DS 4440

Activation Patching

Which (learned) features led to prediction A rather than B?

Approach (general)

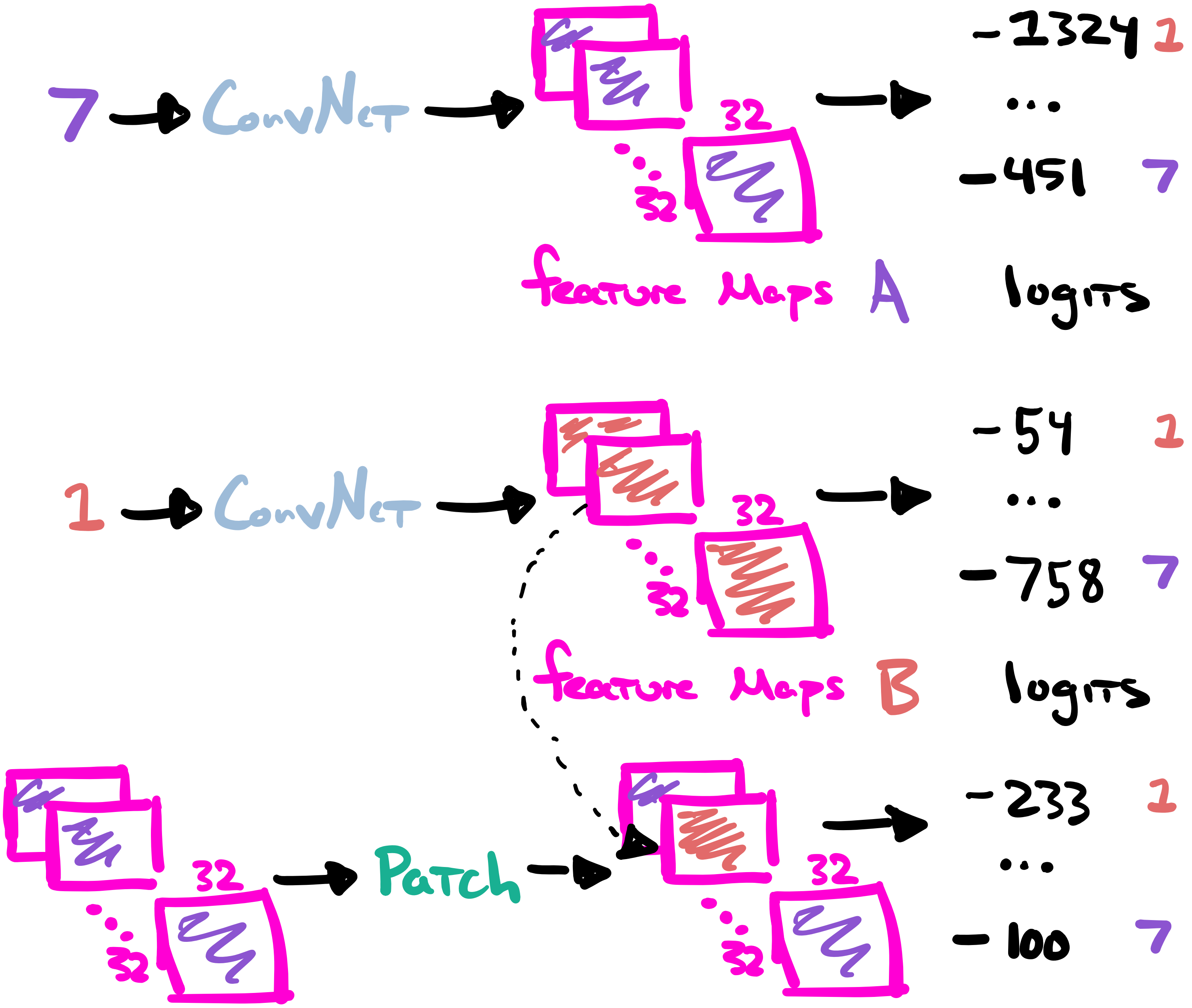
(i) Run A and B fwd through Network

(ii) Collect activations of interest from B pass

(iii) Patch these in to same location in A pass

(iv) Observe change (Δ) in output

For the homework, we consider ConvNet and specifically the last layer feature maps (32 x 32); We have 64 of them. Which is most important for telling 7s from 1s?



Feature Maps A New logits

$\Delta_{7,1} \overset{..}{=} \tilde{y}_7 - \tilde{y}_1$; measure for patched and 1s

$$\Delta_{7,1} - \Delta_{7,1} = \underline{(\overset{\tilde{y}_7}{-100} - (\overset{\tilde{y}_1}{-233}))} - \underline{(-758 - (-54))}$$

$$= 133 - 704 = 837$$

(if $\tilde{y}_1$ smaller say $-1000 \rightarrow 1604$ ↑

bigger say $+1000 \rightarrow -396$ ↓

$\tilde{y}_7$ smaller say $-1000 \rightarrow -63$ ↓

bigger say $+1000 \rightarrow 1937$ ↑)

The range is funny so we normalize (ish) by

$$\underbrace{\Delta_{7,1}} - \Delta_{7,1}$$

for run with 7s ; sort of an upper bound