	Logical Inference						ListOps
Ordered Memory	98 ± 0.4	97 ± 0.5	96 ± 0.8	94 ± 0.8	94 ± 1.5	92 ± 0.7	99.9 ± 0.02
~~cell(·)~~ TreeRNN Op.	69	67	65	61	57	53	63.1
~~Stick-breaking~~ softmax	98	97	96	92	93	92	98
Yogatama et al. (2018)	58	57	58	56	55	50	60

Table 1: We replaced the $\mathrm{cell}(\cdot)$ operator with the RNN operator found in TreeRNN, which is the best performing model that explicitly uses the structure of the logical clause. In this test, we find that the TreeRNN operator results in a large drop across the different tasks. We also replaced the stick-breaking process with masked and scaled softmax: $p_t = \mathrm{softmax}\left(\frac{\hat{\beta}_t}{\sqrt{d}}\right)$ where $\hat{\beta}_t$ is defined in Section 3.1 and d is the dimension of memory slot. The purpose of this is to scale down the logits before softmax is applied, a technique similar to the one seen in Vaswani et al. (2017). Surprisingly, we observed that the masked and scaled softmax results in a more robust model (the model is less sensitive to hyper-parameters, and thus easier to train) while the stick-breaking formulation provides marginally better performance. The reason could be that softmax is more numerically stable for both feedforward and backpropagation. As discussed in Section 3.3 of the paper, the stick-breaking formulation was initially used to reflect the process that a shift-reduce parser would make if the decisions were made one after another.

Thank you all for your detailed review and insightful comments.

Firstly, to address comments on using natural language data: We have indeed found it challenging to learn structure from real language data and associated tasks. We think that a natural language task with a more informative signal (perhaps language modelling) would be able to correct this. This will be the direction in which we take our future work.

We have conducted an ablation test for the Gated Recursive Cell and Stick-breaking Attention. Results of ablation test are shown in above table. We will add more discussion about different attention methods to our paper.

Reviewer 1 You are correct regarding complexity of the model during training time. We will include a description of this in the camera-ready version. We will also fix the bibliography to reflect the conferences/proceedings the arxiv-ed papers were published in.

As for the reproducibility checklist, we thought that checking that box meant that we would release the code *after* the paper has been published.

Reviewer 2 We have actually included the performance of models that learn a tree structure. For ListOps, we also have also listed the results for RL-SPINN, which learns to use a stack. In addition, we have also tested our implementation of the stack-augmented model in Yogatama et al. (2018). We currently have preliminary results using that model (See the above table for detailed results on Logical Inference and ListOps). Note that these results are expected to change as we find better hyperparameters for this model.

We will correct the typos in the paper for the camera-ready version. Our apologies for not catching them before submission.

Reviewer 3 We find that it is difficult to show that our model learns compositionality on real language due to the lack of datasets that explicitly test for this property. The various toy tasks that we have tested our model on were designed to isolate this capability, and so we have used them to demonstrate this to the extent that we can. And because we know the structure of the data in the logical inference task, we were also able to remove clauses from training in order to see how well the model generalises when tested on them during evaluation. In those cases, to generalise to those unseen substructures requires compositionality.

We understand your criticism with respect to ablation studies. We have provided the details of the ablation studies above, and the detailed results are shown in the table.

Also, we apologise for the lengthy discussion of related work, as we thought providing a comprehensive coverage of existing work would show the state of the field much more clearly. We will make amendments to the related work section as we accommodate all the reviewer comments in our camera-ready version.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Dani Yogatama, Yishu Miao, Gabor Melis, Wang Ling, Adhiguna Kuncoro, Chris Dyer, and Phil Blunsom. 2018. Memory architectures in recurrent neural network language models.