Critique : Zoneout: Regularizing RNNS By Randomly Preserving Hidden Activations (accepted as poster at ICLR'17)

Anshuman Suri, 2014021

I. Summary

Well known Regularization techniques exist for neural networks and CNNs: dropout, drop-connect, early stopping, L_n norm, etc. However, not as much extensive research has been done in RNNs. Authors, in this paper, propose a new kind of regularization technique (synonymous to dropout in normal neural networks), which they call *Zoneout*, which is shown to give performance boosts comparable/better than existing regularization techniques. Their method, which they show via experiments on standard benchmarks data-sets for RNNs, gives boosts even when applied directly to existing architectures (with the same parameters). Thus, they pitch it as a method that can be potentially used as a plug-and-use technique for existing models.

Zoneout like dropout, randomly injects noise during training. However, it sets activations as previous activations , unlike 0 in normal dropout. *Zoneout* (shown experimentally as well as intuitive from the way it is defined) also helps overcome the vanishing gradient problem (which is much more pronounced in RNNs), since it preserves information flow in time.

Analysis of related work in this field shows that though research has been done on different regularizers for RNNs, like applying layer-wise dropout (not across time), recurrent dropout (dropping certain activations in an LSTM gate), *rnnDrop*, etc. However, almost all of these techniques either suffer from the vanishing gradient problems, are not feasible for long sequences, or are specific to LSTM/GRU architectures. Another benefit that *Zoneout* has over the others is that it is much more generic and easier to understand/reason. However, it is not discussed why some of the existing literature (like dynamically re-scaling row weights) is not good enough, or where it lacks.

Experiments are three data-sets (standard in RNN related literature) indicate that this form of regularization works, and can be used without changes in hyper-parameters to existing models.The performance boost isn't much in comparison to Recurrent Dropout based regularization, but significant when compared to a vanilla setting (for all three data-sets).

The effect of using this form of regularization on gradient flow is also looked at experimentally, showing that *Zoneout* helps smoothen the exploding gradient problem, which Dropout is prone to. Thus the advantage Dropout has over *Zoneout* in terms of performance over the data-sets shown (though it may be not much) is overshadowed by the gradient problem. One can, thus, conclude that *Zoneout* works much better practically as it is not susceptible to the exploding gradient problems and generates results comparable to the case of Dropout, thus giving the best of both worlds.

II. Limitations/Improvements

Overall the paper is a good one (must be; it is published as a Poster paper at ICLR'17), introducing a simple yet efficient for regularizing recurrent neural networks. However, there are certain limitations which could ave potentially lead to this paper being rejected. Some of them are:

- As pointed out by the reviewers as well, the authors do not use any hyper-parameter search while reporting results with *Zoneout*. This indirectly implies that using hyper-parameter search may yield better results. However, this is not necessary: it may lead worse results as well. Thus, they use the benefit of doubt to make their technique look very robust. This could have been fixed by trying out hyper-parameter tuning: the results may have worked out in their favor.
- Results obtained for the three data-sets are not significantly better than some of already existing regularization techniques. Thus, the motivation to try *Zoneout*, in terms of a boost in performance, is not very strong. Moreover, it is not mentioned whether only one iteration of training/testing was performed, or it was done multiple times. The slight variations in results could be because of variations based on model-randomness as well, which should have been ruled out by averaging results over multiples passes of training/testing.
- Though pitched as a regularization method, it is not shown mathematically how it works as a regularizer, or how it helps the network generalize better. In fact, given the fact that it tackles the gradient problem, it should be considered more of an optimization technique. This description would fit well, as the improvement in results is anyway not significant.
- For analyzing gradient flow, comparison is done only with this technique, the one that worked best for the data-sets (Dropout) and a vanilla model. Moreover, this analysis is done for only one epoch, and that too for just one data-set. Such an analysis is highly biased and no conclusions should be drawn from such an analysis. Other methods which have results comparable to *Zoneout* could have had better gradient flow, but again the authors uses the benefit of doubt to pitch their method as a better one. Gradients for other methods should have been analyzed as well, for a full comprehensive analysis.