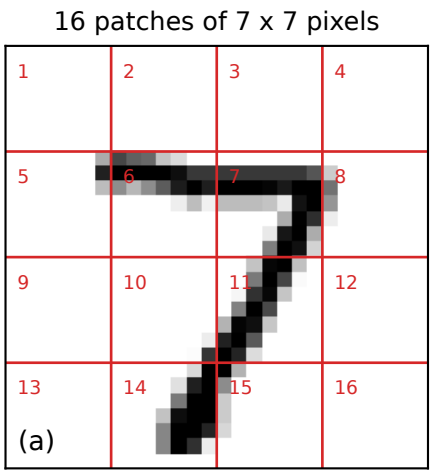
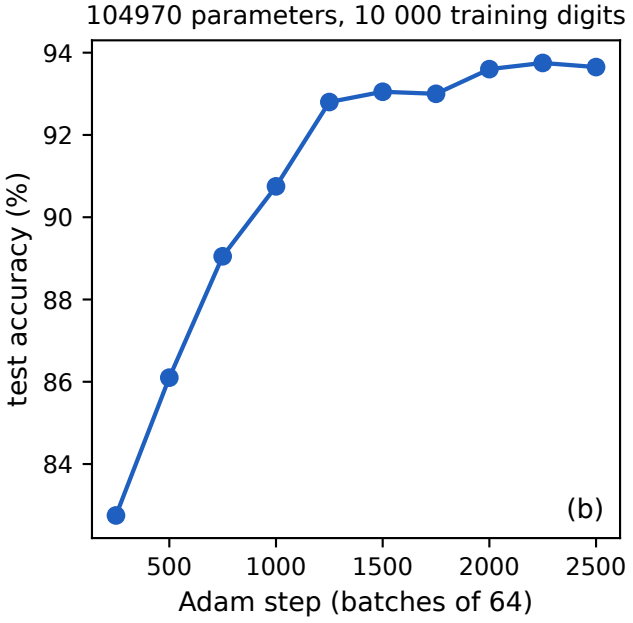
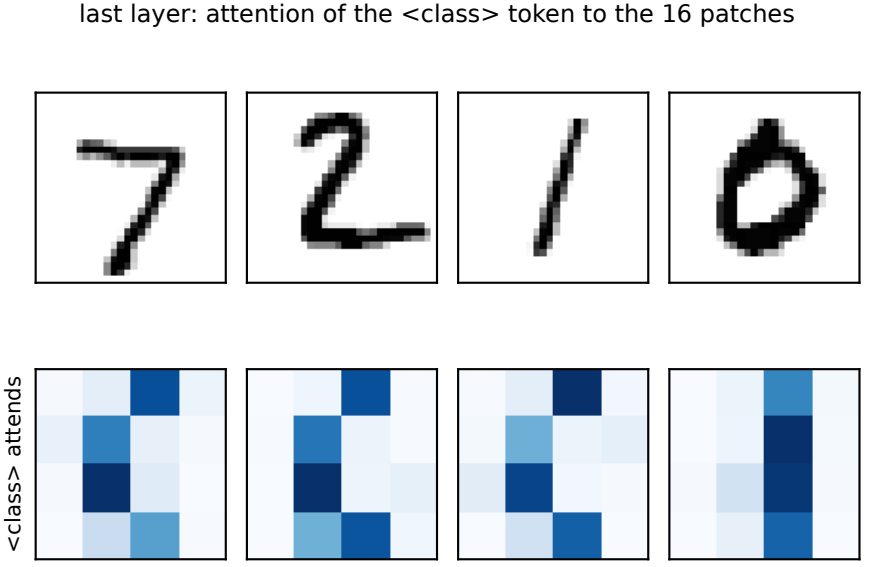

16 patches of 7 x 7 pixels
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
(a)


104970 parameters, 10 000 training digits
test accuracy (%)
94
92
90
88
86
84
500
1000
1500
2000
2500
Adam step (batches of 64)
(b)


last layer: attention of the <class> token to the 16 patches
<class> attends