1. $y = f_2(W_{(2)}^T f_1(W_{(1)}^T x + b_{(1)}) + b_{(2)})$
   $Y_2 = f_2(W_{(ho)} a_2 + b_2$
   $a_2 = f(z_2)$
   $z_2 = W_{(ih)} x_2 + W_{(hh)} + a_1 + b_1$
   $a_1 = f(z_1)$
   $z_1 = W_{(ih)} x_1 + W_{(hh)} + a_0 + b_1$
   $a_0 = h_0$

2. (a) $y_3 - y_{10}$ and $h_3 - h_{10}$
   (b) $h_1 - h_{10}$ and $y_1 - y_{10}$
   (c) $h_0 - h_{10}$ and $y_1 - y_{10}$
   (d) $y_1 - y_{10}$

3. (a) $\delta(2,1) \sim \delta(2,10)$
   (b) $\delta(2,1) \sim \delta(1,5)$
   (c) $\delta(1,10) \rightarrow \delta(1,5)$
   (d) $\delta(1,10) \leftarrow \delta(2,10)$
   (e) $\delta(1,10) \sim \delta(2,7)$
   (f) $\delta(2,7) \rightarrow \delta(1,5)$

4. Doubling *num_steps* would slow down training more because more steps in the backpropagation would have to be calculated. Doubling the minibatch size would also likely make training faster because it has to load fewer batches to train through the model.

5. Gradient clipping means that when we go to take a step through the gradient we can limit the magnitude of the step size to help us avoid erratic or unpredictable steps through the gradient. For example if a gradient cliff exists and as we approach it we might be "thrown off", we can clip the maximum magnitude that the step can take so that we make a more reasonable step.

6. Unidirectional:

   (a) Predictive Policing / Pattern Recognition in human activity
   (b) Haiku generator that takes in a sequence of x-syllable words and produces a haiku

   Bidirectional:

   (a) One direction converts Music to sheet music, and the other direction converts sheet music back to music
   (b) One direction takes a picture of a person as an individual input and show what they would look like aged by a set amount of time. The other direction would reverse age them.

7. The major advantage of LSTMs is they maintain memory cells for long periods of time (thus the long-short term memory). This helps to mitigate the vanishing gradient problem helping to model long-term dependencies, a shortcoming of vanilla RNNs.