Knowledge Transfer in Vision Tasks with Incomplete Data

Zhizhong Li
University of Illinois Urbana Champaign

Chair: Derek Hoiem
Committee: Derek Hoiem Svetlana Lazebnik
Alexander Schwing Linjie Luo
Knowledge in humans

- It comes quite naturally
  - Classify
  - Attribute
  - Applying to new circumstances
  - Inference between attributes
  - Know if answer is uncertain
  - Understand its behavior
  - Know how to interact with it
  - More…
Knowledge in ML models

- Harder for ML algorithms
  - Classify
  - Attribute
  - Applying to new circumstances
  - Inference between attributes
  - Know if answer is uncertain
  - Understand its behavior
  - Know how to interact with it
  - More…

✔️ Easier for models to learn

Knowledge transfer!

Trickier to learn
Help generalization
Transfer learning

source \(\rightarrow\) input + GT \(\rightarrow\) non-i.i.d. (different task or different distribution) \(\rightarrow\) target

In practice: what if we cannot obtain all aspects of necessary data?

* or multi-task learning
Transfer learning in practice

① old data missing
Learning without Forgetting

② label missing
Task-assisted Domain Adaptation

③ domain unknown
Improving Confidence Estimates for Unfamiliar Examples
Learning without Forgetting

Zhizhong Li, Derek Hoiem

In ECCV 2016 (spotlight); PAMI, 2018
Motivation

• Task: extending capability (transfer to new task)
  * closer to multi-task learning

• Constraint:
  • Cannot access original dataset
  • Common in industry settings

• Challenge:
  • Catastrophic forgetting
  • ... but maintain old task performance
Baselines

- Fine-tuning?
- Feature extraction?
- Joint training?

Training data required again?

scene “my bedroom”
Related work

• Fine-tuning, feature extracting, Multi-task learning

• Closely related:
  • Less Forgetting Learning [1]
  • A-LTM [2]

• Other continual learning methods:
  • iCaRL [3]
  • EWC [4], SI [5]

Method

1. Obtain old task responses

“90% bed, 10% lamp” etc.

Serve as reminder of old task

[new image] [old task 1 response $Y_o$]
Method

2. Train on new images

- Fine-tuning: no old task loss
- Feature extraction: freeze old layers
- Joint training (multi-task): use old task image + GT (oracle)

Target:
- “90% bed, 10% lamp” etc.
- [old task 1 response $Y_o$]
- new task ground truth $Y_n$

“bedroom”
Experiments

• AlexNet

1 old task + 1 new task

ILSVRC 2012 Places365

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

(8 combinations)

• Compared Methods:
  • Baselines
  • Less-forgetting Learning
  • Joint training (oracle)
Results: LwF vs. Feature Extraction

- Shown: accuracy (ours) relative to the baseline’s on eight task pairs

Old tasks:

- Dataset pairs
  - Places365 → CUB
  - ImageNet → MNIST

New tasks:

- MNIST → MNIST

vs. Feature Extraction
Results: LwF vs. Fine-tuning

- Old task: actively preserves performance
- New task: mimics joint training
Results: LwF vs. oracle

- Joint training
- Similar performance

Old tasks:

New tasks:

vs. Feature Extraction

vs. Fine-tuning

vs. Joint training
Results

- Old-new trade-off (accuracy / VOC mAP)

![Graph showing trade-off between old and new task performance for Places365→VOC](image)
Limitation

• Worse when old/new images too different
  • How to add as new classes?
    • (a.k.a. class-incremental learning)
Follow-up: Dreaming to Distill

In collaboration with NVIDIA
Hongxu Yin, Pavlo Molchanov, Zhizhong Li, Jose M. Alvarez, Arun Mallya, Derek Hoiem, Niraj K. Jha, Jan Kautz
Accepted in CVPR 2020 as an oral presentation
A better old data proxy

• Network visualization methods
  • e.g. Deep dream, Tensorflow lucid

• Generates images given only class ID or neuron ID
• No data retention required!

• Too different from original data?

• DeepInversion: use pretrained BatchNorm statistics
Image generation

- **DeepDream**
  \[
  \min_{\hat{x}} \mathcal{L}(\hat{x}, y) + \mathcal{R}(\hat{x})
  \]

- **DeepInversion**
  \[
  \min_{\hat{x}} \mathcal{L}(\hat{x}, y) + \mathcal{R}(\hat{x}) + R_{\text{feature}}(\hat{x})
  \]

  \[
  R_{\text{feature}}(\hat{x}) = \sum_{l} || \hat{x}'s \text{ mean/var} - \text{BatchNorm mean/var} ||^2
  \]

  Makes feature distribution similar to training

![Images of DeepDream and DeepInversion results for CIFAR10 dataset]
Quantitative results

- ImageNet $\rightarrow$ CUB, ImageNet $\rightarrow$ Flowers
  - Allow confusion between old/new classes
    - (i.e. class-incremental instead of task-incremental)
  - Report accuracy on each dataset

![Graphs showing accuracy on CUB and Flowers datasets]
Take-away

• Data for existing knowledge can be missing

• Proxy for old task data
  • New task data / DeepInversion data
  • Original network responses

• Outperforms fine-tuning, etc.
Anchor Tasks for Domain Adaptation

Zhizhong Li, Linjie Luo, Sergey Tulyakov, Qieyun Dai, Derek Hoiem

In collaboration with Snap Inc.
Spatial ground truth problems

• Hard to obtain
  Time-consuming

Cannot manually annotate

Estimations: noisy

Estimations: unfaithful

• Use domain adaptation (and synth data) to help!
Unsupervised domain adaptation

• Make distribution between domains match

\[
\text{min} \| \mathcal{D} - \mathcal{G} \|
\]

• Feature space [1,2]
• Input space (Refiner [3], CyCADA [4])
• Output space [5,6]

• Assume distributions *should* be made identical

Semantic and spatial info can help matching

Task-Assisted Domain Adaptation (TADA)

• How about an auxiliary supervised task?

• Pick “anchor task”
  • Easier to obtain
  • Guidance info (e.g. semantic / spatial)
  • On both domains

• No explicit task relationship needed!
Method

- Baselines
  - Single task
  - Multi-task (not shared)

- Multi-task (shared anchor)

(b) Multitask learning (source only with anchor tasks per domain)
Method

• Relationship modeling

(e) FREEZE (ours)
Experiments

• Two datasets

<table>
<thead>
<tr>
<th>Normal</th>
<th>Keypoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main task</td>
<td>Anchor task</td>
</tr>
</tbody>
</table>

- **Source domain**
  - SfSNet [1]
  - SUNCG
- **Target domain**
  - Face-Warehouse
  - NYUdv2

- **No ground truth**

- **State-of-the-art keypoint estimator [2]**

- **Labels**

• Compared methods
  - Baseline, oracle
  - SfSNet [1]

[2] Adrian Bulat and Georgios Tzimiropoulos. “How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)”
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>src main</th>
<th>src anch</th>
<th>tgt main</th>
<th>tgt anch</th>
<th>Faces $&lt; 11.25^\circ$</th>
<th>Faces $&lt; 30^\circ$</th>
<th>RMSE</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>STL</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>0.424</td>
<td>0.929</td>
<td>17.8</td>
<td>14.8</td>
<td>12.8</td>
</tr>
<tr>
<td>DA [1]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>0.456</td>
<td>0.937</td>
<td>17.2</td>
<td>14.2</td>
<td>12.1</td>
</tr>
<tr>
<td>MTL</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.409</td>
<td>0.935</td>
<td>17.7</td>
<td>14.9</td>
<td>13.1</td>
</tr>
<tr>
<td>MTL</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.162</td>
<td>0.791</td>
<td>24.3</td>
<td>21.8</td>
<td>20.4</td>
</tr>
<tr>
<td>MTL</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.492</td>
<td><strong>0.953</strong></td>
<td>16.0</td>
<td>13.3</td>
<td>11.4</td>
</tr>
<tr>
<td>FREEZE (ours)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>0.519</strong></td>
<td><strong>0.954</strong></td>
<td><strong>15.8</strong></td>
<td><strong>12.9</strong></td>
<td><strong>10.9</strong></td>
</tr>
</tbody>
</table>

Qualitative results

<table>
<thead>
<tr>
<th>Input</th>
<th>Ground Truth</th>
<th>STL</th>
<th>DA</th>
<th>MTL (one domain)</th>
<th>MTL (both domains)</th>
<th>Freeze</th>
<th>SfSNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image" /></td>
<td><img src="image2.jpg" alt="Image" /></td>
<td><img src="image3.jpg" alt="Image" /></td>
<td><img src="image4.jpg" alt="Image" /></td>
<td><img src="image5.jpg" alt="Image" /></td>
<td><img src="image6.jpg" alt="Image" /></td>
<td><img src="image7.jpg" alt="Image" /></td>
<td><img src="image8.jpg" alt="Image" /></td>
</tr>
<tr>
<td><img src="image9.jpg" alt="Image" /></td>
<td><img src="image10.jpg" alt="Image" /></td>
<td><img src="image11.jpg" alt="Image" /></td>
<td><img src="image12.jpg" alt="Image" /></td>
<td><img src="image13.jpg" alt="Image" /></td>
<td><img src="image14.jpg" alt="Image" /></td>
<td><img src="image15.jpg" alt="Image" /></td>
<td><img src="image16.jpg" alt="Image" /></td>
</tr>
<tr>
<td><img src="image17.jpg" alt="Image" /></td>
<td><img src="image18.jpg" alt="Image" /></td>
<td><img src="image19.jpg" alt="Image" /></td>
<td><img src="image20.jpg" alt="Image" /></td>
<td><img src="image21.jpg" alt="Image" /></td>
<td><img src="image22.jpg" alt="Image" /></td>
<td><img src="image23.jpg" alt="Image" /></td>
<td><img src="image24.jpg" alt="Image" /></td>
</tr>
<tr>
<td><img src="image25.jpg" alt="Image" /></td>
<td><img src="image26.jpg" alt="Image" /></td>
<td><img src="image27.jpg" alt="Image" /></td>
<td><img src="image28.jpg" alt="Image" /></td>
<td><img src="image29.jpg" alt="Image" /></td>
<td><img src="image30.jpg" alt="Image" /></td>
<td><img src="image31.jpg" alt="Image" /></td>
<td><img src="image32.jpg" alt="Image" /></td>
</tr>
</tbody>
</table>
Take-away

• Matching distributions are not enough for unsupervised domain adaptation

• Easy-to-obtain labels for another task can help

• Modeling task relationship can help
Study: Improving Confidence Estimates for Unfamiliar Examples

Zhizhong Li, Derek Hoiem

Accepted in CVPR 2020 as an oral presentation
Imagine in a face photo gender classification application...

User input

P(female) = 99.9%
P(male) = 99.3%

97% accuracy

(biased) dataset collection

CNN model

99%+ confidence

= <1% error rate?

• 0.5% w/ familiar
• 6.0% w/ unfamiliar
  • 12x errors!

Problems:
• Test data different in unexpected ways
• Underrepresented data get confidently misclassified
Prior work

- Domain adaptation, Domain generalization
  - Needs knowing variations of future domains

I’m adapted to in-the-wild images!

I’ve got you some new stuff

Adapted models
Prior work

Novelty detection

These are not faces I have seen.

Yeah, but aren’t you going to classify them?
Prior work

Modeling epistemic uncertainty
• (i.e. uncertainty due to lack of knowledge)

I am not familiar with these, so I make predictions with adjusted confidence

Desired model
Goal

• Comparative study
  • Which prior work has the most well-behaved confidence on unseen data?

• How to evaluate?

- CNN model
  \[ P(\text{female}) = 99.9\% \]
  \[ P(\text{male}) = 99.3\% \]

- Novelty detector
  These are not faces I have seen.

- Uncertainty model
  I am not familiar with these, so I make predictions with adjusted confidence.
Comparative study

• List of compared works
  • Regularly-trained model (baseline)
  • Modeling uncertainty [2]
  • Calibration with temperature-scaling [1]
  • Ensemble
    • Calibrated ensemble
    • Distilling [3]
    • Distilling [3] (modified)
  • Novelty detection [4] (modified)

[2] Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?."
Compared methods (cont’d)

• Calibration
  • Temperature scaling [1]
    \[ p(x) = \text{softmax}(f(x)) \]
    \[ p'(x) = \text{softmax}(f(x)/T) \]
  • Use a higher temperature in the prediction
  • Calibrate the temperature in a validation set

Experimental setup

- Evaluate confidence: Negative log-likelihood (NLL)

- Get underrepresented data: split by subcategories

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Familiar</th>
<th>Unfamiliar</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFW+ (face gender)</td>
<td>Ages 18-59</td>
<td>Ages 0-17, 60+</td>
</tr>
<tr>
<td>ImageNet superclass*</td>
<td>Some species</td>
<td>Other species</td>
</tr>
<tr>
<td>Pets (cat v. dog)</td>
<td>Some breeds</td>
<td>Other breeds</td>
</tr>
<tr>
<td>VOC-COCO, 20 classes</td>
<td>PASCAL VOC, whole dataset</td>
<td>MSCOCO, ignoring non-VOC classes</td>
</tr>
</tbody>
</table>

* mammals vs. herptiles vs. birds vs. fishes
Results: Negative log-likelihood

- Animal classification (ImageNet subset)
  - Smoothing effect: trade-off familiar / unfamiliar
Results: errors among 99% confident
Take-away

Issue highlight: data underrepresented in training can get confidently misclassified

Best-performing methods

- Calibrated ensembles
  - -32% unfamiliar NLL
- Calibration (T-scaling)
  - -23% unfamiliar NLL

Experimental method

- Split familiar / unfamiliar by subcategories
Story so far

① source $\rightarrow$ input + GT
input + GT
old data missing
Regenerate
(images & labels)

② source $\rightarrow$ input + GT
input + GT
label missing
Guide
(w/ anchor task)

③ source $\rightarrow$ input + GT
input + GT
domain unknown
Calibrate
(w/ val set)
Future work

Open-ended question:
• How to improve knowledge transfer?
• How to circumvent data constraints in industry settings?

What IS knowledge?

① $(x_i, y_i)$? $\hat{y} = f(x)$? $\mathbb{P}(x)$?

② Interaction between two tasks’ $y_1, y_2$?

③ Generalization where $\mathbb{P}(x) \approx 0$?

More?
Future work

Leverage other knowledge in humans

- How different things behave
- Why things behave this way
- Does this new thing behave the same way
- How does this knowledge affect my decisions
- Etc.

Can we extract these from models?
Can these be represented without using data?
Can we use these to improve knowledge transfer?
Acknowledgements

• My advisor

• Collaborators

Linjie Luo  Sergey Tulyakov  Qieyun Dai  Arun Mallya  Hongxu Yin  Pavlo Molchanov
Thanks everyone!
Questions?
## Network pruning results

- Resnet 50; Compared to methods that use real data

<table>
<thead>
<tr>
<th>Image Source</th>
<th>Top-1 acc. (%)</th>
<th>-50% filters</th>
<th>-20% filters</th>
<th>-71% FLOPs</th>
<th>-37% FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>No finetune</td>
<td>1.9</td>
<td>16.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial ImageNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1M images / 0 label</td>
<td>69.8</td>
<td>74.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proxy datasets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS COCO</td>
<td>66.0</td>
<td>73.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PASCAL VOC</td>
<td>54.4</td>
<td>70.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generator, BigGAN</td>
<td>63.0</td>
<td>73.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise (Ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepInversion (DI)</td>
<td>55.9</td>
<td>72.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive DeepInversion (ADI)</td>
<td>60.7</td>
<td>73.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Knowledge transfer results

- Resnet 50 v1.5; from scratch

<table>
<thead>
<tr>
<th>Image source</th>
<th>Data</th>
<th>Top-1 acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td>1.3M, Real</td>
<td>77.26%</td>
</tr>
<tr>
<td>DeepInversion</td>
<td>140K, Dream</td>
<td>73.8%</td>
</tr>
</tbody>
</table>
Existing work with TADA structure

- Focus on known, explicit main-auxiliary label relationships [1,2,3,4]

Results

• + Domain adaptation

<table>
<thead>
<tr>
<th>src</th>
<th>src</th>
<th>tgt</th>
<th>tgt</th>
<th>Faces</th>
<th>SfSsyn→FaceWH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>main</td>
<td>anch</td>
<td>main</td>
<td>anch</td>
<td>&lt; 11.25°</td>
</tr>
<tr>
<td>STL</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.424</td>
</tr>
<tr>
<td>MTL-src</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.409</td>
</tr>
<tr>
<td>MTL-SmTa</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.162</td>
</tr>
<tr>
<td>MTL-a</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.492</td>
</tr>
<tr>
<td>FREEZE (ours)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td><strong>0.519</strong></td>
</tr>
</tbody>
</table>

with unsupervised domain adaptation: [1]

| DA  | ✓    |      |         |         | 0.456    | 0.937 | 17.2   | 14.2  | **12.1**|
| MTL-src | ✓  | ✓    |      |         | 0.402    | 0.932 | 18.0   | 15.1  | 13.3   |
| MTL-SmTa | ✓ | ✓    | ✓    |         | 0.216    | 0.854 | 22.0   | 19.5  | 18.1   |
| MTL-a | ✓    | ✓    | ✓    |         | 0.455    | 0.946 | **16.7**| **13.9**| **12.1**|
| FREEZE (ours) | ✓ | ✓    | ✓    |         | **0.455**| 0.935 | 17.2   | 14.2  | **12.1**|

## Qualitative results

<table>
<thead>
<tr>
<th>Input</th>
<th>Ground Truth</th>
<th>STL</th>
<th>DA</th>
<th>MTL-SmTa</th>
<th>MTL-a</th>
<th>Freeze (ours)</th>
<th>Freeze+DA</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="image4.png" alt="Image 4" /></td>
<td><img src="image5.png" alt="Image 5" /></td>
<td><img src="image6.png" alt="Image 6" /></td>
<td><img src="image7.png" alt="Image 7" /></td>
<td><img src="image8.png" alt="Image 8" /></td>
<td><img src="image9.png" alt="Image 9" /></td>
</tr>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="image4.png" alt="Image 4" /></td>
<td><img src="image5.png" alt="Image 5" /></td>
<td><img src="image6.png" alt="Image 6" /></td>
<td><img src="image7.png" alt="Image 7" /></td>
<td><img src="image8.png" alt="Image 8" /></td>
<td><img src="image9.png" alt="Image 9" /></td>
</tr>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="image4.png" alt="Image 4" /></td>
<td><img src="image5.png" alt="Image 5" /></td>
<td><img src="image6.png" alt="Image 6" /></td>
<td><img src="image7.png" alt="Image 7" /></td>
<td><img src="image8.png" alt="Image 8" /></td>
<td><img src="image9.png" alt="Image 9" /></td>
</tr>
</tbody>
</table>

**Better ceiling / wall**

*(STL works pretty well already)*
Compared methods (2)

• Ensemble

• “Distilling” [1] an ensemble
  • Train single model on soft labels to mimic the ensemble

• G-distill (modified)
  • Use an additional unsupervised dataset
    e.g. Internet pictures

Compared methods (3)

- Novelty detection [1], modified (cannot use directly)
  1. Get original confidence
  2. Run novelty detection procedure
  3. Higher outlier score ▼ more reduction in confidence

"NCR" (Novel Confidence Reduction)

Results: Negative log-likelihood

- Face gender, Pets cat vs. dogs, VOC-COCO

VOC too similar to COCO
Results: errors among 99% confident