big) \\
&\quad \odot \tanh\Big(\mathbf{W}^{(i)}_{\mathbf{xc}}(\dots) + \mathbf{W}^{(i)}_{\mathbf{hc}}\hat{\mathbf{h}}^{\{i\}}_{T+i-2} + \mathbf{b}^{(i)}_{\mathbf{c}}\Big)\Big)
\end{aligned}$$

From the inductive assumption we have that $\hat{\mathbf{h}}^{\{j\}}_{T+j-2} = \mathbf{h}^{(j)}_{T-1}$ and $\hat{\mathbf{c}}^{\{j\}}_{T+j-2} = \mathbf{c}^{(j)}_{T-1}$ for all $1 \le j \le k$, then it follows that

$$\begin{aligned}
&= \sigma\Big(\mathbf{W}^{(i)}_{\mathbf{xo}} \dots \Big\{\sigma\Big(\mathbf{W}^{(j)}_{\mathbf{xo}}[\dots \\
&\quad \sigma\Big(\mathbf{W}^{(2)}_{\mathbf{xo}}\sigma\Big(\mathbf{W}^{(1)}_{\mathbf{xo}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{ho}}\mathbf{h}^{(1)}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{o}}\Big) \\
&\quad \odot \tanh\Big(\sigma\Big(\mathbf{W}^{(1)}_{\mathbf{xf}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{hf}}\mathbf{h}^{(1)}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{f}}\Big) \odot \mathbf{c}^{(1)}_{T-1} \\
&\quad + \sigma\Big(\mathbf{W}^{(1)}_{\mathbf{xe}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{he}}\mathbf{h}^{(1)}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{e}}\Big) \\
&\quad \odot \tanh\Big(\mathbf{W}^{(1)}_{\mathbf{xc}}\mathbf{x}_T + \mathbf{W}^{(1)}_{\mathbf{hc}}\mathbf{h}^{(1)}_{T-1} + \mathbf{b}^{(1)}_{\mathbf{c}}\Big)\Big) \\
&\quad + \mathbf{W}^{(2)}_{\mathbf{ho}}\mathbf{h}^{(2)}_{T-1} + \mathbf{b}^{(2)}_{\mathbf{o}}\Big) \\
&\quad \odot \tanh\Big(\sigma\Big(\mathbf{W}^{(2)}_{\mathbf{xf}}(\dots) + \mathbf{W}^{(2)}_{\mathbf{hf}}\mathbf{h}^{(2)}_{T-1} + \mathbf{b}^{(2)}_{\mathbf{f}}\Big) \odot \mathbf{c}^{(2)}_{T-1} \\
&\quad + \sigma\Big(\mathbf{W}^{(2)}_{\mathbf{xe}}(\dots) + \mathbf{W}^{(2)}_{\mathbf{he}}\mathbf{h}^{(2)}_{T-1} + \mathbf{b}^{(2)}_{\mathbf{e}}\Big) \\
&\quad \odot \tanh\Big(\mathbf{W}^{(2)}_{\mathbf{xc}}(\dots) + \mathbf{W}^{(2)}_{\mathbf{hc}}\mathbf{h}^{(2)}_{T-1} + \mathbf{b}^{(2)}_{\mathbf{c}}\Big)\Big) \\
&\quad \dots] + \mathbf{W}^{(j)}_{\mathbf{ho}}\mathbf{h}^{(j)}_{T-1} + \mathbf{b}^{(j)}_{\mathbf{o}}\Big) \\
&\quad \odot \tanh\Big(\sigma\Big(\mathbf{W}^{(j)}_{\mathbf{xf}}[\dots] + \mathbf{W}^{(j)}_{\mathbf{hf}}\mathbf{h}^{(j)}_{T-1} + \mathbf{b}^{(j)}_{\mathbf{f}}\Big) \odot \mathbf{c}^{(j)}_{T-1} \\
&\quad + \sigma\Big(\mathbf{W}^{(j)}_{\mathbf{xe}}[\dots] + \mathbf{W}^{(j)}_{\mathbf{he}}\mathbf{h}^{(j)}_{T-1} + \mathbf{b}^{(j)}_{\mathbf{e}}\Big) \\
&\quad \odot \tanh\Big(\mathbf{W}^{(j)}_{\mathbf{xc}}[\dots] + \mathbf{W}^{(j)}_{\mathbf{hc}}\mathbf{h}^{(j)}_{T-1} + \mathbf{b}^{(j)}_{\mathbf{c}}\Big)\Big)\Big\} \\
&\quad \dots + \mathbf{W}^{(i)}_{\mathbf{ho}}\mathbf{h}^{(i)}_{T-1} + \mathbf{b}^{(i)}_{\mathbf{o}} \\
&\quad \odot \tanh\Big(\sigma\Big(\mathbf{W}^{(i)}_{\mathbf{xf}}(\dots) + \mathbf{W}^{(i)}_{\mathbf{hf}}\mathbf{h}^{(i)}_{T-1} + \mathbf{b}^{(i)}_{\mathbf{f}}\Big) \odot \mathbf{c}^{(i)}_{T-1} \\
&\quad + \sigma\Big(\mathbf{W}^{(i)}_{\mathbf{xe}}(\dots) + \mathbf{W}^{(i)}_{\mathbf{he}}\mathbf{h}^{(i)}_{T-1} + \mathbf{b}^{(i)}_{\mathbf{e}}\Big) \\
&\quad \odot \tanh\Big(\mathbf{W}^{(i)}_{\mathbf{xc}}(\dots) + \mathbf{W}^{(i)}_{\mathbf{hc}}\mathbf{h}^{(i)}_{T-1} + \mathbf{b}^{(i)}_{\mathbf{c}}\Big)\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \dots \\
&= \sigma\left(\mathbf{W}_{\mathbf{xo}}^{(i)} \dots \left\{\sigma\left(\mathbf{W}_{\mathbf{xo}}^{(j)}\mathbf{h}_T^{(j-1)} + \mathbf{W}_{\mathbf{ho}}^{(j)}\mathbf{h}_{T-1}^{(j)} + \mathbf{b}_{\mathbf{o}}^{(j)}\right)\right.\right. \\
&\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(j)}\mathbf{h}_T^{(j-1)} + \mathbf{W}_{\mathbf{hf}}^{(j)}\mathbf{h}_{T-1}^{(j)} + \mathbf{b}_{\mathbf{f}}^{(j)}\right)\right. \\
&\quad \odot \mathbf{c}_{T-1}^{(j)} + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(j)}\mathbf{h}_T^{(j-1)} + \mathbf{W}_{\mathbf{he}}^{(j)}\mathbf{h}_{T-1}^{(j)} + \mathbf{b}_{\mathbf{e}}^{(j)}\right) \\
&\quad \left.\left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(j)}\mathbf{h}_T^{(j-1)} + \mathbf{W}_{\mathbf{hc}}^{(j)}\mathbf{h}_{T-1}^{(j)} + \mathbf{b}_{\mathbf{c}}^{(j)}\right)\right)\right\} \\
&\quad \dots + \mathbf{W}_{\mathbf{ho}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{o}}^{(i)} \\
&\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{hf}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{f}}^{(i)}\right)\right. \\
&\quad \odot \mathbf{c}_{T-1}^{(i)} + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{he}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\quad \left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}(\dots) + \mathbf{W}_{\mathbf{hc}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{c}}^{(i)}\right)\right) \\
&= \dots \\
&= \sigma\left(\mathbf{W}_{\mathbf{xo}}^{(i)}\mathbf{h}_T^{(i-1)} + \mathbf{W}_{\mathbf{ho}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{o}}^{(i)}\right) \\
&\quad \odot \tanh\left(\sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}\mathbf{h}_T^{(i-1)} + \mathbf{W}_{\mathbf{hf}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{f}}^{(i)}\right)\right. \\
&\quad \odot \mathbf{c}_{T-1}^{(i)} + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}\mathbf{h}_T^{(i-1)} + \mathbf{W}_{\mathbf{he}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\quad \left.\odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}\mathbf{h}_T^{(i-1)} + \mathbf{W}_{\mathbf{hc}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{c}}^{(i)}\right)\right) \\
&= \mathbf{h}_T^{(i)},
\end{aligned}
$$

where we use the recurrent definition of the hidden and cell states in Equations (B.9) and (B.10). In particular, we obtained for $i = k$ that $\hat{\mathbf{h}}_{T+k-1}^{\{k\}} = \mathbf{h}_T^{(k)}$. Applying the same steps as in the d-RNN proof in Eq. (A.2), we obtain $\hat{\mathbf{y}}_{T+k-1} = \mathbf{y}_T$. Last, we obtain for the cell state that

$$
\begin{aligned}
\hat{\mathbf{c}}_{T+i-1}^{\{i\}} &= \sigma\left(\hat{\mathbf{W}}_{\mathbf{xf}}^{\{i\}}\mathbf{x}_{T+i-1} + \hat{\mathbf{W}}_{\mathbf{hf}}^{\{i\}}\hat{\mathbf{h}}_{T+i-2} + \hat{\mathbf{b}}_{\mathbf{f}}^{\{i\}}\right) \\
&\quad \odot \hat{\mathbf{c}}_{T+i-2}^{\{i\}} \\
&\quad + \sigma\left(\hat{\mathbf{W}}_{\mathbf{xe}}^{\{i\}}\mathbf{x}_{T+i-1} + \hat{\mathbf{W}}_{\mathbf{he}}^{\{i\}}\hat{\mathbf{h}}_{T+i-2} + \hat{\mathbf{b}}_{\mathbf{e}}^{\{i\}}\right) \\
&\quad \odot \tanh\left(\hat{\mathbf{W}}_{\mathbf{xc}}^{\{i\}}\mathbf{x}_{T+i-1} + \hat{\mathbf{W}}_{\mathbf{hc}}^{\{i\}}\hat{\mathbf{h}}_{T+i-2} + \hat{\mathbf{b}}_{\mathbf{c}}^{\{i\}}\right) \\
&= \sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}\hat{\mathbf{h}}_{T+i-2}^{\{i-1\}} + \mathbf{W}_{\mathbf{hf}}^{(i)}\hat{\mathbf{h}}_{T+i-2}^{\{i\}} + \mathbf{b}_{\mathbf{f}}^{(i)}\right) \odot \mathbf{c}_{T-1}^{(i)} \\
&\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}\hat{\mathbf{h}}_{T+i-2}^{\{i-1\}} + \mathbf{W}_{\mathbf{he}}^{(i)}\hat{\mathbf{h}}_{T+i-2}^{\{i\}} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\quad \odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}\hat{\mathbf{h}}_{T+i-2}^{\{i-1\}} + \mathbf{W}_{\mathbf{hc}}^{(i)}\hat{\mathbf{h}}_{T+i-2}^{\{i\}} + \mathbf{b}_{\mathbf{c}}^{(i)}\right) \\
&= \sigma\left(\mathbf{W}_{\mathbf{xf}}^{(i)}\mathbf{h}_T^{(i-1)} + \mathbf{W}_{\mathbf{hf}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{f}}^{(i)}\right) \odot \mathbf{c}_{T-1}^{(i)} \\
&\quad + \sigma\left(\mathbf{W}_{\mathbf{xe}}^{(i)}\mathbf{h}_T^{(i-1)} + \mathbf{W}_{\mathbf{he}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{e}}^{(i)}\right) \\
&\quad \odot \tanh\left(\mathbf{W}_{\mathbf{xc}}^{(i)}\mathbf{h}_T^{(i-1)} + \mathbf{W}_{\mathbf{hc}}^{(i)}\mathbf{h}_{T-1}^{(i)} + \mathbf{b}_{\mathbf{c}}^{(i)}\right) \\
&= \mathbf{c}_T^{(i)}
\end{aligned}
$$

Which completes the proof. ■

## C. Weight Constraints and Connections in d-RNN

Figure 6 shows the weight constraints imposed to achieve equivalence between the stacked RNN and single-layer d-RNN, and a visualization of the d-RNN as connections in the stacked RNN. Figure 6(b) depicts the delay (or "shift") of all the hidden states as they would be computed in the stacked RNN. Each layer is equivalent to a shift by one timestep.

## D. Additional Plots for Error Maps

Figure 7 present the standard deviation diagrams for the error maps in Figure 5.

## E. Masked Character-Level Language Modeling: Additional Results

In Table 3, we include additional results for smaller networks of the masked language model task. We sampled more delay values for d-LSTMs, but the general conclusions remain the same: intermediate values of delay achieve the lowest BPC. Forward-pass runtimes across delay values show a small increase with larger delays, but the increment is relatively flat compared to stacked LSTMs or (stacked) Bi-LSTMs as they increase in depth. For these experiments, we also used a batch of 128 sequences, and an embedding of dimension 10.

## F. Part-of-Speech Tagging: Additional Details and Results

In this section, we include more details about the dataset and the results of all the combinations for the Parts-Of-Speech experiment. We used treebanks from Universal Dependencies (UD) (Nivre et al., 2016) version 2.3. We selected the English EWT treebank[1] (Silveira et al., 2014) (254,854 words), French GSD treebank[2] (411,465 words), and German GSD treebank[3] (297,836 words) based on the quality assigned by the UD authors. We follow the partitioning onto training, validation and test datasets as pre-defined in UD. All treebanks use the same POS tag set containing 17 tags. We use the Polyglot project (Al-Rfou' et al., 2013) word embeddings (64 dimensions). We build our own alphabets

[1] https://github.com/UniversalDependencies/UD_English-EWT/tree/r2.3

[2] https://github.com/UniversalDependencies/UD_French-GSD/tree/r2.3

[3] https://github.com/UniversalDependencies/UD_German-GSD/tree/r2.3

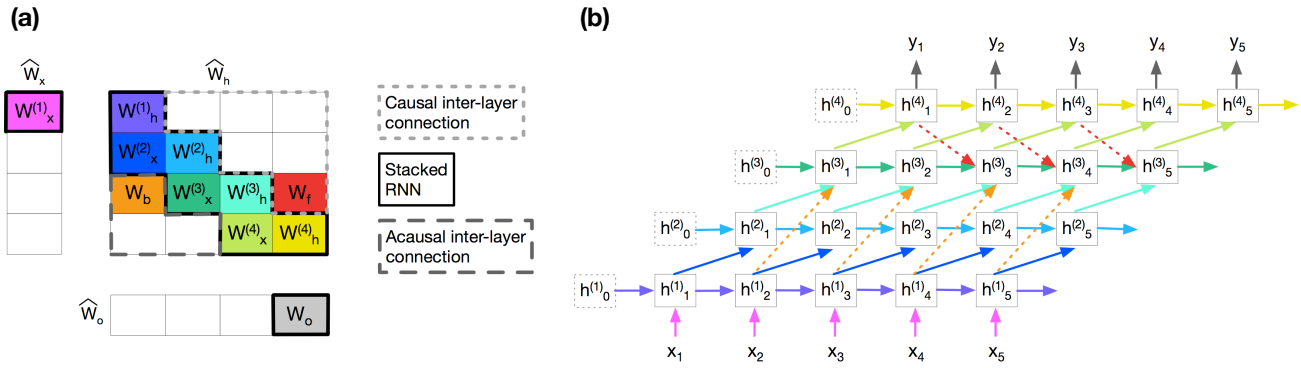



*Figure 6.* (a) Weights of the single-layer and weight constrained d-RNN that are equivalent to connections in the stacked RNN from Figure 2. (b) Connections in the d-RNN based on the weight matrix in (a). The d-RNN is depicted as it would be the stacked RNN. The hidden states are delayed in time with respect to the stacked network.

*Table 3.* Performance for smaller networks on the masked character-level language modeling task. Mean and standard deviation values are computed over 5 repetitions of training and inference runtime on the test set.

| MODEL | LAYERS | DELAY | UNITS | PARAMS. | VAL. BPC | TEST BPC | RUNTIME |
|---|---|---|---|---|---|---|---|
| LSTM | 1 | - | 512 | 1087283 | $2.139 \pm 0.005$ | $\mathbf{2.195 \pm 0.002}$ | $2.85ms \pm 0.14$ |
| LSTM | 2 | - | 298 | 1090689 | $2.156 \pm 0.003$ | $2.215 \pm 0.002$ | $6.69ms \pm 0.27$ |
| LSTM | 5 | - | 172 | 1083735 | $2.199 \pm 0.016$ | $2.255 \pm 0.015$ | $11.32ms \pm 0.05$ |
| BI-LSTM | 1 | - | 360 | 1091107 | $1.130 \pm 0.003$ | $1.187 \pm 0.004$ | $5.82ms \pm 0.18$ |
| BI-LSTM | 2 | - | 182 | 1090487 | $0.800 \pm 0.004$ | $0.846 \pm 0.005$ | $11.08ms \pm 0.59$ |
| BI-LSTM | 5 | - | 102 | 1104151 | $0.796 \pm 0.007$ | $\mathbf{0.841 \pm 0.006}$ | $23.94ms \pm 0.17$ |
| D-LSTM | 1 | 1 | 512 | 1087283 | $1.470 \pm 0.002$ | $1.518 \pm 0.003$ | $2.80ms \pm 0.02$ |
| D-LSTM | 1 | 2 | 512 | 1087283 | $1.162 \pm 0.004$ | $1.208 \pm 0.003$ | $2.81ms \pm 0.01$ |
| D-LSTM | 1 | 3 | 512 | 1087283 | $0.995 \pm 0.002$ | $1.039 \pm 0.002$ | $3.02ms \pm 0.23$ |
| D-LSTM | 1 | 5 | 512 | 1087283 | $0.877 \pm 0.001$ | $0.920 \pm 0.003$ | $3.01ms \pm 0.22$ |
| D-LSTM | 1 | 8 | 512 | 1087283 | $0.859 \pm 0.002$ | $\mathbf{0.905 \pm 0.003}$ | $3.04ms \pm 0.19$ |
| D-LSTM | 1 | 10 | 512 | 1087283 | $0.889 \pm 0.004$ | $0.935 \pm 0.005$ | $3.22ms \pm 0.18$ |
| D-LSTM | 1 | 15 | 512 | 1087283 | $0.971 \pm 0.004$ | $1.014 \pm 0.002$ | $3.17ms \pm 0.05$ |

based on the most frequent 100 characters in the vocabularies. All the networks have a 100-dimensional character-level embedding, which is trained with the network. We use a batch size of 32 sentences.

Results for German, English, and French can be found in Tables 4, 5, and 6, respectively. The best result that does not use a bidirectional network is marked in bold for each language.

## References

Al-Rfou', R., Perozzi, B., and Skiena, S. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-3520.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1262.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on*

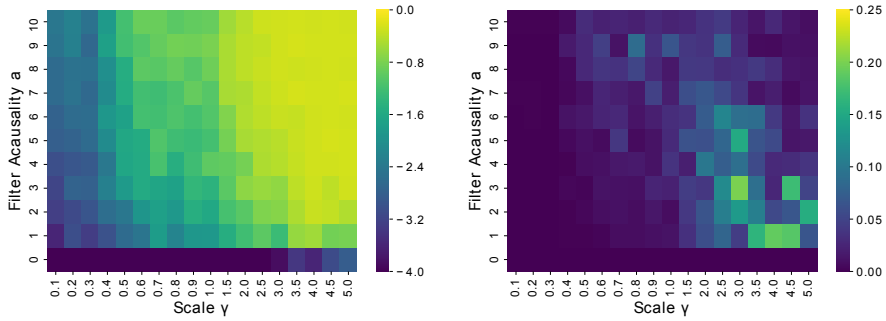
*Table 4.* Parts-of-Speech results for German. The table shows all possible combinations of delays or bidirectional LSTM networks. The best forward-only network is marked in bold.

| CHARACTER-LEVEL NETWORK | WORD-LEVEL NETWORK | VALIDATION ACCURACY | TEST ACCURACY |
|---|---|---|---|
| BI-LSTM | BI-LSTM | $93.88 \pm 0.13$ | $93.15 \pm 0.08$ |
| BI-LSTM | LSTM | $92.00 \pm 0.16$ | $91.50 \pm 0.05$ |
| BI-LSTM | D-LSTM WITH DELAY=1 | $93.32 \pm 0.23$ | $92.81 \pm 0.14$ |
| BI-LSTM | D-LSTM WITH DELAY=2 | $93.15 \pm 0.06$ | $92.67 \pm 0.08$ |
| BI-LSTM | D-LSTM WITH DELAY=3 | $92.82 \pm 0.14$ | $92.25 \pm 0.16$ |
| BI-LSTM | D-LSTM WITH DELAY=4 | $92.41 \pm 0.12$ | $91.95 \pm 0.17$ |
| BI-LSTM | D-LSTM WITH DELAY=5 | $91.86 \pm 0.11$ | $91.57 \pm 0.20$ |
| LSTM | BI-LSTM | $93.96 \pm 0.12$ | $93.43 \pm 0.07$ |
| LSTM | LSTM | $92.05 \pm 0.16$ | $91.58 \pm 0.11$ |
| LSTM | D-LSTM WITH DELAY=1 | $93.46 \pm 0.16$ | $92.71 \pm 0.11$ |
| LSTM | D-LSTM WITH DELAY=2 | $93.13 \pm 0.10$ | $92.61 \pm 0.26$ |
| LSTM | D-LSTM WITH DELAY=3 | $92.91 \pm 0.13$ | $92.38 \pm 0.15$ |
| LSTM | D-LSTM WITH DELAY=4 | $92.56 \pm 0.17$ | $92.06 \pm 0.19$ |
| D-LSTM WITH DELAY=1 | BI-LSTM | $93.93 \pm 0.06$ | $93.39 \pm 0.18$ |
| D-LSTM WITH DELAY=1 | LSTM | $92.04 \pm 0.11$ | $91.58 \pm 0.14$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=1 | $93.48 \pm 0.31$ | $\mathbf{92.87 \pm 0.24}$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=2 | $93.11 \pm 0.18$ | $92.54 \pm 0.08$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=3 | $92.85 \pm 0.14$ | $92.28 \pm 0.19$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=4 | $92.50 \pm 0.12$ | $92.11 \pm 0.19$ |
| D-LSTM WITH DELAY=3 | BI-LSTM | $94.00 \pm 0.17$ | $93.32 \pm 0.18$ |
| D-LSTM WITH DELAY=3 | LSTM | $92.10 \pm 0.24$ | $91.61 \pm 0.18$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=1 | $93.29 \pm 0.09$ | $92.68 \pm 0.09$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=2 | $93.09 \pm 0.21$ | $92.59 \pm 0.16$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=3 | $92.86 \pm 0.24$ | $92.42 \pm 0.16$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=4 | $92.53 \pm 0.17$ | $92.08 \pm 0.18$ |
| D-LSTM WITH DELAY=5 | BI-LSTM | $93.88 \pm 0.17$ | $93.27 \pm 0.06$ |
| D-LSTM WITH DELAY=5 | LSTM | $91.88 \pm 0.18$ | $91.54 \pm 0.11$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=1 | $93.31 \pm 0.14$ | $92.74 \pm 0.10$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=2 | $93.17 \pm 0.13$ | $92.57 \pm 0.17$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=3 | $92.84 \pm 0.19$ | $92.25 \pm 0.10$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=4 | $92.50 \pm 0.22$ | $91.96 \pm 0.19$ |

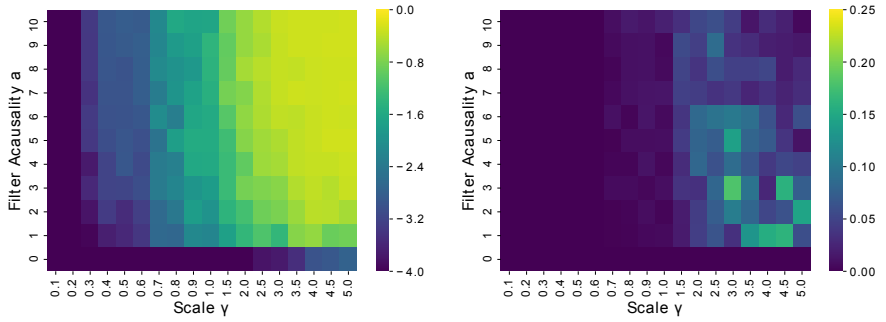
*Language Resources and Evaluation (LREC'14)*, pp. 2897–2904, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1089_Paper.pdf.

*Table 5.* Parts-of-Speech results for English. The table shows all possible combinations of delays or bidirectional LSTM networks. The best forward-only network is marked in bold.

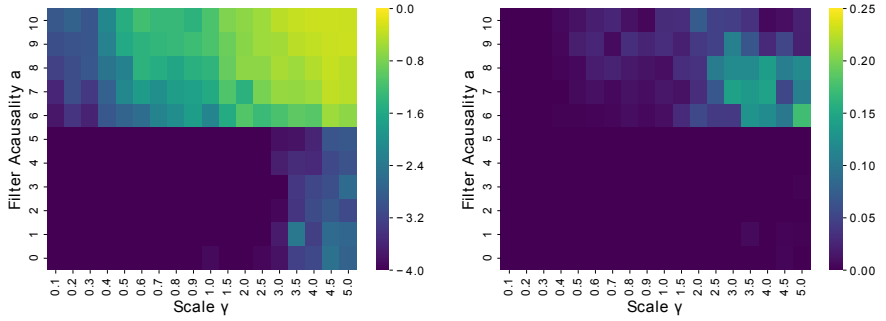
| CHARACTER-LEVEL NETWORK | WORD-LEVEL NETWORK | VALIDATION ACCURACY | TEST ACCURACY |
|---|---|---|---|
| BI-LSTM | BI-LSTM | $94.85 \pm 0.05$ | $94.84 \pm 0.08$ |
| BI-LSTM | LSTM | $91.90 \pm 0.12$ | $92.05 \pm 0.09$ |
| BI-LSTM | D-LSTM WITH DELAY=1 | $94.47 \pm 0.06$ | $94.41 \pm 0.05$ |
| BI-LSTM | D-LSTM WITH DELAY=2 | $94.17 \pm 0.13$ | $94.14 \pm 0.10$ |
| BI-LSTM | D-LSTM WITH DELAY=3 | $93.70 \pm 0.07$ | $93.87 \pm 0.07$ |
| BI-LSTM | D-LSTM WITH DELAY=4 | $93.11 \pm 0.14$ | $93.26 \pm 0.08$ |
| BI-LSTM | D-LSTM WITH DELAY=5 | $92.54 \pm 0.16$ | $92.70 \pm 0.10$ |
| LSTM | BI-LSTM | $95.03 \pm 0.14$ | $94.99 \pm 0.15$ |
| LSTM | LSTM | $92.05 \pm 0.13$ | $92.14 \pm 0.10$ |
| LSTM | D-LSTM WITH DELAY=1 | $94.53 \pm 0.08$ | $94.58 \pm 0.11$ |
| LSTM | D-LSTM WITH DELAY=2 | $94.29 \pm 0.05$ | $94.28 \pm 0.05$ |
| LSTM | D-LSTM WITH DELAY=3 | $93.81 \pm 0.11$ | $93.85 \pm 0.12$ |
| LSTM | D-LSTM WITH DELAY=4 | $93.39 \pm 0.12$ | $93.55 \pm 0.10$ |
| D-LSTM WITH DELAY=1 | BI-LSTM | $94.94 \pm 0.07$ | $94.95 \pm 0.06$ |
| D-LSTM WITH DELAY=1 | LSTM | $91.96 \pm 0.16$ | $92.09 \pm 0.10$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=1 | $94.57 \pm 0.08$ | $\mathbf{94.57 \pm 0.14}$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=2 | $94.29 \pm 0.12$ | $94.37 \pm 0.08$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=3 | $93.86 \pm 0.05$ | $93.84 \pm 0.10$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=4 | $93.35 \pm 0.10$ | $93.56 \pm 0.13$ |
| D-LSTM WITH DELAY=3 | BI-LSTM | $94.98 \pm 0.09$ | $94.91 \pm 0.10$ |
| D-LSTM WITH DELAY=3 | LSTM | $91.96 \pm 0.08$ | $92.08 \pm 0.10$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=1 | $94.47 \pm 0.03$ | $94.51 \pm 0.10$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=2 | $94.21 \pm 0.05$ | $94.18 \pm 0.03$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=3 | $93.80 \pm 0.13$ | $93.88 \pm 0.13$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=4 | $93.23 \pm 0.13$ | $93.38 \pm 0.11$ |
| D-LSTM WITH DELAY=5 | BI-LSTM | $94.90 \pm 0.07$ | $94.87 \pm 0.09$ |
| D-LSTM WITH DELAY=5 | LSTM | $91.84 \pm 0.11$ | $91.98 \pm 0.20$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=1 | $94.36 \pm 0.09$ | $94.44 \pm 0.08$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=2 | $94.05 \pm 0.07$ | $94.19 \pm 0.05$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=3 | $93.61 \pm 0.07$ | $93.76 \pm 0.05$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=4 | $93.14 \pm 0.04$ | $93.27 \pm 0.12$ |



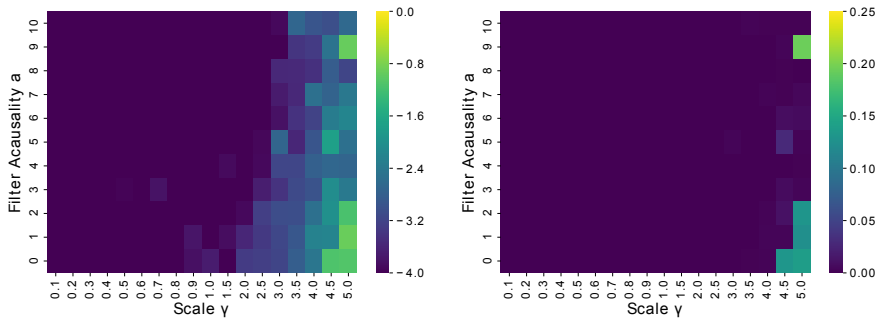

(a) LSTM




(b) Bi-LSTM




(c) d-LSTM with delay=5




(d) d-LSTM with delay=10

*Figure 7.* Error maps presented in Figure 4 (left column) together with their standard deviation figures.

*Table 6.* Parts-of-Speech results for French. The table shows all possible combinations of delays or bidirectional LSTM networks. The best forward-only network is marked in bold.

| CHARACTER-LEVEL NETWORK | WORD-LEVEL NETWORK | VALIDATION ACCURACY | TEST ACCURACY |
|---|---|---|---|
| BI-LSTM | BI-LSTM | $97.63 \pm 0.06$ | $97.22 \pm 0.11$ |
| BI-LSTM | LSTM | $96.67 \pm 0.05$ | $96.15 \pm 0.17$ |
| BI-LSTM | D-LSTM WITH DELAY=1 | $97.48 \pm 0.02$ | $96.98 \pm 0.05$ |
| BI-LSTM | D-LSTM WITH DELAY=2 | $97.41 \pm 0.02$ | $96.91 \pm 0.12$ |
| BI-LSTM | D-LSTM WITH DELAY=3 | $97.31 \pm 0.05$ | $96.84 \pm 0.09$ |
| BI-LSTM | D-LSTM WITH DELAY=4 | $97.12 \pm 0.05$ | $96.61 \pm 0.06$ |
| BI-LSTM | D-LSTM WITH DELAY=5 | $96.88 \pm 0.10$ | $96.20 \pm 0.14$ |
| LSTM | BI-LSTM | $97.70 \pm 0.07$ | $97.19 \pm 0.09$ |
| LSTM | LSTM | $96.67 \pm 0.07$ | $96.10 \pm 0.11$ |
| LSTM | D-LSTM WITH DELAY=1 | $97.49 \pm 0.07$ | $97.03 \pm 0.07$ |
| LSTM | D-LSTM WITH DELAY=2 | $97.49 \pm 0.05$ | $97.00 \pm 0.06$ |
| LSTM | D-LSTM WITH DELAY=3 | $97.34 \pm 0.04$ | $96.89 \pm 0.09$ |
| LSTM | D-LSTM WITH DELAY=4 | $97.16 \pm 0.06$ | $96.66 \pm 0.15$ |
| D-LSTM WITH DELAY=1 | BI-LSTM | $97.67 \pm 0.07$ | $97.23 \pm 0.12$ |
| D-LSTM WITH DELAY=1 | LSTM | $96.66 \pm 0.06$ | $95.97 \pm 0.07$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=1 | $97.49 \pm 0.04$ | $\mathbf{97.04 \pm 0.13}$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=2 | $97.43 \pm 0.05$ | $96.98 \pm 0.05$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=3 | $97.36 \pm 0.08$ | $96.80 \pm 0.10$ |
| D-LSTM WITH DELAY=1 | D-LSTM WITH DELAY=4 | $97.22 \pm 0.06$ | $96.57 \pm 0.10$ |
| D-LSTM WITH DELAY=3 | BI-LSTM | $97.67 \pm 0.08$ | $97.21 \pm 0.08$ |
| D-LSTM WITH DELAY=3 | LSTM | $96.67 \pm 0.07$ | $95.98 \pm 0.14$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=1 | $97.52 \pm 0.04$ | $97.02 \pm 0.09$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=2 | $97.44 \pm 0.02$ | $96.97 \pm 0.12$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=3 | $97.28 \pm 0.04$ | $96.74 \pm 0.07$ |
| D-LSTM WITH DELAY=3 | D-LSTM WITH DELAY=4 | $97.13 \pm 0.05$ | $96.57 \pm 0.09$ |
| D-LSTM WITH DELAY=5 | BI-LSTM | $97.61 \pm 0.03$ | $97.12 \pm 0.06$ |
| D-LSTM WITH DELAY=5 | LSTM | $96.64 \pm 0.06$ | $96.08 \pm 0.08$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=1 | $97.46 \pm 0.02$ | $96.96 \pm 0.13$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=2 | $97.41 \pm 0.06$ | $96.87 \pm 0.06$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=3 | $97.36 \pm 0.05$ | $96.82 \pm 0.07$ |
| D-LSTM WITH DELAY=5 | D-LSTM WITH DELAY=4 | $97.15 \pm 0.05$ | $96.51 \pm 0.07$ |