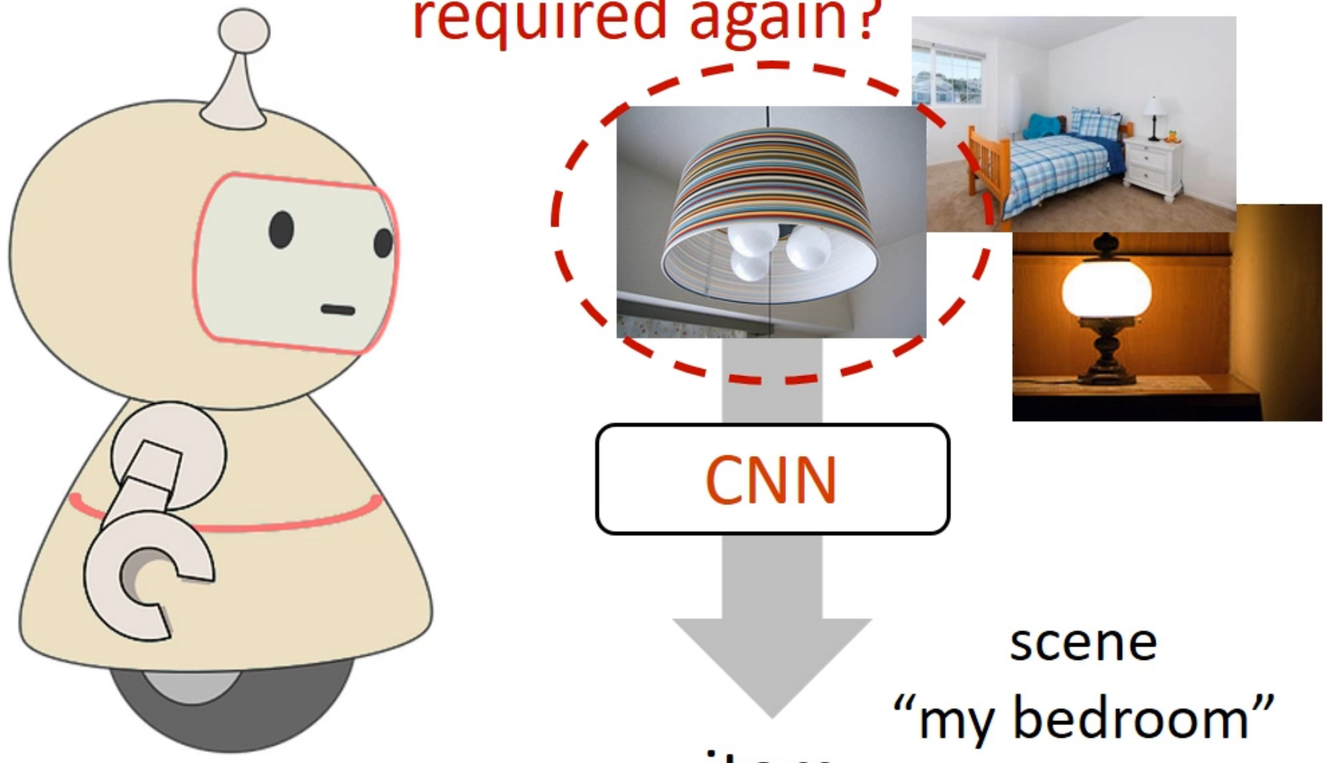
Learning Without Forgetting

Zhizhong Li, Derek Hoiem

University of Illinois, Urbana Champaign

Motivation



- Use as feature extractor?
- Fine-tuning?
- Joint training?

Image credit:
https://pixabay.com/en/alphabet-word-images-antenna-cartoon-1292931/

Goal

- Add new capabilities, keep existing capabilities
- Using *only data from the new task*.

CNN

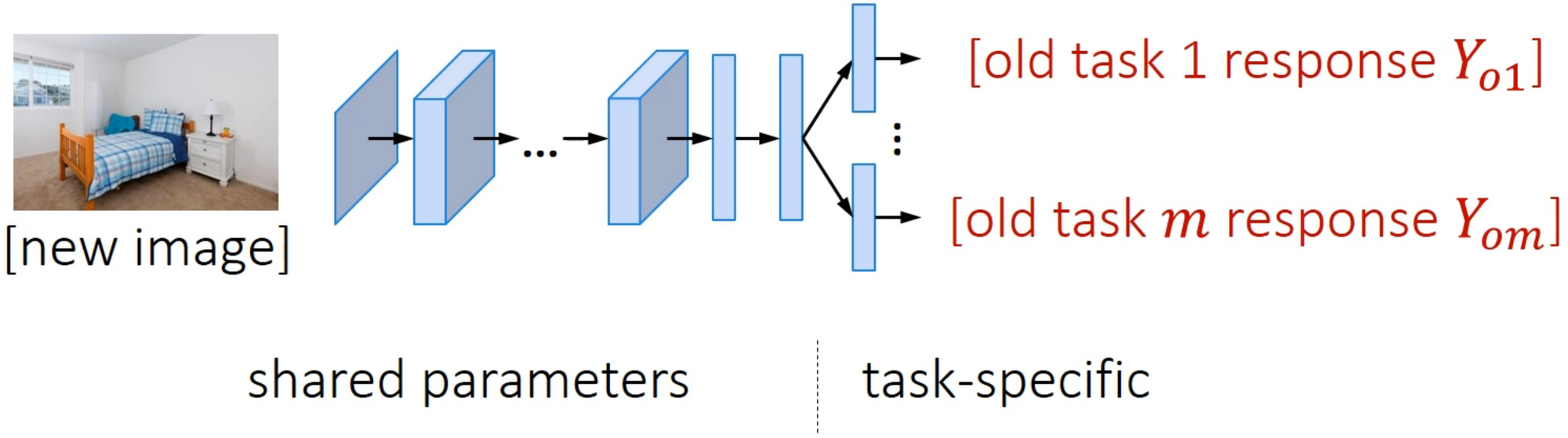
remember the old task

new task: bedroom

- ✓ Outperforms fine-tuning
- ✓ Outperforms feature extraction on new task
- ✓ Simple

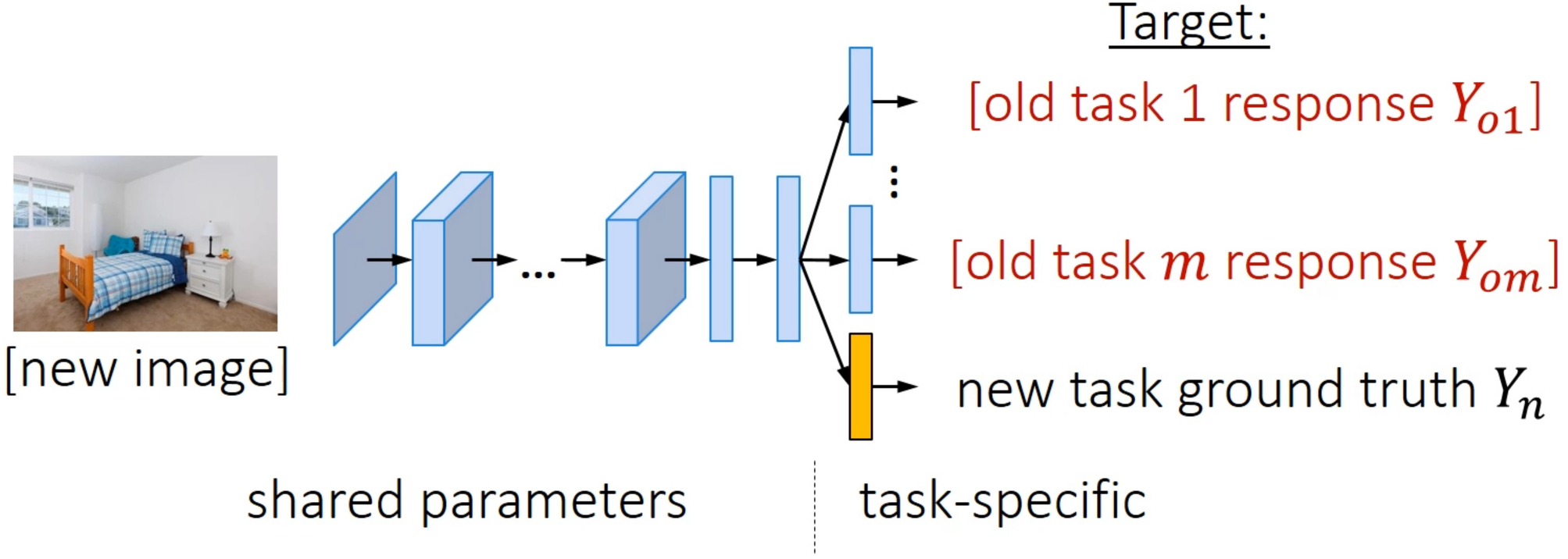
Method

1. Obtain old task responses



Method

2. Train on new images



$$\mathcal{L} = \sum_{i=1}^{m} \mathcal{L}_{old}(Y_{oi}, \hat{Y}_{oi}) + \mathcal{L}_{new}(Y_n, \hat{Y}_n) + \mathcal{R}(\theta)$$

Experiments

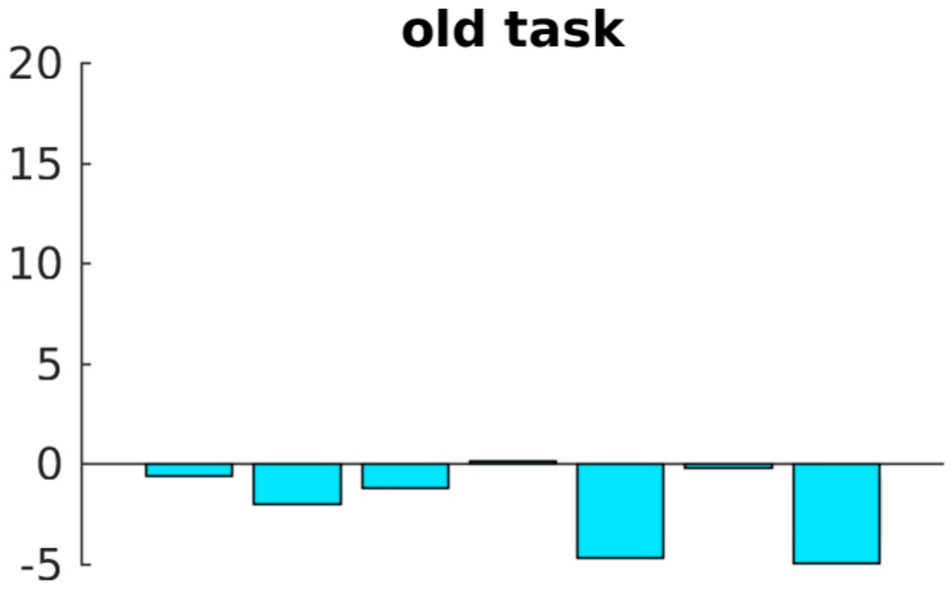
- AlexNet 1 old task + 1 new task

ILSVRC 2012 Places2	+	PASCAL VOC 2012 Caltech-UCSD Birds MIT indoor scenes MNIST

- Compared Methods:
  - Feature extraction (keep original network)
  - Fine-tuning (keep original last layer)
  - Joint training (requires old data)

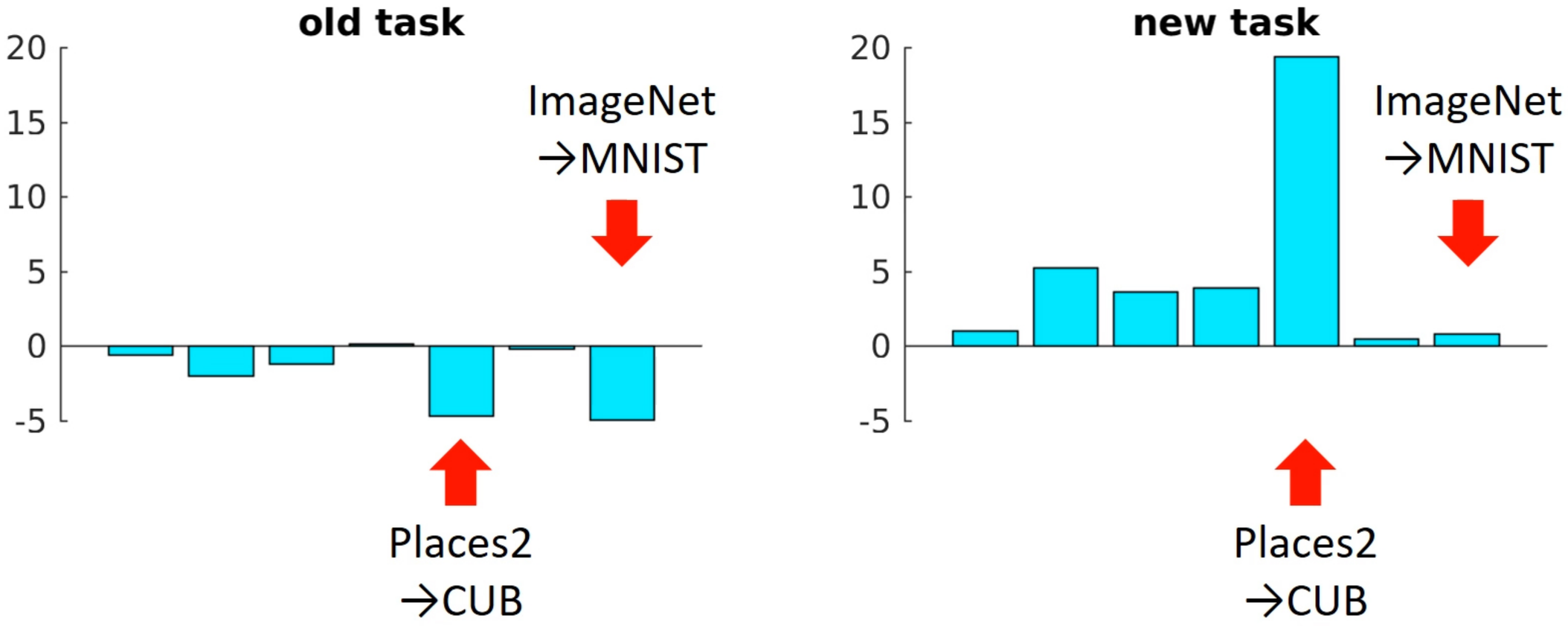
Results (vs. Feature Extraction)

- Shown: accuracy (ours) relative to the baseline's on seven task pairs



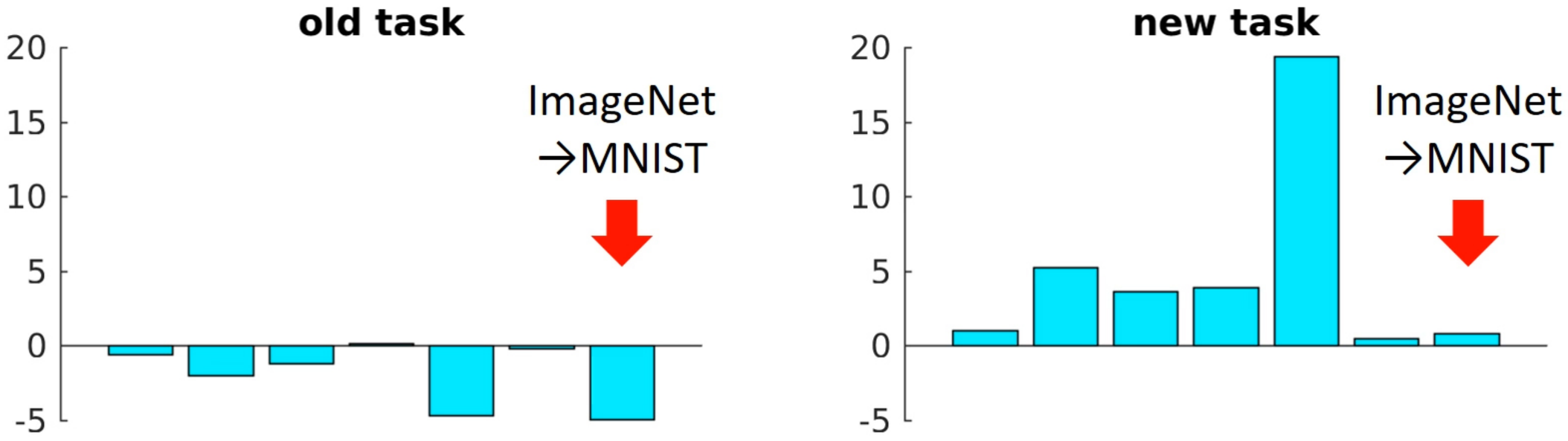
Results (vs. Feature Extraction)

- Shown: accuracy (ours) relative to the baseline's on seven task pairs



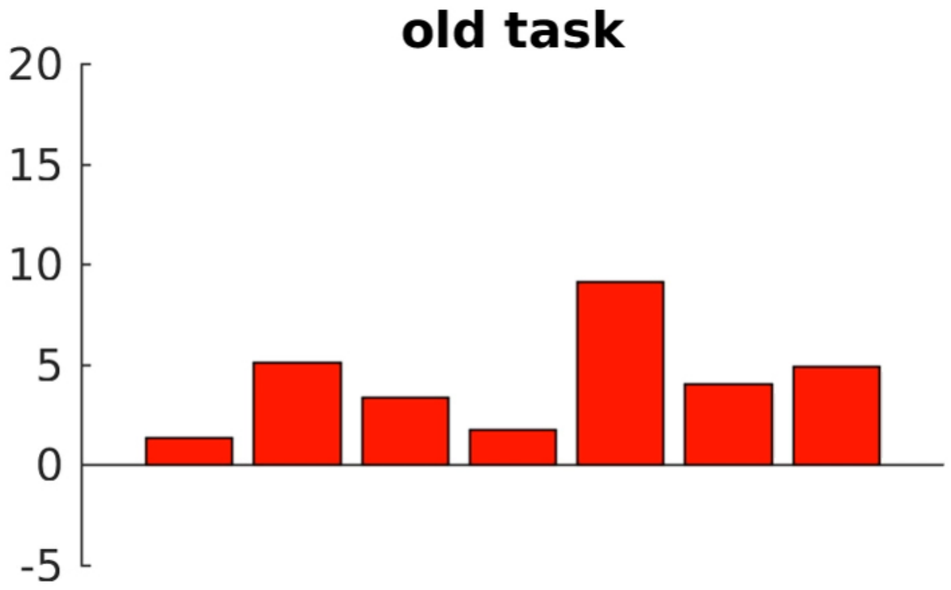
Results (vs. Feature Extraction)

- Shown: accuracy (ours) relative to the baseline's on seven task pairs



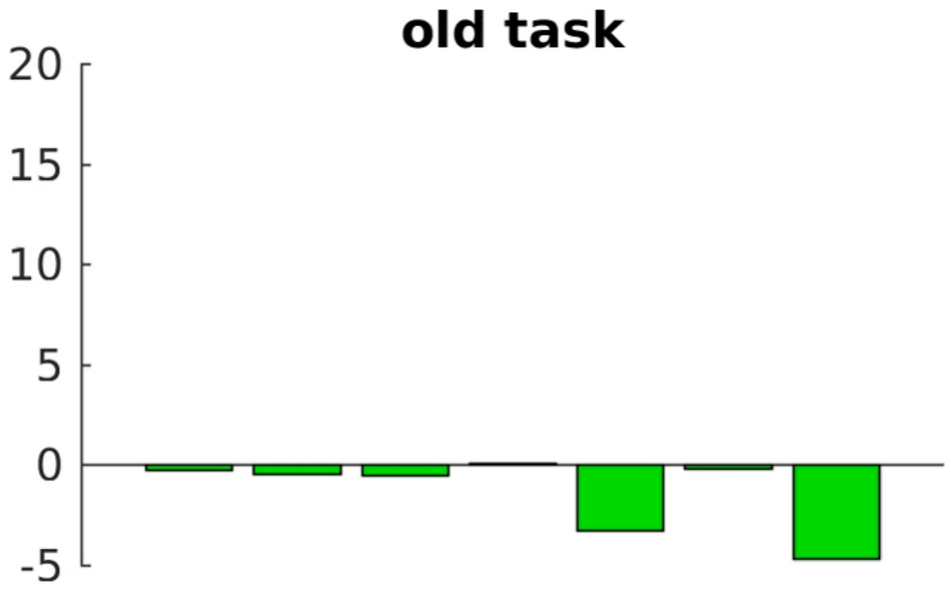
Results (vs. Fine-tuning)

- Old task: actively preserves performance
- New task: mimics joint training



Results (vs. Joint Training)

- Similar performance



Results (vs. Joint Training)

- Similar performance