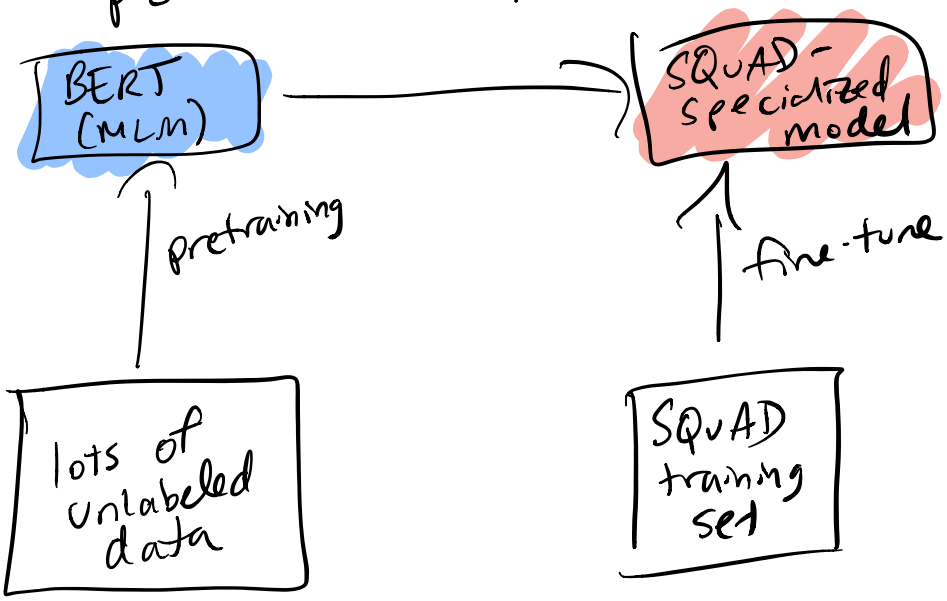
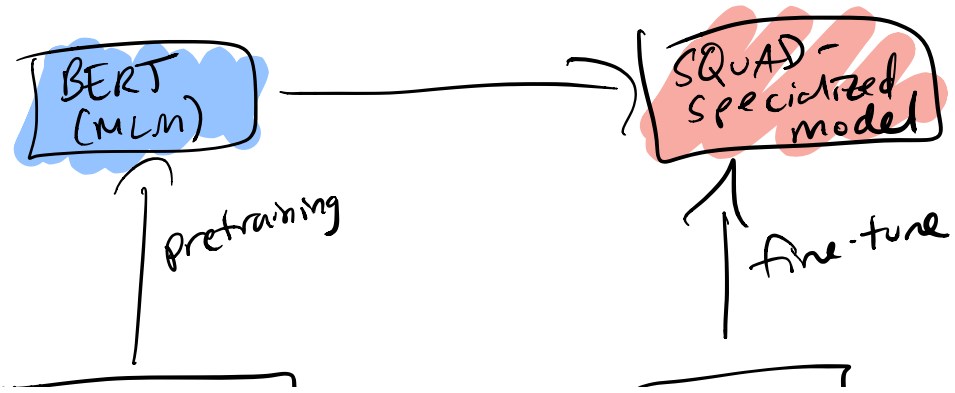
Today: intermediate task fine-tuning

- imagine I am trying to optimize perf. on SQuAD



can we leverage other QA datasets to improve our SQuAD test-time perf?

↳ one way: multi-task learning



lots of unlabeled data

SQuAD + HotpotQA + NewsQA + ...

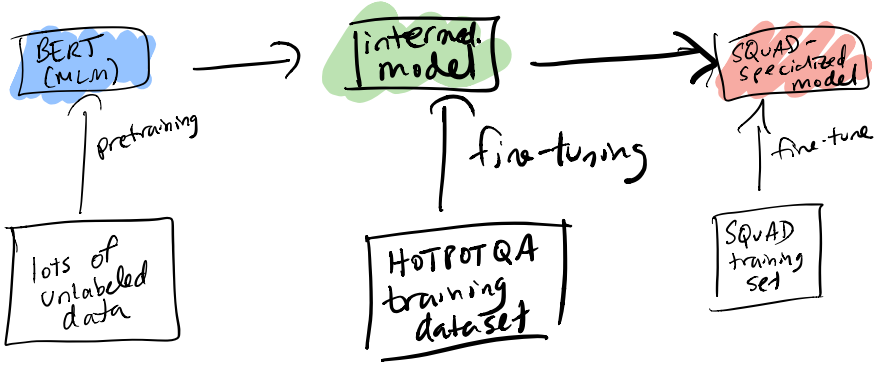
assume we train on SQuAD + HotpotQA

- MTL:

$$L = \lambda L_{SQUAD} + (1-\lambda) L_{HOTPOTQA}$$

if I care about SQuAD, maybe I use a high λ

- how to choose λ?



1. how do we know what intermediate task will result in the biggest downstream improvem.?
   - task similarity (e.g. QA/QA vs. Sentiment/QA)

- size of intermediate dataset
  (e.g. 100 QA examples vs. 100,000 sentiment examples)
- domain similarity
  (e.g. 10,000 QA examples from medical journals vs. 10,000 sentiment examples from Wikipedia)

- Can we predict which task (out of some finite set of tasks) will be most useful as an intermediate task given a specific downstream dataset?