1) Following is my network architecture:

(layers): SequentialLayer:
  (0): ConvLayer: Kernel: (3, 3) In Channels 1 Out Channels 16 Stride 1
  (1): MaxPoolLayer: kernel: 2 stride: 2
  (2): ReLULayer:
  (3): ResNetBlock:
    (conv_layers): SequentialLayer:
      (0): ConvLayer: Kernel: (3, 3) In Channels 16 Out Channels 16 Stride 1
      (1): ReLULayer:
      (2): ConvLayer: Kernel: (3, 3) In Channels 16 Out Channels 16 Stride 1
    (add_layer): AddLayer:
    (relu2): ReLULayer:
  (4): MaxPoolLayer: kernel: 2 stride: 2
  (5): ConvLayer: Kernel: (3, 3) In Channels 16 Out Channels 6 Stride 1
  (6): FlattenLayer:
  (7): LinearLayer: (294, 120)
  (8): ReLULayer:
  (9): LinearLayer: (120, 84)
  (10): ReLULayer:
  (11): LinearLayer: (84, 10)

With this relatively shallow architecture, I was able to get %99 in 27 epochs.

2) Using sigmoid slightly reduced the accuracy to 98% on my best model and using pReLU did not change the accuracy.

3) With the following network I was able to achieve 99.1% in 40 epochs.

(layers): SequentialLayer:
  (0): ConvLayer: Kernel: (5, 5) In Channels 1 Out Channels 6 Stride 1
  (1): ResNetBlock:
    (conv_layers): SequentialLayer:
      (0): ConvLayer: Kernel: (3, 3) In Channels 6 Out Channels 6 Stride 1
      (1): ReLULayer:
      (2): ConvLayer: Kernel: (3, 3) In Channels 6 Out Channels 6 Stride 1
    (add_layer): AddLayer:
    (relu2): ReLULayer:
  (2): MaxPoolLayer: kernel: 2 stride: 2
  (3): ReLULayer:
  (4): ResNetBlock:
    (conv_layers): SequentialLayer:
      (0): ConvLayer: Kernel: (3, 3) In Channels 6 Out Channels 6 Stride 1
      (1): ReLULayer:
      (2): ConvLayer: Kernel: (3, 3) In Channels 6 Out Channels 6 Stride 1
    (add_layer): AddLayer:
    (relu2): ReLULayer:

(5): ConvLayer: Kernel: (5, 5) In Channels 6 Out Channels 16 Stride 1
(6): ReLULayer:
(7): ResNetBlock:
  (conv_layers): SequentialLayer:
    (0): ConvLayer: Kernel: (3, 3) In Channels 16 Out Channels 16 Stride 1
    (1): ReLULayer:
    (2): ConvLayer: Kernel: (3, 3) In Channels 16 Out Channels 16 Stride 1
  (add_layer): AddLayer:
  (relu2): ReLULayer:
(8): MaxPoolLayer: kernel: 2 stride: 2
(9): ReLULayer:
(10): FlattenLayer:
(11): LinearLayer: (784, 120)
(12): ReLULayer:
(13): LinearLayer: (120, 84)
(14): ReLULayer:
(15): LinearLayer: (84, 10)