Architecture	WSJ Accuracy
GRU	96.43
LSTM	96.47
Bidirectional GRU	97.28
b-LSTM	97.25
INN	**97.37**
Stanford POS Tagger	97.33

Table 2: Tagging performance relative to recurrent architectures and Stanford POS Tagger.

4 Time Complexity

The implicit experiments in this paper took approximately 3-5 days to run on a single Tesla K40, while the explicit experiments took approximately 1-3 hours. Running time of the solver is approximately $n_n \times n_b \times t_b$ where n_n is the number of Newton iterations, n_b is the number of BiCG-STAB iterations, and t_b is the time for a single BiCG-STAB iteration. t_b is proportional to the number of non-zero entries in the matrix (Van der Vorst, 1992), in our case $n(2k^2 + 1)$. Newton's method has second order convergence (Isaacson and Keller, 1994), and while the specific bound depends on the norm of $(I - \nabla_H F)^{-1}$ and the norm of its derivatives, convergence is well-behaved. For n_b, however, we are not aware of a bound. For symmetric matrices, the Conjugate Gradient method is known to take $O(\sqrt{\kappa})$ iterations (Shewchuk et al., 1994), where κ is the condition number of the matrix. However, our matrix is nonsymmetric, and we expect κ to vary from problem to problem. Because of this, we empirically estimated the correlation between sequence length and total time to compute a batch of 20 hidden layer states.

For the random walk experiment with $b = 0.5$, we found the the average run time for a given sequence length to be approximately $0.17n^{1.8}$, with $r^2 = 0.994$. Note that the exponent would have been larger had we not truncated the number of BiCG-STAB iterations to 40, as the inner iteration frequently hit this limit for larger n. However, the average number of Newton iterations did not go above 10, indicating that exiting early from the BiCG-STAB loop did not prevent the Newton solver from converging. Run times for the other random walk experiments were very similar, indicating run time does not depend on b; However, for the POS task runtime was $0.29n^{1.3}$, with $r^2 = 0.910$.

5 Conclusion and Future Work

We have introduced a novel, implicitly defined neural network architecture based on the GRU and shown that it outperforms a b-LSTM on an artificial random walk task and slightly outperforms both the Stanford Parser and a baseline bidirectional network on the Penn Treebank Part-of-Speech tagging task.

In future work, we intend to consider implicit variations of other architectures, such as the LSTM, as well as additional, more challenging, and/or data-rich applications. We also plan to explore ways to speed up the computation of $(I - \nabla_H F)^{-1}$. Potential speedups include approximating the hidden state values by reducing the number of Newton and/or BiCG-STAB iterations, using cached previous solutions as initial values, and modifying the gradient update strategy to keep the batch full at every Newton iteration.

6 Acknowledgements

This work would not be possible without the support and funding of the Air Force Research Laboratory. We also acknowledge Nick Malyska, Elizabeth Salesky, and Jonathan Taylor at MIT Lincoln Lab for interesting technical discussions related to this work.

Cleared for Public Release on 29 Jul 2016. Originator reference number: RH-16-115722. Case Number: 88ABW-2016-3809.