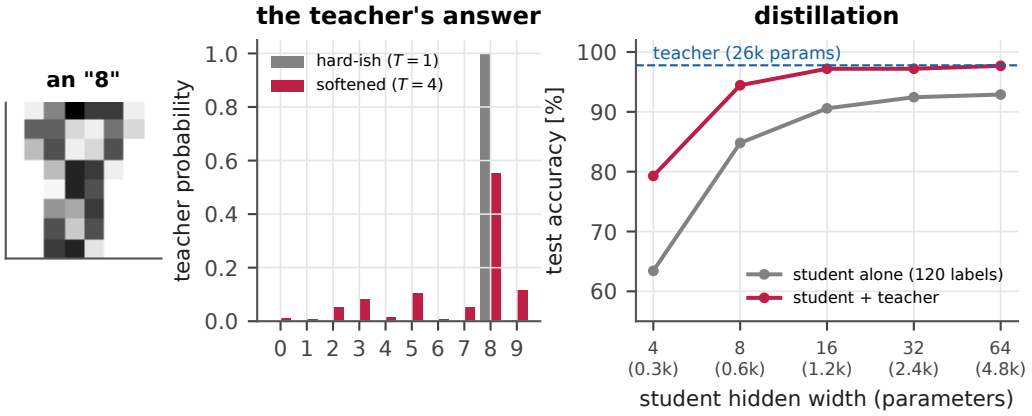

an "8"
the teacher's answer
hard-ish ($T = 1$)
softened ($T = 4$)
teacher probability
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5 6 7 8 9
distillation
teacher (26k params)
test accuracy [%]
100
90
80
70
60
student alone (120 labels)
student + teacher
4 (0.3k)
8 (0.6k)
16 (1.2k)
32 (2.4k)
64 (4.8k)
student hidden width (parameters)