1. the qualitative difference between the functions computed by the hidden layer nodes of the three models, and a brief description of how the network uses these functions to achieve the classification
2. the effect of different values for initial weight size on the speed and success of learning, for both RawNet and ShortNet
3. the relative "naturalness" of the output function computed by the three networks, and the importance of representation for deep learning tasks in general
4. you may like to also experiment with other changes and comment on the result - for example, changing batch size from 97 to 194, using SGD instead of Adam, changing tanh to relu, adding a third hidden layer, etc.