Learning Without Forgetting

Zhizhong Li, Derek Hoiem
University of Illinois, Urbana Champaign
Motivation

Training data required again?

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

Goal

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

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

new task: bedroom

remember the old task
Method

1. Obtain old task responses

[new image]

shared parameters | task-specific

[old task 1 response $Y_{o1}$]

[old task $m$ response $Y_{om}$]
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

**old task**
- ImageNet → MNIST

**new task**
- ImageNet → MNIST
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