(a)

- Input: 32x32x3
- Conv1: 28x28x16
- MaxPool1: 14x14
- Conv2: 12x12x32
- MaxPool2: 6x6x32
- Fully connected: 128
- Output: 10

(b)

- Conv1: 5*5*16(filtes) + 16(biases)=1216
- Conv2: 3*3*32(filters) + 32(biases)=1472
- Fully connected:
  - inputs from MaxPool2 = 6*6*32=1152
  - biases: 128
  - total: 1152*128+128=147584

*Output: 128*10+10(biases) = 1290

Total: 1216 + 1472 + 147584 + 1290 = **151562**

(c)

Forward Pass Normalization:

x as the input to the batch normalization layer (output of Conv1 in this case)
μ_B as the mean of the input batch: $\mu_B = (1/m)*\Sigma(x_i)$ (*m* is the batch size)
σ_B^2 as the variance of the input batch: $\sigma_B^2=(1/m)*\Sigma(x_i-\mu_B)^2$
ϵ as a small constant for numerical stability
γ as the scale parameter (learnable)
β as the shift parameter (learnable)

Input normalization: $\bar{x}=(x_i-\mu_B)/\sqrt{(\sigma_B^2+\epsilon)}$
Scale and shift: $y_i=\gamma*\bar{x}+\beta$

Backpropagation:

Gradient with respect to y_i :
$\partial L/\partial\bar{x}=\partial L/\partial y_i*\gamma \qquad \partial L/\partial\beta=\Sigma(\partial L/\partial y_i)$
Gradient with respect to x_i :
$\partial L/\partial x_i = \partial L/\partial\bar{x}*(1/m)*(1/\sqrt{(\sigma_B^2+\epsilon)}) * (m-(x_i-\mu_B)*(1/m)*\Sigma(\partial L/\partial\bar{x}))$
Gradient with respect to γ:
$\partial L/\partial\gamma=\Sigma(\partial L/\partial y_i*\bar{x})$
Gradient with respect to β:
$\partial L/\partial\beta=\Sigma(\partial L/\partial y_i)$