Today: intermediate task fine-tuning

- imagine I am trying to optimize perf. on SQuAD

BERT (MLM) \(\rightarrow\) SQuAD-specialized model

\(\text{pretraining} \rightarrow \text{finetune}\)

lots of unlabeled data

SQuAD training set

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

\(\rightarrow\) one-way: multi-task learning
lots of unlabeled data

SQUAD
SQUAD + HOTPOTQA
SQUAD + HOTPOTQA + NEWSQA + ...

assume we train on SQUAD + HOTPOTQA

- MTL:
  \[ L = \lambda L_{\text{SQUAD}} + (1 - \lambda) L_{\text{HOTPOTQA}} \]

if I care about SQUAD, maybe I use a high \( \lambda \)

- how to choose \( \lambda \)?

how do we know what intermediate task will result in the biggest downstream improvement?

- 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?