EECS 182 Deep Neural Networks
Fall 2022 Anant Sahai

Discussion 6

1. **Inductive Biases in Standard Deep Learning Building Blocks** In this question, we will compile and review inductive biases that standard deep learning building blocks have. This question is inspired by Battaglia et al. (2018), so we highly recommend you read through this paper.

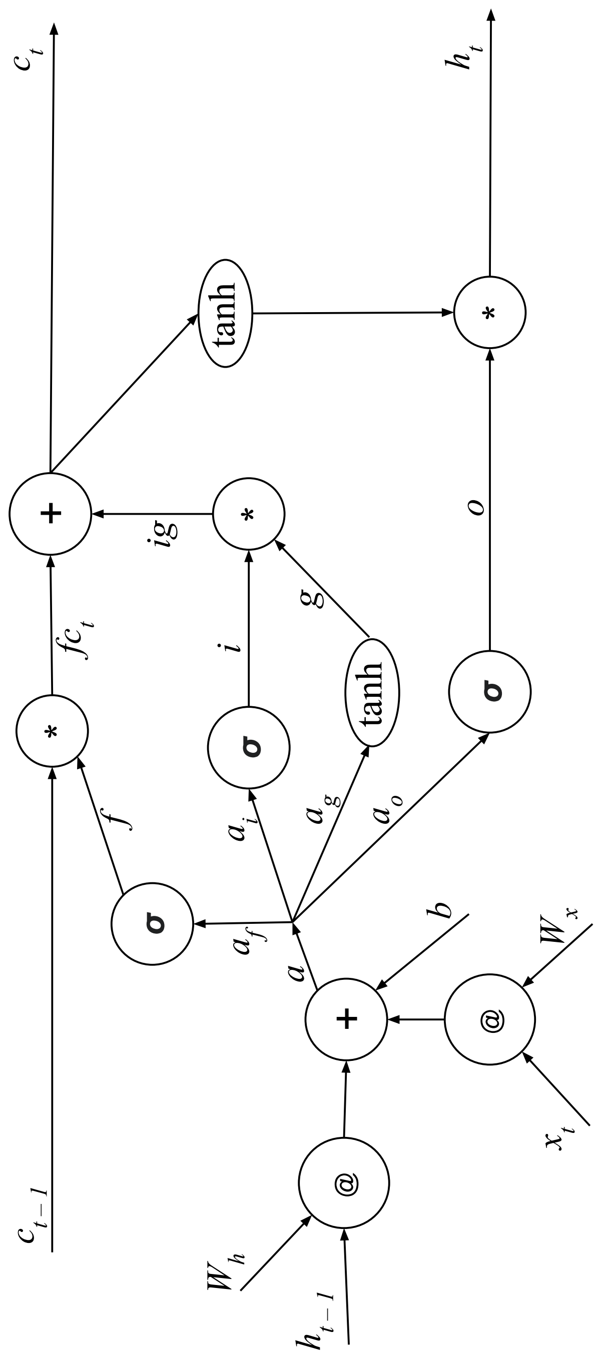
   Inductive biases are any assumptions that learners utilize to learn the world and predict the output. These assumptions can either be related to the date-generation process, the functional family space chosen, or the space of the solutions. Inductive biases could be a regularization term added for better generalization or preventing overfitting, and they could also be embedded in the model function family or model architecture (e.g. CNN vs MLP). Inductive biases generally reduce the amount of data needed to fit the model while constraining the model's flexibility. Ideally, inductive biases will improve training efficiency (small data points, small gradient steps, faster optimization), while maintaining performance and generalizing well. However, if inductive biases are mismatched with the problem domain, this will lead our solutions to be sub-optimal. Another way to think about this is through the lens of the bias-variance tradeoff: inductive biases induce a large bias term in approximation error.

   The first example of inductive bias we'll look at is Ridge Regression. What inductive biases lead us to introduce a L2 Penalty on the ordinary least squares objective?

   Now let's discuss the inductive biases of deep learning building blocks. Fill in the table below.

Blocks	Inputs	Interactions	Invariance	Inductive Bias
Fully Connected				
Convolutional				
Recurrent				

2. Long Short Term Memory (LSTM) Vanilla RNNs suffer from vanishing / exploding gradients. The intuition behind LSTMs is similar to that of skip connections that allowed us to pass gradients forward a few layers to mitigate this problem. LSTMs rely on a different kind of recurrent cell, whose schematic is shown below. The @ symbol represents matrix-vector multiplication.



You may find the following derivatives useful. $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, and $\tanh'(x) = 1 - \tanh^2(x)$

Let's say the upstream gradients are δh_t and δc_{next}. Trace the gradients through the LSTM cell above!

References

Battaglia, P., Hamrick, J. B. C., Bapst, V., Sanchez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G. E., Vaswani, A., Allen, K., Nash, C., Langston, V. J., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., and Pascanu, R. Relational inductive biases, deep learning, and graph networks. *arXiv*, 2018. URL https://arxiv.org/pdf/1806.01261.pdf.

Contributors:

- Anant Sahai.
- Kumar Krishna Agrawal.
- Matthew Lacayo.
- Suhong Moon.
- Shivam Singhal.
- Bryan Wu.
- Saagar Sanghavi.

