AdaShare: Learning What To Share For Efficient Deep Multi-Task Learning

Ximeng Sun\textsuperscript{1}, Rameswar Panda\textsuperscript{2}, Rogerio Feris\textsuperscript{2}, Kate Saenko\textsuperscript{1,2}

\textsuperscript{1} Boston University, \textsuperscript{2} MIT-IBM Watson AI Lab, IBM Research

\{sunxm, saenko\}@bu.edu, \{rpanda, rsferis\}@us.ibm.com

Project page: https://cs-people.bu.edu/sunxm/AdaShare/project.html
Multi-task Learning

Multi-task learning (MTL) is a subfield of machine learning in which multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks.

MTL improves generalization by leveraging the domain-specific information contained in the training signals of related tasks.
Previous Work – Two prevailing trends for MTL

**Hard parameter sharing**

- Task A
- Task B
- Task C

  Task-specific layers

  Shared layers

Examples: Deep Relationship Network, Fully-Adaptive Feature Sharing

**Soft parameter sharing**

- Task A
- Task B
- Task C

  Constrained layers

Examples: Cross Stitch, Sluice, NDDR

*Hand-crafted* tree structure

*Non-scalable* with the increasing number of tasks
AdaShare – Learn non-handcrafted and scalable sharing patterns

- **Each block:** shared or task-specific
- **Adaptive Computation:** Gumbel-Softmax Sampling
- **Loss:** Task-specific Loss, Sparsity Loss, Sharing Loss
- **Training Strategy:** policy learning stage and retraining stage
Experiments – Quantitative Results

Datasets: NYU v2 (2 or 3 tasks), CityScapes (2 tasks), Tiny-Taskonomy (5 tasks), DomainNet (6 tasks), Text-Recognition (10 tasks)

Table 4: Tiny-Taskonomy 5-Task Learning. \( T_1 \): Semantic Segmentation, \( T_2 \): Surface Normal Prediction, \( T_3 \): Depth Prediction, \( T_4 \): Keypoint Estimation, \( T_5 \): Edge Estimation.

| Models       | # Params ↓ | \( \Delta T_1 \) ↑ | \( \Delta T_2 \) ↑ | \( \Delta T_3 \) ↑ | \( \Delta T_4 \) ↑ | \( \Delta T_5 \) ↑ | \( \Delta T \) ↑ |
|--------------|------------|--------------------|--------------------|--------------------|--------------------|--------------------|------------------|
| Multi-Task   | -80.0      | -2.1               | -0.7               | -9.1               | +1.5               | +5.2               | -1.0             |
| Cross-Stitch | 0.0        | +2.6               | -3.3               | **0.0**            | -2.5               | -3.3               | -1.3             |
| Sluice       | 0.0        | -6.1               | -0.7               | -4.6               | **+2.5**           | +6.6               | -0.4             |
| NDDR-CNN     | +8.2       | **+6.3**           | -0.3               | -11.4              | +1.5               | +2.8               | -0.2             |
| MTAN         | -9.8       | -10.8              | -0.7               | -4.5               | **+2.0**           | +4.2               | -2.0             |
| DEN          | -77.6      | -28.2              | -3.0               | -22.7              | **+2.5**           | +4.2               | -9.4             |
| AdaShare     | -80.0      | +1.6               | **0.0**            | -13.6              | **+2.5**           | **+9.0**           | -0.1             |

**Single-Task Learning:** Seg: 0.575; SN: 0.707; Depth: 0.022; Keypoint: 0.197; Edge: 0.212
Experiments – Policy Visualization

**Observations:**
1. Not all blocks contribute to the task equally
2. More blocks shared only among a sub-group of tasks in ResNet’s conv3_x layers, where middle/high-level features (more task-specific) are starting to get captured
3. Similar tasks should have similar execution distribution to share knowledge
Thank you and welcome to our poster!